Arizona's
Instrument to Measure Standards

Bookmark Standard Setting Technical Report 2008

for

Grades 4, 8, and High School Science

Submitted to
Arizona Department of Education
February 2009

Table of Contents

SECTION A

Executive Summary

Executive Summary

In June 2008, staff members from the Arizona Department of Education (ADE) and CTB/McGraw-Hill worked in collaboration to perform standard setting on Arizona's Instrument to Measure Standards (AIMS) in Grades 4, 8, and High School Science. The purpose of the standard setting was to establish cut scores for each assessment and to place students into four performance levels: *Falls Far Below the Standard*, *Approaches the Standard*, *Meets the Standard* and *Exceeds the Standard*, where *Exceeds the Standard* represents the highest level of performance on the test. The Bookmark Standard Setting Procedure (BSSP; Lewis, Mitzel & Green, 1996) was implemented to set performance standards on the assessments.

A committee of educators from across the state of Arizona convened to engage in the standard setting workshop on June 9 – 11, 2008, to recommend a well-articulated set of performance standards. The ADE divided participants into three grade groups, each with approximately 12 participants. Participants were divided into assigned grade groups that were balanced in terms of relevant demographic characteristics (e.g., gender, geographic location). The standard setting consisted of training, orientation, three rounds of judgments, an articulation discussion, and performance level description writing.

Following the standard setting, the cut scores recommended by the standard setting committee were approved by the Arizona State Board of Education. The final cut scores adopted for the AIMS program for Science are shown in Table 1. The impact data associated with these cut scores—the percentage of students classified in each performance level—are shown in Table 2.

Table 1. Final cut scores for Science approved by the Arizona Board of Education.

	Cut Scores		
Grade	***Approaches***	***Meets***	***Exceeds***
4	462	500	547
8	473	500	532
HS	475	500	537

Table 2. Impact data associated with the final cut scores in Table 1.

	Impact Data			
Grade	***Falls Far Below***	***Approaches***	***Meets***	***Exceeds***
4	23%	26%	35%	17%
8	30%	20%	22%	28%
HS	32%	18%	26%	24%

This report summarizes the results of the AIMS Standard Setting for Grades 4, 8, and High School Science. A day-by-day synopsis is included in Section B. The master agenda is included in Section C. The handouts of slides presented to participants during orientation and training are in Section D. The training materials given to participants are provided in Section E. Section F presents details of the participants' Bookmark placements for each group. In Section G, estimates are given of the percentages of students in each performance level at plus/minus one, two, and three standard errors of the participants' recommended final round cut scores. Section H contains graphical representations of participants' judgments. Section I contains the results of the participants' evaluation of the workshop. Section J contains the performance level descriptors (PLDs) presented to participants as well as the final revised ADE approved PLDs. As a reference for the reader, Section K presents *Calculating a Meaningful Standard Error for the Bookmark Cut Score* and *The Bookmark Standard Setting Procedure: Methodology & Recent Implementations* (Lewis, Green, Mitzel, Baum, & Patz, 1998).

SECTION B

Synopsis of the Standard Setting

AIMS Standard Setting: Day-by-Day Synopsis

The Arizona Department of Education (ADE) held a standard setting for Arizona's Instrument to Measure Standards (AIMS) that was facilitated by CTB/McGraw-Hill (CTB). The CTB Standard Setting Team facilitated the AIMS Standard Setting in Phoenix, Arizona on June 9–11, 2008. The purpose of the standard setting was to establish cut scores that placed students into four performance levels: *Falls Far Below the Standard*, *Approaches the Standard*, *Meets the Standard*, and *Exceeds the Standard*, where *Exceeds the Standard* represents the highest level of performance on the assessment. The purpose of this document is to describe the implementation of the AIMS standard setting and provide evidence that may be used to support the procedural validity of the BSSP procedure to recommend cut scores for the AIMS Science assessment.

The Bookmark Standard Setting Procedure (BSSP; Lewis, Mitzel & Green, 1996) was used to set the performance standards in the AIMS Science assessment for Grades 4, 8, and High School. The BSSP is the most commonly used standard setting method, and it has been implemented across the nation to establish performance standards for statewide assessments (Karantonis & Sireci, 2006). The BSSP consists of training and orienting participants, three rounds of judgments, an articulation discussion, and performance level description writing.

The BSSP is considered a test-centered, standard setting methodology because participants study the content of the test. During the procedure, participants are trained to consider a select group of students for each performance level, termed the "target students." A target student can be thought of as a student who "just" meets the expectations of the performance level of interest. Participants develop and use a target- student descriptor to conceptualize and exemplify the student who barely has the knowledge, skills, and abilities (KSAs) to be considered, for example, a student who "just" meets the expectations of the performance level *Exceeds*. Referencing the target-student descriptors serves to assist participants in the articulation of performance levels. Using the performance level descriptors, the Arizona Content Standards, the target student descriptions, and participants' knowledge of students, participants are trained to individually determine how much KSAs of the assessment that students need to demonstrate to be placed in a performance level.

Arizona educators convened to study the AIMS Science assessment, consider the KSAs required of students in each performance level, and to discuss these expectations with their colleagues.

Standard Setting Security

Security was of paramount importance throughout the standard setting process. Participants received secure test materials based upon operational items. Secure test materials used during the workshop were numbered and assembled into packets. Each participant signed out a specific packet and signed his or her name on each piece of secure material in the packet. At all times, CTB staff monitored the standard setting rooms to prevent the removal of secure materials. At the end of each day, each participant's materials were collected and audited for each piece of secure material. The secure materials were stored overnight in a secure room. At the conclusion of the workshop, the secure materials were collected, audited, and assessed against the sign-out lists.

Standard Setting Roles

CTB Staff

The CTB Standard Setting Team, a specialized team within CTB Research, worked with staff from the ADE to design, organize, and facilitate the standard setting activities. The CTB Standard Setting Team was composed of Dr. Steve Ferrara, Principal Research Scientist; Dr. Dong-In Kim, Research Scientist; and Dorothy Tele'a, Standard Setting Specialist.

Prior to the workshop, the CTB Standard Setting Team prepared all materials for the workshop. During the workshop, the team was responsible for facilitating the workshop, training participants, entering participant results into a database, and tracking secure materials. Following the workshop, the team prepared the standard setting technical report.

Leslie Dodge, CTB Program Manager, and Nadia Greer, CTB Program Office Coordinator, attended the standard setting and helped with on-site logistics. Michael Frontz, CTB Development Manager; and Randi Rieman-Johns and Andrina Ortiz, CTB Content Editors, attended the standard setting and served as group leaders.

Group Leaders

At the standard setting, the group leaders from CTB helped to implement the BSSP to set performance standards. Group leaders were staff members from CTB Development with expertise in Science and Science assessment development. A description of the group leader's role follows.

> *Group leader.* The group leader served as a facilitator and was in charge of time management, focusing the participants on the series of standard setting tasks and interacting with the participants. The group leader also facilitated discussions and was in charge of security and data management. The group leader collected the rating forms from participants and communicated with staff from CTB and the ADE. The group leader was a non-voting member.

Participants

Participants were recruited from across the state of Arizona. All participants were selected by the ADE such that the committees were composed of a diverse, experienced group of Arizona educators. The standard setting committee comprised 35 participants.

The committee was divided into three groups: Grades 4, 8, and High School. Each of the groups comprised 12 participants with the exception of High School which comprised 11 participants. Table 1 shows the number of participants for each grade.

Table 1. Number of participants for the standard setting workshop by grade.

Grade	Number of Participants
4	12
8	12
HS	11
Total	**35**

Configuration of the Grade Panels

The ADE assigned participants such that each table was as representative and balanced as possible in regard to the relevant demographic characteristics (e.g., gender, geographic location). In addition, a table leader was selected for each group. A description of the table leaders' role follows.

> *Table leaders.* Table leaders were experienced educators and were chosen from among the participants. Some table leaders had a previous role with the assessment, such as serving as item-writers. The primary role of the table leader was to monitor the group discourse, keep the group focused on the task at hand, and keep time for the group. As needed, table leaders found a diplomatic middle ground between participants or requested assistance from CTB and the ADE. Table leaders were voting members of their panels.

Committee Demographics

Following the workshop, all 35 participants completed written evaluations from which CTB collected self-reported demographic information. This information about the participants has been summarized. Table 2 shows the educational background of the participants at each workshop. Tables 3 and 4 show the occupation and work experience of the participants. All of the participants were teachers or administrators.

Table 5 shows participants' experience teaching English-language learners and students with disabilities. At the standard setting, 20% of participants had experience working with students with disabilities, approximately 49% of participants had experience with ELL/ESL students, 9% of participants had experience with vocational education students, approximately 19% of participants had experience with alternative education students, and approximately 46% of participants had experience with adult education students. Section I contains the complete results of the participant evaluation from the workshop.

Table 2. Educational background of participants by grade.

Grade	N	High School	Bachelor's	Master's	Doctorate
Overall	35	0.0%	17.1%	80.0%	2.9%
4	12	0.0%	25.0%	75.0%	0.0%
8	12	0.0%	25.0%	75.0%	0.0%
HS	11	0.0%	0.0%	90.9%	9.1%

Table 3. Occupation of participants by grade.

Grade	N	Teacher	Administrator	Instructional Assistant	Other
Overall	35	82.9%	17.1%	0.0%	0.0%
4	12	75.0%	25.0%	0.0%	0.0%
8	12	91.7%	8.3%	0.0%	0.0%
HS	11	81.8%	18.2%	0.0%	0.0%

Table 4. Work experience in years of participants by grade.

Grade	N	1-5	6-10	11-15	16-20	21+
Overall	35	5.7%	8.6%	17.1%	20.0%	48.6%
4	12	0.0%	16.7%	16.7%	16.7%	50.0%
8	12	8.3%	8.3%	16.7%	8.3%	58.3%
HS	11	9.1%	0.0%	18.2%	36.4%	36.4%

Table 5. Experience of participants by grade, teaching English-language learners, students with disabilities, and other special groups*.

Grade	N	Special Ed.	N	ELL/ ESL	N	Vocational Ed.	N	Alternative Ed.	N	Adult Ed.
Overall	35	20.0%	35	48.6%	33	9.1%	32	18.8%	35	45.7%
4	12	25.0%	12	41.7%	12	0.0%	11	0.0%	12	50.0%
8	12	8.3%	12	41.7%	11	0.0%	11	9.1%	12	33.3%
HS	11	27.3%	11	63.6%	10	30.0%	10	50.0%	11	54.5%

* **Note**: Some participants did not indicate whether or not they have taught students in selected special groups (e.g., vocational education). For this reason, the N-count in Table 5 is different for some special group: the N-counts reflect the response rate from the workshop evaluation.

Bookmark Materials

Ordered Item Booklets

The Ordered Item Booklets (OIBs) was made up of multiple-choice (MC) items. More items were selected for the OIBs than would be administered to a single student, as shown in Table 6. Items from Forms A and B of the Science assessment were combined to form a pseudo test to represent the content for each grade. The OIBs followed the guidance found in the 2001 text *Setting Performance Standards*, in which Mitzel, Lewis, Patz, and Green noted:

> Ordered item booklets span from about 80 to 110 score points, which exceeds normal test lengths. We view the ability to present a more representative sample of a content domain than a single test form to be a strength of the procedure (p. 252).

The selected items for each grade were ordered according to their scale location using a response probability criterion of 0.67. With this criterion, each MC scored item was located at the ability level (scale score) that students would need in order to have a 0.67 probability of answering the item correctly. The Rasch model was used to scale the MC scored items. For more information about the construction of the OIBs, see Lewis, Green, Mitzel, Baum, & Patz (1998), which is included in Section K. Additionally, Beretvas (2004) includes a discussion of the calculation of response probability-adjusted locations for items scaled with the Rasch model.

Table 6 shows the percentage of items in the OIB and the percentage of items in the test blueprint, for each grade. The last column in Table 6 shows the difference between the percentage of items in each OIB and the percentage of items in the operational test blueprint using Form A for comparison. The content coverage of the OIB closely aligns to the coverage specified in the operational test blueprint, as shown in Table 6.

Table 7 shows the total number of score points—the MC scored items—in each OIB for each grade.

Table 6. Standard setting pseudo form verses test blueprint, by grade.

	Standard Setting Pseudo Form				Test Blueprint				
Grade	Total Number of Test Items by Grade Level	Sub-skill	Number of Test Items by Subskill	% of Test Items by Subskill	Total Number of Test Items by Grade Level	Sub-skill	Number of Test Items by Subskill	% of Test Items by Subskill	Difference Between Coverage
4	86	1.1	10	12%	54	1.1	6	11%	1%
		1.2	10	12%		1.2	6	11%	1%
		1.3	10	12%		1.3	6	11%	1%
		2.1	9	10%		2.1	6	11%	-1%
		3.1	9	10%		3.1	6	11%	-1%
		4.1	10	12%		4.1	6	11%	1%
		5.3	10	12%		5.3	6	11%	1%
		6.2	9	10%		6.2	6	11%	-1%
		6.3	9	10%		6.3	6	11%	-1%
8	89	1.1	10	11%	58	1.1	6	10%	1%
		1.2	6	7%		1.2	4	7%	0%
		1.3	10	11%		1.3	6	10%	1%
		1.4	6	7%		1.4	4	7%	0%
		2.1	9	10%		2.1	6	10%	0%
		3.1	8	9%		3.1	6	10%	-1%
		4.2	13	15%		4.2	8	14%	1%
		5.1	15	17%		5.1	10	17%	0%
		5.2	12	13%		5.2	8	14%	-1%
HS	108	1.1	7	6%	64	1.1	5	8%	-2%
		1.2	10	9%		1.2	6	9%	0%
		1.3	11	10%		1.3	6	9%	1%
		1.4	7	6%		1.4	4	6%	0%
		2.1	10	9%		2.1	6	9%	0%
		3.1	11	10%		3.1	7	11%	-1%
		4.1	10	9%		4.1	6	9%	0%
		4.2	10	9%		4.2	6	9%	0%
		4.3	11	10%		4.3	6	9%	1%

Table 7. Total number of score points in each OIB, by grade.

Grade	Total OIB Score Points
4	86
8	89
HS	108

Item Maps

The item maps summarize the materials in the OIBs. The item map for each grade included the order of difficulty, location, form, item number, score key (correct response for a MC item), and the subskill number. Participants filled in the final two columns as they studied the items in the OIB. The first of these columns asks, "What does this item measure? That is, what do you know about a student who can respond successfully to this item?" The second of these columns asks "Why is this item more difficult than the preceding items?" Figure 1 shows the item map used for training.

Figure 1. Item Map for Training

SAMPLE Mathematics Item Map

*Print Name:*_____ *Group Number:*_____

Order of difficulty (easy to hard)	Location	Form	Item No.	Item Type	Score Key	Content Strand *	What does this item measure? That is, what do you know about a student who can respond successfully to this item?	Why is this item more difficult than the preceding items?
1	220	12	1	MC	B	1		N/A
2	225	9	4	MC	C	4		
3	229	9	3	MC	B	5		
4	240	12	2	MC	D	1		
5	241	12	4	MC	B	4		
6	262	9	5	MC	A	1		
7	303	9	6	MC	B	2		
8	321	9	8	MC	B	2		
9	401	9	9	MC	C	4		

* 1 = Number Sense, Properties, & Operations; 2 = Measurement; 3 = Geometry; 4 = Data Analysis, Statistics, & Probability; 5 = Algebra & Functions

Standard Setting: Day 1

Opening Session
Staff from the ADE and CTB welcomed the participants to the AIMS Science Standard Setting. Ms. Roberta Alley, ADE Deputy Associate Superintendent of Assessment; Mr. Frank Brashear, ADE Director of Test & Item Development; and Ms. Irene Hunting, ADE Director of State Test Administration, gave the welcoming address and described the purpose of the standard setting. The ADE described the expectations for the type of cut scores that the state anticipated from the process.

Training
Following the presentation by the ADE, Dr. Steve Ferrara, a member of the CTB Standard Setting Team, provided an overview of the purpose of the standard setting and described the implementation of the BSSP. Participants were introduced to key concepts and materials of the BSSP, including the OIB and the item map. During this training, it was explained that table leaders would facilitate discussion at their tables and help participants in completing tasks in a timely manner. Participants were given a synopsis of each day's activities. The Master Agenda is included in Section C, and handouts of the training slides are included in Section D.

Participants then engaged in a brief, mock standard setting using sample items from the 1996 publicly released test items of the National Assessment of Educational Progress State Assessment Program in Mathematics. During the mock standard setting, participants reviewed the tools of the BSSP, including a sample OIB and item map. The item map from the mock standard setting was presented previously in Figure 1. Following the mock standard setting, participants were directed to their pre-assigned breakout rooms.

Target Student Descriptions
Participants studied the AIMS Science performance level descriptions. Group leaders worked with participants to review the descriptions and to discuss the KSAs of each target student.

A target student is defined as a student whose performance minimally meets the criteria for entry into a particular performance level, for example, the "just" *Meets* student. For each grade there were three target student descriptions, one for each cut score (*Approaches*, *Meets*, and *Exceeds*). The target student descriptions served as a basis for establishing a common understanding of the type of student that should be considered when placing the bookmark. Participants were encouraged to take notes during the target student discussion and referred to the target student descriptions and performance level descriptions throughout the standard setting

Study the Performance Level Descriptions
Participants reviewed the Arizona Science content standards for their grade, as well as the performance level descriptions. Participants were encouraged to discuss the performance level descriptions and to consider the differences between each level. For example, participants were encouraged to consider the difference in test performance and in the standards which might differentiate students classified as *Meets* or *Exceeds.*

Examine the Test
Participants received a list of 20 items to examine and complete in the OIB for their grade to familiarize themselves with the items.

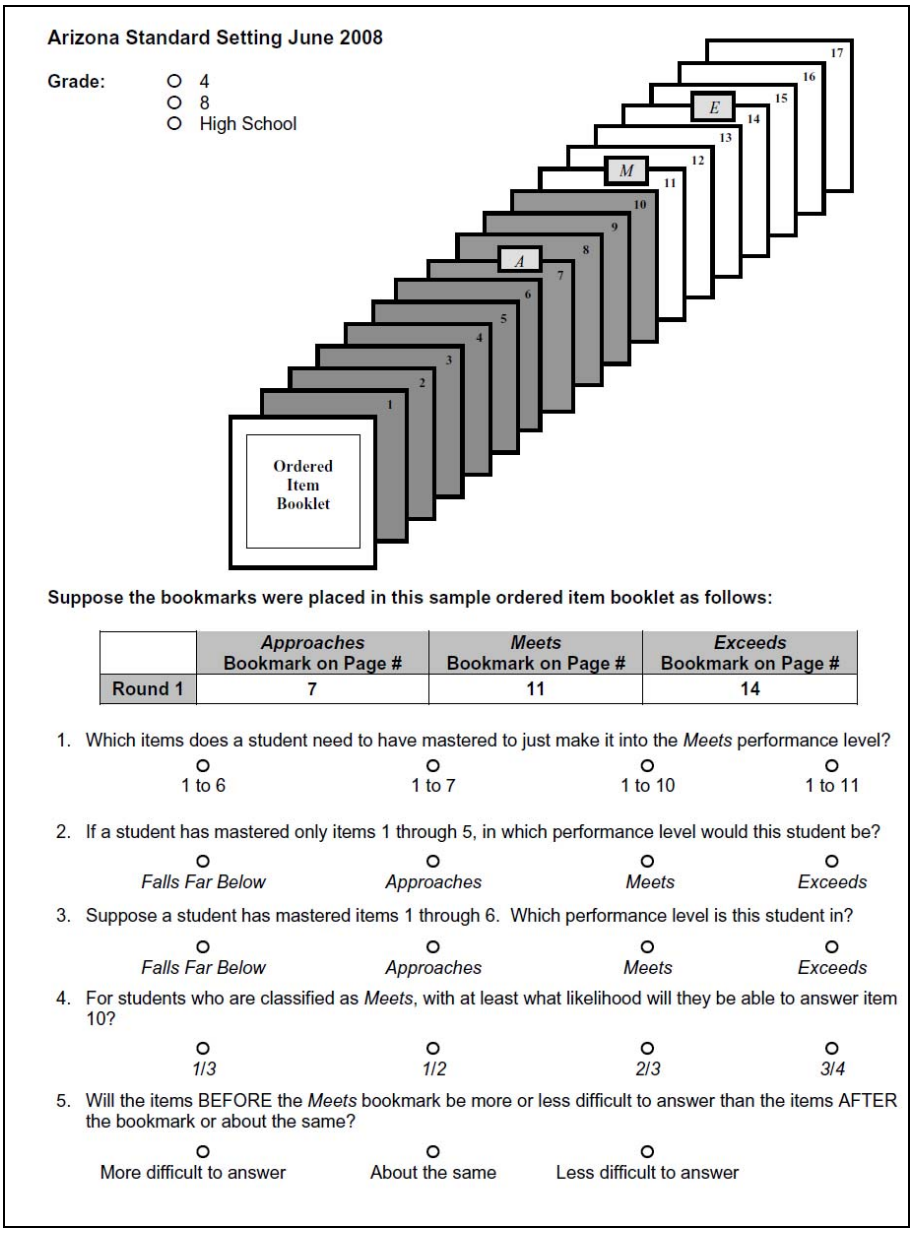
Study Items in the Ordered Item Booklet
Participants at each table studied each of the items in the OIB in terms of what each item measured and why it was more difficult than the items preceding it. Participants recorded their notes about the items on the item maps. At each table, one participant, denoted as the scribe, recorded the group's comments about each item.

Review Bookmark Placement
Prior to setting their Round 1 bookmarks, Dr. Steve Ferrara, presented a refresher of bookmark placement. Participants were instructed to use four tools when placing their bookmarks: the Arizona Science content standards, the target student descriptions, the performance level descriptions, and the KSAs represented by the items.

Participants were given training materials and three explanations of bookmark placement. The training materials titled "Bookmark Placement" and "Frequently Asked Questions about Bookmark Placement" were summarized orally to all participants. The first explanation of bookmark placement demonstrated the mechanics: participants were instructed that all items preceding the bookmark define the KSAs that a "just" *Meets* student, for example, is expected to know. The second explanation of bookmark placement was more conceptual in that participants were instructed to examine each item in terms of its KSAs and to make a judgment about the type of KSAs that a student would need to know in order to be considered, for example, "just" *Meets.* The final explanation discussed the relationship between the bookmarks and the scale scores, as described in the training material titled, "Mastery." The bookmark training materials are included in Section E.

The participants were tested on their understanding of bookmark placement with a short check set. The check set questions are presented in Figure 2. The results of the check set are presented in Table 8. After participants took the check set, Dr. Ferrara provided the correct answers and discussed the rationales for the correct answers. The responses to the check set, shown in Table 8, indicate that participants understood how to place their bookmarks. Note that two additional check sets were included in the results in Table 8. The check set (and its graphic) in Figure 2 is also included in Section E.

Figure 2. Check Set Questions



Arizona Standard Setting June 2008

Grade: ○ 4 ○ 8 ○ High School

Suppose the bookmarks were placed in this sample ordered item booklet as follows:

	Approaches Bookmark on Page #	*Meets* Bookmark on Page #	*Exceeds* Bookmark on Page #
Round 1	7	11	14

1. Which items does a student need to have mastered to just make it into the *Meets* performance level?
   ○ 1 to 6 ○ 1 to 7 ○ 1 to 10 ○ 1 to 11
2. If a student has mastered only items 1 through 5, in which performance level would this student be?
   ○ *Falls Far Below* ○ *Approaches* ○ *Meets* ○ *Exceeds*
3. Suppose a student has mastered items 1 through 6. Which performance level is this student in?
   ○ *Falls Far Below* ○ *Approaches* ○ *Meets* ○ *Exceeds*
4. For students who are classified as *Meets*, with at least what likelihood will they be able to answer item 10?
   ○ *1/3* ○ *1/2* ○ *2/3* ○ *3/4*
5. Will the items BEFORE the *Meets* bookmark be more or less difficult to answer than the items AFTER the bookmark or about the same?
   ○ More difficult to answer ○ About the same ○ Less difficult to answer

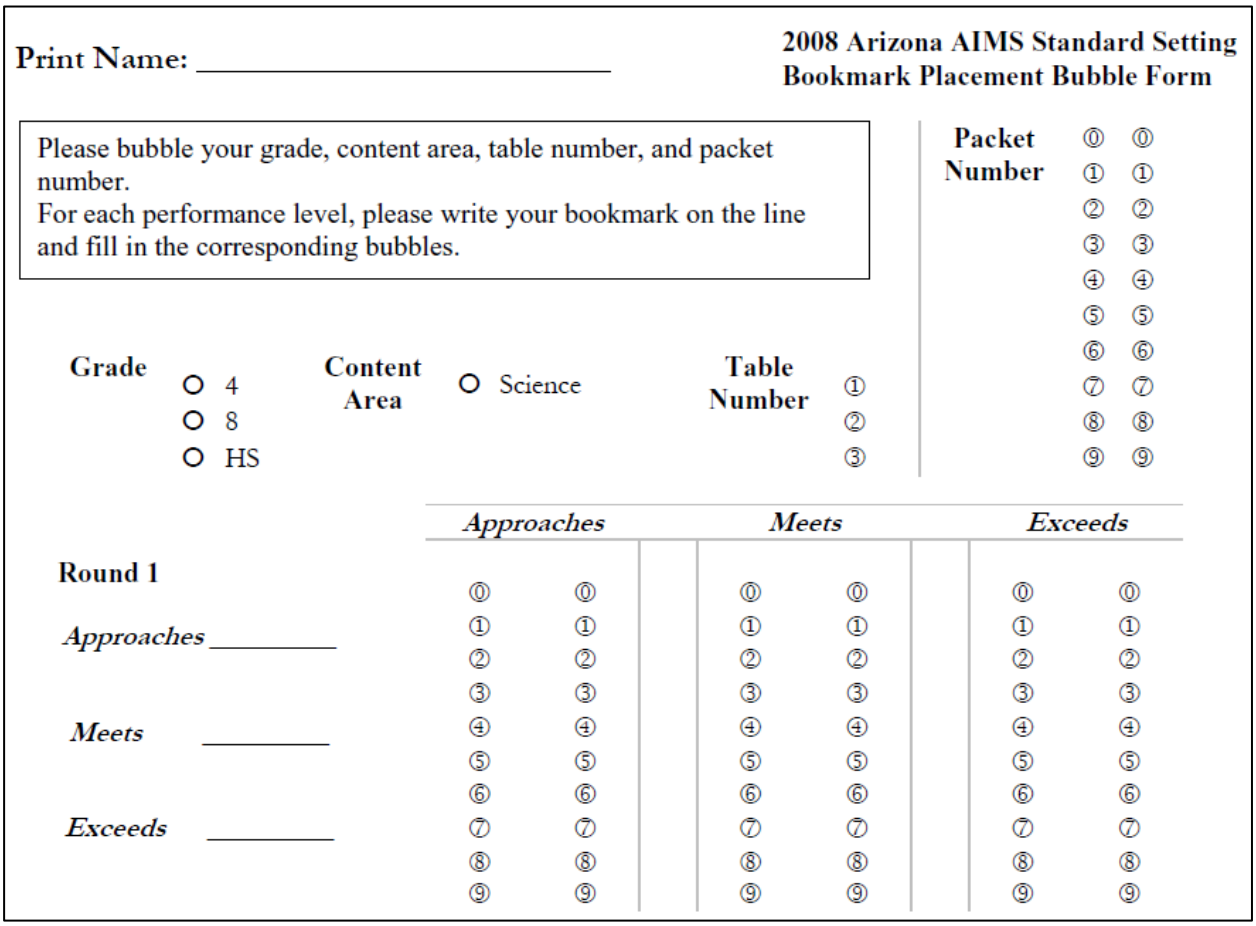
Table 8. Number and percentage of participants that correctly responded to each question on the check set (*N* = 37).

Question	Number Correct	Percent Correct
1	35	95%
2	36	97%
3	31	84%
4	36	97%
5	36	97%

Round 1 Bookmark Placement

Once participants demonstrated that they understood how to place their bookmarks through the check set, participants placed their bookmarks. The training materials indicated that the bookmarks should be placed starting with *Meets* then *Approaches* and lastly, *Exceeds*. Participants recorded their bookmark placements on a bubble form as shown in Figure 3. Participants were instructed to keep the target student descriptions in mind when completing their bookmark placement. Participants were reminded that bookmark placement is always an individual activity.

Figure 3. Sample Bookmark Placement Bubble Form



Participants placed their Round 1 bookmarks for *Approaches*, *Meets*, and *Exceeds*, while keeping in mind the Arizona Science content standards, the target student descriptions, the performance level descriptions, and the KSAs measured by the items on the test.

Round 2 Bookmark Placement
In each grade, Round 2 began with the table leader facilitating a discussion of all the bookmark placements for the table. Participants were encouraged to focus on the differences among their bookmarks by discussing the items between the lowest and highest bookmarks at their table. Participants were then directed back to their OIBs and item maps to continue their discussions of the KSAs expected of students in each performance level. After discussion, participants were reminded to place their bookmarks independently.

Round 3 Bookmark Placement
Participants received feedback based on their Round 2 bookmark placements from a member of the CTB Standard Setting Team in collaboration with an ADE representative. Participants were shown the median bookmark placement for each performance level for their grade. CTB staff answered process-related questions, and the ADE staff answered all policy-related questions.

After the presentation of Round 2 results, participants discussed the rationale of their bookmark placements within their grade. The group leader facilitated the discussion among all participants. After the discussion, participants were instructed to place their bookmarks independently for the final time.

Round 3 Results
Participants received feedback based on their final bookmark placements from a member of the CTB Standard Setting Team in collaboration the ADE. Participants were shown the median bookmarks for each table as well as the medians for their grade and the impact data based on the median final bookmarks. In addition, participants were shown the impact data for all grades as an introduction to the articulation discussion. The impact data came from the AIMS Spring 2008 administration.

Table 9 shows the participant-recommended cut scores and associated impact data based on the final round of bookmark placements. The impact data in Table 9 were shown to the participants at the workshop.

Table 9. Participant-recommended cut scores and associated impact data, based on the final round of bookmark placements.

	Cut Scores			Impact Data			
Grade	*Approaches*	*Meets*	*Exceeds*	*Falls Far Below*	*Approaches*	*Meets*	*Exceeds*
4	460	498	545	22%	25%	35%	18%
8	472	499	531	30%	20%	22%	28%
HS	497	522	559	49%	19%	20%	13%

Section F presents details of the participants' Bookmark judgments for each grade. In Section G, estimates are given of the percentages of students in each performance level at plus/minus one, two, and three standard errors of the participants' recommended final round cut scores for each grade. Section H contains graphical representations of participants' judgments. Section I contains the results of the participants' evaluation of the workshop.

Description Writing

The Group Leader introduced the process for description writing. Participants recommended changes to the existing performance level descriptions that detailed the KSAs needed to be classified in each performance level. CTB Development incorporated the changes recommended by the participants. Section J contains the original performance level descriptions used by participants at the workshop as well as the final ADE approved performance level descriptions.

Articulation (Smoothing) Discussion

Following description writing, all committee members from each grade engaged in an articulation (smoothing) discussion. The purpose of this discussion was to establish a system of cut scores that was coherent across grades while simultaneously, respectful of the committee's original recommendations. The ADE assisted CTB in facilitating these discussions because of the policy-related nature of such a discussion.

The participants of the articulation discussion recommended no changes to the cut scores for Grades 4, 8, and High School Science. Participants felt that their recommended cut scores accurately reflected their expectations for students in each performance level.

Following the standard setting, the ADE and CTB rescaled the three tests such that the *Meets* cut score for each grade was equal to 500. The final ADE approved cut scores, as well as the associated impact data are summarized in Section A. Section F contains a graphical representation of the impact data associated with the ADE approved cut scores.

Evaluations

Following the description writing and articulation discussion, participants were asked to complete an evaluation of the standard setting. Some results are presented in Tables 10–15. Complete results of the evaluation are included in Section I.

Participants were asked to respond to the statement, "Overall, I was satisfied with my group's final bookmarks." The majority of participants agreed or strongly agreed that they were satisfied with their group's final bookmarks, as shown in Table 16.

Table 10. Participants' agreement/disagreement with the statement, "Overall, I was satisfied with my group's final bookmarks."

Grade	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Agree + Strongly Agree
Overall	35	0.0%	0.0%	0.0%	31.4%	68.6%	100.0%
4	12	0.0%	0.0%	0.0%	41.7%	58.3%	100.0%
8	12	0.0%	0.0%	0.0%	16.7%	83.3%	100.0%
HS	11	0.0%	0.0%	0.0%	36.4%	63.6%	100.0%

Evaluation of Training

An indication of the effectiveness of training may be found in the participants' answers to statements and questions on the evaluations. Table 11 shows that all participants agreed or strongly agreed that they understood how to place their bookmarks. Table 12 summarizes that most participants agreed or strongly agreed that the task of bookmark placement was clear.

Table 13 shows that all participants agreed or strongly agreed that the training materials were helpful. Table 14 indicates that most participants agreed or strongly agreed that the Bookmark Procedure was well described. As Table 15 demonstrates, participants agreed or strongly agreed that the goals of the process were clear.

Table 11. Participants' agreement/disagreement with the statement, "I understood how to place my bookmarks."

Grade	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Agree + Strongly Agree
Overall	35	0.0%	0.0%	0.0%	28.6%	71.4%	100.0%
4	12	0.0%	0.0%	0.0%	50.0%	50.0%	100.0%
8	12	0.0%	0.0%	0.0%	8.3%	91.7%	100.0%
HS	11	0.0%	0.0%	0.0%	27.3%	72.7%	100.0%

Table 12. Participants' agreement/disagreement with the statement, "The training on Bookmark placement made the task clear to me."

Grade	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Agree + Strongly Agree
Overall	35	0.0%	0.0%	2.9%	37.1%	60.0%	97.1%
4	12	0.0%	0.0%	0.0%	58.3%	41.7%	100.0%
8	12	0.0%	0.0%	0.0%	25.0%	75.0%	100.0%
HS	11	0.0%	0.0%	9.1%	27.3%	63.6%	90.9%

Table 13. Participants' agreement/disagreement with the statement, "The training materials were helpful."

Grade	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Agree + Strongly Agree
Overall	35	0.0%	0.0%	0.0%	34.3%	65.7%	100.0%
4	12	0.0%	0.0%	0.0%	50.0%	50.0%	100.0%
8	12	0.0%	0.0%	0.0%	16.7%	83.3%	100.0%
HS	11	0.0%	0.0%	0.0%	36.4%	63.6%	100.0%

Table 14. Participants' agreement/disagreement with the statement, "The Bookmark Procedure was well described."

Grade	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Agree + Strongly Agree
Overall	35	0.0%	0.0%	2.9%	22.9%	74.3%	97.2%
4	12	0.0%	0.0%	0.0%	41.7%	58.3%	100.0%
8	12	0.0%	0.0%	0.0%	8.3%	91.7%	100.0%
HS	11	0.0%	0.0%	9.1%	18.2%	72.7%	90.9%

Table 15. Participants' agreement/disagreement with the statement, "The goals for the Bookmark Procedure were clear."

Grade	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Agree + Strongly Agree
Overall	35	0.0%	0.0%	0.0%	37.1%	62.9%	100.0%
4	12	0.0%	0.0%	0.0%	33.3%	66.7%	100.0%
8	12	0.0%	0.0%	0.0%	16.7%	83.3%	100.0%
HS	11	0.0%	0.0%	0.0%	63.6%	36.4%	100.0%

Quality Control Procedures

The CTB Standard Setting Team adhered to many quality control procedures to foster the accuracy of the materials used and the results presented during the workshop. Prior to the workshop, the CTB Standard Setting Team cross-checked the ordering of items in the ordered item booklets, the accuracy of the information in the item maps, and the accuracy of the Microsoft Excel macros and Bookmark Pro software used to generate results and impact data. All data were scanned on-site at the workshop. The CTB Standard Setting Team checked the reasonableness of the data presented to participants.

References

Beretvas, S.N. (2004). Comparison of Bookmark difficulty locations under different item response models. *Applied Psychological Measurement, 28,* 25-47.

Karantonis, A., & Sireci, S. (2006). The bookmark standard setting method: A literature review. *Educational Measurement: Issues and Practice.* 4–12.

Lewis, D.M., Mitzel, H.C., & Green, D.R. (1996). Standard Setting: A bookmark approach. In D.R. Green (Chair), *IRT-based standard-setting procedures utilizing behavioral anchoring*. Symposium conducted at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix, AZ.

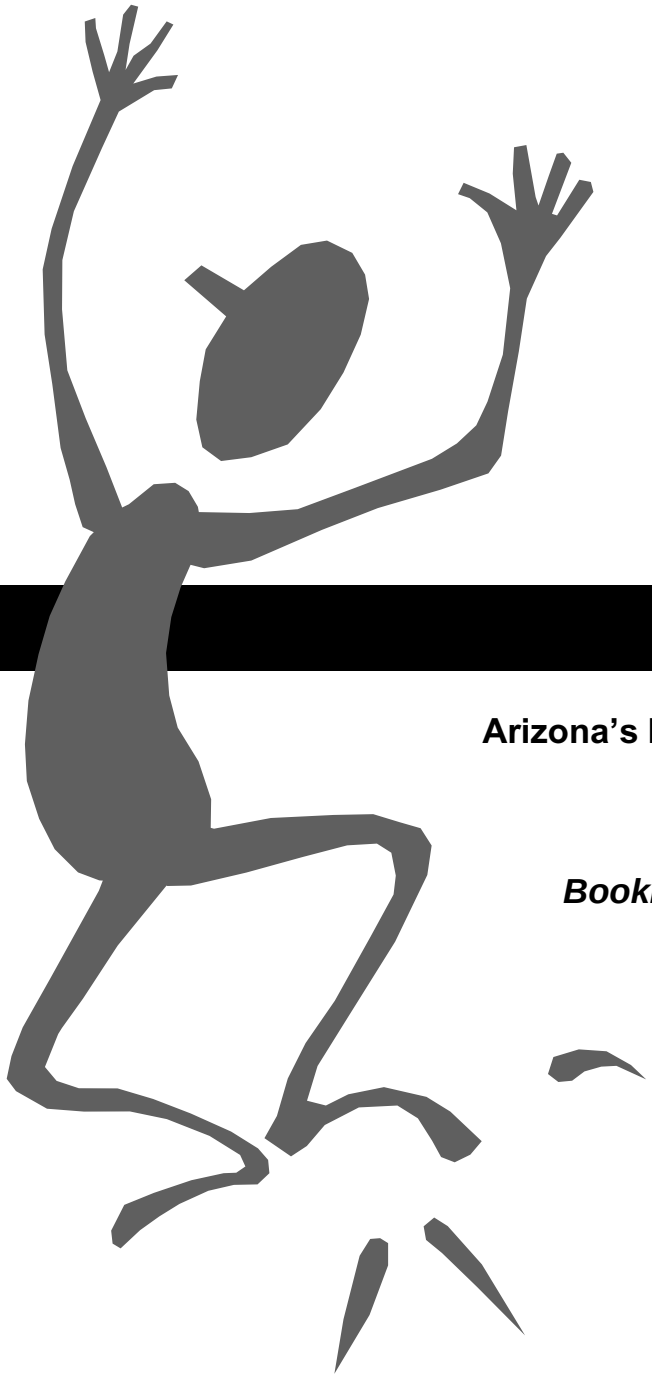
Lewis, D. M., Green, D. R., Mitzel, H. C., Baum, K., & Patz, R. J., (1998, April). The Bookmark Standard Setting Procedure: Methodology and Recent Implementations. Paper presented at the 1998 annual meeting of the National Council of Measurement in Education annual meeting, San Diego, CA.

Mitzel, H.C., Lewis, D.M., Patz, R.J., & Green, D.R. (2001). The Bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives*. (pp. 249–281). Mahwah, NJ: Lawrence Erlbaum Associates.

SECTION C

Bookmark Standard Setting Agenda

CTB McGraw-Hill



Master Agenda

Arizona's Instrument to Measure Standards
Grades 4, 8, and High School
Science

Bookmark Standard Setting Workshop

June 9 – 11, 2008
Phoenix, Arizona

Welcome to the Bookmark Standard Setting Workshop for Arizona's Instrument to Measure Standards for Science for Grades 4, 8, and High School.

The Arizona Department of Education and CTB/McGraw-Hill would like to thank you for your time and expertise during this important process.

Please use this agenda to orient yourself during the workshop. If you have any questions or concerns, please do not hesitate to contact a member of the CTB Standard Setting Team.

Monday, June 9
Welcome!

7:30 AM **Table Leader registration**
Please check in at the reception area to sign a non-disclosure agreement, get your nametag, and collect any other information.

8:00 AM **Table Leader training**
You will receive an overview of the standard setting workshop, learn how the Bookmark Standard Setting Procedure works and practice briefly, and discuss your role and responsibilities during the workshop. You will learn how to "answer the two questions" for items in the ordered item booklet.

Target Student discussion
Table Leaders engage in structured discussions about the knowledge, skills, and abilities they expect to be demonstrated by students in each performance level.

8:30 AM **Participant registration and continental breakfast**[1]
Participants check-in at the reception. Table Leaders need not register again. Continental breakfast is served.

9:00 AM **Opening session (all grade level groups together)**
All participants are formally welcomed by ADE, then CTB, and receive an overview of how the standard setting workshop will work.

Bookmark overview
Participants will be introduced to the Bookmark Standard Setting Procedure. A CTB Standard Setting Team member presents "Opening Session Slides."

[1] *A 15-minute break will be held at 10:30 am and 2:30 pm each day.*

10:45 AM **Grade level groups work in pre-assigned breakout rooms**

The Group Leader welcomes participants to the group and distributes secure materials.

- Ensure that all participants at your table write their name on each of their secure materials. Each participant's packet of secure materials consists of a test book, OIB, item map, performance level descriptors, and standards. All secure materials are printed on colored paper.

Review the test and respond to selected test items (30 minutes)

Participants review and take selected items in the secure test book.

- Although some discussion about individual test items is normal, focus your participants away from prolonged debate and toward responding to test items--independently.
- Participants use provided index cards to record comments about test items.

Review and discuss the PLDs (15 minutes)

- Review with the participants the knowedge and skill demands for the description for *Meets the Standard*. Draw their attention to terminology, knowledge, and skills in the PLDs that correspond to the science standards.
- Do the same for *Exceeds*, *Approaches*, and *Falls Far Below*.

Discuss Target Students (15 minutes)

- There are three Target Students that participants need to think about: Just *Approaches*, Just *Meets*, and Just *Exceeds*. A Target Student is a student who just makes it into a performance level. The group will review the performance level descriptors and discuss the knowledge, skills, and abilities expected of these Target Students.
- Record their ideas about the target students on easel paper.

12:00 PM **Lunch**

1:00 PM **Answer the two questions for items in the Ordered Item Booklet (in breakout rooms)**

The Group Leader introduces this exercise by instructing participants to find the Item Map in their secure materials, then reviewing the purpose of each column.

- Facilitate a discussion amongst everyone at your table about each of the items in the OIB. Start with the first item, and discuss each item in turn, focusing on what each item measures and what makes it harder than the previous items. All participants record these details on their Item Maps.
- Assign a scribe to take a master set of notes for your table.
- Remember to use the index cards, as necessary.
- Ensure that each participant at your table has a chance to speak.

4:15 PM **Secure materials collection and audit**

The Group Leader facilitates collection of the secure materials from all participants. A listing of secure materials to be collected is displayed in the room.

- Supervise the collection of secure materials at your tables. See the "Secure Materials" page in this agenda for more information.

The Group Leader asks the Table Leaders to audit the secure materials at one other's table.

- Order materials numerically by packet number within each table.
- Verify that all signed-out packets are present.
- Stack materials at each table neatly into one pile with the table tent on top, under the top packet's rubber band.
- Place the separate stacks on one table. Do not combine tables' stacks.

4:30 PM **Dismissal of participants and Table Leader debriefing**

Table Leaders discuss the events of the day and plans for the next day. Participants are dismissed.

4:45 PM **Table Leader dismissal**

8:30 AM **Continental breakfast**

9:00 AM **Orientation to bookmark placement (all grade level groups together)**
Participants reconvene in the large group meeting room. A member of the CTB Standard Setting Team introduces bookmark placement, explains and illustrates how bookmarks are placed and what bookmarks mean. After this brief presentation, a short check set is given.

10:15 AM **Round 1 bookmark placements (grade level groups in breakout rooms)**
Group Leaders check in with Table Leaders that everyone is ready for Round 1. Table Leaders assist participants as needed. Participants place bookmarks.

- Each participant should place the *Meets the Standard* bookmark first, followed by *Approaches* and *Exceeds*.
- Remind participants that bookmark placement is always an independent activity.
- Collect your participants' Bookmark Placement Bubble Forms as they complete them, ensuring that each participant has made a single, unambiguous placement for each bookmark.
- Give your participants' Bookmark Placement Bubble Forms to the Group Leader.

12:00 PM **Lunch**

12:30 PM **Discussion of Round 1 as a table**
After results are presented, Table Leaders lead a discussion about the bookmark placements made at your table.

1:30 PM **Round 2 bookmark placements**
Group Leaders check in with Table Leaders that everyone is ready for Round 2. Table Leaders assist participants as needed. Participants place bookmarks.

- Each participant should place the *Meets the Standard* bookmark first, followed by *Approaches* and *Exceeds*.
- Remind participants that bookmark placement is always an independent activity.
- Collect your participants' Bookmark Placement Bubble Forms as they complete them.
- Give your participants' Bookmark Placement Bubble Forms to the Group Leader.

3:00 PM **Discussion of Round 2 as a grade level group**
A member of the CTB Standard Setting Team presents a summary of the bookmark placements from each table to the entire group and the impact data based on the Round 2 cut scores. Then the Group Leader leads a discussion with the entire group about each bookmark, similar to the table-level discussions after Round 1.

4:15 PM **Secure materials collection and audit**

The Group Leader facilitates collection of the secure materials from all participants. A listing of secure materials to be collected is displayed in the room.

- Supervise the collection of secure materials at your tables. See the "Secure Materials" page in this agenda for more information.

The Group Leader asks the Table Leaders to audit the secure materials at one other's table.

- Order materials numerically by packet number within each table.
- Verify that all signed-out packets are present.
- Stack materials at each table neatly into one pile with the table tent on top, under the top packet's rubber band.
- Place the separate stacks on one table. Do not combine tables' stacks.

4:30 PM **Dismissal of participants and Table Leader debriefing**

Table Leaders discuss the events of the day and plans for the next day. Participants are dismissed.

4:45 PM **Table Leader dismissal**

8:30 AM **Continental breakfast**

9:00 AM **Round 3 bookmark placements**
Group Leaders check in with Table Leaders that everyone is ready for Round 3. Table Leaders assist participants as needed. Participants place bookmarks.

- Each participant should place the *Meets the Standard* bookmark first, followed by *Approaches* and *Exceeds*.
- Remind participants that bookmark placement is always an independent activity.
- Collect your participants' Bookmark Placement Bubble Forms as they complete them.
- Give your participants' Bookmark Placement Bubble Forms to the Group Leader.

10:15 AM **Presentation of final recommendations (all grade level groups together)**
A member of the CTB Standard Setting Team presents presents a summary of Round 3 final recommended cut scores.

10:45 AM **Cross-grade articulation discussion (all grade level groups together)**
Participants from all grade levels will discuss their grade level cut scores and impact data. During these discussions, participants will discuss the knowledge, skills, and abilities they expect of students in each performance level.

- As a group, the participants will examine the bookmark placements and impact data as a multi-grade system of performance standards. If needed, the group will make recommendations to adjust some bookmarks or not ot adjust them.
- All decisions will be supported by a brief written rationale.

12:00 PM **Lunch**

1:00 PM **Refinement of performance level descriptors at each grade level**
The Group Leader presents instructions for refining PLDs.

- Your group's descriptors should synthesize the knowledge, skills, and abilities necessary to respond successfully to each of the items mapped to each performance level.

2:00 PM **Refinement of performance level descriptors, across grade levels**
Participants will ensure that PLDs are appropriately articulated across grade levels.

2:45 PM **Workshop evaluations**
Each participant completes an evaluation of the standard setting.

3:15 PM **Secure materials collection and audit**

The Group Leader facilitates collection of the secure materials from all participants. A listing of secure materials to be collected is displayed in the room.

- Supervise the collection of secure materials at your tables. See the "Secure Materials" page in this agenda for more information.

The Group Leader asks the Table Leaders to audit the secure materials at one other's table.

- Order materials numerically by packet number within each table.
- Verify that all signed-out packets are present.
- Stack materials at each table neatly into one pile with the table tent on top, under the top packet's rubber band.
- Place the separate stacks on one table. Do not combine tables' stacks.

3:30 PM **Dismissal**

The Arizona Department of Education and CTB/McGraw-Hill thank you for your time and participation!

Why do we do Secure Materials Collection?

A thorough collection of secure test materials protects both the reliability of the testing program and the substantial monetary investment in the assessment. A structured method of collection has been established to gather effectively all of the secure material at the workshop. Each day as you facilitate secure materials collection at your table, refer to this guide for instructions and suggestions.

During the collection, participants should place each secure item, one at a time, in a pile on the table in front of them. After the process, each participant will have a single stack of materials, each stacked in the same way as everyone else in the room. Please follow these steps to facilitate the process.

How do I do Secure Materials Collection?

1. Get the attention of all the participants at your table. Discourage any side conversations or inattention.

2. Using the list provided, call out each item, one at a time, and watch participants place that item on their stack. Discourage participants from moving ahead. Ensure that participants have placed the item in their stack before moving on.

3. Proceed through the list until each piece of secure material has been collected. Direct participants to place a rubber band around their stack when completed.

4. If any participants wish to leave additional items with their materials overnight, encourage them to place it beneath their stack, inside the rubber band.

5. Table Leaders will audit the secure materials at one other table.

6. Once you have supervised the collection of secure materials and are satisfied that all items have been collected, inform the Group Leader.

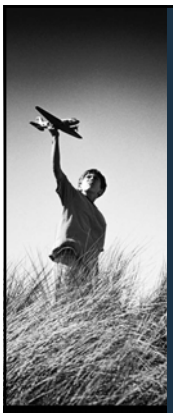
7. The collected materials are stored overnight and will be available in the morning.

What should I expect from Secure Materials Collection?

Generally, secure materials collection goes smoothly. If you have any questions about the collection process, or if you have a concern about test security at the standard setting workshop, please contact your Group Leader or a member of the CTB Standard Setting Team.

SECTION D

Handouts of Slides for Opening Session and Bookmark Training



Setting the Standard

For Arizona's Instrument to Measure Standards (AIMS)

Science Grades 4, 8, and High School

Opening Session

CTB McGraw-Hill

What is standard setting?

- A process that enables experts to make judgments about the knowledge, skills, and abilities that students should know and be able to do to be classified as *Meets the Standard*
  - Also, *Approaches the Standard* and *Exceeds the Standard*

CTB McGraw-Hill

Why standard setting?

- Content standards define what students are tested on.
  - These are things students *should* know and be able to do.
  - Arizona has content standards in Science.
- Performance standards define what students in each performance level *can do*.
  - You will actively discuss your expectations of students in each performance level.

CTB McGraw-Hill

Performance levels

- Performance levels specify what students in Arizona should know and be able to do to be categorized as *Approaches the Standard, Meets the Standard,* or *Exceeds the Standard* (or *Falls Far Below the Standard*).
- Performance Level Descriptors (PLDs) describe knowledge and skills at each level.

CTB McGraw-Hill

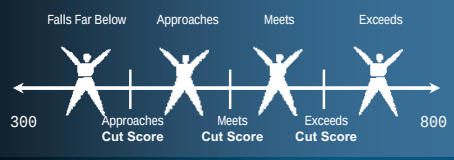
How do we set our standards?

- Percentages
  - Arbitrary
  - Test-specific
  - Does not consider content
- Content
  - Uses pre-established content standards
  - Considers the educational objectives
- Bookmark Standard Setting Procedure

CTB McGraw-Hill

Purpose of the standard setting

- Enables cut scores to be set on the test scale
- The test scale represents performance by students at higher (or lower) performance levels



CTB McGraw-Hill

Purpose of the standard setting

- Set three cut scores on the test scale
- Students who meet or exceed a cut score have demonstrated enough knowledge and skills to be categorized as *Meets the Standard* on the AIMS assessments.
  - Also *Approaches the Standard* and *Exceeds the Standard.*
- Content decisions will be based on Arizona content standards.

CTB McGraw-Hill

Bookmark standard setting

- Item-centered method
- Content-based decisions

CTB McGraw-Hill

Committee roles

- Group Leaders
- Table Leaders
- Participants
- ADE
- CTB

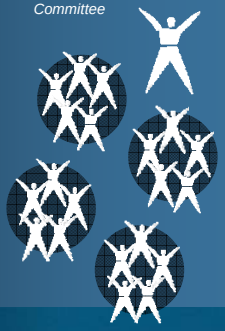
Standard Setting Committee

CTB McGraw-Hill

Committee roles

- Group Leader
  - Facilitator
    - Participants stay focused on task
    - Participants interact with their own group
    - Participants finish in a timely manner
    - Leads discussion
  - Materials collection
    - Secure materials

Standard Setting Committee

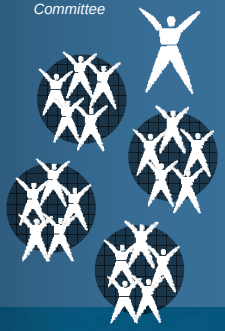


CTB McGraw-Hill

Committee roles

- Table Leaders
  - Lead discussion at the table
  - Standard setters
- Participants
  - Standard setters

Standard Setting Committee



CTB McGraw-Hill

Workshop overview

- Round 1
  - Take the test, study the PLDs, discuss the target students, answer the two questions
  - Place your bookmarks (independent)
- Round 2
  - Discuss bookmarks at tables
  - Place your bookmarks (independent)
- Round 3
  - Discuss bookmarks and impact as a grade group
  - Place your bookmarks (independent)

CTB McGraw-Hill

Ordered Item Booklets

- One item per page
- Easiest item first, hardest item last
- Items increase in difficulty

CTB McGraw-Hill

Item Map

Print Name:_____ Group Number:_____

Order of difficulty (easy to hard)	Location	Form	Item No.	Item Type	Score Key	Content Strand *	What does this item measure? That is, what do you know about a student who can respond successfully to this item?	Why is this item more difficult than the preceding items?
1	220	12	1	MC	B	1		N/A
2	225	9	4	MC	C	4		
3	229	9	3	MC	B	5		
4	240	12	2	MC	D	1		
5	241	12	4	MC	B	4		
6	262	9	5	MC	A	1		
7	303	9	6	MC	B	2		
8	321	9	8	MC	B	2		
9	401	9	9	MC	C	4		

* 1 = Number Sense, Properties, & Operations; 2 = Measurement; 3 = Geometry; 4 = Data Analysis, Statistics, & Probability; 5 = Algebra & Functions

CTB McGraw-Hill

Ordered item 1

1

1. Kitty is taking a trip on which she plans to drive 300 miles each day. Her trip is 1,723 miles long. She has already driven 849 miles. How much farther must she drive?

A. 574 miles

B. 874 miles

C. 1,423 miles

D. 2,872 miles

CTB McGraw-Hill

Item I

Subtraction, operations, eliminate distractors

Print Name:

Order of difficulty (easy to hard)	Location	Form	Item No.	Item Type	Score Key	Content Strand *	student who can respond successfully to this item?	Why is this item more difficult than the preceding item?
1	220	12	1	MC	B	1	Subtraction, operations, eliminate distractors	N/A
2	225	9	4	MC	C	4		
3	229	9	3	MC	B	5		
4	240	12	2	MC	D	1		
5	241	12	4	MC	B	4		
6	262	9	5	MC	A	1		
7	303	9	6	MC	B	2		
8	321	9	8	MC	B	2		
9	401	9	9	MC	C	4		

* 1 = Number Sense, Properties, & Operations; 2 = Measurement; 3 = Geometry; 4 = Data Analysis, Statistics, & Probability; 5 = Algebra & Functions

CTB McGraw-Hill

Ordered item 2

2

CARTONS OF EGGS SOLD LAST MONTH

Farm A ❍ ❍ ❍ ❍

Farm B ❍ ❍ ❍ ❍ ❍ ❍

Farm C ❍ ❍ ❍

Each ❍ = 100 Cartons

4. According to the graph, how many cartons of eggs were sold altogether by farms A, B, and C last month?
   - A. 13
   - B. 130
   - C. 1,300
   - D. 13,000

CTB McGraw-Hill

Agenda

- Opening session
- Take the test (selected items)
  - Individual activity
- Review the PLDs and discuss the target student
  - Table activity
- Study the ordered item booklet—answer the two questions
  - Table activity

CTB McGraw-Hill

Agenda

- Make Round 1 bookmark placements
  - Individual activity—independent!
- Round 2
  - Review Round 1 results in tables
  - Discuss in tables
  - Make new judgments individually—independent!

CTB McGraw-Hill

Agenda

- Round 3
  - Review Round 2 results and impact data as a grade group
  - Discuss as a grade group
  - Make new judgments individually—independent!
- Review final results
- Conduct cross grade articulation
- Refine PLDs
- Evaluate the standard setting workshop

CTB McGraw-Hill

Agenda

- Cross-grade articulation
  - Participants engage in a cross grade discussion to consider articulating (i.e., smoothing) the results
- Refine the PLDs
  - Participants refine the knowledge and skills described for each performance level

CTB McGraw-Hill

Bookmark Training

Arizona's Instrument to Measure Standards
Grades 4, 8, and High School
Science
June 2008

Bookmark Placement

- Items preceding the Bookmark reflect the knowledge and skills that students should know and do to be classified as *Meets the Standard.*
  - For MC items this means that students who reach *Meets the Standard* would most likely know the correct responses.
  - For CR items, they would most likely earn all the score points before the bookmark.

2

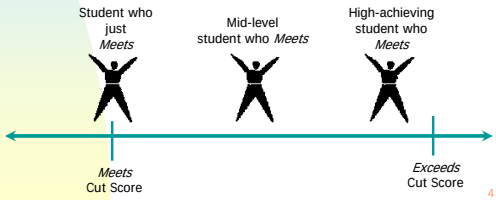
Bookmark Placement

- Place the bookmark on the first page where you judge that a student who has the knowledge and skills to demonstrate mastery of the items before the bookmark would be classified as *Meets the Standard*

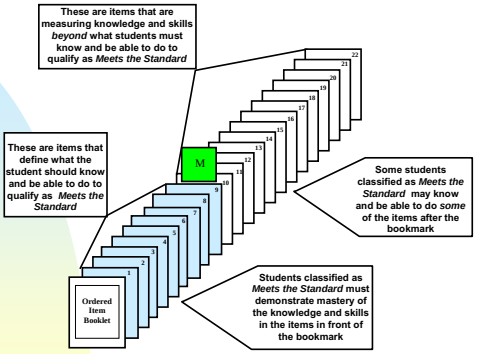
3

Target Student

- We want to describe the skills held in *common* by *all* these students
  - These are the skills of the student who *just meet the standard*

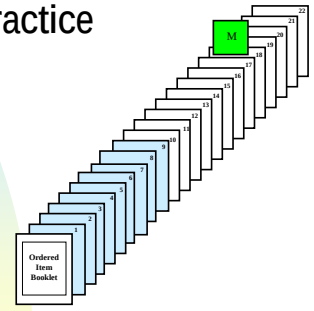


4



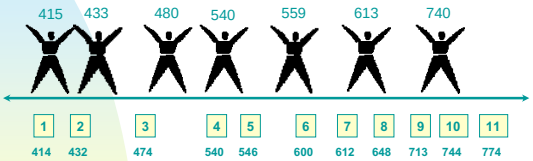
5

Practice



6

Test Scale

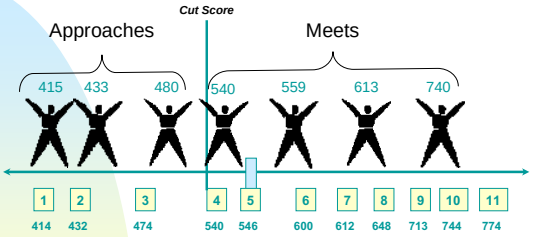


Items ordered by difficulty

Students ordered by achievement

7

The Bookmark and the Cut Score



The bookmark separates items.

The cut score separates students.

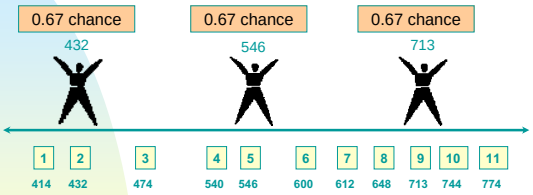
8

Mastery

- Students show mastery when they have at least a 2/3 chance of answering an item correctly.
  - Decision to use 2/3 based on research

9

Item Location

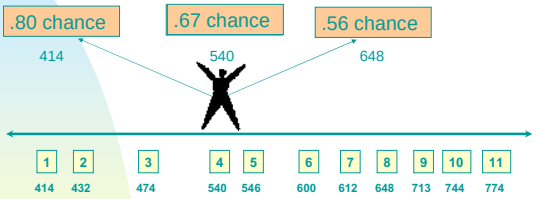


Location is an indication of difficulty.

Location represents knowledge and skills necessary to have a .67 chance of answering the item correctly.

10

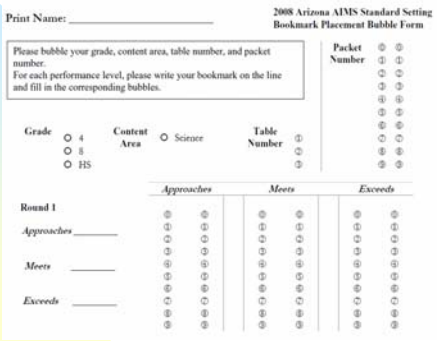
Mastery and the Target Student



A student right at the cut score will have at least a 2/3 chance of answering the items at and below the cut score correctly.

11

Bookmark Placement Bubble Form


Print Name: _____

2008 Arizona AIMS Standard Setting
Bookmark Placement Bubble Form

Please bubble your grade, content area, table number, and packet number.
For each performance level, please write your bookmark on the line and fill in the corresponding bubbles.

Grade ○ 4 ○ 8 ○ HS

Content Area ○ Science

Table Number

Packet Number

Approaches | Meets | Exceeds

Round 1

Approaches _____

Meets _____

Exceeds _____

12

Orange Sheet

Table Leader: After participants have individually recorded their votes on their Rating Forms, please record the votes on this sheet and keep it for Round 2 discussions.

Please turn in the Round 1 bookmark placement bubble forms to the Group Leader. Thank you!

Table Number _____

Participant	*Approaches* Bookmark on Page:	*Meets* Bookmark on Page:	*Exceeds* Bookmark on Page:
1			
2			
3			
4			

13

Sample Results

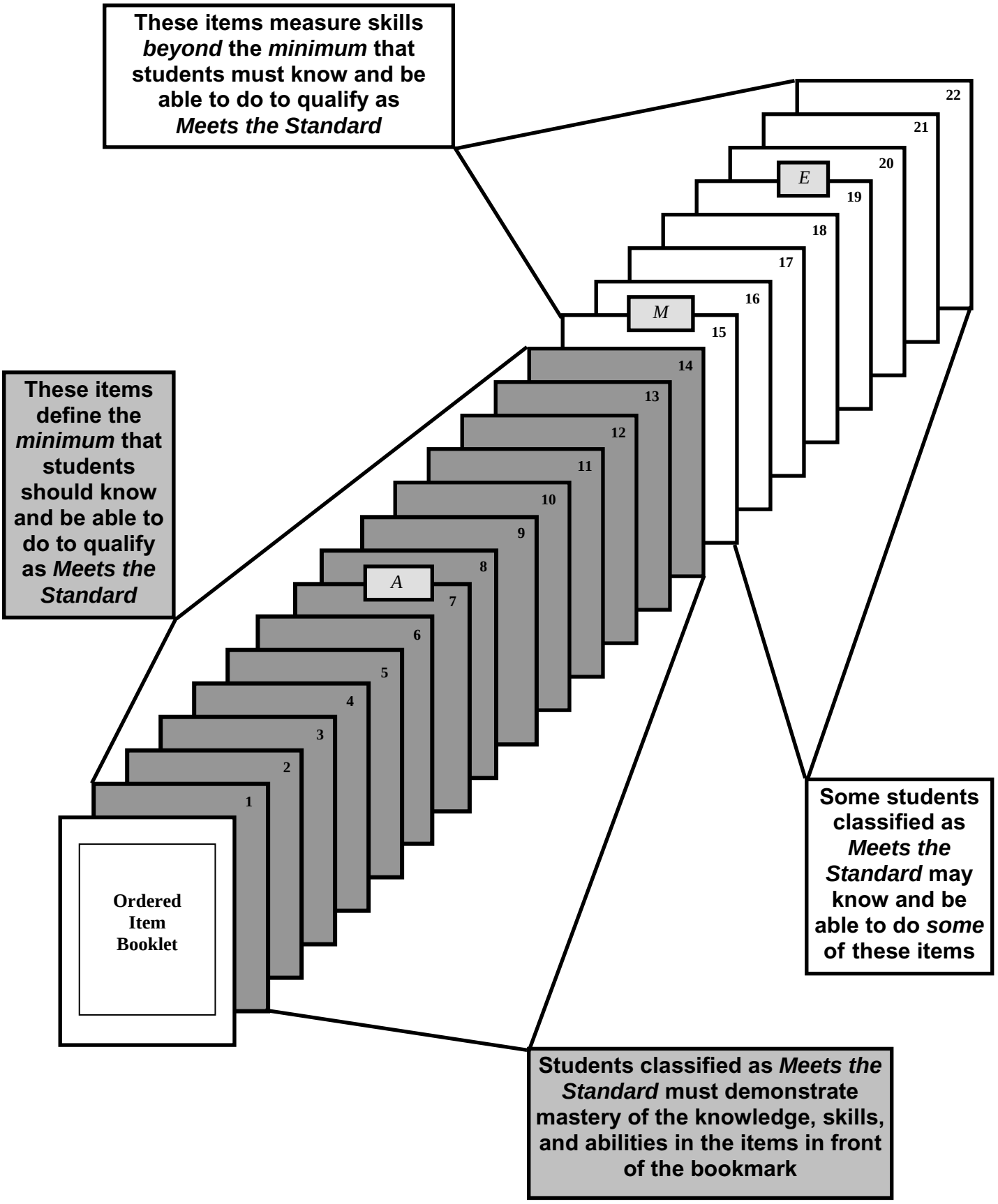
	Approaches Bookmark	*Meets* Bookmark	*Exceeds* Bookmark
Table 1	15	34	86
Table 2	11	37	82
Table 3	14	34	81
Overall Median	13	34	82

Impact Data: estimated percent of students in each performance level based on the current Large Group median			
FFB	Approaches	Meets	Exceeds
0%	0%	0%	0%

14

SECTION E

Training Materials



Bookmark Placement

These directions are written for placing the *Meets the Standard* bookmark and apply analogously to the *Exceeds the Standard* and the *Approaches the Standard* bookmarks.

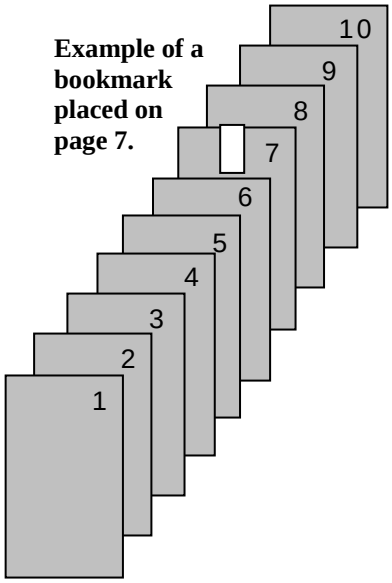
For whom am I placing this bookmark? The Target Student

When you place your *Meets the Standard* bookmark, you are separating students with a higher proficiency level in the *Approaches the Standard* level from students with a lower proficiency level in the *Meets the Standard* level. In other words, you are keeping in mind the Target Student who will just make it into the *Meets the Standard* level.

How do I place my bookmark? The Mechanics

The bookmark is exactly that: a bookmark. It separates the knowledge, skills, and abilities that students are expected to master from the knowledge, skills, and abilities they are *not* expected to master. In the example below, a participant has placed the *Meets the Standard* bookmark on page 7. With this bookmark placement, the participant says that a student must master the knowledge, skills, and abilities represented by items 1 through 6 to be classified as *Meets the Standard*.

To place your bookmark, start at page 1 in the Ordered Item Booklet (OIB). Page through the OIB **looking at the knowledge, skills, and abilities covered** until you find the ***first*** page where you think a student has demonstrated a sufficient body of evidence to indicate that the student is *Meets the Standard* relative to the content standards. This is the knowledge, skills, and abilities you are saying a *Meets the Standard* Target Student needs to master to just make it into the *Meets the Standard* level.



Hold the pages that contain the knowledge, skills, and abilities you expect the student to master in your left hand. Place your bookmark on the page AFTER the last item you expect the student to master. This page number is your bookmark. Write it on your Bookmark Placement Bubble Form.

Hint: *It may be helpful to first identify the interval of items in which you are reasonably certain the bookmark should be placed; then you can place the bookmark within that interval. If you are uncertain about where to place your bookmark, make your best decision; you will have two more rounds of voting to reconsider your bookmark.*

What does my *Meets the Standard* Bookmark mean? Some Answers

- You expect students classified as *Meets the Standard* to master the knowledge, skills, and abilities contained in the items *before* your bookmark.
- Students classified as *Meets the Standard* should know and be able to do the items *before* the bookmark. For multiple-choice items, students classified as *Meets the Standard* should know the correct response.

Is my bookmark the same as a raw score? NO

It is very important to remember that your bookmark placement is *not* equal to a raw score. In the example above, the *Meets the Standard* bookmark was placed on page 7. The participant was *not* saying that a student must get six items correct to be classified as *Meets the Standard*. This participant is saying that a barely *Meets the Standard* student must master the knowledge, skills, and abilities measured by the items on pages 1 through 6. The numbers in the OIB correspond to the rank order of difficulty of each item. These numbers do *not* correspond to a raw score.

Frequently Asked Questions about Bookmark Placement

These questions are written in reference to the *Meets the Standard* bookmark and apply analogously to the *Exceeds the Standard* and the *Approaches the Standard* bookmarks.

How do I know if I placed my bookmark in the "right" place?

The "right" place is a matter of judgment, *your* judgment. You are placing your bookmark based on the knowledge, skills, and abilities you expect students to know and be able to do.

I set my bookmark based on the knowledge, skills, and abilities I expect students to know and be able to do, that is, the knowledge, skills, and abilities I expect students to master. What is the definition of mastery?

We look at mastery by considering the likelihood with which students will respond correctly to the items. This question is answered in more depth in the handout "Mastery."

If a student misses some items before the *Meets the Standard* bookmark and gets some correct after the bookmark, is that student still *Meets the Standard*?

A student does *not* have to get every item before the bookmark correct to be classified as *Meets the Standard*. Students classified as *Meets the Standard* can miss some items *before* the bookmark and correctly respond to some items *after* the bookmark.

Does the page number on which I place my bookmark correspond to the raw score a student must get on the test?

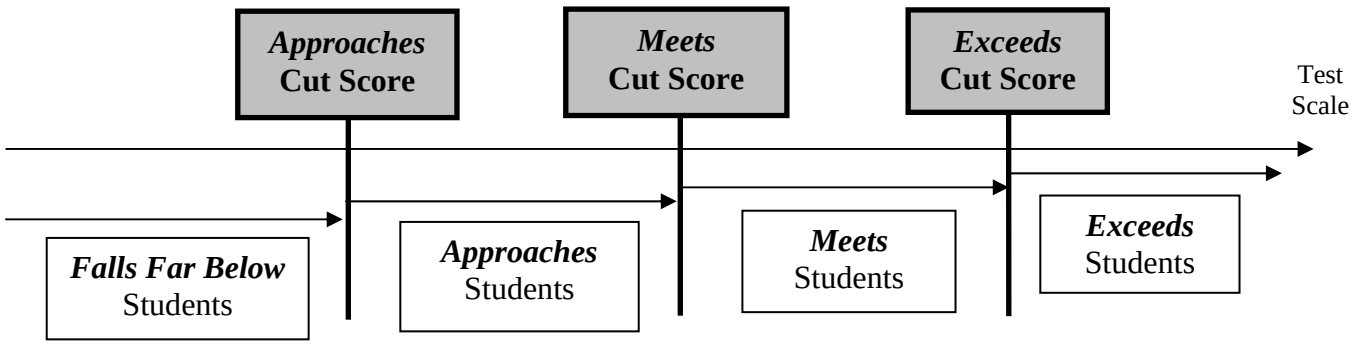
No. Remember, you are placing your bookmark based on the knowledge, skills, and abilities you expect students to master. You are *not* making your decision based on the number of items students must answer correctly. The bookmark is placed on a *page* in the Ordered Item Booklet. This page number corresponds to the difficulty ordering of the item, *not* to the raw score.

Should I place my bookmark in the first place in the Ordered Item Booklet where all the content standards have occurred?

Not necessarily. The test only samples the domain. In some cases, some standards will only be represented by difficult items that would be hard for most students to master.

How many bookmarks do I set?

You set one less bookmark than the number of performance levels. For Arizona's Instrument to Measure Standards tests, you will set three bookmarks to separate students into four performance levels.



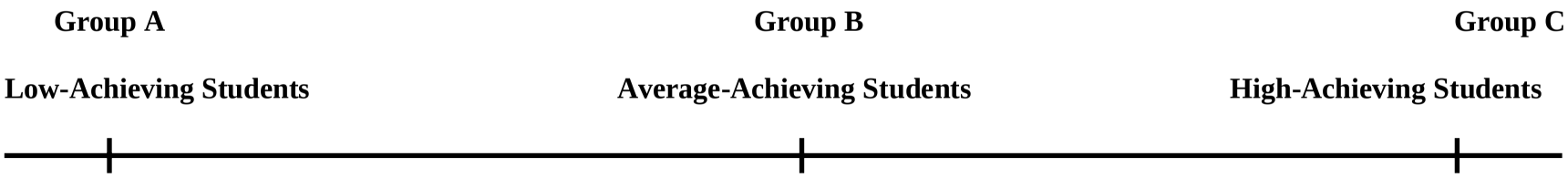
How Participants' Bookmark Judgments Relate to Expected Student Performance within Each Performance Level

You are participating in this standard setting because of your experience with students and your knowledge of the state standards, curriculum, and current instructional practices. You will be making judgments that will operationalize your expectations for the level of performance students must demonstrate in order to place in each performance level. To understand how your judgments relate to expected student performance within each performance level, consider the following examples.

Consider how students at various scale locations might perform on an imaginary assessment that consists of a total of 50 score points. In particular, we will consider the performance of groups of students who are at three specific points on the test scale. Group A consists of 100 low-achieving students, Group B consists of 100 average-achieving students, and Group C consists of 100 high-achieving students. Assume that the students have all taken the assessment and that the 100 students within each group have all obtained the exact same scale score. Note the location of the obtained scale score for each of the three groups on the test scale below.

E4

Test Scale



The following three figures show how students in each of the three groups might perform on the assessment.

Figure A shows how many students in Group A responded correctly to each item in the ordered item booklet. Observe that the students in Group A performed well on the items that appear early in the ordered item booklet but performed poorly on the items that appear later in the ordered item booklet. This makes sense, because the items appear in order of difficulty, with the easiest item first and the hardest item last. For example, 99 of the 100 Group A students responded correctly to item 1, 67 of the Group A students responded correctly to item 10, but only 1 of the Group A students responded correctly to item 50.

We say that *a group of like students have demonstrated mastery of the skills represented by an item if at least 2/3 of the students (about 67 out of 100) in the group can be expected to respond successfully to the item.* According to Figure A, Group A students have demonstrated mastery of items 1 through 10, but have not demonstrated mastery of items 11 through 50.

Figure A. The number (or percent) of Group A students who responded correctly to each item in the ordered item booklet.

item 1	item 2	item 3	item 4	item 5	item 6	item 7	item 8	item 9	item 10
99 / 100	95 / 100	93 / 100	87 / 100	83 / 100	82 / 100	78 / 100	74 / 100	69 / 100	67 / 100

item 11	item 12	item 13	item 14	item 15	item 16	item 17	item 18	item 19	item 20	item 21	item 22	item 23	item 24	item 25	item 26	item 27	item 28	item 29	item 30
63 / 100	60 / 100	59 / 100	58 / 100	57 / 100	53 / 100	52 / 100	50 / 100	50 / 100	49 / 100	49 / 100	48 / 100	47 / 100	43 / 100	41 / 100	39 / 100	37 / 100	35 / 100	34 / 100	31 / 100

item 31	item 32	item 33	item 34	item 35	item 36	item 37	item 38	item 39	item 40	item 41	item 42	item 43	item 44	item 45	item 46	item 47	item 48	item 49	item 50
30 / 100	29 / 100	25 / 100	22 / 100	20 / 100	18 / 100	17 / 100	14 / 100	11 / 100	10 / 100	9 / 100	7 / 100	5 / 100	5 / 100	4 / 100	3 / 100	2 / 100	2 / 100	1 / 100	1 / 100

Definition of Mastery

We say that a group of like students have demonstrated mastery of the skills represented by an item if at least 2/3 (67/100) of the students in the group can be expected to respond successfully to the item.

Figure B shows how many students in Group B responded correctly to each item in the ordered item booklet. Observe that the students in Group B performed much better than students in Group A. That makes sense because Group B students are average-achieving students while Group A students are low-achieving students.

Before you read further, use Figure B and the definition of mastery stated in the box above to determine which items Group B has mastered.

Group B students have demonstrated mastery of the skills reflected in items 1 through 30 of the ordered item booklet, but have not demonstrated mastery of the skills reflected by items 31 through 50. This is true according to the definition, because at least 67 of the 100 Group B students responded successfully to each of items 1 through 30, but fewer than 67 of them responded correctly to items 31 through 50.

Figure B. The number (or percent) of Group B students who responded correctly to each item in the ordered item booklet.

item 1	item 2	item 3	item 4	item 5	item 6	item 7	item 8	item 9	item 10
99 / 100	99 / 100	99 / 100	99 / 100	99 / 100	98 / 100	98 / 100	98 / 100	97 / 100	97 / 100

item 11	item 12	item 13	item 14	item 15	item 16	item 17	item 18	item 19	item 20	item 21	item 22	item 23	item 24	item 25	item 26	item 27	item 28	item 29	item 30
96 / 100	96 / 100	95 / 100	93 / 100	89 / 100	85 / 100	84 / 100	83 / 100	83 / 100	81 / 100	79 / 100	79 / 100	78 / 100	73 / 100	72 / 100	72 / 100	71 / 100	70 / 100	69 / 100	67 / 100

item 31	item 32	item 33	item 34	item 35	item 36	item 37	item 38	item 39	item 40	item 41	item 42	item 43	item 44	item 45	item 46	item 47	item 48	item 49	item 50
65 / 100	63 / 100	63 / 100	61 / 100	58 / 100	57 / 100	57 / 100	55 / 100	55 / 100	54 / 100	53 / 100	53 / 100	52 / 100	51 / 100	44 / 100	41 / 100	39 / 100	37 / 100	35 / 100	33 / 100

Definition of Mastery

We say that a group of like students have demonstrated mastery of the skills represented by an item if at least 2/3 (67/100) of the students in the group can be expected to respond successfully to the item.

Figure C shows how many students in Group C responded correctly to each item in the ordered item booklet. Observe that Group C performed much better than Groups A or B. That makes sense because Group C consists of high-achieving students while Groups A and B consist of low-and average-achieving students, respectively.

Before you read further, use Figure C and the definition of mastery stated in the box above to determine which items Group C has mastered. Group C students have demonstrated mastery of the skills reflected in items 1 through 45 of the ordered item booklet, but have not demonstrated mastery of the skills reflected by items 46 through 50. This is true according to the definition, because at least 67 of the 100 Group C students responded successfully to each of items 1 through 45, but fewer than 67 of them responded correctly to items 46 through 50.

Figure C. The number (or percent) of Group C students who responded correctly to each item in the ordered item booklet.

item 1	item 2	item 3	item 4	item 5	item 6	item 7	item 8	item 9	item 10
99	99	99	99	99	99	99	99	97	97
100	100	100	100	100	100	100	100	100	100

item 11	item 12	item 13	item 14	item 15	item 16	item 17	item 18	item 19	item 20	item 21	item 22	item 23	item 24	item 25	item 26	item 27	item 28	item 29	item 30
97	97	95	95	94	93	92	92	91	89	89	89	88	88	88	87	87	86	85	84
100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100

item 31	item 32	item 33	item 34	item 35	item 36	item 37	item 38	item 39	item 40	item 41	item 42	item 43	item 44	item 45	item 46	item 47	item 48	item 49	item 50
83	81	81	81	80	80	79	78	77	75	74	72	70	68	67	64	58	53	49	46
100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100

You have seen from the above examples that by using a specific definition of mastery, we can identify the skills in the ordered item booklet that students at any location of the test scale have mastered.

Also, if *you* identify a set of items in the ordered item booklet, the specific point on the test scale at which students have demonstrated mastery of the skills you have specified can be determined. This is how the various cut scores are ascertained.

As experts, you will first specify the skills in the ordered item booklet that you expect students to master in order to be classified as ***Meets***. This means that you will identify the items that reflect the knowledge, skills, and abilities you expect all ***Meets*** students to master. When you have made that judgment, the point on the scale at which students achieve that level of mastery can be identified.

SAMPLE Mathematics Item Map

Print Name:_____ ***Group Number:***_____

Order of difficulty (easy to hard)	Location	Form	Item No.	Item Type	Score Key	Content Strand *	What does this item measure? That is, what do you know about a student who can respond successfully to this item?	Why is this item more difficult than the preceding items?
1	220	12	1	MC	B	1		**N/A**
2	225	9	4	MC	C	4		
3	229	9	3	MC	B	5		
4	240	12	2	MC	D	1		
5	241	12	4	MC	B	4		
6	262	9	5	MC	A	1		
7	303	9	6	MC	B	2		
8	321	9	8	MC	B	2		
9	401	9	9	MC	C	4		

* 1 = Number Sense, Properties, & Operations; 2 = Measurement; 3 = Geometry; 4 = Data Analysis, Statistics, & Probability; 5 = Algebra & Functions

SAMPLE

Standard Setting Workshop

Grade 4

Mathematics

Ordered Item Booklet

Publicly released items from the National Assessment of Educational Progress 1996 State Assessment Program in Mathematics.

1. Kitty is taking a trip on which she plans to drive 300 miles each day. Her trip is 1,723 miles long. She has already driven 849 miles. How much farther must she drive?

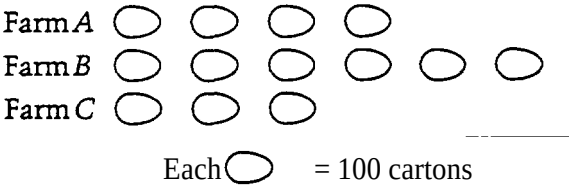
Ⓐ 574 miles

Ⓑ 874 miles

Ⓒ 1,423 miles

Ⓓ 2,872 miles

CARTONS OF EGGS SOLD LAST MONTH



4. According to the graph, how many cartons of eggs were sold altogether by farms *A*, *B*, and C last month?

Ⓐ 13

Ⓑ 130

Ⓒ 1,300

Ⓓ 13,000

3. *N* stands for the number of stamps John had. He gave 12 stamps to his sister. Which expression tells how many stamps John has now?

Ⓐ N+12

Ⓑ N–12

Ⓒ 12- N

Ⓓ 12 x N

2. A whole number is multiplied by 5. Which of these could be the result?

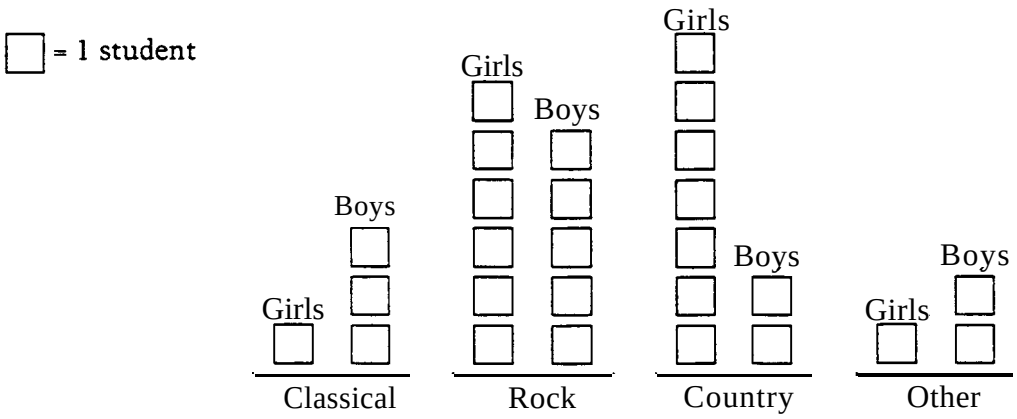
Ⓐ 652

Ⓑ 562

Ⓒ 526

Ⓓ 265

4. Each boy and girl in the class voted for his or her favorite kind of music. Here are the results.



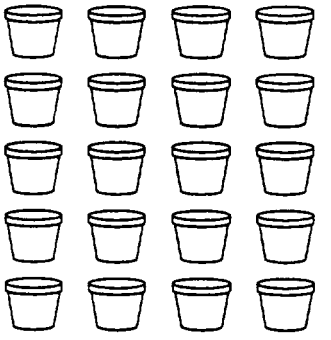
Which kind of music did most students in the class prefer?

Ⓐ Classical

Ⓑ Rock

Ⓒ Country

Ⓓ Other



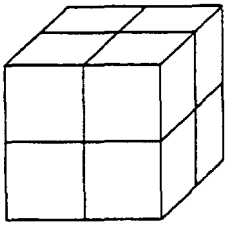
5. The picture shows the flowerpots in which Kevin will plant flower seeds. He needs 3 seeds for each pot. Which of the following number sentences shows how many seeds Kevin will need for all of the pots?

Ⓐ 5 x 4 x 3 = ☐

Ⓑ (5 x 4) + 3 = ☐

Ⓒ (5 + 4) x 3 = ☐

Ⓓ 5 + 4 + 3 = ☐



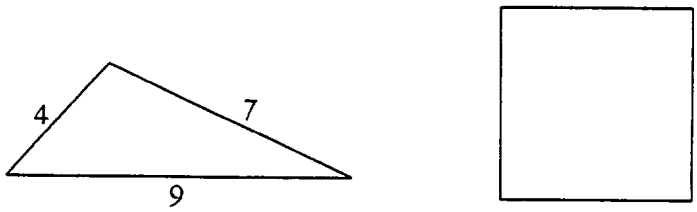
6. In this figure, how many small cubes were put together to form the large cube?

Ⓐ 7

Ⓑ 8

Ⓒ 12

Ⓓ 24



8. If both the square and the triangle above have the same perimeter, what is the length of each side of the square?

Ⓐ 4

Ⓑ 5

Ⓒ 6

Ⓓ 7

9. There are 3 fifth graders and 2 sixth graders on the swim team. Everyone's name is put in a hat and the captain is chosen by picking one name. What are the chances that the captain will be a fifth grader?

Ⓐ l out of 5

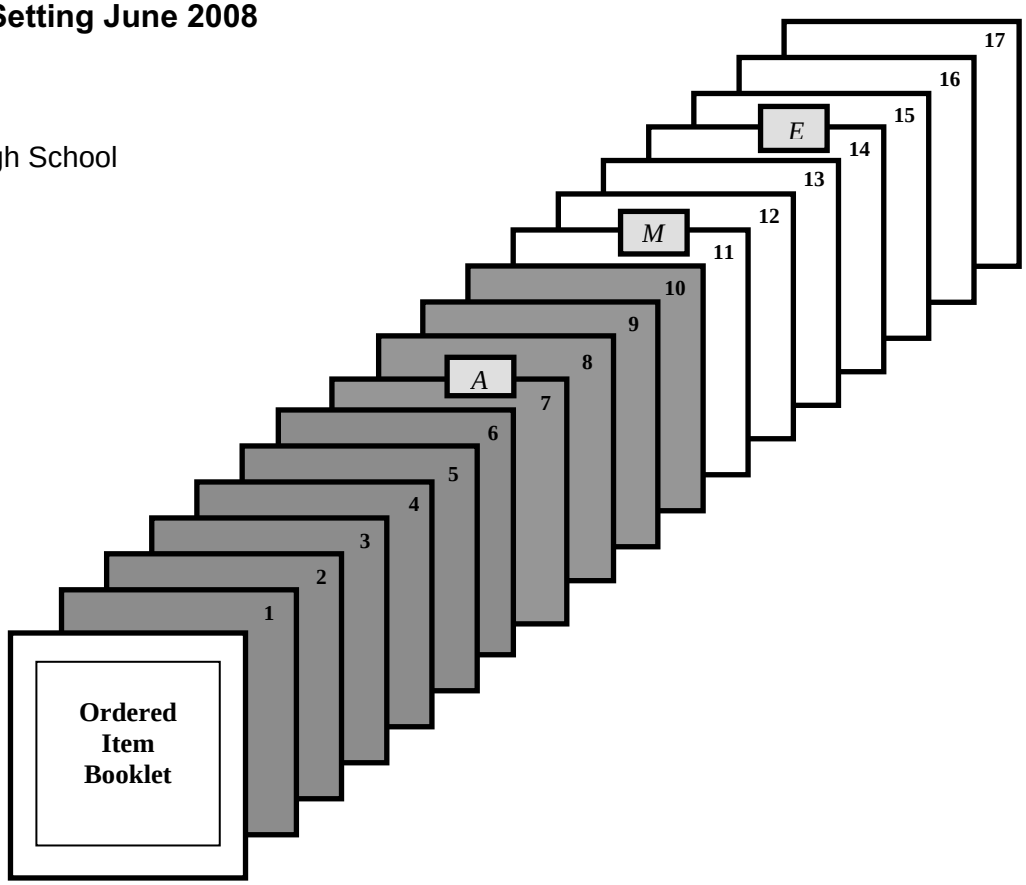
Ⓑ l out of 3

Ⓒ 3 out of 5

Ⓓ 2 out of 3

Arizona Standard Setting June 2008

Grade:
- ○ 4
- ○ 8
- ○ High School



Suppose the bookmarks were placed in this sample ordered item booklet as follows:

	Approaches **Bookmark on Page #**	***Meets*** **Bookmark on Page #**	***Exceeds*** **Bookmark on Page #**
Round 1	**7**	**11**	**14**

1. Which items does a student need to have mastered to just make it into the *Meets* performance level?

   ○ 1 to 6 ○ 1 to 7 ○ 1 to 10 ○ 1 to 11

2. If a student has mastered only items 1 through 5, in which performance level would this student be?

   ○ *Falls Far Below* ○ *Approaches* ○ *Meets* ○ *Exceeds*

3. Suppose a student has mastered items 1 through 6. Which performance level is this student in?

   ○ *Falls Far Below* ○ *Approaches* ○ *Meets* ○ *Exceeds*

4. For students who are classified as *Meets*, with at least what likelihood will they be able to answer item 10?

   ○ *1/3* ○ *1/2* ○ *2/3* ○ *3/4*

5. Will the items BEFORE the *Meets* bookmark be more or less difficult to answer than the items AFTER the bookmark or about the same?

   ○ More difficult to answer ○ About the same ○ Less difficult to answer

SECTION F

Detailed Bookmark Placement Tables and Graphs

AIMS Standard Setting Grade 4 Science
Round 1 Bookmark Placements

Table	Participant	Approaches	Meets	Exceeds
1	1	19	27	59
1	2	18	39	49
1	3	6	37	59
1	4	26	41	66
2	5	10	19	46
2	6	19	44	78
2	7	18	46	71
2	8	19	41	50
3	9	7	58	69
3	10	18	31	47
3	11	10	27	59
3	12	18	57	71

Overall	Median	18	40	59
	Minimum	6	19	46
	Maximum	26	58	78
	SD	5.99	11.77	10.75

AIMS Standard Setting Grade 4 Science
Round 1 Cut Scores

Table	Participant	Approaches	Meets	Exceeds
1	1	464	474	540
1	2	460	486	513
1	3	412	485	540
1	4	473	487	545
2	5	440	464	503
2	6	464	498	562
2	7	460	503	552
2	8	464	487	516
3	9	419	537	550
3	10	460	477	505
3	11	440	474	540
3	12	460	525	552

Overall	Median	460	487	540
	Minimum	412	464	503
	Maximum	473	537	562
	SD	19.38	21.50	20.20

AIMS Standard Setting Grade 4 Science
Round 1 Summary of Bookmark Placements

Statistic	Table	Approaches	Meets	Exceeds
Median	1	18.5	38	59
Median	2	18.5	42.5	60.5
Median	3	14	44	64
Median	Overall	18	40	59
Minimum	1	6	27	49
Minimum	2	10	19	46
Minimum	3	7	27	47
Minimum	Overall	6	19	46
Maximum	1	26	41	66
Maximum	2	19	46	78
Maximum	3	18	58	71
Maximum	Overall	26	58	78
SD	1	8.30	6.22	6.99
SD	2	4.36	12.50	15.65
SD	3	5.62	16.54	11.00
SD	Overall	5.99	11.77	10.75

Overall	Median	18	40	59
	Minimum	6	19	46
	Maximum	26	58	78
	SD	5.99	11.77	10.75

AIMS Standard Setting Grade 4 Science
Round 1 Summary of Cut Scores

Statistic	Table	Approaches	Meets	Exceeds
Median	1	462	486	540
Median	2	462	493	534
Median	3	450	501	545
Median	Overall	460	487	540
Minimum	1	412	474	513
Minimum	2	440	464	503
Minimum	3	419	474	505
Minimum	Overall	412	464	503
Maximum	1	473	487	545
Maximum	2	464	503	562
Maximum	3	460	537	552
Maximum	Overall	473	537	562
SD	1	27.38	6.06	14.53
SD	2	11.49	17.34	28.23
SD	3	19.59	32.44	21.81
SD	Overall	19.38	21.50	20.20

Overall	Median	460	487	540
	Minimum	412	464	503
	Maximum	473	537	562
	SD	19.38	21.50	20.20

AIMS Standard Setting Grade 4 Science
Round 1 Median Bookmark Summary

Table	Approaches	Meets	Exceeds
1	18.5	38	59
2	18.5	42.5	60.5
3	14	44	64
Overall	18	40	59

Impact Data

	Falls Far Below	Approaches	Meets	Exceeds
Overall	21.6	17.0	39.4	22.0

AIMS Standard Setting Grade 4 Science
Round 2 Bookmark Placements

Table	Participant	Approaches	Meets	Exceeds
1	1	19	39	59
1	2	18	39	59
1	3	18	37	61
1	4	22	40	66
2	5	10	19	46
2	6	19	41	61
2	7	17	46	73
2	8	19	46	61
3	9	14	46	62
3	10	18	38	59
3	11	14	33	59
3	12	15	43	61

Overall	Median	18	39.5	61
	Minimum	10	19	46
	Maximum	22	46	73
	SD	3.18	7.44	6.10

AIMS Standard Setting Grade 4 Science
Round 2 Cut Scores

Table	Participant	Approaches	Meets	Exceeds
1	1	464	486	540
1	2	460	486	540
1	3	460	485	542
1	4	467	487	545
2	5	440	464	503
2	6	464	487	542
2	7	459	503	553
2	8	464	503	542
3	9	456	503	542
3	10	460	485	540
3	11	456	481	540
3	12	458	498	542

Overall	Median	460	486	542
	Minimum	440	464	503
	Maximum	467	503	553
	SD	6.92	11.44	12.02

AIMS Standard Setting Grade 4 Science
Round 2 Summary of Bookmark Placements

Statistic	Table	Approaches	Meets	Exceeds
Median	1	18.5	39	60
Median	2	18	43.5	61
Median	3	14.5	40.5	60
Median	Overall	18	39.5	61
Minimum	1	18	37	59
Minimum	2	10	19	46
Minimum	3	14	33	59
Minimum	Overall	10	19	46
Maximum	1	22	40	66
Maximum	2	19	46	73
Maximum	3	18	46	62
Maximum	Overall	22	46	73
SD	1	1.89	1.26	3.30
SD	2	4.27	12.88	11.06
SD	3	1.89	5.72	1.50
SD	Overall	3.18	7.44	6.10

Overall	Median	18	39.5	61
	Minimum	10	19	46
	Maximum	22	46	73
	SD	3.18	7.44	6.10

AIMS Standard Setting Grade 4 Science
Round 2 Summary of Cut Scores

Statistic	Table	Approaches	Meets	Exceeds
Median	1	462	486	541
Median	2	462	495	542
Median	3	457	492	541
Median	Overall	460	486	542
Minimum	1	460	485	540
Minimum	2	440	464	503
Minimum	3	456	481	540
Minimum	Overall	440	464	503
Maximum	1	467	487	545
Maximum	2	464	503	553
Maximum	3	460	503	542
Maximum	Overall	467	503	553
SD	1	3.40	0.82	2.36
SD	2	11.41	18.45	21.95
SD	3	1.91	10.44	1.15
SD	Overall	6.92	11.44	12.02

Overall	Median	460	486	542
	Minimum	440	464	503
	Maximum	467	503	553
	SD	6.92	11.44	12.02

AIMS Standard Setting Grade 4 Science
Round 2 Median Bookmark Summary

Table	Approaches	Meets	Exceeds
1	18.5	39	60
2	18	43.5	61
3	14.5	40.5	60
Overall	18	39.5	61

Impact Data

	Falls Far Below	Approaches	Meets	Exceeds
Overall	21.6	17.0	41.2	20.2

AIMS Standard Setting Grade 4 Science
Round 3 Bookmark Placements

Table	Participant	Approaches	Meets	Exceeds
1	1	19	39	66
1	2	18	39	61
1	3	18	39	69
1	4	28	48	66
2	5	16	43	64
2	6	19	45	79
2	7	22	44	69
2	8	19	44	74
3	9	14	44	62
3	10	18	42	61
3	11	15	37	63
3	12	18	54	71

Overall	Median	18	43.5	66
	Minimum	14	37	61
	Maximum	28	54	79
	SD	3.60	4.65	5.57

AIMS Standard Setting Grade 4 Science
Round 3 Cut Scores

Table	Participant	Approaches	Meets	Exceeds
1	1	464	486	545
1	2	460	486	542
1	3	460	486	550
1	4	474	508	545
2	5	458	498	544
2	6	464	500	562
2	7	467	498	550
2	8	464	498	554
3	9	456	498	542
3	10	460	496	542
3	11	458	485	542
3	12	460	521	552

Overall	Median	460	498	545
	Minimum	456	485	542
	Maximum	474	521	562
	SD	5.00	10.56	6.19

AIMS Standard Setting Grade 4 Science
Round 3 Summary of Bookmark Placements

Statistic	Table	Approaches	Meets	Exceeds
Median	1	18.5	39	66
Median	2	19	44	71.5
Median	3	16.5	43	62.5
Median	Overall	18	43.5	66
Minimum	1	18	39	61
Minimum	2	16	43	64
Minimum	3	14	37	61
Minimum	Overall	14	37	61
Maximum	1	28	48	69
Maximum	2	22	45	79
Maximum	3	18	54	71
Maximum	Overall	28	54	79
SD	1	4.86	4.50	3.32
SD	2	2.45	0.82	6.45
SD	3	2.06	7.14	4.57
SD	Overall	3.60	4.65	5.57

Overall	Median	18	43.5	66
	Minimum	14	37	61
	Maximum	28	54	79
	SD	3.60	4.65	5.57

AIMS Standard Setting Grade 4 Science
Round 3 Summary of Cut Scores

Statistic	Table	Approaches	Meets	Exceeds
Median	1	462	486	545
Median	2	464	498	552
Median	3	459	497	542
Median	Overall	460	498	545
Minimum	1	460	486	542
Minimum	2	458	498	544
Minimum	3	456	485	542
Minimum	Overall	456	485	542
Maximum	1	474	508	550
Maximum	2	467	500	562
Maximum	3	460	521	552
Maximum	Overall	474	521	562
SD	1	6.61	11.00	3.32
SD	2	3.77	1.00	7.55
SD	3	1.91	15.12	5.00
SD	Overall	5.00	10.56	6.19

Overall	Median	460	498	545
	Minimum	456	485	542
	Maximum	474	521	562
	SD	5.00	10.56	6.19

AIMS Standard Setting Grade 4 Science
Round 3 Median Bookmark Summary

Table	Approaches	Meets	Exceeds
1	18.5	39	66
2	19	44	71.5
3	16.5	43	62.5
Overall	18	43.5	66

Impact Data

	Falls Far Below	Approaches	Meets	Exceeds
Overall	21.6	25.4	34.7	18.3

AIMS Standard Setting Grade 8 Science
Round 1 Bookmark Placements

Table	Participant	Approaches	Meets	Exceeds
1	2	21	40	74
1	3	15	30	62
1	4	21	35	74
1	5	24	35	62
2	6	13	30	63
2	7	31	40	54
2	8	13	30	64
2	9	13	31	62
3	10	18	27	62
3	11	22	42	71
3	12	12	32	70
3	13	21	41	71

Overall	Median	19.5	33.5	63.5
	Minimum	12	27	54
	Maximum	31	42	74
	SD	5.74	5.18	6.14

AIMS Standard Setting Grade 8 Science
Round 1 Cut Scores

Table	Participant	Approaches	Meets	Exceeds
1	2	473	501	540
1	3	469	488	526
1	4	473	499	540
1	5	483	499	526
2	6	468	488	527
2	7	490	501	515
2	8	468	488	527
2	9	468	490	526
3	10	471	486	526
3	11	482	502	532
3	12	465	492	531
3	13	473	501	532

Overall	Median	471	495	527
	Minimum	465	486	515
	Maximum	490	502	540
	SD	7.67	6.54	6.87

AIMS Standard Setting Grade 8 Science
Round 1 Summary of Bookmark Placements

Statistic	Table	Approaches	Meets	Exceeds
Median	1	21	35	68
Median	2	13	30.5	62.5
Median	3	19.5	36.5	70.5
Median	Overall	19.5	33.5	63.5
Minimum	1	15	30	62
Minimum	2	13	30	54
Minimum	3	12	27	62
Minimum	Overall	12	27	54
Maximum	1	24	40	74
Maximum	2	31	40	64
Maximum	3	22	42	71
Maximum	Overall	31	42	74
SD	1	3.77	4.08	6.93
SD	2	9.00	4.86	4.57
SD	3	4.50	7.23	4.36
SD	Overall	5.74	5.18	6.14

Overall	Median	19.5	33.5	63.5
	Minimum	12	27	54
	Maximum	31	42	74
	SD	5.74	5.18	6.14

AIMS Standard Setting Grade 8 Science
Round 1 Summary of Cut Scores

Statistic	Table	Approaches	Meets	Exceeds
Median	1	473	499	533
Median	2	468	489	527
Median	3	472	497	532
Median	Overall	471	495	527
Minimum	1	469	488	526
Minimum	2	468	488	515
Minimum	3	465	486	526
Minimum	Overall	465	486	515
Maximum	1	483	501	540
Maximum	2	490	501	527
Maximum	3	482	502	532
Maximum	Overall	490	502	540
SD	1	5.97	5.91	8.08
SD	2	11.00	6.24	5.85
SD	3	7.04	7.63	2.87
SD	Overall	7.67	6.54	6.87

Overall	Median	471	495	527
	Minimum	465	486	515
	Maximum	490	502	540
	SD	7.67	6.54	6.87

AIMS Standard Setting Grade 8 Science
Round 1 Median Bookmark Summary

Table	Approaches	Meets	Exceeds
1	21	35	68
2	13	30.5	62.5
3	19.5	36.5	70.5
Overall	19.5	33.5	63.5

Impact Data

	Falls Far Below	Approaches	Meets	Exceeds
Overall	30.3	17.0	21.5	31.2

AIMS Standard Setting Grade 8 Science
Round 2 Bookmark Placements

Table	Participant	Approaches	Meets	Exceeds
1	2	21	35	74
1	3	20	32	66
1	4	21	35	70
1	5	22	35	66
2	6	13	30	70
2	7	21	35	67
2	8	13	32	68
2	9	19	35	66
3	10	22	42	80
3	11	22	39	70
3	12	16	32	87
3	13	20	35	70

Overall	Median	20.5	35	70
	Minimum	13	30	66
	Maximum	22	42	87
	SD	3.33	3.25	6.39

AIMS Standard Setting Grade 8 Science
Round 2 Cut Scores

Table	Participant	Approaches	Meets	Exceeds
1	2	473	499	540
1	3	472	492	528
1	4	473	499	531
1	5	482	499	528
2	6	468	488	531
2	7	473	499	529
2	8	468	492	530
2	9	471	499	528
3	10	482	502	552
3	11	482	501	531
3	12	470	492	564
3	13	472	499	531

Overall	Median	472	499	531
	Minimum	468	488	528
	Maximum	482	502	564
	SD	5.10	4.60	11.31

AIMS Standard Setting Grade 8 Science
Round 2 Summary of Bookmark Placements

Statistic	Table	Approaches	Meets	Exceeds
Median	1	21	35	68
Median	2	16	33.5	67.5
Median	3	21	37	75
Median	Overall	20.5	35	70
Minimum	1	20	32	66
Minimum	2	13	30	66
Minimum	3	16	32	70
Minimum	Overall	13	30	66
Maximum	1	22	35	74
Maximum	2	21	35	70
Maximum	3	22	42	87
Maximum	Overall	22	42	87
SD	1	0.82	1.50	3.83
SD	2	4.12	2.45	1.71
SD	3	2.83	4.40	8.30
SD	Overall	3.33	3.25	6.39

Overall	Median	20.5	35	70
	Minimum	13	30	66
	Maximum	22	42	87
	SD	3.33	3.25	6.39

AIMS Standard Setting Grade 8 Science
Round 2 Summary of Cut Scores

Statistic	Table	Approaches	Meets	Exceeds
Median	1	473	499	530
Median	2	470	496	530
Median	3	477	500	542
Median	Overall	472	499	531
Minimum	1	472	492	528
Minimum	2	468	488	528
Minimum	3	470	492	531
Minimum	Overall	468	488	528
Maximum	1	482	499	540
Maximum	2	473	499	531
Maximum	3	482	502	564
Maximum	Overall	482	502	564
SD	1	4.69	3.50	5.68
SD	2	2.45	5.45	1.29
SD	3	6.40	4.51	16.34
SD	Overall	5.10	4.60	11.31

Overall	Median	472	499	531
	Minimum	468	488	528
	Maximum	482	502	564
	SD	5.10	4.60	11.31

AIMS Standard Setting Grade 8 Science
Round 2 Median Bookmark Summary

Table	Approaches	Meets	Exceeds
1	21	35	68
2	16	33.5	67.5
3	21	37	75
Overall	20.5	35	70

Impact Data

	Falls Far Below	Approaches	Meets	Exceeds
Overall	30.3	20.0	21.7	28.0

AIMS Standard Setting Grade 8 Science
Round 3 Bookmark Placements

Table	Participant	Approaches	Meets	Exceeds
1	2	21	35	74
1	3	20	35	70
1	4	20	35	70
1	5	22	35	66
2	6	21	35	70
2	7	18	32	67
2	8	18	35	70
2	9	19	35	68
3	10	22	35	80
3	11	25	42	74
3	12	16	32	81
3	13	19	37	73

Overall	Median	20	35	70
	Minimum	16	32	66
	Maximum	25	42	81
	SD	2.35	2.53	4.74

AIMS Standard Setting Grade 8 Science
Round 3 Cut Scores

Table	Participant	Approaches	Meets	Exceeds
1	2	473	499	540
1	3	472	499	531
1	4	472	499	531
1	5	482	499	528
2	6	473	499	531
2	7	471	492	529
2	8	471	499	531
2	9	471	499	530
3	10	482	499	552
3	11	485	502	540
3	12	470	492	555
3	13	471	500	537

Overall	Median	472	499	531
	Minimum	470	492	528
	Maximum	485	502	555
	SD	5.18	3.05	8.92

AIMS Standard Setting Grade 8 Science
Round 3 Summary of Bookmark Placements

Statistic	Table	Approaches	Meets	Exceeds
Median	1	20.5	35	70
Median	2	18.5	35	69
Median	3	20.5	36	77
Median	Overall	20	35	70
Minimum	1	20	35	66
Minimum	2	18	32	67
Minimum	3	16	32	73
Minimum	Overall	16	32	66
Maximum	1	22	35	74
Maximum	2	21	35	70
Maximum	3	25	42	81
Maximum	Overall	25	42	81
SD	1	0.96	0.00	3.27
SD	2	1.41	1.50	1.50
SD	3	3.87	4.20	4.08
SD	Overall	2.35	2.53	4.74

Overall	Median	20	35	70
	Minimum	16	32	66
	Maximum	25	42	81
	SD	2.35	2.53	4.74

AIMS Standard Setting Grade 8 Science
Round 3 Summary of Cut Scores

Statistic	Table	Approaches	Meets	Exceeds
Median	1	473	499	531
Median	2	471	499	531
Median	3	477	500	546
Median	Overall	472	499	531
Minimum	1	472	499	528
Minimum	2	471	492	529
Minimum	3	470	492	537
Minimum	Overall	470	492	528
Maximum	1	482	499	540
Maximum	2	473	499	531
Maximum	3	485	502	555
Maximum	Overall	485	502	555
SD	1	4.86	0.00	5.20
SD	2	1.00	3.50	0.96
SD	3	7.62	4.35	8.83
SD	Overall	5.18	3.05	8.92

Overall	Median	472	499	531
	Minimum	470	492	528
	Maximum	485	502	555
	SD	5.18	3.05	8.92

AIMS Standard Setting Grade 8 Science
Round 3 Median Bookmark Summary

Table	Approaches	Meets	Exceeds
1	20.5	35	70
2	18.5	35	69
3	20.5	36	77
Overall	20	35	70

Impact Data

	Falls Far Below	Approaches	Meets	Exceeds
Overall	30.3	20.0	21.7	28.0

AIMS Standard Setting Grade 10 Science
Round 1 Bookmark Placements

Table	Participant	Approaches	Meets	Exceeds
1	3	41	68	96
1	4	40	56	82
1	5	13	39	85
1	6	24	36	67
2	7	41	65	80
2	8	17	35	72
2	9	34	64	88
2	10	35	55	78
3	11	39	52	102
3	12	38	63	80
3	13	33	64	79

Overall	Median	35	56	80
	Minimum	13	35	67
	Maximum	41	68	102
	SD	9.87	12.31	9.97

AIMS Standard Setting Grade 10 Science
Round 1 Cut Scores

Table	Participant	Approaches	Meets	Exceeds
1	3	510	541	570
1	4	510	524	557
1	5	467	509	560
1	6	488	505	540
2	7	510	540	556
2	8	472	505	543
2	9	501	536	560
2	10	505	523	552
3	11	509	520	577
3	12	508	536	556
3	13	497	536	553

Overall	Median	505	524	556
	Minimum	467	505	540
	Maximum	510	541	577
	SD	15.66	13.84	10.71

AIMS Standard Setting Grade 10 Science
Round 1 Summary of Bookmark Placements

Statistic	Table	Approaches	Meets	Exceeds
Median	1	32	47.5	83.5
Median	2	34.5	59.5	79
Median	3	38	63	80
Median	Overall	35	56	80
Minimum	1	13	36	67
Minimum	2	17	35	72
Minimum	3	33	52	79
Minimum	Overall	13	35	67
Maximum	1	41	68	96
Maximum	2	41	65	88
Maximum	3	39	64	102
Maximum	Overall	41	68	102
SD	1	13.48	15.02	11.96
SD	2	10.31	13.91	6.61
SD	3	3.21	6.66	13.00
SD	Overall	9.87	12.31	9.97

Overall	Median	35	56	80
	Minimum	13	35	67
	Maximum	41	68	102
	SD	9.87	12.31	9.97

AIMS Standard Setting Grade 10 Science
Round 1 Summary of Cut Scores

Statistic	Table	Approaches	Meets	Exceeds
Median	1	499	517	559
Median	2	503	530	554
Median	3	508	536	556
Median	Overall	505	524	556
Minimum	1	467	505	540
Minimum	2	472	505	543
Minimum	3	497	520	553
Minimum	Overall	467	505	540
Maximum	1	510	541	570
Maximum	2	510	540	560
Maximum	3	509	536	577
Maximum	Overall	510	541	577
SD	1	20.63	16.36	12.47
SD	2	17.07	15.77	7.27
SD	3	6.66	9.24	13.08
SD	Overall	15.66	13.84	10.71

Overall	Median	505	524	556
	Minimum	467	505	540
	Maximum	510	541	577
	SD	15.66	13.84	10.71

AIMS Standard Setting Grade 10 Science
Round 1 Median Bookmark Summary

Table	Approaches	Meets	Exceeds
1	32	47.5	83.5
2	34.5	59.5	79
3	38	63	80
Overall	35	56	80

Impact Data

	Falls Far Below	Approaches	Meets	Exceeds
Overall	54.3	12.9	19.0	13.8

AIMS Standard Setting Grade 10 Science
Round 2 Bookmark Placements

Table	Participant	Approaches	Meets	Exceeds
1	3	24	56	87
1	4	41	59	83
1	5	36	53	90
1	6	23	55	83
2	7	38	59	80
2	8	34	48	80
2	9	34	55	80
2	10	36	56	78
3	11	33	52	83
3	12	33	57	83
3	13	33	52	79

Overall	Median	34	55	83
	Minimum	23	48	78
	Maximum	41	59	90
	SD	5.38	3.29	3.59

AIMS Standard Setting Grade 10 Science
Round 2 Cut Scores

Table	Participant	Approaches	Meets	Exceeds
1	3	488	524	560
1	4	510	530	559
1	5	505	522	567
1	6	488	523	559
2	7	508	530	556
2	8	501	514	556
2	9	501	523	556
2	10	505	524	552
3	11	497	520	559
3	12	497	528	559
3	13	497	520	553

Overall	Median	501	523	559
	Minimum	488	514	552
	Maximum	510	530	567
	SD	7.30	4.79	4.09

AIMS Standard Setting Grade 10 Science
Round 2 Summary of Bookmark Placements

Statistic	Table	Approaches	Meets	Exceeds
Median	1	30	55.5	85
Median	2	35	55.5	80
Median	3	33	52	83
Median	Overall	34	55	83
Minimum	1	23	53	83
Minimum	2	34	48	78
Minimum	3	33	52	79
Minimum	Overall	23	48	78
Maximum	1	41	59	90
Maximum	2	38	59	80
Maximum	3	33	57	83
Maximum	Overall	41	59	90
SD	1	8.91	2.50	3.40
SD	2	1.91	4.65	1.00
SD	3	0.00	2.89	2.31
SD	Overall	5.38	3.29	3.59

Overall	Median	34	55	83
	Minimum	23	48	78
	Maximum	41	59	90
	SD	5.38	3.29	3.59

AIMS Standard Setting Grade 10 Science
Round 2 Summary of Cut Scores

Statistic	Table	Approaches	Meets	Exceeds
Median	1	497	524	560
Median	2	503	524	556
Median	3	497	520	559
Median	Overall	501	523	559
Minimum	1	488	522	559
Minimum	2	501	514	552
Minimum	3	497	520	553
Minimum	Overall	488	514	552
Maximum	1	510	530	567
Maximum	2	508	530	556
Maximum	3	497	528	559
Maximum	Overall	510	530	567
SD	1	11.44	3.59	3.86
SD	2	3.40	6.60	2.00
SD	3	0.00	4.62	3.46
SD	Overall	7.30	4.79	4.09

Overall	Median	501	523	559
	Minimum	488	514	552
	Maximum	510	530	567
	SD	7.30	4.79	4.09

AIMS Standard Setting Grade 10 Science
Round 2 Median Bookmark Summary

Table	Approaches	Meets	Exceeds
1	30	55.5	85
2	35	55.5	80
3	33	52	83
Overall	34	55	83

Impact Data

	Falls Far Below	Approaches	Meets	Exceeds
Overall	51.4	15.8	20.1	12.7

AIMS Standard Setting Grade 10 Science
Round 3 Bookmark Placements

Table	Participant	Approaches	Meets	Exceeds
1	3	33	56	88
1	4	40	56	85
1	5	33	53	86
1	6	23	50	83
2	7	32	56	84
2	8	34	48	80
2	9	34	55	84
2	10	34	54	76
3	11	33	53	85
3	12	30	52	83
3	13	33	48	79

Overall	Median	33	53	84
	Minimum	23	48	76
	Maximum	40	56	88
	SD	4.01	3.03	3.44

AIMS Standard Setting Grade 10 Science
Round 3 Cut Scores

Table	Participant	Approaches	Meets	Exceeds
1	3	497	524	560
1	4	510	524	560
1	5	497	522	560
1	6	488	517	559
2	7	497	524	559
2	8	501	514	556
2	9	501	523	559
2	10	501	522	550
3	11	497	522	560
3	12	495	520	559
3	13	497	514	553

Overall	Median	497	522	559
	Minimum	488	514	550
	Maximum	510	524	560
	SD	5.17	3.62	3.30

AIMS Standard Setting Grade 10 Science
Round 3 Summary of Bookmark Placements

Statistic	Table	Approaches	Meets	Exceeds
Median	1	33	54.5	85.5
Median	2	34	54.5	82
Median	3	33	52	83
Median	Overall	33	53	84
Minimum	1	23	50	83
Minimum	2	32	48	76
Minimum	3	30	48	79
Minimum	Overall	23	48	76
Maximum	1	40	56	88
Maximum	2	34	56	84
Maximum	3	33	53	85
Maximum	Overall	40	56	88
SD	1	6.99	2.87	2.08
SD	2	1.00	3.59	3.83
SD	3	1.73	2.65	3.06
SD	Overall	4.01	3.03	3.44

Overall	Median	33	53	84
	Minimum	23	48	76
	Maximum	40	56	88
	SD	4.01	3.03	3.44

AIMS Standard Setting Grade 10 Science
Round 3 Summary of Cut Scores

Statistic	Table	Approaches	Meets	Exceeds
Median	1	497	523	560
Median	2	501	523	558
Median	3	497	520	559
Median	Overall	497	522	559
Minimum	1	488	517	559
Minimum	2	497	514	550
Minimum	3	495	514	553
Minimum	Overall	488	514	550
Maximum	1	510	524	560
Maximum	2	501	524	559
Maximum	3	497	522	560
Maximum	Overall	510	524	560
SD	1	9.06	3.30	0.50
SD	2	2.00	4.57	4.24
SD	3	1.15	4.16	3.79
SD	Overall	5.17	3.62	3.30

Overall	Median	497	522	559
	Minimum	488	514	550
	Maximum	510	524	560
	SD	5.17	3.62	3.30

AIMS Standard Setting Grade 10 Science
Round 3 Median Bookmark Summary

Table	Approaches	Meets	Exceeds
1	33	54.5	85.5
2	34	54.5	82
3	33	52	83
Overall	33	53	84

Impact Data

	Falls Far Below	Approaches	Meets	Exceeds
Overall	48.6	18.6	20.1	12.7

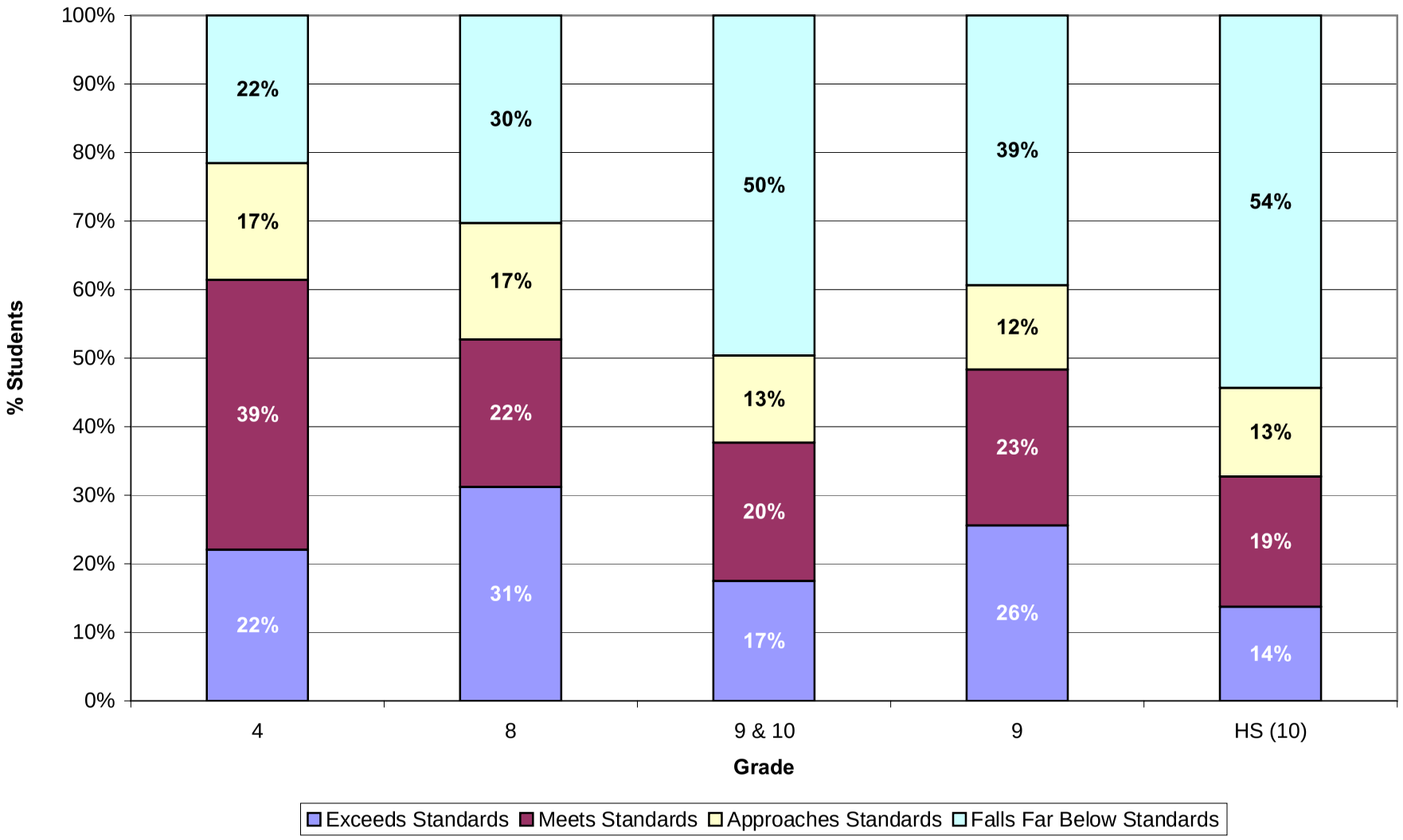
Science

Based on Participants' Round 1 Bookmark Recommendations
Standard setting workshop held June 9-11, 2008

	4	8	9 & 10	9	HS (10)	
Falls Far Below Standards	22%	30%	50%	39%	54%	**Impact**
Approaches Standards	17%	17%	13%	12%	13%	
Meets Standards	39%	22%	20%	23%	19%	
Exceeds Standards	22%	31%	17%	26%	14%	
Meets Standards & Above	**61%**	**53%**	**38%**	**48%**	**33%**	

Approaches Standards	460	471	505	505	505	**Cut Scores**
Meets Standards	487	495	524	524	524	
Exceeds Standards	540	527	556	556	556	

Arizona's Instrument to Measure Standards
Science Round 1 Results: Percent of Students by Performance Level



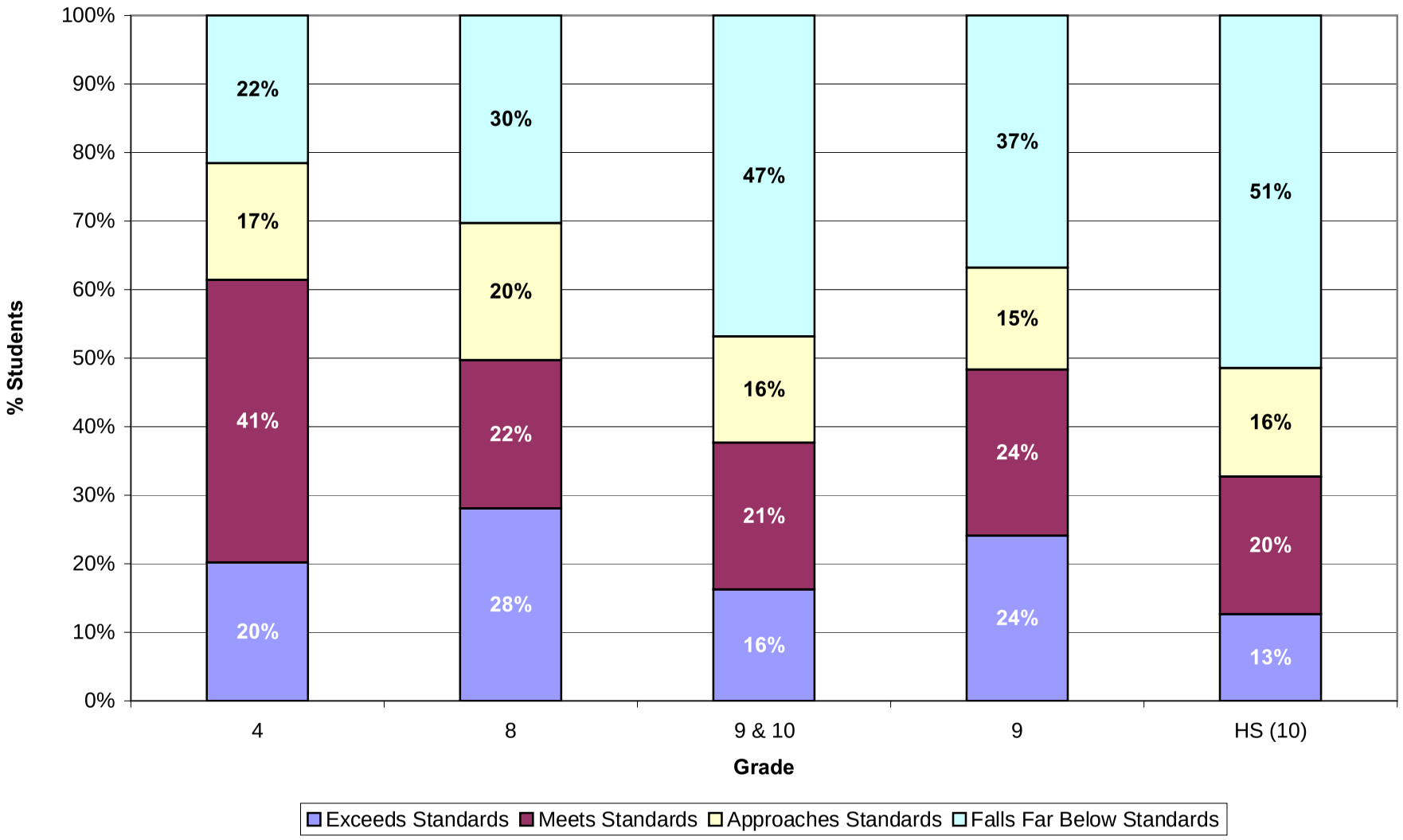
Science

Based on Participants' Round 2 Bookmark Recommendations
Standard setting workshop held June 9-11, 2008

	4	8	9 & 10	9	HS (10)	
Falls Far Below Standards	22%	30%	47%	37%	51%	**Impact**
Approaches Standards	17%	20%	16%	15%	16%	
Meets Standards	41%	22%	21%	24%	20%	
Exceeds Standards	20%	28%	16%	24%	13%	
Meets Standards & Above	**61%**	**50%**	**38%**	**48%**	**33%**	

Approaches Standards	460	472	501	501	501	**Cut Scores**
Meets Standards	486	499	523	523	523	
Exceeds Standards	542	531	559	559	559	

Arizona's Instrument to Measure Standards
Science Round 2 Results: Percent of Students by Performance Level



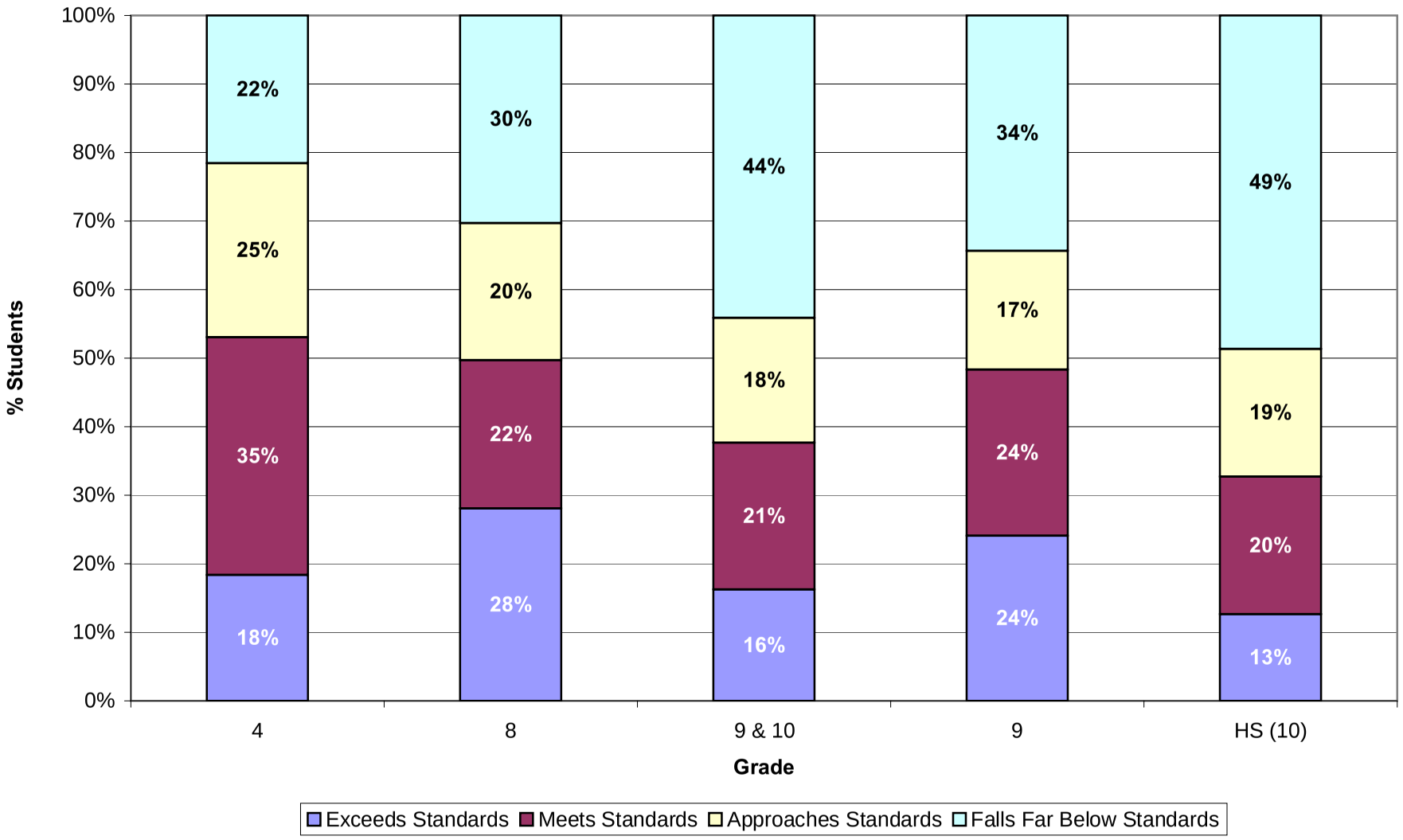
Science

Based on Participants' Final Round Bookmark Recommendations

Standard setting workshop held June 9-11, 2008

	4	8	9 & 10	9	HS (10)	
Falls Far Below Standards	22%	30%	44%	34%	49%	**Impact**
Approaches Standards	25%	20%	18%	17%	19%	
Meets Standards	35%	22%	21%	24%	20%	
Exceeds Standards	18%	28%	16%	24%	13%	
Meets Standards & Above	**53%**	**50%**	**38%**	**48%**	**33%**	
Approaches Standards	460	472	497	497	497	**Cut Scores**
Meets Standards	498	499	522	522	522	
Exceeds Standards	545	531	559	559	559	

Arizona's Instrument to Measure Standards
Science Final Round Results: Percent of Students by Performance Level



Science

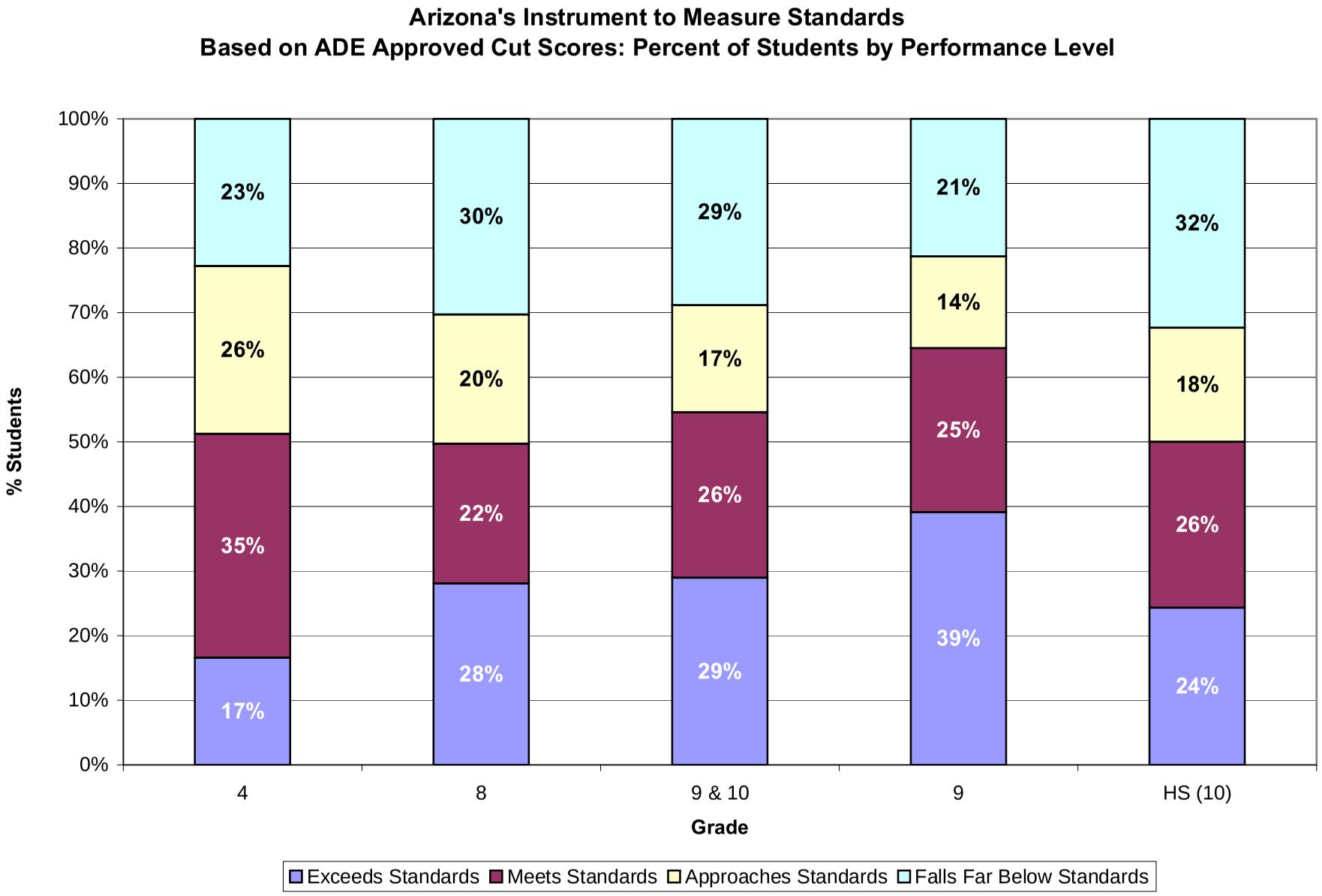
Based on ADE Approved Changes on June 30, 2008

Standard setting workshop held June 9-11, 2008

	4	8	9 & 10	9	HS (10)	
Falls Far Below Standards	23%	30%	29%	21%	32%	**Impact**
Approaches Standards	26%	20%	17%	14%	18%	
Meets Standards	35%	22%	26%	25%	26%	
Exceeds Standards	17%	28%	29%	39%	24%	
Meets Standards & Above	**51%**	**50%**	**55%**	**64%**	**50%**	

Approaches Standards	462	473	475	475	475	**Cut Scores**
Meets Standards	500	500	500	500	500	
Exceeds Standards	547	532	537	537	537	

F 53


Arizona's Instrument to Measure Standards
Based on ADE Approved Cut Scores: Percent of Students by Performance Level
% Students
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
23%
26%
35%
17%
30%
20%
22%
28%
29%
17%
26%
29%
21%
14%
25%
39%
32%
18%
26%
24%
4
8
9 & 10
9
HS (10)
Grade
Exceeds Standards
Meets Standards
Approaches Standards
Falls Far Below Standards

SECTION G

Participant Judgments
Plus/Minus 1, 2, and 3 Standard Errors

AIMS Standard Setting Grade 4 Science

Recommended Cut Points* Plus/Minus Selected Standard Errors (SEs) of the Cut Score

Performance Level	Falls Far Below	Approaches	Meets	Exceeds	
SE (cut score)		3.24	4.52	4.90	
Recommended Cut Point* + 3 SE		469	512	560	+ 3 SE
Percent of Students in Each Level	26.6	31.5	30.2	11.7	
Recommended Cut Point* + 2 SE		466	507	555	+ 2 SE
Percent of Students in Each Level	25.3	29.1	32.4	13.2	
Recommended Cut Point* + 1 SE		463	503	550	+ 1 SE
Percent of Students in Each Level	24.0	26.6	32.8	16.6	
Recommended Cut Point*		460	498	545	Recommended Cut Points*
Percent of Students in Each Level	21.6	25.4	34.7	18.3	
Recommended Cut Point* -1 SE		457	494	541	-1 SE
Percent of Students in Each Level	20.4	23.2	36.3	20.1	
Recommended Cut Point* -2 SE		453	489	536	-2 SE
Percent of Students in Each Level	18.1	22.2	35.6	24.1	
Recommended Cut Point* -3 SE		450	485	531	-3 SE
Percent of Students in Each Level	17.0	20.0	35.0	28.0	

* Participants' Large Group Medians

AIMS Standard Setting Grade 4 Science

Recommended Cut Points* Plus/Minus Selected Standard Errors (SEs) of Measurement

Performance Level	Falls Far Below	Approaches	Meets	Exceeds	
Standard Error (SE) measurement		15.00	15.00	18.00	
Recommended Cut Point* + 3 SE		505	543	599	+ 3 SE
Percent of Students in Each Level	52.4	27.4	18.1	2.1	
Recommended Cut Point* + 2 SE		490	528	581	+ 2 SE
Percent of Students in Each Level	41.9	28.2	24.6	5.3	
Recommended Cut Point* + 1 SE		475	513	563	+ 1 SE
Percent of Students in Each Level	30.8	27.3	31.7	10.2	
Recommended Cut Point*		460	498	545	Recommended Cut Points*
Percent of Students in Each Level	21.6	25.4	34.7	18.3	
Recommended Cut Point* -1 SE		445	483	527	-1 SE
Percent of Students in Each Level	14.7	22.2	33.1	30.0	
Recommended Cut Point* -2 SE		430	468	509	-2 SE
Percent of Students in Each Level	8.9	17.7	29.6	43.8	
Recommended Cut Point* -3 SE		415	453	491	-3 SE
Percent of Students in Each Level	4.5	13.6	23.8	58.1	

* Participants' Large Group Medians

AIMS Standard Setting Grade 4 Science

Recommended Cut Points* Plus/Minus Selected Standard Errors (SEs) of Measurement and the Cut Score

Performance Level	Falls Far Below	Approaches	Meets	Exceeds	
Standard Error (SE) measurement + cutscore		15.34	15.66	18.65	
Recommended Cut Point* + 3 SE		506	545	601	+ 3 SE
Percent of Students in Each Level	52.4	29.2	16.3	2.1	
Recommended Cut Point* + 2 SE		490	529	583	+ 2 SE
Percent of Students in Each Level	41.9	28.2	24.6	5.3	
Recommended Cut Point* + 1 SE		475	514	564	+ 1 SE
Percent of Students in Each Level	30.8	29.2	29.7	10.3	
Recommended Cut Point*		460	498	545	Recommended Cut Points*
Percent of Students in Each Level	21.6	25.4	34.7	18.3	
Recommended Cut Point* -1 SE		444	482	527	-1 SE
Percent of Students in Each Level	13.7	21.7	34.7	29.9	
Recommended Cut Point* -2 SE		429	467	508	-2 SE
Percent of Students in Each Level	8.1	17.3	29.1	45.5	
Recommended Cut Point* -3 SE		414	451	490	-3 SE
Percent of Students in Each Level	4.0	14.1	23.8	58.1	

* Participants' Large Group Medians

AIMS Standard Setting Grade 8 Science

Recommended Cut Points* Plus/Minus Selected Standard Errors (SEs) of the Cut Score

Performance Level	Falls Far Below	Approaches	Meets	Exceeds	
SE (cut score)		2.74	2.07	6.25	
Recommended Cut Point* + 3 SE		481	506	550	+ 3 SE
Percent of Students in Each Level	35.7	20.6	27.9	15.8	
Recommended Cut Point* + 2 SE		478	504	544	+ 2 SE
Percent of Students in Each Level	35.7	17.6	26.4	20.3	
Recommended Cut Point* + 1 SE		475	502	538	+ 1 SE
Percent of Students in Each Level	33.0	20.4	23.3	23.3	
Recommended Cut Point*		472	499	531	Recommended Cut Points*
Percent of Students in Each Level	30.3	20.0	21.7	28.0	
Recommended Cut Point* -1 SE		470	497	525	-1 SE
Percent of Students in Each Level	27.7	19.6	21.5	31.2	
Recommended Cut Point* -2 SE		467	495	519	-2 SE
Percent of Students in Each Level	27.7	19.6	18.4	34.3	
Recommended Cut Point* -3 SE		464	493	513	-3 SE
Percent of Students in Each Level	25.2	19.3	15.0	40.5	

* Participants' Large Group Medians

AIMS Standard Setting Grade 8 Science

Recommended Cut Points* Plus/Minus Selected Standard Errors (SEs) of Measurement

Performance Level	Falls Far Below	Approaches	Meets	Exceeds	
Standard Error (SE) measurement		14.00	14.00	15.00	
Recommended Cut Point* + 3 SE		514	541	576	+ 3 SE
Percent of Students in Each Level	61.0	17.4	15.1	6.5	
Recommended Cut Point* + 2 SE		500	527	561	+ 2 SE
Percent of Students in Each Level	50.3	18.5	19.6	11.6	
Recommended Cut Point* + 1 SE		486	513	546	+ 1 SE
Percent of Students in Each Level	41.5	18.0	21.9	18.6	
Recommended Cut Point*		472	499	531	Recommended Cut Points*
Percent of Students in Each Level	30.3	20.0	21.7	28.0	
Recommended Cut Point* -1 SE		458	485	516	-1 SE
Percent of Students in Each Level	20.4	18.1	24.1	37.4	
Recommended Cut Point* -2 SE		444	471	501	-2 SE
Percent of Students in Each Level	14.1	16.2	20.0	49.7	
Recommended Cut Point* -3 SE		430	457	486	-3 SE
Percent of Students in Each Level	8.0	12.4	21.1	58.5	

* Participants' Large Group Medians

AIMS Standard Setting Grade 8 Science

Recommended Cut Points* Plus/Minus Selected Standard Errors (SEs) of Measurement and the Cut Score

Performance Level	Falls Far Below	Approaches	Meets	Exceeds	
Standard Error (SE) measurement + cutscore		14.26	14.15	16.25	
Recommended Cut Point* + 3 SE		515	542	580	+ 3 SE
Percent of Students in Each Level	62.5	15.8	16.2	5.5	
Recommended Cut Point* + 2 SE		501	528	564	+ 2 SE
Percent of Students in Each Level	50.3	20.0	19.5	10.2	
Recommended Cut Point* + 1 SE		487	514	548	+ 1 SE
Percent of Students in Each Level	41.5	19.5	20.4	18.6	
Recommended Cut Point*		472	499	531	Recommended Cut Points*
Percent of Students in Each Level	30.3	20.0	21.7	28.0	
Recommended Cut Point* -1 SE		458	485	515	-1 SE
Percent of Students in Each Level	20.4	18.1	24.1	37.4	
Recommended Cut Point* -2 SE		444	471	499	-2 SE
Percent of Students in Each Level	14.1	16.2	20.0	49.7	
Recommended Cut Point* -3 SE		430	457	483	-3 SE
Percent of Students in Each Level	8.0	12.4	18.1	61.5	

* Participants' Large Group Medians

AIMS Standard Setting Grade 10 Science

Recommended Cut Points* Plus/Minus Selected Standard Errors (SEs) of the Cut Score

Performance Level	Falls Far Below	Approaches	Meets	Exceeds	
SE (cut score)		3.55	1.96	2.50	
Recommended Cut Point* + 3 SE		508	528	567	+ 3 SE
Percent of Students in Each Level	55.8	14.5	20.7	9.0	
Recommended Cut Point* + 2 SE		505	526	564	+ 2 SE
Percent of Students in Each Level	54.3	15.9	19.0	10.8	
Recommended Cut Point* + 1 SE		501	524	562	+ 1 SE
Percent of Students in Each Level	51.4	15.8	22.0	10.8	
Recommended Cut Point*		497	522	559	Recommended Cut Points*
Percent of Students in Each Level	48.6	18.6	20.1	12.7	
Recommended Cut Point* -1 SE		494	520	557	-1 SE
Percent of Students in Each Level	45.8	18.8	22.8	12.6	
Recommended Cut Point* -2 SE		490	518	554	-2 SE
Percent of Students in Each Level	43.1	21.5	20.7	14.7	
Recommended Cut Point* -3 SE		487	516	552	-3 SE
Percent of Students in Each Level	41.6	20.1	23.5	14.8	

* Participants' Large Group Medians

AIMS Standard Setting Grade 10 Science

Recommended Cut Points* Plus/Minus Selected Standard Errors (SEs) of Measurement

Performance Level	Falls Far Below	Approaches	Meets	Exceeds	
Standard Error (SE) measurement		14.00	15.00	16.00	
Recommended Cut Point* + 3 SE		539	567	607	+ 3 SE
Percent of Students in Each Level	78.2	12.7	6.7	2.4	
Recommended Cut Point* + 2 SE		525	552	591	+ 2 SE
Percent of Students in Each Level	70.3	14.9	10.6	4.2	
Recommended Cut Point* + 1 SE		511	537	575	+ 1 SE
Percent of Students in Each Level	58.7	17.0	16.9	7.4	
Recommended Cut Point*		497	522	559	Recommended Cut Points*
Percent of Students in Each Level	48.6	18.6	20.1	12.7	
Recommended Cut Point* -1 SE		483	507	543	-1 SE
Percent of Students in Each Level	37.7	18.1	24.9	19.3	
Recommended Cut Point* -2 SE		469	492	527	-2 SE
Percent of Students in Each Level	28.5	15.8	26.0	29.7	
Recommended Cut Point* -3 SE		455	477	511	-3 SE
Percent of Students in Each Level	20.0	13.7	25.1	41.2	

* Participants' Large Group Medians

AIMS Standard Setting Grade 10 Science

Recommended Cut Points* Plus/Minus Selected Standard Errors (SEs) of Measurement and the Cut Score

Performance Level	Falls Far Below	Approaches	Meets	Exceeds	
Standard Error (SE) measurement + cutscore		14.44	15.12	16.19	
Recommended Cut Point* + 3 SE		541	567	608	+ 3 SE
Percent of Students in Each Level	78.2	12.7	7.1	2.0	
Recommended Cut Point* + 2 SE		526	552	592	+ 2 SE
Percent of Students in Each Level	70.3	14.9	11.1	3.7	
Recommended Cut Point* + 1 SE		512	537	575	+ 1 SE
Percent of Students in Each Level	58.7	17.0	16.9	7.4	
Recommended Cut Point*		497	522	559	Recommended Cut Points*
Percent of Students in Each Level	48.6	18.6	20.1	12.7	
Recommended Cut Point* -1 SE		483	507	543	-1 SE
Percent of Students in Each Level	37.7	18.1	24.9	19.3	
Recommended Cut Point* -2 SE		469	492	527	-2 SE
Percent of Students in Each Level	28.5	15.8	26.0	29.7	
Recommended Cut Point* -3 SE		454	477	511	-3 SE
Percent of Students in Each Level	18.8	14.8	25.1	41.3	

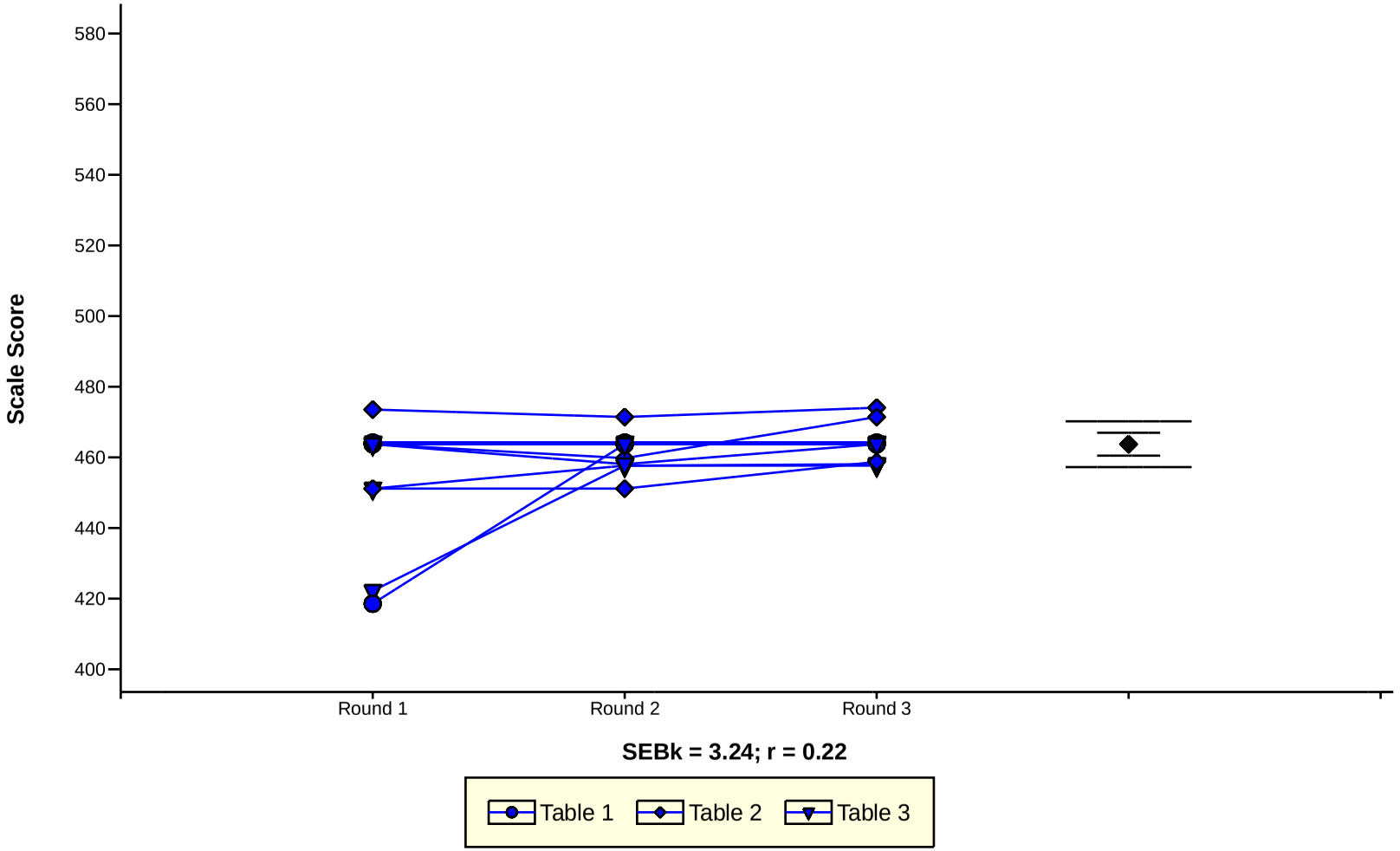
* Participants' Large Group Medians

SECTION H

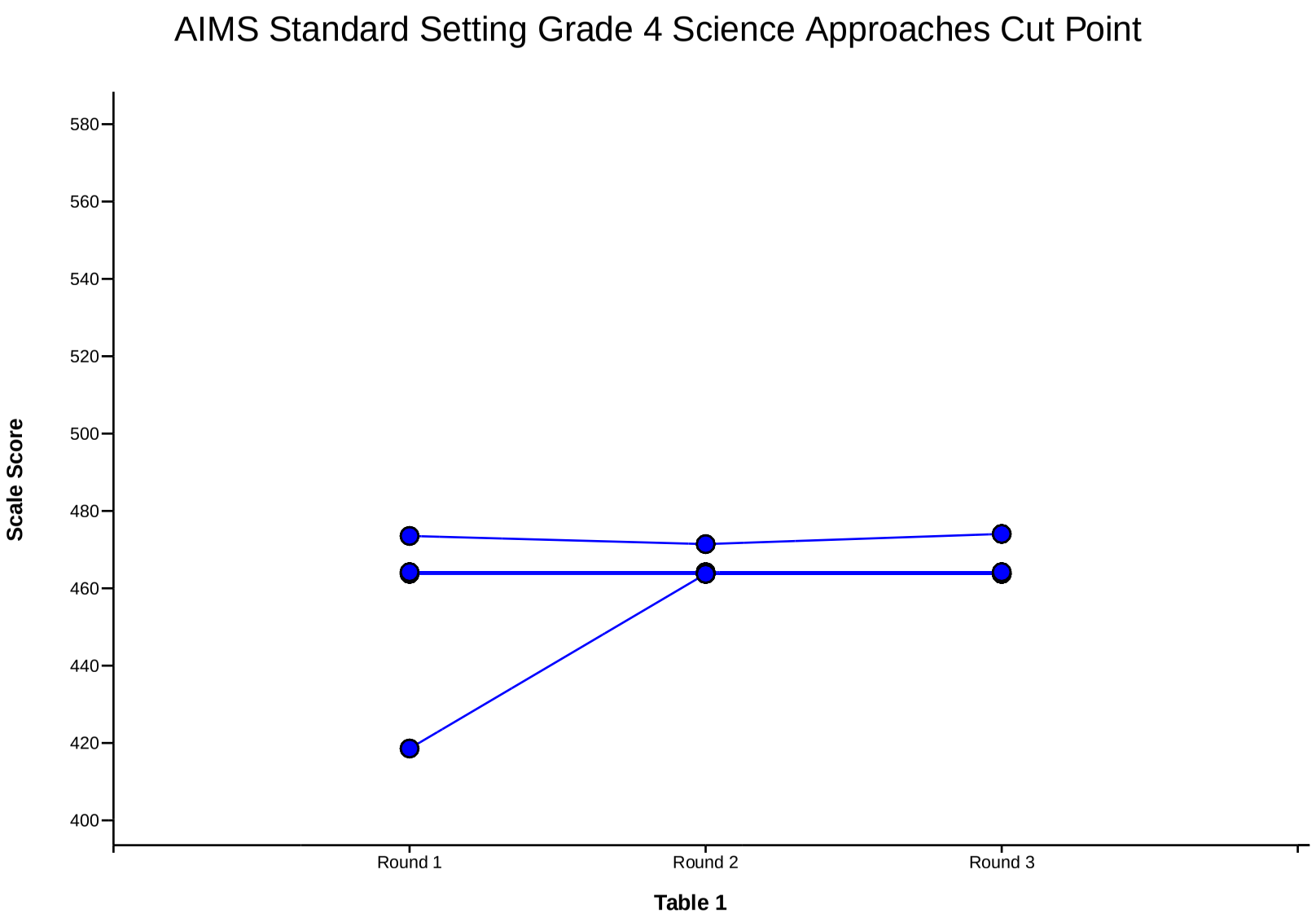
Graphical Representations of Participants' Judgments

H 1

AIMS Standard Setting Grade 4 Science Approaches Cut Point



H 2


AIMS Standard Setting Grade 4 Science Approaches Cut Point
Scale Score
580
560
540
520
500
480
460
440
420
400
Round 1
Round 2
Round 3
Table 1

H 3

AIMS Standard Setting Grade 4 Science Approaches Cut Point

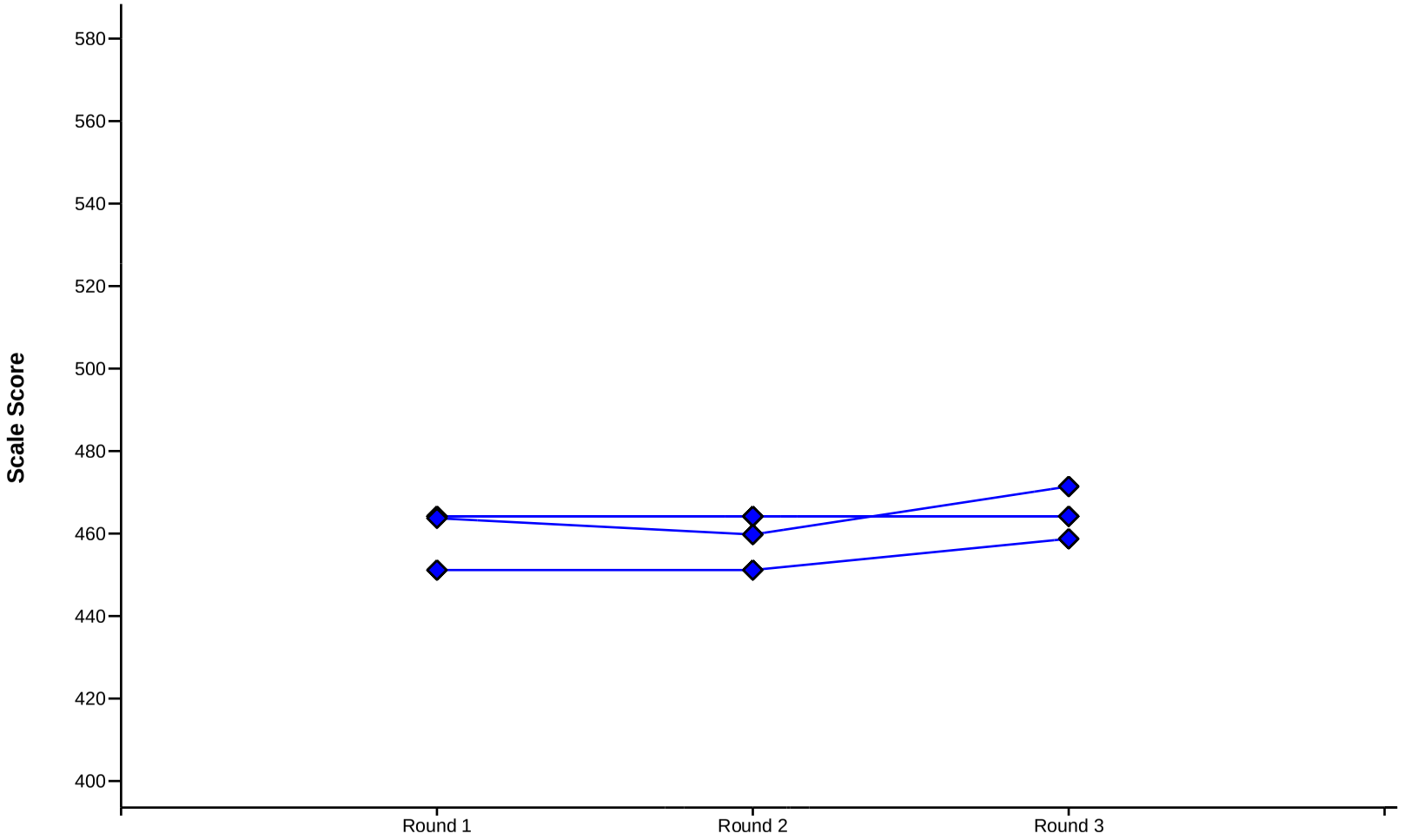


Table 2

H 4

AIMS Standard Setting Grade 4 Science Approaches Cut Point

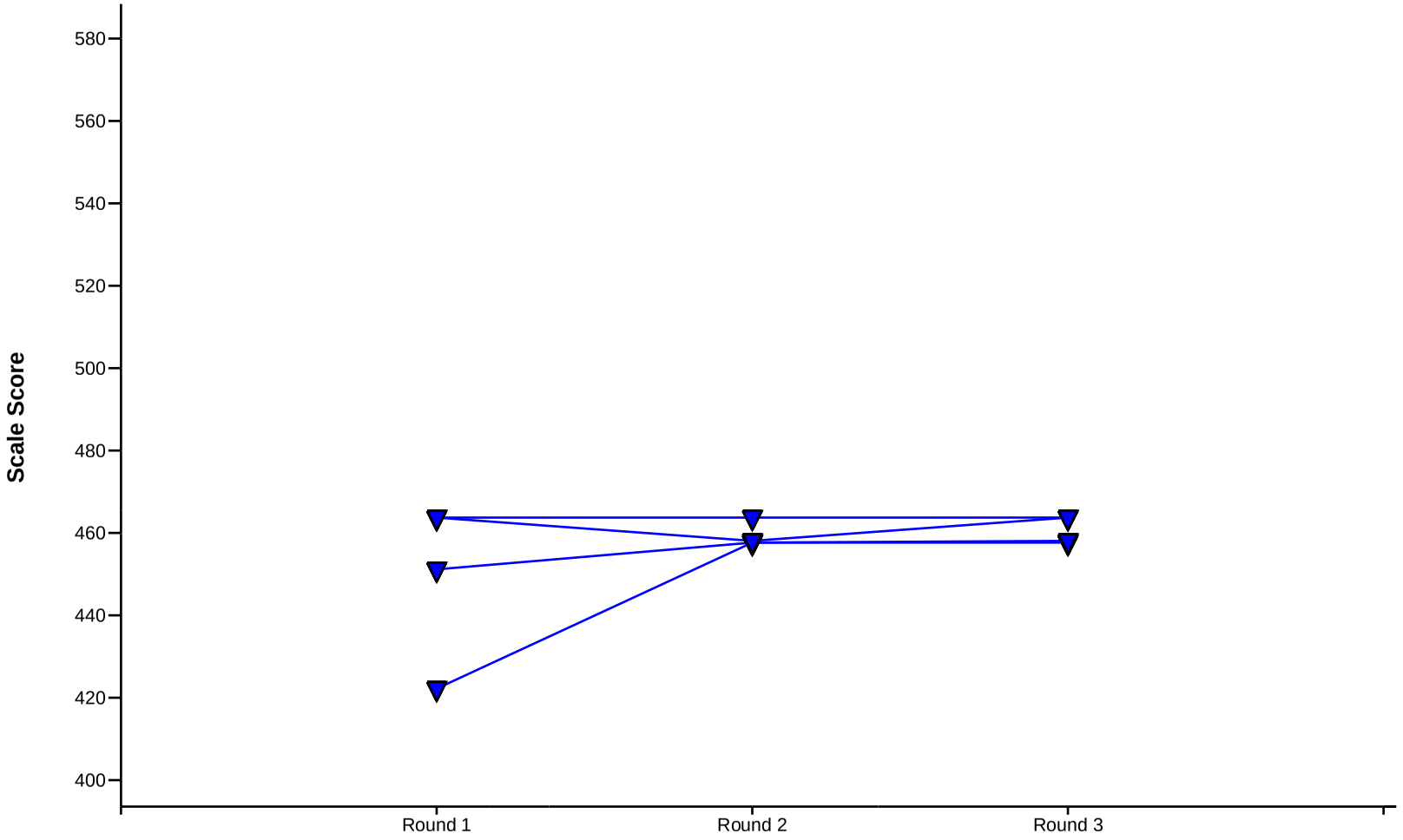
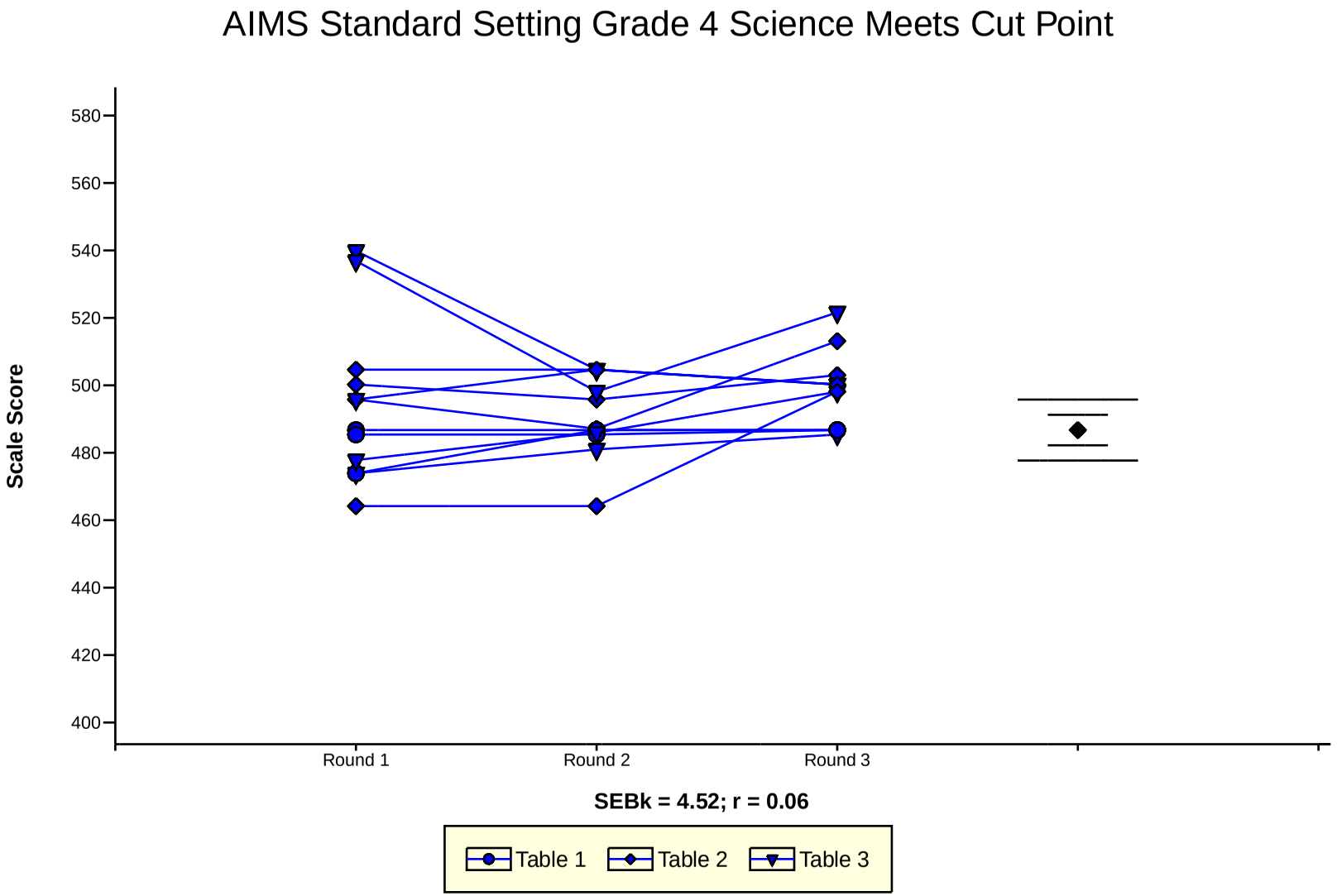
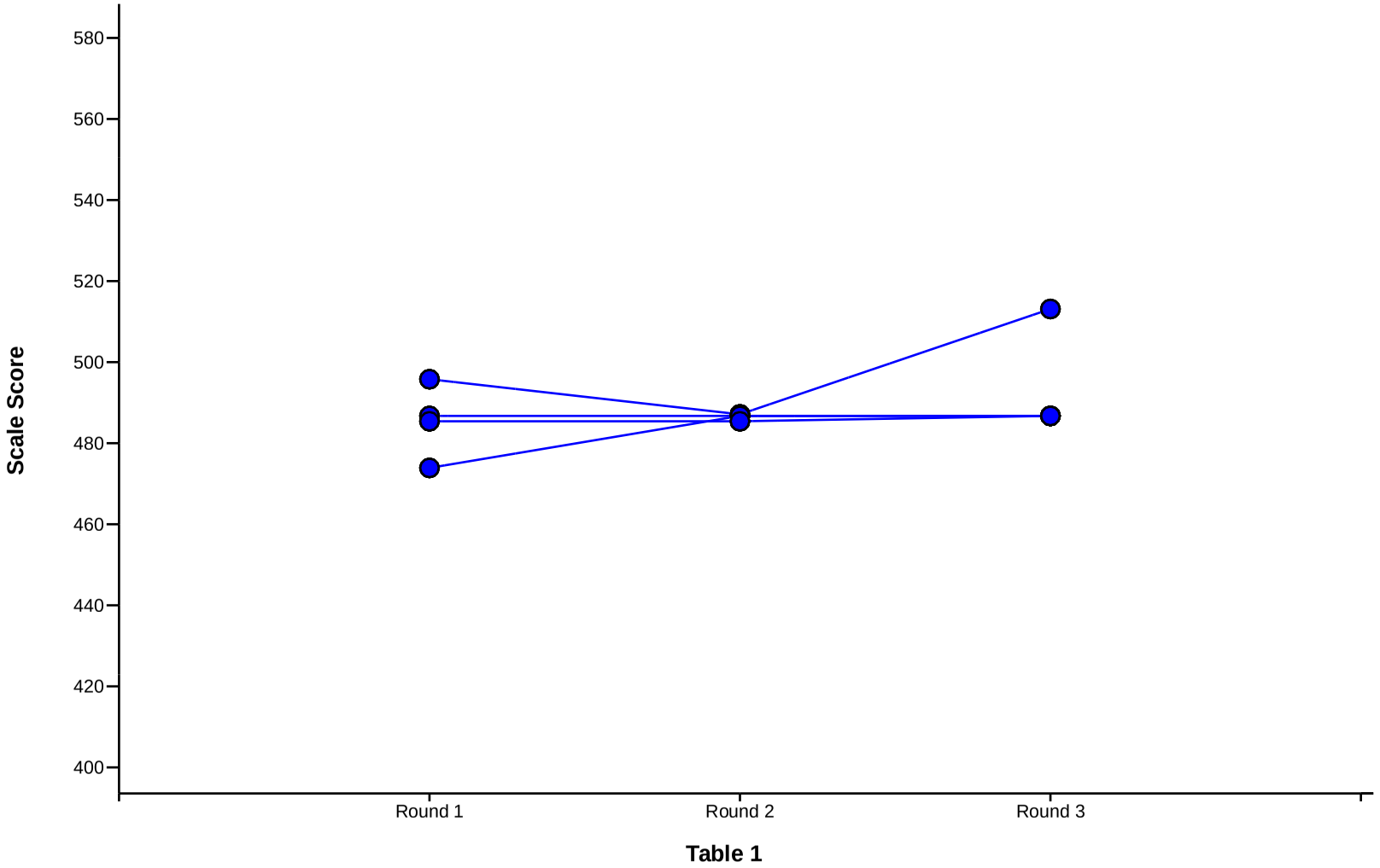


Table 3

H 5


AIMS Standard Setting Grade 4 Science Meets Cut Point
Scale Score
580
560
540
520
500
480
460
440
420
400
Round 1
Round 2
Round 3
SEBk = 4.52; r = 0.06
Table 1
Table 2
Table 3

AIMS Standard Setting Grade 4 Science Meets Cut Point



AIMS Standard Setting Grade 4 Science Meets Cut Point

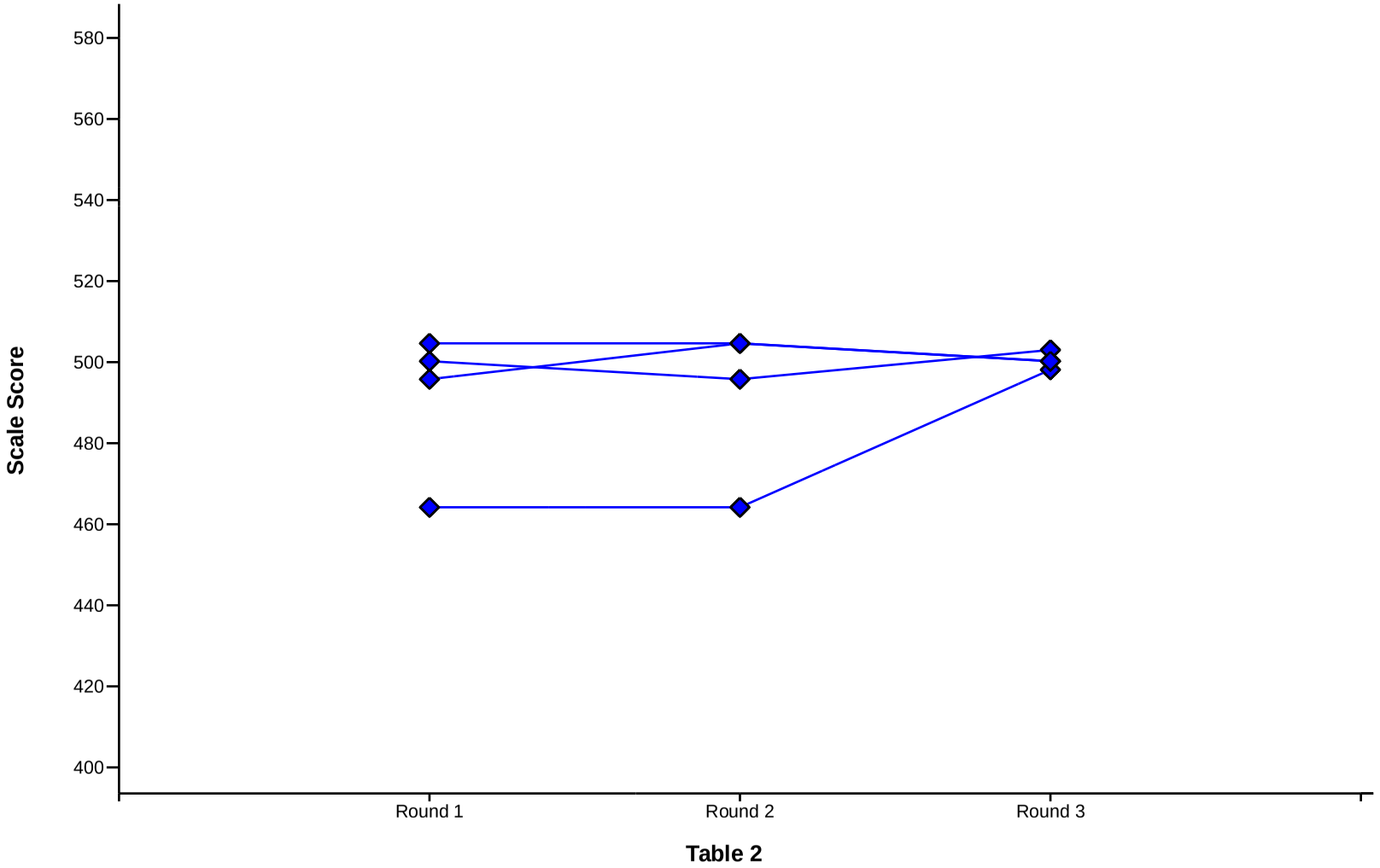
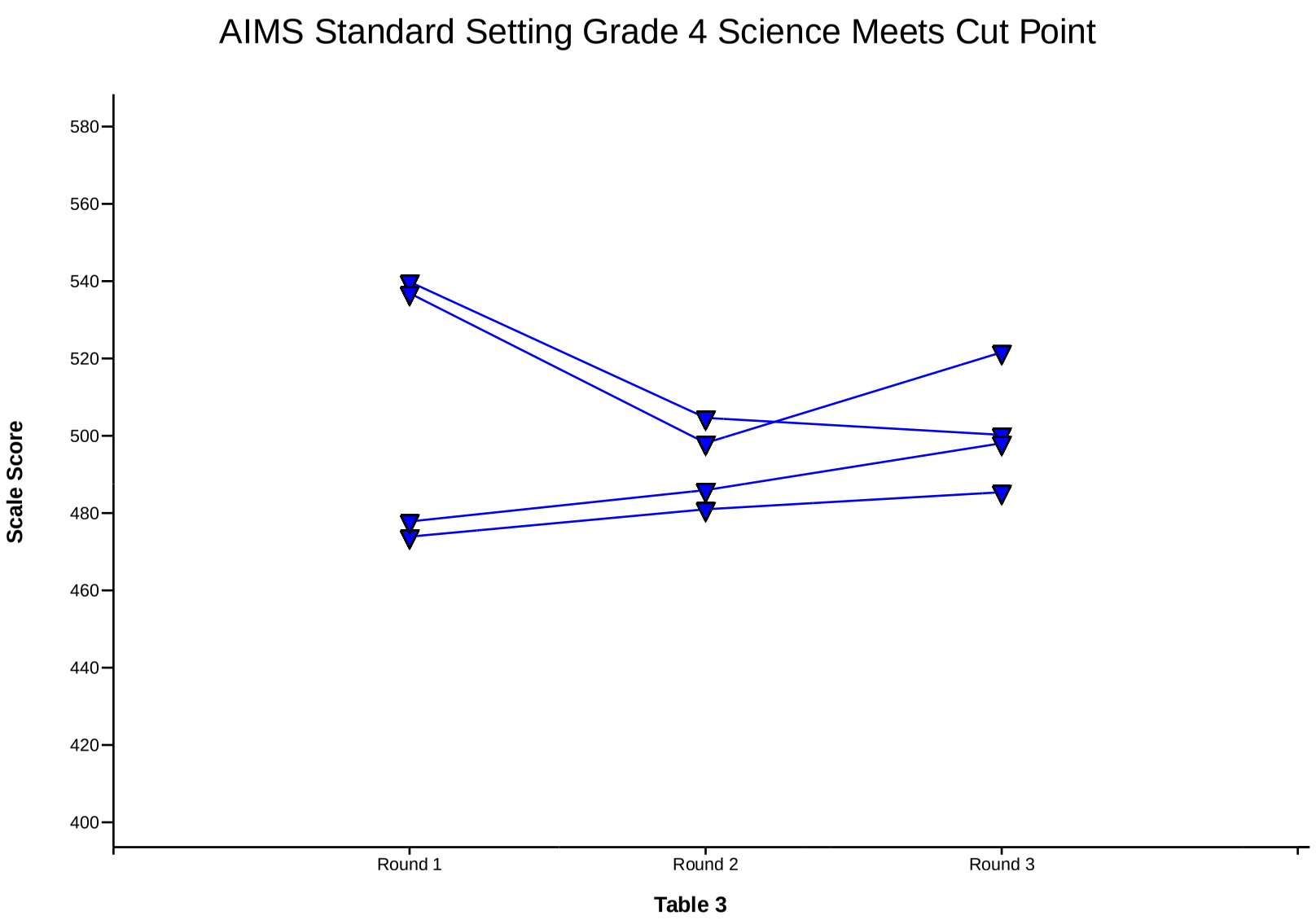
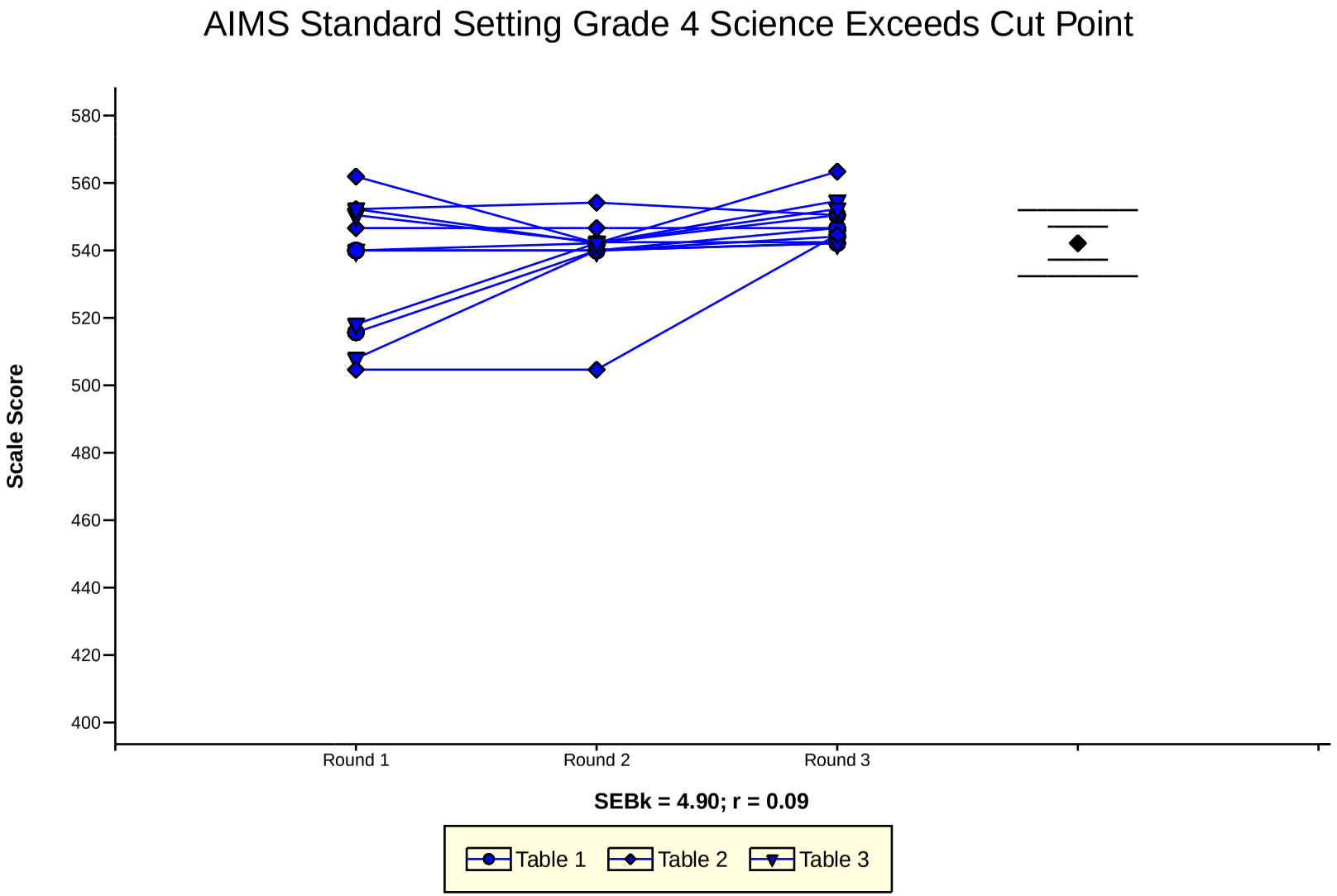


Table 2

8 H

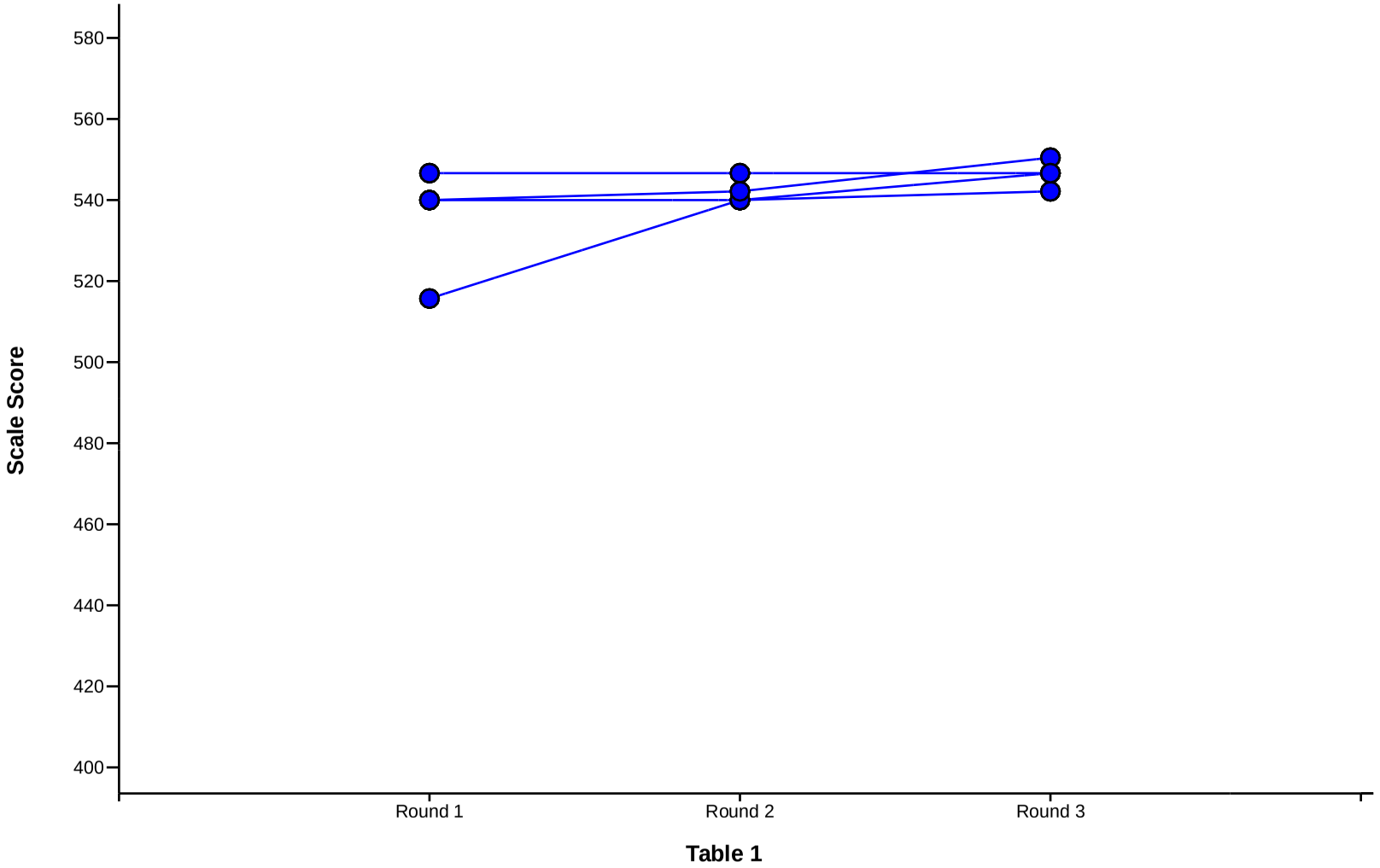

AIMS Standard Setting Grade 4 Science Meets Cut Point
Scale Score
580
560
540
520
500
480
460
440
420
400
Round 1
Round 2
Round 3
Table 3

6 H


AIMS Standard Setting Grade 4 Science Exceeds Cut Point
Scale Score
580
560
540
520
500
480
460
440
420
400
Round 1
Round 2
Round 3
SEBk = 4.90; r = 0.09
Table 1
Table 2
Table 3

H 10

AIMS Standard Setting Grade 4 Science Exceeds Cut Point



H 11

AIMS Standard Setting Grade 4 Science Exceeds Cut Point

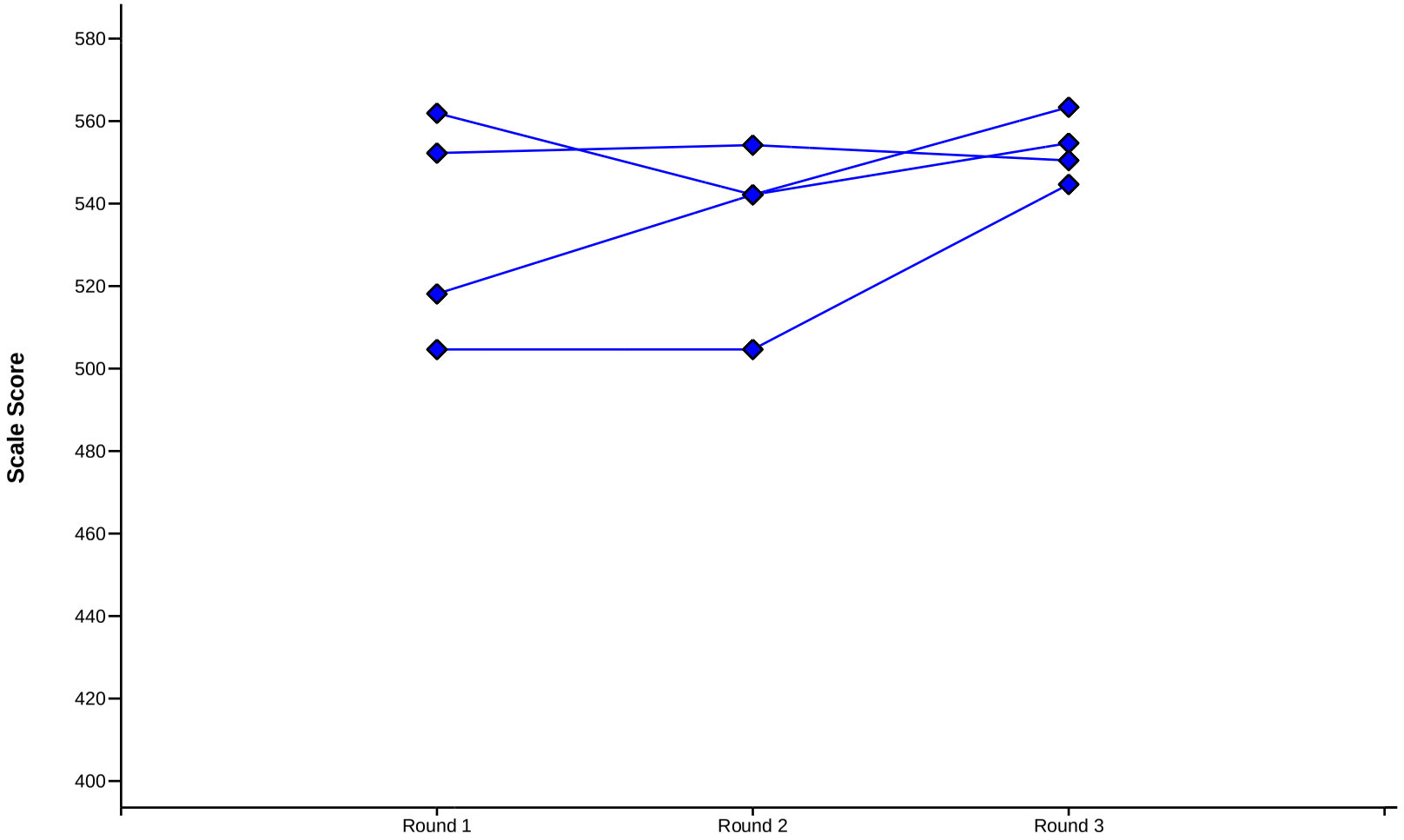


Table 2

H 12

AIMS Standard Setting Grade 4 Science Exceeds Cut Point

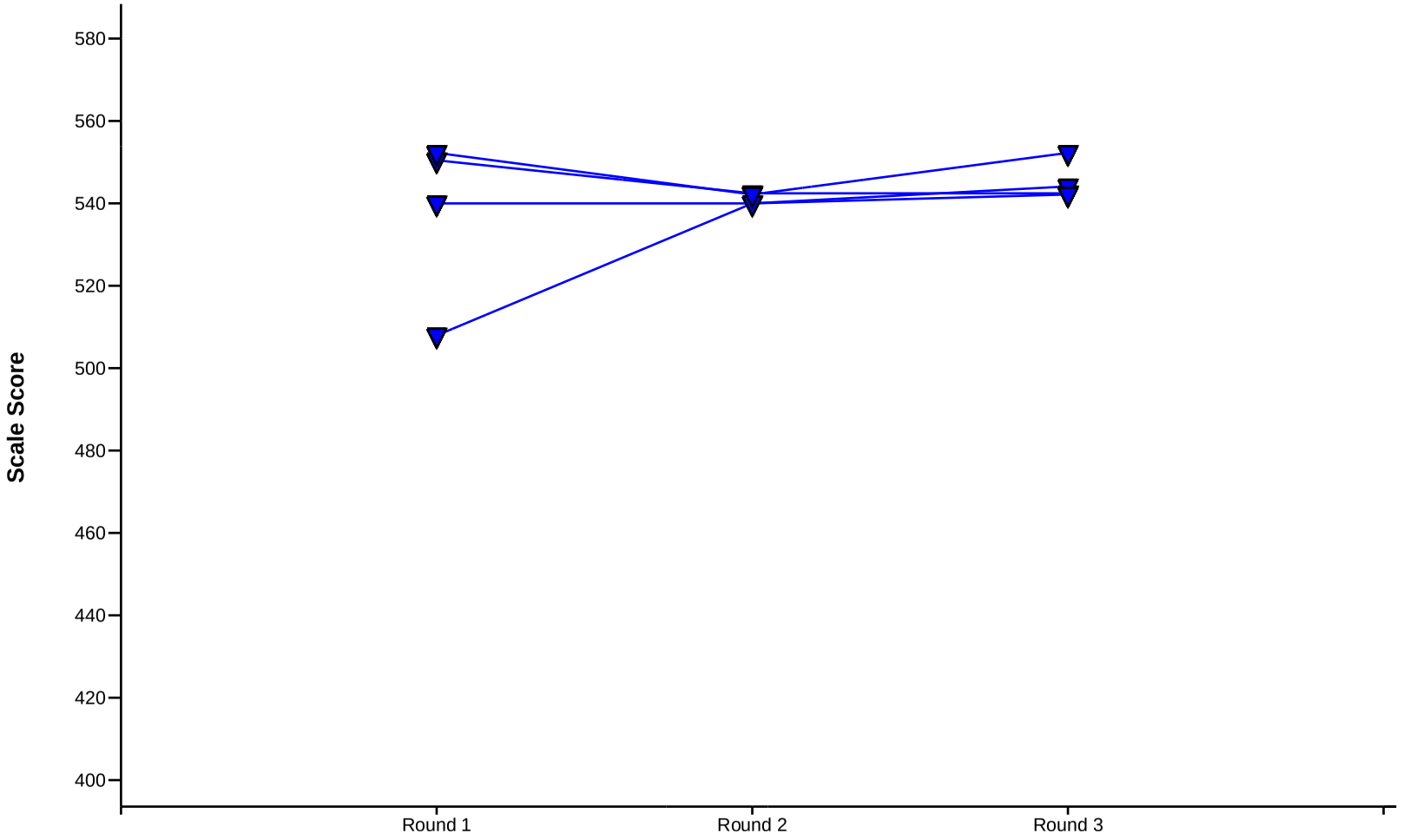
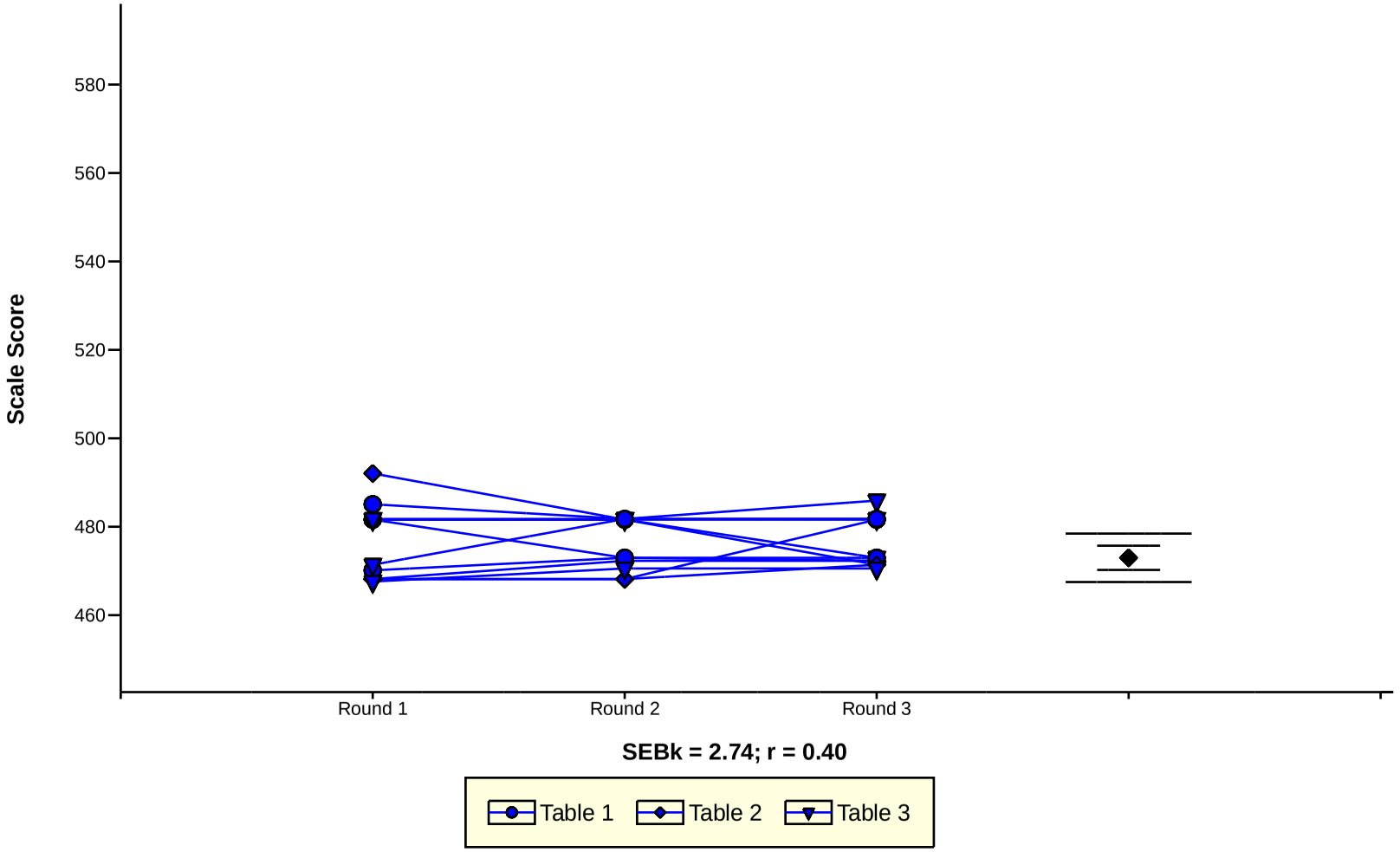


Table 3

H 13

AIMS Standard Setting Grade 8 Science Approaches Cut Point



H 14

AIMS Standard Setting Grade 8 Science Approaches Cut Point

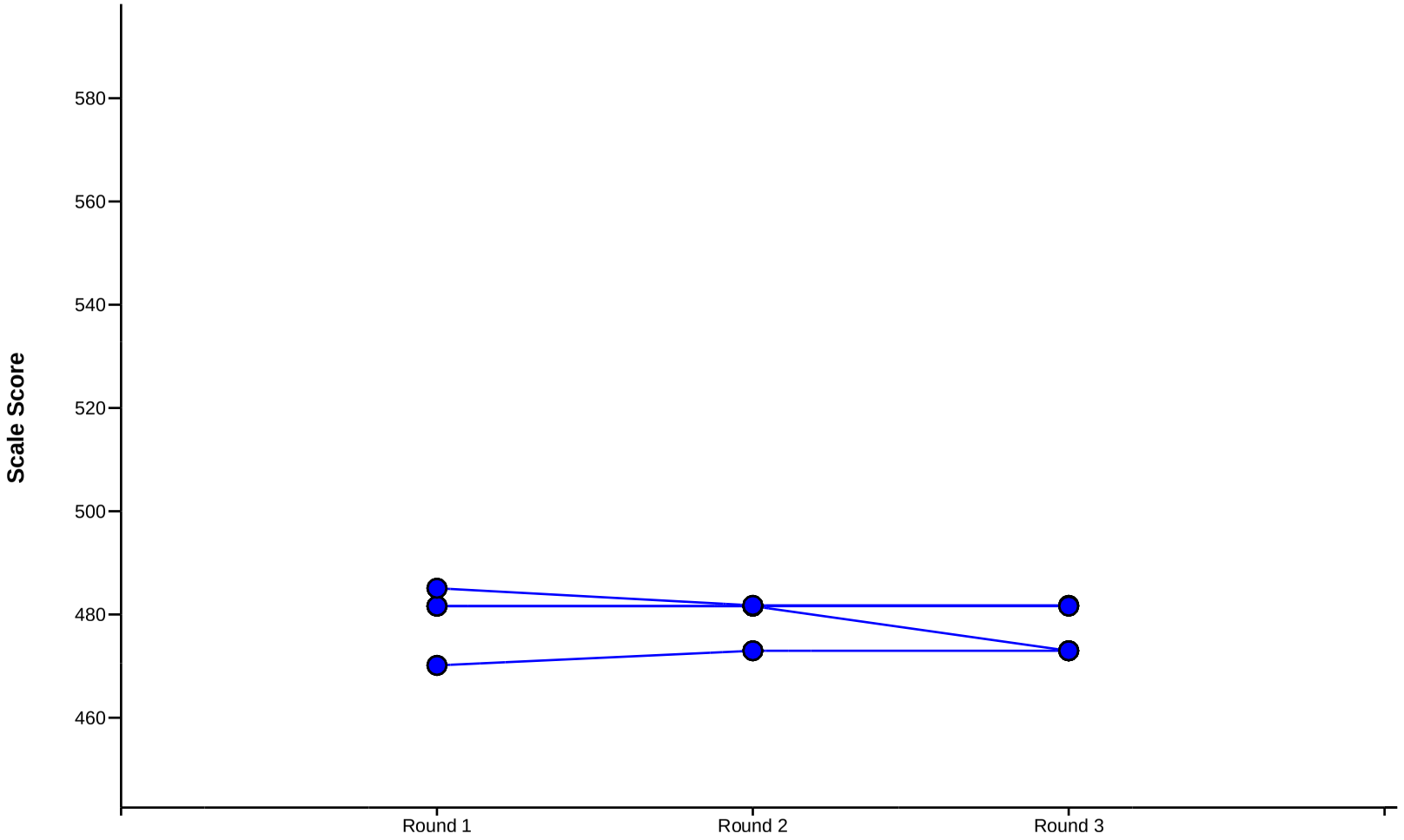


Table 1

H 15

AIMS Standard Setting Grade 8 Science Approaches Cut Point

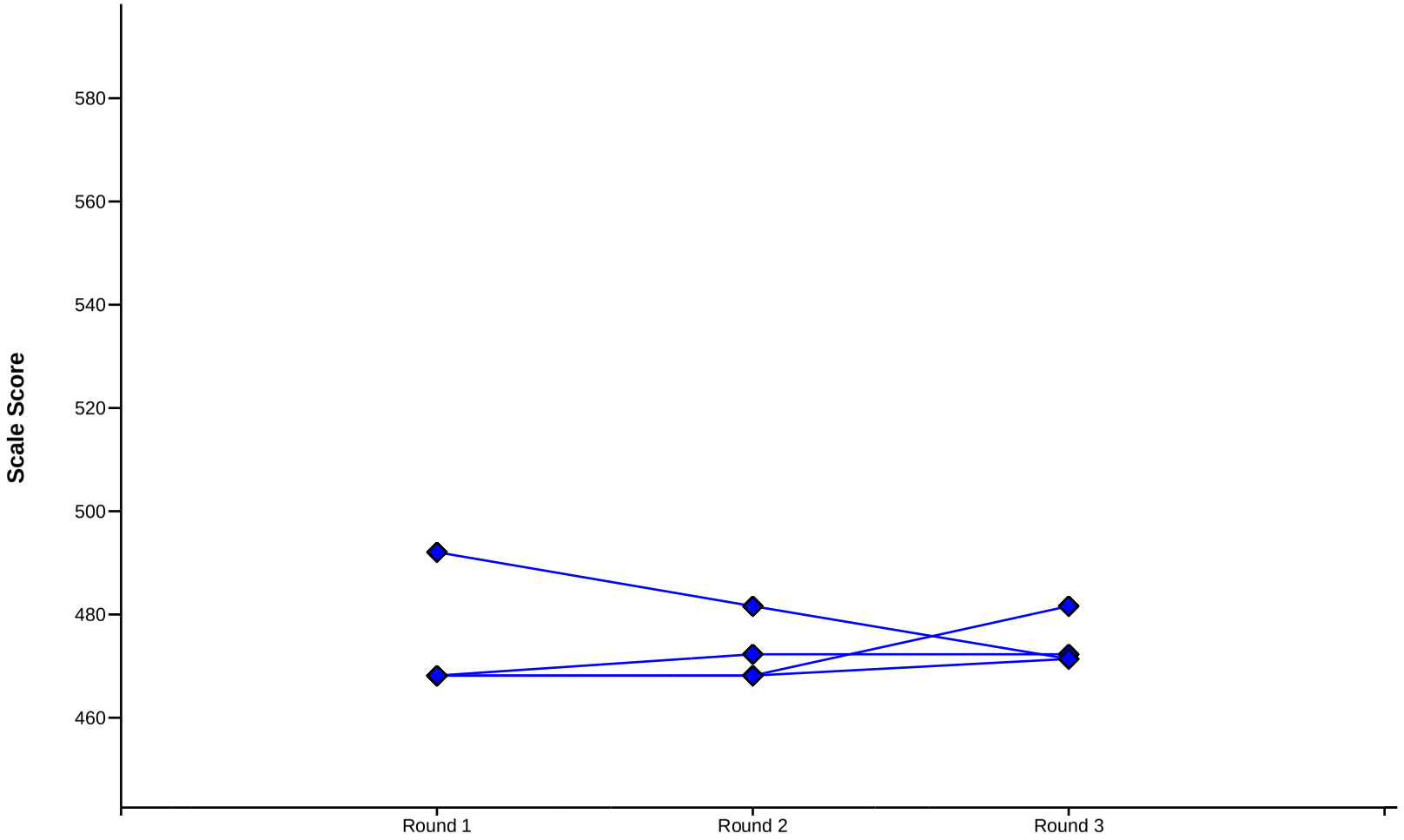


Table 2

AIMS Standard Setting Grade 8 Science Approaches Cut Point

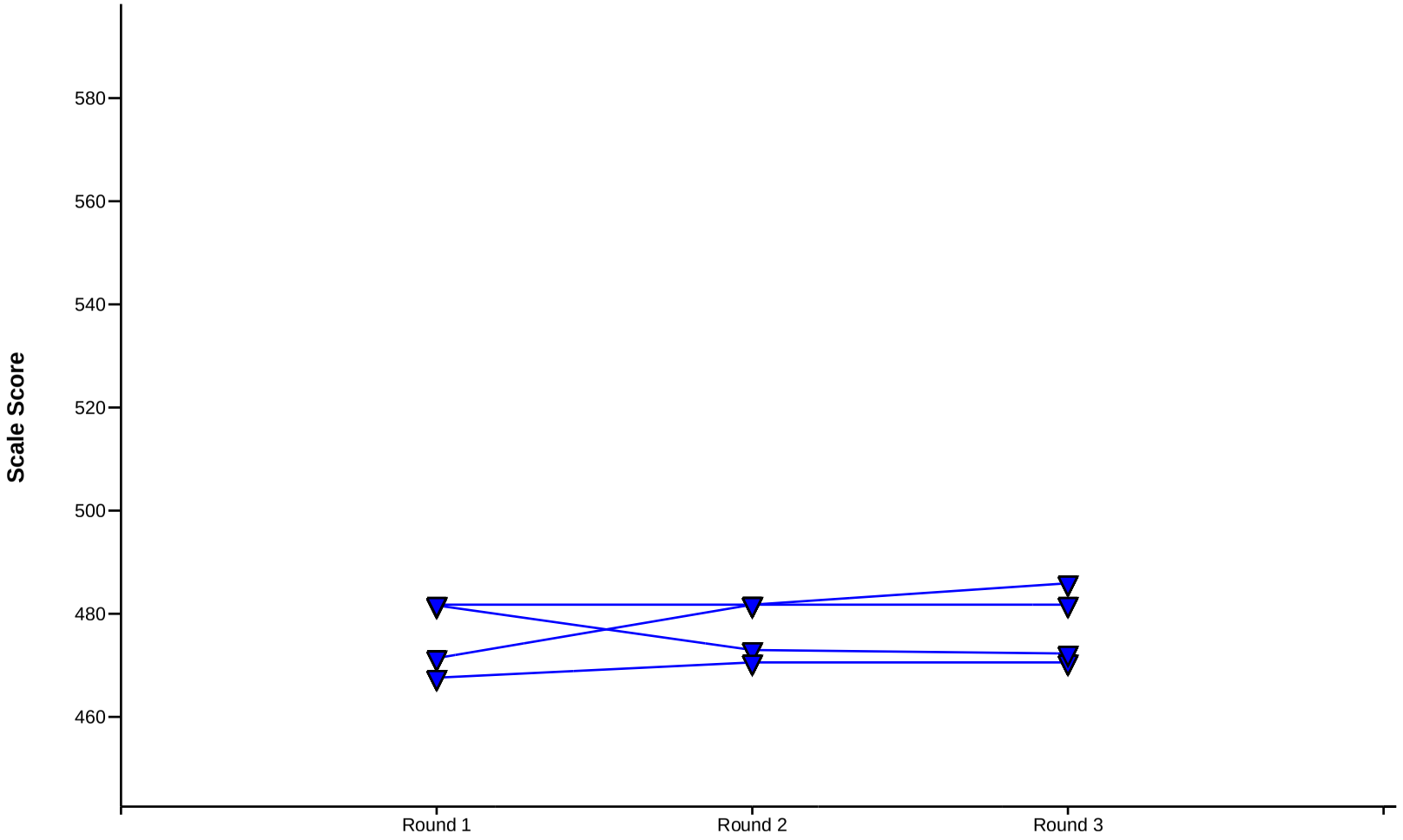
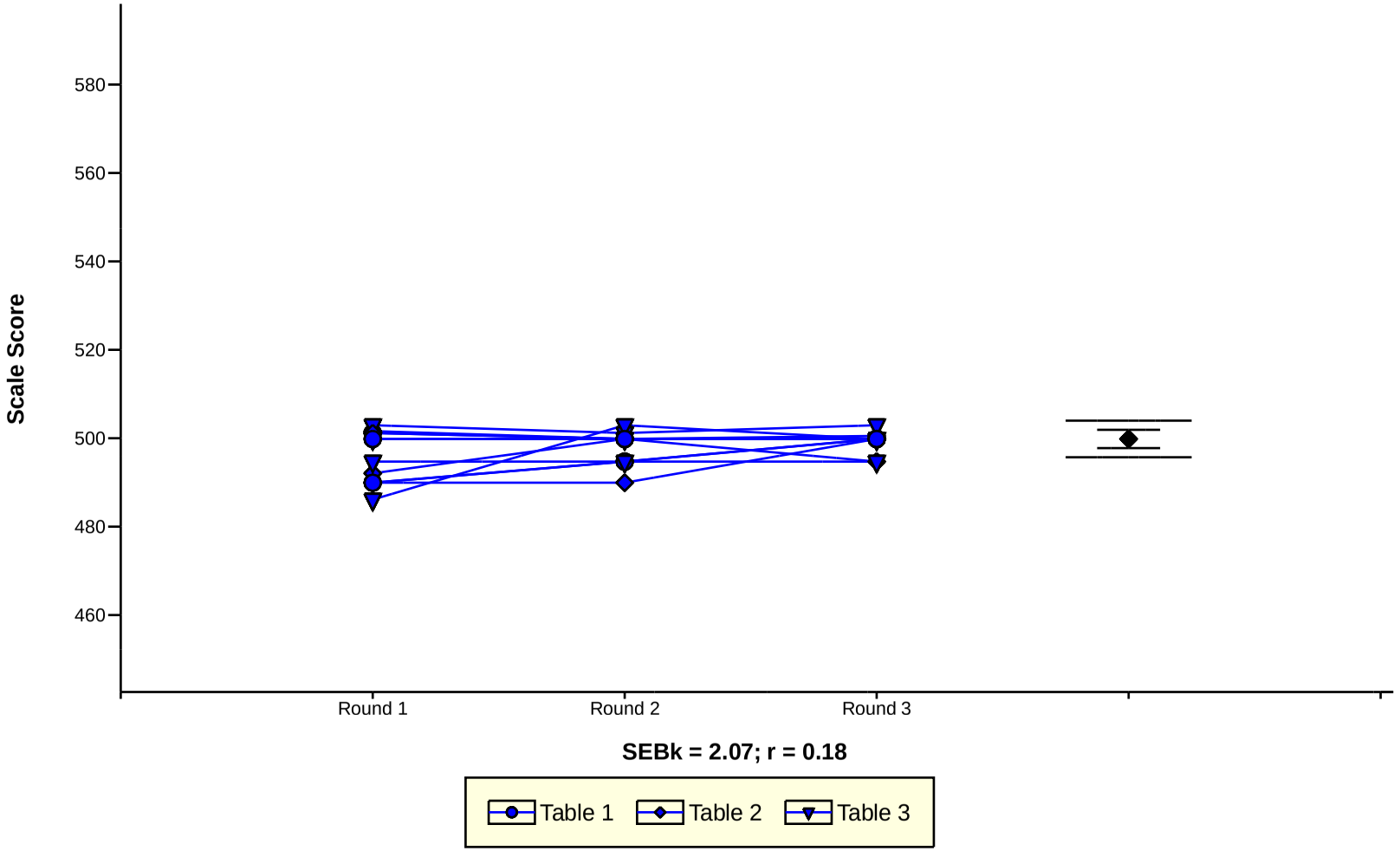


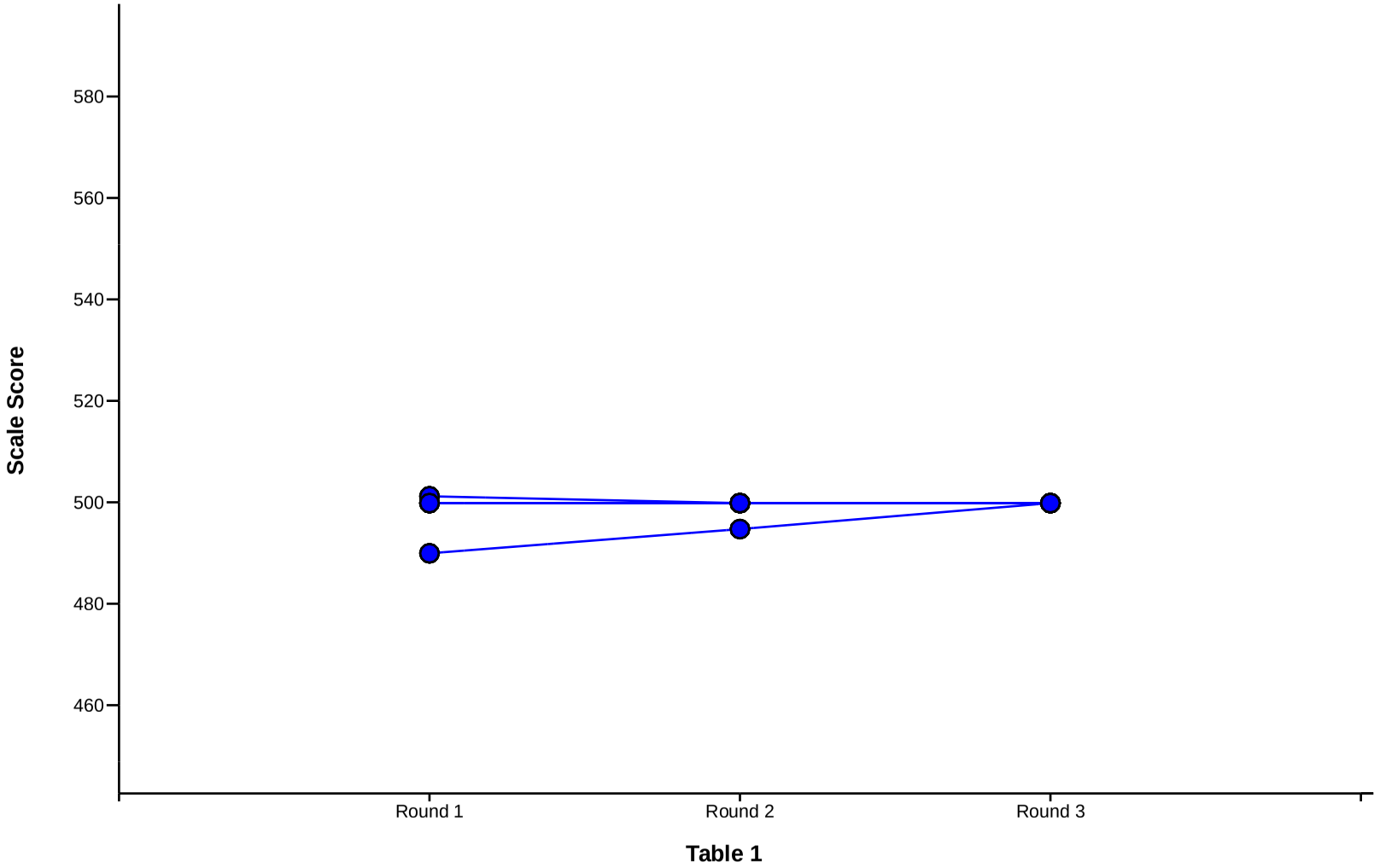
Table 3

AIMS Standard Setting Grade 8 Science Meets Cut Point



H 18

AIMS Standard Setting Grade 8 Science Meets Cut Point



AIMS Standard Setting Grade 8 Science Meets Cut Point

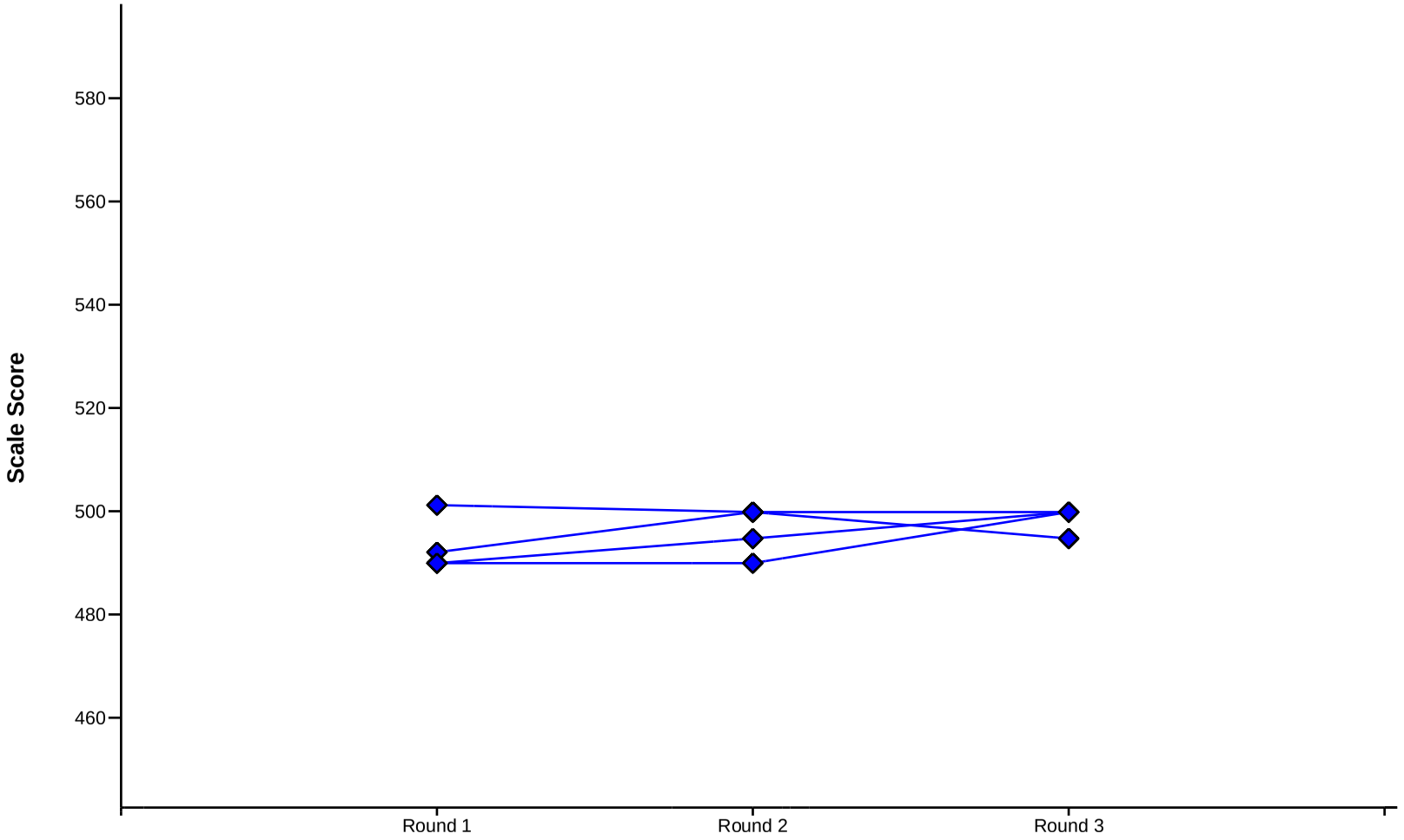


Table 2

H 20

AIMS Standard Setting Grade 8 Science Meets Cut Point

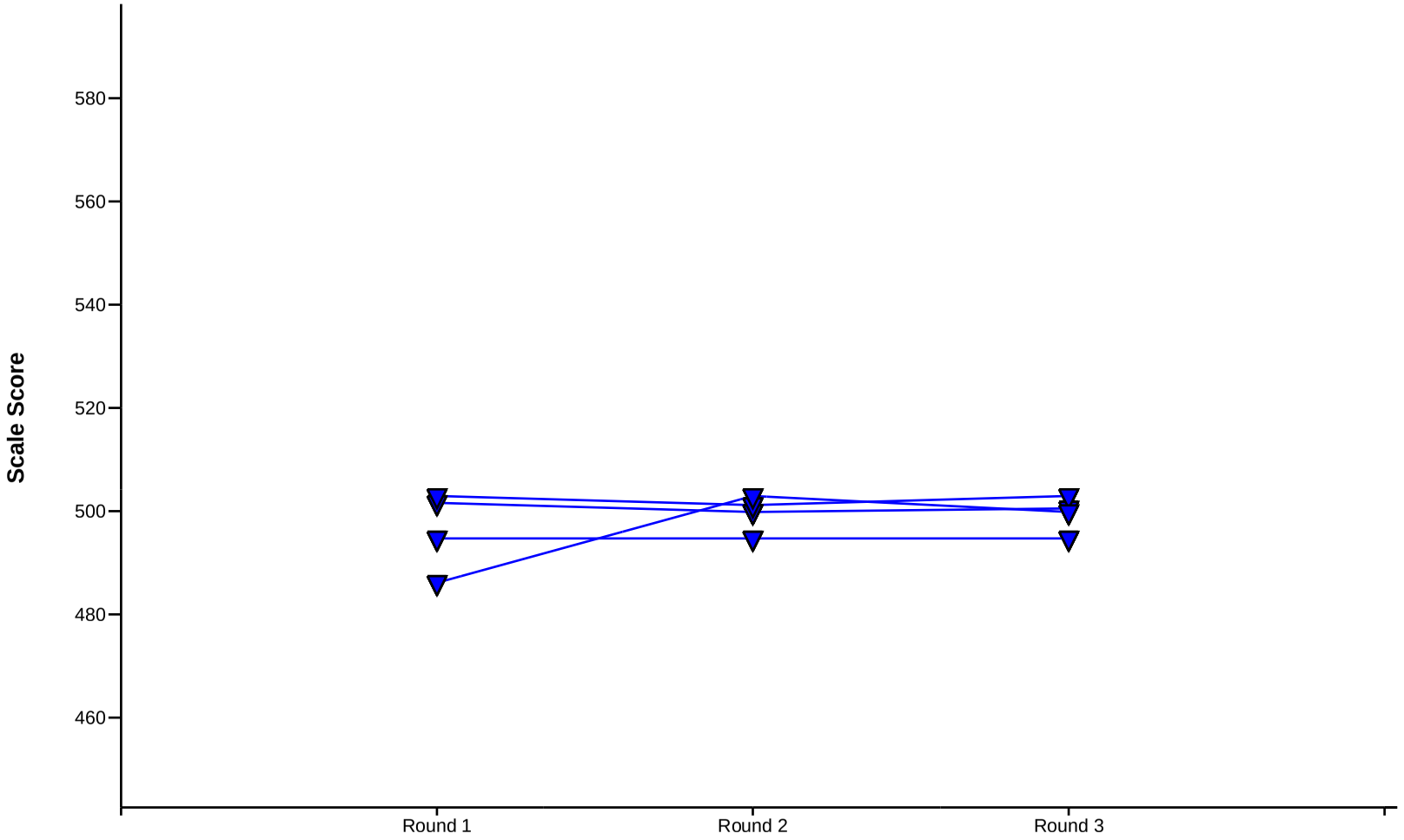
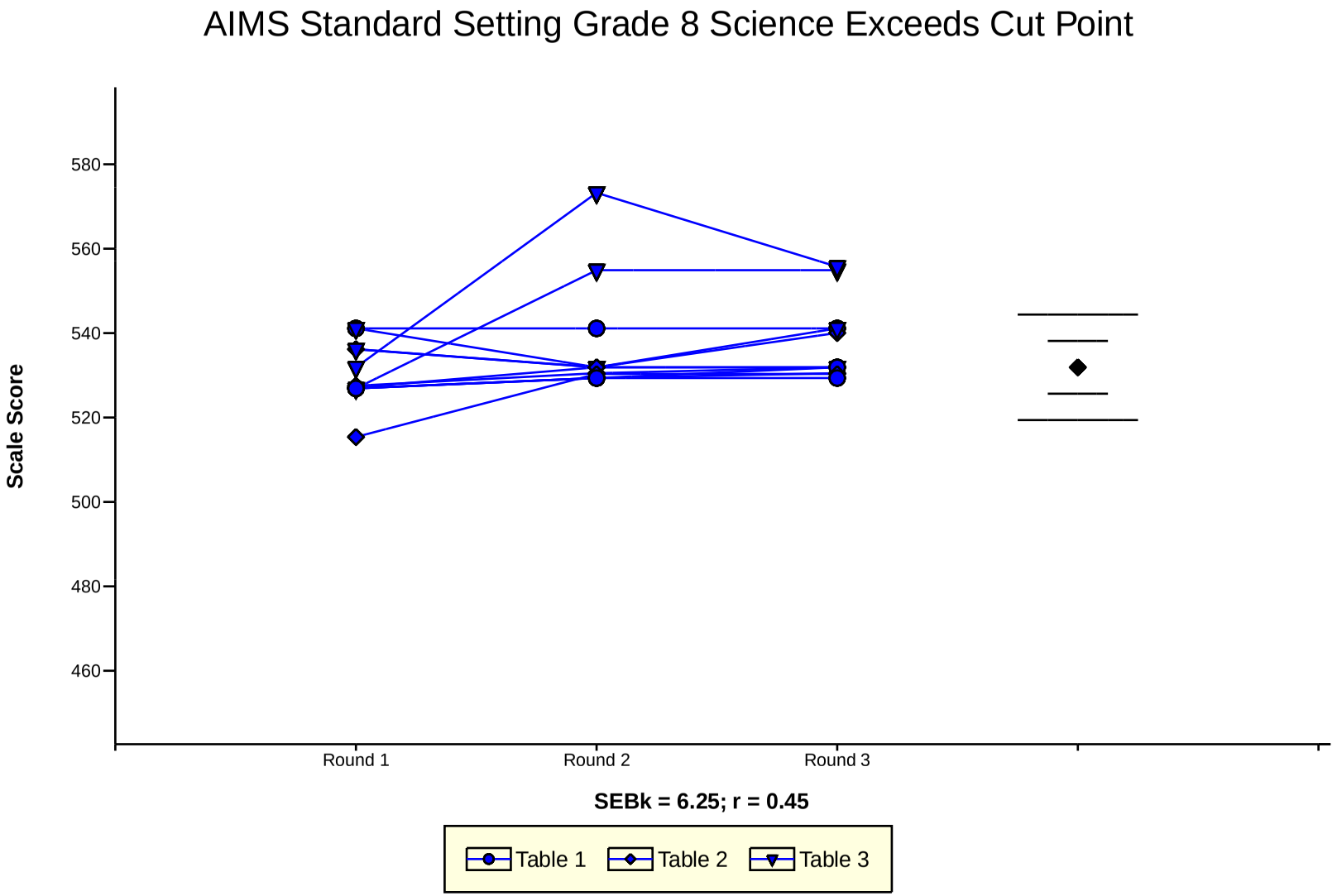


Table 3

H 21


AIMS Standard Setting Grade 8 Science Exceeds Cut Point
Scale Score
580
560
540
520
500
480
460
Round 1
Round 2
Round 3
SEBk = 6.25; r = 0.45
Table 1
Table 2
Table 3

H 22

AIMS Standard Setting Grade 8 Science Exceeds Cut Point

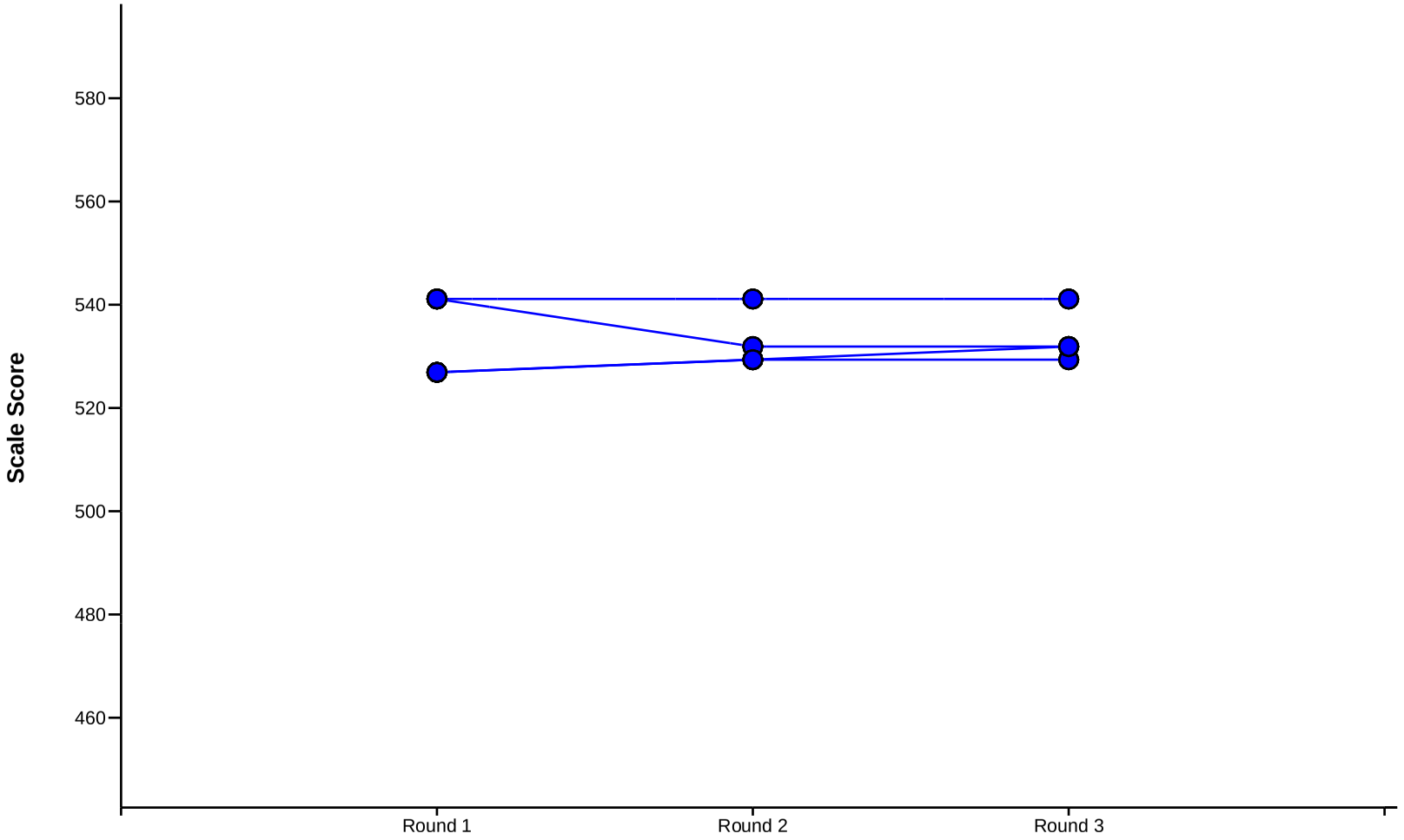


Table 1

H 23

AIMS Standard Setting Grade 8 Science Exceeds Cut Point

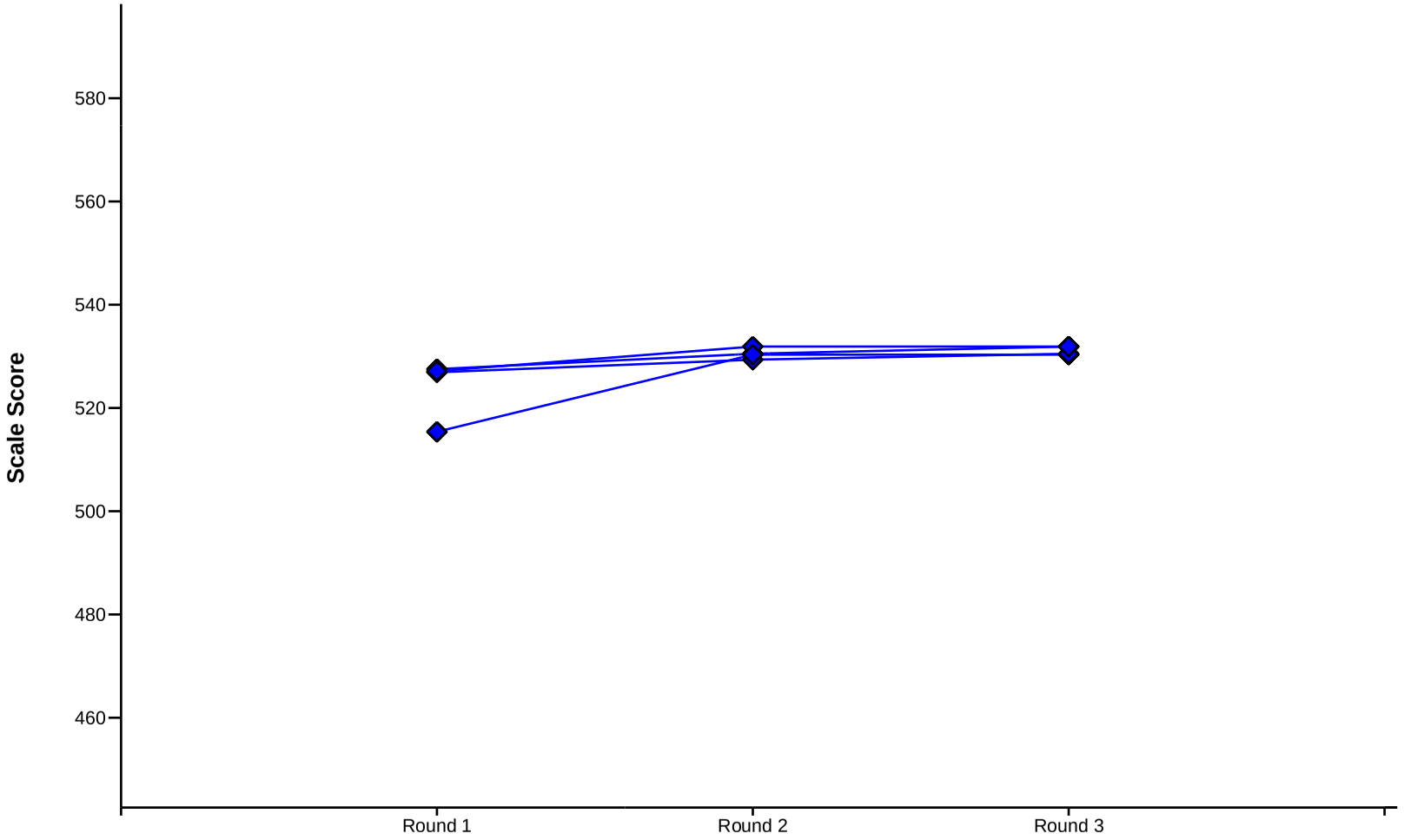


Table 2

H 24

AIMS Standard Setting Grade 8 Science Exceeds Cut Point

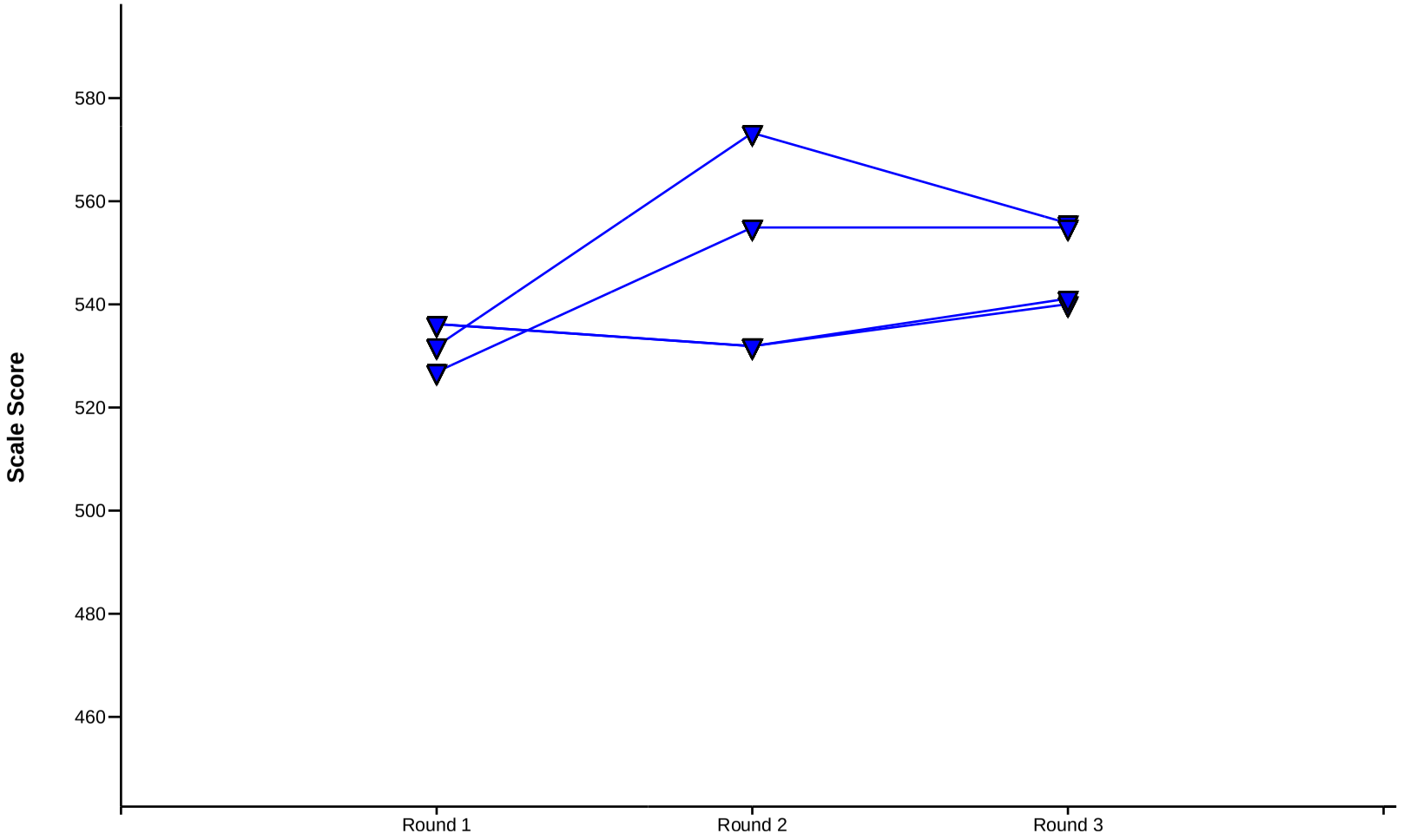
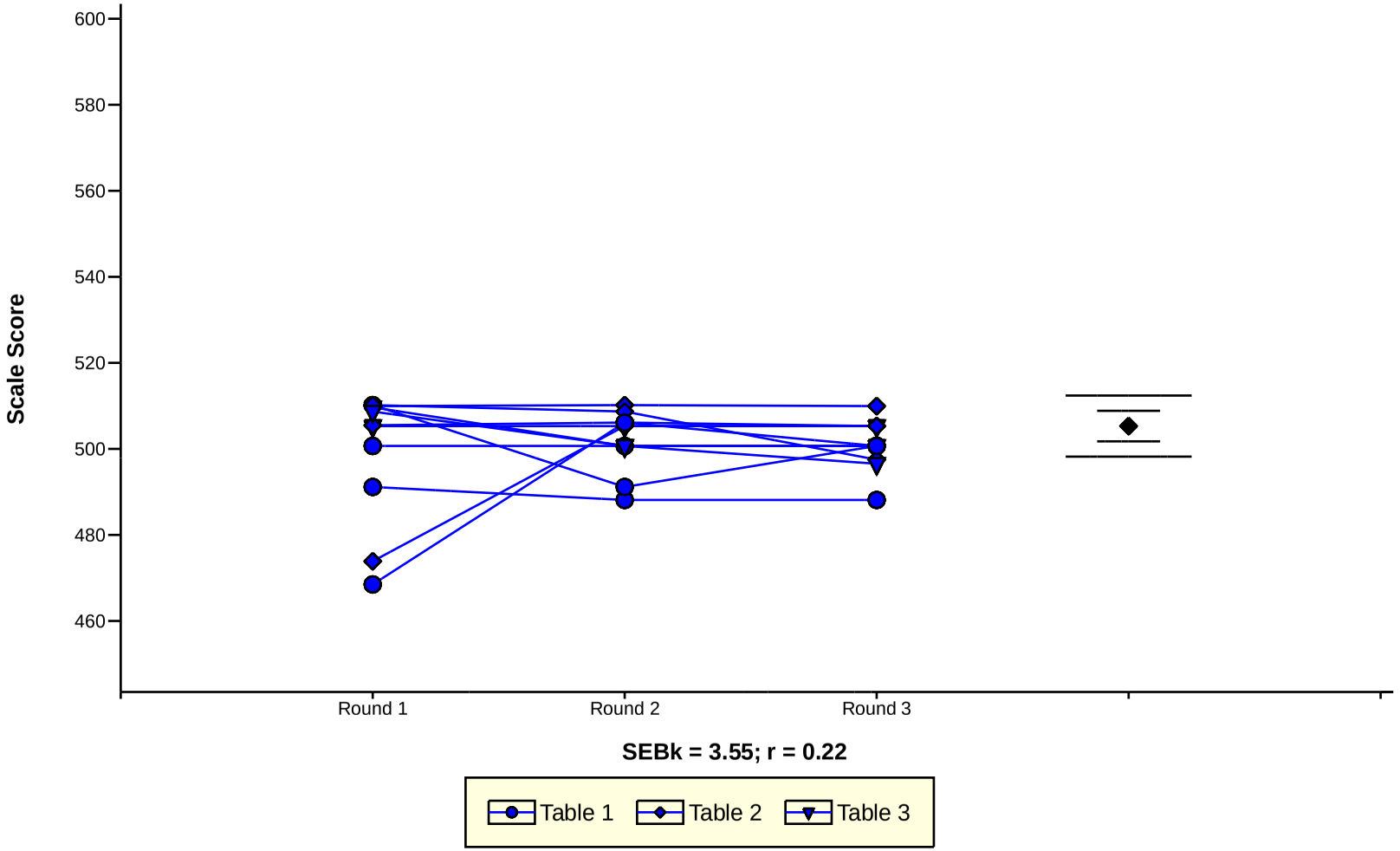


Table 3

H 25

AIMS Standard Setting Grade 10 Science Approaches Cut Point



H 26

AIMS Standard Setting Grade 10 Science Approaches Cut Point

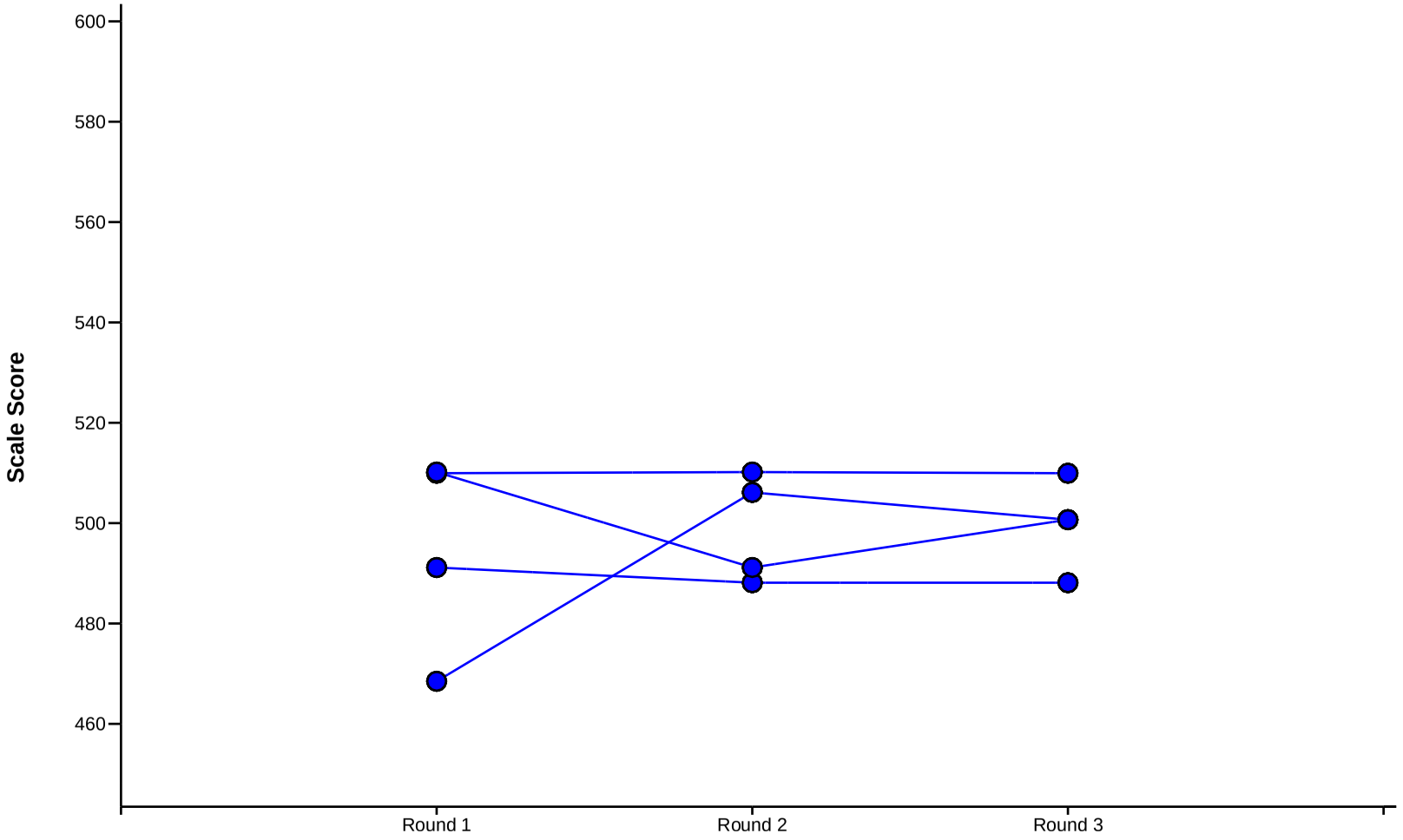
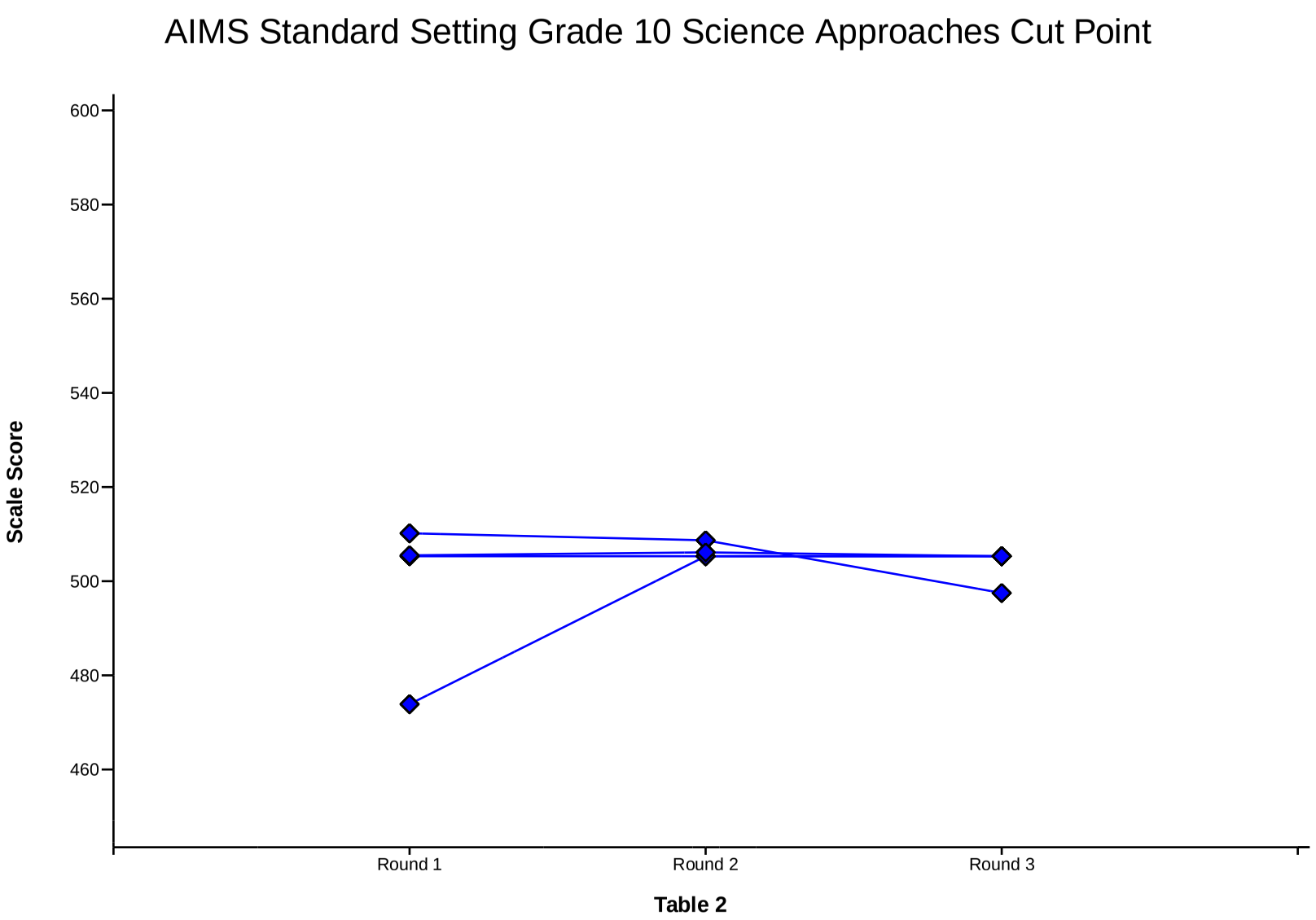


Table 1

H 27


AIMS Standard Setting Grade 10 Science Approaches Cut Point
Scale Score
600
580
560
540
520
500
480
460
Round 1
Round 2
Round 3
Table 2

H 28

AIMS Standard Setting Grade 10 Science Approaches Cut Point

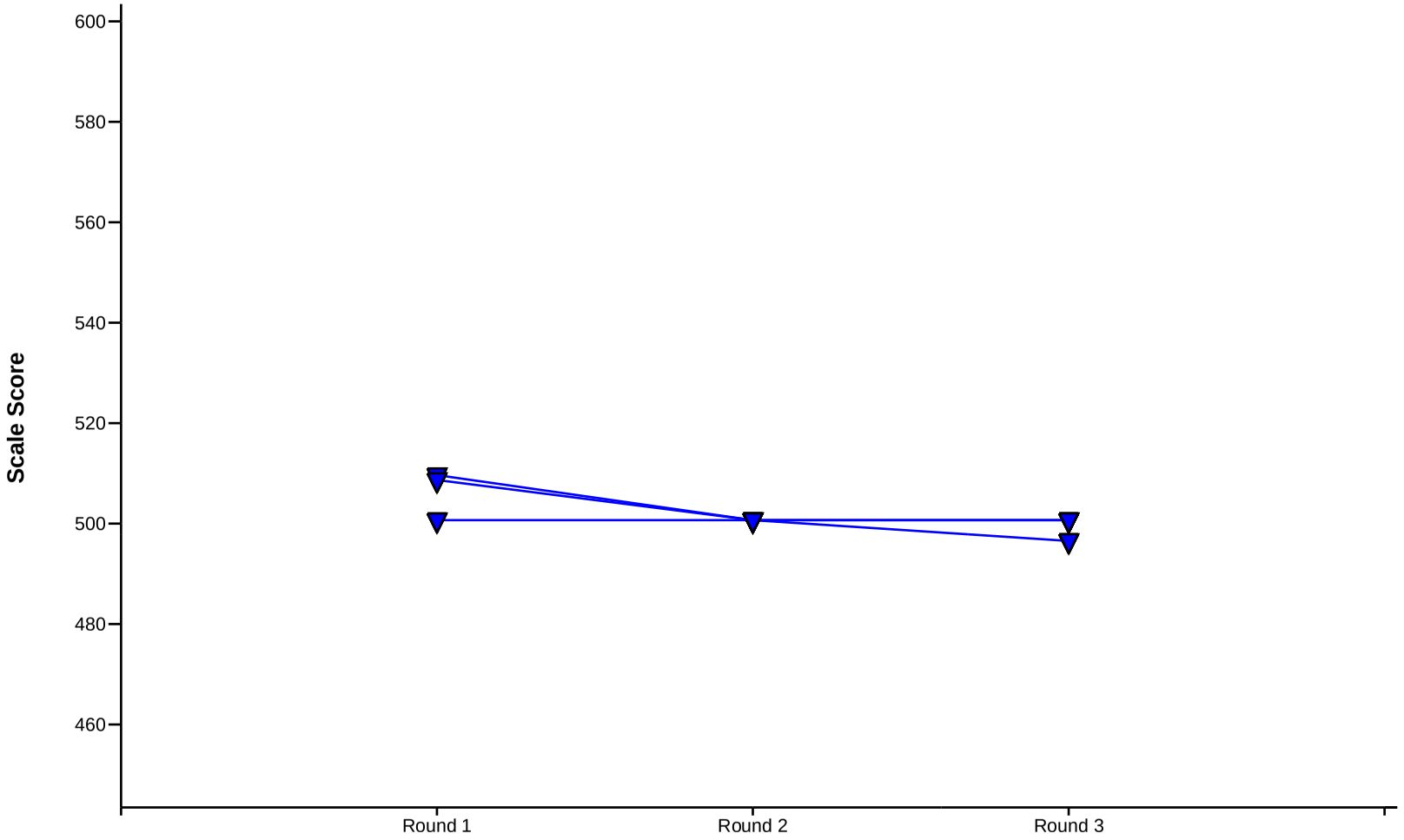
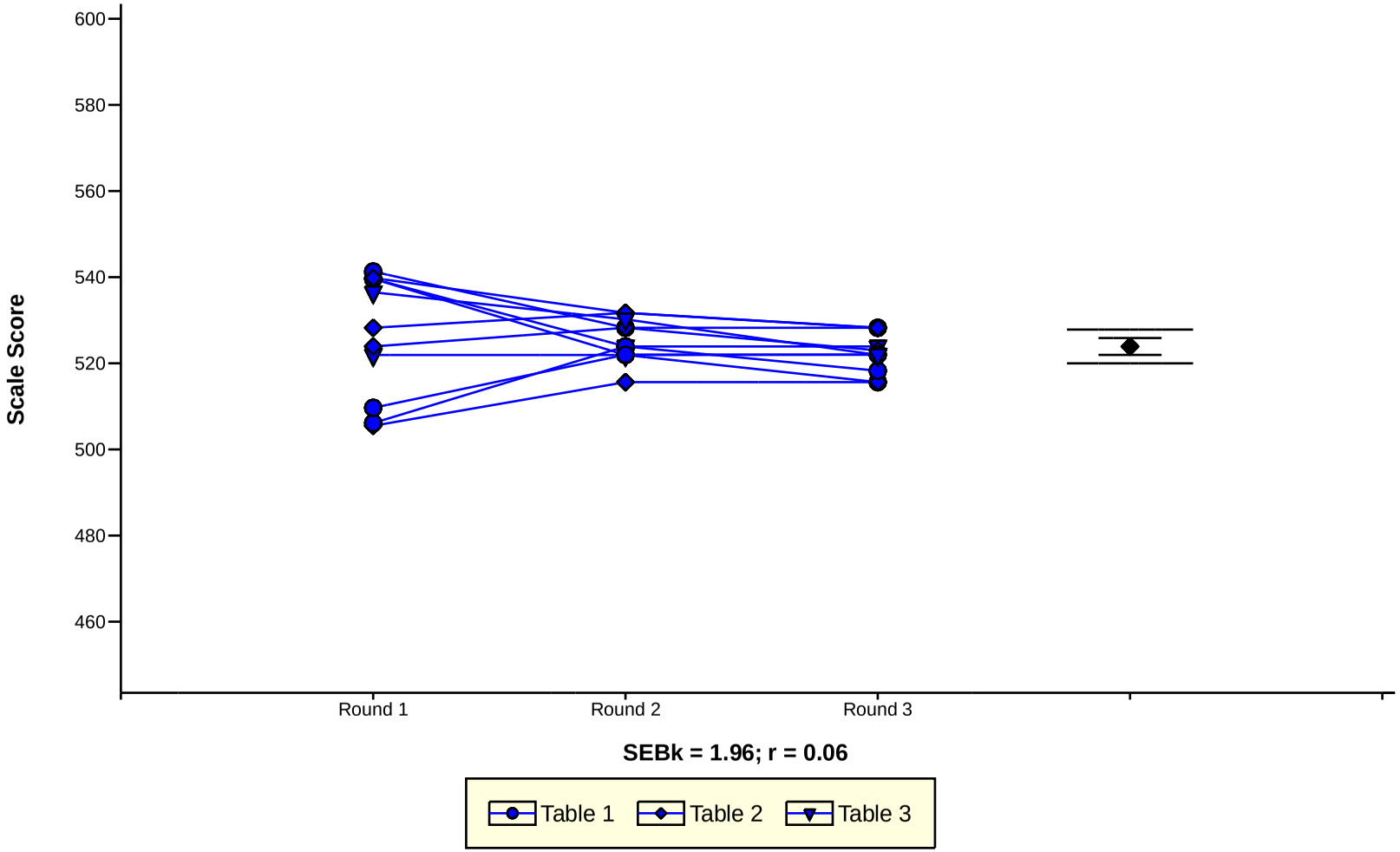


Table 3

H 29

AIMS Standard Setting Grade 10 Science Meets Cut Point



H 30

AIMS Standard Setting Grade 10 Science Meets Cut Point

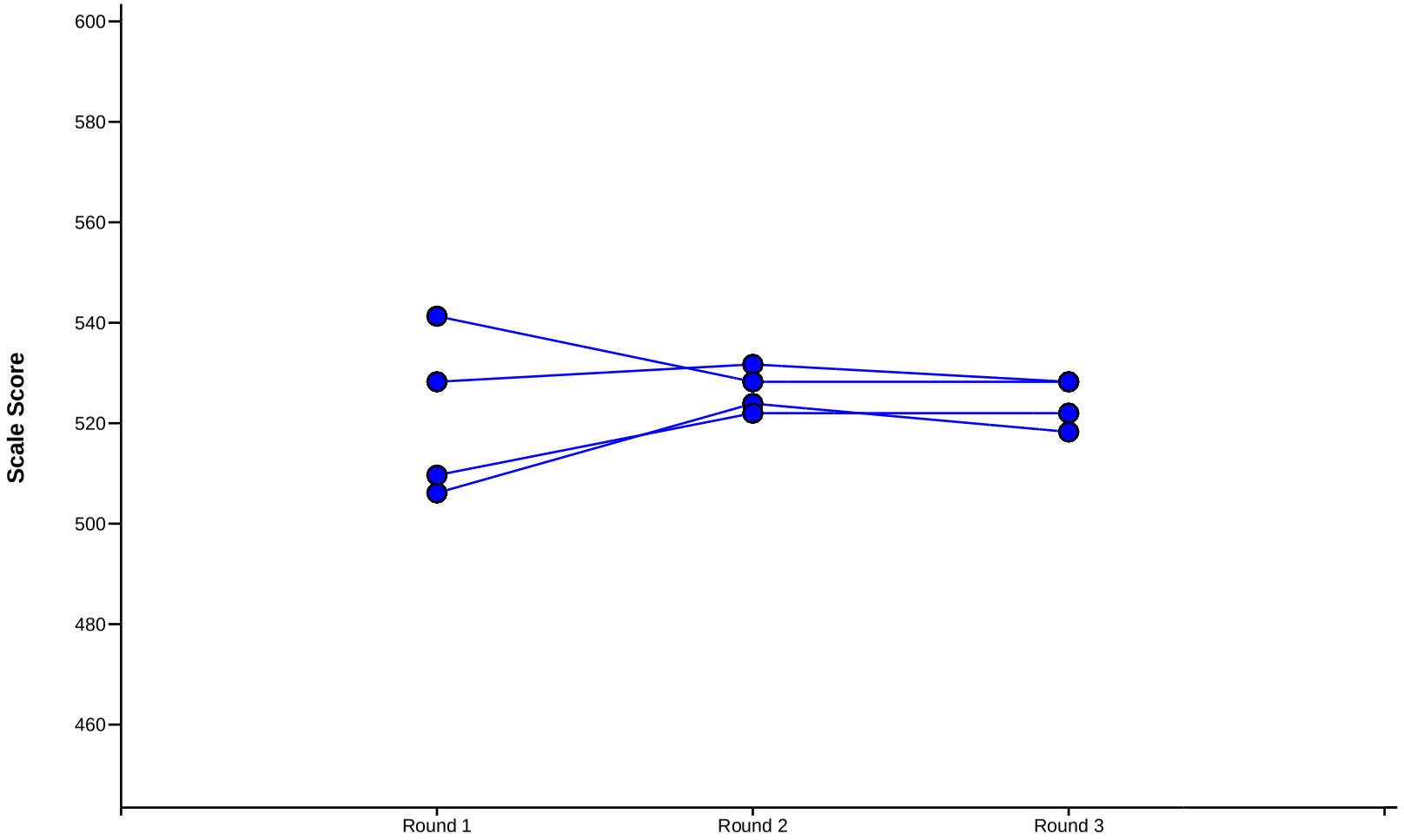


Table 1

AIMS Standard Setting Grade 10 Science Meets Cut Point

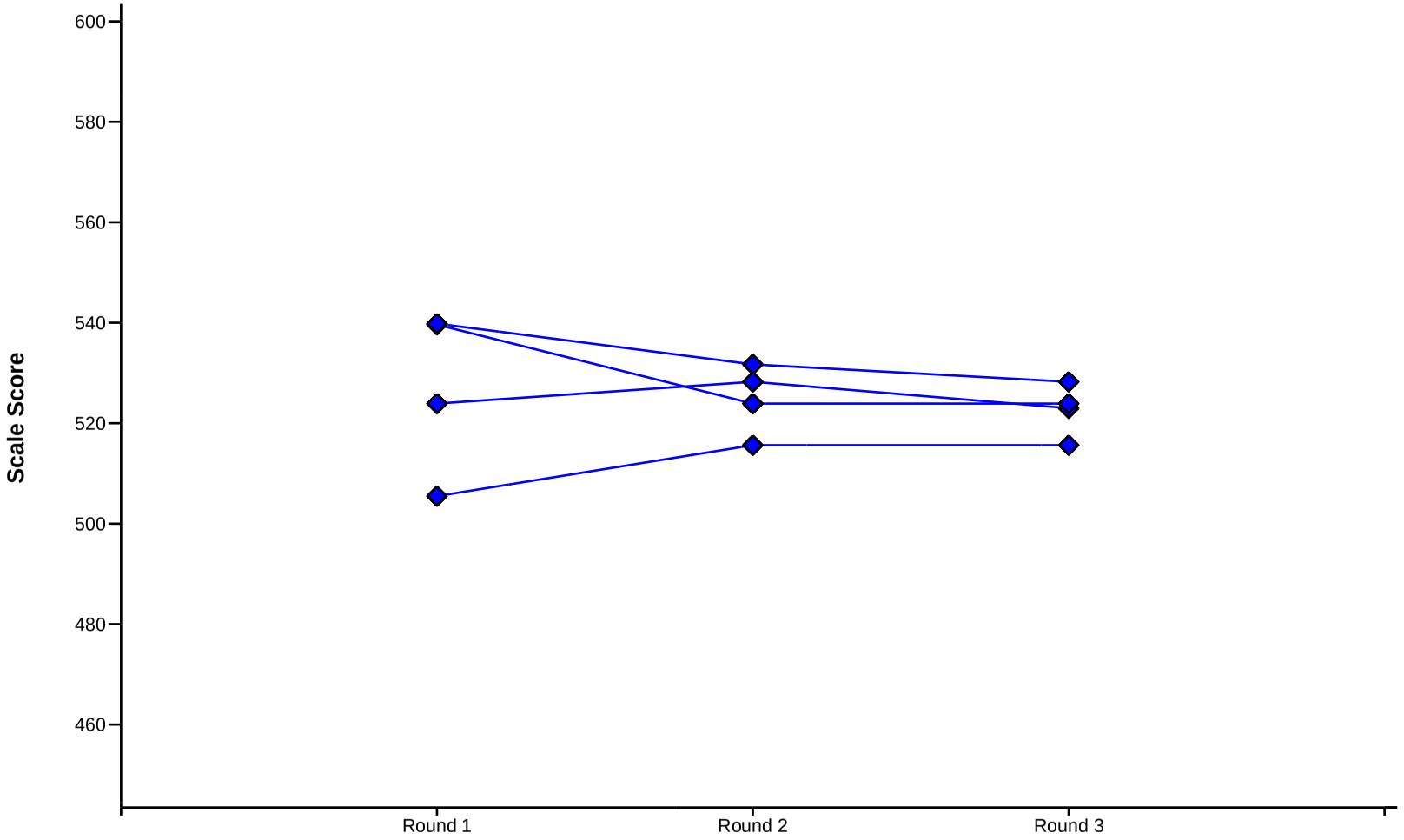


Table 2

H 32

AIMS Standard Setting Grade 10 Science Meets Cut Point

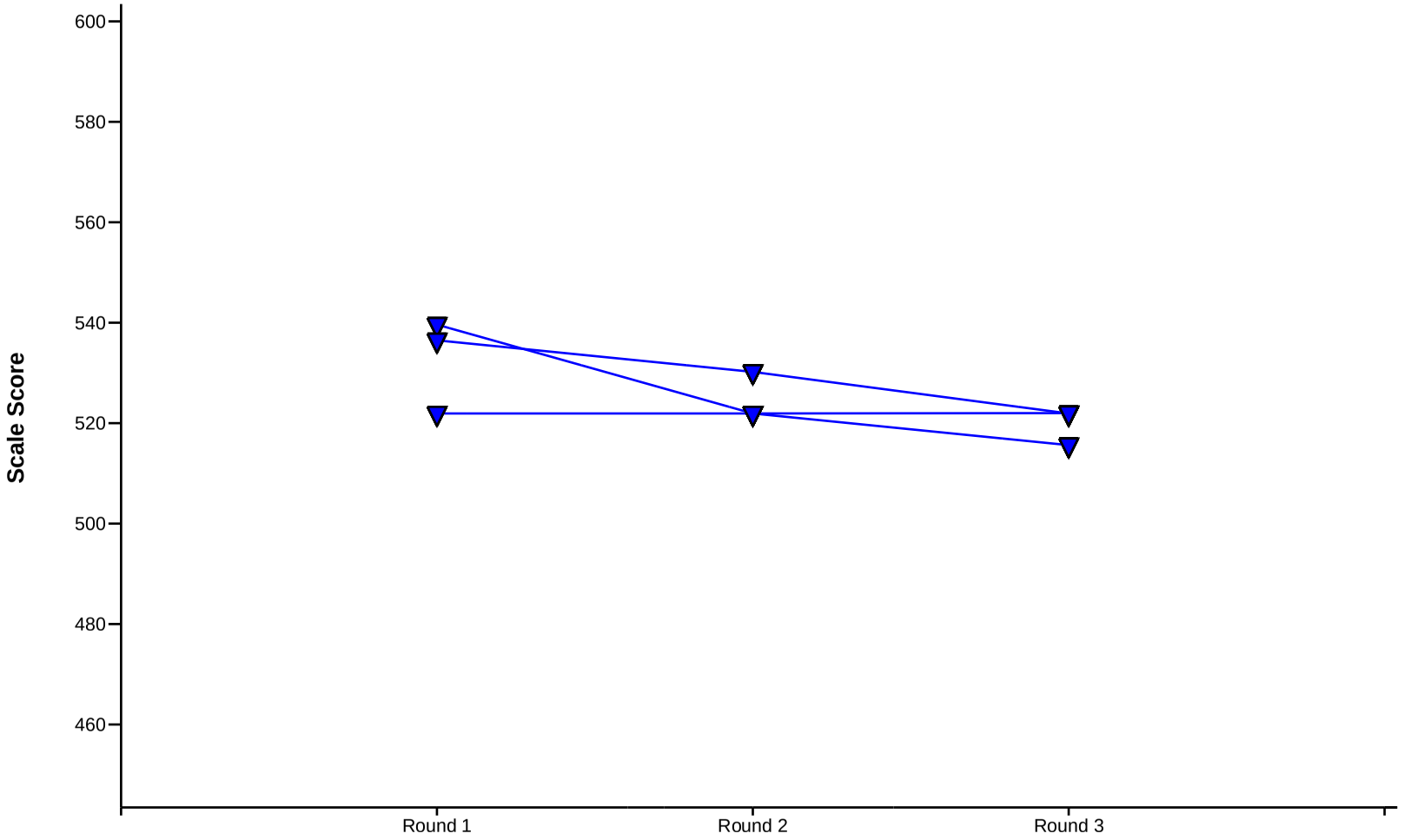
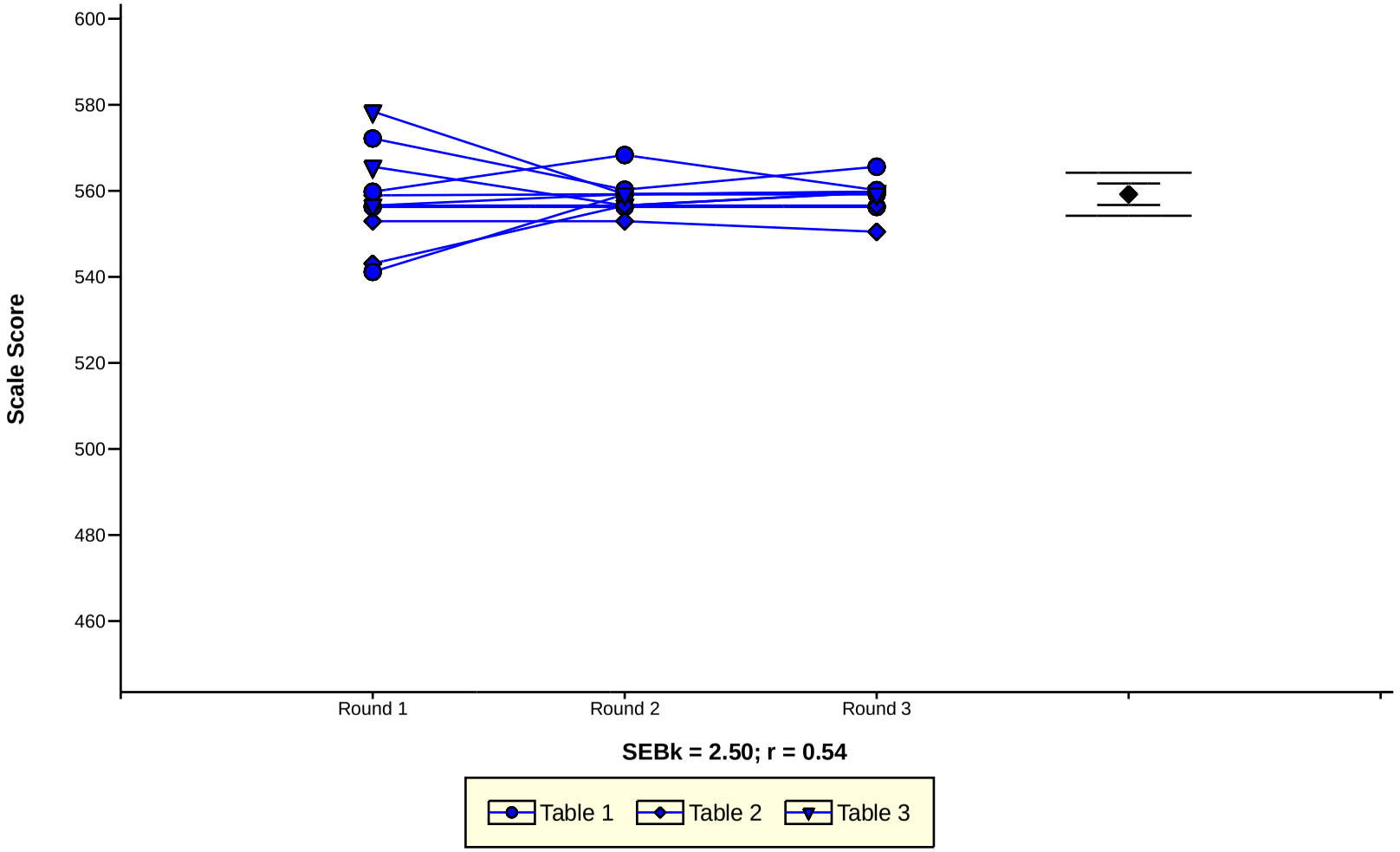


Table 3

H 33

AIMS Standard Setting Grade 10 Science Exceeds Cut Point



H 34

AIMS Standard Setting Grade 10 Science Exceeds Cut Point

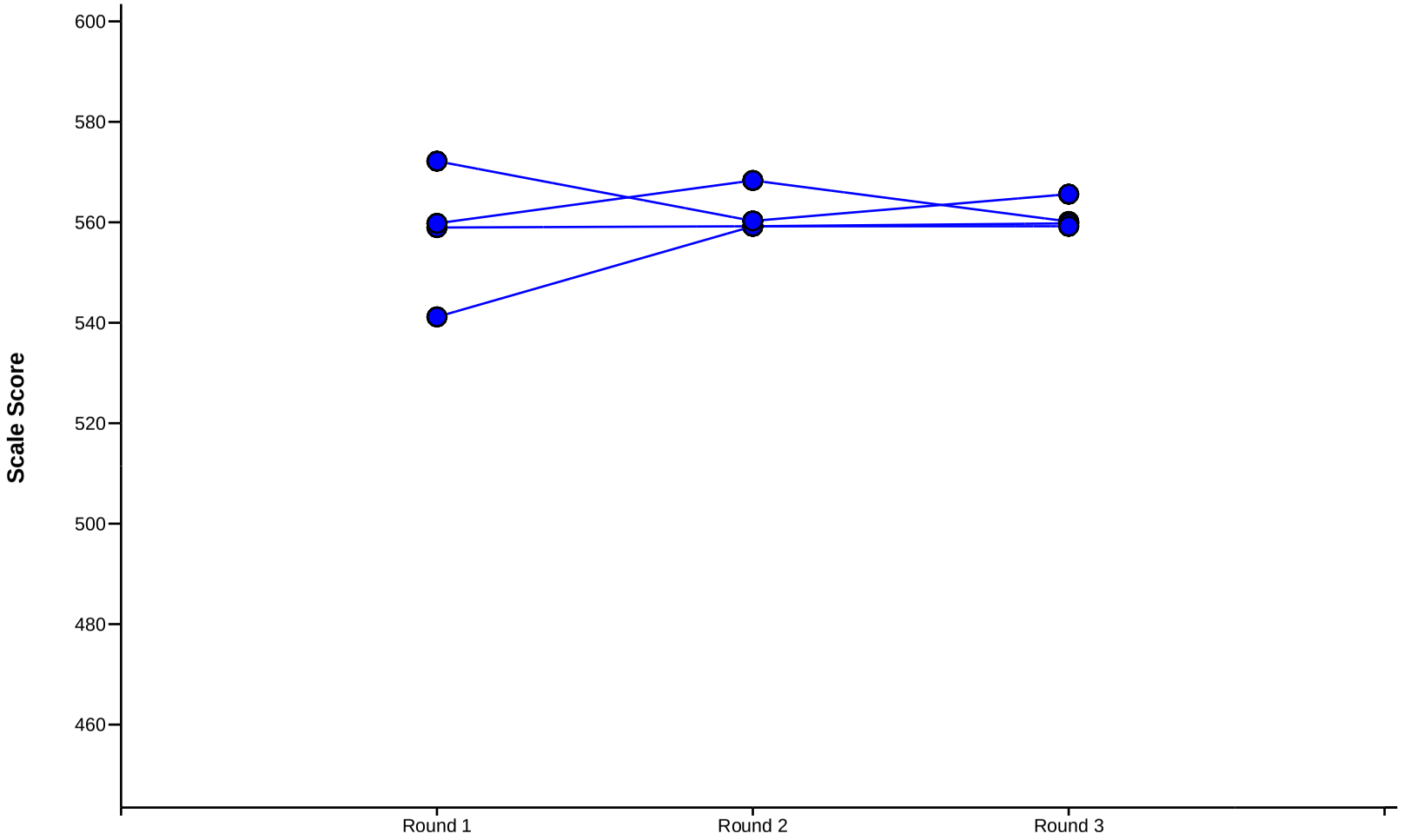


Table 1

H 35

AIMS Standard Setting Grade 10 Science Exceeds Cut Point

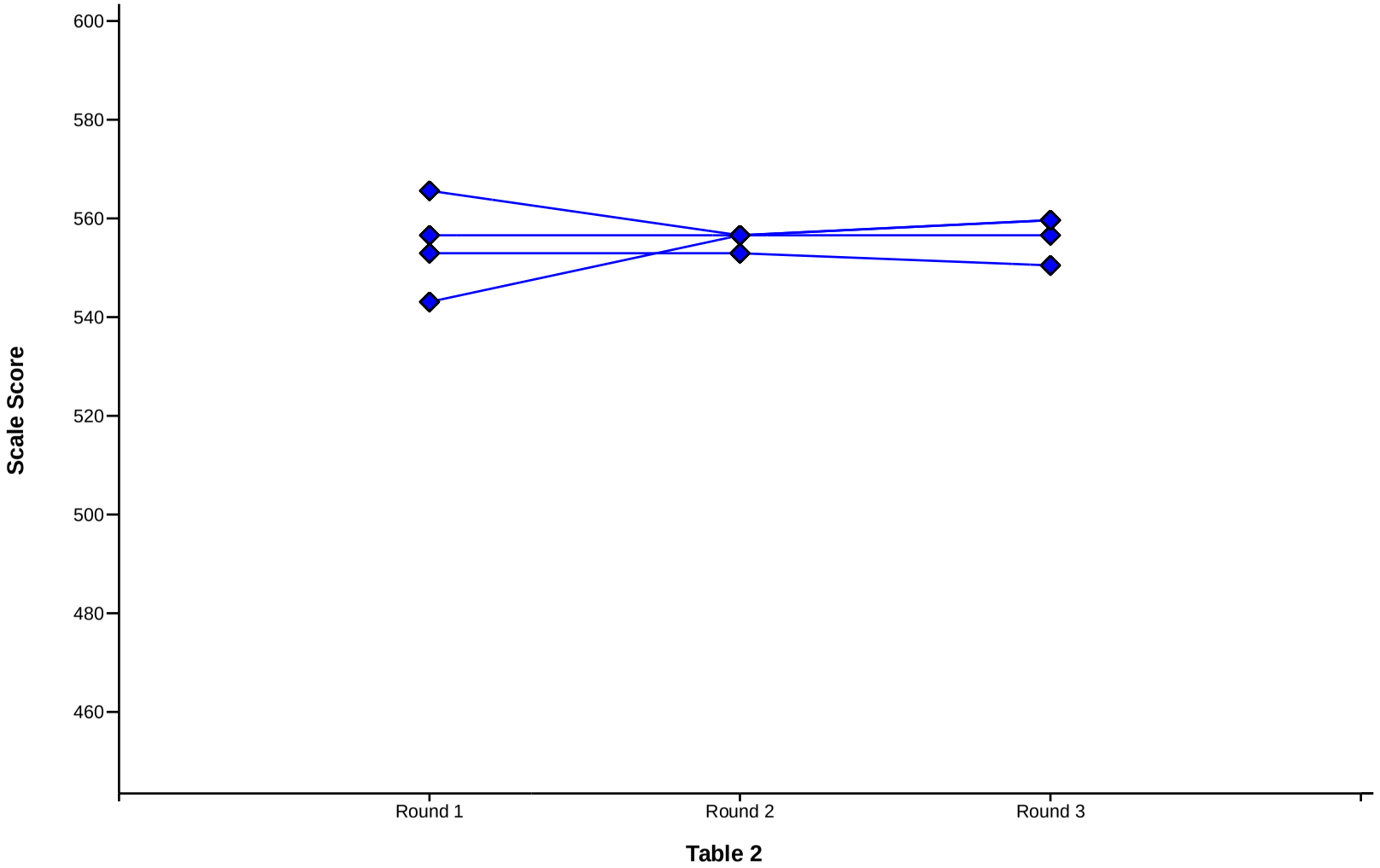
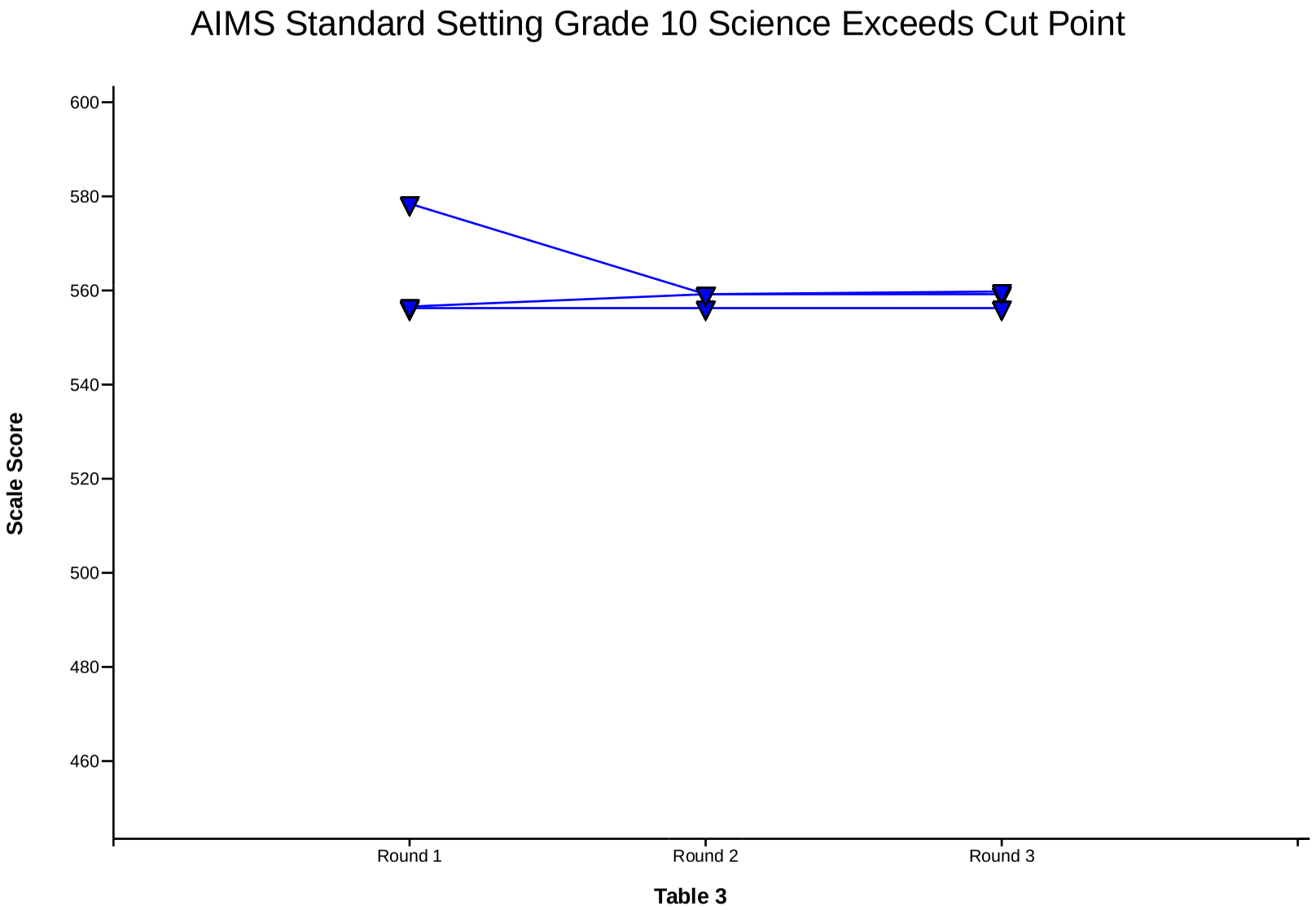


Table 2


AIMS Standard Setting Grade 10 Science Exceeds Cut Point
Scale Score
600
580
560
540
520
500
480
460
Round 1
Round 2
Round 3
Table 3

SECTION I

Participant Evaluation

Arizona's Instrument to Measure Standards
Bookmark Standard Setting
Evaluation Results

About these results

Each question is shown, along with its answer choices and associated response percentages. For Likert-type questions, there are five possible responses: "Strongly Disagree," "Disagree," "Neutral," "Agree," and "Strongly Agree." For each question, the number of respondents is shown in the column labeled "N."

PART I: ABOUT THE CONFERENCE

Question 1

The Bookmark Standard Setting Procedure was well described.

Content Area	Grade	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Agree + Strongly Agree
Overall		35	0.0%	0.0%	2.9%	22.9%	74.3%	97.2%
Science	4	12	0.0%	0.0%	0.0%	41.7%	58.3%	100.0%
	8	12	0.0%	0.0%	0.0%	8.3%	91.7%	100.0%
	HS	11	0.0%	0.0%	9.1%	18.2%	72.7%	90.9%

Question 2

The goals of this procedure were clear.

Content Area	Grade	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Agree + Strongly Agree
Overall		35	0.0%	0.0%	0.0%	37.1%	62.9%	100.0%
Science	4	12	0.0%	0.0%	0.0%	33.3%	66.7%	100.0%
	8	12	0.0%	0.0%	0.0%	16.7%	83.3%	100.0%
	HS	11	0.0%	0.0%	0.0%	63.6%	36.4%	100.0%

Question 3
I felt that this procedure was fair.

Content Area	Grade	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Agree + Strongly Agree
Overall		35	0.0%	0.0%	0.0%	34.3%	65.7%	100.0%
Science	4	12	0.0%	0.0%	0.0%	50.0%	50.0%	100.0%
	8	12	0.0%	0.0%	0.0%	16.7%	83.3%	100.0%
	HS	11	0.0%	0.0%	0.0%	36.4%	63.6%	100.0%

Question 4
Participating in the Bookmark Standard Setting increased my understanding of the test.

Content Area	Grade	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Agree + Strongly Agree
Overall		35	0.0%	0.0%	2.9%	5.7%	91.4%	97.1%
Science	4	12	0.0%	0.0%	0.0%	8.3%	91.7%	100.0%
	8	12	0.0%	0.0%	0.0%	0.0%	100.0%	100.0%
	HS	11	0.0%	0.0%	9.1%	9.1%	81.8%	90.9%

Question 5
The workshop was well organized.

Content Area	Grade	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Agree + Strongly Agree
Overall		35	0.0%	0.0%	2.9%	14.3%	82.9%	97.2%
Science	4	12	0.0%	0.0%	0.0%	8.3%	91.7%	100.0%
	8	12	0.0%	0.0%	0.0%	8.3%	91.7%	100.0%
	HS	11	0.0%	0.0%	9.1%	27.3%	63.6%	90.9%

Question 6
The training materials were helpful.

Content Area	Grade	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Agree + Strongly Agree
Overall		35	0.0%	0.0%	0.0%	34.3%	65.7%	100.0%
Science	4	12	0.0%	0.0%	0.0%	50.0%	50.0%	100.0%
	8	12	0.0%	0.0%	0.0%	16.7%	83.3%	100.0%
	HS	11	0.0%	0.0%	0.0%	36.4%	63.6%	100.0%

Question 7
The training on bookmark placement made the task clear to me.

Content Area	Grade	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Agree + Strongly Agree
Overall		35	0.0%	0.0%	2.9%	37.1%	60.0%	97.1%
Science	4	12	0.0%	0.0%	0.0%	58.3%	41.7%	100.0%
	8	12	0.0%	0.0%	0.0%	25.0%	75.0%	100.0%
	HS	11	0.0%	0.0%	9.1%	27.3%	63.6%	90.9%

Question 8
Taking the test helped me place my bookmark.

Content Area	Grade	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Agree + Strongly Agree
Overall		35	0.0%	2.9%	14.3%	40.0%	42.9%	82.9%
Science	4	12	0.0%	0.0%	25.0%	41.7%	33.3%	75.0%
	8	12	0.0%	0.0%	8.3%	50.0%	41.7%	91.7%
	HS	11	0.0%	9.1%	9.1%	27.3%	54.5%	81.8%

Question 9
During Round 1, I placed my bookmark without consulting other participants.

Content Area	Grade	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Agree + Strongly Agree
Overall		35	0.0%	0.0%	0.0%	8.6%	91.4%	100.0%
Science	4	12	0.0%	0.0%	0.0%	8.3%	91.7%	100.0%
	8	12	0.0%	0.0%	0.0%	8.3%	91.7%	100.0%
	HS	11	0.0%	0.0%	0.0%	9.1%	90.9%	100.0%

Question 10
I considered the Arizona Content Standards when I placed my bookmark.

Content Area	Grade	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Agree + Strongly Agree
Overall		35	0.0%	0.0%	2.9%	25.7%	71.4%	97.1%
Science	4	12	0.0%	0.0%	8.3%	25.0%	66.7%	91.7%
	8	12	0.0%	0.0%	0.0%	25.0%	75.0%	100.0%
	HS	11	0.0%	0.0%	0.0%	27.3%	72.7%	100.0%

Question 11
I understood how to place my bookmark.

Content Area	Grade	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Agree + Strongly Agree
Overall		35	0.0%	0.0%	0.0%	28.6%	71.4%	100.0%
Science	4	12	0.0%	0.0%	0.0%	50.0%	50.0%	100.0%
	8	12	0.0%	0.0%	0.0%	8.3%	91.7%	100.0%
	HS	11	0.0%	0.0%	0.0%	27.3%	72.7%	100.0%

Question 12
I had enough time to consider my Round 1 bookmark.

Content Area	Grade	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Agree + Strongly Agree
Overall		35	0.0%	0.0%	0.0%	20.0%	80.0%	100.0%
Science	4	12	0.0%	0.0%	0.0%	16.7%	83.3%	100.0%
	8	12	0.0%	0.0%	0.0%	8.3%	91.7%	100.0%
	HS	11	0.0%	0.0%	0.0%	36.4%	63.6%	100.0%

Question 13
I understood how to do bookmark placement from the beginning, so my earlier bookmarks are comparable to my later bookmarks.

Content Area	Grade	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Agree + Strongly Agree
Overall		35	0.0%	2.9%	14.3%	34.3%	48.6%	82.9%
Science	4	12	0.0%	8.3%	25.0%	33.3%	33.3%	66.6%
	8	12	0.0%	0.0%	0.0%	25.0%	75.0%	100.0%
	HS	11	0.0%	0.0%	18.2%	45.5%	36.4%	81.9%

Question 14
Overall, I was satisfied with my group's final bookmarks.

Content Area	Grade	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Agree + Strongly Agree
Overall		35	0.0%	0.0%	0.0%	31.4%	68.6%	100.0%
Science	4	12	0.0%	0.0%	0.0%	41.7%	58.3%	100.0%
	8	12	0.0%	0.0%	0.0%	16.7%	83.3%	100.0%
	HS	11	0.0%	0.0%	0.0%	36.4%	63.6%	100.0%

Question 15
I would defend the *Exceeds* cut scores against criticism that they are too high.

Content Area	Grade	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Agree + Strongly Agree
Overall		35	2.9%	5.7%	0.0%	25.7%	65.7%	91.4%
Science	4	12	0.0%	8.3%	0.0%	33.3%	58.3%	91.6%
	8	12	8.3%	8.3%	0.0%	8.3%	75.0%	83.3%
	HS	11	0.0%	0.0%	0.0%	36.4%	63.6%	100.0%

Question 16
I would defend the *Exceeds* cut scores against criticism that they are too low.

Content Area	Grade	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Agree + Strongly Agree
Overall		35	5.7%	2.9%	0.0%	34.3%	57.1%	91.4%
Science	4	12	0.0%	8.3%	0.0%	33.3%	58.3%	91.6%
	8	12	8.3%	0.0%	0.0%	33.3%	58.3%	91.6%
	HS	11	9.1%	0.0%	0.0%	36.4%	54.5%	90.9%

Question 17
I would defend the *Approaches* cut scores against criticism that they are too high.

Content Area	Grade	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Agree + Strongly Agree
Overall		35	2.9%	0.0%	2.9%	37.1%	57.1%	94.2%
Science	4	12	0.0%	0.0%	0.0%	58.3%	41.7%	100.0%
	8	12	8.3%	0.0%	0.0%	16.7%	75.0%	91.7%
	HS	11	0.0%	0.0%	9.1%	36.4%	54.5%	90.9%

Question 18
I would defend the *Approaches* cut scores against criticism that they are too low.

Content Area	Grade	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Agree + Strongly Agree
Overall		35	5.7%	5.7%	0.0%	31.4%	57.1%	88.5%
Science	4	12	0.0%	16.7%	0.0%	41.7%	41.7%	83.4%
	8	12	8.3%	0.0%	0.0%	16.7%	75.0%	91.7%
	HS	11	9.1%	0.0%	0.0%	36.4%	54.5%	90.9%

Question 19
I would defend the *Meets* cut scores against criticism that they are too high.

Content Area	Grade	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Agree + Strongly Agree
Overall		35	2.9%	2.9%	2.9%	31.4%	60.0%	91.4%
Science	4	12	0.0%	8.3%	0.0%	41.7%	50.0%	91.7%
	8	12	8.3%	0.0%	0.0%	16.7%	75.0%	91.7%
	HS	11	0.0%	0.0%	9.1%	36.4%	54.5%	90.9%

Question 20
I would defend the *Meets* cut scores against criticism that they are too low.

Content Area	Grade	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Agree + Strongly Agree
Overall		35	5.7%	2.9%	0.0%	31.4%	60.0%	91.4%
Science	4	12	0.0%	8.3%	0.0%	41.7%	50.0%	91.7%
	8	12	8.3%	0.0%	0.0%	16.7%	75.0%	91.7%
	HS	11	9.1%	0.0%	0.0%	36.4%	54.5%	90.9%

Question 21
Overall, I believe my opinions were considered and valued by my group.

Content Area	Grade	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Agree + Strongly Agree
Overall		35	0.0%	0.0%	2.9%	11.4%	85.7%	97.1%
Science	4	12	0.0%	0.0%	0.0%	8.3%	91.7%	100.0%
	8	12	0.0%	0.0%	0.0%	8.3%	91.7%	100.0%
	HS	11	0.0%	0.0%	9.1%	18.2%	72.7%	90.9%

Question 22
I am confident that the Bookmark Procedure produced valid standards.

Content Area	Grade	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Agree + Strongly Agree
Overall		35	2.9%	0.0%	0.0%	22.9%	74.3%	97.2%
Science	4	12	0.0%	0.0%	0.0%	41.7%	58.3%	100.0%
	8	12	0.0%	0.0%	0.0%	8.3%	91.7%	100.0%
	HS	11	9.1%	0.0%	0.0%	18.2%	72.7%	90.9%

Question 23
The ordering of the items in the order item booklet agreed with my perception of the relative difficulty of the items.

Content Area	Grade	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Agree + Strongly Agree
Overall		35	2.9%	14.3%	14.3%	40.0%	28.6%	68.6%
Science	4	12	0.0%	8.3%	8.3%	58.3%	25.0%	83.3%
	8	12	0.0%	16.7%	8.3%	41.7%	33.3%	75.0%
	HS	11	9.1%	18.2%	27.3%	18.2%	27.3%	45.5%

Question 24
Overall, my table's discussions were open and honest.

Content Area	Grade	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Agree + Strongly Agree
Overall		35	0.0%	0.0%	0.0%	17.1%	82.9%	100.0%
Science	4	12	0.0%	0.0%	0.0%	25.0%	75.0%	100.0%
	8	12	0.0%	0.0%	0.0%	8.3%	91.7%	100.0%
	HS	11	0.0%	0.0%	0.0%	18.2%	81.8%	100.0%

Question 25
The presentation of impact data was helpful to me.

Content Area	Grade	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Agree + Strongly Agree
Overall		35	0.0%	0.0%	5.7%	31.4%	62.9%	94.3%
Science	4	12	0.0%	0.0%	0.0%	50.0%	50.0%	100.0%
	8	12	0.0%	0.0%	0.0%	16.7%	83.3%	100.0%
	HS	11	0.0%	0.0%	18.2%	27.3%	54.5%	81.8%

Question 26
You participated in making final recommendations for cut scores in the cross-grade discussion.

Did you understand the purpose for considering adjusting cut scores?
Please explain your response.

35 out of 35 respondents answered "yes." Selected explanations:

- *"After listening to everyone, I adjusted my cut scores based on the content."*
- *"Important to realize how the standard strands developed across grade levels, and what expectations were for cross-grade levels."*
- *"The purpose of the discussion was to hear points of view and evaluate them, then make necessary adjustments."*

What comments do you have regarding rationales that other participants gave for adjusting or not adjusting cut scores?

Selected comments:

- *"I thought the group grade level discussion (all grade levels) was very focused on the cut scores and the overall feeling was that they built on each other."*
- *"Difference rationales helped me confirm or adjust my thinking, but did not cause me to change my mind based on my beliefs of what is important to teach and to always focus on the student and learning rather than the teacher and teaching."*
- *"Comments were very enlightening because participants brought their experiences and different perspectives. This enabled one to gain a broader view."*

Do you think that the discussion addressed all considerations adequately (e.g., placement of the bookmarks, rationales for adjusting or not adjusting cut scores, impact data)? Please explain your response.

31 out of 33 respondents answered "yes." Selected explanations:

- *"Yes, because the groups were confident on their placement and articulated what concerns they did have resulting in a common conclusion of the cut scores."*
- *"Yes, discussion open. Everyone had the option to express an opinion. Discussion was more than adequate with no placed limitations."*
- *"Mostly. Sometimes discussion got into pedagogy and supplies and support available for teaching. I'm frustrated when we use the test data to*

judge ourselves as teachers and not our students as learners and a combination of both."

Question 27
You participated in making final recommendations for cut scores in the cross-grade discussion.

Are you generally satisfied with the final recommendations for the cut scores?

Content Area	Performance Level	Grade	N	Yes	No
Overall					
Science	**Exceeds**	4	35	97.1%	2.9%
		8	12	91.7%	8.3%
		HS	12	100.0%	0.0%
	Meets	4	35	100.0%	0.0%
		8	12	100.0%	0.0%
		HS	12	100.0%	0.0%
	Approaches	4	35	100.0%	0.0%
		8	12	100.0%	0.0%
		HS	12	100.0%	0.0%
	Falls Far Below	4	35	100.0%	0.0%
		8	12	100.0%	0.0%
		HS	12	100.0%	0.0%

Question 27 (cont.)
If you are not satisfied, in which direction would you move the placement of a cut score and by how much?

Exceeds the Standard	_____ I would leave it where it is	I would move it _____ pages before the final page I would move it _____ pages after the final page
Meets the Standard	_____ I would leave it where it is	I would move it _____ pages before the final page I would move it _____ pages after the final page
Approaches the Standard	_____ I would leave it where it is	I would move it _____ pages before the final page I would move it _____ pages after the final page
Falls Far Below the Standard	_____ I would leave it where it is	I would move it _____ pages before the final page I would move it _____ pages after the final page

12 participants responded to this question, and one additional participant left a comment.

Exceeds the Standard:

- *11 out of 12 would leave it where it is.*
- *1 out of 12 would move it before the final page (3 pages).*
- *0 out of 12 would move it after the final page.*
- *One additional participant would move it "to Page 74."*

Meets the Standard:

- *12 out of 12 would leave it where it is.*

Approaches the Standard:

- *12 out of 12 would leave it where it is.*

Falls Far Below the Standard:

- *12 out of 12 would leave it where it is.*

Question 28
The training on performance level descriptors (PLDs) made the task clear to me.

Content Area	Grade	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Agree + Strongly Agree
Overall		35	0.0%	0.0%	5.7%	51.4%	42.9%	94.3%
Science	4	12	0.0%	0.0%	0.0%	50.0%	50.0%	100.0%
	8	12	0.0%	0.0%	0.0%	50.0%	50.0%	100.0%
	HS	11	0.0%	0.0%	18.2%	54.5%	27.3%	81.8%

Question 29
Examining the test items helped me to draft the PLDs.

Content Area	Grade	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Agree + Strongly Agree
Overall		35	0.0%	2.9%	8.6%	40.0%	48.6%	88.6%
Science	4	12	0.0%	0.0%	0.0%	50.0%	50.0%	100.0%
	8	12	0.0%	0.0%	8.3%	33.3%	58.3%	91.6%
	HS	11	0.0%	9.1%	18.2%	36.4%	36.4%	72.8%

Question 30
I considered the content standards when drafting the PLDs.

Content Area	Grade	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Agree + Strongly Agree
Overall		35	0.0%	0.0%	2.9%	31.4%	65.7%	97.1%
Science	4	12	0.0%	0.0%	0.0%	25.0%	75.0%	100.0%
	8	12	0.0%	0.0%	8.3%	25.0%	66.7%	91.7%
	HS	11	0.0%	0.0%	0.0%	45.5%	54.5%	100.0%

Question 31
I considered the cognitive rigor of items when drafting the PLDs.

Content Area	Grade	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Agree + Strongly Agree
Overall		35	0.0%	0.0%	8.6%	31.4%	60.0%	91.4%
Science	4	12	0.0%	0.0%	0.0%	25.0%	75.0%	100.0%
	8	12	0.0%	0.0%	16.7%	25.0%	58.3%	83.3%
	HS	11	0.0%	0.0%	9.1%	45.5%	45.5%	91.0%

Question 32
Overall, I valued the workshop as a professional development experience.

Content Area	Grade	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Agree + Strongly Agree
Overall		35	0.0%	0.0%	5.7%	11.4%	82.9%	94.3%
Science	4	12	0.0%	0.0%	0.0%	25.0%	75.0%	100.0%
	8	12	0.0%	0.0%	0.0%	8.3%	91.7%	100.0%
	HS	11	0.0%	0.0%	18.2%	0.0%	81.8%	81.8%

Question 33
This experience will help me target instruction for the students in my classroom.

Content Area	Grade	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Agree + Strongly Agree
Overall		34	0.0%	0.0%	2.9%	8.8%	88.2%	97.0%
Science	4	12	0.0%	0.0%	8.3%	0.0%	91.7%	91.7%
	8	12	0.0%	0.0%	0.0%	16.7%	83.3%	100.0%
	HS	10	0.0%	0.0%	0.0%	10.0%	90.0%	100.0%

PART II: ABOUT YOU

Question 34
What is your occupation?

Content Area	Grade	N	Teacher	Education, Non-Teacher	Other, Non-Education
Overall		35	82.9%	17.1%	0.0%
Science	4	12	75.0%	25.0%	0.0%
	8	12	91.7%	8.3%	0.0%
	HS	11	81.8%	18.2%	0.0%

Question 35
How many years in your current profession?

Content Area	Grade	N	1-5	6-10	11-15	16-20	21+
Overall		35	5.7%	8.6%	17.1%	20.0%	48.6%
Science	4	12	0.0%	16.7%	16.7%	16.7%	50.0%
	8	12	8.3%	8.3%	16.7%	8.3%	58.3%
	HS	11	9.1%	0.0%	18.2%	36.4%	36.4%

Question 36
What is your highest level of education?

Content Area	Grade	N	High School	Bachelor's	Master's	Doctorate
Overall		35	0.0%	17.1%	80.0%	2.9%
Science	4	12	0.0%	25.0%	75.0%	0.0%
	8	12	0.0%	25.0%	75.0%	0.0%
	HS	11	0.0%	0.0%	90.9%	9.1%

Question 37
What is your race/ethnicity?

Content Area	Grade	N	American Indian	Asian/ Pacific Islander	Black/ African-American	Hispanic	White	Other
Overall		35	0.0%	0.0%	0.0%	8.6%	80.0%	11.4%
Science	4	12	0.0%	0.0%	0.0%	8.3%	91.7%	0.0%
	8	12	0.0%	0.0%	0.0%	16.7%	75.0%	8.3%
	HS	11	0.0%	0.0%	0.0%	0.0%	72.7%	27.3%

Question 38
What is your gender?

Content Area	Grade	N	Male	Female
Overall		34	17.6%	82.4%
Science	4	12	8.3%	91.7%
	8	12	16.7%	83.3%
	HS	10	30.0%	70.0%

Question 39
Have you taught Special Education?

Content Area	Grade	N	Yes	No
Overall		35	20.0%	80.0%
Science	4	12	25.0%	75.0%
	8	12	8.3%	91.7%
	HS	11	27.3%	72.7%

Question 40
Have you taught ELL/ESL?

Content Area	Grade	N	Yes	No
Overall		35	48.6%	51.4%
Science	4	12	41.7%	58.3%
	8	12	41.7%	58.3%
	HS	11	63.6%	36.4%

Question 41
Have you taught Vocational Education?

Content Area	Grade	N	Yes	No
Overall		33	9.1%	90.9%
Science	4	12	0.0%	100.0%
	8	11	0.0%	100.0%
	HS	10	30.0%	70.0%

Question 42
Have you taught Alternative Education?

Content Area	Grade	N	Yes	No
Overall		32	18.8%	81.3%
Science	4	11	0.0%	100.0%
	8	11	9.1%	90.9%
	HS	10	50.0%	50.0%

Question 43
Have you taught Adult Education?

Content Area	Grade	N	Yes	No
Overall		35	45.7%	54.3%
Science	4	12	50.0%	50.0%
	8	12	33.3%	66.7%
	HS	11	54.5%	45.5%

Question 44
Which grade did you work on during the standard setting?

Content Area	Grade	N	Overall
Overall		35	100.0%
Science	4	12	34.3%
	8	12	34.3%
	HS	11	31.4%

Arizona AIMS Science
Bookmark Standard Setting 2008

PART I: ABOUT THE CONFERENCE Please consider the statements below and fill in the bubble for the level of agreement or disagreement you have with each statement. A 5-point rating scale ranging from Strongly Disagree to Strongly Agree is provided. Please bubble only 1 of the 5 options for each statement.	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
1. The Bookmark Standard Setting Procedure was well described.	O	O	O	O	O
2. The goals of this procedure were clear.	O	O	O	O	O
3. I felt that this procedure was fair.	O	O	O	O	O
4. Participating in the Bookmark Standard Setting increased my understanding of the test.	O	O	O	O	O
5. The workshop was well organized.	O	O	O	O	O
6. The training materials were helpful.	O	O	O	O	O
7. The training on bookmark placement made the task clear to me.	O	O	O	O	O
8. Taking the test helped me place my bookmark.	O	O	O	O	O
9. During Round 1, I placed my bookmark without consulting other participants.	O	O	O	O	O
10. I considered the Arizona Content Standards when I placed my bookmark.	O	O	O	O	O
11. I understood how to place my bookmark.	O	O	O	O	O
12. I had enough time to consider my Round 1 bookmark.	O	O	O	O	O
13. I understood how to do bookmark placement from the beginning, so my earlier bookmarks are comparable to my later bookmarks.	O	O	O	O	O
14. Overall, I was satisfied with my group's final bookmarks.	O	O	O	O	O
15. I would defend the *Exceeds* cut scores against criticism that they are too high.	O	O	O	O	O
16. I would defend the *Exceeds* cut scores against criticism that they are too low.	O	O	O	O	O
17. I would defend the *Approaches* cut scores against criticism that they are too high.	O	O	O	O	O
18. I would defend the *Approaches* cut scores against criticism that they are too low.	O	O	O	O	O
19. I would defend the *Meets* cut scores against criticism that they are too high.	O	O	O	O	O
20. I would defend the *Meets* cut scores against criticism that they are too low.	O	O	O	O	O
21. Overall, I believe my opinions were considered and valued by my group.	O	O	O	O	O
22. I am confident that the Bookmark Procedure produced valid standards.	O	O	O	O	O
23. The ordering of the items in the order item booklet agreed with my perception of the relative difficulty of the items.	O	O	O	O	O
24. Overall, my table's discussions were open and honest.	O	O	O	O	O
25. The presentation of impact data was helpful to me.	O	O	O	O	O

26. You participated in making final recommendations for cut scores in the cross-grade discussion.

Did you understand the purpose for considering adjusting cut scores? Please explain your response.

What comments do you have regarding rationales that other participants gave for adjusting or not adjusting cut scores?

Do you think that the discussion addressed all considerations adequately (e.g., placement of the bookmarks, rationales for adjusting or not adjusting cut scores, impact data)? Please explain your response.

27. You participated in making final recommendations for cut scores in the cross-grade discussion.

Are you generally satisfied with the final recommendations for the cut scores?

	Yes	No
Exceeds the Standard	**O**	**O**
Meets the Standard	**O**	**O**
Approaches the Standard	**O**	**O**
Falls Far Below the Standard	**O**	**O**

If you are not satisfied, in which direction would you move the placement of a cut score and by how much?

Exceeds the Standard	_____ I would leave it where it is	I would move it _____ pages before the final page I would move it _____ pages after the final page
Meets the Standard	_____ I would leave it where it is	I would move it _____ pages before the final page I would move it _____ pages after the final page
Approaches the Standard	_____ I would leave it where it is	I would move it _____ pages before the final page I would move it _____ pages after the final page
Falls Far Below the Standard	_____ I would leave it where it is	I would move it _____ pages before the final page I would move it _____ pages after the final page

PART I: ABOUT THE CONFERENCE (cont'd)

Please consider the statements below and fill in the bubble for the level of agreement or disagreement you have with each statement.

A 5-point rating scale ranging from Strongly Disagree to Strongly Agree is provided.
Please bubble only 1 of the 5 options for each statement.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
28. The training on performance level descriptors (PLDs) made the task clear to me.	**O**	**O**	**O**	**O**	**O**
29. Examining the test items helped me to draft the PLDs.	**O**	**O**	**O**	**O**	**O**
30. I considered the content standards when drafting the PLDs.	**O**	**O**	**O**	**O**	**O**
31. I considered the cognitive rigor of items when drafting the PLDs.	**O**	**O**	**O**	**O**	**O**
32. Overall, I valued the workshop as a professional development experience.	**O**	**O**	**O**	**O**	**O**
33. This experience will help me target instruction for the students in my classroom.	**O**	**O**	**O**	**O**	**O**

PART II: ABOUT YOU

Please tell us about yourself. This information will be used for classification purposes and allows us to better understand the Bookmark Standard Setting Procedure. Please bubble only 1 for each question.

34. What is your occupation?

O Teacher
O Education, Non-Teacher
O Other, Non-Education:
_____

35. How many years in your current profession?

O 1–5
O 6–10
O 11–15
O 16–20
O 21+

36. What is your highest level of education?

O High School
O Bachelor's
O Master's
O Doctorate

37. What is your race/ethnicity?

O American Indian
O Asian/Pacific Islander
O African American
O Hispanic
O White
O Other (please specify below)
_____

38. What is your gender?

O Male
O Female

39. Have you taught Special Education?

O Yes
O No

40. Have you taught ELL/ESL?

O Yes
O No

41. Have you taught Vocational Education?

O Yes
O No

42. Have you taught Alternative Education?

O Yes
O No

43. Have you taught Adult Education?

O Yes
O No

44. Which grade did you work on during the standard setting?

O 4
O 8
O HS

PART III: YOUR TURN

Please feel free to add comments on any of your responses above, make suggestions to improve future standard settings, and/or tell us what you liked and did not like about this workshop on the back of this evaluation. Thank you!

SECTION J

Performance Level Descriptors

Arizona Science Standard Performance Level Descriptors
Grade 4

Exceeds the Standard – Students who score in this level illustrate a superior academic performance as evidenced by achievement that is substantially beyond the goal for all students. Students who perform at this level demonstrate a wealth of knowledge, skills, and abilities in fulfillment of the science standard. They can plan simple investigations identifying variables, analyze data to determine trends, formulate conclusions based upon the data, and explain the role of experimentation in scientific inquiry.

Meets the Standard – Students who score in this level demonstrate a solid academic performance on subject matter as reflected by the science standard. Students who perform at this level are able to measure using appropriate tools, identify and compare structures in plants/animals noting their different functions in growth and survival, show that electricity flowing in circuits can produce light, heat, sound, and magnetic effects, measure changes in weather, and interpret the symbols on a weather map.

Approaches the Standard – Students who score in this level show partial understanding of the knowledge and application of the skills that are fundamental for proficient work. Students who perform at this level show some understanding of the science standard's concepts and procedures by demonstrating safe behavior and appropriate procedures in all science inquiry, classifying animals by their traits, investigating the characteristics of magnets, and identifying elements of the Earth's erosion. Some gaps in knowledge and skills are evident and may require additional instruction and remediation in order to achieve a satisfactory level of understanding.

Falls Far Below the Standard – Students who score in this level may have significant gaps and limited knowledge and skills that are necessary to satisfactorily meet the state's science standard. Students will usually require a considerable amount of additional instruction and remediation in order to achieve a satisfactory level of understanding.

Students at the "Exceeds the Standard" level generally know the skills required at the "Meets" and "Approaches" levels and are able to:	Students at the "Meets the Standard" level generally know the skills required at the "Approaches" level and are able to:	Students at the "Approaches the Standard" level generally know and are able to:
• Differentiate inferences from observations • Plan a simple investigation that identifies the variables to be controlled. • Analyze data obtained in a scientific investigation to identify trends. • Formulate conclusions based upon identified trends in data • Explain the role of experimentation in scientific inquiry. • Evaluate the consequences of environmental occurrences that happen either rapidly (e.g., fire, flood, tornado) or over a long period of time • Analyze the effect that limited resources (e.g., natural gas, minerals) may have on an environment. • Construct series and parallel electric circuits. • Compare rapid and slow processes that change the Earth's surface • Analyze evidence that indicates life and environmental conditions have changed • Differentiate between weather and climate as they relate to the southwestern United States	• Measure using appropriate tools (e.g., ruler, scale, balance) and units of measure (i.e., metric, U.S. customary). • Describe how natural events and human activities have positive and negative impacts on environments. • Describe how science and technology (e.g., computers, air conditioning, medicine) have improved the lives of many people. • Describe benefits (e.g., easy communications, rapid transportation) and risks (e.g., pollution, destruction of natural resources) related to the use of technology. • Compare structures in plants (e.g., roots, stems, leaves, flowers) and animals (e.g., muscles, bones, nerves) that serve different functions in growth and survival. • Describe ways various resources (e.g., air, water, plants, animals, soil) are utilized to meet the needs of a population. • Recognize that successful characteristics of populations are inherited traits that are favorable in a particular environment. • Demonstrate that electricity flowing in circuits can produce light, heat, sound, and magnetic effects. • Describe the role that water plays in the following processes that alter the Earth's surface features: • Measure changes in weather • Interpret the symbols on a weather map.	• Demonstrate safe behavior and appropriate procedures (e.g., use and care of technology, materials, organisms) in all science inquiry. • Classify animals by identifiable group characteristics. • Describe ways in which resources can be conserved • Investigate the characteristics of magnets • Identify the Earth processes that cause erosion. ***Separate process bullets from content bullets in each level.***

These descriptors do not include all the skills and knowledge as contained in the Science Standard.

Arizona Science Standard Performance Level Descriptors
Grade 8

DRAFT

Exceeds the Standard – Students who score in this level illustrate a superior academic performance as evidenced by achievement that is substantially beyond the goal for all students. Students who perform at this level demonstrate a wealth of knowledge, skills, and abilities in fulfillment of the science standard. They can formulate questions based upon observations, generate a hypothesis that can be tested, analyze data to identify trends, compose new questions based upon the results of a previous investigation, identify matter based upon its characteristics, and describe the intent of Newton's 2nd and 3rd Laws of Motion.

Meets the Standard – Students who score in this level demonstrate a solid academic performance on subject matter as reflected by the science standard. Students who perform at this level are able to interpret data to determine relationships between variables, identify potential investigation error, display data in an appropriate graphic, and write clear instructions for conducting investigations. They can distinguish between dominant and recessive traits in humans, classify matter as elements, compounds, or mixtures, and identify conditions under which an object will continue in its state of motion (Newton's 1st Law of Motion).

Approaches the Standard – Students who score in this level show partial understanding of the knowledge and application of the skills that are fundamental for proficient work. Students who perform at this level show some understanding of the science standard's concepts and procedures by demonstrating appropriate procedures during scientific inquiry, performing measurements using appropriate tools, and explaining the purposes of cell division and how an organism's behavior allows it to survive in an environment. Some gaps in knowledge and skills are evident and may require additional instruction and remediation in order to achieve a satisfactory level of understanding.

Falls Far Below the Standard – Students who score in this level may have significant gaps and limited knowledge and skills that are necessary to satisfactorily meet the state's science standard. Students will usually require a considerable amount of additional instruction and remediation in order to achieve a satisfactory level of understanding.

Students at the "Exceeds the Standard" level generally know the skills required at the "Meets" and "Approaches" levels and are able to:	Students at the "Meets the Standard" level generally know the skills required at the "Approaches" level and are able to:	Students at the "Approaches the Standard" level generally know and are able to:
• Formulate questions based on observations • Generate a hypothesis that can be tested • Analyze data to identify trends • Form a logical argument about a correlation between variables or sequence of events • Explain how evidence supports the validity and reliability of a conclusion • Critique scientific reports from periodicals, television, or other media • Formulate new questions based on the results of a previous investigation • Identify matter based on state, density, boiling point, melting point and solubility • Describe how the acceleration of a body is dependent on its mass and the net applied force (Newton's 2nd Law of Motion) • Describe forces as interactions between bodies (Newton's 3rd Law of Motion)	• Interpret data to determine relationships between variables • Identify potential investigational error • Choose an appropriate graphic representation for collected data • Write clear, step-by-step instructions for conducting investigations or operating equipment • Apply the scientific processes of prediction, comparison, inference, data organization and identification of variables to problem solving or decision making situations • Compare solutions to best address an identified need or problem • Distinguish between dominant and recessive traits in humans • Identify matter based on reactivity, pH, and oxidation • Identify evidence that a chemical reaction has occurred including formation of precipitate, generation of gas, color change and absorption or release of heat • Classify matter as elements, compounds, or mixtures • Identify conditions under which an object will continue in its state of motion (Newton's 1st Law of Motion). • Create position-time and velocity-time graphs from measurements of moving objects.	• Demonstrate safe behavior and appropriate procedures during scientific inquiry • Perform measurements using appropriate scientific tools • Explain the purposes of cell division for growth and repair, and reproduction • Explain how an organism's behavior allows it to survive in an environment ***Separate process bullets from content bullets in each level.***

These descriptors do not include all the skills and knowledge as contained in the Science Standard.

Arizona Science Standard Performance Level Descriptors
High School

Exceeds the Standard – Students who score in this level illustrate a superior academic performance as evidenced by achievement that is substantially beyond the goal for all students. Students who perform at this level demonstrate a wealth of knowledge, skills, and abilities in fulfillment of the science standard. They can specify the requirements of a valid, scientific theory, evaluate the effectiveness of conservation practices and preservation techniques, and describe the molecular basis of heredity in viruses and living things, including DNA replication and protein synthesis.

Meets the Standard – Students who score in this level demonstrate a solid academic performance on subject matter as reflected by the science standard. Students who perform at this level are able to develop questions from observations that transition into testable hypotheses, predict the outcome of an investigation, design an appropriate written plan of action for testing a hypothesis, interpret data, and evaluate whether the data supports a proposed hypothesis. They can describe the purposes and processes of cellular reproduction, analyze the relationships among nucleic acids (DNA, RNA), genes, and chromosomes, analyze the degree of relatedness among various species, and explain how genotypic and phenotypic variations can result in adaptations that influence an organism's success in an environment.

Approaches the Standard – Students who score in this level show partial understanding of the knowledge and application of the skills that are fundamental for proficient work. Students who perform at this level show some understanding of the science standard's concepts and procedures by evaluating scientific information for relevance, demonstrate safe and ethical procedures, produce graphs that communicate data, identify the relationships among organisms, and describe the levels of organization of living things. Some gaps in knowledge and skills are evident and may require additional instruction and remediation in order to achieve a satisfactory level of understanding.

Falls Far Below the Standard – Students who score in this level may have significant gaps and limited knowledge and skills that are necessary to satisfactorily meet the state's science standard. Students will usually require a considerable amount of additional instruction and remediation in order to achieve a satisfactory level of understanding.

Students at the "Exceeds the Standard" level generally know the skills required at the "Meets" and "Approaches" levels and are able to:	Students at the "Meets the Standard" level generally know the skills required at the "Approaches" level and are able to:	Students at the "Approaches the Standard" level generally know and are able to:
• Specify the requirements of a valid, scientific explanation (theory), including that it be: logical, subject to peer review, public, and respectful of rules of evidence. • Evaluate the effectiveness of conservation practices and preservation techniques on environmental quality and biodiversity. • Analyze the costs, benefits, and risks of various ways of dealing with the following needs or problems: various forms of alternative energy, storage of nuclear waste, abandoned mines, greenhouse gasses, and hazardous wastes. • Describe the molecular basis of heredity, in viruses and living things, including DNA replication and protein synthesis. ***Separate process bullets from content bullets in each level.***	• Develop questions from observations that transition into testable hypotheses. • Predict the outcome of an investigation based on prior evidence, probability, and/or modeling. • Design an appropriate protocol (written plan of action) for testing a hypothesis. • Interpret data that show a variety of possible relationships between variables. • Evaluate whether investigational data supports the proposed hypothesis. • Evaluate the design of an investigation to identify possible sources of procedural error. • Propose further investigations based on the findings of a conducted experiment. • Explain the process by which accepted ideas are challenged or extended by scientific innovation. • Analyze the use of renewable and nonrenewable resources in Arizona. • Analyze mechanisms of transport of materials into and out of cells. • Describe the purposes and processes of cellular reproduction.	• Evaluate scientific information for relevance to a given problem. • Demonstrate safe and ethical procedures in all science inquiry. • Identify the resources needed to conduct an investigation. • Use descriptive statistics to analyze data. • For a specific investigation, choose an appropriate method for communicating the results. • Produce graphs that communicate data. • Evaluate how the processes of natural ecosystems affect, and are affected by, humans. • Describe the environmental effects of the natural and/or human-caused hazards. pollution, extreme weather • Support a position on a science or technology issue.

Students at the “Exceeds the Standard” level generally know the skills required at the “Meets” and “Approaches” levels and are able to:	Students at the “Meets the Standard” level generally know the skills required at the “Approaches” level and are able to:	Students at the “Approaches the Standard” level generally know and are able to:
	• Analyze the relationships among nucleic acids (DNA, RNA), genes, and chromosomes. • Explain how genotypic variation occurs and results in phenotypic diversity. • Assess how the size and the rate of growth of a population are determined by birth rate, death rate, immigration, emigration, and carrying capacity of the environment. • Identify components of natural selection. • Explain how genotypic and phenotypic variation can result in adaptations that influence an organism’s success in an environment. • Analyze how patterns in the fossil record, nuclear chemistry, geology, molecular biology, and geographical distribution give support to the theory of organic evolution through natural selection over billions of years and the resulting present day biodiversity. • Analyze, using a biological classification system, the degree of relatedness among various species. • Compare the processes of photosynthesis and cellular respiration. • Describe the role of organic and inorganic chemicals important to living things.	• Identify the relationships among organisms within populations, communities, ecosystems, and biomes. • Predict how a change in an environmental factor can affect the number and diversity of species in an ecosystem. • Diagram the energy flow in an ecosystem through a food chain. • Describe the levels of organization of living things.

These descriptors do not include all the skills and knowledge as contained in the Science Standard.

Arizona Science Standard Performance Level Descriptors
Grade 4

Exceeds the Standard – Students who score in this level illustrate a superior academic performance as evidenced by achievement that is substantially beyond the goal for all students. Students who perform at this level demonstrate a comprehensive range of knowledge, skills, and abilities in fulfillment of the science standard. They are able to plan simple investigations that control variables, formulate conclusions based upon data, explain the role of experiments in scientific inquiry, evaluate the consequences of environmental occurrences, and construct electric circuits.

Meets the Standard – Students who score in this level demonstrate a solid academic performance on subject matter as reflected by the science standard. Students who perform at this level are generally able to analyze data to determine trends, formulate predictions based on cause and effect relationships, describe benefits and risks related to the use of technology, compare structures and their functions in plants and animals, and measure changes in weather.

Approaches the Standard – Students who score in this level show partial understanding of the knowledge and application of the skills that are fundamental for proficient work. Students who perform at this level generally show some understanding of the science standard's concepts and procedures by being able to demonstrate safe behavior and appropriate procedures in science inquiry, measure using appropriate tools, describe the interaction of components in a system, classify animals by their traits, investigate the characteristics of magnets, and interpret the symbols on weather maps. Some gaps in knowledge and skills are evident and may require additional instruction and remediation in order to achieve a satisfactory level of understanding.

Falls Far Below the Standard – Students who score in this level may have significant gaps and limited knowledge and skills that are necessary to satisfactorily meet the state's science standard. Students will usually require a considerable amount of additional instruction and remediation in order to achieve a satisfactory level of understanding.

J 5

Arizona Science Standard Performance Level Descriptors
Grade 4

Students at the "Exceeds the Standard" level have demonstrated proficiency in the skills at the "Approaches" and "Meets" levels and also have a range of the following knowledge and skills.	Students at the "Meets the Standard" level have demonstrated proficiency in the skills at the "Approaches" level and also have a range of the following knowledge and skills.	Students at the "Approaches the Standard" level have a range of the following knowledge and skills.
Process • Plan a simple investigation that identifies the variables to be controlled. • Formulate conclusions based upon identified trends in data • Explain the role of experimentation in scientific inquiry. • Evaluate the consequences of environmental occurrences that happen either rapidly (e.g., fire, flood, tornado) or over a long period of time • Analyze the effect that limited resources (e.g., natural gas, minerals) may have on an environment. **Content** • Construct series and parallel electric circuits. • Compare rapid and slow processes that change the Earth's surface • Analyze evidence that indicates life and environmental conditions have changed • Differentiate between weather and climate as they relate to the southwestern United States	**Process** • Differentiate inferences from observations • Formulate predictions in the realm of science based on observed cause and effect relationships. • Analyze data obtained in a scientific investigation to identify trends. • Determine whether the data supports the prediction for an investigation. • Describe how natural events and human activities have positive and negative impacts on environments. • Describe benefits (e.g., easy communications, rapid transportation) and risks (e.g., pollution, destruction of natural resources) related to the use of technology. **Content** • Compare structures in plants (e.g., roots, stems, leaves, flowers) and animals (e.g., muscles, bones, nerves) that serve different functions in growth and survival. • Describe ways various resources (e.g., air, water, plants, animals, soil) are utilized to meet the needs of a population. • Recognize that successful characteristics of populations are inherited traits that are favorable in a particular environment. • Describe the role that water plays in the following processes that alter the Earth's surface features: erosion, deposition, weathering. • Measure changes in weather • Identify the Earth processes that cause erosion. • Give examples of adaptations that allow plants and animals to survive: camouflage – horned lizards, coyotes; mimicry – Monarch and Viceroy butterflies; physical – cactus spines; mutualism – species of acacia that harbor ants, which repel other harmful insects	**Process** • Demonstrate safe behavior and appropriate procedures (e.g., use and care of technology, materials, organisms) in all science inquiry. • Measure using appropriate tools (e.g., ruler, scale, balance) and units of measure (i.e., metric, U.S. customary). • Describe how science and technology (e.g., computers, air conditioning, medicine) have improved the lives of many people. • Describe the interaction of components in a system (e.g., flashlight, radio). **Content** • Classify animals by identifiable group characteristics. • Describe ways in which resources can be conserved • Investigate the characteristics of magnets • Interpret the symbols on a weather map.

These descriptors do not include all the skills and knowledge as contained in the Science Standard.

Arizona Science Standard Performance Level Descriptors
Grade 8

Exceeds the Standard – Students who score in this level illustrate a superior academic performance as evidenced by achievement that is substantially beyond the goal for all students. Students who perform at this level demonstrate a wealth of knowledge, skills, and abilities in fulfillment of the science standard. They can generate a hypothesis that can be tested, analyze data to identify trends, compose new questions based upon the results of a previous investigation, explain the basic principles of heredity, and describe the intent of Newton's 3rd Law of Motion.

Meets the Standard – Students who score in this level demonstrate a solid academic performance on subject matter as reflected by the science standard. Students who perform at this level are able to interpret data to determine relationships between variables, identify potential investigation error, display data in an appropriate graphic, and write clear instructions for conducting investigations. They can distinguish between dominant and recessive traits in humans, determine changes in characteristics of organisms over generations, classify matter as elements, compounds, or mixtures, identify matter based upon its characteristics, and identify conditions under which an object will continue in its state of motion (Newton's 1st and 2nd Laws of Motion).

Approaches the Standard – Students who score in this level show partial understanding of the knowledge and application of the skills that are fundamental for proficient work. Students who perform at this level show some understanding of the science standard's concepts and procedures by being able to formulate questions based upon observations, demonstrate appropriate procedures during scientific inquiry, perform measurements using appropriate tools, and explain the purposes of cell division and how an organism's behavior allows it to survive in an environment. Some gaps in knowledge and skills are evident and may require additional instruction and remediation in order to achieve a satisfactory level of understanding.

Falls Far Below the Standard – Students who score in this level may have significant gaps and limited knowledge and skills that are necessary to satisfactorily meet the state's science standard. Students will usually require a considerable amount of additional instruction and remediation in order to achieve a satisfactory level of understanding.

J7

Arizona Science Standard Performance Level Descriptors
Grade 8

Students at the "Exceeds the Standard" level have demonstrated proficiency in the skills at the "Approaches" and "Meets" levels and also have a range of the following knowledge and skills.	Students at the "Meets the Standard" level have demonstrated proficiency in the skills at the "Approaches" level and also have a range of the following knowledge and skills.	Students at the "Approaches the Standard" level have a range of the following knowledge and skills.
Process • Generate a hypothesis that can be tested • Analyze data to identify trends • Form a logical argument about a correlation between variables or sequence of events • Explain how evidence supports the validity and reliability of a conclusion • Critique scientific reports from periodicals, television, or other media • Formulate new questions based on the results of a previous investigation **Content** • Identify matter based on state, density, boiling point, melting point and solubility • Describe forces as interactions between bodies (Newton's 3rd Law of Motion) • Explain the basic principles of heredity	**Process** • Interpret data to determine relationships between variables • Identify potential investigational error • Choose an appropriate graphic representation for collected data • Write clear, step-by-step instructions for conducting investigations or operating equipment • Apply the scientific processes of prediction, comparison, inference, data organization and identification of variables to problem solving or decision making situations • Compare solutions to best address an identified need or problem **Content** • Distinguish between dominant and recessive traits in humans • Identify matter based on reactivity, pH, and oxidation • Identify matter based on state, density, boiling point, melting point and solubility • Identify evidence that a chemical reaction has occurred including formation of precipitate, generation of gas, color change and absorption or release of heat • Classify matter as elements, compounds, or mixtures • Identify conditions under which an object will continue in its state of motion (Newton's 1st Law of Motion) • Describe how the acceleration of a body is dependent on its mass and the net applied force (Newton's 2nd Law of Motion) • Create position-time and velocity-time graphs from measurements of moving objects. • Determine characteristics of organisms that could change over several generations	**Process** • Formulate questions based on observations that lead to the development of a hypothesis • Demonstrate safe behavior and appropriate procedures during scientific inquiry • Perform measurements using appropriate scientific tools **Content** • Explain the purposes of cell division for growth and repair, and reproduction • Explain how an organism's behavior allows it to survive in an environment

J8

These descriptors do not include all the skills and knowledge as contained in the Science Standard.

Arizona Science Standard Performance Level Descriptors
High School

Exceeds the Standard – Students who score at this level demonstrate superior academic performance and knowledge at all levels in fulfillment of the science standard. They can specify the requirements of a valid, scientific theory, evaluate the effectiveness of conservation practices and preservation techniques, and describe the molecular basis of heredity in viruses and living things, including DNA replication and protein synthesis.

Meets the Standard – Students who score in this level demonstrate a solid academic performance on subject matter as reflected by the science standard. Students who perform at this level are able to develop questions from observations that transition into testable hypotheses, predict the outcome of an investigation, design an appropriate written plan of action for testing a hypothesis, interpret data, and evaluate whether the data supports a proposed hypothesis. They can describe the purposes and processes of cellular reproduction, analyze the relationships among nucleic acids (DNA, RNA), genes, and chromosomes, analyze the degree of relatedness among various species, and explain how genotypic and phenotypic variations can result in adaptations that influence an organism's success in an environment.

Approaches the Standard – Students who score in this level show partial understanding of the knowledge and application of the skills that are fundamental for proficient work. Students who perform at this level show some understanding of the science standard's concepts and procedures by being able to evaluate scientific information for relevance, demonstrate safe and ethical procedures, produce graphs that communicate data, identify the relationships among organisms, and describe the levels of organization of living things. Some gaps in knowledge and skills are evident and may require additional instruction and remediation in order to achieve a satisfactory level of understanding.

Falls Far Below the Standard – Students who score in this level may have significant gaps and limited knowledge and skills that are necessary to satisfactorily meet the state's science standard. Students will usually require a considerable amount of additional instruction and remediation in order to achieve a satisfactory level of understanding.

Arizona Science Standard Performance Level Descriptors
High School

Students at the "Exceeds the Standard" level have demonstrated proficiency in the skills at the "Approaches" and "Meets" levels and also have a range of the following knowledge and skills.	Students at the "Meets the Standard" level have demonstrated proficiency in the skills at the "Approaches" level and also have a range of the following knowledge and skills.	Students at the "Approaches the Standard" level have a range of the following knowledge and skills.
Process • Specify the requirements of a valid, scientific explanation (theory), including that it be: logical, subject to peer review, public, and respectful of rules of evidence. • Evaluate the effectiveness of conservation practices and preservation techniques on environmental quality and biodiversity. • Analyze the costs, benefits, and risks of various ways of dealing with the following needs or problems: various forms of alternative energy, storage of nuclear waste, abandoned mines, greenhouse gasses, and hazardous wastes. **Content** • Describe the molecular basis of heredity, in viruses and living things, including DNA replication and protein synthesis.	**Process** • Develop questions from observations that transition into testable hypotheses. • Predict the outcome of an investigation based on prior evidence, probability, and/or modeling. • Design an appropriate protocol (written plan of action) for testing a hypothesis. • Interpret data that show a variety of possible relationships between variables. • Evaluate whether investigational data supports the proposed hypothesis. • Evaluate the design of an investigation to identify possible sources of procedural error. • Propose further investigations based on the findings of a conducted experiment. • Explain the process by which accepted ideas are challenged or extended by scientific innovation. • Analyze the use of renewable and nonrenewable resources in Arizona. **Content** • Analyze mechanisms of transport of materials into and out of cells. • Describe the purposes and processes of cellular reproduction. • Analyze the relationships among nucleic acids (DNA, RNA), genes, and chromosomes. • Explain how genotypic variation occurs and results in phenotypic diversity. • Assess how the size and the rate of growth of a population are determined by birth rate, death rate, immigration, emigration, and carrying capacity of the environment.	**Process** • Evaluate scientific information for relevance to a given problem. • Demonstrate safe and ethical procedures in all science inquiry. • Identify the resources needed to conduct an investigation. • Use descriptive statistics to analyze data. • For a specific investigation, choose an appropriate method for communicating the results. • Produce graphs that communicate data. • Evaluate how the processes of natural ecosystems affect, and are affected by, humans. • Describe the environmental effects of the natural and/or human-caused hazards. pollution, extreme weather • Support a position on a science or technology issue. **Content** • Identify the relationships among organisms within populations, communities, ecosystems, and biomes. • Diagram the energy flow in an ecosystem through a food chain. • Describe the levels of organization of living things.

J 10

	• Identify components of natural selection. • Explain how genotypic and phenotypic variation can result in adaptations that influence an organism's success in an environment. • Predict how a change in an environmental factor can affect the number and diversity of species in an ecosystem. • Analyze how patterns in the fossil record, nuclear chemistry, geology, molecular biology, and geographical distribution give support to the theory of organic evolution through natural selection over billions of years and the resulting present day biodiversity. • Analyze, using a biological classification system, the degree of relatedness among various species. • Compare the processes of photosynthesis and cellular respiration. • Describe the role of organic and inorganic chemicals important to living things.	

These descriptors do not include all the skills and knowledge as contained in the Science Standard.

J 11

Calculating a Meaningful Standard Error for the Bookmark Cut Score

The Bookmark Standard Setting Procedure: Methodology and Recent Implementations

Calculating a Meaningful Standard Error for the Bookmark Cut Score

In the Bookmark Standard Setting Procedure for a given grade and content area, participants are assigned to roughly equivalent small groups that work independently through Round 2. Thus, the set of Round 2 cut scores provide some information about the stability of consensus in Bookmark cut scores across independent small group replications. To quantify this degree of consensus, we calculate the cluster sample standard error (Cochran, 1963, p. 210) of the Round 2 mean cut score. Cluster sample standard errors are appropriate when, as may be reasonably assumed here, data are collected from groups and independence can be assumed between groups but not within groups.

For the Bookmark Procedure, the standard error of the Bookmark cut score (SE_{cut}) is based on the cluster sample standard error of the Round 2 mean cut score. Because the final Bookmark cut scores are based on the *median* of the group instead of the mean, this cluster sample standard error (SE_{cut}) is adjusted by $\sqrt{\frac{\pi}{2}}$ (Huynh, 2003). The standard error of the Bookmark cut score is:

$$SE_{cut} = \left(\sqrt{\frac{\pi}{2}}\right)\left(\sqrt{\frac{S^2}{N}\left[1+\left(\frac{N}{n}-1\right)r\right]}\right),$$

where S^2 is the sample variance of individual Round 2 cut scores, r is the Round 2 intraclass correlation, N is the number of participants, and n is the number of groups. To be precise, if Y_{ik} is the cut score from the i^{th} participant in the k^{th} group, $\overline{Y}_k$ is the average cut score for group k, and $\overline{\overline{Y}}$ is the average of all Round 2 cut scores, then

$$r = \frac{Var(\overline{Y}_k)}{Var(\overline{Y}_k) + Var(Y_{ik} - \overline{Y}_k)} \quad \text{and} \quad S^2 = \frac{1}{N-1}\sum_{n,k}\left(Y_{nk} - \overline{\overline{Y}}\right)^2$$

If we have only two groups ($n=2$) and perfect dependence (agreement) within groups ($r=1$), then the cluster sample standard error simplifies to $SE_{cut} = \left(\sqrt{\frac{\pi}{2}}\right)\left(\frac{|Y_1 - Y_2|}{2}\right)$, which is the standard error formula employed by NAEP for two independent replications of a modified Angoff procedure (ACT, 1983, pp. 4-8). If, on the other hand, individual participants acted independently of their groups ($r=0$), then the cluster sample standard error simplifies to the traditional standard error of the mean for independent observations, $SE_{cut} = \left(\sqrt{\pi/2}\right)\left(\sqrt{S^2/N}\right)$. In this manner, SE_{cut} provides a simple, flexible, and general way to quantify the amount of uncertainty associated with final Bookmark cut scores.

It is appropriate (if statistically imprecise) to say that repeated replications of this very standard setting procedure with different judges sampled from the same population of potential judges would result in a range of cut scores, most of which would fall in a band of width 4* SE_{cut}. In the graphical displays of participant data, we depict such an interval centered at the median of the Round 3 cut score. The purpose of calculating statistics like SE_{cut} and producing graphs of the types displayed here is to effectively communicate the complex information that is gathered during a Bookmark Standard Setting Procedure.

References

ACT (1993). Setting achievement levels on the 1992 National Assessment of Educational Progress in Mathematics, Reading, and Writing: A technical report on reliability and validity.

Cochran, W. G. (1963). *Sampling techniques*. New York: John Wiley & Sons.

Huynh, H. (2003, August). Technical Memorandum for Computing Standard Error in Bookmark Standard Setting. (The South Carolina PACT 2003 Standard Setting Support Project). Columbia: University of South Carolina.

The Bookmark Standard Setting Procedure: Methodology and Recent Implementations

Daniel M. Lewis, Donald Ross Green, Howard C. Mitzel,

Katherine Baum, Richard J. Patz

CTB/McGraw-Hill

Paper presented at the 1998 Annual Meeting of the National Council on Measurement in Education

1. Introduction

Setting performance standards has become commonplace due to the standards-based education reform movement, Title 1 requirements, and public demands for educational accountability. However, standard setting—the determination of the cut scores for an assessment used to measure students' progress towards performance standards—remains a controversial topic. Recent trends in standards and assessments have presented challenges for standard setting techniques. First, there is a need for a standard setting procedure that efficiently accommodates multiple cut scores. Title 1 requires the demonstration of growth through at least three performance levels—Partially Proficient, Proficient, and Advanced. Second, there is a need for a standard setting procedure that accommodates multiple item types—selected-response (SR) and constructed-response (CR). The development of new standard setting procedures has been driven in part because the widely used Angoff procedure (Angoff, 1971) does not accommodate these trends effectively and has been criticized as being seriously flawed (National Academy of Education, 1993; Mitzel, 1996).

The Bookmark Standard Setting Procedure (Lewis, Mitzel, and Green, 1996) is an item response theory-based item mapping procedure developed to address these trends in standards and assessment and to simplify the cognitive tasks required of the participants setting the cut scores. This paper presents the methodology used to conduct the Bookmark Procedure. Section 2 reviews item response theory (IRT) based standard setting procedures. Section 3 describes the Bookmark Procedure in detail. The results of recent implementations of the Bookmark Procedure are presented in Section 4. The paper closes with a discussion of these results in Section 5 and conclusions in Section 6.

2. Review of IRT-Based Item Mapping Procedures

Item mapping, sometimes referred to as "behavioral anchoring," has been used for over a decade to help identify what students at various scale locations know and are able to do. NAEP (ETS, 1987) used scale anchoring to help interpret what students know and are able to do by mapping selected "anchor" points on the scale for the NAEP reading assessment. They selected items that discriminated well according to the criteria, "(a) eighty percent or more of the students at that [anchor] point could answer the item correctly; (b) less than 50 percent of the students at the next lower [anchor] point could answer the item correctly…" (ETS, 1987, p. 386). Item mapping, then, refers to the general approach of mapping items to locations on the IRT scale such that students with scale scores near the location of specific items can be inferred to hold the knowledge, skills, and abilities required to respond successfully to those items. NAEP continued to use scale anchoring to help interpret their results for later assessments, but the discrimination criteria applied to anchor items was modified.

The 1991 Maryland School Performance Assessment Program (MSPAP) used an item mapping procedure to set proficiency levels (CTB Macmillan/McGraw-Hill, 1992). For this purpose, score points for performance assessment items were mapped to the scale at the IRT maximum information location. The proficiency levels were set by identifying interpretable clusters of item locations on the scale and the items falling within each cluster were analyzed by content experts to interpret what students in each proficiency level knew and were able to do.

Both the NAEP anchor points and the 1991 MSPAP proficiency levels were intended to help interpret what students at various points on a scale knew and were able to do. Neither was a "true" standard setting procedure in the sense that no judgments were made concerning what students should know and be able to do; instead, both used item mapping as a means to interpret what students did know and could do at various scale locations.

NAEP conducted a bona fide standard setting for the 1992 math and reading assessments using a modified Angoff procedure (Angoff, 1971). An item mapping study was conducted as part of the review of the achievement level setting (National Academy of Education, 1993). Content experts evaluated the appropriateness of the cut scores and the quality of the achievement level descriptions. Item maps, in which items were located at the point where 80% of students in the appropriate grade could answer the items correctly (after allowing for guessing), were provided to facilitate the evaluation. Although the approach used was not intended as a new or alternative standard setting method, several positive features of the item mapping approach were noted and contrasted with the Angoff procedure that was used to set cut scores. For example, it was noted that participants using the item mapping approach had "...a more systematic understanding of the item pool as a whole than did participants using the Angoff approach (National Academy of Education, 1993, p. 110)."

One drawback of the method was also reported—the lack of clear guidelines for the probability level at which to map items to the scale. It was noted that the 80-percent-correct level possibly contributed to the experts setting very high cut scores for some of the achievement levels, and that different cut scores would possibly have resulted had a 65-percent-correct mapping criterion been used.

An "item matching" procedure was used to set proficiency levels for the 1993 MSPAP (Westat, 1994). Participants studied proficiency level descriptions and conceptualized what students at a higher level could do that students at the next lower level could not do. Initial cut scores were determined by having participants match items to the proficiency level descriptions. For example, to determine the level 2 cut score, participants examined items in order of scale location and identified the items as "clearly level 1," "clearly level 2," or "borderline." When participants identified a "run" of "clearly level 1" items followed by a "run" of "clearly level two" items, the scale locations of the items constituting the two runs were used to identify the level 2 cut score. Initial cut scores for higher levels were determined in an analogous manner, and final cut scores were determined after several rounds of discussion and consensus building.

Lewis and Mitzel (1995) developed an "IRT-Modified Angoff Procedure" for which SR items were mapped onto the IRT scale at the location at which a student would have a .5 probability of a correct response, with guessing factored out. Each positive CR item score point was mapped onto the same IRT scale at the location at which a student would have a .5 probability of obtaining at least the given score point. To determine a proficient cut score, participants conceptualized "just barely proficient" students, studied the test items in order of scale location, and classified each item according to whether a just barely proficient student should have greater than, less than, or equal to a .5 likelihood of success on the item. The cut score was determined by averaging the locations of items that participants classified at the "equal to .5" level.

Under both the Maryland 1993 standard setting procedure (Westat, 1994) and the Lewis and Mitzel (1995) procedure participants could, and did, classify items such that the participants' classifications were not consistent with the scale locations. Under the Maryland procedure, participants classified some items with higher scale locations as being associated with lower proficiency levels than other items with lower scale locations. Under the Lewis and Mitzel procedure, participants judged that Proficient students should have greater success on some items with higher scale locations than on other items with lower scale locations. This inconsistency might in part be explained by noting that the scaling of items is based on empirical student performance data, that is, what students do know and can do, and that participant judgments were based on expected student performance, that is, what students should know and be able to do. However, making judgments based on "what students should know and be able to do" without conditioning those judgments based on "what students do know and can do" can lead to serious problems in 1) interpreting the results of the assessments to which standards are applied and 2) assessing student growth relative to content standards. These problems are discussed by Lewis and Green (1997).

In 1995, the Bookmark Standard Setting Procedure was developed and used to set standards for CTB/McGraw-Hill's new standardized assessment TerraNova® and has been used to set standards in 18 states or districts from 1996 to 1998. The Bookmark Procedure evolved from Lewis and Mitzel's IRT-Modified Angoff Procedure and was designed to remove the inconsistency noted above between participants' item level judgments and the items' scale locations. This was accomplished by moving the level of judgment from the item level to the cut score level, that is, instead of making judgments about each item, participants considered all the items together to make judgments about each cut score.

Several aspects of the IRT-Modified Angoff Procedure that were particularly successful were retained in the Bookmark Procedure. Most notable are 1) the use of the ordered item booklet to help participants understand how items work together to measure student achievement relative to specified content standards and 2) the common framework for interpreting SR and CR items by mapping them to the same scale and at the same probability level. These two components were central to the primary goals of the Bookmark Procedure—to provide a standard setting procedure that treats SR and CR items in a unified manner and that is based on judgments that ease the cognitive load on participants by drawing primarily on the participants' expertise, that is, their understanding of content standards, the curriculum, teaching practices, the assessment, and student performance. The fundamental tasks required of participants in the Bookmark Procedure are analyzing items to determine what they are measuring and specifying which items students in the various performance levels should be expected to respond to successfully. We next consider the Bookmark Procedure in detail, first providing information about basic assumptions underlying the structure of the procedure.

3. Basic Assumptions and Overview of The Bookmark Procedure

3.1 Mapping Items to the IRT Scale

Item response theory (IRT, Lord 1980) provides a framework that simultaneously characterizes the proficiency of examinees and the difficulty of test items. Each IRT-scaled item has an estimated item characteristic curve (ICC) that describes how the probability of success on the item depends on the proficiency or "scale score" of the examinee. Just as it is possible to order examinees by estimated proficiency, IRT enables items to be ordered by the proficiency needed to have a specified probability of success. The facility to order items on the IRT proficiency scale is fundamental to the Bookmark Procedure.

Selected-response (SR) items can be scaled under a variety of models, for example, the Rasch (1960) model, or the 2- and 3-parameter logistic models (Birnbaum, 1968). Constructed-response (CR) items can be scaled using polytomous models, for example, the 2-parameter or generalized partial credit model (Yen, 1993; Muraki, 1992). The 3-parameter logistic (3PL) model and the 2-parameter partial credit (2PPC) model are the default models used by CTB for SR and CR items, respectively.

Scaling SR and CR items together brings significant advantages to the standard setting process, most importantly, the ability to order the CR score points with the SR items. This joint scaling allows participants to consider all items on which the standard is to be set, regardless of item format, and to directly set a single cut score for each performance level. The joint scaling of CR and SR items can be accomplished using commercially available computer programs (e.g., PARDUX, Burket, 1996; PARSCALE, Muraki & Bock, 1991).

For the purpose of standard setting, SR and CR items are located on the IRT scale such that the location of each item type is directly interpretable and conceptually similar.

Selected-Response Items. The location of an SR item is defined as the point on the ability scale at which a student would have a .67 (2/3) probability of success, with guessing factored out. We remove consideration of guessing as a factor because participants are asked to make complex judgments about what students should know and be able to do, and the consideration of guessing unnecessarily complicates those judgments. We also note that this approach was used for the item mapping studies that followed the 1992 NAEP achievement level setting (National Academy of Education, 1993).

For the 3PL model, the probability that a student with trait or scale score θ will respond correctly to SR item *j* is given by

$$P_j(\theta) = c_j + (1 - c_j) / [1 + \exp(-1.7 a_j(\theta - b_j))].$$

where a_j is the item discrimination, b_j is the item difficulty, and c_j is the probability of a correct response by a very low-scoring student. We estimate the probability, P_j^*, of a correct response with guessing removed using the formula

$$P_j^*(\theta) = (P_j(\theta) - c_j)/(1 - c_j).$$

The location of SR item *j* is θ, such that $P_j^*(\theta) = .67$.

Constructed-Response Items. Each CR score point has a unique location on the scale. The location of a given CR score point is defined as the position on the ability scale for which students have a .67 probability of achieving at least that score point, that is, that score point or higher. This criteria was selected so that the location of the CR score point could be interpreted in a manner similar to the location of a SR item and in a way that is conceptually useful to the participants in setting the cut score.

Using the 2PPC model for CR items, the probability that a student with trait or scale score θ will respond at score level *k* to CR item *j* is given by

$$P_{jk}(\theta) = \exp(z_{jk}) / \sum_{i=1}^{m_j} \exp(z_{ji}),$$

where $z_{jk} = (k-1)\alpha_j - \sum_{i=0}^{k-1} \gamma_{ji}$, α_j and γ_{ji} , i = 1, 2, …m_{j-1}, are the parameters estimated during calibration, $\gamma_{j0} = 0$ for all *j*, and m_j is the number of levels for item *j*.

For the purpose of standard setting, the location of score point *k* for constructed response item *j*, is the scale score θ, such that $P_{jk}^*(\theta) = .67$, where

$$P_{jk}^*(\theta) = \sum_{i=k}^{m_j} P_{jk}(\theta) .$$

Although the selection of .67 as the probability level used to map items to the scale is somewhat arbitrary, this value was not selected capriciously. First, because the probability level must be considered by the participants when making their judgments, a familiar value was desired. That is, using a probability level of .5823 would not be useful, but values such as .5 (1/2), .67 (2/3), or .75 (3/4) would be. Second, other item mapping procedures and research have provided some precedent. Huynh (1998) showed that for the 3PL model, the item information function is maximized at θ for which P(θ) = (c + 2)/3. This corresponds to the value of 2/3 when guessing is factored out. Thus, the choice of 2/3 for mapping SR items corresponds to the maximum information location. Huynh states that the maximum information location associated with a correct response "…might serve as a signal that an examinee located at this place would be 'expected' to have the skills underlying the item."

3.2 Bookmark Standard Setting Materials

Many of the materials used for Bookmark Standard Settings are commonly used within other standard setting procedures, such as operational test booklets, student exemplar papers, and scoring guides. The following materials are unique to Bookmark Standard Settings and other item mapping procedures.

Ordered Item Booklets. Ordered item booklets are typically assembled using all items on which the standards are to be based, in order of scale location. The ordered item booklet focuses the participants' attention on one item per page, with the "easiest" item (lowest scale location) first and the "hardest" item (highest scale location) last. The purpose of the ordered item booklets is to help participants' foster an integrated conceptualization of what the test measures, as well as to serve as a vehicle to make cut score judgments. Studying the items one by one, from easiest to hardest, discussing what each item measures and why each item is more difficult than items that precede it in the book, is intended to provide participants with an understanding of how the trait increases in complexity as the items ascend the scale, and of the knowledge, skills, and abilities students must hold in order to respond successfully to items.

The items used in the ordered item booklets can be items from single or multiple forms of an operational test or items on a common scale from an item pool that is representative in content and difficulty of a single form of the operational test. The use of items beyond those of a single operational form is recommended when possible, to increase the generalizability of the standards to other forms to which the standards may be applied in future years.

Item Map Rating Forms. The item map rating form is a guide to the ordered item booklet, and lists all items ascending by location, that is, in the same order in which they appear in the ordered item booklets. Associated item information is also included on the item map rating form, such as the items' scale location, item number in the operational or field test booklet, the standard or objective the item was written to measure, space for the participants to make notes about the items, and the cut score judgments they are considering for each round.

3.3 Determining Cut Scores Under the Bookmark Procedure

The cut score for a given performance level, for example, Proficient, can be identified by a bookmark placed between two items in the ordered item booklet such that from the judge's perspective, the items preceding the bookmark represent content that all proficient students should be expected to know and be able to do (with at least a 2/3 likelihood of knowing the correct response for SR items or of obtaining at least the given score point for CR item score points). By placing the bookmark at the furthest most item for which this is true, a location on the ability scale can be estimated as the cut score. This is computed as the scale location of the item that appears immediately prior to the bookmark. Judgments are made at the cut score level, that is, participants consider all the items when they place their bookmarks, but the bookmarks define cut scores.

To set two cut scores defining three performance levels, for example, Partially Proficient, Proficient, and Advanced, each judge considers the items in the ordered booklet and places two bookmarks that define the two cut scores. The items that precede the first bookmark should represent content that all proficient students are expected to know and be able to do. The items that precede the second bookmark should represent content that all advanced students are expected to know and be able to do.

When an item precedes a judge's bookmark, the judge is stating that all proficient students should have ability sufficient to have at least a 2/3 likelihood of responding correctly to the SR item or of obtaining at least that score point for a CR item score point. This probability level is held only by students with scale ability locations as high or higher than the scale location of the item. Thus, all proficient students must have ability level at least as high as the scale location of each item before the bookmark. On the other hand, when an item falls after the bookmark, the judge is stating that a student could be classified as proficient, yet have less than a 2/3 likelihood of success on the item. This means that a student could have ability lower than the location of the first item after the bookmark and still be classified as proficient. Thus, the proficient cut score is at least the location of the item immediately prior to the bookmark but less than the location of the item following the bookmark. The location of the item immediately prior to the bookmark is used as the operational cut score.

3.4 Writing Performance Level Descriptors

Performance level descriptors are intended to be valid descriptions of the knowledge, skills, and abilities held by students that place in the various performance levels. Performance level descriptors emerge as an outcome of setting cut scores under the Bookmark Procedure. For example, suppose two cut scores are set defining the three performance levels Partially Proficient, Proficient, and Advanced. Items prior to the Proficient bookmark reflect content that all Proficient students are expected to know and be able to do, and therefore, the knowledge, skills, and abilities required to respond successfully to these items are synthesized to form descriptors of the Proficient student. Similarly, the items following the Proficient bookmark and prior to the Advanced bookmark are used to yield descriptors of the additional knowledge, skills, and abilities a student must hold to be considered Advanced.

The estimated probability of a successful response for a student in a given performance level is at least .2/3 for the items used to write the performance level descriptors. Thus, descriptors written with this approach are valid to the degree that participants can communicate the knowledge, skills, and abilities required to successfully complete the items attributed to the respective performance levels. Of course, because they are based on probabilities, not every student will have mastered all the skills attributed to them by the descriptors. The validity of performance level descriptors written in this manner is discussed more fully by Lewis and Green (1997).

3.5 Bookmark Standard Setting Panel Composition and the Use of Multiple Panels

Operationally, the composition of a standard setting panel results from the sponsoring agency's selection criteria and availability of participants. We recommend at least 18 participants per panel. The panel of participants for a given grade and content area are typically divided into three small groups. One participant within each small group is predesignated to act as a small group facilitator for the process, and receives training prior to the standard setting. Small-group facilitators are selected from the pool of participants based on experience with the students, curriculum, instruction, assessment, and the ability to facilitate groups. The small-group facilitators are voting members of their small group. The sponsoring agency makes recommendations for the assignment of participants to small groups such that the three small groups are roughly balanced in terms of the educational background and geographic location of the participants. The use of small groups facilitates having all participants actively involved in the discussion of items and expectations for student performance. A Bookmark standard setting is typically facilitated by a single large group leader who is responsible for monitoring the process for a given grade and content area and the small group facilitators who monitor the process within their small groups.

The use of multiple small groups is integrated into the structure of the judgment process. Prior to the first round of judgments, participants study the ordered item booklets within their small groups, and discuss what each item measures and why each item is more difficult than the preceding items in the booklet. Following discussion, participants make individual and independent Round 1 judgments, that is they place bookmarks that indicate the items that reflect content they expect students in each performance level to know and be able to do.

In Round 2, each small group discusses the items for which there was not consensus according to the small group's Round 1 judgments. For a given performance level, these are the items in the ordered item booklet between the first and last of the small group participants' bookmarks. This appropriately narrows the discussion only to the items for which participants have differing opinions relative to expected student performance for a given performance level. Following discussion, Round 1 judgments may be modified with Round 2 judgments.

Prior to Round 3, a small-group judgment is computed for each small group as the median of the small group's bookmark placements. In Round 3, the large group is presented with each small group's Round 2 judgments and the estimated percent of students in each performance level based on the current large group median. The large group discusses the reasonableness of the impact data and the items for which their was not consensus among the small groups. Following discussion, Round 2 judgments may be modified with Round 3 judgments.

The Bookmark Procedure is structured so that each small group works independently of the other small groups until the third round. The standard error estimated from each small groups' independent Round 2 results provides a measure of the stability of the cut scores, as discussed in the next section.

3.6 Capturing and Communicating Degrees of Consensus

The Bookmark Standard Setting Procedure is a collaborative enterprise that fosters consensus among participants as to the standards to which we hold our students accountable. However, consensus is not forced. In the results discussed in Section 4, varying degrees of consensus were attained. It is important that the degree of consensus be measured and reported with the recommended cut scores to the governing bodies who make final cut score decisions.

The degree of consensus is quantified by calculating a standard error for each cut score arrived at through the multiple-group, three-round process. Because the small groups act independently through the first two rounds, an appropriate standard error can be calculated by treating individual Round 2 scores as if sampled from independent clusters. Formulas for the cluster sample standard error (Cochran, 1963, p. 210) are presented in Appendix 1.

Data arising in standard setting contexts have complex dependency structures and reflect many sources of error. It is important to appreciate this complexity and avoid making strong conclusions based on statistical procedures whose assumptions can not be satisfied. In Bookmark standard settings we use appropriately general statistics such as the cluster sample standard error, as well as graphics to help inform these judgments.

4. Recent Implementations of the Bookmark Procedure

4.1 Background

Table 1 summarizes the grades, content areas, test scales, test formats, and numbers of participants associated with four state and one district Bookmark standard settings facilitated by CTB in 1996 and 1997. A total of twenty panels set cut scores in grades ranging from 3 to 10 in Reading, Language Arts, and Mathematics.

For thirteen of the twenty grade/content areas, the ordered item booklets used to set cut scores included more items than were on the operational test forms. As Table 1 indicates, the operational test forms had an average of 67 score points and the ordered item booklets used to set cut scores had an average of 111 score points. The operational tests were all composed of a mixture of SR and CR items with an average of 76 percent SR items and 24 percent CR items. On average 59 percent of the total score points were from SR items and 41 percent were from CR items. The ordered item booklets used to set standards had an average of 73 percent SR items and 27 percent CR items. On average, 54 percent of the total score points in the ordered item booklets were from SR items and 46 percent were from CR items.

Table 1 also shows the number of cut scores, number of small groups, and total number of judges per grade/content area.

4.2 An Illustrative Example

Figures 1-4 illustrate the Bookmark Standard Setting Procedure for an example selected from the recent implementations. In this case, three cut scores were set for a Grade 8 Language Arts assessment. Figures 1, 2, and 3 show the individual participants' Proficient cut score ratings for Small Groups 1, 2, and 3, respectively. The vertical axes indicate the test scale referenced to a mean of 0 and standard deviation of 1. The horizontal axes indicate the round (1, 2, or 3).

Figure 1 shows the Proficient cut score ratings for the four participants in Small Group 1. Note that there is a reasonable amount of variability in the first round, with Group 1 participants' cut scores ranging from .05 to .44 on the scale. The observed variability reflects the fact that in the first round, participants make individual and independent judgments.

In the second round, the small group participants discuss and debate the rationale and perspective that lead to each of their Round 1 judgments. This tends to decrease the variability within each small group. In the case of Group 1 (Figure 1), a high degree of consensus has been reached in Round 2, with participants' cut scores ranging from .41 to .44 on the scale. Three of the four Group 1 participants raised their cut scores, apparently strongly influenced by the fourth participant's perspective.

In the third round, small-group cut scores are computed for each small group (based on small-group medians). Each small group presents the rationale and perspective that lead to their Round 2 judgments, and impact data is presented. In the example indicated in Figure 1, all participants in Group 1 maintained their Round 2 judgments in Round 3. This was probably due to the fact that Small Groups 2 and 3 both made Round 2 judgments that were very similar to those of Small Group 1, as can be observed in Figures 2 and 3.

Figures 2 and 3 illustrate the three rounds of judgments for Small Groups 2 and 3, respectively. Figure 2 indicates that Group 2 made judgments for each round that were very similar to those of Group 1. Figure 3 shows a different pattern of ratings for Small Group 3. There is a reasonable amount of variability in the Round 1 ratings for Small Group 3, with the five participants' cut scores ranging from .31 to .61. In the second round, we see the results of consensus building, however in this case, the participants tended toward the group's median cut score. The range of the participants' cut scores (.41 to .46) has decreased considerably from that of Round 1. In the third round, Small Group 3 reached consensus, with all five participants rating the Proficient cut score at .44.

Figure 4 illustrates the judgments for all participants, by round, for all three cut scores (Partially Proficient, Proficient, and Advanced). The middle set of lines indicate the Proficient judgments examined in Figures 1-3. It can easily be seen that in Round 2, each of the three groups independently arrived at the same median cut score (.44). However, this does not occur routinely. The reader need only look at the patterns for the Advanced and Partially Proficient cut scores to observe that although Round 2 does typically bring a degree of consensus, it is not as uniform for these cut scores as for the Proficient cut score.

Also depicted in Figure 4 are confidence bands centered at the Round 3 median cut score with a width of two Round 2 standard errors. The Round 3 median best captures the consensus cut score from the entire Bookmark Procedure. Round 2 standard errors are used to quantify the degree of consensus obtained across independent groups, as discussed in Section 3.6 Capturing and Communicating Degrees of Consensus. The type of information exemplified in Figure 4, is valuable to decision makers who must act on the recommendations of the standard setting panels. In the example depicted in Figure 4, the participants' recommended cut scores were adopted by the sponsoring agency.

4.3 Results

The results for the proficient cut score by round for each of the 20 examples are located in Table 2 (Summary data for all performance level cut scores are provided in Tables 3 and 4.). All statistics that are derived from the participants cut score judgments are presented in standardized units, that is, referenced to the standard deviation units of the scale. This allows statistics across scales to be compared.

The column labeled "Range (Cut)" indicates the magnitude of the range of the participants' scale score cut scores for each round and each cut score in scale standard deviation units (computed as the difference between the maximum and minimum of the participants' cut scores divided by the scale standard deviation). The column "SD (Cut)" indicates the standard deviation of the participants' scale score cut scores for each round in scale standard deviation units.

The columns labeled "Intra Class Corr" [Intraclass Correlations] and "Round 2 SE (Cut)" [standard errors] provide information about the replicability of the participants' judgments across groups. These are explained in detail in Appendix 1. The standard error is reported in scale standard deviation units.

Table 3 presents the mean SD of the participants' cut score judgments for each cut score and round (in standardized units), as well as the standard deviation, minimum, and maximum of these standard deviations. For the Advanced cut scores, the mean SDs decreased from .35 (Round 1) to .16 (Round 2) to .15 (Round 3). For the Proficient cut scores, the mean standard deviations decreased from .32 (Round 1) to .14 (Rounds 2 and 3). For the Partially Proficient cut scores, the mean standard deviations decreased from .27 (Round 1) to .16 (Round 2) to .13 (Round 3).

Table 3 also presents the mean Round 2 standard errors and intraclass correlations of the participants' cut score judgments for each cut score. The mean Round 2 standard errors are .07, .08, and .07, and the mean Round 2 intraclass correlations are .67, .69, and .70 for the Advanced, Proficient, and Partially Proficient cut scores, respectively.

Table 4 presents the mean difference in median cut scores between successive rounds, as well as the standard deviation, minimum, and maximum of the mean differences. The mean differences between the median Round 2 and Round 1 cut scores were .22, .16, and .10, for the Advanced, Proficient, and Partially proficient cut scores, respectively. The mean differences between the median Round 3 and Round 2 cut scores were .04, .00, and .04, for the Advanced, Proficient, and Partially Proficient cut scores, respectively.

5. Discussion

As would be expected in a consensus building process, the variability of participants' judgments tended to decrease in successive rounds for each cut score. The magnitude of the variability was similar for the three performance levels in each round. This is indicated by the mean standard deviations (Table 3) for the Advanced, Proficient, and Partially Proficient cut scores of .35, .32, and .27, respectively, in Round 1; .16, .14, and .16, respectively in Round 2; and .15, .14, and .13, respectively, in Round 3. This suggests a consistency in the degree to which participants are able to translate their qualitative conceptualizations of each performance level operationally into expected performance on test items. The ability for participants to be able to clearly conceptualize the knowledge, skills, and abilities of students within each performance level is fundamental to any standard setting process. These results indicate that participants seem to be able to do so to a similar degree for three performance levels. This may not hold when there are more than three performance levels.

A pattern of decreasing variability in participants' judgments from each round to the next is also consistent for the three performance levels. The mean standard deviations decreased from .35 (Round 1) to .16 (Round 2) to .15 (Round 3) for the Advanced performance level; from .32 to .14 to .14 for the Proficient performance level; and from .27 to .16 to .13 for the Partially Proficient performance level. A considerable reduction in variability occurs from

Round 1 to Round 2, but there is only a nominal reduction from Round 2 to Round 3. This indicates that the participants perspectives change considerably from the interactions within their small groups during Round 2, but do not change as much from the interactions between the small groups or the consideration of impact data in Round 3. This is desirable from the perspective that participants should feel more confident of their judgments with each round, and therefore, should be less likely to modify their judgments in subsequent rounds. However, the results may not only reflect an increase in confidence in participants' judgments, but also the support of other members within the small group to maintain their judgments in spite of differences between the small groups.

The mean standard errors computed from Round 2 provide an estimate of the variability of the cut scores across panels. The mean standard errors of .07, .08, and .07 for the Advanced, Proficient, and Partially Proficient cut scores are of similar magnitude to those reported for Math and Reading in the NAEP 1992 standard setting (ACT, 1993). It is important to remember that these are estimated from the small groups' independent Round 2 results.

The mean Round 2 intraclass correlations of .67, .69, and .70 for the Advanced, Proficient, and Partially Proficient cut scores, respectively, indicate that an appropriate degree of within-group consensus occurred in Round 2, and that individual judgments should not be treated as independent once group discussions have taken place.

Several conclusions can be drawn from looking at the mean differences between the median of the participants' cut scores between Rounds 2 and 1 and between Rounds 3 and 2. The mean differences in medians between Rounds 2 and 1 of .22, .16, and .10, for the Advanced, Proficient, and Partially Proficient cut scores, respectively, indicate that participants' cut scores tend to rise considerably from Round 1 to Round 2. This is somewhat surprising, as one might expect participants' judgments to tend toward the median, but leave the median relatively unchanged. The rise may be attributable to social pressure for high standards. For example, suppose one participant enters Round 2 having placed his/her bookmark in the ordered item booklet at say, page 50, and a second participant has placed his/her bookmark on page 60. In Round 2, the participants discuss items 50-59 in terms of whether a student should be expected to master these items to be considered proficient. It may be that under these circumstances, a psychological advantage exists for "higher standards." It is interesting to note that the increase in median cut scores from Round 1 to Round 2 is greatest for the Advanced cut score, and the least for the Partially Proficient cut score. Thus, the increase is positively correlated with the performance level, suggesting that this social pressure is greatest when the standards are expected to be highest.

The mean differences between the median of the participants' cut scores between Round 3 and Round 2 are .04, .00, and .04, for the Advanced, Proficient, and Partially Proficient cut scores, respectively. Thus, the increase in median cut scores from Round 2 to Round 3 tends not to be large. This must be considered in light of the two new pieces of information that are provided to participants in the third round. First, the participants view and discuss the results from the other small groups. Second, the participants discuss impact data associated with the median cut score computed from all participants' bookmarks. The results indicate that although these factors can affect participants judgments, they are not systematic. Again, it seems that by Round 3, participants are well grounded in their judgments.

6. Conclusions

In sum, the results indicate that the participants are making judgments as would be expected and desired, given the structure of the Bookmark Procedure. The patterns of variability are particularly encouraging. The highest variability occurs in the first round, when participants make independent ratings, and decreases significantly from Round 1 to Round 2, but does not decrease significantly from Round 2 to Round 3. This indicates that participants listen to each others' perspectives and in many cases find the arguments persuasive and therefore modify their judgments in Round 2. The stability of the small group median scores from Round 2 to Round 3 suggest that participants have developed a stable perspective by the third round. They do not react strongly to the new information provided in the third and final round as they did to that of the second round.

Setting standards is a complex process involving educational, psychological, statistical, and ultimately, political considerations. We have observed that the Bookmark Procedure facilitates the standard setting process by providing a framework through which informed educators come to understand how a particular test measures the skills the students are expected to master, and by providing a structure that fosters rational consensus building regarding expected student performance. Participants judgments are based on well defined criteria—which items students be expected to respond successfully to be classified in the various performance levels.

Further studies are required to determine the degree to which cut scores arrived at through the Bookmark Procedure are consistent with other measures of student proficiency such as teacher judgment or cut scores set concurrently with other procedures. There is no "gold standard" for cut scores or standard setting procedures. Research has shown that different standard setting procedures will likely lead to somewhat different cut scores (National Academy of Education, 1993). However, several aspects of the Bookmark Procedure have lead CTB to make it their default standard setting method.

First, participants leave the Bookmark Standard Setting with a strong understanding of what their final cut scores mean in terms of expected student performance for each performance level, as measured by the assessment. This understanding is fostered by the use of the ordered item booklets and the structure provided by item mapping procedures in general. Observations during the item mapping studies that followed the 1992 NAEP standard setting have also been observed following each Bookmark standard setting:

> "*...the experts or judges using the item-mapping approach had a much more direct understanding of the continuum for which they were attempting to devise levels*...by engaging in discussions and studying the item maps, *participants had a more systematic understanding of the item pool as a whole* than did participants using the Angoff approach.... (National Academy of Education, 1993, p. 110)."

Second, Bookmark Standard Setting participants are able to translate this "understanding" to communicate what students in each performance level know and are able to do by writing performance level descriptors based on empirical data. Teachers, parents, and students are able to use the performance level descriptors to understand the level of achievement required for students to place in each performance level. The sponsoring agency and the public can use the performance level descriptors and the percent of students in each performance level to better understand the current state of student achievement relative to the standards.

Third, Bookmark Standard Setting participants frequently comment on how instruction would improve if every teacher could go through a similar process. Their comments suggest that they have a unique awareness of how the assessment relates to the content standards, curriculum, and instruction. CTB is currently experimenting with methods of capturing the participants' perspectives to provide information to the sponsoring agency that may improve the alignment of content standards, curriculum, instruction, and assessment. This topic is more fully discussed in Lewis and Green (1998).

Send requests for information to: Daniel M. Lewis
Research Department
CTB/McGraw-Hill
Monterey, CA 93940

References

ACT (1993). Setting achievement levels on the 1992 National Assessment of Educational Progress in Mathematics, Reading, and Writing: A technical report on reliability and validity.

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

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395-479). Reading, MA: Addison-Wesley.

ETS. (1987). The NAEP 1983-84 Technical Report. Princeton, NJ: Educational Testing Service.

Burket, G. R. (1996). PARDUX [Computer program]. Monterey, CA: CTB/McGraw-Hill.

Cochran, W. G. (1963). *Sampling techniques*. New York: John Wiley & Sons.

CTB Macmillan/McGraw-Hill. (1992). Final technical report: Maryland School Performance Assessment Program, 1991. (Available from the Maryland State Department of Education, Baltimore, MD)

Huynh, H. (1998). On score locations of binary and partial credit items and their applications to item mapping and criterion-referenced interpretation. *Journal of Educational and Behavioral Statistics, 23,* 37-58.

Lewis, D. M., & Mitzel, H. C. (September 1995). An item response theory based standard setting procedure. In D. R. Green (Chair), Some uses of item response theory in standard setting. Symposium presented at the annual meeting of the California Educational Research Association, Lake Tahoe, NV.

Lewis, D. M., Mitzel, H. C., & Green, D. R. (June 1996). Standard setting: A Bookmark approach. In D. R. Green (Chair), IRT-based standard setting procedures utilizing behavioral anchoring. Symposium presented at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix, AZ.

Lewis, D. M., & Green, D. R. (June 1997). The validity of performance level descriptors. Paper presented at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix, AZ.

Lewis, D. M., & Green, D. R. (June 1998). Assessing the state of the standards: Linking content standard, curriculum & instruction, and assessment. Paper presented at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Colorado Spring, CO.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. New York: Erlbaum.

Mitzel, H. C. (1996). Standard setting as a judgment task. In D. R. Green (Chair), IRT-based standard setting procedures utilizing behavioral anchoring. Symposium presented at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix, AZ.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Muraki, E., & Bock, R. D. (1991). PARSCALE: Parameter scaling of rating data [Computer program]. Chicago: Scientific Software, Inc.

National Academy of Education. (1993). Setting performance standards for student achievement. Stanford: Author.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research.

Westat. (1994). Establishing proficiency levels and descriptions for the 1993 Maryland School Performance Assessment Program (MSPAP). Technical Report. Rockville, MD.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local independence. *Journal of Educational Measurement, 30*, 187-213.

Appendix 1

Calculating a Meaningful Standard Error for the Bookmark Cut Score

In the Bookmark Standard Setting Procedure for a given grade and content area, participants are assigned to roughly equivalent small groups that work independently through Round 2. Thus, the set of Round 2 cut scores provide some information about the stability of consensus in Bookmark cut scores across independent small group replications. To quantify this degree of consensus, we calculate the cluster sample standard error (Cochran, 1963, p. 210) of the Round 2 mean cut score. Cluster sample standard errors are appropriate when, as may be reasonably assumed here, data are collected from groups and independence can be assumed between groups but not within groups.

For the Bookmark Procedure, the standard error of the Bookmark cut score (SE_{cut}) is given by the cluster sample standard error of the Round 2 mean cut score:

$$SE_{cut} = \sqrt{\frac{S^2}{N}\left[1+(n-1)r\right]},$$

where S^2 is the sample variance of individual Round 2 cut scores, r is the Round 2 intraclass correlation, N is the number of participants, and n is the number of groups. To be precise, if Y_{ik} is the cut score from the i^{th} participant in the k^{th} group, $\overline{Y}_k$ is the average cut score for group k, and $\overline{\overline{Y}}$ is the average of all Round 2 cut scores, then

$$r = \frac{Var(\overline{Y}_k)}{Var(\overline{Y}_k)+Var(Y_{ik}-\overline{Y}_k)} \quad \text{and} \quad S^2 = \frac{1}{N-1}\sum_{n,k}\left(Y_{nk}-\overline{\overline{Y}}\right)^2$$

If we have only two groups ($n=2$) and perfect dependence (agreement) within groups ($r=1$), then the cluster sample standard error simplifies to $SE_{cut} = \left|\overline{Y}_1 - \overline{Y}_2\right|/2$, which is the standard error formula employed by NAEP for two independent replications of a modified Angoff procedure (ACT, 1983, pp. 4-8). If, on the other hand, individual participants acted independently of their groups ($r=0$), then the cluster sample standard error simplifies to the traditional standard error of the mean for independent observations, $SE_{cut} = \sqrt{S^2/N}$. In this manner, SE_{cut} provides a simple, flexible, and general way to quantify the amount of uncertainty associated with final Bookmark cut scores.

It is appropriate (if statistically imprecise) to say that repeated replications of this very standard setting procedure with different judges sampled from the same population of potential judges would result in a range of cut scores, most of which would fall in a band of width 4* SE_{cut}. In Figures 1-4 we depict such an interval centered at the median of the Round 3 cut score. The purpose of calculating statistics like SE_{cut} and producing graphs of the types displayed here is to effectively communicate the complex information that is gathered during a Bookmark Standard Setting Procedure.

Table 1. Background Information for Recent Implementations

		Operational Test			Ordered Item Booklet						
Grade	Content Area	# of SR Items	# of CR Items	Total Score Points	# of SR Items	# of CR Items	Total Score Points	# of Cut Points	# of Small Groups	# of Judges per Small Group	Total # of Judges
3	Reading	30	13	51	74	18	104	3	2	4	8
	Language	20	4	31	56	10	83	3	1	6	6
	Math	30	8	48	96	16	132	3	2	4-5	9
6	Reading	31	9	55	98	23	159	3	2	4-5	9
	Language	24	5	43	67	12	114	3	1	6	6
	Math	31	11	58	95	33	176	3	1	8	8
8	Reading	34	6	51	67	12	98	3	2	6	12
	Language	21	5	41	43	12	88	3	3	4-5	14
	Math	31	10	54	63	20	116	3	2	5-6	11
4	Reading	46	21	95	46	21	95	3	3	6-7	19
4	Writing	32	13	59	42	34	112	3	3	6-7	20
4	Math	32	23	77	31	62	146	4	6	6-8	42
8	Math	31	20	73	31	56	145	4	7	5-6	39
10	Math	25	19	73	25	43	133	4	8	4-5	35
3	ELA	46	2	55	46	2	55	1	3	6	18
6	ELA	55	13	87	55	13	87	1	3	6-7	20
8	ELA	72	18	113	72	18	113	1	3	5-6	17
10	ELA	69	11	100	69	11	100	1	3	6	18
10	ELA	67	13	98	67	13	98	1	4	6-7	25
10	Math	59	22	74	59	22	74	1	4	6	24

Note. ELA = English/Language Arts, SR = selected response, CR = constructed response.

Table 2. Results

Grade	Content Area	Cut	Round	Range (Cut)*	SD (Cut)*	Intra Class Corr	Round 2 SE (Cut)*
3	Reading	Proficient	1	0.45	0.15		
			2	0.53	0.25	0.96	0.17
			3	0.31	0.11		
3	Language	Proficient	1	0.29	0.11		
			2	0.19	0.07	NA	NA
			3	0.00	0.00		
3	Math	Proficient	1	1.09	0.37		
			2	0.24	0.08	0.37	0.04
			3	0.00	0.00		
6	Reading	Proficient	1	0.72	0.26		
			2	0.05	0.02	0.50	0.01
			3	0.00	0.00		
6	Language	Proficient	1	0.41	0.16		
			2	0.27	0.11	NA	NA
			3	0.27	0.11		
6	Math	Proficient	1	1.32	0.36		
			2	0.67	0.19	NA	NA
			3	0.00	0.00		
8	Reading	Proficient	1	0.55	0.13		
			2	0.11	0.03	0.70	0.02
			3	0.00	0.00		
8	Language	Proficient	1	0.56	0.18		
			2	0.05	0.01	0.09	0.00
			3	0.05	0.01		
8	Math	Proficient	1	0.89	0.23		
			2	0.38	0.15	0.81	0.10
			3	0.28	0.13		
4	Reading	Proficient	1	0.97	0.25		
			2	0.32	0.13	0.72	0.06
			3	2.07	0.56		
4	Writing	Proficient	1	1.52	0.69		
			2	0.51	0.12	0.16	0.04
			3	2.13	0.55		
4	Math	Proficient	1	2.52	0.52		
			2	1.07	0.25	0.63	0.08
			3	1.05	0.20		
8	Math	Proficient	1	2.37	0.44		
			2	1.32	0.24	0.65	0.08
			3	1.32	0.24		
10	Math	Proficient	1	1.33	0.28		
			2	0.29	0.08	0.73	0.02
			3	0.42	0.10		
3	ELA**	Proficient	1	0.89	0.25		
			2	0.12	0.06	1.00	0.03
			3	0.10	0.02		
6	ELA	Proficient	1	1.53	0.29		
			2	0.18	0.08	1.00	0.05
			3	0.17	0.07		
8	ELA	Proficient	1	2.66	0.56		
			2	0.59	0.23	0.94	0.14
			3	0.09	0.02		
10	ELA	Proficient	1	1.45	0.43		
			2	1.13	0.43	0.98	0.25
			3	1.05	0.34		
10	ELA	Proficient	1	1.74	0.41		
			2	1.06	0.19	0.60	0.08
			3	1.04	0.18		
10	Math	Proficient	1	1.54	0.34		
			2	0.60	0.17	0.41	0.06
			3	0.58	0.17		

* Values are in scale standard deviation units.

** ELA = English/Language Arts.

Table 3. Summary Statistics: Meaure of Variability in Participants' Cut Score Judgments

	Standardized Standard Deviation				Standardized Standard Error				Intra Class Correlation			
	Mean	**SD**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**
Advanced												
Round 1	0.35	0.16	0.17	0.73								
Round 2	0.16	0.12	0.02	0.46	0.07	0.05	0.02	0.15	0.67	0.20	0.37	0.99
Round 3	0.15	0.15	0.00	0.51								
Proficient												
Round 1	0.32	0.16	0.11	0.69								
Round 2	0.14	0.10	0.01	0.43	0.08	0.07	0.00	0.25	0.69	0.27	0.09	1.00
Round 3	0.14	0.17	0.00	0.56								
Partially Proficient												
Round 1	0.27	0.20	0.05	0.68								
Round 2	0.16	0.14	0.03	0.53	0.07	0.04	0.03	0.13	0.70	0.30	0.11	1.00
Round 3	0.13	0.10	0.00	0.28								

Table 4. Summary Statistics: Difference Between Successive Round Medians

	Round 2 - Round 1				Round 3 - Round 2			
	Mean	**SD**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**
Advanced	0.22	0.26	-0.16	0.78	0.04	0.15	-0.11	0.52
Proficient	0.16	0.23	-0.13	0.81	0.00	0.22	-0.73	0.24
Partially Proficient	0.10	0.20	-0.11	0.66	0.04	0.16	-0.14	0.55

Note. Standardized scale score units are used.

Figure 1. Group 1 Proficient Cutscores for Rounds 1, 2, and 3

(Vertical axis in scale standard deviation units: Mean = 0, SD = 1.)

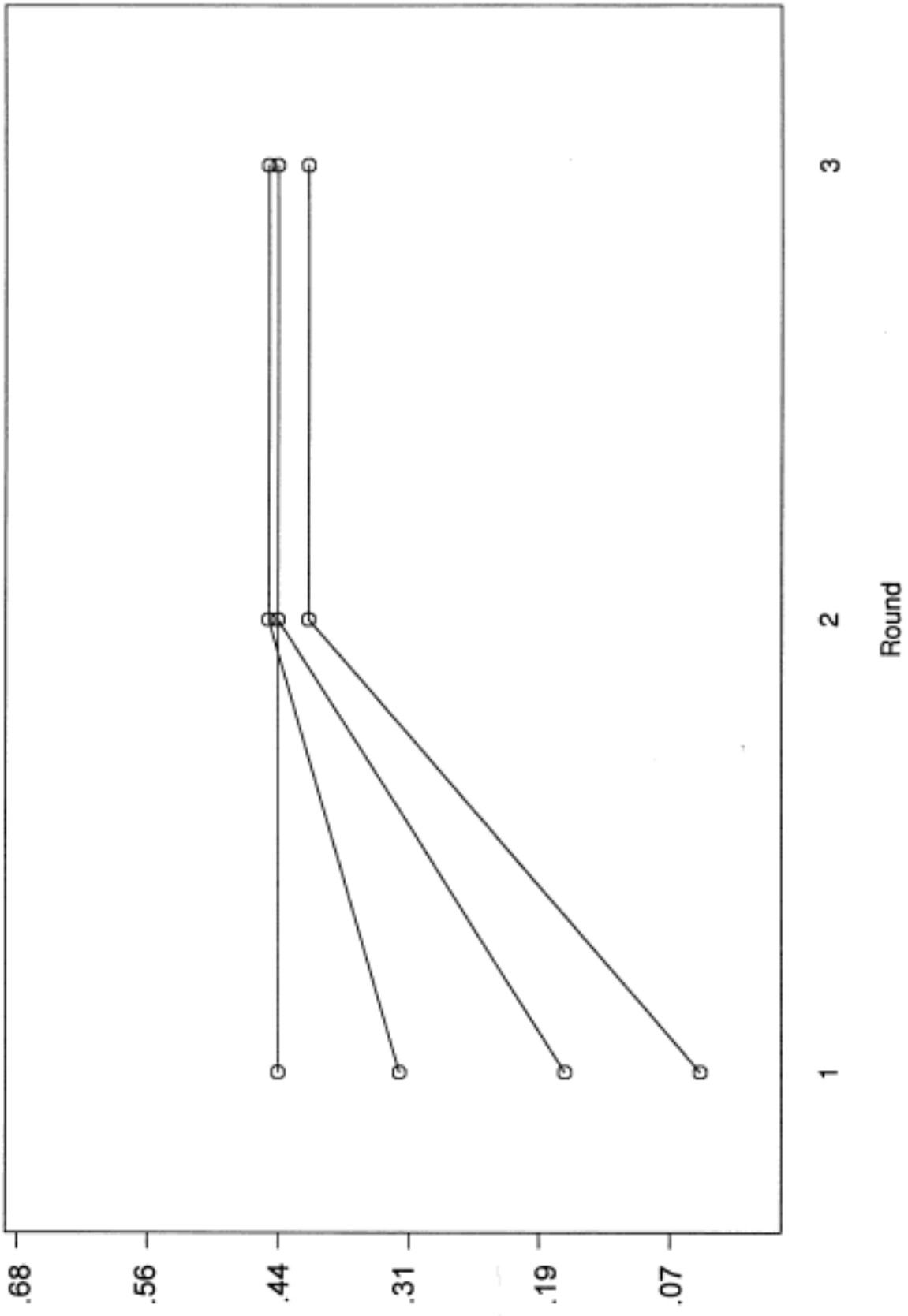


Figure 2. Group 2 Proficient Cutscores for Rounds 1, 2, and 3

(Vertical axis in scale standard deviation units: Mean = 0, SD = 1.)

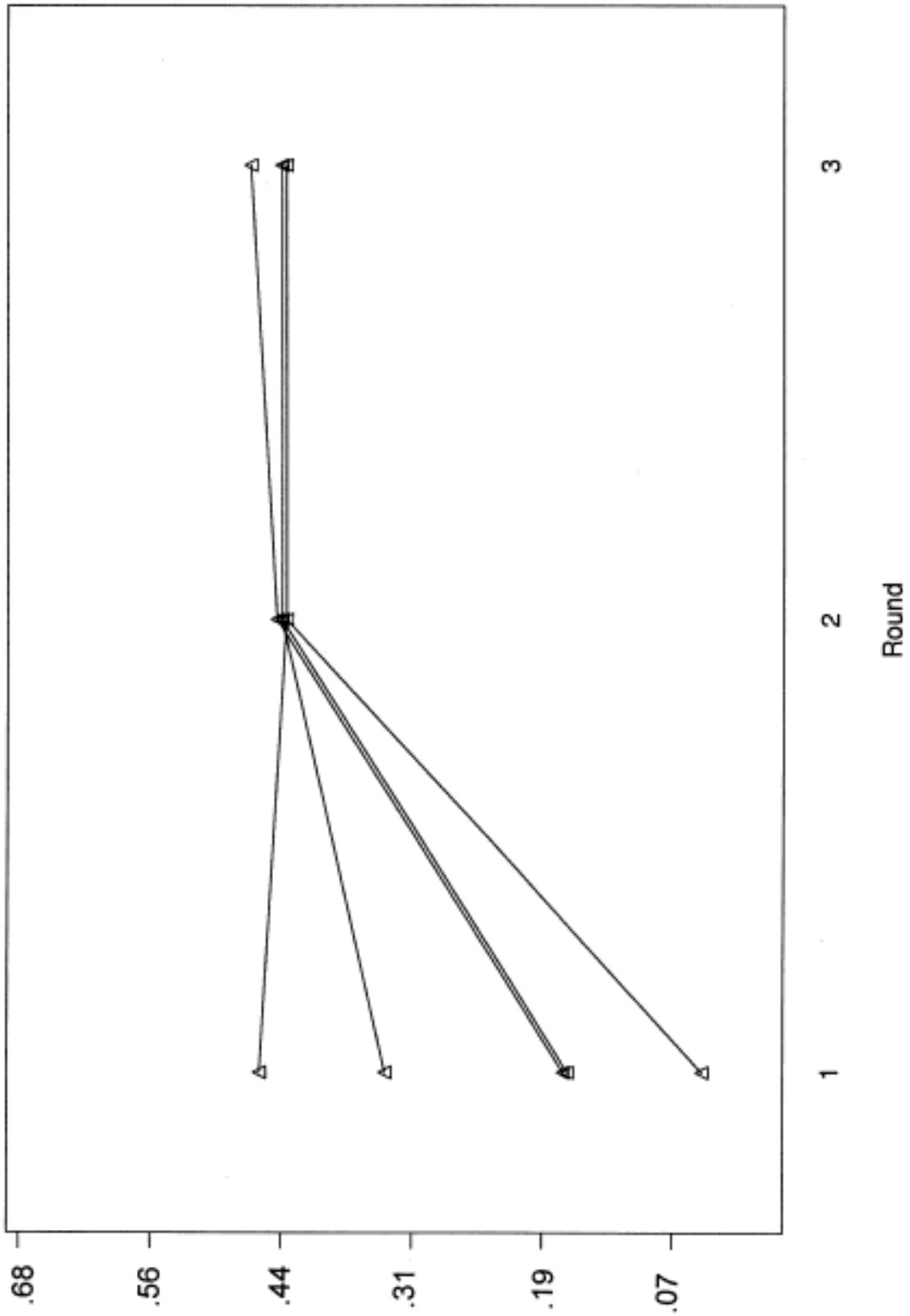


Figure 3. Group 3 Proficient Cutscores for Rounds 1, 2, and 3

(Vertical axis in scale standard deviation units: Mean = 0, SD = 1.)

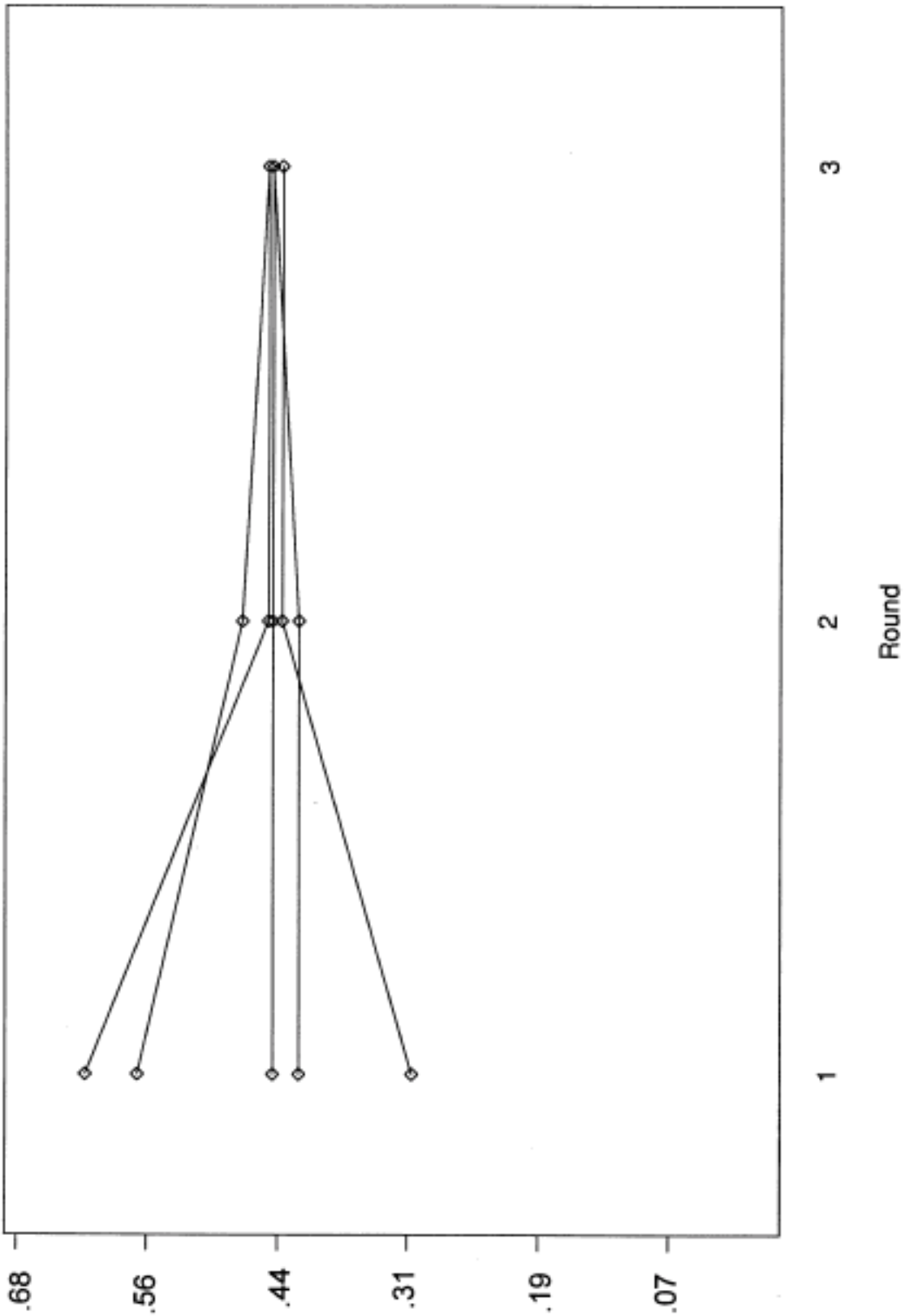
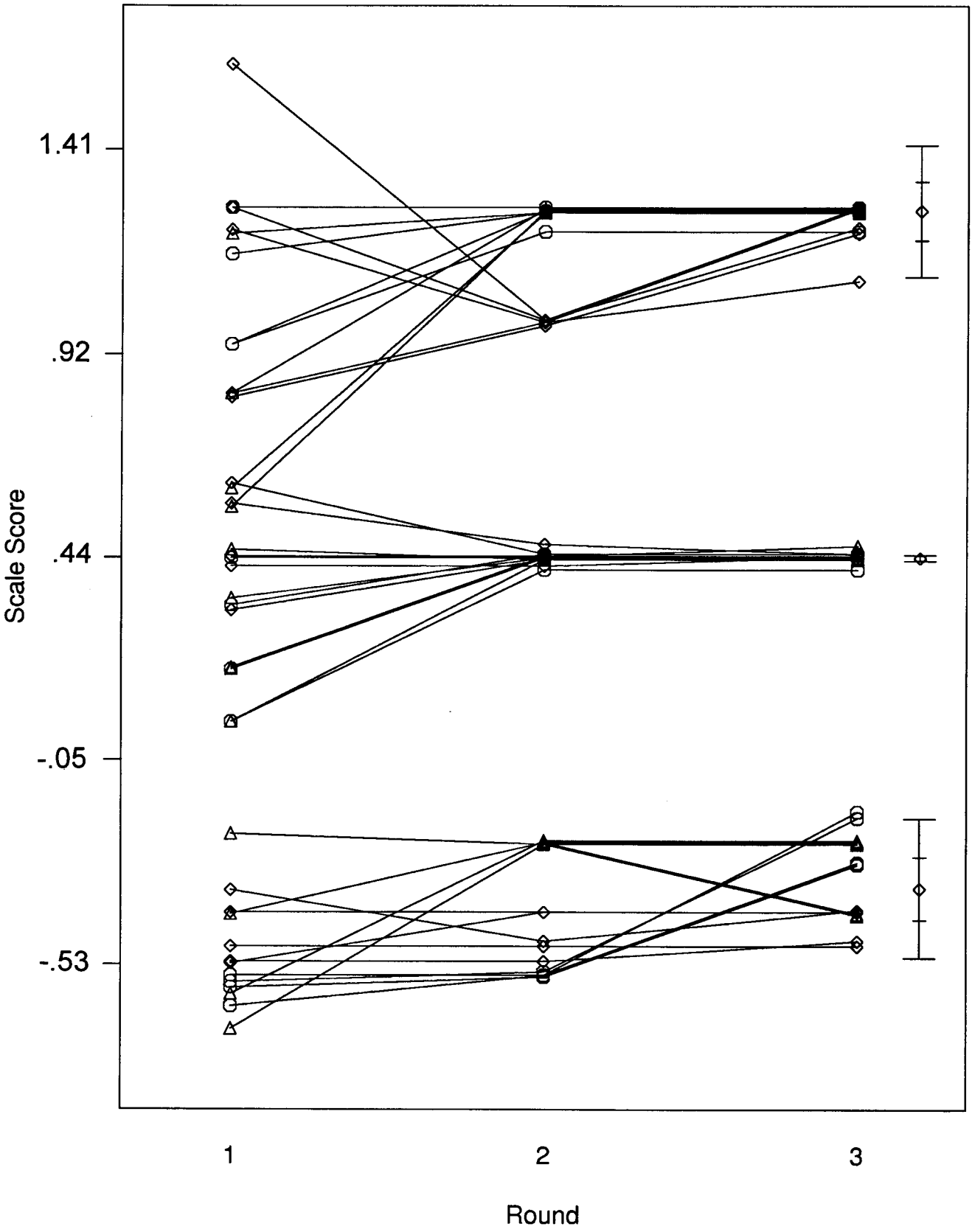


Figure 4. Advanced, Proficient, and Partially Proficient Cutscores of All Participants



K22