PROCESS USED TO

DETERMINE CUT SCORES FOR THE

ALABAMA HIGH SCHOOL

GRADUATION EXAM

(READING, LANGUAGE, MATHEMATICS, SCIENCE, SOCIAL STUDIES, AND BIOLOGY)

PROCESS USED TO DETERMINE CUT SCORES FOR THE ALABAMA HIGH SCHOOL GRADUATION EXAM

PERFORMANCE LEVELS

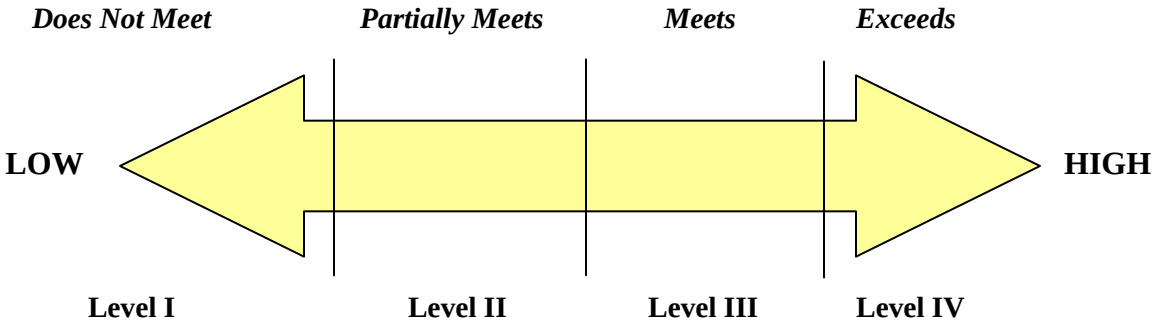
For the *Alabama High School Graduation Exam* (AHSGE), the Alabama State Department of Education (ALSDE) required four performance levels that corresponded to three cut scores. The four performance levels are:

Exceeds Academic Content Standards (*Level IV*)
Meets Academic Content Standards (*Level III*)
Partially Meets Academic Content Standards (*Level II*)
Does Not Meet Academic Content Standards (*Level I*)

The three cut scores were for:

Exceeds (for those who are at the threshold of *Exceeds* and barely pass the *Exceeds* performance level)
Meets (for those who are at the threshold of *Meets* and barely pass the *Meets* performance level)
Partially Meets (for those who are at the threshold of *Partially Meets* and barely pass the *Partially Meets* performance level)

The performance levels and cut scores may be represented by the following figure. The two-headed arrow symbolizes the continuum of scores on the tests, ranging from low to high.



Note that the vertical lines indicate the location of cut scores.

Level III and Level IV Cut Scores for Reading, Language, Mathematics, Science, and Social Studies

The ALSDE and CTB McGraw-Hill (CTB) conducted standard settings for the AHSGE. To set the Level III and Level IV cut points, the Bookmark method (described below) was used. Two standard setting committees established cut scores for Level III and Level IV for the AHSGE Reading, Language, Mathematics, Science, and Social Studies subject-area tests. The standard setting committees were comprised of classroom teachers representing the demographics and state board districts of Alabama. In each standard setting committee meeting, a CTB psychometrician, a member of the ALSDE staff, and one of the Technical Advisory Committee members worked as facilitators.

For each subject-area test, the groups worked separately, determining performance levels for their subject. After three rounds of suggestions and recommendations for each subject-area test, the committees provided findings regarding Level III and Level IV cut scores.

STANDARD SETTING

Bookmark

There are several well-established methods available for establishing performance standards. The Bookmark standard setting procedure was used for producing the suggested cut scores for Level III and Level IV for the AHSGE. The Bookmark standard setting procedure was developed by CTB. The standard setting activity for each subject took approximately 16 hours spread across two days.

Standard Setting Process

The AHSGE judges were Alabama teachers recommended by their school districts to participate in the standard setting meeting. The judges were placed on panels each of which had three facilitators: one from the ALSDE, one from CTB, and one from the Alabama Technical Advisory Committee. The Technical Advisory Committee members are tenured at Alabama universities.

The first activity during the standard setting meeting was an orientation of the table leaders. The next activity was an orientation of all committee members to the standard setting process. The purpose of the orientation about the procedures for establishing cut scores for Level III and Level IV was to ensure the appropriate operation of the standard setting. It is likely that the standard setting activity would be unfamiliar to most of the judges, so acquainting them with the expectations for their performance served to increase their confidence in the task.

The judges were informed that the standard setting was not a forum to address the quality of the content standards, the test, or the policies related to the administration of the test. The orientation concentrated on helping judges to become familiar with two substantive aspects of the standard setting procedure. First, using an Ordered Item Booklet (OIB), ordered by difficulty, the judges

were asked to place the bookmark at the place where the committee member felt that a student who had mastered content reflected by the items before the bookmark should have sufficient skills to infer that the student merited each proficiency level. This was to be established at the point where students who are just at the threshold for each proficiency level *should* perform rather than how they do or will perform. This important distinction was emphasized on numerous occasions.

Second, judges were assured that their ratings would remain confidential. The recommended cut points would be based on the group's ratings, and individual ratings would not be released in the technical documentation. Although an important goal of the process was for judges to approach consensus or convergence in ratings, it was integral to the process for judges to feel free to maintain a rating that they personally believed was correct, whether or not it was consistent with ratings made by other judges. As Fitzpatrick (1989) noted, preserving the anonymity of judges may make it easier for them to revise an initial item rating after they have learned more about the item, because the judges have not been publicly committed to their initial rating of the item. In contrast, encouraging judges to maintain their initial ratings, if they believe them to be appropriate, may be desirable if it enables judges to resist pressures from other panel members to conform. Fitzpatrick suggests that conformity due to social pressure is not desirable in standard setting. Items with disparate ratings were discussed in order to educate the judges about other judges' rationale behind their ratings. Any potential effects of undue social pressure would be moderated through the group process skills of the facilitators of the standard setting.

Independent Ratings of Each Item

At the beginning of the breakout session, the group facilitators led the judges in developing a shared concept of the threshold student at each proficiency level in their respective subject. Each committee member was then given a copy of their respective test and they worked individually to answer the items. Once all committee members completed the test, each committee member was given a copy of the OIB, Item Map, and Key. The OIB contains the test items, ordered by difficulty, with the easiest item appearing first and the hardest item appearing last. The ordering is determined by student performance on the items. The Item Map provides information about each item in the OIB and has space for participants to record their thoughts about the items.

In small groups (tables of six or seven), judges examined each item in the OIB, discussing what each item measures and what makes it harder than the items before it. After this discussion, each participant set a cut score for Level III and Level IV by placing a bookmark in the OIB according to his or her own judgment of what content, for example Level IV, students should know and be able to do.

In Round 2, judges discussed the rationale behind their original bookmark placements with the participants at their table. After review and discussion of Round 1 results, judges were free to keep or adjust their bookmark placements.

At the beginning of Round 3, judges were shown a summary of each small group's bookmark placement and impact data (percent of students in each performance level based on Round 2 results). After a review and discussion of Round 2 results for all three small groups and a review

of the impact of the results of Round 2, judges were free to keep or adjust their bookmark placements. After the final round of bookmark placement, the recommended cut score for each performance level was established as the median of the bookmark placements in the final record.

On the final day, the Technical Advisory Committee, CTB, and ALSDE staff met to discuss the recommended cut scores. After considering technical aspects and impact data, recommended cut scores were presented to the State Superintendent of Education for approval. Cut scores for proficiency were then presented to the State Board of Education for approval.

Level II Cut Score for Reading, Language, Mathematics, Science, and Social Studies

The ALSDE and Data Recognition Corporation (DRC) conducted standard settings for the Level II cut scores for the AHSGE. The initial cut scores were obtained based on the panelists' judgments. After review by the ALSDE, the recommended cut scores for Level II were determined considering psychometric perspectives (e.g., standard error of measurement of the total scores).

To set the Level II cut points, the modified-Angoff method (described below) was used. The Standard Setting Committees were comprised of classroom teachers representing the demographics and state board districts of Alabama. In each standard setting, a DRC psychometrician, a member of the ALSDE staff, and one of the Technical Advisory Committee members worked as facilitators.

For each subject-area test, the groups worked separately, determining performance levels for their subject. After three rounds of suggestions and recommendations for each subject-area test, the committees provided findings regarding Level II cut scores.

STANDARD SETTING

Modified-Angoff

There are several well-established methods available for establishing performance standards. A modified-Angoff procedure (Angoff, 1984) was used for producing the suggested Level II cut scores for the AHSGE. This procedure has a long and successful history in similar applications for both educational and professional certification assessments. The modified-Angoff procedure provides a systematic technique for eliciting judgments from panels of experts, producing consensus among these experts, and quantifying the results of the judgments. It is widely recognized as the simplest method to use (Norcini, et al., 1987; Shepard, 1980). Moreover, research has shown that the modified-Angoff method produces ratings with better reliability and smaller variability among the ratings of judges than other standard setting procedures (Andrew and Hecht, 1976; Brennan and Lockwood, 1980; Cross, et al., 1984; Poggio, Glasnapp, and Eros, 1981; Skakun and Kling, 1980). This procedure represents an appropriate balance between statistical rigor and informed opinion. However, this method can be somewhat time-consuming due to the necessity of rating each item, going through three rounds of ratings, and processing all

of the data. The standard setting activity for each subject-area test took approximately 16 hours spread across two days.

Standard Setting Process

The AHSGE judges were Alabama teachers recommended by their school districts to participate in the standard setting conference. The panelists were placed in committees each of which had three facilitators: one from the ALSDE, one from DRC, and one from the Alabama Technical Advisory Committee. The Technical Advisory Committee members are tenured at Alabama universities.

The first activity during the standard setting was an orientation of the committee members to the standard setting process. The orientation about the procedures for establishing cut scores for the proficiency level was to ensure the appropriate operation of the standard setting. It is likely that the standard setting activity would be unfamiliar to most of the panel members, so acquainting them with the expectations for their performance served to increase their confidence in the task.

At the outset, judges were reminded that their task was to review the items for their respective subject-area test and to estimate the minimal acceptable performance for students at proficiency Level II on each item. They were informed that the standard setting was not a forum to address the quality of the content standards, the test, or the policies related to the administration of the test. The orientation concentrated on helping judges to become familiar with two substantive aspects of the standard setting procedure. First, the judges were asked to estimate how students who are just at the threshold for proficiency Level II *should* perform rather than how they do or will perform. This important distinction was emphasized on numerous occasions.

Second, judges were assured that their ratings would remain confidential. The recommended cut points would be based on the group's ratings, and individual ratings would not be released in the technical documentation. Although an important goal of the process was for judges to approach consensus or convergence in ratings, it was integral to the process for judges to feel free to maintain a rating that they personally believed was correct, whether or not it was consistent with ratings made by other judges. As Fitzpatrick (1989) noted, preserving the anonymity of judges may make it easier for them to revise an initial item rating after they have learned more about the item, because the judges have not been publicly committed to their initial rating of the item. In contrast, encouraging judges to maintain their initial ratings, if they believe them to be appropriate, may be desirable if it enables judges to resist pressures from other panel members to conform. Fitzpatrick suggests that conformity due to social pressure is not desirable in standard setting. Items with disparate ratings will be discussed in order to educate the judges about other judges' rationale behind their ratings. Any potential effects of undue social pressure will be moderated through the group process skills of the facilitators of the standard setting.

Independent Ratings of Each Item

At the beginning of the session, the group facilitators led the panelists in developing a shared concept of the threshold student at proficiency Level II in their respective subject area. Each committee member was then given a copy of their respective tests and worked individually to

answer the items. Once all committee members completed the test, answer keys were provided and their tests were scored. Committee members were given sufficient time (approximately 60-90 minutes) to independently rate each item on the test (Round 1). They were encouraged to read each item, consider the skills being assessed and the importance of those skills, think of 100 threshold students (at proficiency Level II), and record an estimate of how many, or what percentage, of those 100 threshold students (at proficiency Level II) *should* correctly answer the item. Upon completion of the first round of ratings, all secure materials were collected and inventoried before committee members were dismissed from the meeting.

During the evening, the individual ratings of the judges were aggregated by the DRC research analysts. Statistics for each judge and for the entire panel were also computed. To obtain an overall estimate of the cut point for proficiency Level II from the total group of judges, the initial item ratings provided by the judges were treated as *p*-values (in multiple choice items) and summed across items by level. The result of this summation is a number-correct value for each judge. The number-correct values were then averaged across judges to obtain the judges' estimate of the cut point for proficiency Level II.

Collection and Discussion of Data

The following morning, the panelists were shown the frequency distributions of their individual item ratings and cut scores, along with the average cut score given by their group. Discussion followed. Once discussion of the results of the initial ratings concluded, the judges were asked to review the entire set of items that they rated in Round 1, to reconsider these ratings in light of the discussion and the data they had been shown, and to revise any of their ratings, if necessary. The judges' focus was again directed toward thinking about 100 threshold students at the proficiency Level II and how they *should* perform on the items. The Round 2 ratings were collected and inventoried along with the secure materials. As with Round 1, the judges' Round 2 ratings were aggregated. Statistics for each judge and for the entire panel were also computed.

Judges were shown the frequency distributions of their Round 2 individual item ratings and cut scores, along with the average Round 2 cut score and impact data based on their Round 2 judgment. In Round 3, judges had another opportunity to alter their estimates of the Round 2 cut point if they felt that their Round 2 cut point was too high or too low. Again, the judges rated individual items in Round 3, and then the Round 3 cut scores were collected and tabulated. After Round 3, the initial cut score for Level II for each subject was determined, based on the results of the judges.

On the final day, the Technical Advisory Committee, DRC, and ALSDE staff met to discuss the recommended cut scores. After considering technical aspects and impact data, recommended cut scores were presented to the State Superintendent of Education for approval.

Level II, III, and IV Cut Scores for Biology

The ALSDE and DRC conducted standard setting for the Level II, III, and IV cut scores for the biology subject-area test of the AHSGE. The initial cut scores were obtained based on the panelists' judgments. After review by the ALSDE, the recommended cut scores for each level

were determined considering psychometric perspectives (e.g., standard error of measurement of the total scores).

To set the cut points, the modified-Angoff method (described below) was used. The Standard Setting Committee was comprised of classroom teachers representing the demographics and state board districts of Alabama. A DRC psychometrician, a member of the ALSDE staff, and the Technical Advisory Committee members worked as facilitators.

After three rounds of suggestions and recommendations, the committee provided findings regarding Level II, III and IV cut scores.

STANDARD SETTING

Modified-Angoff

There are several well-established methods available for establishing performance standards. A modified-Angoff procedure (Angoff, 1984) was used for producing the suggested cut scores for the biology subject-area test of the AHSGE. This procedure has a long and successful history in similar applications for both educational and professional certification assessments. The modified-Angoff procedure provides a systematic technique for eliciting judgments from panels of experts, producing consensus among these experts, and quantifying the results of the judgments. It is widely recognized as the simplest method to use (Norcini, et al., 1987; Shepard, 1980). Moreover, research has shown that the modified-Angoff method produces ratings with better reliability and smaller variability among the ratings of judges than other standard setting procedures (Andrew and Hecht, 1976; Brennan and Lockwood, 1980; Cross, et al., 1984; Poggio, Glasnapp, and Eros, 1981; Skakun and Kling, 1980). This procedure represents an appropriate balance between statistical rigor and informed opinion. However, this method can be somewhat time-consuming due to the necessity of rating each item, going through three rounds of ratings, and processing all of the data. The standard setting activity for each subject-area test took approximately 16 hours spread across two days.

Standard Setting Process

The AHSGE judges were Alabama teachers recommended by their school districts to participate in the standard setting conference. The panelists were placed in a committee which had three facilitators: one from the ALSDE, one from DRC, and one from the Alabama Technical Advisory Committee. The Technical Advisory Committee members are tenured at Alabama universities.

The first activity during the standard setting was an orientation of the committee members to the standard setting process. The orientation about the procedures for establishing cut scores for the achievement levels was to ensure the appropriate operation of the standard setting. It is likely that the standard setting activity would be unfamiliar to most of the panel members, so acquainting them with the expectations for their performance served to increase their confidence in the task.

At the outset, judges were reminded that their task was to review the items for their respective subject-area test and to estimate the minimal acceptable performance for students at each achievement level (Level II, III, and IV) on each item. They were informed that the standard setting was not a forum to address the quality of the content standards, the test, or the policies related to the administration of the test. The orientation concentrated on helping judges to become familiar with two substantive aspects of the standard setting procedure. First, the judges were asked to estimate how students who are just at the threshold for each achievement level *should* perform rather than how they do or will perform. This important distinction was emphasized on numerous occasions.

Second, judges were assured that their ratings would remain confidential. The recommended cut points would be based on the group's ratings, and individual ratings would not be released in the technical documentation. Although an important goal of the process was for judges to approach consensus or convergence in ratings, it was integral to the process for judges to feel free to maintain a rating that they personally believed was correct, whether or not it was consistent with ratings made by other judges. As Fitzpatrick (1989) noted, preserving the anonymity of judges may make it easier for them to revise an initial item rating after they have learned more about the item, because the judges have not been publicly committed to their initial rating of the item. In contrast, encouraging judges to maintain their initial ratings, if they believe them to be appropriate, may be desirable if it enables judges to resist pressures from other panel members to conform. Fitzpatrick suggests that conformity due to social pressure is not desirable in standard setting. Items with disparate ratings will be discussed in order to educate the judges about other judges' rationale behind their ratings. Any potential effects of undue social pressure will be moderated through the group process skills of the facilitators of the standard setting.

The *borderline student* was defined for the panelists as the lowest possible student who would be classified as Level II, III, or IV. They were provided with a short definition of each achievement level in terms of the student's degree of mastery of fundamental Biology content and through a short case study of hypothetical students for each level. These descriptions are consistent with the materials used in prior standard setting sessions in Alabama.

Standard setting involved three rounds of deliberations. In round one, each panelist recorded a rating for each item, with no discussion among the panelists about the items. Panelists were asked to determine their Level III ratings for each item, followed by Level II ratings for each item, and ending with their Level IV ratings for each item. They proceeded through the items in test booklet order and the ratings were recorded on a rating form in the same order. The rating form provided for each panelist included the item identification, the item key, the content standard assignment, space for recording the panelist's ratings, and space for recording notes.

Round two began with the presentation of the round one ratings for each panelist and for the group as a whole. The panelists were shown a diagram indicating each panelist's cut points, as well as a chart for each item, showing all panelists' ratings on each item. The panelists were instructed to review the items as a large group and discuss any differences in understanding of what was required by the item. Panelists could revise their own ratings if they felt it appropriate. The primary difference from round one to round two was the group discussion about the items.

At the start of round three, panelists were again presented with their individual ratings, as well as the group's ratings from round two. They were also shown impact data for grade 11 students from the recent live administration (i.e., the percent of students in each achievement level) based on round two. This round provided a second opportunity to discuss and revise after viewing the impacts and how their ratings fell alongside other panelists' ratings.

On the final day, the Technical Advisory Committee, DRC, and ALSDE staff met to discuss the recommended cut scores. After considering technical aspects and impact data, recommended cut scores were presented to the State Superintendent of Education for approval.