2017 SAT Standard Setting for Illinois
Final Report

Deanna L. Morgan
Lei Wan
Pamela Kaliski
Priyank Patel
Jonathan Beard
Janeen McCullough
Caiyan Zhang
YanYan Fu

Table of Contents

List of Tables

List of Figures

2017 SAT Standard Setting for Illinois Final Report

Deanna L. Morgan
Lei Wan
Pamela Kaliski
Priyank Patel
Jonathan Beard
Janeen McCullough
Caiyan Zhang
YanYan Fu

Executive Summary

On September 27-28, 2017, College Board Psychometrics facilitated a Standard Setting for the Math and Evidence-Based Reading and Writing (ERW) Sections of the SAT for the state of Illinois. The purpose was to obtain three recommended cut scores for each of these exam sections that can be used for accountability purposes. A panel of subject matter experts was convened for each subject area with 24 panelists on the ERW panel and 25 panelists on the Math panel. Panelists received an overview of the exam, experienced the test by actually taking the exam under timed conditions, reviewed the Performance Level Descriptors (PLDs) that were provided, and received training on the Modified Angoff method used to make judgments. Following an opportunity to practice making ratings on a small sample of items, panelists completed a Ready to Proceed form indicating they understood the task and were ready to make operational ratings. All panelists in both groups indicated they understood and were ready to proceed.

Each panel used the PLDs to guide them in making ratings for each test item at each of the three performance levels using a Google Docs spreadsheet to record their ratings. After their ratings were completed, panelists completed an evaluation form as a check on their understanding and comfort with the task. Panelists' ratings indicated they agreed or strongly agreed with most statements and considered most materials or information provided to be influential in their decisions. The next morning panelists received feedback on the ratings that had been provided by other panelists sitting at their tables and on the operational difficulty of the items. After a discussion of these results and comparison of their ratings with others at their table, panelists were given an opportunity to adjust their ratings based on any insight they may have gained during the feedback and discussion time.

Following the Round 2 ratings, panelists received additional feedback on the ratings at their table and the ratings for the room as a whole. Discussions were held at the table level and then at the large group level before the final piece of feedback was shared. After the large group discussion, panelists were provided with the impact data at their table showing the percentage of examinees expected to be placed into each performance level if the results from their table were used to report scores. Simultaneously, the same type of

impact data based on the whole room was shared on the screen at the front of the room. Panelists were asked to discuss the results and any concerns they had at their tables.

Following this discussion, a third and final round of adjustments could be made. These final adjustments result in the recommended cut scores that would be brought to the policy meeting on October 4, 2017. The final recommended cut scores for the respective content area were shared with each panel, and the panel had an opportunity for discussion and to share any reactions with Illinois State Board of Education (ISBE). The final activity for panelists was the completion of the final evaluation form as part of the procedural validity of the process and to give ISBE feedback on panelists' impressions of the process and results for consideration at the policy meeting.

Results from the final evaluations across ERW and Math indicated that panelists agreed or strongly agreed that they understood the purpose of the study (Q1), and the instructions, explanations (Q2), and training provided (Q3) were clear and sufficient for them to complete the rating task. The mean ratings for these questions are almost always above 3.5 (the scale is 1-4; 1 indicates strongly disagree and 4 indicates strongly agree). Also, panelists expressed that they understood the concept of borderline examinees (Q5), and agreed it was beneficial to review feedback and have table or room discussion between rounds (Q7, Q8, Q9, Q10). They liked the opportunity to make more than one round of ratings (Q11). The mean ratings for these questions are mostly above 3.5. Furthermore, the survey results showed that panelists were engaged throughout the process (Q12) and they were very comfortable to speak up and share their opinions during the discussion (Q13). The mean ratings for these questions are close to 4 in both subjects. Additionally, the results indicated that across the panels, panelists found the standard setting materials and activities to be influential and useful, and in general felt enough time was provided for all the activities, though some panelists did indicate they would have liked more time taking the ERW test.

In Math, many panelists pointed out that the PLDs seemed to be too rigorous ---they might be appropriate for strong or average students in a performance level, but not an accurate description of borderline students. Therefore, the mean ratings for PLD-related questions (Q4 and Q6) are relatively low, i.e. around 2.6, and this might have influenced the panelists' confidence for the final Math cut score recommendations (Q14). In contrast, in the ERW meeting, panelists' ratings for the PLDs-related questions are high, and their confidence for the final ERW cut score recommendations is high as well.

Overall, panelists across both of the panels had very positive feedback about the meeting procedures, materials, and training, and they were very comfortable and engaged in the process. In the Math panel, PLDs seemed to be an issue and its impact may need to be considered in finalizing the cut scores.

The recommended cut scores in ERW from Round 3 were 430 for Approaching Standards, 540 for Meets Standards, and 640 for Exceeds Standards which would result in the following distributions of students : 22.40% Partially Meets Standards, 37.31% Approaching Standards, 26.99% Meets Standards, and 13.31% Exceeds Standards. In Math,

the cut scores were 450 for Approaching Standards, 540 for Meets Standards, and 690 for Exceeds Standards which would result in the following distributions of students: 30.40% Partially Meets Standards, 32.75% Approaching Standards, 30.70% Meets Standards, and 6.15% Exceeds Standards. Standard Error of Judgement values were very small (less than one scale score point) indicating a high degree of consensus within each panel.

On October 4th, a policy meeting was held to review the cut score recommendations from Round 3. At this meeting, the Exceeds cut score for Math was lowered to 670 (mean rating) from 690 (median rating). As a result, 7.8% of students placed in the Exceeds Standards category, and 28.5% placed in the Meets Standards category in Math. Other than these changes, all the other cut scores and student distribution data remained unchanged. The State Board met on October 18th and voted to accept the cut scores recommended by the policy meeting.

Final Report on the 2017 SAT Standard Setting for Illinois

In September 2017, College Board Psychometric and Assessment Design and Development staff conducted panel-based standard setting meetings for the new SAT. The meetings were held concurrently on September 27th – 28th, 2017 in Springfield, IL at the Wyndham Springfield City Center. The purpose of the standard setting meeting was to produce recommended cut scores on the SAT Math section and the SAT Evidence-Based Reading and Writing section (ERW) for classifying students into four performance levels to be used by Illinois for accountability purposes. This report summarizes the procedures used to collect recommended cut scores from the standard setting panelists, along with the results from the meetings. First, the instrument and participants are described. Then, procedures used during the standard setting meeting are presented, followed by a description of the results.

The New SAT

The new SAT consists of 3 sections: Math, ERW, and an optional essay. The essay was not part of the process being described in this report. The Math section has two parts: a No-Calculator part with 20 items where examinees are allotted 25 minutes and a Calculator part with 38 items and 55 minutes allotted. Both parts contain 4-option multiple choice (MC) items and student produced response (SPR) items. The ERW section also has two parts: a Reading part with 52 MC items where examinees are allotted 65 minutes, and a Writing and Language part with 44 MC items and 35 minutes. The Math and ERW sections are each on the 200-800 scale score metric. The new SAT is rights-scored meaning there is not penalty or correction for incorrect answers or guessing.

Subject Matter Experts

Two standard setting panels of subject matter experts (SMEs) were convened, one for Math and one for ERW. The Illinois State Board of Education was responsible for recruiting participants in each content area. The primary requirements for participation were teaching experience of at least 3 years and content expertise in the subject area. In addition, guidance was provided in terms of other characteristics that should be taken into consideration to obtain a panel with as much diversity as possible, to bring multiple perspectives to the meeting, and to aid in the generalizability of the results. The panels convened for Math and ERW contained 25 and 24 SMEs, respectively. The panels for ERW and Math were broken out in five tables, with five panelists at each table (one ERW table had four). A table lead was appointed for each table by the state. Table leads received a brief training on their responsibilities to take notes and keep the table discussion on track.

At the beginning of the meetings, following a brief introduction, the SMEs were asked to complete a biographical data form for use in summarizing panelist characteristics as evidence of procedural validity (Kane, 2001; Pitoniak and Morgan, 2012, 2017). These self-report biographical data are summarized in Tables 1 and 2. In this report, the term SMEs is used interchangeably with *panelists*.

Standard Setting Meeting Procedures

In this section, the procedures used to collect standard setting ratings from the SMEs are described. The sequence of activities in this report matches that used in the September standard setting meeting. The agenda used to guide the meeting is presented in Appendix A. The agenda for only one meeting is provided but the activities were the same for both Math and ERW.

Introduction to the Standard Setting Meeting

The standard setting meeting began with a large group session led by Dr. Deanna Morgan, Senior Director of Psychometrics at the College Board, to welcome panelists and provide a short introduction to the work that would occur during the meeting. Representatives from the Illinois State Board of Education (ISBE) and College Board also made welcoming comments during this opening session. Following the opening session, panelists moved into the content specific breakout rooms.

At the start of the subject specific meeting, panelists completed a short biographical data form (see Appendix B), for the purpose of contributing to the documentation of the procedural validity (Kane, 2001; Hambleton, Pitoniak, & Copella, 2012; Pitoniak and Morgan, 2012, 2017) of the standard setting process. Findings are summarized in Tables 1 and 2. Additionally, panelists were required to sign a confidentiality form, though the test form was a released form it was important to keep other information presented and results confidential since they were not final or released publicly yet.

At the start of the meeting, the facilitators, Dr. Lei Wan, Psychometrician at the College Board, for Math and Dr. Pamela Kaliski, Psychometrician at the College Board, for ERW, provided an introduction to the concept of *cut scores*, which are values used to classify student exam performance into distinct categories. Three cut scores were used to assign examinees to one of four performance levels where Partially Meets Standard is low and Exceeds Standard is high (See Figure 1).

After a brief introduction of the purpose of the meeting, panelists were given an overview of the exam format by College Board Assessment Design and Development staff who remained in the room during the meeting to address any content related questions that were raised. Jim Patterson, Executive Director for English Language Arts, presented to

the ERW panel and Bill Trapp, Executive Director for Math and Science, presented to the Math panel.

Experiencing the Exam

In order to provide panelists with a frame of reference for considering student performance, the panelists took the relevant sections of the SAT in a timeframe that was reduced from that allowed operationally. Panelists did not have access to answer keys during the exam administration. Operationally, students are allotted approximately 100 minutes to complete the ERW section with 65 minutes devoted to the Reading items and 35 minutes devoted to the Writing and Language items; standard setting panelists took the ERW section in 70 minutes. Operationally, students are allotted approximately 80 minutes for the Math section with 25 minutes for the No Calculator items and 55 minutes for the Calculator items; standard setting panelists took the Math section in 60 minutes. This activity was designed to familiarize panelists with the exam questions, as well as, the rigor and time constraints experienced by students who take the exam.

Following completion of the exam, an answer key was distributed to panelists so they could grade themselves on the items. No record of SME performance on the exam was kept, and panelists were free to share their performance with the other members of the panel at their own discretion.

Review of items

Panelists then had an opportunity to review and discuss items that they found especially difficult or confusing with an emphasis on characteristics of specific items. Panelists were reminded that the purpose of this activity was to discuss their perceived difficulty of items in the context of the entire experience, rather than to critique the items or the test. Any comments of a critical nature or editorial type beyond the scope of the standard setting task were to be collected and shared with College Board Assessment Design and Development staff. In general, the group of standard setting panelists was positive about the exam.

Performance Level Descriptors

Following the exam experience and discussion, Assessment Design and Development staff reviewed the Performance Level Descriptors (PLDs) written in collaboration with the content experts in Illinois. PLDs describe the borderline knowledge, skills, and abilities that are required for a student to be placed into each performance level. Discussion of PLDs prior to assigning standard setting ratings helps to establish a common understanding across standard setting panelists of the meaning of the borderline of each

performance level in terms of what students at the borderline know and are able to do. In essence, the PLDs serve as anchors during the rating task. The PLDs used for the standard setting are provided in Appendix C.

Borderline Examinee

Borderline examinees are students whose knowledge, skills, and abilities represent the minimal level of competence required for placement in each performance level. This concept is illustrated in Figure 2. The PLDs provided to the panelists were written with the Borderline Examinee in mind. The concept of the Borderline Examinee was reviewed thoroughly with the standard setting panelists because understanding this concept is integral to the standard setting task.

Rating Task for Standard Setting

A variety of methods have been proposed for setting performance standards on educational assessments. Despite procedural similarity across many standard setting techniques (Hambleton, Pitoniak, & Copella, 2012), Cizek (2012) describes at least ten separate standard setting processes with a host of modifications that yield even more methods that can be used to collect ratings from panelists. In spite of the numerous methods, various modifications described as Angoff standard setting procedures remain among the most widely used (Angoff, 1971; Plake & Cizek, 2012). It should be noted that the Angoff methods derive from a brief description and footnote in the 2nd Edition of Educational Measurement and is typically not implemented as originally described, thus most of the methods are more accurately referenced as Modified Angoff methods. The Angoff method and its variations are criterion-referenced standard setting methods that require panelists to estimate the probability that a "minimally acceptable person" (i.e., a borderline examinee) will answer an item correctly. These probabilities are then summed to produce recommended cut scores. While it is typical of many state-wide standard setting meetings to use an item mapping procedure such as the Bookmark method to make cut score recommendations, the SAT uses classical test theory and therefore a method that aligned with that model was desirable. A Modified Angoff standard setting method (Plake & Cizek, 2012) was used to collect panelist ratings for this Standard Setting.

Training on the Modified Angoff Method

Panelists were trained to provide Modified Angoff ratings. The training was facilitated by the standard setting facilitators, and a variety of activities was used in order to train panelists, evaluate understanding, and provide feedback prior to the collection of actual standard setting ratings. Training materials were presented orally and visually, using PowerPoint displays and handouts.

First, the concept of a borderline examinee was reviewed. Using PLDs to represent borderline examinees in each performance level, panelists were asked to provide expected

probabilities for correctly answering each item. In order to ease the cognitive demand during rating, panelists were asked to imagine a group of 100 borderline students at the threshold of each performance level, and estimate the number who could correctly answer each item given the knowledge, skills, and abilities required for a correct response and what the PLDs indicate about the knowledge, skills, and abilities of the borderline examinee at that level. Students in borderline groups were described in terms of "cuts" that distinguish between performance levels. These groups were described as follows:

- **Examinees at Approaching Standard**
- **Examinees at Meets Standard**
- **Examinees at Exceeds Standard**

Panelists were restricted to ratings between 10 and 95 in intervals of 5 for MC items and between 5 and 95 in intervals of 5 for SPR items. They were not allowed to provide ratings below 10 on MC items in order to prevent a cut score that could allow for a student to be placed into a performance level above Partially Meets Standards by random guessing. The MC items on both ERW and Math sections have four responses, so in theory the chance of guessing is 25% for MC items. However, it was later decided to allow panelists ratings to go to 10 for all MC items, since inspection of the operational item P-values indicated extremely low performance (e.g. p-values close to 0.10) on some items, thus it was a realistic expectation that the borderline examinee at the lower cut scores may do even worse than chance. Similarly, although the SPR items require examinees to write a response rather than select from a few options, and chance should not be a factor in the performance of these items, the inspection of operational student performance helped the state to decide that the lowest probability for SPR items should be 5.

On the high end, panelists were not allowed to provide ratings greater than 95 in recognition that perfect performance is not common, nor a reasonable expectation of the borderline examinees. Additionally, this helped to control for examinees being required to earn a perfect score to be placed into the highest score category.

Because three cut scores were needed to assign students to the four performance levels, panelists provided ratings for the three borderline groups simultaneously on each item. Specifically, panelists examined each item, and provided ratings for each of the three borderline groups using the following directions (e.g. MC items) to guide the judgmental process:

> *For each item indicate the number of borderline examinees out of 100 at each cut score that would answer the item correctly. Use only numbers between **10** and **95** in **5 point** intervals. The following list provides the possible values that can be used:*
>
> **10 15 20 25 30 35 40 45 50**
> **55 60 65 70 75 80 85 90 95**

Emphasis was placed on providing estimates of what students in each group *would* do on an exam, rather than using the ratings to describe their opinions of what students *should* do.

Next, instructions were provided for recording ratings. In order to facilitate data analyses and the provision of feedback between rounds of ratings, panelists entered ratings on individual provided laptops using customized Google Docs spreadsheets. Panelists were also asked to also complete a paper rating form for documentation, and to serve as backup in case of technology failure. An example of a Google Docs Excel file format is given in Appendix D. The Excel worksheet included a row for each item, with three columns used to enter ratings for each borderline examinee groups. Cells for ratings were restricted to the appropriate values for each item (e.g. 10-95 in intervals of 5 for MC). Panelists could either enter ratings manually, or select a value from a pull-down menu. Panelists were encouraged to use the entire range of values for the items. However, the worksheet was programmed to warn panelists of chance ratings on MC items (any value equal to or less than 25) using a yellow-colored cell. Rating cells were also programmed to prevent decreasing ratings across categories. In other words, ratings were required to either increase or remain the same as borderline examinee groups increased. As a result, panelists were prevented from providing cut scores that allowed higher achievement levels to be associated with lower performance.

Instructions for completing the rating form were given and the SMEs were asked to practice the method using four items from the SAT exam they experienced earlier. The four items were selected to represent different types of items (e.g., MC or SPR) and different content or testing conditions (e.g., Calculator and No Calculator, Reading and Writing). When the training round of ratings was complete, selected panelists were asked to volunteer their ratings. Then a brief discussion was held to discuss any discrepancies in ratings with emphasis in relating the discussion to the PLDs. Panelists were reminded that the standard setting panel was designed to represent a variety of perspectives that were needed to inform decisions about cut scores and that consensus was not a goal.

Following this discussion, panelists were asked to complete the Training Evaluation Form (Appendix E) which required them to indicate whether they were ready to proceed to the operational task. Ratings on the evaluation form were reviewed and all SMEs indicated that they understood the task and were ready to proceed. After receiving confirmation that the panelists were ready to proceed, the operational standard setting task began.

Ratings and Discussion

Next, three rounds of ratings were collected. Discussion was facilitated and feedback was provided between rounds. A description of the procedures used to collect ratings is provided below, followed by a summary of the results from the three rounds of ratings. Results from panelist evaluations of the procedures are provided in the next section.

Round 1

Round 1 ratings were provided for the entire set of items in the same fashion as was used to assign the practice ratings. When the panelists were completing their ratings, Google Docs saved data every few seconds allowing the progress to be monitored in real time by the data specialist. At the end of Round 1 ratings, panelists completed a Round 1 Evaluation Form (See Appendix F) to describe their level of understanding and comfort with the standard setting task, and to identify any areas where additional discussion or training may be necessary prior to Round 2.

Feedback and Discussion for Round 1

Files from all the panelists were saved in real time allowing for immediate analysis when the last panelists finished rating. Using prepared SAS code, the ratings were summarized at the individual and group level for each item and across items to provide feedback data for Round 1. All results received an independent QC before being released. Appendix H displays an example of the individual feedback forms provided to each panelist, and Appendix I shows the item-level feedback form, which includes median ratings from each table along with observed difficulty values (p-values) from the operational administration for each item.

Discussion following Round 1 judgments utilized the observed *p*-value (see Appendix I) for each item to describe item difficulty for the exam-taking population in April 2017. Items were chosen for discussion based on those which were most discrepant across panelists, as well as any additional items that panelists wished to discuss. The Round 1 discussion was held at table level.

Round 2

During Round 2, panelists were instructed to review each item to confirm their rating provided in Round 1 or to provide new ratings as they deemed appropriate based on the information that was presented during the discussion. Panelists were asked to change any ratings they desired from those assigned during Round 1 in their Google Docs Spreadsheet for Round 2. Data from Round 1 of ratings were saved to a master file for each panelist before Round 2 ratings began so that the results from Round 1 were available for documentation.

Feedback and Discussion for Round 2

Feedback for Round 2 was presented to the panelists in a table discussion first and then large-group discussion. The feedback included again, the median cut scores from the table and entire group of panelists, and the median ratings for each item from the table, along with the operational item p-values. Panelists discussed these results at their table and then shared opinions in the large group. Next, impact data, which describes the distribution of students for each performance level, was presented at the table level using recommended cut scores within the table from Round 2. In the meantime, the impact data

based on recommended cut scores by the entire group of panelists was projected on screen. Panelists were given time to discuss the impact data at their tables, but large Group discussion of the impact data was kept at a minimum and discouraged.

Round 3

Following the Round 2 discussion, a third and final round of ratings was conducted. Again, panelists reviewed each item, and they could either confirm or modify the ratings provided in Round 2 in their Google Docs Spreadsheet.

Feedback and Discussion for Round 3

Feedback for Round 3 was presented to the panelists in a large-group setting. The feedback only included the median (i.e. recommended) cut scores from the entire group of panelists and the associated impact data. The panelists were encouraged to share their thoughts about the impact data. It was reminded that the results from Round 3 would serve as the recommendations submitted to the state, but the cut scores were not final yet --- the policy meeting to follow or the state board meeting could modify the cut scores after considering other pieces of information.

Following Round 3, panelists were asked to complete a final evaluation form (see Appendix G) to provide additional evidence of the procedural validity of the standard setting meeting. All meeting materials were then collected, and this was the end of the standard setting meetings.

Evaluating the Standard Setting Procedures

Panelists were asked to complete a rating form after Round 1 and a final evaluation form at the end of the standard setting meeting. Evaluation forms are shown in Appendices E-G, and results are summarized in Tables 3 and 4 for Round 1 and Tables 5 – 12 for the final evaluation.

Round 1 Evaluation

The Round 1 Evaluation Form (Appendix F) was completed by panelists immediately following Round 1 ratings and prior to any discussion of the ratings or results. The primary purpose of this evaluation was to gather evidence about panelist confidence and comfort with the rating task. In addition, panelists were asked to indicate what factors they were finding influential in making their ratings and provided an opportunity to ask for additional information about the process. The completed Round 1 Evaluation Forms were examined prior to the Round 1 discussion in order to determine if any retraining was needed.

Findings from the Round 1 Evaluation are provided in Tables 3 and 4. The first set of evaluation questions asked panelists to respond on a scale from *Strongly Disagree* (1) to *Strongly Agree* (4). The mean ratings for these six questions were generally high, and ranged from 3.1 to 3.7 for Math and 3.5 to 3.8 for ERW. The two relatively low rating questions in Math were related to PLDs (Q4 and Q6). The second set of Round 1 evaluation questions asked panelists to describe influences on their Round 1 ratings using a scale from *Not Influential* (1) to *Very Influential* (3). Relatively high average ratings at or above 2.5 were received in this section, though a lower mean rating was observed related to the influence of test consequences for students (Q10, mean =1.8 for Math and 1.9 for ERW).

Panelists were also asked to provide comments about additional factors that they considered when making the Round 1 ratings (Q12). The comments given were mostly related to item layout, style of questions, text complexity, time constraint, and distractors. The last question on the Round 1 Evaluation asked panelists to leave any other comment that they would like to share at the time (Q13). Most comments provided showed that panelists thought the standard setting process was informative, interesting and beneficial. In the Math panel, one comment raised some concern about the PLDs. Detailed comments for Q12 and Q13 are recorded at the bottom of Tables 3 and 4.

Final Evaluation

Following Round 3, panelists completed a Final Evaluation Form, which was comprised of 4 sections. Summaries of findings from the final evaluation (see Appendix G) are presented in Tables 5 - 12.

Section 1.

First, panelists were asked to indicate the extent to which they agree or disagree with a series of statements about the standard setting process. The rating scale is from *Strongly Disagree* (1) to *Strongly Agree* (4). Findings for Section 1 are summarized in Tables 5 and 6.

Results across ERW and Math indicated that panelists agreed or strongly agreed that they understood the purpose of the study (Q1), and the instructions, explanations (Q2), and training provided (Q3) were clear and sufficient for them to complete the rating task. The mean ratings for these questions were almost always above 3.5. Also, panelists expressed that they understood the concept of borderline examinees (Q5), and agreed it was beneficial to review feedback and have table or room discussion between rounds (Q7, Q8, Q9, Q10). They liked the opportunity to make more than one round of ratings (Q11). The mean ratings for Q7-Q11 were mostly above 3.5. Furthermore, the survey results showed that panelists were engaged throughout the process (Q12) and they were very comfortable to speak up and share their opinions during the discussion (Q13). The mean ratings for Q12-Q13 were close to 4 in both subjects. Ratings for Q14 and Q15 in Math were relatively lower (2.9 for Q14 and 3.2 for Q15) than ratings for other questions, which might

be explained by the comments given by panelists for this part (see below). Nevertheless, ratings for Q14-Q15 in ERW were not lower than others.

Q16 asked panelists to leave comments if they answered *Disagree* or *Strongly Disagree* for any previous question. In Math, the comments primarily focused on the rigor and use of the PLDs ---- many panelists commented that the math PLDs set the performance expectations too high, thus the cut scores recommended were too high. In ERW, the comments cover various topics. Specific comments are recorded at the bottom of Tables 5 and 6.

Section 2.

Section 2 of the Final Evaluation Form is summarized in Tables 7 and 8. This set of questions asked panelists to indicate how influential they found specific activities or pieces of information using a rating scale that ranged from *Not Influential* (1) to *Very Influential* (3). Mean ratings ranged from 2.2 to 2.9 for Math and 1.9 to 2.9 for ERW indicating the majority of examinees rated the tasks/materials as Influential or Very Influential. "Taking the test" (Q17, mean = 2.9) received the highest ratings on both Math and ERW and "Consequences of the exam" (Q27, mean = 2.2 for Math and 1.9 for ERW) received the lowest, though it was still in the Influential to Very Influential range. Lower mean ratings were also seen in ERW for "the average ratings of all panelists in the room" (Q25, mean=2.2) and "Distributions of students" (Q20, mean=2.2). Panelists were asked to provide feedback about additional influential factors; responses are listed at the end of each table and are varied with no consistent primary focus.

Section 3.

Section 3 asked panelists about the usefulness of specific activities or pieces of information that were provided with a rating scale that ranges from *Not Useful* (1) to *Very Useful* (3). Findings are summarized in Tables 9 and 10. Mean ratings of usefulness ranged from 2.6 to 2.9 for Math and 2.3 to 2.9 for ERW. The ratings were generally high, indicating that panelists agreed that most standard setting materials and activities were helpful. "Taking the exam" (Q30, Mean = 2.9) received the highest ratings for Math and ERW and "Test Overview" (Q35, Mean = 2.3) received the highest rating for ERW. An opportunity to provide feedback on other information that would have been useful to the panelists was given on the evaluation form and those comments are summarized at the end of each table. The comments for ERW did not have a consistent theme, yet a few comments in Math mentioned the omitted response data at item level.

Section 4.

The final section of the Final Evaluation Form asked panelists to describe the amount of time devoted to each activity using a scale ranging from *Too Little Time* (1) to *Too Much Time* (3). Findings are summarized in Tables 11 and 12. Mean ratings ranged from 1.9 to 2.4 for Math and 1.5 to 2.1 for ERW. Most mean ratings centered around 2.0

indicating that the timing was generally about right for most panelists. In ERW, some panelists seemed to feel the testing time was a bit too short (Q40).

Finally, panelists were asked to provide additional comments about the standard setting process and logistics; these comments are provided at the end of Tables 11 and 12. Again the PLD theme emerged in the comments for Math. Other than that, lots of positive feedback was provided to this last open-ended question.

Setting the Final Cuts

Ratings from all 3 rounds were provided to the state, along with the Standard Error of Judgment (SEJ) from each round as an indicator of the variability in panelist ratings. The Standard Error of Measurement (SEM), 30.00, was also provided as another measure of variation. The SEJ should get smaller across rounds as the agreement among panelists increases. The SEJ and the SEM may be used to make adjustments to the recommended cut scores when there is a clear rationale for doing so. Generally, adjustments to the cut scores should stay within +/- 1.5 SEJ or SEM to maintain the integrity of the panel.

Table 13 provides a summary of the cut scores (mean, median, min, and max) based on the three rounds of standard setting ratings. In addition to the room-level statistics, table level summary of the cut scores are also reported in this table. The SEJ values for ratings during each round are provided in Table 14. SEJ values are on the SAT scale score metric which increments by 10. All the SEJ values were below 10 indicating they were relatively small and there was general agreement among panelists. The SEJ values for ERW in general decreased over the rounds. The SEJ values for Math had slight increases in Round 2 and Round 3 from Round 1, but they remained very small. This trend is consistent with feedback from Math panelists that they were having difficulty in using PLDs to guide their ratings, hence lack of consensus about expectations for borderline students.

Table 15 summarizes the expected student impact based on the recommended cut scores from Round 3. Table 16 provides the final cut scores agreed upon by the IL Policy meeting and the expected impact. The Exceed Standard cut score for Math decreased from 690 to 670 in the policy meeting, and all other cut score recommendations remained unchanged from Round 3. Finally, Table 17 reports the ratings on the raw score scale from each individual panelist in Round 3, which provides another way to show how close the agreement was between the panelists in the final round.

More procedural validity evidence is provided in appendices J through M. Appendix J presents the slides for the opening session. Appendix K and Appendix L, respectively present the slides used to facilitate the Math and ERW meeting. Appendix M is a memo requesting the State Board to adopt the cut scores from the Policy meeting.

Concluding Comments

Overall, the standard setting meeting went very well. Procedurally, everything went as scheduled and expected. It was shown in the evaluation surveys that the panelists were in general satisfied with the procedures, activities, materials and training, and many of them felt the standard setting meeting was an interesting, educational, and beneficial professional development opportunity. The cut scores recommended by the panel during the third round of ratings were provided to the state departments of education staff for consideration along with other information they considered relevant. On Oct 4th, 2017, the state had the policy meeting and came to agreement on the set of cut scores as listed in Table 16, which accepts the panelists recommendations with the exception of the Exceeds Standards cut score for Math which was lowered slightly in response to panelists concerns about the very high level of expectations presented in the PLDs at that cut score. Though it should be noted the change from 690 to 670 was a change from using the Median panelists' recommendation at the Exceeds Cut for Math to using the Mean panelists' rating and therefore, still in line with the recommendation from the panel. The results from the Policy meeting were moved forward to be shared with the State Board of Education on October 18, and the Board approved the cut score recommendations.

References

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed*.), Educational Measurement*, (2nd ed., pp. 508-560), Washington, DC: American Council on Education.

Cizek, G.J. (2012). *Setting performance standards: Foundations, Methods, and Innovations*. New York: Routledge

Hambleton, R.K., Pitoniak, M.J., & Copella, J.M. (2012). Essential steps in setting performance standards on educational tests and strategies for assessing the reliability of results. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, Methods, and Innovations*, (pp. 47-76). New York: Routledge.

Kane, M. (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: Lawrence Erlbaum Associates.

Pitoniak, M. J., & Morgan, D.L. (2017). Setting and validating cutscores for tests. In C. Secolsky & D.B. Denison (Eds.), *Handbook on Measurement, Assessment, and Evaluation in Higher Education, 2nd edition* (pp.235-258). New York, NY: Routledge.

Pitoniak, M. J., & Morgan, D.L. (2012). Setting and validating cutscores for tests. In C. Secolsky & D.B. Denison (Eds.), *Handbook on Measurement, Assessment, and Evaluation in Higher Education* (pp.343-366). New York, NY: Routledge.

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: Concepts, methods, and perspectives*,(pp. 181-200). Mahwah, NJ: Lawrence Erlbaum Associates.

Table 1 Biographical Information for SAT Math Standard Setting Panel

Survey Topic	Response Categories	*N* (Total *N* = 25)	%
Demographic Characteristics			
Gender	Male	12	48.00
	Female	13	52.00
Race/ethnicity	American Indian or Alaska Native	0	0
	Asian, Asian American, or Pacific Islander	1	4.00
	Black or African American	1	4.00
	Mexican or Mexican American	0	0
	Puerto Rican	0	0
	Other Hispanic, Latino, or Latin American	1	4.00
	White	22	88.00
	Other	0	0
	Multicultural	0	0
Education	Undergraduate Degree	4	16.00
	Master's Degree	21	84.00
	Specialist's Degree	0	0
	Doctoral Degree	0	0
	Other Education Level	0	0
Teaching Experience			
Years of Experience Teaching	1-3 years	0	0
	4-6 years	2	8.00
	7-12 years	4	16.00
	12+ years	19	76.00
	No Experience Teaching	0	0
Current Teaching Level	High School	24	96.00
	College	0	0
	Both College & High School	1	4.00
	Other	0	0
Special Interest Groups*			
	Special Education	11	44.00
	English Language Learners	5	20.00
	Gifted and talented	6	24.00
	Community member	4	16.00
	Postsecondary	2	8.00
	Other	2	8.00

*Categories are not necessarily mutually exclusive.

Table 2 Biographical Information for SAT Math Standard Setting Panel

Survey Topic	Response Categories	*N* (Total *N* =24)	%
Demographic Characteristics			
Gender	Male	5	20.83
	Female	19	79.17
Race/ethnicity	American Indian or Alaska Native	0	0
	Asian, Asian American, or Pacific Islander	0	0
	Black or African American	3	12.50
	Mexican or Mexican American	0	0
	Puerto Rican	0	0
	Other Hispanic, Latino, or Latin American	0	0
	White	19	79.17
	Other	2	8.33
	Multicultural	0	0
Education	Undergraduate Degree	4	16.67
	Master's Degree	17	70.83
	Specialist's Degree	1	4.17
	Doctoral Degree	2	8.33
	Other Education Level	0	0
Teaching Experience			
Years of Experience Teaching	1-3 years	1	4.17
	4-6 years	2	8.33
	7-12 years	3	12.50
	12+ years	18	75.00
	No Experience Teaching	0	0
Current Teaching Level	High School	18	75.00
	College	0	0
	Both College & High School	5	20.83
	Other	1	4.17
Special Interest Groups*			
	Special Education	6	25.00
	English Language Learners	3	12.50
	Gifted and talented	5	20.83
	Community member	0	0
	Postsecondary	2	8.33
	Other	0	0

*Categories are not necessarily mutually exclusive.

Table 3 Summary of Round 1 Evaluation – Math

Rating Scale	Questions		*Mean Rating*
Strongly Disagree (1) to Strongly Agree (4)	*1*	I understand the purpose of the study.	3.7
	2	The instructions and explanations provided by the facilitator were clear.	3.6
	3	The training in the standard setting method gave me the information I needed to complete my assignment.	3.6
	4	The PLDs that were developed prior to the meeting were accurate.	3.1
	5	I understand the concept of the borderline examinee.	3.6
	6	The PLDs helped me to determine how to rate each item.	3.3
Not Influential (1) to Very Influential (3)			
	7	Completing the test before beginning the task.	2.9
	8	My perception of the difficulty of the items	2.6
	9	The PLDs	2.5
	10	The consequences of the test for students	1.8
	11	My experience with students in my classroom	2.7
Extended Response	12	Identify additional factors:	

Responses were as follows:

- The layout of the test, time of test
- experience with student responses to similar questions
- Test fatigue, distractor items, past student success/failures on similar items
- Reading level
- Distracting choices, heavily worded problems, using the initial solution to answer a different question.
- Communication and ideas with the group.
- There were two statistics questions that were printed on two pages. This formatting of the test made answering the second question more challenging (whether or not the question itself was challenging).
- Multiple levels of complexity of the problem. For example a problem may contain 3 different PLDs from level 2. This would be a more challenging problem for level 3 students.
- Group conversations was helpful
- The distractors that were presented. The number of "layers" needed to solve each problem.
- I considered that the test would be taken AFTER the ELA and Reading tests, and the individual placement of test items (in some circumstances).
- I also considered the way the question was asked and the potential factors beyond the mathematics that may impact a student's response.
- my misconceptions of test items
- amount of reading involved to answer question
- Where the questions was in terms of the overall test (was it at the beginning or end for example). Were there distractors in the MC section. Could students plug in numbers to find solutions.

Extended Response	13	Additional Comments:

Responses were as follows:

- When the comparison that the PLD described the student and the rating described the question, that was very helpful
- It seems like 2 days might be too short but I appreciate being a part of this and am interested to see how it works. The calibration is interesting to me.
- I thought some of the practice and preliminary information was repetitive.
- I think the way the PLDs are designed around content is really flawed and can be harmful to the most marginalized groups that we teach. I would appreciate a conversation around the assumption that the borderline examinee is someone who can do all the content in each of the performance levels. That does not match up with the contexts I have taught in and so to dismiss that is to dismiss the experiences of entire populations of marginalized people. I do not like using this theoretical student as a baseline for defining approaching, meeting, or exceeding the standards. I think there are healthier ways to define those that incorporate the diversity of our contexts.
- I liked taking the test before rating it
- I really found today informational. All presenters were able to articulate and convey the necessary information. Thank you all!
- I would like to see specific examples of questions that would accompany the PLD's so I would have a more concrete understanding of what they are asking for.

Table 4 Summary of Round 1 Evaluation – Evidence-Based Reading and Writing

Rating Scale		Questions	*Mean Rating*
Strongly Disagree (1) to Strongly Agree (4)	*1*	I understand the purpose of the study.	3.7
	2	The instructions and explanations provided by the facilitator were clear.	3.6
	3	The training in the standard setting method gave me the information I needed to complete my assignment.	3.8
	4	The PLDs that were developed prior to the meeting were accurate.	3.5
	5	I understand the concept of the borderline examinee.	3.8
	6	The PLDs helped me to determine how to rate each item.	3.5
Not Influential (1) to Very Influential (3)			
	7	Completing the test before beginning the task.	2.8
	8	My perception of the difficulty of the items	2.8
	9	The PLDs	2.6
	10	The consequences of the test for students	1.9
	11	My experience with students in my classroom	2.7
Extended Response	12	Identify additional factors:	
Responses were as follows: • Previous work with the ACT, SAT, and AP tests. • Practice was helpful • Style of question, reliance on student prior knowledge/exposure to content-specific material • discussion • text complexity • The presentation of the PLDs was high to low; the presentation of the Angoff ratings were from low to high. This was an additional translation step in considering test items and ratings. • The importance of the performance level descriptors and the crucial role they play in the cut score determination. • I thought of real kids of mine that were symbolic of the borderline examinees.			
Extended Response	13	Additional Comments:	
Responses were as follows: • I am fairly certain my brain feels like my students after their day of testing. It is so full of information and exhausted. • I am interested in knowing about the second group who will convene after this group's work is completed. • I did my best with the information I had. I would like to have had the presenter model the percentages and how they relate to the test exactly. I know we practiced as a group, but I would have like to see a possible correct range from a previous test. • This has been a very interesting day. Thank you for selecting me to be a part of it. • This session was extremely informative and beneficial to our schools and how we look at material when choosing curriculum and best practices. • This process really makes me examine how I teach and what scaffolding I need to include such that even if my students cannot get the answer correct, they have the tools to THINK at the level 4 type of questioning. This is an eye opener regarding the quality of thinking they must do and how			

low my (and others) expectations of our students have been. My students will have a new teacher when I get back.

- I would like to have had time to completely finish the test. The questions I did not finish--where I just used the answer to determine difficulty--I don't believe I had the full experience the students would have in answering. Just a minor adjustment--overall, everything was very well explained.

Table 5 Summary of Final Evaluation – Section 1 - Math

Rating Scale		Questions	*Mean Rating*
Strongly Disagree (1) to Strongly Agree (4)	1	I understood the purpose of the study.	3.8
	2	The instructions and explanations provided by the facilitator were clear.	3.6
	3	The training on the standard setting method gave me the information I needed to complete my assignment.	3.4
	4	The PLDs that were developed prior to the meeting were accurate.	2.6
	5	I understood the concept of the borderline examinee.	3.4
	6	The PLDs helped me to determine how to rate each item.	2.7
	7	It was beneficial to have an opportunity for table discussions between rounds.	3.8
	8	It was beneficial to have an opportunity to review feedback between rounds.	3.8
	9	The impact data showing the percent of students expected to place into each category **based on my table's** cut scores made a difference in how I rated the items in round 3.	3.4
	10	The impact data showing the percent of students expected to place into each category **based on all panelists in the room's** cut scores made a difference in how I rated the items in round 3.	3.5
	11	The opportunity to make more than 1 round of ratings (i.e., round 2) helped me to feel more confident about my final ratings.	3.6
	12	I felt engaged in the process.	3.8
	13	I was comfortable sharing my ideas with the other panelists during the discussions.	3.8
	14	I am confident this standard setting process will produce fair cut scores.	2.9
	15	I would be comfortable defending this process to my peers.	3.2
Extended Response	16	Comments if *Disagree* or *Strongly Disagree*	

Responses were as follows:

- Please see my comments (panelist 6) on my issues with the PLDs. That is where my discomfort with the process comes from.
- I perceive the level 3 descriptors to be too in depth and heavy. I think some of the level 3 should be in the level 4 descriptors.
- It was difficult overall because I feel that the test was written for a purpose other than what we were attempting to measure. We were using the PLDs but the PLDs were confounded with other aspects of the test and how it math is at the end of a 3-hour test. We as content experts were expected to know this impact but I still feel it was a difficult task overall. It would be nice to have the complete item analysis to see which wrong answers students selected.
- I am not sure the PLDs were clearly written. It would have been helpful is they had been organized by CCSS area. Also, we were given conflicting information by different individuals about "who" they described
- I feel that the PLDs did not correctly represent the borderline student. I think the process of creating the cut points was appropriate, but the baseline provided by the PLDs skewed our scoring of individual items.
- I felt that the score cut offs were way to large. Also the PDL's are not accurate to represent a boarder line student. It skewed the cut off making them higher than I feel they should be.
- The PLD is tricky. High expectations are VERY important. The reality is not always college bound. The

world is an advancing place where technical and advanced skills are important. Education can make those skills accessible to the population of all students

- In my professional opinion, the PLD's are much closer to the average student in each category than the borderline student.
- The PLD's were not an accurate measure of what an borderline student is in the Junior Year.
- I believe the PLDs are written for an average student not a borderline student.
- I think our cut score for “meets” is still too high. I don't think our cut score should be higher than what the College Board deems as college readiness. The PLDs were written with a mathematics curriculum in mind as to what we think is important for further math and did not align to the SAT (nor should they) - but they did influence us and caused concern when we couldn't match questions to levels."
- I feel that the PLDs do not paint an accurate picture of what the borderline student would know when taking the SAT in April of their junior year. I think that the PLDs mostly reflect what a strong student would be able to demonstrate on the SAT in each of the levels.
- My Round 2 cut scores were lower than what the entire group ended up coming around to, there for there was not much need for me to make adjustments. I still think the last round of cut scores are too high and will not accurately represent the students of Illinois. I believe they should be closer to 400, 500 and 600 as the breaks. The exceeds score of 690 is way to high. I believe the when these proposed scores are presented to the next committee, adjustments will need to be made. I initially used the PLD to guide my decision, but after round 1 they became very insignificant as I was concentrating on the layers and distractors provided in each problem as solely the topic that was presented in the problem

Table 6 Summary of Final Evaluation – Section 1 – Evidence-Based Reading and Writing

Rating Scale		Questions	*Mean Rating*
Strongly Disagree (1) to Strongly Agree (4)	1	I understood the purpose of the study.	3.8
	2	The instructions and explanations provided by the facilitator were clear.	3.8
	3	The training on the standard setting method gave me the information I needed to complete my assignment.	3.8
	4	The PLDs that were developed prior to the meeting were accurate.	3.5
	5	I understood the concept of the borderline examinee.	3.8
	6	The PLDs helped me to determine how to rate each item.	3.5
	7	It was beneficial to have an opportunity for table discussions between rounds.	3.8
	8	It was beneficial to have an opportunity to review feedback between rounds.	3.8
	9	The impact data showing the percent of students expected to place into each category **based on my table's** cut scores made a difference in how I rated the items in round 3.	2.9
	10	The impact data showing the percent of students expected to place into each category **based on all panelists in the room's** cut scores made a difference in how I rated the items in round 3.	3.0
	11	The opportunity to make more than 1 round of ratings (i.e., round 2) helped me to feel more confident about my final ratings.	3.5
	12	I felt engaged in the process.	3.8
	13	I was comfortable sharing my ideas with the other panelists during the discussions.	3.6
	14	I am confident this standard setting process will produce fair cut scores.	3.2
	15	I would be comfortable defending this process to my peers.	3.4
Extended Response	16	Comments if *Disagree* or *Strongly Disagree*	

Responses were as follows:

- I am not certain that providing more than one round to discuss and change my cut scores increased my confidence in those ratings. In some ways, it raised more questions than answers.
- I did not feel that it was beneficial to see and discuss the percentage breakdown, they did not affect my rating.
- Given the national benchmark for ELA is 480, I don't believe that all members of the ELA panel were cognizant of the implications of this process nor did they understand a marginal student.
- Our "Meets" cut score is much higher than that of the SAT "Meets" cut score. Although my personal cut score for "Meets" is more in line with that of SAT's
- I made very few changes from Round 2-3 and those results didn't influence me because my Round 2 results were pretty consistent.
- While I understood the process and am appreciative that instructions were given multiple times, I would feel more confident doing this a second time. Knowing what to expect and being more knowledgeable would have an effect on my confidence. Please reiterate again and again to NOT focus on reaching a number, but on using the PDLs. Also, I was so sleepy during the reading test. Not finishing and not doing an item analysis of the questions presented a challenge that kept coming back over and over again with each round. It felt like having to take the test over and over again. I wish I had known that I would be taking the SAT test myself. That way, I would have been focused on what skills I was being assessed.

Table 7 Summary of Final Evaluation – Section 2 – Math

Rating Scale	Questions		*Mean Rating*
Not Influential(1) to Very Influential (3)	*17*	Completing the test before beginning the task.	2.9
	18	My perception of the difficulty of the items	2.8
	19	The actual item difficulty provided for each item	2.6
	20	Distributions of students expected to earn each performance level	2.3
	21	Table discussion after Round 1	2.8
	22	Table discussion after Round 2	2.7
	23	Large group discussion after Round 2	2.7
	24	The average ratings of other panelists at my table	2.4
	25	The average ratings of all panelists in the room	2.3
	26	The PLDs	2.4
	27	The consequences of the exam for students	2.2
	28	My experience with students in my classroom	2.8
Extended Response	*29*	Other factors that influenced decisions:	

Responses were as follows:

- I also considered if the question was MC or grid in, if there were distractor choices.
- The college board's benchmarks for college readiness
- Reading level
- I considered the ELL population and how many of them would be successful on the test items
- No, all considerations I used were listed above.
- The length of the test and when it is given in relation to the rest of the test. Also the grid in questions and non-calculator questions gave students extra challenges that didn't really measure whether or not they met standards.
- My ratings were also influenced by my experience in the classroom, specifically with student motivation, perseverance, and outside influences.
- The questions that provided multiple layers were a struggle to rate. I ended up thinking of this as “partial" credit thinking what my borderline students would be able to get out of 5 points.

Table 8 Summary of Final Evaluation – Section 2 – Evidence-Based Reading and Writing

Rating Scale	Questions		*Mean Rating*
Not Influential (1) to Very Influential (3)	*17*	Completing the test before beginning the task.	2.9
	18	My perception of the difficulty of the items	2.8
	19	The actual item difficulty provided for each item	2.8
	20	Distributions of students expected to earn each performance level	2.2
	21	Table discussion after Round 1	2.5
	22	Table discussion after Round 2	2.3
	23	Large group discussion after Round 2	2.3
	24	The average ratings of other panelists at my table	2.4
	25	The average ratings of all panelists in the room	2.2
	26	The PLDs	2.8
	27	The consequences of the exam for students	1.9
	28	My experience with students in my classroom	2.8
Extended Response	*29*	Other factors that influenced decisions:	

Responses were as follows:

- As you asked, I kept picturing the borderline student for that PLD.
- I think slightly our test fatigue from the morning could have altered Round 1 ratings a bit. It seemed like we were very exhausted at the end of Day 1 with the first round of ratings. So, it is probably a good idea to do ratings 2 and 3.
- My perception of question validity--some questions seemed to be worded with a high degree of difficulty or relied on unclear graphs and charts.

Table 9 Summary of Final Evaluation – Section 3 – Math

Rating Scale	Questions		*Mean Rating*
Not Useful(1) to Very Useful(3)	*30*	Taking the exam prior to beginning the task	2.9
	31	Practicing the procedure with real items prior to beginning the actual rating task	2.6
	32	Referencing the PLDs	2.6
	33	Table discussion after round 1	2.9
	34	Table discussion after round 2	2.8
	35	Test overview	2.6
	36	Actual item difficulty values	2.8
	37	Distribution of students earning each performance level	2.6
	38	Large group discussion after round 2	2.8
Extended Response	*39*	Other information that would have been useful during ratings:	

Responses were as follows:

- Current mean and standard deviation of scores.
- The percent of each question that was left blank.
- Item analysis to see how many omitted and how many chose certain answers.
- A possibility of changing tables to hear other opinions not only the large group discussion.
- item analysis - how many problems were omited
- Knowing the SAT cut scores would have helped keep things in perspective. At least on the 3rd round. Our borderline students are supposed to be above the College's Board readiness standard. That shows that we are too high for the borderline scores.
- I thought the three stage process worked extremely well for the task at hand.
- We wondered what percentage of students skipped answers and how that affected the item difficulty.
- It may have been beneficial to know how many students actually answered (especially the grid-in) questions, versus just left them blank.
- I would have liked to see the number of students who left questions blank be removed from the level of difficulty, just to see how the kids who answered the question did.

Table 10 Summary of Final Evaluation – Section 3 – Evidence-Based Reading and Writing

Rating Scale	Questions		*Mean Rating*
Not Useful(1) to Very Useful(3)	*30*	Taking the exam prior to beginning the task	2.9
	31	Practicing the procedure with real items prior to beginning the actual rating task	2.8
	32	Referencing the PLDs	2.8
	33	Table discussion after round 1	2.8
	34	Table discussion after round 2	2.6
	35	Test overview	2.3
	36	Actual item difficulty values	2.7
	37	Distribution of students earning each performance level	2.4
	38	Large group discussion after round 2	2.4
Extended Response	*39*	Other information that would have been useful during ratings:	

Responses were as follows:

- Possibly deliver feedback with my ratings side by side with group
- I would've liked to know the actual difficulty of the texts; for instance: the Lexile score
- SAT's national ratings. SAT has given this test for many years. That information is valuable.
- Knowing ahead of time that I would be taking the exam myself.

Table 11 Summary of Final Evaluation – Section 4 – Math

Rating Scale	Questions		*Mean Rating*
Too Little Time (1) to Too Much Time (3)	*40*	Taking the test	1.9
	41	Reviewing the PLDs	2.0
	42	Training on the rating task before Round 1	2.0
	43	Round 1 of the rating task	2.2
	44	Table discussion after round 1	2.2
	45	Round 2 of the rating task	2.4
	46	Table discussion after round 2	2.4
	47	Large group discussion after round 2	2.1
	48	Review of impact data for the total group	2.1
	49	Review of impact data for my table	2.2
Extended Response	*50*	Any additional comments:	

Responses were as follows:

- Well organized activities overall.
- I think the cut scores probably better represent an average student in each category rather than a borderline student. I think the final cut scores should be lower. I think the PLD's better represent an average student in each category.
- The process works well. I am unsure of the PLD process. I would question that
- This was an interesting experience. Education is so interesting and the changes happening continually are somewhat frustrating. I do think change is the only way improvement can happen. I appreciate the opportunity to be a part of this process.
- I am not comfortable with our results and do not stand behind the results reached here.
- I believe the PLD's for meeting standards (level 3) covers way too much information, which is reflected in our high cut score for meeting standards. Overall, I believe a student should be able to meet standards and not be able to do everything included on that list.
- We were curious about who developed the PDL's and set the levels for the skills.
- The PLDs are not written for that borderline student but an average student - I think this influenced our group greatly. The percent of partially and approaching are still very high - I think we did a good job for the average student but is it overall a good representation of the borderline student
- I consider the PLDs to be too rigorous for the borderline student at each level. I rated our cut scores as appropriate for the PLD descriptors, but I would not say they are appropriate for the borderline student in my own description.
- More space would have been helpful. Otherwise, everything went very smoothly!
- I thought the entire process was very interesting and I feel honored to have been able to be a participant! Thank you for your time and I feel that you facilitated the process WONDERFULLY!
- I think some of the tables relied too much on the PLD's instead of taking in the other factors of the test. The PLD's were a guideline of where students should be as a average or above average. The PLD's in my opinion was not representative of the borderline student. I do believe that the cut off for Meets is still too high in the spectrum. I did enjoy being part of this process. I feel very honored.

- This process was very interesting. I think we, as educators of the students tested, have high expectations for our students and we tend to forget the many factors that effect their scores on the SAT. I had trouble coming down from the expectation of high percentages for my "borderline exceeds" students. After much deliberation with my team
- I think more discussion should have happened before round 3. The impact data was shocking and demonstrated how far apart each table was from what SAT suggests as college ready.

Table 12 Summary of Final Evaluation – Section 4 – Evidence-Based Reading and Writing

Rating Scale	Questions		*Mean Rating*
Too Little Time (1) to Too Much Time (3)	*40*	Taking the test	1.5
	41	Reviewing the PLDs	2.1
	42	Training on the rating task before Round 1	2.2
	43	Round 1 of the rating task	2.0
	44	Table discussion after round 1	2.0
	45	Round 2 of the rating task	2.1
	46	Table discussion after round 2	2.1
	47	Large group discussion after round 2	2.0
	48	Review of impact data for the total group	2.0
	49	Review of impact data for my table	2.0
Extended Response	*50*	Any additional comments:	

Responses were as follows:

- I would like to have been explained how you arrived at the data.
- This process has really opened my eyes into all of the work that goes into creating the SAT test
- A shuttle bus from one location to the other may have been helpful. Rae and Pamela did a fantastic job!!!
- This was a very enlightening process! Thank you for selecting me to be a part of it!
- Wonderful to see this side of the process, would be great to pull in teachers and students when selecting texts for the exam.

Table 13 Cut Score Summary by Round

Math Overall

Round	Approaching Standard					Meets Standard					Exceeds Standard				
	Mean	Median	Min	Max	SD	Mean	Median	Min	Max	SD	Mean	Median	Min	Max	SD
1	502	**500**	460	540	26.30	610	**610**	550	650	26.22	714	**720**	650	770	28.27
2	466	**480**	400	530	32.64	568	**580**	490	630	35.24	691	**700**	590	740	40.34
3	445	**450**	350	520	36.53	542	**540**	480	600	31.13	676	**690**	590	730	39.57

Math Table 1

Round	Approaching Standard					Meets Standard					Exceeds Standard				
	Mean	Median	Min	Max	SD	Mean	Median	Min	Max	SD	Mean	Median	Min	Max	SD
1	498	**510**	470	520	21.68	598	**590**	580	640	24.90	696	**690**	650	750	37.15
2	432	**430**	400	480	34.21	528	**540**	490	570	32.71	628	**620**	590	680	37.01
3	428	**430**	390	480	35.64	516	**510**	490	560	29.66	620	**610**	590	660	33.17

Math Table 2

Round	Approaching Standard					Meets Standard					Exceeds Standard				
	Mean	Median	Min	Max	SD	Mean	Median	Min	Max	SD	Mean	Median	Min	Max	SD
1	514	**520**	470	540	29.66	620	**620**	600	640	20.00	730	**730**	710	740	12.25
2	486	**480**	460	530	27.02	580	**580**	540	630	36.74	724	**730**	710	730	8.94
3	476	**470**	450	520	27.02	568	**568**	540	600	26.83	712	**710**	700	730	13.04

Math Table 3

Round	Approaching Standard					Meets Standard					Exceeds Standard				
	Mean	Median	Min	Max	SD	Mean	Median	Min	Max	SD	Mean	Median	Min	Max	SD
1	496	**490**	470	540	27.02	612	**610**	580	650	28.64	708	**700**	650	770	43.24
2	486	**480**	480	500	8.94	594	**590**	590	610	8.94	708	**700**	690	740	21.68
3	462	**460**	450	480	13.04	566	**560**	550	590	15.17	694	**700**	660	730	28.81

Math Table 4

Round	Approaching Standard					Meets Standard					Exceeds Standard				
	Mean	Median	Min	Max	SD	Mean	Median	Min	Max	SD	Mean	Median	Min	Max	SD
1	494	**500**	460	530	32.86	606	**610**	550	640	35.07	712	**710**	690	750	22.80
2	440	**440**	420	460	15.81	554	**560**	530	580	19.49	686	**690**	650	720	25.10
3	404	**400**	350	450	38.47	520	**530**	480	540	25.50	682	**690**	660	690	13.04

Math Table 5

Round	Approaching Standard					Meets Standard					Exceeds Standard				
	Mean	Median	Min	Max	SD	Mean	Median	Min	Max	SD	Mean	Median	Min	Max	SD
1	508	**510**	470	530	24.90	612	**630**	570	630	26.83	722	**720**	720	730	4.47
2	488	**480**	470	510	16.43	584	**590**	530	610	31.30	710	**710**	700	720	7.07
3	454	**450**	440	470	15.17	542	**540**	520	560	17.89	674	**680**	620	710	33.62

ERW Overall

Round	Approaching Standard					Meets Standard					Exceeds Standard				
	Mean	Median	Min	Max	SD	Mean	Median	Min	Max	SD	Mean	Median	Min	Max	SD
1	449	**440**	400	540	39.49	554	**555**	480	620	39.32	653	**655**	560	710	36.05
2	432	**430**	400	480	22.65	536	**540**	500	570	21.23	641	**640**	590	700	32.34
3	430	**430**	390	480	23.93	534	**540**	500	570	20.41	639	**640**	580	700	31.75

*Values have been rounded to the closest reportable score

ERW Table 1

Round	Approaching Standard					Meets Standard					Exceeds Standard				
	Mean	Median	Min	Max	SD	Mean	Median	Min	Max	SD	Mean	Median	Min	Max	SD
1	454	**430**	410	520	48.27	578	**570**	550	620	31.14	674	**670**	630	710	32.09
2	444	**450**	420	480	25.10	556	**560**	540	570	11.40	664	**670**	610	700	39.12
3	446	**440**	420	480	21.91	552	**560**	540	560	10.95	662	**680**	610	700	36.33

ERW Table 2

Round	Approaching Standard					Meets Standard					Exceeds Standard				
	Mean	Median	Min	Max	SD	Mean	Median	Min	Max	SD	Mean	Median	Min	Max	SD
1	464	**460**	400	540	51.77	566	**540**	530	620	40.99	658	**660**	600	710	39.62
2	426	**420**	400	460	27.93	522	**530**	500	540	16.43	624	**640**	590	650	27.02
3	422	**420**	390	460	32.71	518	**510**	500	540	16.43	622	**640**	580	650	30.33

ERW Table 3

Round	Approaching Standard					Meets Standard					Exceeds Standard				
	Mean	Median	Min	Max	SD	Mean	Median	Min	Max	SD	Mean	Median	Min	Max	SD
1	434	**430**	420	450	15.17	526	**520**	490	560	32.86	628	**640**	560	660	39.62
2	428	**430**	420	440	8.37	518	**520**	500	540	17.89	618	**620**	590	650	23.87
3	428	**430**	420	430	5.00	515	**515**	500	530	12.91	615	**615**	600	630	12.91

ERW Table 4

Round	Approaching Standard					Meets Standard					Exceeds Standard				
	Mean	Median	Min	Max	SD	Mean	Median	Min	Max	SD	Mean	Median	Min	Max	SD
1	472	**480**	440	510	31.14	564	**560**	510	610	38.47	654	**640**	620	700	34.35
2	448	**440**	430	470	16.43	550	**550**	530	570	15.81	642	**640**	610	690	29.50
3	442	**440**	410	460	20.49	546	**550**	520	570	18.17	636	**640**	590	680	32.09

ERW Table 5

Round	Approaching Standard					Meets Standard					Exceeds Standard				
	Mean	Median	Min	Max	SD	Mean	Median	Min	Max	SD	Mean	Median	Min	Max	SD
1	413	**410**	400	430	15.00	530	**540**	480	560	38.30	650	**655**	610	680	31.62
2	410	**405**	400	430	14.14	535	**535**	520	550	17.32	663	**660**	650	680	15.00
3	408	**405**	400	420	9.57	538	**540**	520	550	15.00	660	**660**	650	670	11.55

Table 14 Standard Error of Judgment by Round (SAT Scale Score Metric)

Math

Standard Error of Judgment	Approaching Standard	Meets Standard	Exceeds Standard
Round 1	5.26	5.24	5.65
Round 2	6.53	7.05	8.07
Round 3	7.31	6.23	7.91

Evidence-Based Reading and Writing

Standard Error of Judgment	Approaching Standard	Meets Standard	Exceeds Standard
Round 1	8.06	8.03	7.36
Round 2	4.62	4.33	6.60
Round 3	4.99	4.26	6.62

Table 15 Impact data for Round 3 – Percent in Category Using Median Cut Scores

Math

Group	Partially Meets Standard	Approaching Standard	Meets Standard	Exceeds Standard
Overall	30.40	32.75	30.70	6.15
Table 1	23.12	28.20	31.55	17.13
Table 2	37.57	37.65	20.18	4.60
Table 3	33.98	36.80	23.88	5.34
Table 4	15.93	44.48	33.44	6.15
Table 5	30.40	32.75	29.79	7.06

Evidence-Based Reading and Writing

Group	Partially Meets Standard	Approaching Standard	Meets Standard	Exceeds Standard
Overall	22.40	37.31	26.99	13.31
Table 1	25.45	40.70	27.08	6.76
Table 2	18.82	31.00	36.88	13.31
Table 3	22.40	30.83	29.30	17.47
Table 4	25.45	37.57	23.67	13.31
Table 5	15.92	43.79	30.56	9.74

Table 16 Final Cut Scores from Policy Meeting and Impact Data

Math

Performance Level	Cut Score	Percent of Examinees in Level
Exceeds Standard	670	7.98
Meets Standard	540	28.87
Approaching Standard	450	32.75
Partially Meets Standard	--	30.40

Evidence-Based Reading and Writing

Performance Level	Cut Score	Percent of Examinees in Level
Exceeds Standard	640	13.31
Meets Standard	540	26.99
Approaching Standard	430	37.31
Partially Meets Standard	--	22.40

Table 17 Individual Panelist Ratings (on the raw score scale) from Round 3

Math (the number of items = 58)

Panelist	Approaching Standard	Meets Standard	Exceeds Standard
1	18.00	24.85	44.25
2	15.20	23.00	35.60
3	22.40	32.30	42.00
4	16.65	27.70	37.60
5	13.90	22.55	36.15
6	21.90	33.80	50.90
7	25.65	36.90	47.95
8	19.15	30.20	49.95
9	20.05	29.35	48.50
10	21.00	34.00	48.05
11	19.50	32.05	43.55
12	21.90	34.90	48.80
13	19.05	33.35	47.55
14	20.65	31.30	50.90
15	18.60	31.75	45.05
16	10.50	22.00	47.10
17	18.60	29.70	44.40
18	17.05	27.50	46.55
19	14.40	25.10	46.50
20	15.30	28.70	46.25
21	19.25	26.35	47.45
22	17.65	27.70	45.95
23	18.00	28.60	45.40
24	20.70	31.75	48.50
25	20.60	31.60	39.10

ERW (the number of items = 96)

Panelist	Approaching Standard	Meets Standard	Exceeds Standard
1	42.70	57.45	67.40
2	33.75	54.65	73.10
3	35.55	55.05	78.90
4	37.65	57.85	79.45
5	36.95	56.40	81.70
6	29.90	48.65	72.05
7	29.95	46.20	73.10
8	40.90	52.45	65.05
9	39.55	53.20	73.80
10	33.20	47.90	62.10
11*	-	-	-
12	33.85	48.15	69.20
13	34.95	50.15	71.40
14	35.40	48.80	65.55
15	36.70	51.90	66.50
16	40.45	52.75	63.85
17	33.65	55.85	72.70
18	38.00	55.65	73.05
19	38.00	50.40	69.80
20	39.60	58.25	79.10
21	30.95	50.95	78.30
22	31.80	50.35	73.60
23	34.00	55.50	77.30
24	32.25	55.00	74.35

*Panelist 11 completed the first two rounds of rating, but did not participate in the third round due to a family emergency.

Figure 1 Diagram of Performance Levels

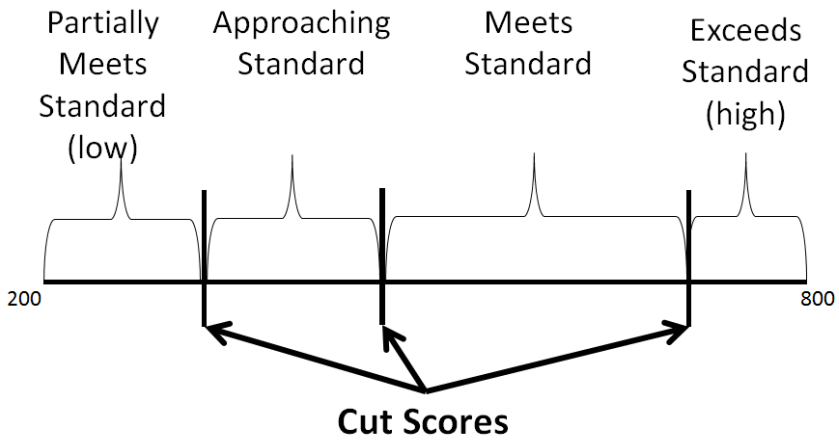
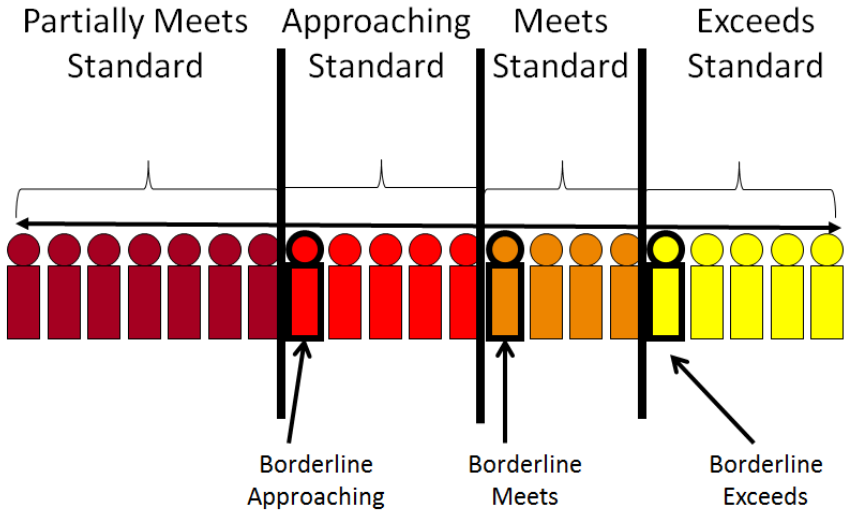


Figure 2 Borderline Examinees



Note. This figure depicts the concept of a *borderline examinee,* and was used during training for the standard setting described in this report. The proportions of examinees shown in each category are not intended to reflect the distribution of examinees within the four categories, actual or expected.

Appendix A: Illinois SAT ERW Standard Setting Agenda

September 27-28, 2017

Wyndham Springfield City Center
700 E. Adams Street
Springfield, IL 62701

Day 1

7:15 – 8:25am	Breakfast and Check in (Prairie Room on the Mezzanine Level)
8:25 – 9:00am	Welcome, Introductions, Announcements, Overview (Prairie Room)
9:00 – 9:05am	Proceed to Break Out Room (Conference Center #4)
9:05 – 9:45am	Welcome and Introduction, Overview of SAT ERW Section
9:45 – 11:15am	Experience the ERW Section of the SAT
11:15 – 11:20	Break (Outside the meeting room)
11:20 – 12:00pm	Review of MCQ items
12:00 – 12:45pm	Lunch (Prairie Room)
12:45 – 2:00pm	Review of Performance Level Descriptors (PLDs)
2:00 – 2:10pm	Break (Outside the meeting room)
2:10 – 3:15pm	Training and Practice on Modified Angoff method
3:15 – 6:00pm	Provide Round 1 Ratings, Evaluation Form

****

6:30 – 8:00pm	Dinner on your own

Day 2

7:15 – 8:25am	Breakfast (Prairie Room)
8:25 – 8:30am	Check In – Conference Center #4

8:30 – 10:00am	Feedback and Table Discussion of Round 1 Ratings
10:00 – 10:15am	Break (Outside meeting room)
10:15 – 11:45am	Provide Round 2 Ratings
11:45 – 12:45pm	Lunch (Prairie Room)
12:45 – 1:30pm	Feedback and Table Discussion of Round 2 Ratings and Impact Data
1:30 – 2:00pm	Large Group Discussion of Round 2 Ratings
2:00 – 2:30pm	Presentation of Impact Data and Discussion at your tables
2:30 – 2:45pm	Break (Outside meeting room)
2:45 – 4:00pm	Provide Round 3 Ratings
4:00 – 4:15pm	Start Material Check In
4:15 – 4:45pm	Final Debrief
4:45 – 5:00pm	Final Evaluation Form, Check in Materials, Dismiss

Appendix B: Biographical Data Form

Biographical Data Form

Please circle the letter of the answer choice that most represents you. This information is for the reporting of panel member diversity as a measure of the generalizability and validity of the results and will be reported in aggregate form only. Data will be used for research purposes only.

1. **Gender:**

   A. Female
   B. Male

2. **How do you describe yourself? (**please choose ONE option, if more than 1 apply then please choose Multicultural)

   A. American Indian or Alaska Native
   B. Asian, Asian American or Pacific Islander
   C. Black or African American
   D. Mexican or Mexican American
   E. Puerto Rican
   F. Other Hispanic, Latino, or Latin American
   G. White
   H. Other
   I. Multicultural

3. **Years of experience teaching Reading and/or Writing at the high school level or a corresponding course at the college level (including this year):**

   A. 1 - 3 years
   B. 4 - 6 years
   C. 7 - 12 years
   D. More than 12 years

4. **Levels of Classes You Teach:**

   A. High School
   B. College
   C. Both College & High School

D. Other:_____

5. Please list the name of the institution where you teach and the city, state where the institution is located:

Name of Institution: _____

City, State of Institution: _____

6. My highest level of education completed is:

A. Undergraduate Degree
B. Master's Degree
C. Specialist Degree or ABD
D. Doctoral Degree
E. Other:_____

7. Do you represent any of the following special interest groups?

A. Special Education
B. English Language Learners
C. Gifted and Talented
D. Community Member
E. Postsecondary
F. Other: (Please specify)_____

Appendix C: Performance Level Descriptors

ILLINOIS STATE BOARD OF EDUCATION

Performance Level Descriptors

Grade 11: Evidence-Based Reading & Writing

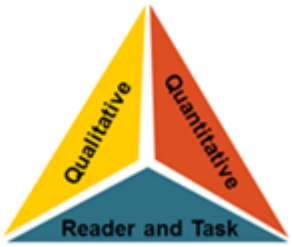
Level 4 Exceeds Standards	Level 3 Meets Standards	Level 2 Approaching Standards	Level 1 Partially Meets Standards
The student has exceeded the proficiency level & demonstrates a **thorough** understanding of the knowledge & skills needed relative to the Common Core ELA/Literacy content standards. The student can understand highly complex texts, perform complex analyses of and across texts, and create compositions that demonstrate mastery of all compositional elements.	The student has met the proficiency level & demonstrates **adequate** understanding of the knowledge & skills needed relative to the Common Core ELA/Literacy content standards. The student can perform analyses of complex texts that identify or infer abstract themes and core information and make connections across texts. The student can create compositions that include sophisticated sentence structure and appropriate application of all compositional elements.	The student is approaching the proficiency level & demonstrates an **incomplete** understanding of the knowledge & skills needed relative to the Common Core ELA/Literacy content standards. Students at this level can understand moderately complex texts, passages, & pairs of passages & important concepts within in simple ways. The student can engage in writing tasks with simple organizational structures of at least one basic paragraph.	The student has only partially met standards & demonstrates a **minimal** understanding of the knowledge & skills needed relative to the Common Core ELA/Literacy content standards. Students at this level can read low-complexity & moderately complex texts to identify basic facts or ideas, showing shallow comprehension of passages and/or pairs of passages. The student can write & edit basic ideas observing some standard English conventions, such as appropriate grammar, punctuation, capitalization, & spelling, but few organizational or compositional elements.

The primary differences between the performance levels are the complexity of the text(s) and complexity of the task in interaction with each other, as well as overall mastery. Students at lower levels may be able to successfully perform some higher-level tasks if the text complexity is low enough or may be able to successfully interact with higher-complexity texts if the task is simple enough. At the higher levels, in addition to the interaction between text and task complexity, mastery of the breadth of literacy knowledge and skills is also a consideration.

Level of Text	Description
Low-Complexity Text	These are texts whose information & ideas as well as structure, purpose, & language are relatively simple & direct & require relatively little analysis. They use common, simple vocabulary and have shorter sentences that can be read and understood even by slow readers.
Moderately Complex Text	These are texts whose information & ideas as well as structure, purpose, & language require some analysis. They use relatively common but more advanced vocabulary and have longer and more sophisticated sentences. A slow reader can understand moderately complex texts but may need to reread or use other strategies to fully comprehend.
Complex Text	These are texts that can be difficult to understand at first because the information, ideas, structure, purpose, & language may be advanced or unfamiliar. Complex texts use advanced and unfamiliar vocabulary, including regular use of technical or academic vocabulary. The sentence and passage structures are complex, and slow readers would struggle to comprehend complex text even with significant use of support strategies.
Highly Complex Text	These are texts that can be challenging even for skilled readers. Often, the information, ideas, structure, purpose, & language are highly advanced or unfamiliar.

Table 1a: Text Complexity Definition

The redesigned SAT's passages/passage pair represent a specified range of text complexity from grades 9–10 to postsecondary entry. Text complexity is defined as:

1. Quantitative measures – readability & other scores of text difficulty; often best measured by computer software. 2. Qualitative measures – levels of meaning, structure, & language conventionality & clarity, & knowledge demands; often best measured by an attentive human reader. 3. Reader & task considerations – background knowledge of reader, motivation, interests, & complexity generated by tasks assigned; often best assessed by educators employing their professional judgment.	**CCSS TEXT COMPLEXITY MODEL** 

PERFORMANCE LEVEL 1

Students at this level can read low-complexity & moderately complex texts to identify basic facts or ideas, showing shallow comprehension of passages &/or pairs of passages. The student can write & edit basic ideas observing some standard English conventions, such as appropriate grammar, punctuation, capitalization, & spelling, but few organizational or compositional elements.

Students at this level can

1. Read a low-complexity or moderately complex passage closely to identify explicitly stated information & ideas or draw a simple reasonable inference
2. Determine the best textual evidence for a simple inference
3. Identify or infer the central idea or theme of a passage or that has a single, clear purpose
4. Identify an accurate summary of a passage or of key information & ideas in a passage
5. Determine a simple relationship between information, ideas, or people depicted in a passage (e.g., recognizing a basic comparison, contrast, or sequence)
6. Determine the meaning of relatively common words or phrases using clear context clues
7. Determine the main purpose of a low-complexity or moderately complex passage
8. Identify evidence to support a claim or counterclaim in a low-complexity passage
9. Identify a similarity or difference in a pair of low-complexity or moderately complex passages (e.g., recognizing that a particular detail appears in one passage but not the other)
10. Recognize a straightforward similarity or difference in a pair of low-complexity to moderately complex passages (e.g., in reading passages on the same topic, recognizing basic similarities & differences in how an event is depicted)
11. Locate data or make a simple accurate interpretation of data in an informational graphic, such as a table, graph, or chart
12. Delete information or ideas that are obviously irrelevant to the main focus of a paragraph or passage
13. Order the sentences in a paragraph to achieve a simple purpose (e.g., grouping related information together; establishing a basic chronology)
14. Use a transitional word or phrase to establish a simple logical relationship between sentences
15. Make an effective word or phrase choice in a straightforward situation
16. Eliminate obvious wordiness or redundancy within a portion of a sentence
17. Combine sentences in a relatively simple way (e.g., making a second sentence into a relative clause of the first) or to achieve a relatively simple purpose
18. Recognize & correct an obviously inappropriate shift in verb tense
19. Recognize & maintain or correct subject-verb or pronoun-antecedent agreement in a straightforward situation
20. Use conventional expression in a straightforward situation
21. Use standard English conventions (e.g., distinguish between singular & plural possessive nouns & between plural & possessive nouns, appropriately punctuate items in a series, eliminate obviously unnecessary & disruptive punctuation)

PERFORMANCE LEVEL 2

Students at this level can understand moderately complex texts, passages, & pairs of passages & important concepts within in simple ways. The student can engage in writing tasks with simple organizational structures of at least one basic paragraph.

Students at this level can (in addition to level 1)

1. Read a moderately complex passage closely to identify explicitly stated information or ideas or draw a simple reasonable inference
2. Determine the best textual evidence for an inference when the evidence requires some interpretation or analysis
3. Determine the central idea or theme of a moderately complex or complex passage
4. Determine the meaning of a relatively common high-utility academic word or phrase in context; determine the meaning of a straightforward figurative expression
5. Determine the main purpose or effect of an author's word choice in a complex passage or in a simpler passage when the purpose or effect is somewhat subtle (e.g., an author using words to convey a particular emotion)
6. Determine a clear relationship between a particular part of a passage (e.g., a sentence) & the whole passage
7. Draw a straightforward reasonable inference about point of view or perspective in a moderately complex passage (e.g., identifying the impact of a technique the author uses to shape point of view in a literary passage; distinguishing among the multiple perspectives in an informational passage)
8. Determine the main purpose of a moderately complex passage or of one of its paragraphs
9. Determine a claim or counterclaim in a moderately complex argument
10. Synthesize information & ideas from a pair of moderately complex passages
11. Locate data or make an accurate interpretation of data in an informational graphic, such as a table, graph, or chart (e.g., drawing a valid conclusion based on an understanding of a bar graph's overall purpose; summarizing a clear trend from several data points)
12. Draw a straightforward supportable connection between a graphic & its accompanying passage (e.g., determining a graphic's clear main purpose & finding a matching assertion in the passage)
13. Compose a paragraph that has a clear, well-defined focus
14. Use supporting information to achieve a straightforward purpose
15. Use a moderately complex informational graphic, such as a table, graph, or chart, to revise a low-complexity or moderately complex passage
16. Order the sentences in a paragraph to achieve a straightforward purpose (e.g., repositioning a supporting detail immediately after a sentence that makes a claim)
17. Introduce or conclude a paragraph or passage based on a general understanding of its content & purpose
18. Make an effective word or phrase choice based on vocabulary knowledge & an understanding of the context
19. Eliminate wordiness or redundancy within sentences & moderately complex passages
20. Maintain a basic consistency in style & tone within a passage
21. Form conventional, complete sentences, recognizing & correcting a disruption in structure (e.g., eliminating an obvious comma splice or fragment; correcting or maintain a lack of parallelism in a simple series or series of phrases; replacing nonstandard terms with standard

PERFORMANCE LEVEL 3

Students at this level can perform analyses of complex texts that identify or infer abstract themes, core information and make connections across texts. The student can create compositions that include complex sentence structures and appropriate application of all compositional elements.

Students at this level can (in addition to level 2)

1. Extrapolate in a reasonable way from the information & ideas in a complex passage or apply information & ideas in such a passage to a new, analogous situation
2. Determine the best textual evidence for an inference when the evidence requires some interpretation or analysis & the inference requires close reading
3. Determine the central idea or theme of a complex passage that features several important ideas
4. Identify an accurate summary of a complex passage or of key information & ideas in a complex passage
5. Determine a relationship between information, ideas, or people depicted in a complex passage
6. Determine the meaning of a relatively uncommon high-utility academic word or phrase in context; determine the meaning of a moderately challenging figurative expression
7. Determine the main purpose or effect of an author's word choice in a complex passage or in a simpler passage when the purpose or effect is fairly subtle or complex (e.g., an author using wordplay or parody)
8. Determine the main purpose of a particular part of a passage (e.g., a detail or a metaphor) in relation to the passage as a whole
9. Draw a reasonable inference about point of view or perspective in a complex passage (e.g., identifying where point of view switches in a literary passage; distinguishing among conflicting perspectives in an informational passage)
10. Determine the main purpose of a complex passage or of one of its paragraphs
11. Determine a claim or counterclaim in a complex argument
12. Analyze a subtle argumentative technique or flaw (e.g. an author using weak reasoning in support of a claim)
13. Synthesize information & ideas from a pair of complex passages
14. Make an accurate, somewhat subtle or complex interpretation of data in an informational graphic, such as a table, graph, or chart (e.g., comparing results in terms of two variables; recognizing an implication of the values represented on a table)
15. Draw a supportable connection between a graphic & its accompanying passage
16. Interpret, paraphrase, or summarize data in an informational graphic, such as a table, graph, or chart, & incorporate them in a passage in an accurate, relevant way (e.g., encompassing multiple data points in a single relevant general statement)
17. Compose a passage based on an understanding of its content & purpose
18. Establish & clarify the structure of a paragraph or passage
19. Use supporting information to develop a point or claim logically
20. Sharpen the focus of a paragraph or passage by making a thoughtful decision about adding, revising, or deleting information or ideas
21. Revise a paragraph to address a critical issue of logic or cohesion
22. Make a nuanced word or phrase choice based on well-developed vocabulary knowledge & an understanding of the context

23. Draw logical comparisons using appropriate terms (e.g. using allusions, analogies, metaphors, similes).
24. Eliminate relatively subtle wordiness or redundancy within a sentence or between sentences (e.g., recognizing when information over explains a concept & correcting accordingly; deleting repetition involving fairly sophisticated language)
25. Make a thoughtful decision about style & tone in a passage based on an understanding of the context (e.g., revising language that is too informal or formal in a fairly challenging context; achieving a particular rhetorical aim, such as establishing a particular sentence pattern or choosing language that sets a contextually appropriate mood)
26. Observe standard English and rhetorical and compositional conventions when composing or correcting complex texts and passages in complex ways, for example:
    a. eliminating a rhetorically inappropriate fragment or eliminating a conjunction based on an understanding of the syntax of a relatively sophisticated or long sentence
    b. making careful distinctions among the possessive determiners *its* & *their*, the contractions *it's* & *they're*, & the adverb *there*
    c. maintaining subject-verb or pronoun-antecedent agreement in a challenging situation
    d. making careful distinctions among singular, singular possessive, plural, & plural possessive nouns
    e. making informed decisions on whether to use punctuation based on an understanding of context
    f. eliminate unnecessary punctuation in a challenging situation)
    g. Using conventional expression in a challenging situation (e.g., selecting appropriately between relatively uncommon words that are frequently confused, such as *discrete* & *discreet*)

PERFORMANCE LEVEL 4

Students at this level can understand highly complex texts, perform complex analyses of and across texts, and create compositions that demonstrate mastery of all compositional and conventional expression elements.

Students at this level can (in addition to level 3)

1. Read a complex or highly complex passage to identify explicitly stated information & ideas or draw a reasonable inference
2. Extrapolate in a reasonable way from the information & ideas in a complex or highly complex passage or apply information & ideas in such a passage to a new, analogous situation
3. Determine the best textual evidence for an inference when the evidence is subtle, abstract, or figurative & the inference requires multiple steps
4. Determine the central idea or theme of a highly complex passage, or identify an accurate summary of key information & ideas
5. Determine a relationship between information, ideas, or people depicted in a highly complex passage
6. Determine the meaning of an uncommon high-utility academic word or phrase in context, including an archaic usage found in a passage from an earlier time period; determine the meaning of a subtle or complex figurative expression
7. Determine the main purpose or effect of an author's word choice in a highly complex passage or in a simpler passage when the purpose or effect is subtle or complex
8. Determine the main purpose of a particular part of a passage in relation to the passage as a whole when the purpose is subtle or complex (e.g., the author using rhetorical questions to indicate self- evident truths)
9. Draw a nuanced inference about point of view or perspective in a complex or highly complex passage (e.g., tracing a subtle shift in point of view in a literary passage; associating particular opinions with the individuals who hold them in an informational passage)
10. Determine the main purpose of a highly complex passage or of one of its paragraphs
11. Determine a claim or counterclaim in a highly complex argument
12. Synthesize information & ideas from a pair of highly complex passages
13. Make an accurate subtle or complex interpretation of data in an informational graphic, such as a table, graph, or chart
14. Draw a subtle or complex supportable connection between a graphic & its accompanying passage
15. Make a sophisticated decision relating to the structure of a paragraph or passage (e.g., using a clause to set up information when the content is complex, the language is challenging, & the linkage is subtle)
16. Use supporting information to develop a point or claim logically on the basis of a thorough understanding of a challenging context (e.g., indicating the last step in a complex sequence; including an example that is similar in content to one or more other examples in a paragraph)
17. Sharpen the focus of a paragraph or passage by making a sophisticated decision about adding, revising, or deleting information or ideas (e.g., adding or retaining optional but relevant material because it enhances meaning & clarity)
18. Interpret, paraphrase, or summarize data in a complex informational graphic, such as a table, graph, or chart, & incorporate them in a passage in an accurate, relevant way
19. Make a sophisticated word or phrase choice based on highly developed vocabulary

knowledge & a thorough understanding of a challenging context (e.g., distinguishing among uncommon words that have similar denotations but differing connotations or uses when the distinctions are subtle)

20. Make a sophisticated decision about style & tone in a passage based on a thorough understanding of the context (e.g., achieving a subtle rhetorical aim, such as closely matching a sentence pattern already established in a passage)
21. Combine sentences to accomplish a subtle or complex purpose (e.g., drawing on an understanding of the context to place a blended sentence's emphasis on its most important idea)
22. Draw logical comparisons using complex terms (e.g. using complex allusions, analogies, metaphors, similes).
23. Use conventional expression in a highly challenging situation (e.g., selecting appropriately between uncommon words that are frequently confused, such as *defuse* & *diffuse*)
24. Compose or correct highly complex sentences, paragraphs, passages and texts using conventional English and rhetorical and composition elements including but not limited to:
    a. Using complex sentence structure
    b. Maintaining subject-verb or pronoun-antecedent agreement in a highly challenging situations
    c. Using a semicolon and colon effectively
    d. Ordering sentences in a paragraph to convey or address a subtle or complex issue of logic and cohesion
    e. Using a transitional word, phrase, clause, or sentence effectively to establish a subtle or complex logical relationship between sentences or paragraphs
    f. Minimizing wordiness or redundancy within a sentence or between sentences & paragraphs

ILLINOIS STATE BOARD OF EDUCATION

PERFORMANCE LEVEL DESCRIPTORS

Introduction to the Performance Level Descriptors for Mathematics

The Mathematics portion of the SAT focuses on a range of knowledge and skills needed to formulate and solve problems with and without context. The SAT Math Test measures students' range of ability in the areas of fluency, conceptual understanding, and application. Fluency requires students to solve problems accurately, efficiently, and strategically. Conceptual understanding requires students to demonstrate their understanding of mathematics concepts, operations, and relations. Application requires students to analyze situations and to represent and solve problems mathematically.

The Mathematics Performance Level Descriptors (PLDs) are aligned to the Common Core State Standards (CCSS) and are therefore based on the progression of those standards. The PLDs are text descriptions of the fundamental skills and knowledge demonstrated by students in each category of achievement.

PERFORMANCE LEVEL DESCRIPTORS

GRADE 11: Mathematics

Level 4 Exceeds Standards	Level 3 Meets Standards	Level 2 Approaching Standards	Level 1 Partially Meets Standards
The student has exceeded the proficiency level and demonstrates a **thorough** understanding of and ability to apply the mathematics knowledge and skills relative to the CCSS Mathematics content standards. The student can solve problems that call for a range of strategies, accurate and insightful reasoning, and connecting different areas of mathematics.	The student has met the proficiency level and demonstrates an **adequate** understanding of and ability to apply the mathematics knowledge and skills needed relative to the CCSS Mathematics content standards. The student can solve problems that call for use of strategies and accurate reasoning accurately applied in different areas of mathematics.	The student is approaching the proficiency level and demonstrates an **incomplete** understanding of and ability to apply the mathematics knowledge and skills relative to the CCSS Mathematics content standards. The student can solve problems that call for simple strategies and reasoning applied to basic areas of mathematics.	The student has partially met standards and demonstrates a **minimal** understanding of and ability to apply the mathematics knowledge and skills needed relative to the CCSS Mathematics content standards. The student can solve some problems that require applying simple strategies to basic areas of mathematics but without an understanding of the reasoning behind the strategies.

PERFORMANCE LEVEL 1

Students at this level demonstrate knowledge of simple linear equations, mostly in one-step problems in context, and solve problems with given data displayed in graphs or tables. They can solve problems arising from familiar contexts, identify important quantities, and begin developing models. They can solve problems that involve simple logical reasoning with basic abstract concepts.

Students in this that level can:

1. Solve problems involving proportional relationships, ratios, rates, and units in a variety of contexts
2. Understand and use the fact that when two quantities are in a proportional relationship, if one changes by a scale factor, then the other also changes by the same scale factor
3. Solve problems involving unit conversion within measurement systems
4. Solve problems involving derived units or unit conversion between different measurement systems
5. Solve problems involving percentages in a variety of contexts. Examples include, but are not limited to, discounts, interest, taxes, tips, and percent increases and decreases for many different quantities
6. Analyze and solve systems two of linear equations given algebraically and graphically
7. Solve problems involving a system of two linear equations in a familiar context
8. Create and solve simple linear equations in one variable, but make only limited strategic use of algebraic structure, including writing equivalent equations
9. Solve problems involving a figure or an object that can be modeled by a geometric figure using given information such as length, area, surface area, or volume
10. Fluently select the correct area or volume formula for a given figure and calculate a specified value
11. Read and use information presented in simple tables or simple graphs
12. Understand and recognize linear and nonlinear functions but cannot differentiate between quadratic and exponential functions
13. Use the relationship between variables shown in a graph or a scenario to make predictions and conclusions given a simple context
14. Given an appropriate data set, determine the measures of center (mean, median) and spread (range, standard deviation) to summarize one set of data

PERFORMANCE LEVEL 2

Students at this level demonstrate some knowledge linear relationships in one and two variables, and systems of linear equations in two variables. They can apply, inconsistently, their knowledge of multiple representations of these relationships and the interpretation of these representations. They have some difficulty with multistep problems that require the use of several skills and concepts. Overall, students use a limited range of strategies needed to solve different types of problems. They can solve problems that require identifying key quantities as well as recognizing the need to supply and developing missing information. They can begin to identify logical assumptions within a model and produce partial justifications and explanations with the model.

Students in this level can (in addition to level 1):

1. Understand and use the relationship between percent change and growth factor (e.g., 5% and 1.05); include percentages greater than or equal to 100%
2. Solve multistep problems using percentages
3. Analyze data to make predictions and calculate probability concepts with context
4. Create and use linear equations in one and two variables to model and solve problems in a context
5. Interpret a term in a linear equation in one variable
6. Solve a linear equation in one variable, making strategic use of algebraic structure
7. Use structure and reasoning to solve simple rational, radical polynomial, and absolute value equations
8. Interpret the meaning of a term in a linear function in two variables that represents a context and/or explain how the variables are related
9. Write an equation for a line given different conditions (e.g., two points on the line or one point and the slope of the line)
10. Make connections between tabular, algebraic, or graphical representations of a linear equation in two variables, deriving one representation from the others
11. Create and use a system of two linear equations in two variables to solve problems in a context
12. Apply knowledge that changing by a geometric figure by a scale factor of (x) changes all lengths by a factor of (x) but does not change angle measures
13. Add, subtract, multiply, and factor polynomials
14. Use some concepts related to congruence and similarity of triangles to solve familiar problems
15. Apply knowledge of the relationship of angles (complementary, supplementary, corresponding, vertical, etc.) formed when a transversal cuts parallel lines
16. Interpret slope (rate of change) and intercept (constant term) of a linear model in the context of the data
17. Make strategic use of algebraic structure and the properties of operations to identify and create equivalent expressions (linear, quadratic, and exponential)
18. Read and interpret contextual information presented in a graph or table
19. Choose an appropriate graphical representation for a given data set
20. Interpret the effect of outliers on measures of center and spread
21. Interpret one- and two-way tables, tree diagrams, area models, and other representations to find relative frequency, probabilities, and conditional probabilities
22. Identify when sample data can be used to make inferences about the corresponding population
23. Construct chains of reasoning to justify a model used

PERFORMANCE LEVEL 3

Students at this level demonstrate some mastery of linear and nonlinear relationships and can apply linear and nonlinear relationships in different contexts. They can solve problems that require them to construct, justify, and reason with mathematical models in a variety of settings. They can compare the strengths and weaknesses of different models for the same setting, and they can support their reasoning in constructing and using mathematical models. They solve problems that require the use of strong problem solving methods (standard algorithms, mathematical reasoning) as well as questions that can be answered with weaker problem solving methods (trial and error).

Students in this level can (in addition to level 2):

1. Use properties of the number system to solve equations with rational exponents and make sense of the algebraic structure to solve these problems in multiple representations
2. Construct and use linear equations and inequalities in one or two variables to solve problems and interpret the solutions in a variety of contexts
3. Solve problems utilizing systems of linear equations and inequalities in a variety of contexts
4. Given a figure and a transformation, identify a transformed figure
5. Solve problems involving properties of right triangles
6. Use given geometric theorems and properties of rigid motions, lines, angles, triangles, and parallelograms to solve routine problems and to prove statements about angle measurement, triangles, distance, line properties, and congruence
7. Make strategic use of algebraic structure, the properties of operations, and reasoning to solve quadratic equations in one variable presented in a variety of forms (e.g., standard form, completing the square)
8. Derive, create, and solve equations or functions to model problems in a variety of contexts with multiple variables
9. Fluently solve quadratic equations in one variable, either written as a quadratic expression in standard form equal to zero or by completing the square
10. Determine the conditions under which a quadratic equation has no solution, one real solution, or two real solutions
11. Use structure and reasoning to solve rational, radical, quadratic, exponential, polynomial, and absolute value equations in one variable
12. Use function notation to represent and interpret input/output pairs in terms of a context and points on the graph
13. Interpret the meaning of an input/output pair, constant, variable, factor, or term of both linear and non-linear relationships based on a context
14. Solve problems involving quadratic or exponential relationships by identifying the equation or function, or by creating and using the function
15. Make connections between tabular, algebraic, and graphical representations of equations, functions and inequalities by:
    - when given one representation, selecting another representation
    - identifying features of one representation given another representation, including maximum and minimum values of the function
    - determining how a graph is affected by a change to its equation, including a vertical shift or scaling of the graph
16. Solve multi-step problems that require factoring a polynomial or simple rational function
17. Represent and interpret input/output pairs of a linear or non-linear function in terms of a context and points on a graph; and present the solutions, intercepts, and key features in terms of a context
18. Use the mean and standard deviation of a data set to fit the set to a normal distribution
19. Represent data on two quantitative variables in graphs, tables, and scatterplots, and describe how the variables are related
20. Construct and interpret one- and two-way tables, tree diagrams, area models, and other representations to find relative frequency, probabilities, and conditional probabilities

21. When given a linear, exponential, or quadratic relationship between two variables, fit a function to the data in order to solve multistep contextual problems
22. Determine whether a sample survey, experiment, or observational study is most appropriate
23. Determine appropriate representations of categorical and quantitative data, summarizing and interpreting the data and characteristics of the representations
24. Describe and interpret possible associations and trends in given data
25. Recognize and determine conditional probability and independence in contextual problems
26. Know, apply, and prove relevant theorems such as
    - the vertical angle theorem
    - triangle similarity and congruence criteria
    - triangle angle sum theorem
    - the relationship of angles formed when a transversal cuts parallel lines
27. Determine which statements may be required to prove certain relationships or to satisfy a given theorem

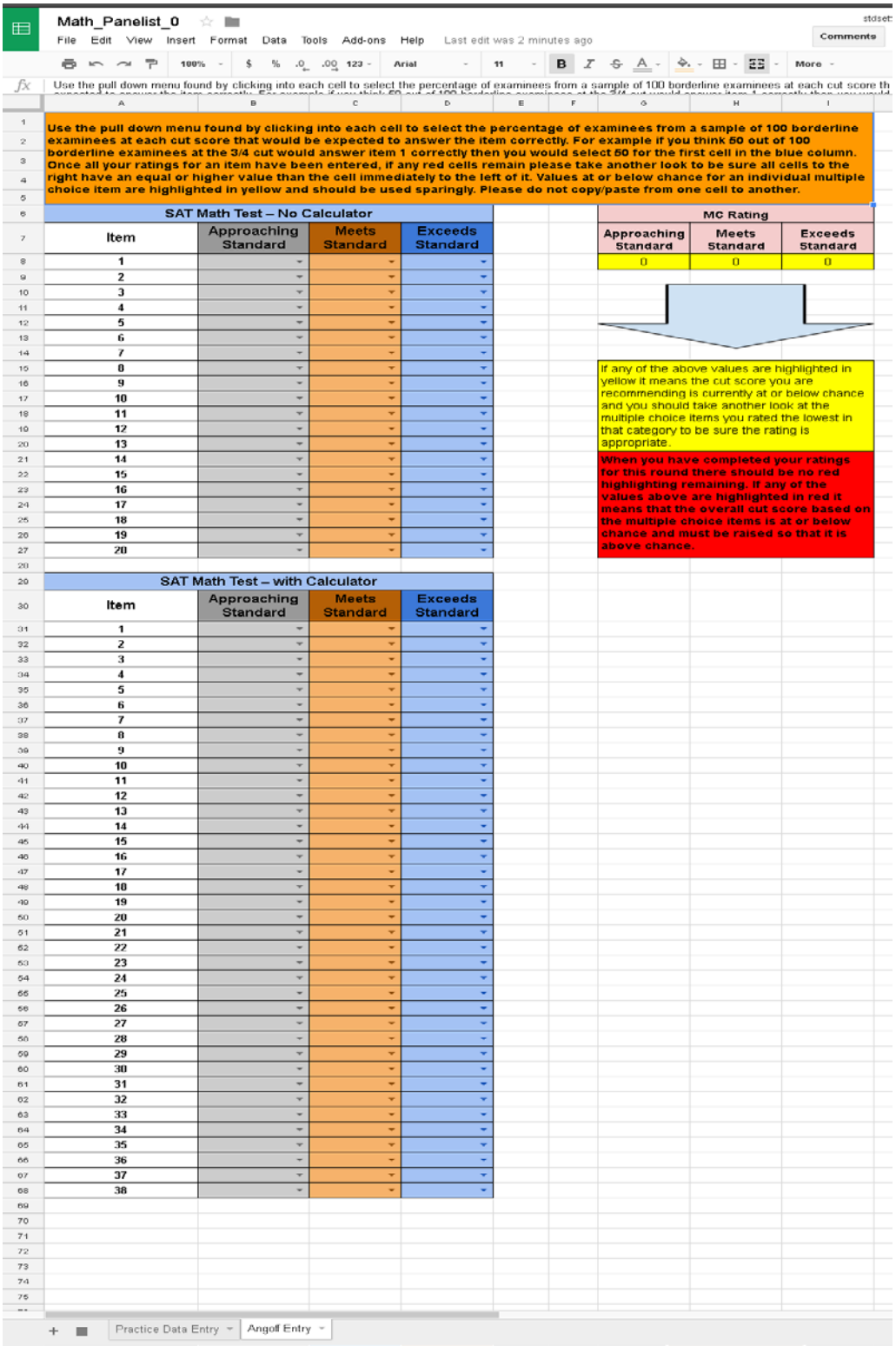
PERFORMANCE LEVEL 4

Students at this level demonstrate a strong command of reasoning and of the previous mathematics relationships and skills described in the first three performance levels. Students can fluently solve multistep contextualized problems that integrate more than one standard. Overall, students effectively use a range of strategies and reasoning to solve a variety of problem types. They can solve unfamiliar problems by insightful, creative use of models. They can identify the logical assumptions in models; they can analyze, apply, interpret, and justify models with accurate, careful reasoning; and they can compare multiple plausible approaches to modeling in a setting and thoughtfully choose the most appropriate model.

Students at this level can (in addition to level 3):

1. Solve problems involving the Pythagorean theorem, right triangle trigonometry, trigonometric ratios, and properties of special right triangles in a variety of contexts
2. Solve problems involving definitions, properties, and theorems relating to circles and parts of circles, such as radii, diameters, tangents, angles, arcs, arc lengths, and sector areas
3. Create an equation of the form $(x - h)^2 + (y - k)^2 = r^2$ to represent a circle in the xy-plane, and describe how the change to the equation representing a circle in the xy-plane affects the graph of the circle or vice versa
4. Complete the square in an equation representing a circle or a parabola to determine properties of the circle when graphed in the xy-plane
5. Use the distance formula in problems related to circles and related to focus and directrix in parabolas
6. Understand geometric constructions: copying a segment, copying an angle, bisecting an angle, bisecting a segment, and including the perpendicular bisector of a line segment
7. Specify a sequence of transformations that will carry a geometric figure onto another figure
8. Given a trigonometric value and quadrant for an angle, utilize the structure and relationships of trigonometry, including relationships in the unit circle, to identify other trigonometric values for the given angle
9. Apply knowledge and understanding of the complex number system to add, subtract, multiply, and divide with complex numbers and solve problems
10. Estimate and interpret the slope and intercepts of the line of best fit for a given scatterplot in a context
11. Fluently apply linear and nonlinear relationships (quadratic and exponential) to model and solve problems, and make estimates that do not involve an exact number for either growth or decay
12. Create, solve, and interpret nonlinear functions and demonstrate how changes in parameters can affect their models, both algebraically and graphically
13. Make strategic use of algebraic structure and the properties of operations to identify and create equivalent expressions (rational, radicals and rational expressions)
14. Identify, construct, and use logarithmic and trigonometric functions to model a relationship between quantities and solve problems
15. Solve systems of nonlinear equations in two variables
16. Identify the effects of transformations on graphs of linear and non-linear functions and determine the end behavior of polynomials
    Apply the addition rule of probability

Appendix D: Operational Rating Sheet



Math_Panelist_0

File Edit View Insert Format Data Tools Add-ons Help Last edit was 2 minutes ago

Comments

Use the pull down menu found by clicking into each cell to select the percentage of examinees from a sample of 100 borderline examinees at each cut score that would be expected to answer the item correctly. For example if you think 50 out of 100 borderline examinees at the 3/4 cut would answer item 1 correctly then you would select 50 for the first cell in the blue column. Once all your ratings for an item have been entered, if any red cells remain please take another look to be sure all cells to the right have an equal or higher value than the cell immediately to the left of it. Values at or below chance for an individual multiple choice item are highlighted in yellow and should be used sparingly. Please do not copy/paste from one cell to another.

SAT Math Test – No Calculator

Item	Approaching Standard	Meets Standard	Exceeds Standard
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			
11			
12			
13			
14			
15			
16			
17			
18			
19			
20			

SAT Math Test – with Calculator

Item	Approaching Standard	Meets Standard	Exceeds Standard
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			
11			
12			
13			
14			
15			
16			
17			
18			
19			
20			
21			
22			
23			
24			
25			
26			
27			
28			
29			
30			
31			
32			
33			
34			
35			
36			
37			
38			

MC Rating

Approaching Standard	Meets Standard	Exceeds Standard
0	0	0

If any of the above values are highlighted in yellow it means the cut score you are recommending is currently at or below chance and you should take another look at the multiple choice items you rated the lowest in that category to be sure the rating is appropriate.

When you have completed your ratings for this round there should be no red highlighting remaining. If any of the values above are highlighted in red it means that the overall cut score based on the multiple choice items is at or below chance and must be raised so that it is above chance.

Practice Data Entry | Angoff Entry

Appendix E: Training Evaluation Form

Panelist's ID#:_____

Training /Ready to Proceed Evaluation

SAT Evidence-Based Reading and Writing

The purpose of this form is to verify whether you understand the general purpose of the standard setting study and believe that you have received sufficient information and explanation to make your standard setting judgments.

	Yes	No
I understand the purpose of the standard setting study.		
I understand the steps I am to follow to make my standard setting judgments.		
I understand the concept of the borderline examinee.		
I am ready to complete my standard setting judgments.		

If you responded "No" to any of these statements, please indicate what additional information or explanations you need.

_____

_____

_____

_____

_____ _____

(Date) (Signature)

_____

(Print Name)

Appendix F: Round 1 Evaluation

Round 1 Evaluation – SAT Evidence-Based Reading and Writing Standard Setting

For questions 1-6, please indicate below the degree to which you agree with each of the following statements.

	Strongly Disagree	Disagree	Agree	Strongly Agree
1. I understand the purpose of the study.				
2. The instructions and explanations provided by the facilitator were clear.				
3. The training in the standard setting method gave me the information I needed to complete my assignment.				
4. The ALDs that were developed prior to the meeting were accurate.				
5. I understand the concept of the borderline examinee.				
6. The ALDs helped me to determine how to rate each item.				

For questions 7-11, indicate how **influential** each of the following factors was in completing the item ratings.

	Not Influential	Influential	Very Influential
7. Completing the test before beginning the task.			
8. My perception of the difficulty of the items			
9. The ALDs			
10. The consequences of the test for students			
11. My experience with students in my classroom			

12. Please identify any additional factors you considered in making your Round 1 ratings that were not included above._____

_____

13. Do you have other comments that you would like to share at this time (use back of paper if needed)?_____

_____

Appendix G: Final Evaluation Form

Final Evaluation Form for SAT Evidence-Based Reading and Writing

For questions 1-15, please indicate below the degree to which you agree with each of the following statements.

	Strongly Disagree	Disagree	Agree	Strongly Agree
1. I understood the purpose of the study.				
2. The instructions and explanations provided by the facilitator were clear.				
3. The training on the standard setting method gave me the information I needed to complete my assignment.				
4. The PLDs that were developed prior to the meeting were accurate.				
5. I understood the concept of the borderline examinee.				
6. The PLDs helped me to determine how to rate each item.				
7. It was beneficial to have an opportunity for table discussions between rounds.				
8. It was beneficial to have an opportunity to review feedback between rounds.				
9. The impact data showing the percent of students expected to place into each category **based on my table's** cut scores made a difference in how I rated the items in round 3.				
10. The impact data showing the percent of students expected to place into each category **based on all panelists in the room's** cut scores made a difference in how I rated the items in round 3.				
11. The opportunity to make more than 1 round of ratings (i.e., round 2) helped me to feel more confident about my final ratings.				
12. I felt engaged in the process.				
13. I was comfortable sharing my ideas with the other panelists during the discussions.				
14. I am confident this standard setting process will produce fair cut scores.				
15. I would be comfortable defending this process to my peers.				

(OVER)

16. If you answered Disagree or Strongly Disagree to any of the above, please give us feedback on what could have been done differently to change this outcome. Please attach additional pages, if needed.

_____

_____

_____

For questions 17-28 indicate how **influential** each of the following factors was in completing your ratings.

	Not Influential	**Influential**	**Very Influential**
17. Completing the test before beginning the task.			
18. My perception of the difficulty of the items			
19. The actual item difficulty provided for each item			
20. Distributions of students expected to earn each performance level			
21. Table discussion after Round 1			
22. Table discussion after Round 2			
23. Large group discussion after Round 2			
24. The average ratings of other panelists at my table			
25. The average ratings of all panelists in the room			
26. The PLDs			
27. The consequences of the exam for students			
28. My experience with students in my classroom			

(Next Page)

29. Were there other factors that influenced your ratings in any of the three rounds that were not described above? If yes, please describe the other factors. Please attach additional pages, if needed.

_____

_____

_____

_____

For questions 30-38, first indicate how **useful** each of the following materials or procedures was in completing the exercises over the past couple of days.

	Not Useful	Useful	Very Useful
30. Taking the exam prior to beginning the task			
31. Practicing the procedure with real items prior to beginning the actual rating task			
32. Referencing the PLDs			
33. Table discussion after round 1			
34. Table discussion after round 2			
35. Test overview			
36. Actual item difficulty values			
37. Distribution of students earning each performance level			
38. Large group discussion after round 2			

(OVER)

39. Can you think of other information that would have been useful to you during the standard setting process in helping you make your ratings? If yes, Please indicate what other information you would have found useful. Please attach additional pages, if needed.

_____

_____

For questions 40-49, please indicate how appropriate the amount of time was to complete the different components of the standard setting task.

	Too Little Time	About Right	Too Much Time
40. Taking the test			
41. Reviewing the PLDs			
42. Training on the rating task before Round 1			
43. Round 1 of the rating task			
44. Table discussion after round 1			
45. Round 2 of the rating task			
46. Table discussion after round 2			
47. Large group discussion after round 2			
48. Review of impact data for the total group			
49. Review of impact data for my table			

(Next Page)

50. Please share any comments that you may have on the items, standard setting process, meeting logistics, or other topic of interest.

Appendix H: Example Individual-Panelist Feedback Form

2017 SAT ERW - Round 1 Results - Panelist 1								
		Your Table			Total Group			
Multiple Choice Questions	Your Recommendation	Recommendation	MIN	MAX	Recommendation	MIN	MAX	Total Number of Items on Test
Number of MCQ items (out of 96) the student must get correct to be placed into the Approaching category:	49.50	34.95	33.30	49.50	37.00	30.25	52.65	96
Number of MCQ items (out of 96) the student must get correct to be placed into the Meets category:	67.20	59.40	56.05	67.20	56.83	43.85	70.50	96
Number of MCQ items (out of 96) the student must get correct to be placed into the Exceeds category:	76.00	79.00	71.65	85.45	74.90	58.00	85.45	96

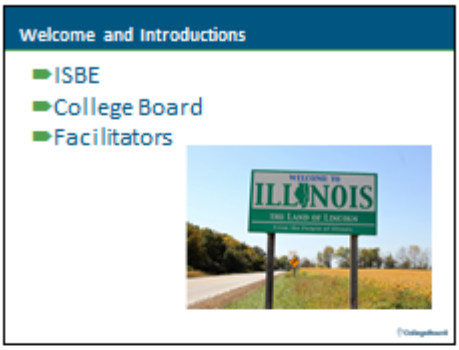
Appendix I: Item-Level Feedback Round 1 - ERW

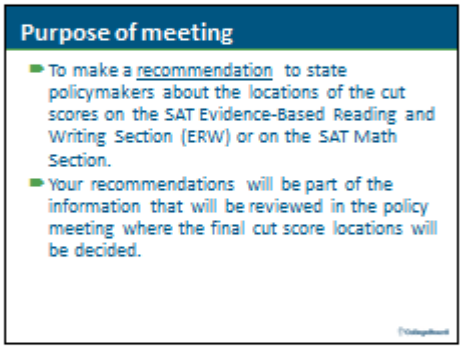
Round 1 Individual Questions Feedback for the SAT ERW				
	Median Rating at Each Cut Score - Total Group			
Item No.	Approaching	Meets	Exceeds	Item Difficulty (Larger = Easier)
RD 1	52.5	75	90	0.92
RD 2	50	72.5	90	0.74
RD 3	40	60	85	0.64
RD 4	30	52.5	80	0.54
RD 5	30	50	72.5	0.48
RD 6	35	50	75	0.61
RD 7	45	65	87.5	0.65
RD 8	37.5	57.5	80	0.64
RD 9	37.5	62.5	85	0.67
RD 10	50	75	90	0.82
RD 11	40	60	80	0.67
RD 12	30	55	75	0.54
RD 13	32.5	50	75	0.52
RD 14	30	50	72.5	0.50
RD 15	40	60	82.5	0.36
RD 16	35	60	75	0.35
RD 17	40	60	80	0.36
RD 18	40	62.5	85	0.57
RD 19	30	47.5	70	0.38
RD 20	30	42.5	65	0.46
RD 21	35	60	80	0.60
RD 22	40	60	75	0.55
RD 23	32.5	52.5	75	0.34
RD 24	35	62.5	80	0.62
RD 25	45	65	85	0.38
RD 26	37.5	50	75	0.44
RD 27	32.5	50	70	0.47
RD 28	45	67.5	87.5	0.70
RD 29	40	60	80	0.70
RD 30	60	75	90	0.89
RD 31	32.5	50	75	0.53
RD 32	30	50	70	0.43
RD 33	35	52.5	75	0.24
RD 34	40	57.5	80	0.69
RD 35	40	57.5	77.5	0.38
RD 36	40	60	82.5	0.73
RD 37	35	52.5	77.5	0.23
RD 38	30	50	75	0.39
RD 39	40	60	77.5	0.38

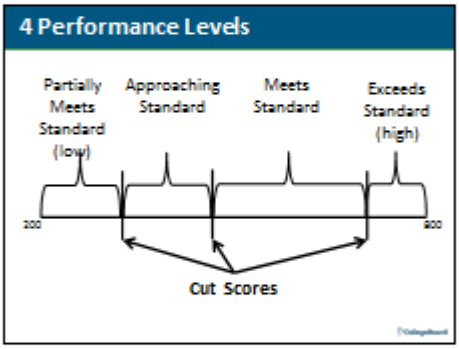
Round 1 Individual Questions Feedback for the SAT ERW				
	Median Rating at Each Cut Score - Total Group			
Item No.	Approaching	Meets	Exceeds	Item Difficulty (Larger = Easier)
RD 40	30	50	72.5	0.41
RD 41	30	50	72.5	0.36
RD 42	30	47.5	67.5	0.38
RD 43	40	60	75	0.34
RD 44	30	50	72.5	0.33
RD 45	30	50	70	0.29
RD 46	30	45	70	0.39
RD 47	30	45	72.5	0.15
RD 48	52.5	75	90	0.75
RD 49	40	55	75	0.42
RD 50	37.5	52.5	72.5	0.41
RD 51	50	70	85	0.48
RD 52	35	47.5	72.5	0.36
WL 1	50	72.5	90	0.88
WL 2	40	55	80	0.56
WL 3	40	60	80	0.63
WL 4	50	70	90	0.60
WL 5	50	72.5	85	0.69
WL 6	42.5	62.5	85	0.67
WL 7	42.5	65	82.5	0.58
WL 8	45	65	85	0.79
WL 9	45	65	90	0.58
WL 10	35	50	75	0.37
WL 11	40	65	87.5	0.60
WL 12	50	65	85	0.72
WL 13	35	55	80	0.43
WL 14	30	50	72.5	0.43
WL 15	45	65	87.5	0.70
WL 16	40	60	77.5	0.76
WL 17	42.5	65	85	0.46
WL 18	45	70	87.5	0.54
WL 19	45	67.5	85	0.71
WL 20	35	57.5	75	0.35
WL 21	35	55	75	0.50
WL 22	30	50	75	0.32
WL 23	50	75	90	0.56
WL 24	50	72.5	85	0.45
WL 25	47.5	65	85	0.69
WL 26	35	65	85	0.50

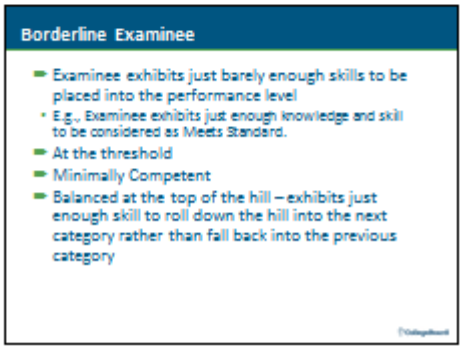
Round 1 Individual Questions Feedback for the SAT ERW				
	Median Rating at Each Cut Score - Total Group			
Item No.	Approaching	Meets	Exceeds	Item Difficulty (Larger = Easier)
WL 27	40	60	80	0.54
WL 28	37.5	60	80	0.72
WL 29	45	67.5	85	0.61
WL 30	30	52.5	75	0.41
WL 31	40	57.5	80	0.28
WL 32	45	67.5	85	0.44
WL 33	35	57.5	80	0.66
WL 34	35	50	75	0.30
WL 35	40	60	80	0.52
WL 36	37.5	60	80	0.53
WL 37	40	65	80	0.65
WL 38	40	60	80	0.56
WL 39	32.5	55	80	0.24
WL 40	35	57.5	75	0.57
WL 41	40	60	80	0.41
WL 42	35	55	75	0.31
WL 43	35	60	77.5	0.24
WL 44	42.5	65	85	0.70

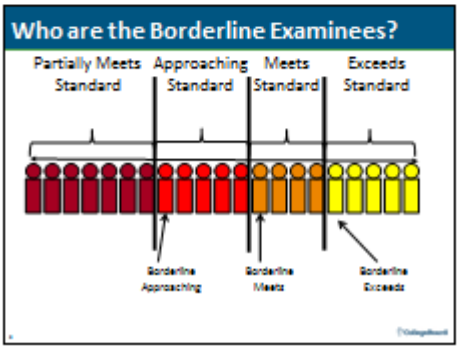
Appendix J: Opening Session Slides











Performance Level Descriptors –PLDs

- Describe what students in each performance level should reasonably know and be able to do
- Relate directly to the knowledge, skills, and abilities a student should know and be able to demonstrate to earn each performance level
- Distinguish clearly from one level (Approaching) to the next (Meets)
- Represent the performance of the **Borderline Examinee** at each performance level (except for PLD 1 which is not at the Borderline)

Modified Angoff Method (the standard setting task)

- For each item
  - Consider 100 borderline examinees at each Performance Level.
  - Consider how difficult you believe the item will be for each group of 100 borderline examinees and what about the item makes it more or less difficult
  - For each group of 100 borderline examinees separately, record the number of examinees that you would expect to answer the item correctly based on the PLDs.

Considerations

- When making ratings remember to take into consideration and refer back to the following information:
  - The Performance Level Descriptors (PLDs) for each performance level
  - The knowledge, skills, and abilities required to answer the item correctly.
  - The Borderline Examinee at each performance level

You are important!

- Each of you was invited because we value your opinion, expertise, and experience. Feel free to share your thoughts with your group during the designated discussion times but remember that the actual assignment of item ratings is an individual task and should be completed independently.
- Discussion is valuable to ensure that all perspectives are represented. Diversity is valued and consensus is not a goal of the discussion.

Agenda

- Start time is 8:25am
- Breakfast will be available beginning at 7:15
- Other than start time, the times found on the agenda are estimates and will remain fluid as needed to accomplish the work that is needed.
- Please remain flexible about the time we end each day as it is hard to predict how long many activities will take and participation from everyone is needed at each step.
- Dress comfortably and consider bringing a sweater or dressing in layers in case the room temperature does not match your preference – we will try, but it is hard to please everyone, so be prepared.
- Please take notes during the meeting to help us document the process.

Breakout Rooms

- ELA – Conference Center #4 – Pamela Kaliski
- Math – Conference Center #2 – Lei Wan

- Lunch – Prairie Room
- Breaks – Outside Meeting Rooms

THANK YOU!!

Appendix K: Math Slides

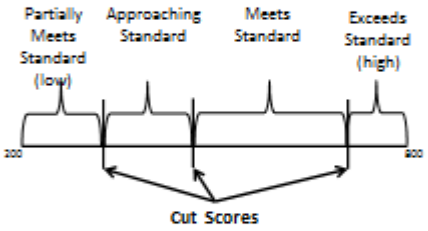
Meeting Logistics

- Introduction of College Board staff members
- Please introduce yourself to the group
  - Name, Position, etc.
- Agenda – all times are approximate (white)
- Bathrooms, Breaks, etc.
- Confidentiality of materials; Panelist unique ID
- Discussion outside the room
- Cell phones should be off or silenced and used only outside of this room.
- Table leaders
- Complete confidentiality form (white) and bio form (online)

Purpose of meeting

- To make a recommendation to state policymakers about what the locations of the cut scores should be on the SAT Math
- 4 performance levels (Partially Meets Standard, Approaching Standard, Meets Standard, and Exceeds Standard); 3 cut score locations will need to be recommended.

4 Performance Levels



The next 2 days…

- Overview of SAT
- Experience the SAT Math Test
- Introduce the Borderline Examinee
- Familiarize with Performance Level Descriptions (PLDs)
- Training and practice on the Standard Setting Rating Task
- Make your ratings/discussion (Round 1, Round 2, Round 3)
- Final Evaluation and wrap-up

Overview of SAT

Experience the Exam

- You have 1 hour to take the SAT Math section (operational time is 80 minutes)
  - Plan to spend 15 minutes on the No Calculator items, and 45 minutes on Calculator items
  - Time will be announced at 15 minutes and again at 5 minutes before time is done.
  - Do your best but don't agonize over any one question since you need to experience all the questions in the time allowed.
- When done let us know by standing up name tent and we will distribute the answer key so you can score yourself. Also, if everyone completes early this will let us move on to the next activity.

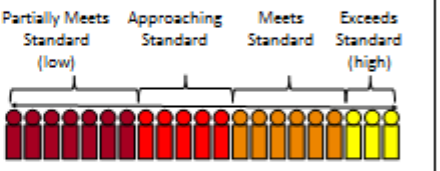
Review of Items
(Feel free to take notes in your test booklet)

- How did you do on the items? Any perfect scores?
- What are your general thoughts on the exam?
- Did you find any items easier or more difficult than you expected? What do you think causes the items to be easier or more difficult?
- What knowledge, skills, and abilities are needed to answer the items correctly?

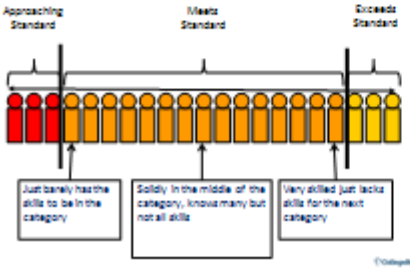
Performance Level Descriptors –PLDs

- Describe what students in each performance level should reasonably know and be able to do
- Relate directly to the knowledge, skills, and abilities a student should know and be able to demonstrate to earn each performance level
- Distinguish clearly from one level (Approaching) to the next (Meets)
- Represent the performance of the **Borderline Examinee** at each performance level (except for PLD 1 which is not at the Borderline)

Range of Performance on the Test

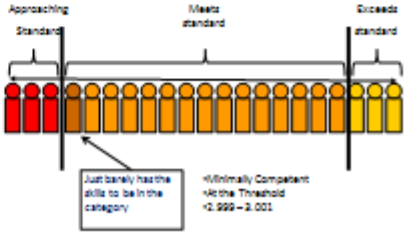


Range of Performance Within Level

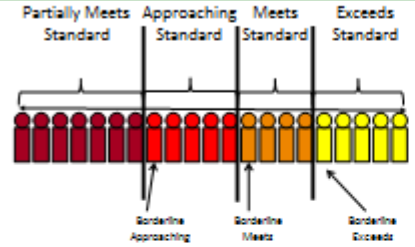


Borderline Student–

Lowest Level of Knowledge, Skills & Abilities That Still Belongs In the Level



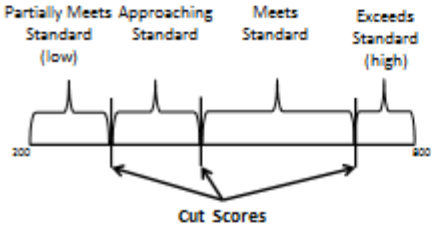
Borderline Examinee at Each Level



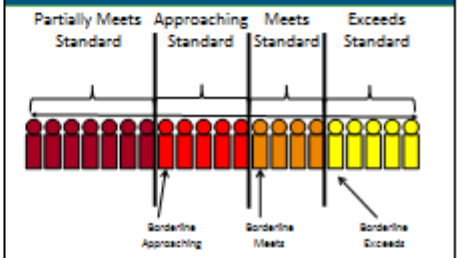
Turn over to…

- You are familiar with Math section of the exam at this point, and have begun to conceptualize the borderline examinee.
- Now you will hear more about the knowledge, skills, and abilities that are necessary for students receiving each SAT performance level to demonstrate on the Math section.
- Bill will review PLDs with panelists

4 Performance Levels



Borderline Examinee at Each Level



Modified Angoff Method (the standard setting task)

- For each item
  - Consider 100 borderline examinees at each performance Level.
  - Consider how difficult you believe the item will be for each group of 100 borderline examinees and what about the item makes it more or less difficult – vocabulary, content, distractor, etc.
  - For each group of 100 borderline examinees separately, record the number of examinees that you would expect to answer the item correctly based on the PLDs.

Thought Process

- Given the item, how difficult it is, and the PLDs, what proportion of borderline examinees at the **Meets Standard** cut would answer the item correctly?
- Given the item, how difficult it is, and the PLDs, what proportion of borderline examinees at the **Approaching Standard** cut would answer the item correctly?
- Given the item, how difficult it is, and the PLDs, what proportion of borderline examinees at the **Exceeds Standard** cut would answer the item correctly?

Acceptable Range of Responses

- Ratings on MC items can be any number between 10 and 95 in increments of 5:
  10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95
- Ratings on Grid-in items can be 5 to 95 by 5
- Ratings for higher cut scores should be equal to or higher than the ratings for lower cut scores
- Use the entire range as appropriate. It may help to think about the distractors and how many you think each group of borderline examinees would be able to eliminate based on the PLDs.

Chance – MC items

- The SAT MC items are 4 choice items.
- Students would be expected to get 25% correct by chance if they were strictly guessing.
- To be placed into Approaching Standard, students should have to do more than just guess so the number of students that you select should usually be higher than 25.
- A value lower than 25 (20, 15, and 10) can be used when you believe it is absolutely necessary but should be used sparingly since it indicates the students are expected to do even worse than guessing the answer correctly.
- Values of 25 and lower will turn yellow on your data entry sheet as a reminder that these are values at or below chance.

Chance – Grid-in items

- Students produce responses for SAT Grid-in items.
- Students would be expected to have no chance of guessing.
- However, to recognize that students with partial knowledge may have low chance of getting the item correctly, we allow the lowest probability to be 5 for a Grid-in item.

High End Rating

- For both MC and Grid-in items, the highest rating possible is 95, in recognition that perfect performance is not common, not a reasonable expectation for borderline students.

Recording the ratings

- First write your rating for each item at each cut score on the paper form provided.
- Double check that numbers to the right are equal to or larger than numbers to the left and no cells are left blank. (practice data entry tab in Google docs workbook)
- Use the **pull down menus** or **type directly into the cell** to enter your ratings into the spreadsheet.
- Do not copy and paste.
- When finished you should have no bright red cells or blank cells in the spreadsheet.
- Let room staff know when you are done entering your ratings by standing your name tent on end.

Make Practice Ratings

- Using the practice rating form (green), make ratings for items 1 and 16 in the no calculator section and items 1 and 31 in the calculator section.
- Remember to reference your PLDs for the borderline examinee at each cut score.
- Write your ratings on the paper form and then enter those ratings into the practice data entry form on your laptop.
- Please let us know if you have any questions or need help accessing the data entry form.
- Please stand your name tent on end when you are finished.

Discuss Practice Ratings

- Identify the range of ratings based on the practice.
- Discuss discrepant ratings and rationales
- Additional questions about the task of providing ratings?
- Training/Ready to Proceed Evaluation Form (blue)

Provide Round 1 Ratings

- **Multiple Choice Items**
  - For each cut score, provide the number of borderline examinees out of 100 that would be expected to answer each item correctly.
  - **Reference the PLDs to make Ratings.**
- After data entry is completed, stand name tent on end.
- Complete Round 1 Evaluation Form (online)
- Check materials back in and be sure facilitator checks that all ratings are present and dismisses you before leaving.

Welcome Back

Day 2

Round 1 Table Discussion

- Review item level feedback form
- Identify items where your table differed most.
- Discuss items with the Most Discrepant ratings first. Provide rationales for why one person rated higher while another rated lower.
- Consider the item difficulty rating and how it compares to your perception of the item.
- Review individual summary and discuss how your cut score recommendations differ from each other and from the total group.

Round 2 Ratings

- Round 2 gives you an opportunity to change your ratings based on the additional information you have received during the table discussion.
- You may change any or all of your ratings. You may also choose to change none of your ratings. Only record in the Round 2 column the ratings that you are changing. Be sure to make the same changes to your data entry spreadsheet!
- Remember to always reference the PLDs
- When finished please stand your name tent on end.

Round 2 Table Discussion

- Review item level feedback form for your table and compare to the form for the room.
- Identify items where your table differed most in how they rated the items.
- Review individual summary and discuss how your cut score recommendations differ from each other and from the total group.
- Keep notes of key points to share in large group discussion

Round 2 Large Group Discussion

- Discuss most discrepant placements and share discussion from table groups.
- Be sure to reference PLDs in the rationales.

Round 2 Discussion

- Present Impact Data

Achievement Level	Percent	
Exceeds	5.34	24.78%
Meets	19.44	
Approaching	34.01	75.22%
Partially Meets	41.21	

- Discuss impact data in tables.

Round 3 Ratings

- Round 3 gives you a final opportunity to change your ratings based on the additional information you have received during the Round 2 discussion.
- You may change any or all of your ratings. You may also choose to change none of your ratings. Only record in the Round 3 column the ratings that you are changing. Be sure to make the same changes to your data entry spreadsheet!
- Remember to always reference the PLDs
- When finished please stand your name tent on end.

Round 3 Debrief

- Present Impact Data

Achievement Level	Cut Score	Percent	
Exceeds	690	6.15	36.85%
Meets	540	30.70	
Approaching	450	32.75	63.15
Partially Meets	NA	30.40	

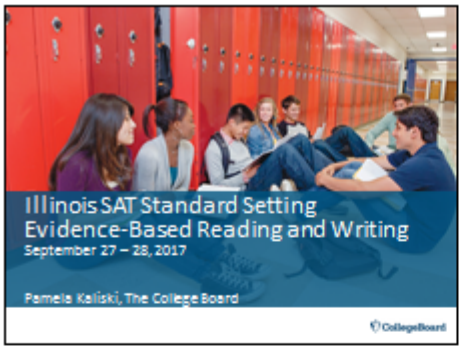
Final Thoughts

- Are there any questions or comments that you would like to share at this time before we end the meeting?

Check out process

- Have you completed your final evaluation form? If not, please do so (link in pdf on your laptop)
- Turn in materials at the "check out station" at the front of the room:
  - Angoff rating form (yellow)
  - Item-level feedback forms (white paper with p-values)
  - Individual cut score feedback (white paper with cut scores)
  - Impact data feedback
  - PLDs
- Check the sign out sheet. Don't leave until your materials have been officially checked back in.
- You can either leave everything else on your desk or take it with you.

Appendix L: ERW Slides



Meeting Logistics

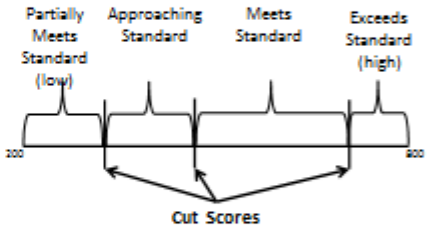
- Introduction of College Board staff members
- Please introduce yourself to the group
  - Name, Position, etc.
- Agenda – all times are approximate (white)
- Bathrooms, Breaks, etc.
- Confidentiality of materials
- Discussion outside the room
- Cell phones should be off or silenced and used only outside of this room.
- Complete confidentiality form (white) and bio form (online/green)
- Panelist unique ID number, table leaders

Purpose of meeting

- To make a recommendation to state policymakers about what the locations of the cut scores should be on the SAT Evidence-Based Reading and Writing Section (ERW).
  - Use your expertise, knowledge, and experience with students along with background and reference information that we will provide to you over the course of the two days.
- 4 Achievement levels (Partially Meets Standard, low – Exceeds Standard, high); 3 cut score locations will need to be recommended.

(Note: Please be mindful of parking lot issues; record them separately)

4 Performance Levels



The next 2 days…

- Overview of SAT
- Experience the SAT Evidence-based Reading & Writing section (ERW)
- Introduce the Borderline Examinee
- Familiarize with Performance Level Descriptions (PLDs)
- Training and Practice on the Standard Setting Rating Task
- Make your ratings/discussion (Round 1, Round 2, Round 3)
- Final Evaluation and wrap-up

Jim Patterson will provide test overview

Experience the Exam

- You have 1.5 hours to take the SAT ERW section (operational time is 100 minutes)
  - Plan to spend 60 minutes on Reading, and 30 minutes on Writing and Language
  - Time will be announced at 1 hour and again at 10 minutes before time is done.
  - Do your best but don't agonize over any one question since you need to experience all the questions in the time allowed.
- When done let us know by standing up name tent and we will distribute the answer key so you can score yourself. Also, if everyone completes early this will let us move on to the next activity.

Review of Items
(Feel free to take notes in your test booklet)

- How did you do on the items? Any perfect scores?
- What are your general thoughts on the exam?
- Did you find any items easier or more difficult than you expected? What do you think causes the items to be easier or more difficult?
- What knowledge, skills, and abilities are needed to answer the items correctly?
- Reminder of parking lot topics and note taking

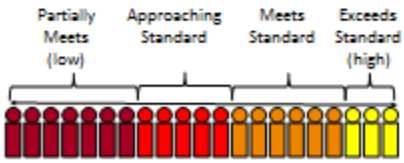
The next 2 days…

- ~~Overview of SAT~~
- ~~Experience the SAT Evidence-based Reading & Writing section (ERW)~~
- Introduce the Borderline Examinee
- Familiarize with Performance Level Descriptions (PLDs)
- Training and Practice on the Standard Setting Rating Task
- Make your ratings/discussion (Round 1, Round 2, Round 3)
- Final Evaluation and wrap-up

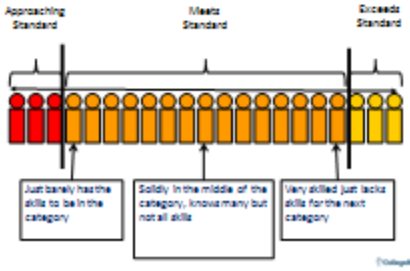
Performance Level Descriptors –PLDs

- Describe what students in each performance level should reasonably know and be able to do
- Relate directly to the knowledge, skills, and abilities a student should know and be able to demonstrate to earn each performance level
- Distinguish clearly from one level (Approaching) to the next (Meets)
- Represent the performance of the **Borderline Examinee** at each performance level (except for PLD 1 which is not at the Borderline)

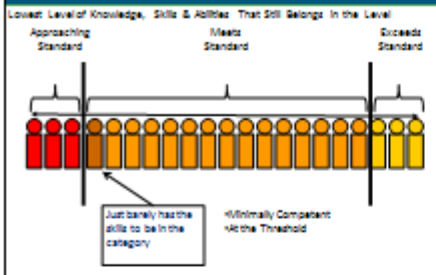
Range of Performance on the Test



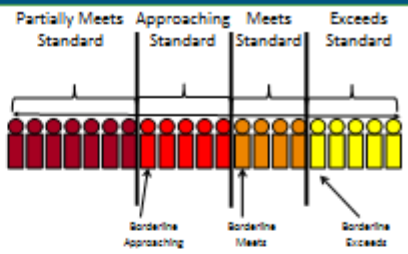
Range of Performance Within Level



Borderline Student–



Borderline Examinee at Each Level



Turn over to Jim Patterson for review of PLDs with panelists

- You are familiar with ERW section of the exam at this point, and have begun to conceptualize the borderline examinee.
- Now you will hear more about the knowledge, skills, and abilities that are necessary for students receiving each SAT level to demonstrate on the ERW section.

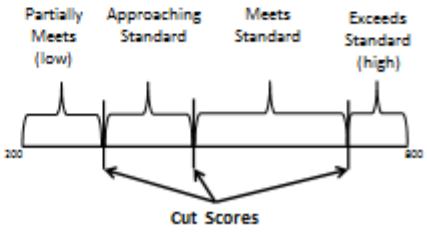
The next 2 days…

- ~~Overview of SAT~~
- ~~Experience the SAT Evidence-based Reading & Writing section (ERW)~~
- ~~Introduce the Borderline Examinee~~
- ~~Familiarize with Performance Level Descriptions (PLDs)~~
- Training and Practice on the Standard Setting Rating Task
- Make your ratings/discussion (Round 1, Round 2, Round 3)
- Final Evaluation and wrap-up

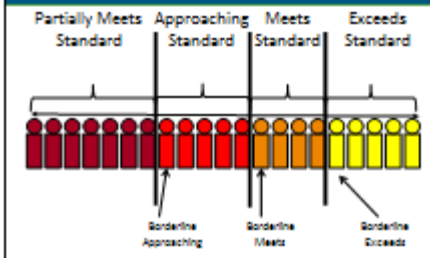
REMINDER: Purpose of meeting

- To make a recommendation to state policymakers about what the locations of the cut scores should be on the SAT Evidence-Based Reading and Writing Section (ERW).
- 4 Achievement levels (Partially Meets Standard, low – Exceeds Standard, high); 3 cut score locations will need to be recommended to classify students into one of the 4 levels.

4 Performance Levels



Borderline Examinee at Each Level



Modified Angoff Method (the standard setting task)

- For each item
  - Consider 100 borderline examinees at each Performance Level.
  - Consider how difficult you believe the item will be for each group of 100 borderline examinees and what about the item makes it more or less difficult – text complexity of passage, wording, vocabulary, content, inference required, etc.
  - For each group of 100 borderline examinees separately, record the number of examinees that you would expect to answer the item correctly based on the PLDs.

Thought Process

- Given the item, how difficult it is, and the PLDs, what proportion of borderline examinees at the **Meets Standard** cut would answer the item correctly?
- Given the item, how difficult it is, and the PLDs, what proportion of borderline examinees at the **Approaching Standard** cut would answer the item correctly?
- Given the item, how difficult it is, and the PLDs, what proportion of borderline examinees at the **Exceeds Standard** cut would answer the item correctly?

Acceptable Range of Responses

- Ratings can be any number between 10 and 95 in increments of 5:

  10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95
- Ratings for higher cut scores to the right on the rating form should be equal to or higher than the ratings that are in the columns to the left.
- Use the entire range as appropriate. It may help to think about the distractors and how many you think each group of borderline examinees would be able to eliminate based on the PLDs.

Chance

- The SAT items are 4 choice items.
- Students would be expected to get 25% correct by chance if they were strictly guessing.
- To be placed into Approaching students should have to do more than just guess so the number of students that you select should usually be higher than 25.
- A value of 10, 15, or 20 can be used when you believe it is absolutely necessary but should be used sparingly since it indicates the students are expected to do even worse than guessing the answer correctly.
- Values of 10, 15, 20 and 25 will turn yellow on your data entry sheet as a reminder that these are values at or below chance.

High End Rating

- The highest rating possible is 95, in recognition that perfect performance is not common, not a reasonable expectation for borderline students.

Recording the ratings

- First write your rating for each item at each cut score on the paper form provided.
- Double check that numbers to the right are equal to or larger than numbers to the left and no cells are left blank. (practice data entry tab in Google docs workbook)
- Then, copy ratings into google spreadsheet on your laptop. Use the **pull down menus** or **type directly into the cell** to enter your ratings into the spreadsheet.
- When finished you should have no bright red cells or blank cells in the spreadsheet.
- Yellow will appear when you select cells of 10, 15, 20 or 25 (to remind of below chance)
- DO NOT COPY AND PASTE
- Let room staff know when you are done entering your ratings by standing your name tent on end.

Make Practice Ratings

- We will now pass out Practice materials
- Using the practice rating form (green), make ratings for items 1 and 2 in the Reading section and items 1 and 2 in the Writing and Language section.
- Remember to reference your PLDs for the borderline examinee at each cut score.
- Write your ratings on the paper form and then enter those ratings into the practice data entry form on your laptop.
- Please let us know if you have any questions or need help accessing the data entry form.
- Please stand your name tent on end when you are finished.

Discuss Practice Ratings

- Identify the range of cut scores based on the practice.
- Discuss discrepant ratings and rationales
- Additional Questions about the task of providing ratings?
- Training/Ready to Proceed Evaluation form (blue)
- We will be coming around to pick up your training rating form.

The next 2 days…

- ~~Overview of SAT~~
- ~~Experience the SAT Evidence based Reading & Writing section (ERW)~~
- ~~Introduce the Borderline Examinee~~
- ~~Familiarize with Performance Level Descriptions (PLDs)~~
- ~~Training and Practice on the Standard Setting Rating Task~~
- Make your ratings/discussion (Round 1, Round 2, Round 3)
- Final Evaluation and wrap-up

Provide Round 1 Ratings

- **Multiple Choice Items**
  - For each cut score, provide the number of borderline examinees out of 100 that would be expected to answer each item correctly.
  - **Reference the PLDs to make Ratings.**
- After data entry is completed close the tab containing the spreadsheet.
- Complete Round 1 Evaluation using the online link
- Stand name tent on end.
- Check materials back in and be sure facilitator checks that all ratings are present and dismisses you before leaving.
  - Take up test materials (operational and practice)
  - PLDs
  - Take up paper rating forms

Round 1 Table Group Discussion

- Review item level feedback form
- Identify items where your group differed most in how they rated the items.
- Discuss Items with the Most Discrepant Ratings first. Provide rationales for why one person rated higher while another rated lower.
- Consider the Item Difficulty Rating and how it compares to your perception of the item
- Review individual summary and discuss how your cut score recommendations differ from each other and from the total group.

Round 2 Ratings

- Round 2 gives you an opportunity to change your ratings based on the additional information you have received during the table group discussions.
- You may change any or all of your ratings. You may also choose to change none of your ratings. Only record in the Round 2 column the ratings that you are changing. Be sure to make the same changes to your data entry spreadsheet!
- Remember to always reference the PLDs
- When finished please stand your name tent on end.

Round 2 Table Group Discussion

- Review item level feedback form
  - Two sets—table and large group
- Identify items where your group differed most in how they rated the items.
- Consider the Item Difficulty Rating and how it compares to your perception of the item
- Review individual summary and discuss how your cut score recommendations differ from each other and from the total group.
- Note the key points for large group discussion

Round 2 Large Group Discussion

- Discuss most discrepant placements and share discussion from table groups.
- Be sure to reference PLDs in the rationales.

Round 2 Discussion

- Present Impact Data

Achievement Level	Percent	
Exceeds		
Meets		
Approaching		
Partially Meets		

- Discuss impact data in table groups

Round 3 Ratings

- Round 3 gives you a final opportunity to change your ratings based on the additional information you have received during the Round 2 group discussion.
- You may change any or all of your ratings. You may also choose to change none of your ratings. Only record in the Round 3 column the ratings that you are changing. Be sure to make the same changes to your data entry spreadsheet!
- Remember to always reference the PLDs
- When finished please stand your name tent on end.

Round 3 Debrief

- Present Impact Data

Achievement Level	Cut Score	Percent	
Exceeds			
Meets			
Approaching			
Partially Meets			

Final Thoughts

- Are there any questions or comments that you would like to share at this time before we end the meeting?

Check out process

- Have you completed your final evaluation form? If not, please do so (link in pdf on your laptop)
- Turn in materials at the "check out station" at the front of the room:
  - Angoff rating form (not practical)
  - Item-level feedback forms (white paper with p-values)
  - Individual cut score feedback (white paper with cut scores)
  - PLDs
- Check the sign out sheet. Don't leave until your materials have been officially checked back in.
- You can either leave everything else on your desk or take it with you.

Thank you!

Safe Travels Home!

Appendix M: ISBE Standard Setting Information for Board Approval

ILLINOIS STATE BOARD OF EDUCATION MEETING
October 18, 2017

TO:	**Illinois State Board of Education**
FROM:	**Tony Smith, Ph.D., State Superintendent of Education** **Libi Gil, Ph.D., Chief Performance Officer**
Agenda Topic:	**SAT School Day Performance Levels and Threshold Scores**
Materials:	**Panelist Demographics**
Staff Contact(s):	**Mary Reynolds, Executive Director of Innovation and Secondary Transformation** **A. Rae Clementz, Director, Assessment and Accountability**

Purpose of Agenda Item
The Center for Teaching and Learning requests the Board to authorize the State Superintendent to adopt the following performance levels and cut scores for the SAT School Day test for the purposes of federal and state accountability.

Relationship to the State Board's Strategic Plan and Implications for the Agency and School Districts
The SAT School Day test provides feedback on academic success in English/language arts (ELA) and mathematics in support of the following area of the State Board's Strategic Plan:

Every child in each public school system in the State of Illinois deserves to attend a system wherein…

- Ninety percent or more of students graduate from high school ready for college and career.

Background Information
The Every Student Succeeds Act requires Illinois to administer an assessment in ELA and mathematics at least once in high school. ISBE administered the SAT School Day test in 2016-17 for the first time to all 11th-grade students attending public schools. Illinois offered the SAT at no cost during the school day to every student in 11th grade. This approach was both the appropriate course of action to ensure that all Illinois students had access to a college entrance exam and to comply with federal assessment requirements. This was the first step to removing Illinois from high-risk status for the receipt of Title I Part A funds.

A rigorous standard-setting process to determine performance levels and threshold scores aligned to Illinois Learning Standards is necessary to meet a second condition for removal of high-risk status. Illinois will provide documentation of this process as part of the evidence needed for peer review of Illinois' standards and assessment system during the February submission window. ISBE convened a panel of 49 educators with subject matter expertise (24 ELA, 25 math) on September 27 and 28, 2017, in Springfield to conduct a standard setting and make recommendations on performance levels. The multi-phase process produced recommended cut scores for each subject area on four performance levels: Exceeds Standards, Meets Standards, Approaching Standards, and Partially Meets Standards. The recommended cut scores are identified in Table One (Table One: Recommended Cut Scores).

Table One: Recommended Cut Scores

	Partially Meets Standards	Approaching Standards	Meets Standards	Exceeds Standards
ELA	430	540	640	
Math	450	540	670	

Summary of the Process

ISBE presented the methodology used for this standard setting to the Board on September 13, 2107. ISBE solicited panelist nominations from school and district administrators in advance of the convening. The nominees, who were primarily practicing teachers and district personnel, were selected to be panelists based on their content area expertise, familiarity with the Illinois Learning Standards, and experience working with diverse student populations, as well as other areas of expertise, such as college and career readiness or the development of large-scale assessments programs (Attachment A). A panel of subject matter experts was convened for each subject area, with 24 panelists on the ELA panel and 25 panelists on the math panel.

Performance level descriptors (PLDs) and associated qualitative skill profiles aligned to the Illinois Learning Standards were drafted by a writing team composed of ISBE staff and external content and assessment experts (Table Two: Performance Level Descriptors). The PLDs were reviewed by College Board psychometric and content experts.

The Modified Angoff Method was used for the SAT performance level-setting process. This method has been used with assessments such as the American College Test (ACT), the Partnership for Assessment of Readiness for College and Careers (PARCC), and the National Assessment of Educational Progress (NAEP). It also meets the *Standards for Educational and Psychological Testing*. The multi-step method relies on panelists estimating the difficulty of each item for a hypothetical group of "borderline students" (students who have just entered the proficiency level from the one below it).

Panelists received an overview of the exam. They then actually took the exam under timed conditions. Next, they reviewed PLDs that were provided so they could understand the key differences between performance levels. Finally, they received training on the Modified Angoff Method used to make judgments.

Panelists used the PLDs to guide them in making ratings for each test item at three points. Three rounds of ratings were conducted, and panelists were provided feedback and conducted discussions at their tables or with the whole group after each round. After the first round, panelists were provided information on how they rated each item compared to the others at their table. After the second round, they received the same type of item-rating information, but were also provided with the resulting cut scores and percentage of students who would fall into each level if those cuts were adopted. After the third round, they were provided only with the resulting cut scores and student distributions. Table Three provides information on the impact data of the recommended cut scores (Table Three: Impact Data) and Table Four places this information in context by showing the percentage of students that met and exceeded standards on the 2017 SAT and the 2017 PARCC (Table Four: Context Data by Subject Area).

Each panelist's estimated passing cut score for the total test was calculated by combining their estimates for each individual item. The mean of all panelists' final test score recommendations was presented as the overall panel-based cut score recommendation. ISBE leadership carefully considered this recommendation and then made its final recommendation of the cuts in Table

One. This final recommendation adopts all of the cuts recommended by the panelists, with the exception of the cut score for "Math Exceeds."

The final recommended cut score in math lowers the cut from the panelist-recommended cut score of 690 to 670, which is the next achievable score below the mean panelist-recommended cut of 676.4. Even in round three, panelists recommended cut scores at this highest level ranging from 590 to a max 730. This change is equivalent to correctly responding to two fewer questions and consistent with qualitative and evaluative feedback given by math panelists. The result of this change is a statewide Exceeds rate of 7.8 percent as opposed to 6.1 percent (the rate at the panelist-recommended median cut score of 690) and does not change the overall Meets/Exceeds rate of 36.3 percent, which is the rate used for purposes of accountability.

The convening was overseen by Assessment and Accountability staff; facilitated by College Board psychometricians and content experts; and evaluated for validity by an external third party, Dr. John Olson of Assessment Solutions Group.

Table Two: Performance Level Descriptors

4	Exceeds Standards	The student has exceeded the proficiency level & demonstrates a **thorough** understanding of the knowledge & skills needed relative to the Illinois Learning Standards.
3	Meets Standards	The student has met the proficiency level & demonstrates **adequate** understanding of the knowledge & skills needed relative to the Illinois Learning Standards.
2	Approaching Standards	The student is approaching the proficiency level & demonstrates an **incomplete** understanding of the knowledge & skills needed relative to the Illinois Learning Standards.
1	Partially Meets Standards	The student has only partially met standards & demonstrates a **minimal** understanding of the knowledge & skills needed relative to the Illinois Learning Standards.

Table Three: Impact Data

ELA					MATH				
Performance Level		Cut	% At	Meets/Not	Performance Level		Cut	% At	Meets/Not
4	Exceeds	640	13.0%	39.7%	4	Exceeds	670	7.8%	36.3%
3	Meets	540	26.7%		3	Meets	540	28.5%	
2	Approaching	430	37.4%	60.3%	2	Approaching	450	32.6%	63.7%
1	Partially Meets	--	22.9%		1	Partially Meets	--	31.1%	

Table Four: Context Data by Subject Area

ELA										
Test	SAT	PARCC	NAEP		SAT	PARCC	NAEP			
Year	2017	2017	2015		2017	2017	2015			
Grade	11	8	8		11	8	8			
Level 4	13.31%	6.2%	4%	Meets	40.30%	37.40%	35%			
Level 3	26.99%	31.2%	31%							
Level 2	37.31%	26.1%	42%	Does Not Meet	59.71%	62.60%	65%			
Level 1	22.40%	19.9%	23%							
Level 0		16.6%								

ELA							
Test	SAT	PARCC	NAEP		SAT	PARCC	NAEP
Year	2017	2017	2015		2017	2017	2015
Grade	11	8	8		11	8	8
Level 4	13.31%	6.2%	4%	Meets	40.30%	37.40%	35%
Level 3	26.99%	31.2%	31%				
Level 2	37.31%	26.1%	42%	Does Not Meet	59.71%	62.60%	65%
Level 1	22.40%	19.9%	23%				
Level 0		16.6%					

Financial Background

None.

Analysis and Implications for Policy, Budget, Legislative Action, and Communications

Policy Implications: The adoption of appropriate performance levels and cut scores that are reflective of the Illinois Learning Standards and in alignment with the broader accountability system are crucial for the accurate identification of schools in need of support and improvement. These performance levels will impact eligibility for supports and services, as well as eligibility for recognition as a highly effective school or district.

Budget Implications: None anticipated.

Legislative Action: No action is required.

Communication: Communication regarding the standard-setting process and the relationship between the standards and their impact has been ongoing since spring 2017. Clear communication regarding the meaning and appropriate interpretation of the performance levels and cut scores will be conducted in conjunction with the release of the Illinois Report Card.

Pros and Cons of Various Actions

Pros: The approval of appropriate performance levels that are reflective of the Illinois Learning Standards and in alignment with the broader accountability system will ensure accurate identification of schools in need of support and improvement. Additionally, adoption of performance levels established via a rigorous standard-setting process will ensure that Illinois complies with federal assessment requirements. The subsequent peer review process in February will give Illinois the opportunity to be removed from high-risk status regarding allocations of Title I Part A funds.

Cons: Failure to approve the performance levels and cut scores would result in lack of compliance, retention or escalation of high-risk status, possibly result in the loss of Title I Part A funds, delay the release of the Report Card, and represent a significant cost to the state because the standard-setting process would have to be repeated.

Superintendent's Recommendation
I recommend that the following motion be adopted:

> The State Board of Education hereby authorizes the State Superintendent to adopt the recommended performance levels and cut scores for the SAT School Day test.

Next Steps
Upon Board authorization, agency staff will proceed with preparations for release of the performance level data via the Illinois Report Card and communicate results to schools and districts.