Colorado
Student Assessment Program

Technical Report for the
Cut Score Review
2008

for

Grades 5, 8, and 10
Science

Submitted to
Colorado Department of Education
July 2008

CTB
McGraw-Hill

Developed and published under contract with Colorado Department of Education by CTB/McGraw-Hill LLC, a subsidiary of The McGraw-Hill Companies, Inc., 20 Ryan Ranch Road, Monterey, California 93940-5703. Copyright © 2008 by Colorado Department of Education. All rights reserved. No part of this publication may be reproduced or distributed in any form or by any means, or stored in a database or retrieval system, without the prior written permission of Colorado Department of Education and the publisher. This work is based on the Bookmark Standard Setting Procedure, copyright © 1999 CTB/McGraw-Hill LLC. Bookmark Standard Setting Procedure is a trademark of The McGraw-Hill Companies, Inc.

Table of Contents

Section K

Section A

Executive Summary

Executive Summary

Staff from CTB/McGraw-Hill and the Colorado Department of Education (CDE) collaborated to conduct a cut score review for Grades 5, 8, and 10 Science of the Colorado Student Assessment Program (CSAP). The workshop was held in Denver, Colorado, on May 14-16, 2008. A modification of the Bookmark Standard Setting Procedure (BSSP; Lewis, Mitzel, & Green, 1996) was implemented to review the performance standards established in 2006. Participants in each grade participated in three rounds of activities to recommend three cut scores that define four performance levels: *Unsatisfactory, Partially Proficient, Proficient,* and *Advanced.*

Participants were recruited from across the state of Colorado to review the cut scores. Each grade group had 8 to 9 participants. Within each grade group, the CDE divided participants into two tables that were balanced in terms of relevant demographic characteristics (e.g., geographic location, school size).

The CSAP Science Cut Score Review consisted of training, orientation, presentation of preliminary bookmarks, three rounds of judgments, presentation of Round 3 results to all participants, and a smoothing discussion. The workshop lasted three days, with the first half-day devoted to Table Leader training and the remainder for the cut score review.

Table 1 summarizes the cut scores and associated impact data recommended by participants in each grade group in the final round of discussion and voting. The impact data shown to participants at the time of the cut score review were based on the student data available at the time of the workshop, comprising approximately 93% of the data from the Spring 2008 administration of the CSAP Science tests.

Table 1. Participant-Recommended Cut Scores and Associated Impact Data Based on the Final Round (Round 3)

	Cut Scores			Impact Data			
Grade	***Partially Proficient***	***Proficient***	***Advanced***	***Unsatis-factory***	***Partially Proficient***	***Proficient***	***Advanced***
5	429	508	569	14.8%	40.8%	33.6%	10.8%
8	459	512	579	23.9%	31.7%	37.0%	7.3%
10	469	507	581	26.0%	24.0%	43.9%	6.1%

Table Leaders met for the smoothing discussion. The purpose of this smoothing discussion was to establish a system of cut scores that was well-articulated and, at the same time, considerate of the participants' original recommendations. After discussion, the Table Leaders recommended changing the Grade 8 *Proficient* cut score from 512 to 507 to promote better cross-grade articulation of the impact data.

Table 2 shows the cut scores developed during the smoothing discussions, as well as the associated impact data. The impact data shown to at the time of the cut score review were based the 93% dataset from the Spring 2008 administration of the CSAP Science tests.

Table 2. Cut Scores and Associated Impact Data after the Smoothing Discussion

	Cut Scores			Impact Data			
Grade	***Partially Proficient***	***Proficient***	***Advanced***	***Unsatis-factory***	***Partially Proficient***	***Proficient***	***Advanced***
5	429	508	569	14.8%	40.8%	33.6%	10.8%
8	459	507	579	23.9%	28.3%	40.5%	7.3%
10	469	507	581	26.0%	24.0%	43.9%	6.1%

Table 3 shows the cut scores and associated impact data, calculated from the complete (100%) data from the Spring 2008 administration of the Science tests.

Table 3. Final Cut Scores and Associated Impact Data from the Complete (100%) Data

	Cut Scores			Impact Data			
Grade	***Partially Proficient***	***Proficient***	***Advanced***	***Unsatis-factory***	***Partially Proficient***	***Proficient***	***Advanced***
5	429	508	569	12.5%	40.8%	35.3%	11.5%
8	459	507	579	23.2%	28.4%	41.0%	7.4%
10	469	507	581	26.7%	24.0%	43.2%	6.1%

This report summarizes the results of the Colorado Cut Score Review for Grades 5, 8, and 10 Science. A round-by-round synopsis is included in Section B. The Master Agenda is included in Section C. The overhead slides presented to participants during Table Leader training, the opening session, and bookmark training are included in Section D. In Section E, detailed results are presented of the participants' judgments for each grade. In Section F, estimates are given of the percentage of students in each performance level at plus and minus one, two, and three standard errors of the participants' recommended final round cut scores for each grade group. Section G contains graphical representations of participants' final round judgments and standard errors. All training materials given to participants are provided in Section H. Section I contains the performance level descriptors the participants used during the cut score review. Section J contains the results of the participant evaluation of the Colorado Standard Setting. Section K contains two papers: "Calculating a Meaningful Standard Error for the Bookmark Cut Score" and "The Bookmark Standard Setting Procedure: Methodology and Recent Implementations." These papers are provided for reference.

Section B

Overview of the Cut Score Review

Overview of the CSAP Science Cut Score Review

Staff from CTB/McGraw-Hill and the Colorado Department of Education (CDE) collaborated to conduct a cut score review for Grades 5, 8, and 10 Science of the Colorado Student Assessment Program (CSAP). The workshop was held in Denver, Colorado, on May 14-16, 2008. A modification of the Bookmark Standard Setting Procedure (BSSP; Lewis, Mitzel, & Green, 1996) was implemented to review the performance standards established in 2006. Participants in each grade participated in three rounds of activities to recommend three cut scores that define four performance levels: *Unsatisfactory, Partially Proficient, Proficient,* and *Advanced.*

Participants were recruited from across the state of Colorado to review the cut scores. Each grade group had 8 to 9 participants. Within each grade group, the CDE divided participants into two tables that were balanced in terms of relevant demographic characteristics (e.g., geographic location, school size).

The CSAP Science Cut Score Review consisted of training, orientation, presentation of preliminary bookmarks, three rounds of judgments, presentation of Round 3 results to all participants, and a smoothing discussion. The workshop lasted three days, with the first half-day devoted to Table Leader training and the remainder for the cut score review.

Security at the Cut Score Review

Security was of paramount importance throughout the workshop. 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. CTB staff monitored the breakout rooms to prevent the removal of secure materials. At the end of each day, each participant’s materials were collected and audited. The secure materials were stored overnight in a secure room. At the conclusion of the workshop, the secure materials were collected, audited, and confirmed against the sign-out lists.

Bookmark Roles

CTB Staff

The CTB Standard Setting Team worked with staff from CDE to design, organize, and conduct the CSAP Science Cut Score Review. The CTB Standard Setting Team comprised Thakur Karkee, Ph.D., Ricardo Mercado, and Adele Brandstrom. Dr. Karkee is the CTB Research Scientist for the CSAP contract. Mr. Mercado, a CTB Research Project Manager, conducted the cut score review and trained participants in the process. Ms. Brandstrom is a CTB Standard Setting Specialist.

Prior to the CSAP Science Cut Score Review, this team prepared all materials for the workshop. During the workshop, this team was responsible for facilitating the workshop, training participants, entering participant results into a database, and tracking secure materials. Following the workshop, this team prepared the technical report for the cut score review.

Angelica Gordon, CTB Program Office Coordinator for the Colorado contract, coordinated the logistics for the workshop.

Group Leaders

The Group Leaders administrated the cut score review for those major portions in which participants were working. In each grade group, the Group Leader served as a facilitator and was in charge of time management, focusing the participants on the task at hand, and interacting with the participants. The Group Leader also facilitated large-group discussions and was in charge of security and data management. The Group Leader collected the rating forms from participants and communicated with CTB Research and CDE staff. The Group Leaders did not vote. The Group Leaders were provided by CTB and are listed in Table 1.

Table 1. Group Leader for Each Grade

Grade	Group Leader
5	Bevin Flaherty
8	Marie-Lise Bouscaren
10	Andrina Ortiz

Participants

Approximately 25 participants from across the state of Colorado attended the 2008 workshop to review the performance standards established in 2006 for CSAP Science in Grades 5, 8, and 10. Participants were full, voting members of their grade committees and drew upon their expertise to help review the performance standards. Table 2 shows the number of participants in each grade.

Table 2. Number of Participants in Each Grade

Grade	Number of Participants
5	8
8	9
10	8

One participant from Grade 8 Science left the workshop after Round 1; however, the participant's ratings are still included in the results for Round 1. In addition, one Grade 5 participant left prior to bookmark training. She observed the process and did not vote, but did complete an evaluation.

Following the cut score review, participants completed evaluations from which demographic information was summarized. Tables 3, 4, 5, 6, 7, and 8 show the gender, ethnicity, highest educational level, profession, work experience in years, and other type of teaching experience, respectively, of the participants in each grade group, as self-reported on the participant evaluations. Section J contains the complete results of the participant evaluation of the CSAP Science Cut Score Review.

Table 3. Gender of Participants in Each Grade

Grade	N	Male	Female
Overall	25	24.0%	76.0%
5	9	11.1%	88.9%
8	8	25.0%	75.0%
10	8	37.5%	62.5%

Table 4. Ethnicity of Participants in Each Grade

Grade	N	Asian/Pacific Islander	Black/African-American	American Indian	Hispanic	White	Other
Overall	25	4.0%	0.0%	0.0%	0.0%	96.0%	0.0%
5	9	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
8	8	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
10	8	12.5%	0.0%	0.0%	0.0%	87.5%	0.0%

Table 5. Highest Educational Level of Participants in Each Grade

Grade	N	High School	Bachelor’s	Master’s	Doctorate
Overall	25	0.0%	8.0%	84.0%	8.0%
5	9	0.0%	11.1%	88.9%	0.0%
8	8	0.0%	12.5%	87.5%	0.0%
10	8	0.0%	0.0%	75.0%	25.0%

Table 6. Profession of Participants in Each Grade

Grade	N	Teacher	Administrator	Other
Overall	25	80.0%	8.0%	12.0%
5	9	77.8%	11.1%	11.1%
8	8	87.5%	0.0%	12.5%
10	8	75.0%	12.5%	12.5%

Table 7. Work Experience in Years of Participants in Each Grade

Grade	N	1-5	6-10	11-15	16-20	21+
Overall	25	4.0%	36.0%	16.0%	16.0%	28.0%
5	9	0.0%	44.4%	22.2%	0.0%	33.3%
8	8	12.5%	37.5%	25.0%	12.5%	12.5%
10	8	0.0%	25.0%	0.0%	37.5%	37.5%

Table 8. Other Types of Teaching Experience of Participants in Each Grade

Grade	N	Special Education	Adult Education	Alternative Education	Vocational Education	English Language Learners
Overall	25	8%	28.0%	0.0%	0.0%	16.0%
5	9	22.2%	22.2%	0.0%	0.0%	22.2%
8	8	0.0%	25.0%	0.0%	0.0%	12.5%
10	8	0.0%	37.5%	0.0%	0.0%	12.5%

Table Leaders

Within each grade group, CDE divided participants into two tables that were balanced in terms of relevant demographic characteristics (e.g., geographic location, school size). Each of the two tables in a grade group had a Table Leader. Their primary role was to monitor the group discourse, keep the group focused on the task, facilitate discussions, and help maintain the schedule.

Bookmark Materials

Ordered Item Booklets

The Ordered Item Booklets (OIBs) comprised the items from the Spring 2008 assessments. The items were ordered according to their scale locations using a response probability (RP) of 0.67. Table 9 lists the number of score points in each OIB for each grade.

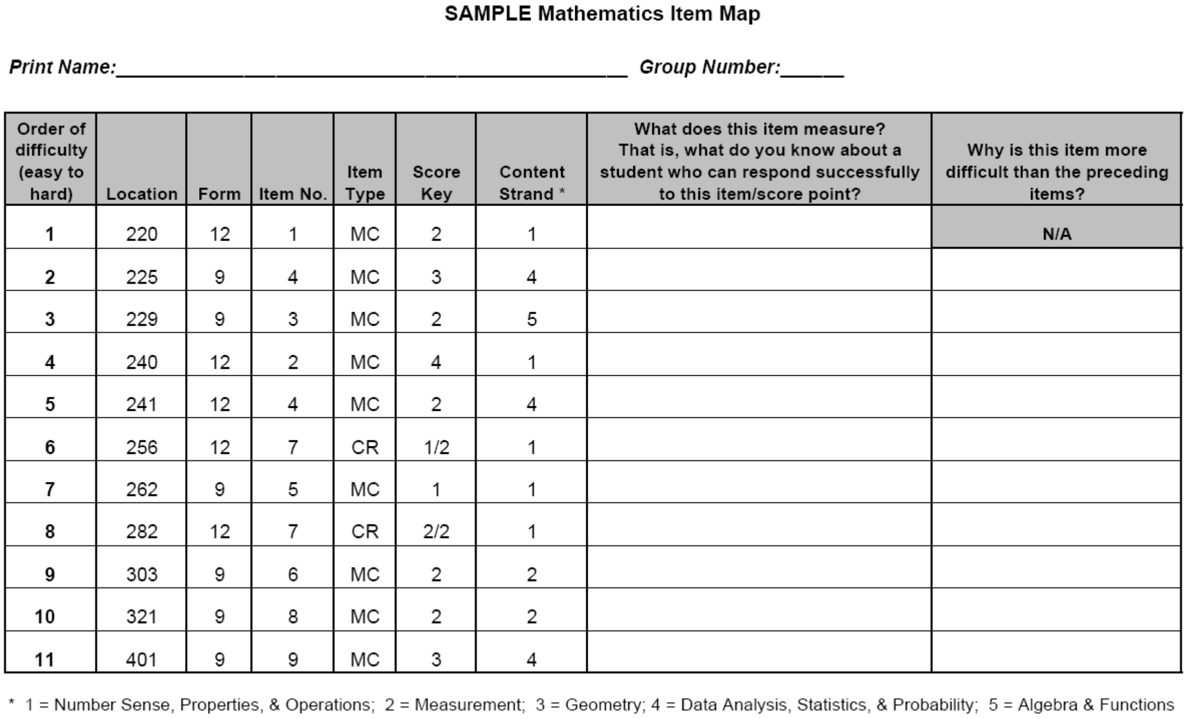
Table 9. Number of Score Points in Ordered Item Booklet for Each Grade

Grade	Number of Score Points
5	86
8	100
10	99

Item Maps

The item maps summarize the material in the OIB. The item maps consisted of nine columns: the first column indicated the item's order of difficulty, the second column indicated the location, the third column indicated the test session, the fourth column indicated the item number, the fifth column reported the score key (the correct response for a multiple-choice item and the number of score points for a constructed-response item), the sixth column showed the item type (MC for a multiple-choice item and CR for a constructed-response item), and the seventh column reported the benchmark the item measures. 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/score point?" The second of these columns asks "Why is this item more difficult than the preceding items?" Figure 1 show the item map used for training. All training materials are included in Section H.

Figure 1. Sample Item Map



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/score point?	Why is this item more difficult than the preceding items?
1	220	12	1	MC	2	1		**N/A**
2	225	9	4	MC	3	4		
3	229	9	3	MC	2	5		
4	240	12	2	MC	4	1		
5	241	12	4	MC	2	4		
6	256	12	7	CR	1/2	1		
7	262	9	5	MC	1	1		
8	282	12	7	CR	2/2	1		
9	303	9	6	MC	2	2		
10	321	9	8	MC	2	2		
11	401	9	9	MC	3	4		

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

Calculating Cut Scores

The cut score associated with a participant's bookmark placement was the average of the item locations immediately before and after the bookmark. For example, if a participant placed a bookmark between Page 9 and Page 10 in the OIB, then the resultant cut score would be the average of the locations of the items on Pages 9 and 10.

For all three grades, the locations of the items were not uniformly distributed across the test scale. There were gaps between the locations of some adjacent test items, especially in the lower and upper location ranges of the OIB. If no provision were made for these gaps, the cut scores associated with consecutive bookmark placements could have been markedly different from each other; that is, a small change in a bookmark placement could result in a large difference in the associated cut score.

By calculating cut scores as the average of the locations of the items before and after a bookmark placement, the impact of the gaps in item locations on the item map is lessened. It should be noted that this method was also used in the initial implementations of the Bookmark Standard Setting Procedure (Lewis, Mitzel, & Green, 1996).

Cut Score Review: Morning of Day 1

Table Leader Training

Table Leaders were trained on the morning of the first day of the CSAP Science Cut Score Review. During this training session, which lasted about three-and-a-half hours, Table Leaders were given an overview of the reasons for standard setting and cut score review and were trained specifically on the Bookmark Standard Setting Procedure (BSSP). They were given a synopsis of each day's activities, as well as their responsibilities during those activities. The Master Agenda is included in Section C, and the training overheads presented to the Table Leaders are included in Section D.

The Table Leaders then participated in a mock standard setting and cut score review using a sample OIB. This sample OIB is included in Section H. During the mock standard setting and cut score review, the Table Leaders practiced all activities that would occur in each round of the BSSP. The Group Leaders acted as Table Leaders during the mock standard setting to demonstrate the type of behavior expected of Table Leaders.

Cut Score Review: Afternoon of Day 1

Opening Session

Staff from the CDE and CTB welcomed the participants to the CSAP Science Cut Score Review. Elizabeth Celva, Director, Unit of Student Assessment, welcomed participants and gave a brief introduction to the week. Ms. Celva also provided a brief overview of the history of the testing program and described the review procedures that would follow the workshop. CTB Research Project Manager Ricardo Mercado then delivered an overview of standard setting and the cut score review process. He also introduced the BSSP to all participants.

The participants were trained on the use of their OIBs and item maps. The training materials are included in Section H. The training overheads from the opening session are included in Section D. Following the opening session, participants proceeded to their breakout rooms. Each grade worked in a separate breakout room.

Taking the Test

Participants spent approximately one hour taking the test for their respective grade.

Study Constructed-Response Items

The Group Leader lead participants in an introduction to the constructed-response items, their scoring guides, and anchor papers, focusing on the knowledge, skills, and abilities required to achieve each score point. Participants referred the scoring guides and anchor papers throughout the workshop as needed.

Discuss Target Student Definitions

A Target Student is defined as a student whose performance is equivalent to the minimum score required for entry into a particular performance level. For their assigned grade, participants examined the performance level descriptors and the Colorado's Standards CSAP Science Assessment Frameworks. The Group Leader in each grade group then lead a discussion of the knowledge, skills, and abilities expected of the Target Students for the *Partially Proficient, Proficient,* and *Advanced* performance levels. The performance level descriptors are included in Section I.

Study Items in the Ordered Item Booklet

For some grades, participants at each table began an examination of each item in the OIB in terms of what the item measures and why it is more difficult than the items preceding it.

Cut Score Review: Day 2

Study Items in the Ordered Item Booklet

Participants completed the examination of the items in the OIB in terms of what each item measures and why it is more difficult than the items preceding it.

Bookmark Training

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 read aloud. The first explanation of bookmark placement demonstrated the mechanics: participants were instructed that all items preceding the bookmark define the knowledge, skills, and abilities that a *just Proficient* 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 content and to make a judgment about the type of content a student would need to know in order to be considered *just Proficient.* The final explanation discussed the relationship between the bookmarks and the scale scores. Participants were also provided a document that explains mastery. The bookmark training overheads are included in Section D, and the training materials are included in Section H.

Participants were then tested on their understanding of bookmark placement with a short check set. The check set questions and the results are presented in Table 10 and Table 11, respectively. Participants were given the correct answers for the check set, as well as explanations of those answers. The check set (and the graphic that appears with it) is included in Section H.

Table 10. Questions in the Check Set that Followed Bookmark Training

	Question
1.	Which items does a student need to master to just make it into the *Partially Proficient* performance level?
2.	If a student mastered only items 1 through 5, in which performance level would this student be?
3.	Suppose a student mastered items 1 through 6. Which performance level is this student in?
4.	For students who are classified as *Partially Proficient*, with at least what likelihood will they be able to answer item 6?
5.	Will the items BEFORE the *Partially Proficient* bookmark be more or less difficult to answer than the items AFTER the bookmark or about the same?

Table 11. Results of the Check Set

	N = 25	
Question	**Count Correct**	**Percent Correct**
1	23	92%
2	24	96%
3	24	96%
4	25	100%
5	25	100%

Preliminary Bookmarks

Once participants demonstrated that they understood bookmark placement through the check set, the preliminary bookmarks were presented to participants.

To calculate the preliminary bookmarks, the percentages of students in each performance level for the 2007 Science assessments were first obtained. The number of students in each performance level and the associated percentages for the 2007 Science tests are shown in Table 12. Cut scores on the 2008 frequency distribution of student scores that most closely yielded the percentages from 2007 shown in Table 12 were then identified. The bookmarks in the OIBs associated with these cut scores were determined; these were the preliminary bookmarks for the 2008 CSAP Science Cut Score Review. The preliminary scale scores and associated bookmarks in the OIBs are shown in Table 13 and Table 14, respectively.

Table 12. N Counts and Percentages in Performance Levels for 2007 CSAP Science

	Unsatisfactory		*Partially Proficient*		*Proficient*		*Advanced*	
	N	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Grade 5	10,908	19.23%	22,054	38.87%	16,223	28.59%	7552	13.31%
Grade 8	11,640	20.20%	15,866	27.53%	25,600	44.42%	4525	7.85%
Grade 10	14,191	25.80%	13,688	24.89%	24,726	44.95%	2398	4.36%

Table 13. Preliminary Scale Scores for the Cut Score Review

	Preliminary Scale Scores		
	Partially Proficient	***Proficient***	***Advanced***
Grade 5	442	512	562
Grade 8	450	501	577
Grade 10	469	508	588

Table 14. Preliminary Bookmarks for the Cut Score Review

	Preliminary Bookmarks		
	Partially Proficient	***Proficient***	***Advanced***
Grade 5	23	53	72
Grade 8	5	21	64
Grade 10	12	31	71

Round 1

After participants examined their preliminary bookmarks, they placed their Round 1 bookmarks for *Partially Proficient*, *Proficient*, and *Advanced*, while keeping in mind the Target Student definitions and Colorado's Standards CSAP Science Assessment Frameworks. Participants were instructed that bookmark placement is always an individual activity.

Cut Score Review: Day 3

Round 2

In Round 2, the Table Leader led a discussion of the bookmarks placed by the participants at the table in Round 1. Participants were instructed to discuss those items for which there was disagreement within the small group; thus, they discussed the range of items between the lowest and highest bookmarks for each performance level.

After this discussion, participants again placed their bookmarks, while keeping in mind the Target Student definitions and Colorado’s Standards CSAP Science Assessment Frameworks. Participants were reminded that bookmark placement is always an individual activity.

Round 3

At the beginning of Round 3 a member of the CTB Standard Setting Team, working with a CDE representative, presented participants with impact data based on their Round 2 bookmark placements. Impact data are the percentages of students who would be classified in each performance level based on the median bookmarks. CTB staff answered process-related questions, and CDE staff answered all policy-related questions concerning the impact data. It was emphasized to the participants that the impact data were being presented as a “reality check.”

For each grade group, the Group Leader facilitated a large-group discussion among the participants of their bookmarks. After discussion, participants again placed bookmarks, while keeping in mind the Target Student definitions and Colorado’s Standards CSAP Science Assessment Frameworks. Participants were reminded that bookmark placement is always an individual activity.

Presentation of Round 3 Results

Participants were shown their final median bookmarks and associated impact data. Table 15 shows the participant-recommended cut scores and associated impact data based on the final round. The impact data shown to participants at the time of the cut score review were based on all student data available at the time of the workshop, which comprised approximately 93% of the data from the Spring 2008 administration of the CSAP Science tests. Detailed results of all rounds are included in Section E. Section F contains estimates of the percentage of students in each performance level at plus and minus one, two, and three standard errors of the participants’ recommended final round cut scores for each grade. Section G contains graphical representations of the participants' final round judgments and the standard errors.

Section K contains two papers: “Calculating a Meaningful Standard Error for the Bookmark Cut Score” and “The Bookmark Standard Setting Procedure: Methodology and Recent Implementations (Lewis, Green, Mitzel, Baum, & Patz, 1998).” These papers are provided for reference.

Table 15. Participant-Recommended Cut Scores and Associated Impact Data Based on the Final Round (Round 3)

	Cut Scores			Impact Data			
Grade	*Partially Proficient*	*Proficient*	*Advanced*	*Unsatis-factory*	*Partially Proficient*	*Proficient*	*Advanced*
5	429	508	569	14.8%	40.8%	33.6%	10.8%
8	459	512	579	23.9%	31.7%	37.0%	7.3%
10	469	507	581	26.0%	24.0%	43.9%	6.1%

Evaluation of the CSAP Science Cut Score Review

Following the presentation of final results, participants were asked to complete an evaluation of the CSAP Sceince Cut Score Review. The evaluation and complete results are included in Section J.

Smoothing Discussion

Table Leaders met for the smoothing discussion. In addition, based on participant feedback, all groups were informed that additional participants could observe the smoothing discussion, if desired. The purpose of this smoothing discussion was to establish a system of cut scores that was well-articulated and, at the same time, considerate of the participants' original recommendations.

The Grade 8 group recommended changing its *Proficient* cut score, from 512 to 507. The group recommended this change to promote better articulation in terms of the impact data across the three grades. In the Grade 8 OIB, the same bookmark placement was associated with a cut score of 507 and 512; that is, the knowledge, skills, and abilities expected of students classified as *Proficient*, as evidenced by the items in the OIB, were not altered by this change.

Table 16 shows the cut scores developed during the smoothing discussions, as well as the associated impact data. The impact data are based on the 93% dataset from the Spring 2008 administration of the CSAP Science tests.

Table 16. Cut Scores and Associated Impact Data after the Smoothing Discussion

	Cut Scores			Impact Data			
Grade	*Partially Proficient*	*Proficient*	*Advanced*	*Unsatis-factory*	*Partially Proficient*	*Proficient*	*Advanced*
5	429	508	569	14.8%	40.8%	33.6%	10.8%
8	459	507	579	23.9%	28.3%	40.5%	7.3%
10	469	507	581	26.0%	24.0%	43.9%	6.1%

Effectiveness of Training

An indication of the effectiveness of training may be found in the participants' answers to statements and questions on the evaluation. The evaluation is included in Section J. Table 17 shows the percentage of participants who agreed or disagreed that they understood how to place a bookmark. Most participants agreed or strongly agreed that they understood how to place their bookmarks. Table 18 summarizes the percentage of participants who agreed or disagreed that bookmark training made the task of bookmark placement clear. Most participants agreed or strongly agreed that the task of bookmark placement was clear. Table 19 summarizes the percentage of participants in each grade who agreed or disagreed that the training materials were helpful. Most participants agreed or strongly agreed that the training materials were helpful. Table 20 shows the percentage of participants who agreed or disagreed that the standard setting facilitator described the Bookmark Procedure well. All participants agreed or strongly agreed that the Bookmark Procedure was well described. The percentage of participants who agreed or disagreed that the goals of the procedure were clear to them is summarized in Table 21. Most participants agreed or strongly agreed that the goals of the standard setting procedure were clear.

Table 17. Participants' Agreement/Disagreement with the Statement, "I understood how to place my bookmarks."

Grade	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Overall	24	0.0%	0.0%	4.2%	33.3%	62.5%
5	9	0.0%	0.0%	0.0%	33.3%	66.7%
8	8	0.0%	0.0%	12.5%	25.0%	62.5%
10	7	0.0%	0.0%	0.0%	42.9%	57.1%

Table 18. 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
Overall	25	0.0%	0.0%	4.0%	28.0%	68.0%
5	9	0.0%	0.0%	11.1%	11.1%	77.8%
8	8	0.0%	0.0%	0.0%	50.0%	50.0%
10	8	0.0%	0.0%	0.0%	25.0%	75.0%

Table 19. Participants' Agreement/Disagreement with the Statement, "The training materials were helpful."

Grade	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Overall	25	0.0%	0.0%	20.0%	28.0%	52.0%
5	9	0.0%	0.0%	11.1%	44.4%	44.4%
8	8	0.0%	0.0%	37.5%	25.0%	37.5%
10	8	0.0%	0.0%	12.5%	12.5%	75.0%

Table 20. Participants' Agreement/Disagreement with the Statement, "The Bookmark Procedure was well described."

Grade	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Overall	25	0.0%	0.0%	0.0%	32.0%	68.0%
5	9	0.0%	0.0%	0.0%	33.3%	66.7%
8	8	0.0%	0.0%	0.0%	37.5%	62.5%
10	8	0.0%	0.0%	0.0%	25.0%	75.0%

Table 21. Participants' Agreement/Disagreement with the Statement, "The goals for the Bookmark Procedure were clear."

Grade	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Overall	25	0.0%	0.0%	4.0%	32.0%	64.0%
5	9	0.0%	0.0%	11.1%	22.2%	66.7%
8	8	0.0%	0.0%	0.0%	50.0%	50.0%
10	8	0.0%	0.0%	0.0%	25.0%	75.0%

Evidence of Perceived Validity

Evidence of perceived validity can be found in the responses of participants to the evaluation. Most participants agreed or strongly agreed that they were satisfied with their final bookmarks as shown in Table 22. Furthermore, all of the participants agreed or strongly agreed that they perceived the Bookmark Procedure to be a fair process, as shown in Table 23.

Table 22. Participants' Agreement/Disagreement with the Statement, "Overall, I am satisfied with my group's final bookmarks."

Grade	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Overall	25	0.0%	0.0%	4.0%	48.0%	48.0%
5	9	0.0%	0.0%	0.0%	55.6%	44.4%
8	8	0.0%	0.0%	12.5%	62.5%	25.0%
10	8	0.0%	0.0%	0.0%	25.0%	75.0%

Table 23. Participants' Agreement/Disagreement with the Statement, "I felt this procedure was fair."

Grade	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Overall	25	0.0%	0.0%	0.0%	48.0%	52.0%
5	9	0.0%	0.0%	0.0%	33.3%	66.7%
8	8	0.0%	0.0%	0.0%	75.0%	25.0%
10	8	0.0%	0.0%	0.0%	37.5%	62.5%

Quality Control Procedures

The CTB Standard Setting Team adheres 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. During the workshop, all participant data were scanned. Any results that appeared to be questionable were further investigated in consultation with the CTB Standard Setting Team and CTB Research staff.

After the Workshop

Grade 10 Science

During the cut score review, the CDE identified an item in Grade 10 Science with possible multiple correct answer choices and decided to suppress the item. After the workshop, the remaining items for Grade 10 Science were recalibrated and updated item parameters were produced. The locations for the remaining items and score points were recalculated, using the updated parameters, and a new OIB was produced.

The ordering of the items and score points in the new Grade 10 OIB was compared to the original ordering used at the time of the cut score review, excluding the suppressed item. The rank-order correlation of the items and score points between the original and new OIB was 0.9997. The majority of items and score points, 82 of 98, did not change rank-order between the original and new OIBs. Table 24 shows the number of items and score points that moved 0, 1, 2, and 4 positions, expressed in absolute value, between the original OIB and the new OIB.

Table 24. Number of Items and Score Points Moving 0, 1, 2, and 4 Positions Between the Original OIB and the New OIB for Grade 10 Science

Number of Item Positions Moved	Number of Items
4	1
2	3
1	12
0	82

The items and score points also were examined to determine which may have changed performance levels when the items were recalibrated. A single item originally had an RP67 location of 580.58, and when recalibrated had an RP67 location of 580.33. These values are very near the *Advanced* cut score of 581. With rounding to a whole number, this item appeared directly after the median *Advanced* bookmark in the original OIB and before the *Advanced* bookmark in the new OIB. However, both before and after recalibration, this item had a location value less than the *Advanced* cut score and was considered a *Proficient* item in both cases.

Final Cut Scores and Impact Data

Table 25 shows the final cut scores and associated impact data. The cut scores are from the recommendations of the smoothing discussion, and the associated impact data are calculated from the complete (100%) data from the Spring 2008 administration of the CSAP Science tests, including the recalibration for Grade 10.

Table 25. Final Cut Scores and Associated Impact Data from the Complete (100%) Data

	Final Cut Scores			Impact Data			
Grade	***Partially Proficient***	***Proficient***	***Advanced***	***Unsatis-factory***	***Partially Proficient***	***Proficient***	***Advanced***
5	429	508	569	12.5%	40.8%	35.3%	11.5%
8	459	507	579	23.2%	28.4%	41.0%	7.4%
10	469	507	581	26.7%	24.0%	43.2%	6.1%

References

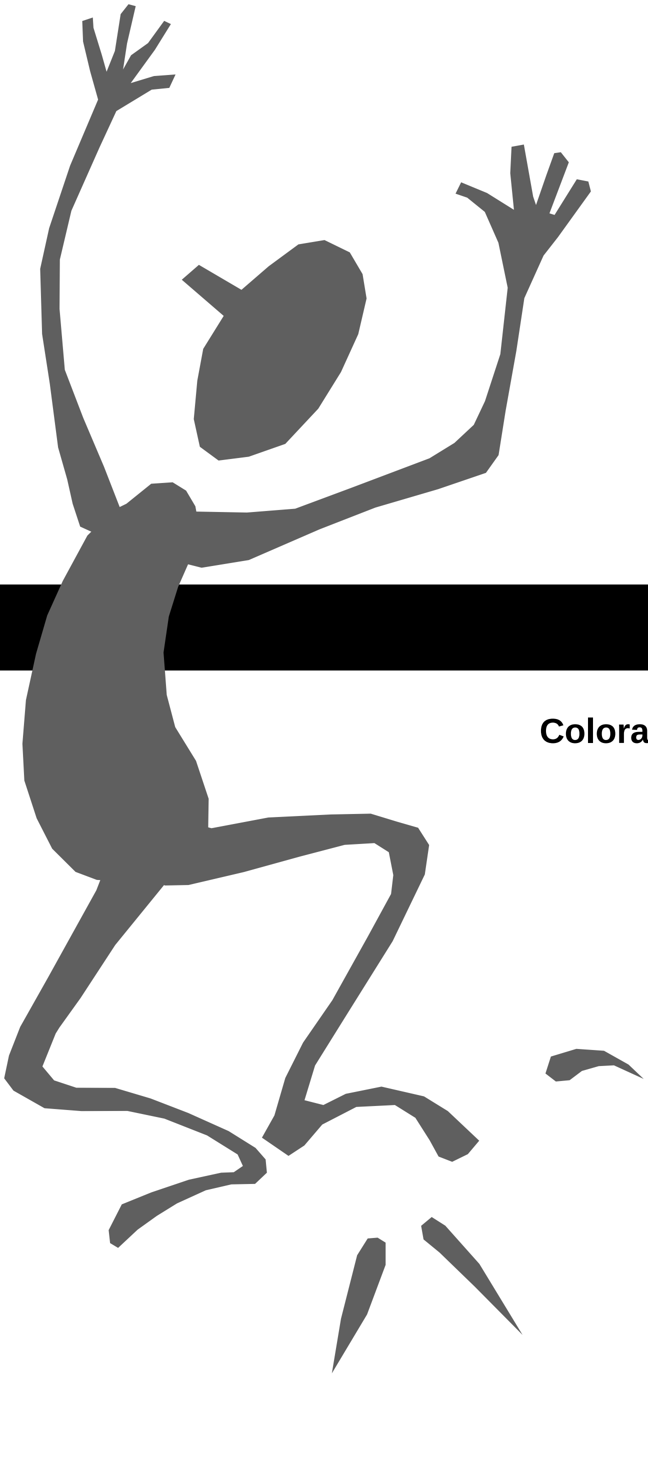
Lewis, D.M., Mitzel, H.C., & Green, D.R. (1996). Standard Setting: A bookmark approach. In D.R. Green (Chair), *IRT-based standard-setting procedures utilizing behavioral anchoring*. Symposium conducted at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix,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

Section C

Master Agenda

CTB McGraw-Hill

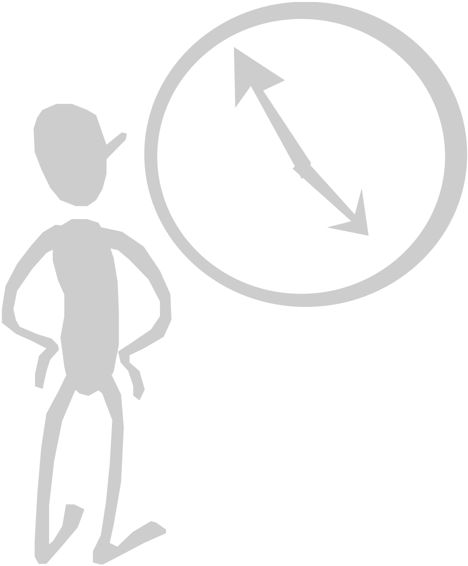


Master Agenda

Colorado Student Assessment Program
Grades 5, 8, and 10 Science

Cut Score Review

May 14–16, 2008
Denver, Colorado



Welcome to the cut score review workshop
for the Colorado Student Assessment Program
for Grades 5, 8, and 10 Science!

The Colorado 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.

Wednesday, May 14

Welcome! Table Leader Training

8:00 AM **Table Leader registration & continental breakfast**
Please check in at the reception area to sign a non-disclosure agreement, get your nametag, and collect any other information.

8:30 AM **Table Leader training**
You are presented an overview of the Colorado Student Assessment Program, learn about the cut score review process, and discuss your role and responsibilities during the workshop.

10:30 AM **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.

12:00 PM **Lunch for Table Leaders**
Table Leaders break for a one-hour lunch .

*****Throughout the workshop, morning and afternoon breaks will occur about 10 AM and 2 PM.*****

12:30 PM **Participant registration**
Participants check in at the reception table. Table Leaders need not register again.

1:00 PM **Opening session**
All participants are formally welcomed and receive an overview of the Colorado Student Assessment Program. Participants are introduced to the cut score review process. After this session, participants break into their assigned tables in their breakout rooms.

2:30 PM **Take the operational test**
Participants sign out secure materials. Participants take the operational test under conditions similar to those experienced by students.

- Ensure that all participants at your table write their name on ***each*** piece of their secure materials. Secure materials are printed on colored paper.
- Although some discussion about individual test items is normal, focus your participants away from prolonged debate and toward taking the test.
- Use the provided index cards to record comments about test items.

3:30 PM **Study constructed-response items**
The Group Leader leads an examination of the constructed-response items, scoring rubrics, and anchor papers, focusing on the knowledge, skills, and abilities required to achieve each score point.

4:00 PM **Target Student discussion**
The group discusses the knowledge, skills, and abilities expected of students in each performance level.

4:45 PM **Secure materials collection**
The Group Leader facilitates collection of the test materials from all participants.

- The Table Leaders supervise the collection of secure materials at their tables. See the “Secure Materials” page in this agenda for more information.

5:00 PM **Secure materials audit**
Table Leaders audit materials at one other table. After all secure materials are accounted for, participants are dismissed by the Group Leader,

5:15 PM **Table Leader debrief**
Table Leaders discuss the events of the day and plans for the next day with the Group Leader.

5:30 PM **Table Leaders dismissed**

Thursday, May 15
Discussion & Bookmark Ratings

8:00 AM **Continental breakfast**
Continental breakfast is served.

8:30 AM **Begin discussion of each item in the Ordered Item Booklet (OIB)**
Facilitate a discussion among everyone at your table of each item 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. Participants record these details on their Item Maps.

- Remember to use the index cards, as necessary.
- Ensure that each participant at your table has a chance to speak.

12:00 PM **Lunch**
The group breaks for a one-hour lunch.

1:00 PM **Complete discussion of each item in the Ordered Item Booklet (OIB)**
Groups continue the discussion of each of the items in the OIB.

- Remember to use the index cards, as necessary.
- Ensure each participant at your table has a chance to speak.

3:00 PM **Orientation to bookmark placement and Round 1 ratings**
A member of the CTB Standard Setting Team introduces bookmark placement, explaining how bookmarks are placed and what bookmarks mean. After this presentation, a short checkset is given, followed by Round 1 bookmark placement. Participants are provided an explanation of how the preliminary cut scores were determined and how participants may adjust them to align better with the Colorado Standards.

- See the bookmark training materials for more info.
- Remind your participants that bookmark placement is always an individual activity.
- Collect your participants' rating forms as they complete them, ensuring that each participant has made a single, unambiguous rating for each bookmark.
- Fill out your orange sheet and begin Round 2 discussions.
- Give your participants' rating forms to the Group Leader.

4:45 PM **Secure materials collection**
The Group Leader facilitates collection of the test materials from all participants.

- The Table Leaders supervise the collection of secure materials at their tables. See the "Secure Materials" page in this agenda for more information.

5:00 PM **Secure materials audit**
Table Leaders audit materials at one other table. After all secure materials are accounted for, participants are dismissed by the Group Leader.

5:15 PM **Table Leader debrief**
Table Leaders discuss the events of the day and plans for the next day with the Group Leader.

5:30 PM **Table Leaders dismissed**

Friday, May 16
Bookmark Ratings

8:00 AM **Continental breakfast**
Continental breakfast is served.

8:30 AM **Discuss Round 1 as a table**
Use the orange sheet to lead a discussion about the ratings made at your table.

10:00 AM **Round 2 ratings**
After your Round 1 discussion, begin Round 2 bookmark placement.

- Remind your participants that bookmark placement is always an individual activity.

10:45 AM **Discuss Round 2 as a large group**
The Group Leader presents a summary of the voting from each table to the entire group. Afterwards, s/he leads a discussion with the entire group of each bookmark, similar to the table-level discussions of Round 2.

12:00 PM **Lunch**
The group breaks for a one-hour lunch.

1:00 PM **Round 3 ratings**
The Group Leader directs all participants to make their Round 3 bookmark placements.

- Remind your participants that bookmark placement is always an individual activity.
- Collect your participants' rating forms as they complete them.
- You need *not* complete another orange sheet.

2:00 PM **Presentation of final recommendations**
A summary of the Round 3 voting is presented to the entire group.

2:30 PM **Smoothing discussion for Table Leaders**
Table Leaders examine the cross-grade data for smoothing purposes.

3:30 PM **Evaluations**
Each participant completes an evaluation of the cut score review.

3:45 PM **Secure materials collection**
The Group Leader facilitates collection of the test materials from all participants.

- The Table Leaders supervise the collection of secure materials at their tables. See the "Secure Materials" page in this agenda for more information.

3:55 PM **Secure materials audit**
Table Leaders audit materials at one other table. After all secure materials are accounted for, participants are dismissed by the Group Leader,

4:00 PM **Table Leaders dismissed**

The Colorado Department of Education and CTB 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 effectively gather all 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.

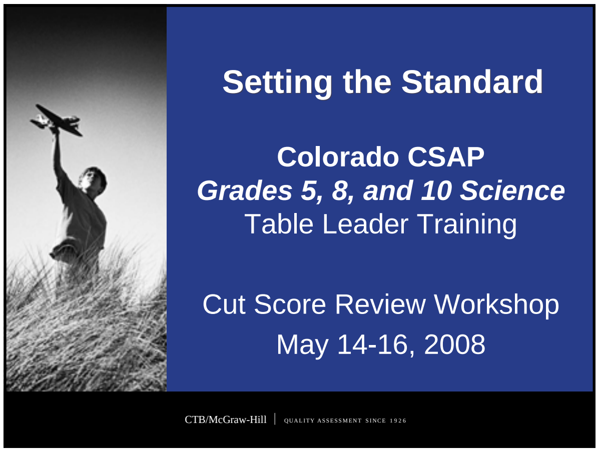
7. The collected materials are stored overnight and will be available in the morning.

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

Training Overheads



Setting the Standard

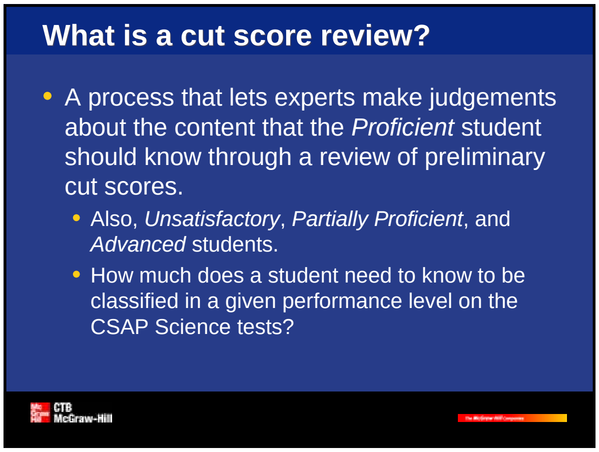
Colorado CSAP
Grades 5, 8, and 10 Science
Table Leader Training

Cut Score Review Workshop
May 14-16, 2008

CTB/McGraw-Hill | QUALITY ASSESSMENT SINCE 1926

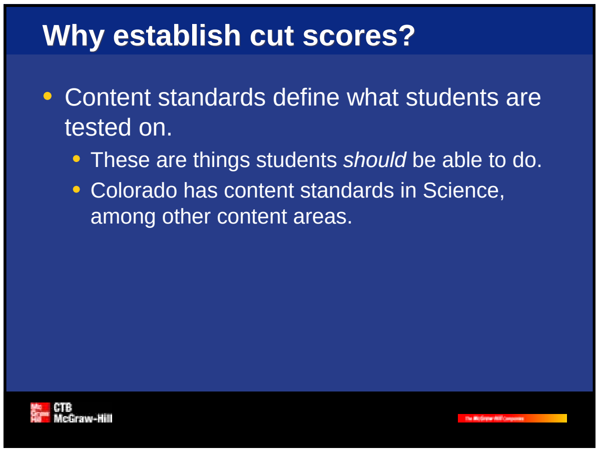
CTB Standard Setting Team

- Rick Mercado
- Thakur Karkee
- Adele Brandstrom
- Dennis Allion
- Cynthia Fischer
- Angelica Gordon
- Marie-Lise Bouscaren
- Bevin Flaherty
- Andrina Ortiz



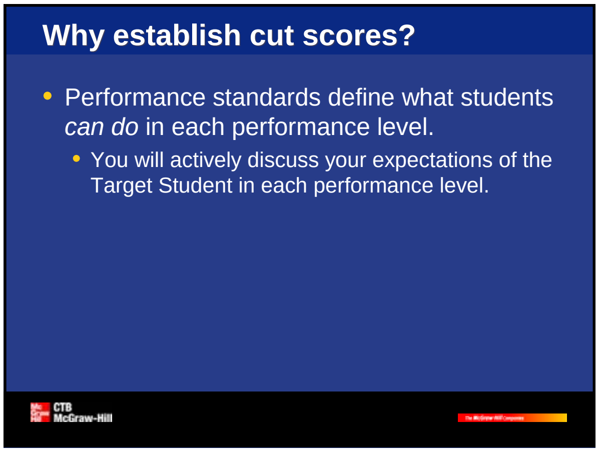
What is a cut score review?

- A process that lets experts make judgements about the content that the *Proficient* student should know through a review of preliminary cut scores.
  - Also, *Unsatisfactory*, *Partially Proficient*, and *Advanced* students.
  - How much does a student need to know to be classified in a given performance level on the CSAP Science tests?



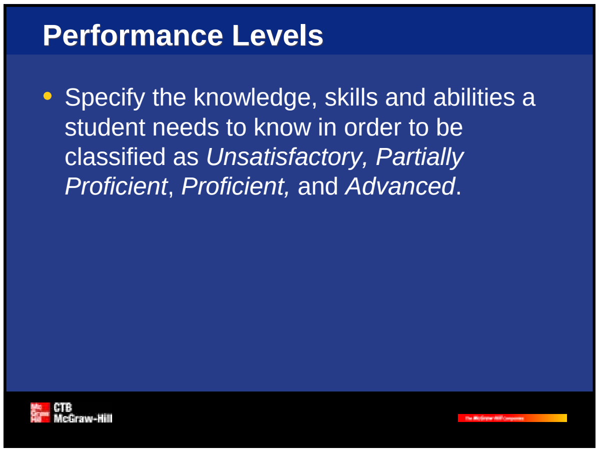
Why establish cut scores?

- Content standards define what students are tested on.
  - These are things students *should* be able to do.
  - Colorado has content standards in Science, among other content areas.



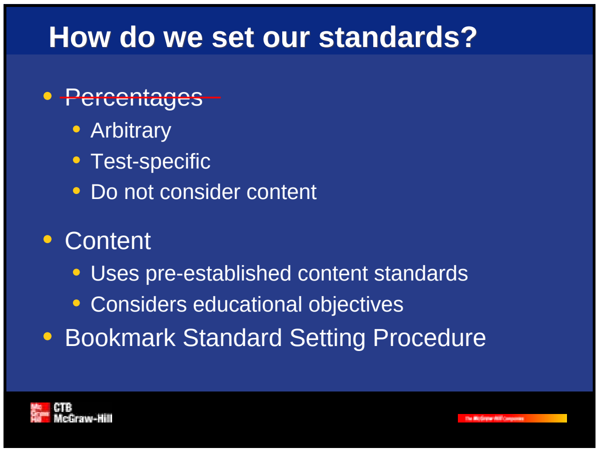
Why establish cut scores?

- Performance standards define what students *can do* in each performance level.
  - You will actively discuss your expectations of the Target Student in each performance level.



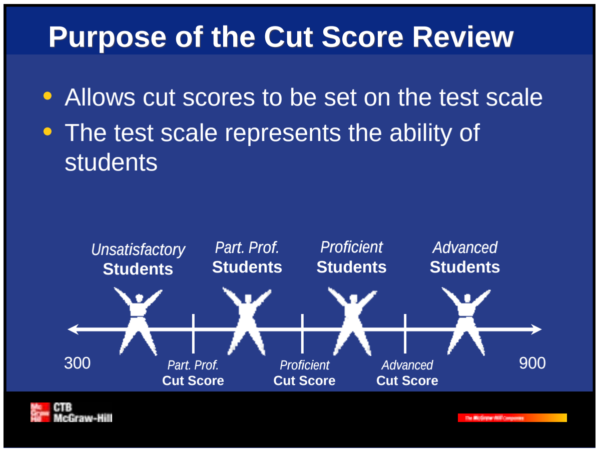
Performance Levels

- Specify the knowledge, skills and abilities a student needs to know in order to be classified as *Unsatisfactory, Partially Proficient*, *Proficient,* and *Advanced*.



How do we set our standards?

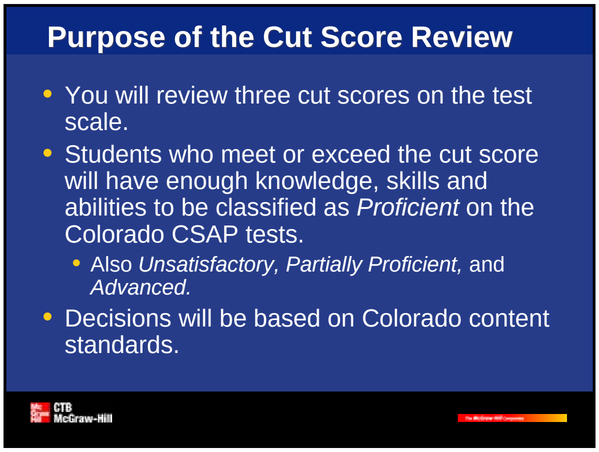
- ~~Percentages~~
  - Arbitrary
  - Test-specific
  - Do not consider content
- Content
  - Uses pre-established content standards
  - Considers educational objectives
- Bookmark Standard Setting Procedure



Purpose of the Cut Score Review

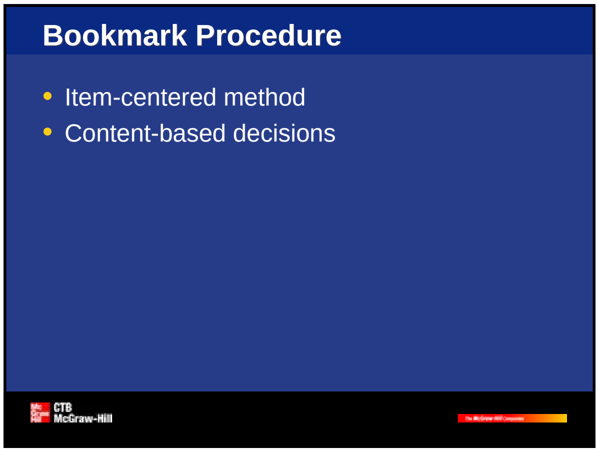
- Allows cut scores to be set on the test scale
- The test scale represents the ability of students

Unsatisfactory **Students**
Part. Prof. **Students**
Proficient **Students**
Advanced **Students**
300
Part. Prof. **Cut Score**
Proficient **Cut Score**
Advanced **Cut Score**
900



Purpose of the Cut Score Review

- You will review three cut scores on the test scale.
- Students who meet or exceed the cut score will have enough knowledge, skills and abilities to be classified as *Proficient* on the Colorado CSAP tests.
  - Also *Unsatisfactory, Partially Proficient,* and *Advanced.*
- Decisions will be based on Colorado content standards.



Bookmark Procedure

- Item-centered method
- Content-based decisions

Committee Roles

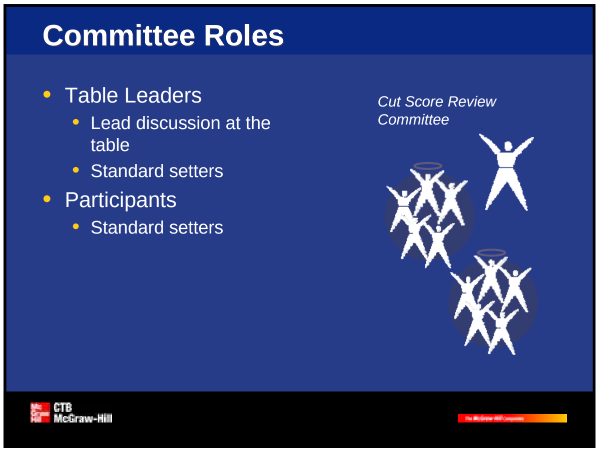
- Group Leaders
- Table Leaders
- Participants
- CDE
- CTB

Cut Score Review Committee

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

Cut Score Review Committee

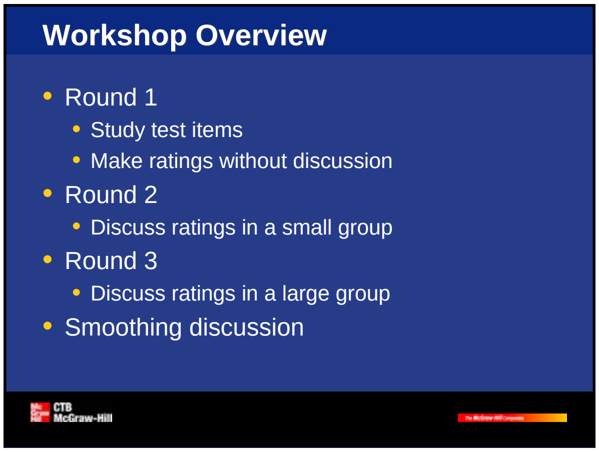


Committee Roles

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

Cut Score Review Committee

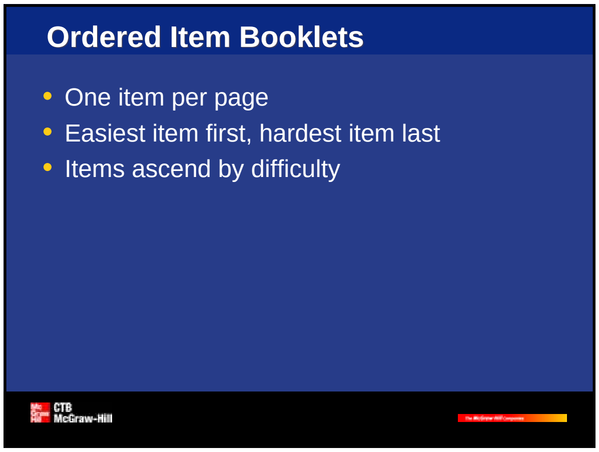
CTB McGraw-Hill



Workshop Overview

- Round 1
  - Study test items
  - Make ratings without discussion
- Round 2
  - Discuss ratings in a small group
- Round 3
  - Discuss ratings in a large group
- Smoothing discussion

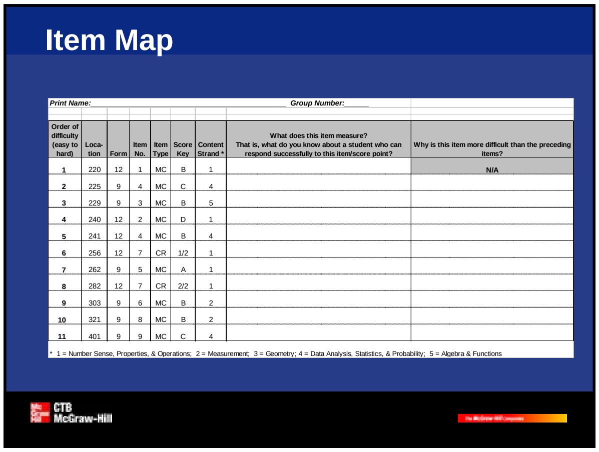
CTB McGraw-Hill



Ordered Item Booklets

- One item per page
- Easiest item first, hardest item last
- Items ascend by difficulty

CTB McGraw-Hill

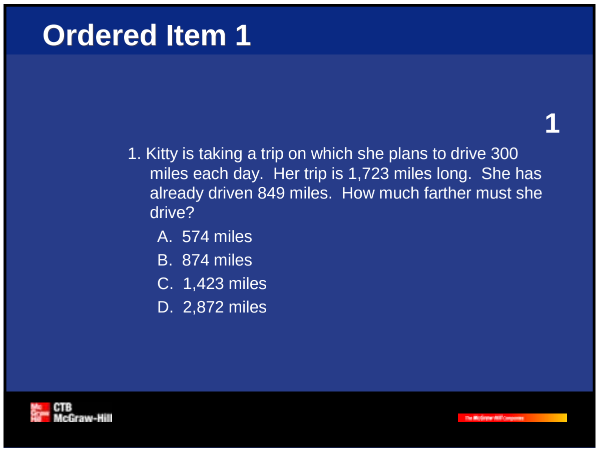


Item Map

Print Name:							Group Number:______	
Order of difficulty (easy to hard)	Loca- tion	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/score point?	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	256	12	7	CR	1/2	1		
7	262	9	5	MC	A	1		
8	282	12	7	CR	2/2	1		
9	303	9	6	MC	B	2		
10	321	9	8	MC	B	2		
11	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

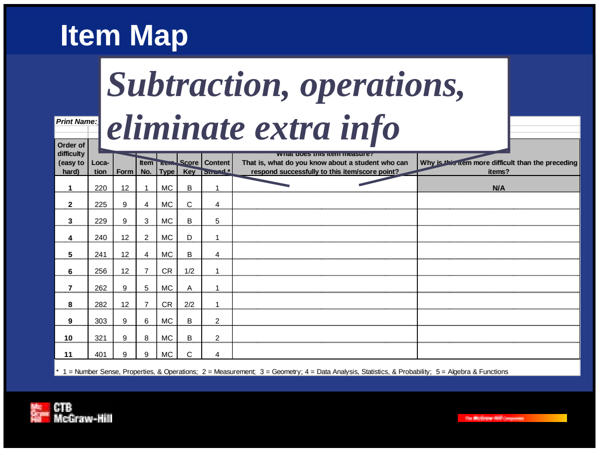


Ordered Item 1

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



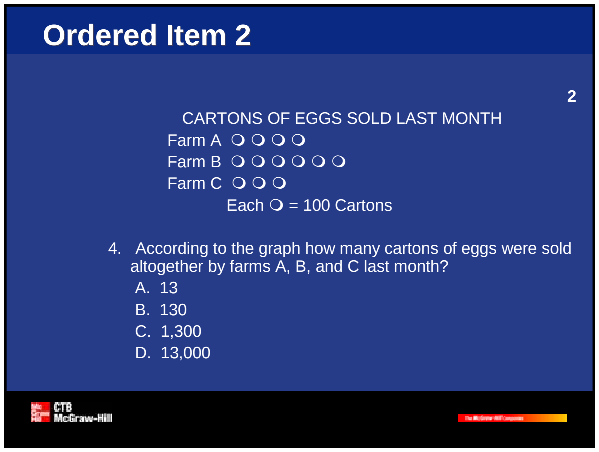
Item Map

Subtraction, operations, eliminate extra info

Print Name:								
Order of difficulty (easy to hard)	Loca- tion	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/score point?	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	256	12	7	CR	1/2	1		
7	262	9	5	MC	A	1		
8	282	12	7	CR	2/2	1		
9	303	9	6	MC	B	2		
10	321	9	8	MC	B	2		
11	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



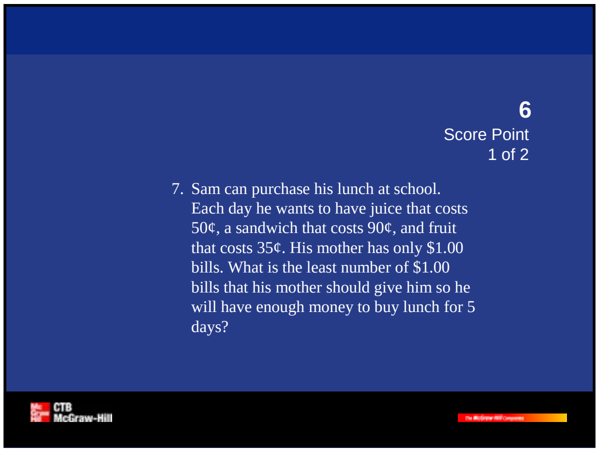
Ordered Item 2

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

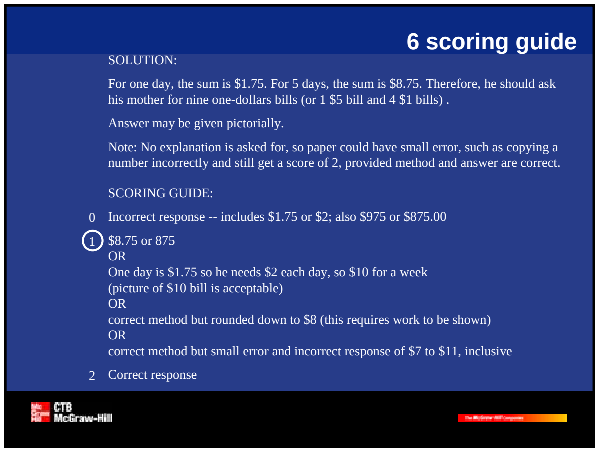
CTB McGraw-Hill



6
Score Point
1 of 2

7. Sam can purchase his lunch at school. Each day he wants to have juice that costs 50¢, a sandwich that costs 90¢, and fruit that costs 35¢. His mother has only $1.00 bills. What is the least number of $1.00 bills that his mother should give him so he will have enough money to buy lunch for 5 days?

CTB McGraw-Hill



6 scoring guide

SOLUTION:

For one day, the sum is $1.75. For 5 days, the sum is $8.75. Therefore, he should ask his mother for nine one-dollars bills (or 1 $5 bill and 4 $1 bills) .

Answer may be given pictorially.

Note: No explanation is asked for, so paper could have small error, such as copying a number incorrectly and still get a score of 2, provided method and answer are correct.

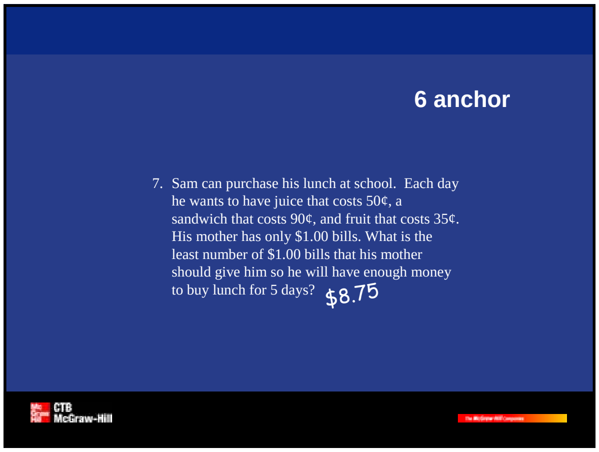
SCORING GUIDE:

0 Incorrect response -- includes $1.75 or $2; also $975 or $875.00

[✓ 1] $8.75 or 875
OR
One day is $1.75 so he needs $2 each day, so $10 for a week
(picture of $10 bill is acceptable)
OR
correct method but rounded down to $8 (this requires work to be shown)
OR
correct method but small error and incorrect response of $7 to $11, inclusive

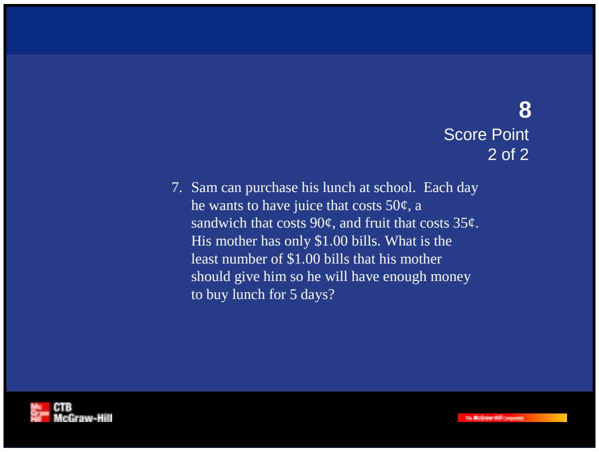
2 Correct response

CTB McGraw-Hill



6 anchor

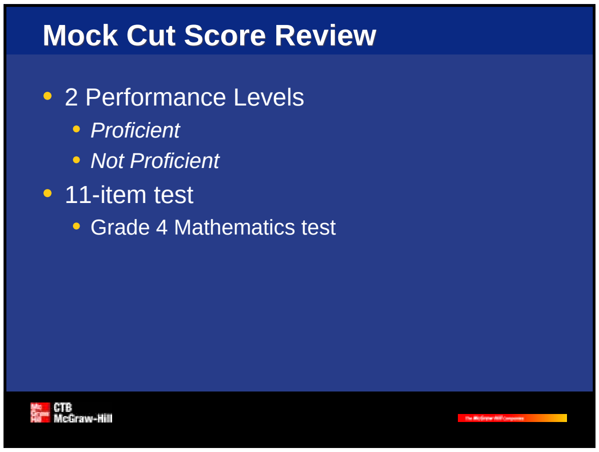
7. Sam can purchase his lunch at school. Each day he wants to have juice that costs 50¢, a sandwich that costs 90¢, and fruit that costs 35¢. His mother has only $1.00 bills. What is the least number of $1.00 bills that his mother should give him so he will have enough money to buy lunch for 5 days? $8.75



8

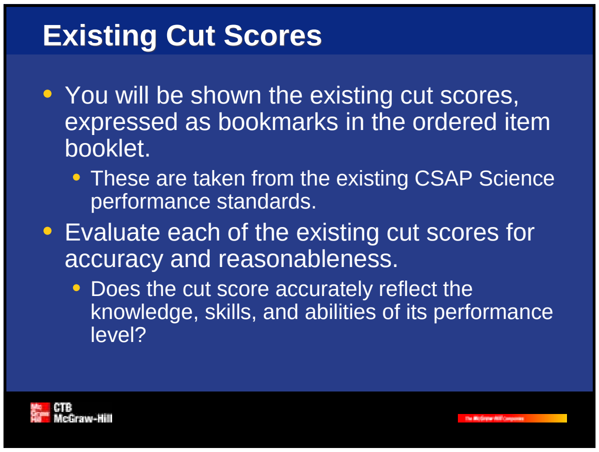
Score Point
2 of 2

7. Sam can purchase his lunch at school. Each day he wants to have juice that costs 50¢, a sandwich that costs 90¢, and fruit that costs 35¢. His mother has only $1.00 bills. What is the least number of $1.00 bills that his mother should give him so he will have enough money to buy lunch for 5 days?



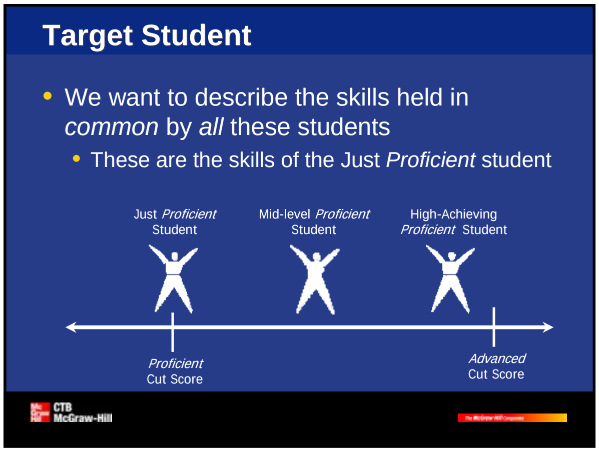
Mock Cut Score Review

- 2 Performance Levels
  - *Proficient*
  - *Not Proficient*
- 11-item test
  - Grade 4 Mathematics test



Existing Cut Scores

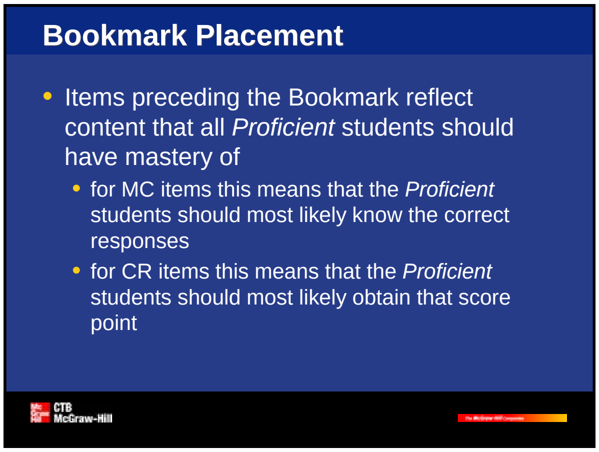
- You will be shown the existing cut scores, expressed as bookmarks in the ordered item booklet.
  - These are taken from the existing CSAP Science performance standards.
- Evaluate each of the existing cut scores for accuracy and reasonableness.
  - Does the cut score accurately reflect the knowledge, skills, and abilities of its performance level?



Target Student

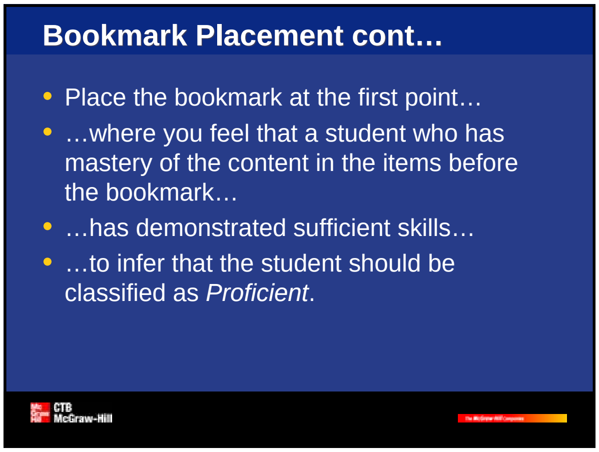
- We want to describe the skills held in *common* by *all* these students
  - These are the skills of the Just *Proficient* student

Just *Proficient* Student
Mid-level *Proficient* Student
High-Achieving *Proficient* Student
Proficient Cut Score
Advanced Cut Score



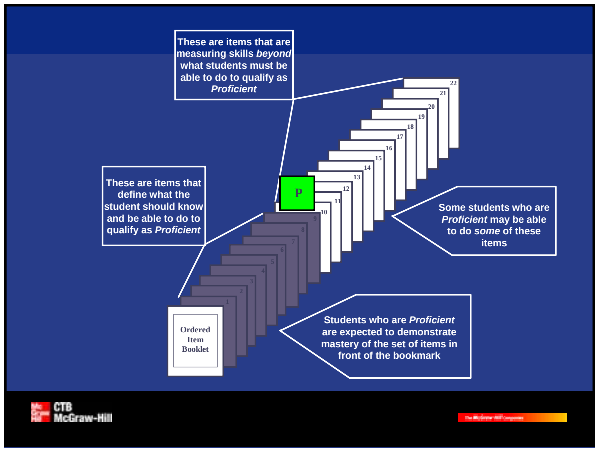
Bookmark Placement

- Items preceding the Bookmark reflect content that all *Proficient* students should have mastery of
  - for MC items this means that the *Proficient* students should most likely know the correct responses
  - for CR items this means that the *Proficient* students should most likely obtain that score point



Bookmark Placement cont…

- Place the bookmark at the first point…
- …where you feel that a student who has mastery of the content in the items before the bookmark…
- …has demonstrated sufficient skills…
- …to infer that the student should be classified as *Proficient*.

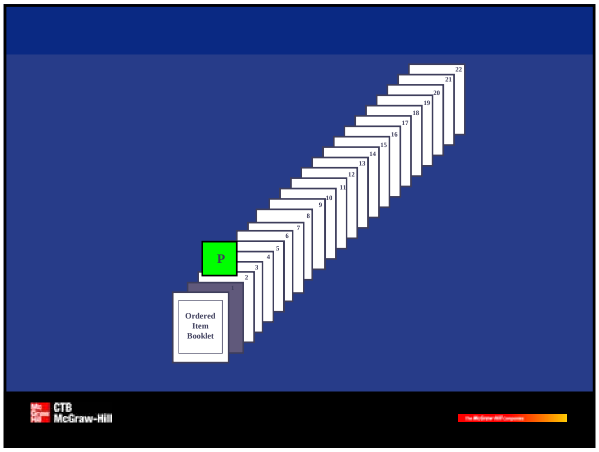


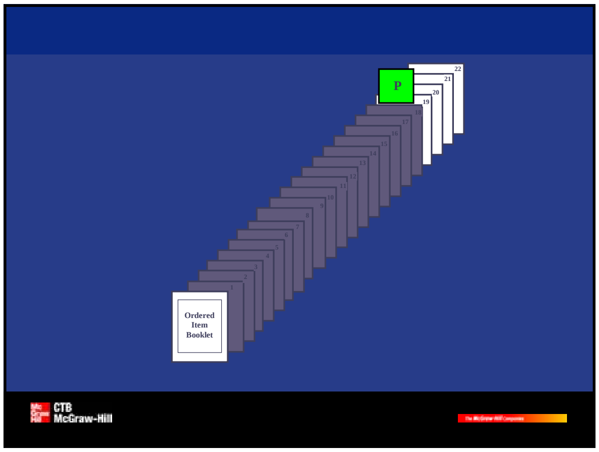
These are items that are measuring skills *beyond* what students must be able to do to qualify as *Proficient*

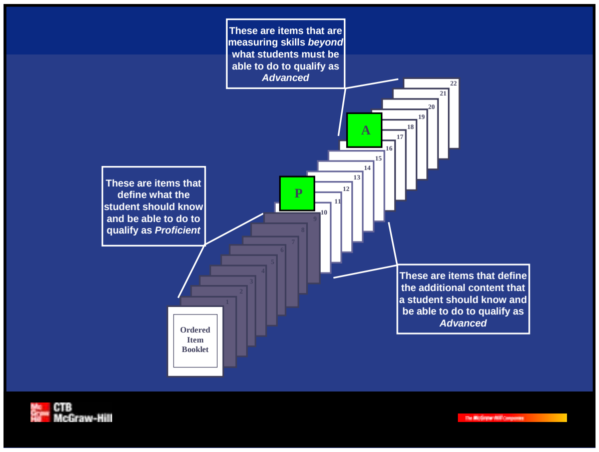
These are items that define what the student should know and be able to do to qualify as *Proficient*

Some students who are *Proficient* may be able to do *some* of these items

Students who are *Proficient* are expected to demonstrate mastery of the set of items in front of the bookmark

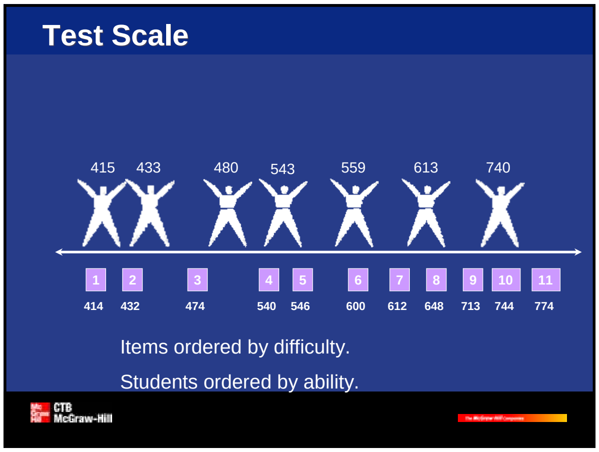





These are items that are measuring skills *beyond* what students must be able to do to qualify as *Advanced*

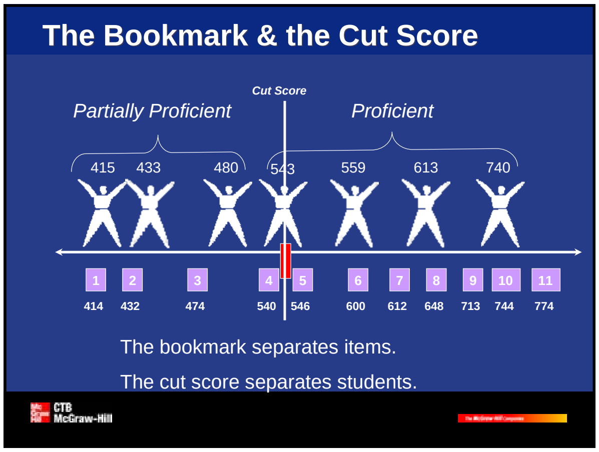
These are items that define what the student should know and be able to do to qualify as *Proficient*

These are items that define the additional content that a student should know and be able to do to qualify as *Advanced*


Test Scale

Items ordered by difficulty.

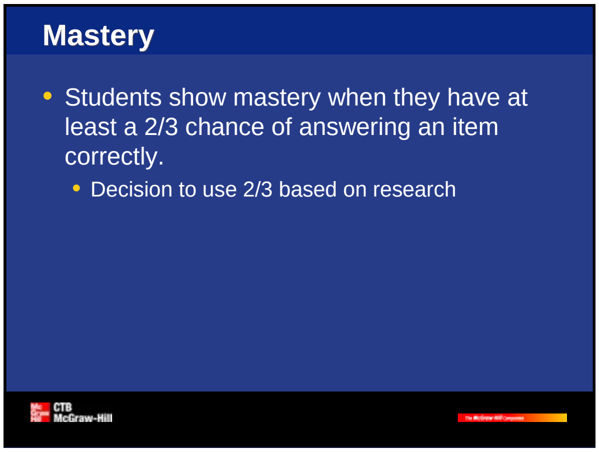
Students ordered by ability.



The Bookmark & the Cut Score

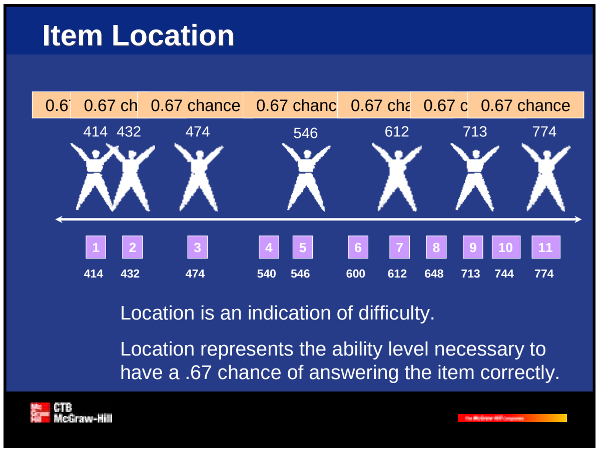
The bookmark separates items.

The cut score separates students.



Mastery

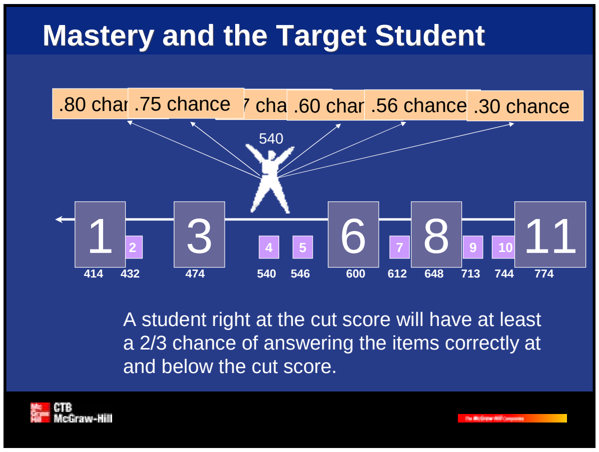
- Students show mastery when they have at least a 2/3 chance of answering an item correctly.
  - Decision to use 2/3 based on research



Item Location

Location is an indication of difficulty.

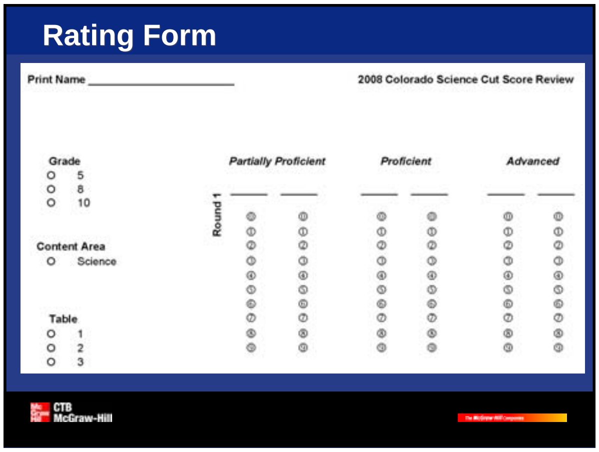
Location represents the ability level necessary to have a .67 chance of answering the item correctly.



Mastery and the Target Student

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

CTB McGraw-Hill



Rating Form

Print Name ____________________

2008 Colorado Science Cut Score Review

Grade
○ 5
○ 8
○ 10

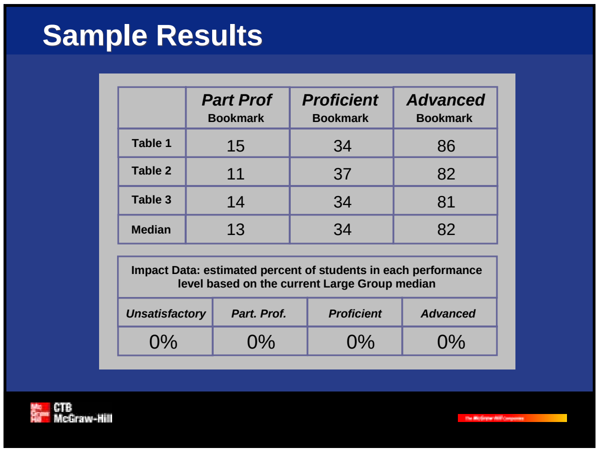
Content Area
○ Science

Table
○ 1
○ 2
○ 3

Round 1

Partially Proficient		Proficient		Advanced	
___	___	___	___	___	___
⓪	⓪	⓪	⓪	⓪	⓪
①	①	①	①	①	①
②	②	②	②	②	②
③	③	③	③	③	③
④	④	④	④	④	④
⑤	⑤	⑤	⑤	⑤	⑤
⑥	⑥	⑥	⑥	⑥	⑥
⑦	⑦	⑦	⑦	⑦	⑦
⑧	⑧	⑧	⑧	⑧	⑧
⑨	⑨	⑨	⑨	⑨	⑨

CTB McGraw-Hill



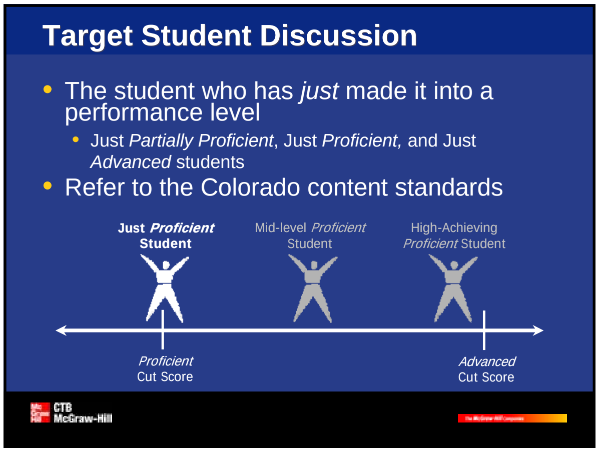
Sample Results

	Part Prof Bookmark	*Proficient* Bookmark	*Advanced* Bookmark
Table 1	15	34	86
Table 2	11	37	82
Table 3	14	34	81
Median	13	34	82

Impact Data: estimated percent of students in each performance level based on the current Large Group median

Unsatisfactory	*Part. Prof.*	*Proficient*	*Advanced*
0%	0%	0%	0%

CTB McGraw-Hill



Target Student Discussion

- The student who has *just* made it into a performance level
  - Just *Partially Proficient*, Just *Proficient,* and Just *Advanced* students
- Refer to the Colorado content standards

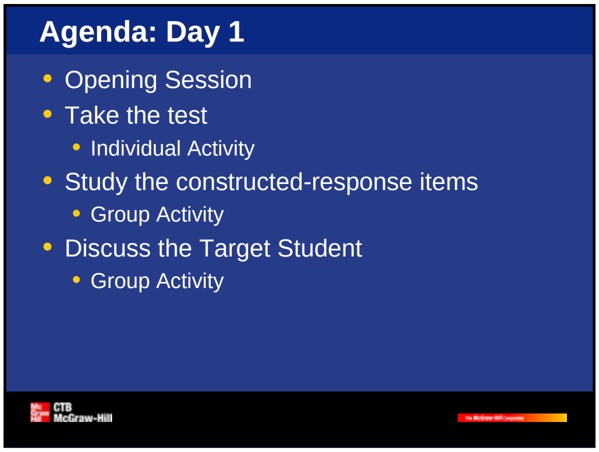
Just *Proficient* Student

Mid-level *Proficient* Student

High-Achieving *Proficient* Student

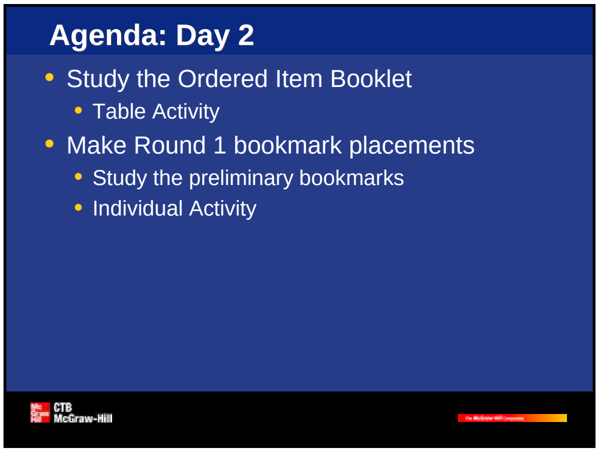
Proficient Cut Score

Advanced Cut Score



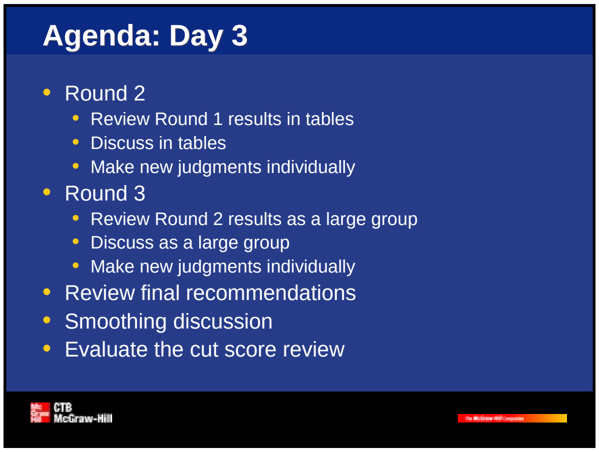
Agenda: Day 1

- Opening Session
- Take the test
  - Individual Activity
- Study the constructed-response items
  - Group Activity
- Discuss the Target Student
  - Group Activity



Agenda: Day 2

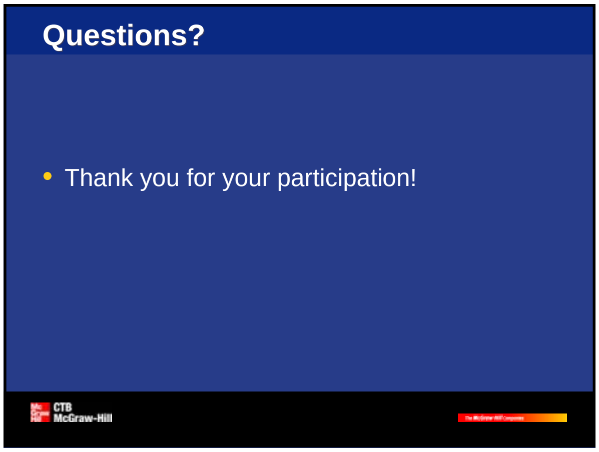
- Study the Ordered Item Booklet
  - Table Activity
- Make Round 1 bookmark placements
  - Study the preliminary bookmarks
  - Individual Activity



Agenda: Day 3

- Round 2
  - Review Round 1 results in tables
  - Discuss in tables
  - Make new judgments individually
- Round 3
  - Review Round 2 results as a large group
  - Discuss as a large group
  - Make new judgments individually
- Review final recommendations
- Smoothing discussion
- Evaluate the cut score review

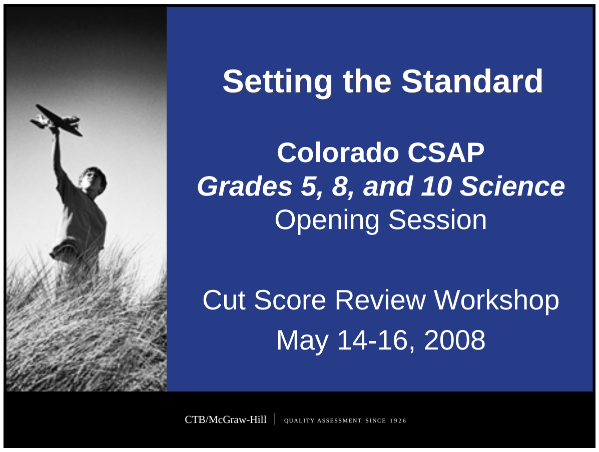
CTB McGraw-Hill



Questions?

- Thank you for your participation!

CTB McGraw-Hill



Setting the Standard

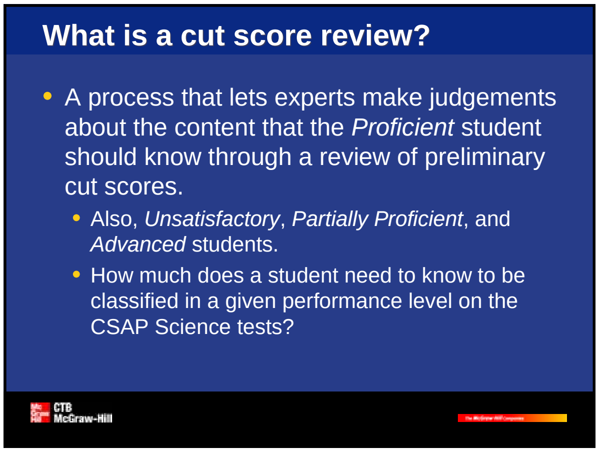
Colorado CSAP
Grades 5, 8, and 10 Science
Opening Session

Cut Score Review Workshop
May 14-16, 2008

CTB/McGraw-Hill | QUALITY ASSESSMENT SINCE 1926

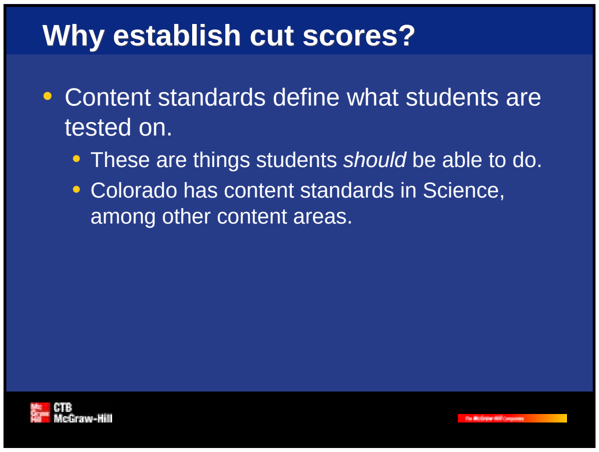
CTB Standard Setting Team

- Rick Mercado
- Thakur Karkee
- Adele Brandstrom
- Dennis Allion
- Cynthia Fischer
- Angelica Gordon
- Marie-Lise Bouscaren
- Bevin Flaherty
- Andrina Ortiz



What is a cut score review?

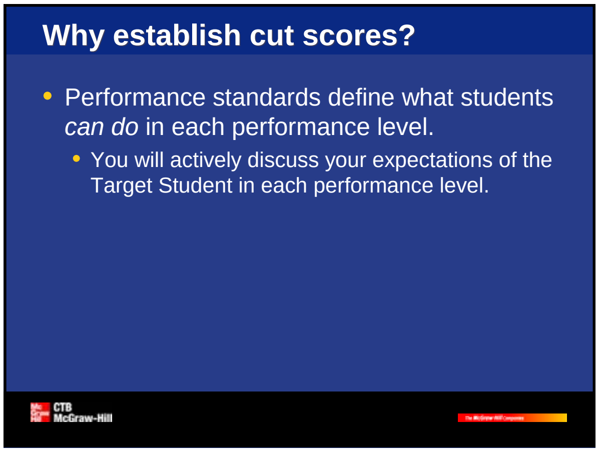
- A process that lets experts make judgements about the content that the *Proficient* student should know through a review of preliminary cut scores.
  - Also, *Unsatisfactory*, *Partially Proficient*, and *Advanced* students.
  - How much does a student need to know to be classified in a given performance level on the CSAP Science tests?



Why establish cut scores?

- Content standards define what students are tested on.
  - These are things students *should* be able to do.
  - Colorado has content standards in Science, among other content areas.

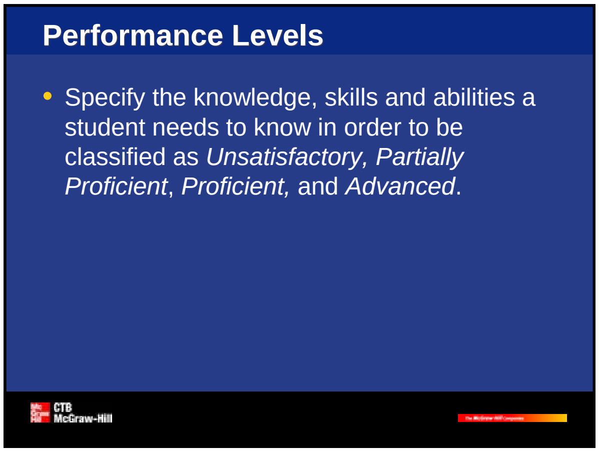
CTB McGraw-Hill



Why establish cut scores?

- Performance standards define what students *can do* in each performance level.
  - You will actively discuss your expectations of the Target Student in each performance level.

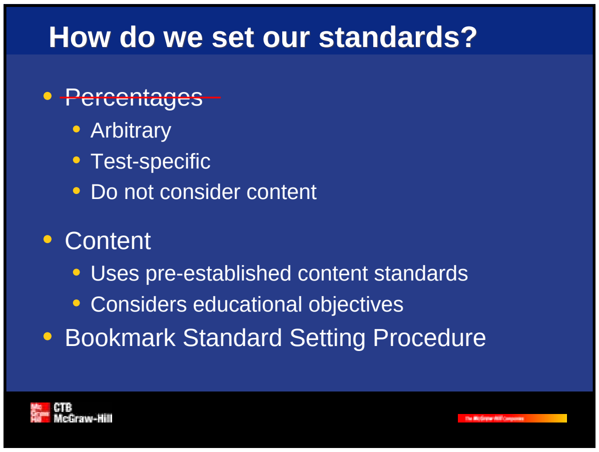
CTB McGraw-Hill



Performance Levels

- Specify the knowledge, skills and abilities a student needs to know in order to be classified as *Unsatisfactory, Partially Proficient*, *Proficient,* and *Advanced*.

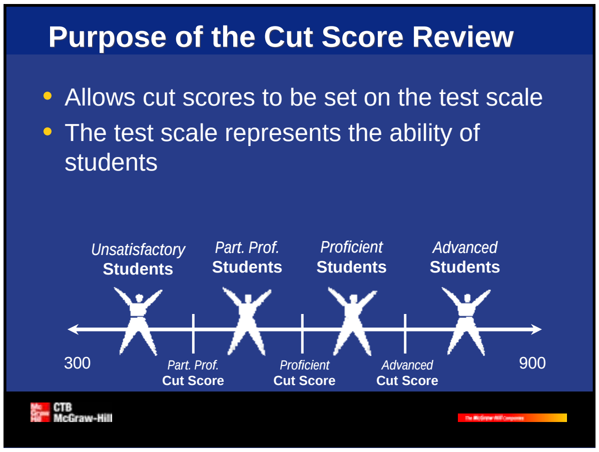
CTB McGraw-Hill



How do we set our standards?

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

CTB McGraw-Hill

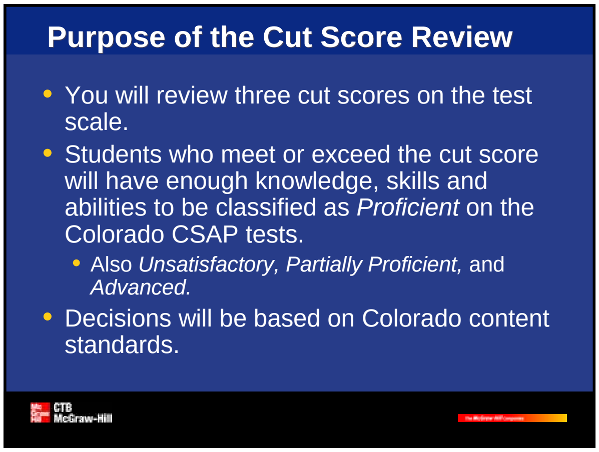


Purpose of the Cut Score Review

- Allows cut scores to be set on the test scale
- The test scale represents the ability of students

Unsatisfactory **Students**
Part. Prof. **Students**
Proficient **Students**
Advanced **Students**
300
Part. Prof. **Cut Score**
Proficient **Cut Score**
Advanced **Cut Score**
900

CTB McGraw-Hill



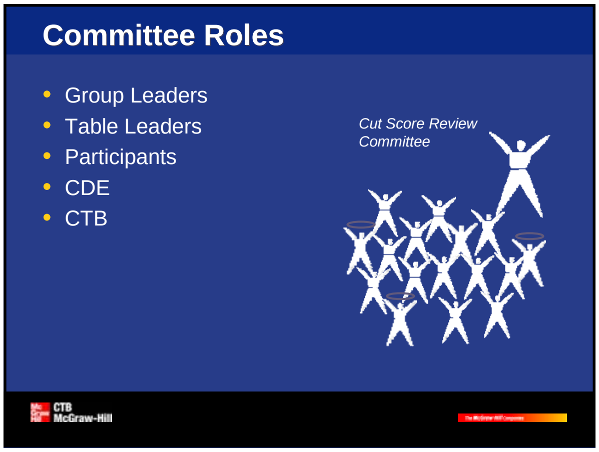
Purpose of the Cut Score Review

- You will review three cut scores on the test scale.
- Students who meet or exceed the cut score will have enough knowledge, skills and abilities to be classified as *Proficient* on the Colorado CSAP tests.
  - Also *Unsatisfactory, Partially Proficient,* and *Advanced.*
- Decisions will be based on Colorado content standards.

CTB McGraw-Hill

Bookmark Procedure

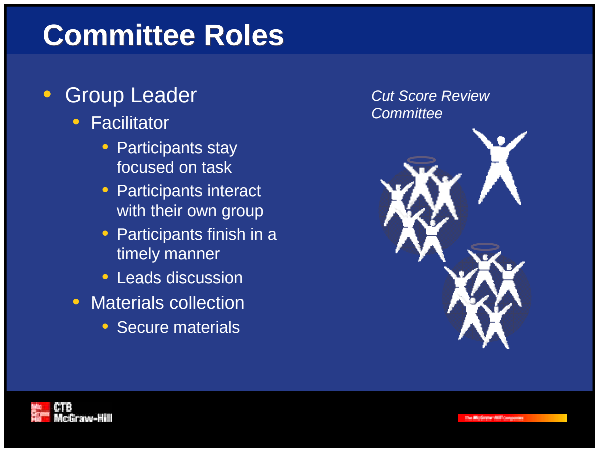
- Item-centered method
- Content-based decisions



Committee Roles

- Group Leaders
- Table Leaders
- Participants
- CDE
- CTB

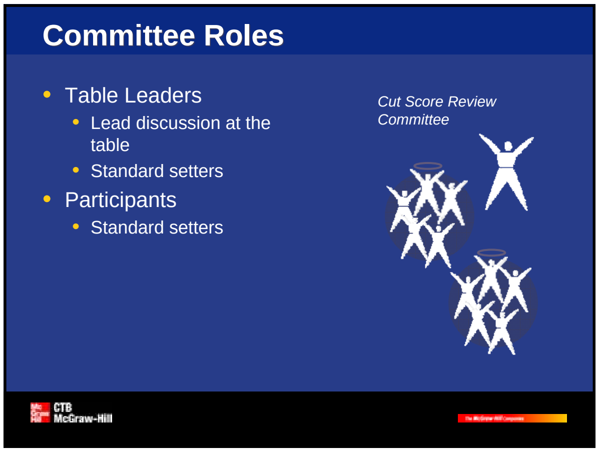
Cut Score Review Committee



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

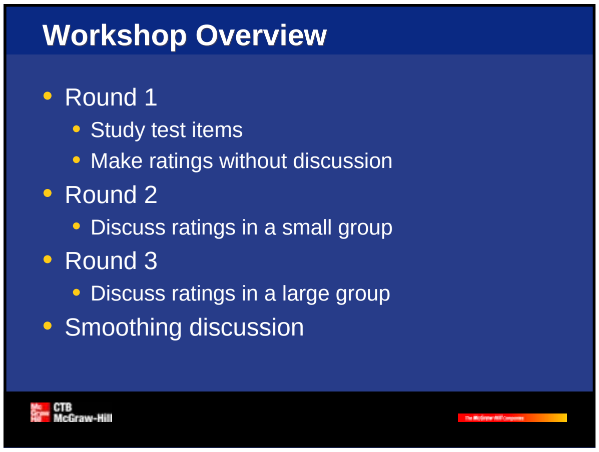
Cut Score Review Committee



Committee Roles

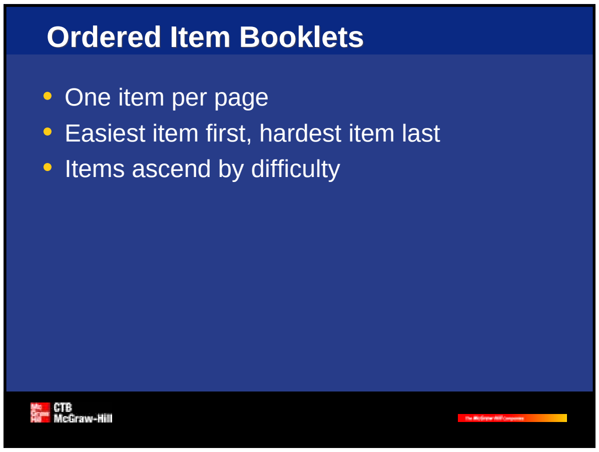
- Table Leaders
  - Lead discussion at the table
  - Standard setters
- Participants
  - Standard setters

Cut Score Review Committee



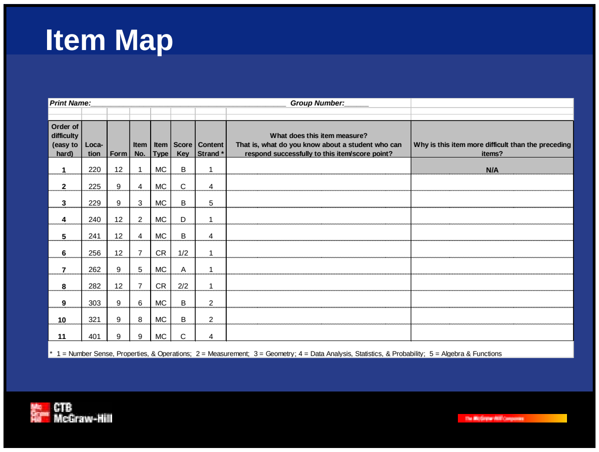
Workshop Overview

- Round 1
  - Study test items
  - Make ratings without discussion
- Round 2
  - Discuss ratings in a small group
- Round 3
  - Discuss ratings in a large group
- Smoothing discussion



Ordered Item Booklets

- One item per page
- Easiest item first, hardest item last
- Items ascend by difficulty



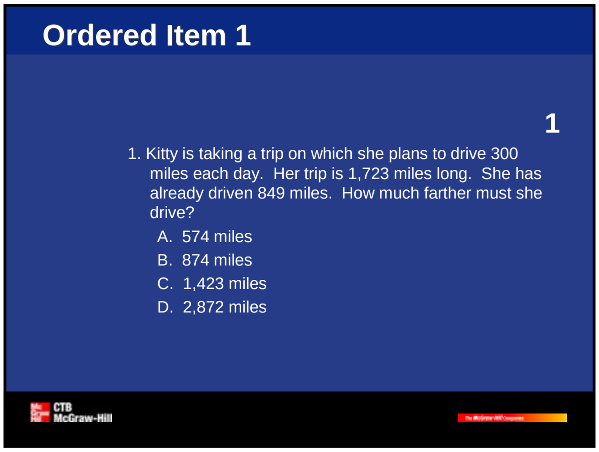
Item Map

*Print Name:*____ *Group Number:*______

Order of difficulty (easy to hard)	Loca-tion	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/score point?	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	256	12	7	CR	1/2	1		
7	262	9	5	MC	A	1		
8	282	12	7	CR	2/2	1		
9	303	9	6	MC	B	2		
10	321	9	8	MC	B	2		
11	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

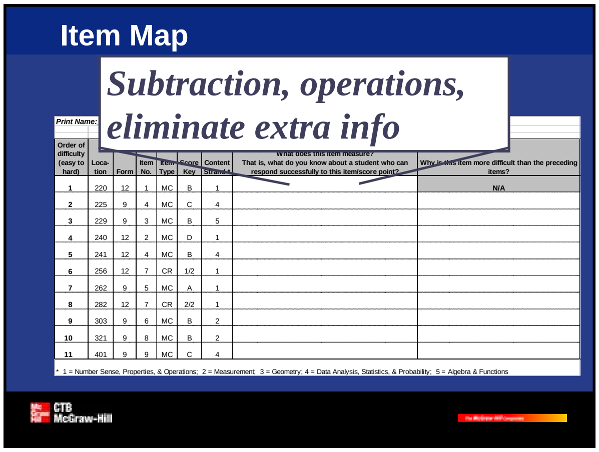


Ordered Item 1

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 Map

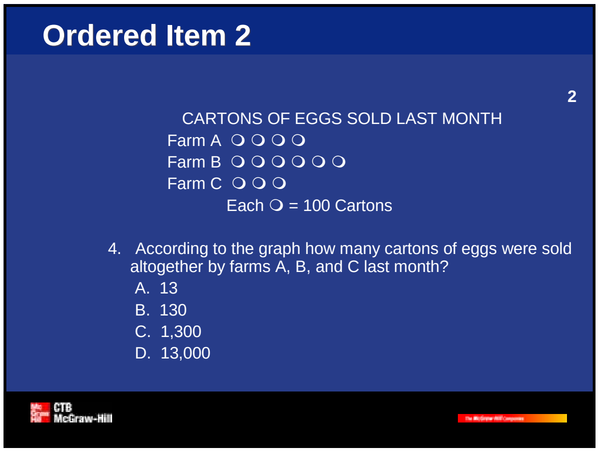
Subtraction, operations, eliminate extra info

Print Name:

Order of difficulty (easy to hard)	Loca-tion	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/score point?	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	256	12	7	CR	1/2	1		
7	262	9	5	MC	A	1		
8	282	12	7	CR	2/2	1		
9	303	9	6	MC	B	2		
10	321	9	8	MC	B	2		
11	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



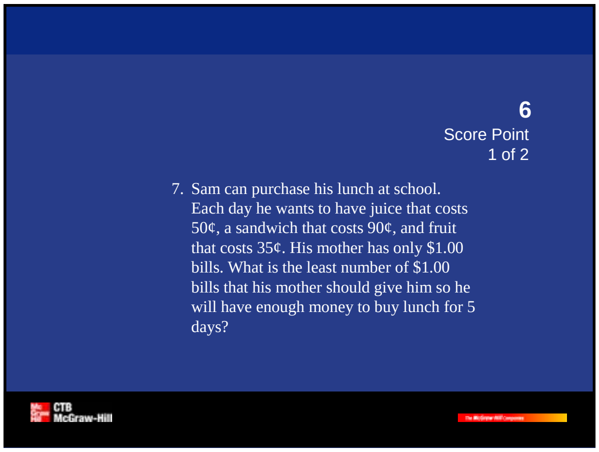
Ordered Item 2

2

CARTONS OF EGGS SOLD LAST MONTH

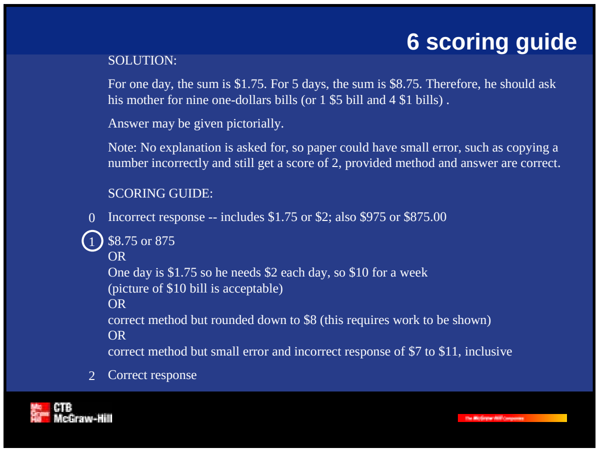
Farm A ❍ ❍ ❍ ❍
Farm B ❍ ❍ ❍ ❍ ❍ ❍
Farm C ❍ ❍ ❍
Each ❍ = 100 Cartons

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



6
Score Point
1 of 2

7. Sam can purchase his lunch at school. Each day he wants to have juice that costs 50¢, a sandwich that costs 90¢, and fruit that costs 35¢. His mother has only $1.00 bills. What is the least number of $1.00 bills that his mother should give him so he will have enough money to buy lunch for 5 days?



6 scoring guide

SOLUTION:

For one day, the sum is $1.75. For 5 days, the sum is $8.75. Therefore, he should ask his mother for nine one-dollars bills (or 1 $5 bill and 4 $1 bills) .

Answer may be given pictorially.

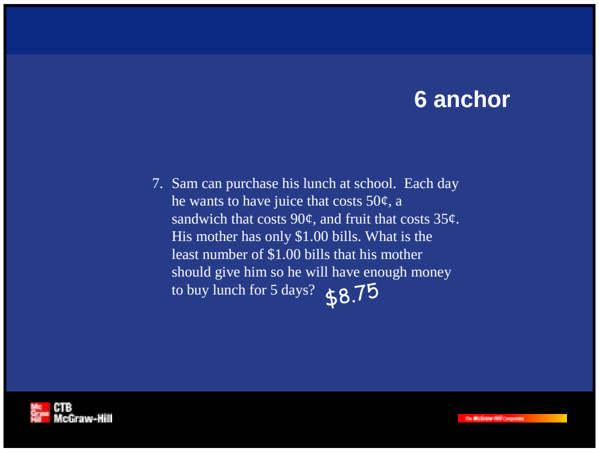
Note: No explanation is asked for, so paper could have small error, such as copying a number incorrectly and still get a score of 2, provided method and answer are correct.

SCORING GUIDE:

0 Incorrect response -- includes $1.75 or $2; also $975 or $875.00

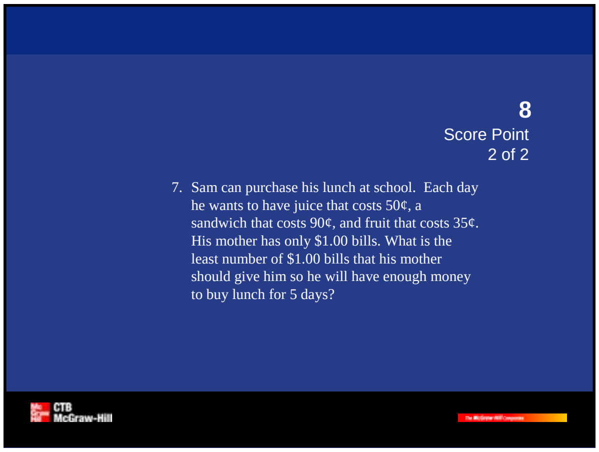
[✓ 1] $8.75 or 875
OR
One day is $1.75 so he needs $2 each day, so $10 for a week
(picture of $10 bill is acceptable)
OR
correct method but rounded down to $8 (this requires work to be shown)
OR
correct method but small error and incorrect response of $7 to $11, inclusive

2 Correct response



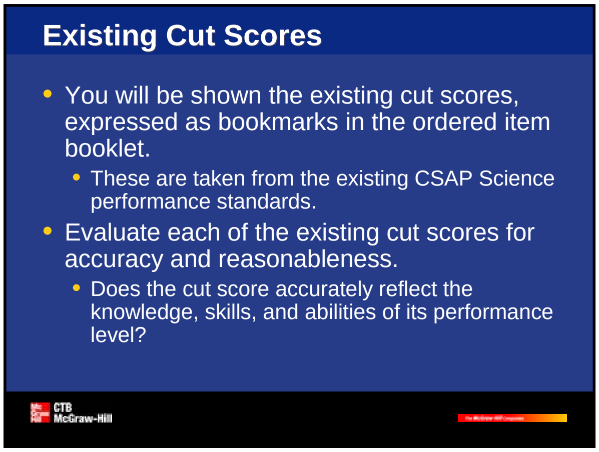
6 anchor

7. Sam can purchase his lunch at school. Each day he wants to have juice that costs 50¢, a sandwich that costs 90¢, and fruit that costs 35¢. His mother has only $1.00 bills. What is the least number of $1.00 bills that his mother should give him so he will have enough money to buy lunch for 5 days? $8.75



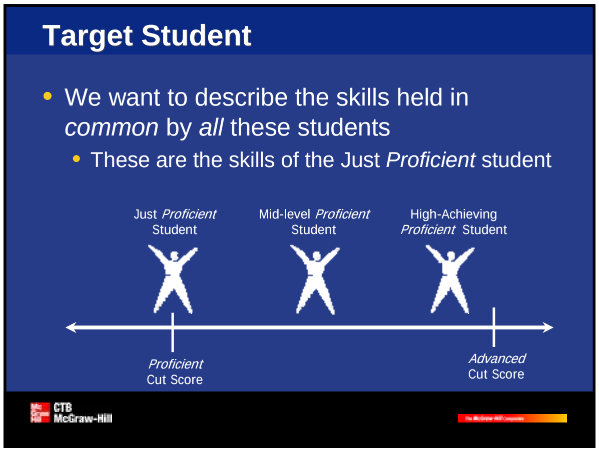
8
Score Point
2 of 2

7. Sam can purchase his lunch at school. Each day he wants to have juice that costs 50¢, a sandwich that costs 90¢, and fruit that costs 35¢. His mother has only $1.00 bills. What is the least number of $1.00 bills that his mother should give him so he will have enough money to buy lunch for 5 days?



Existing Cut Scores

- You will be shown the existing cut scores, expressed as bookmarks in the ordered item booklet.
  - These are taken from the existing CSAP Science performance standards.
- Evaluate each of the existing cut scores for accuracy and reasonableness.
  - Does the cut score accurately reflect the knowledge, skills, and abilities of its performance level?



Target Student

- We want to describe the skills held in *common* by *all* these students
  - These are the skills of the Just *Proficient* student

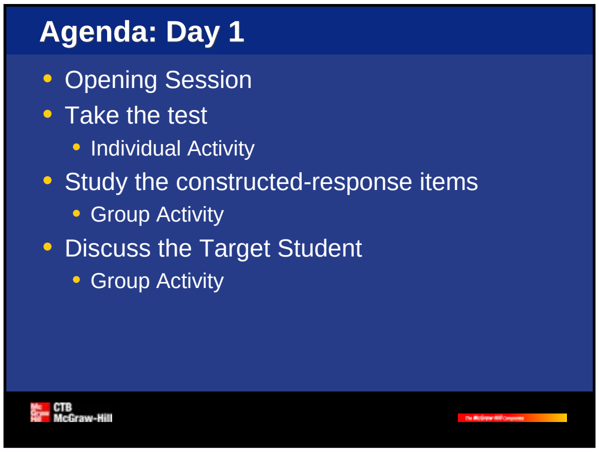
Just *Proficient* Student

Mid-level *Proficient* Student

High-Achieving *Proficient* Student

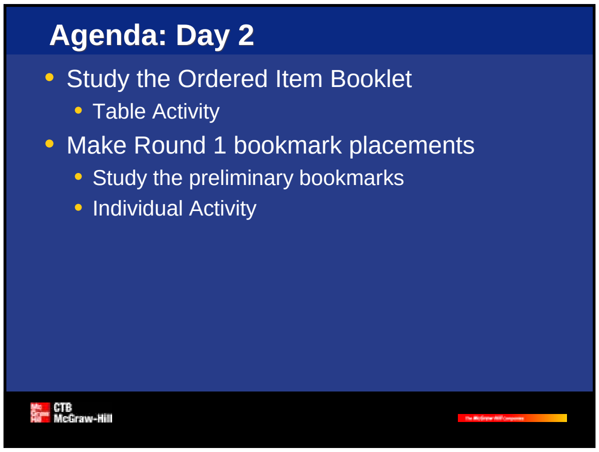
Proficient Cut Score

Advanced Cut Score



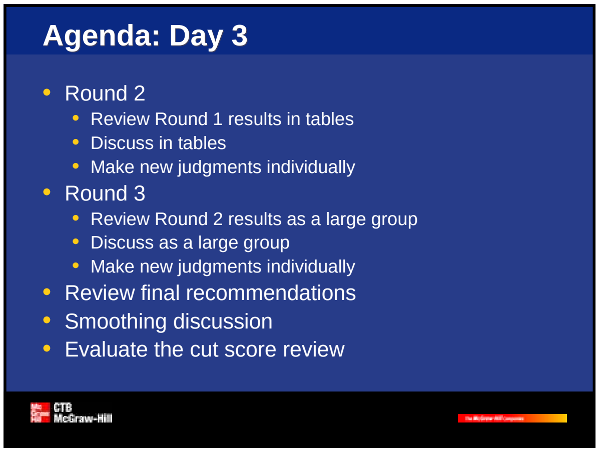
Agenda: Day 1

- Opening Session
- Take the test
  - Individual Activity
- Study the constructed-response items
  - Group Activity
- Discuss the Target Student
  - Group Activity



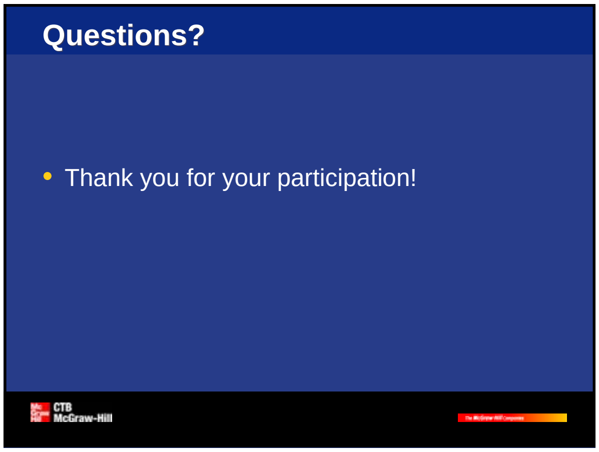
Agenda: Day 2

- Study the Ordered Item Booklet
  - Table Activity
- Make Round 1 bookmark placements
  - Study the preliminary bookmarks
  - Individual Activity



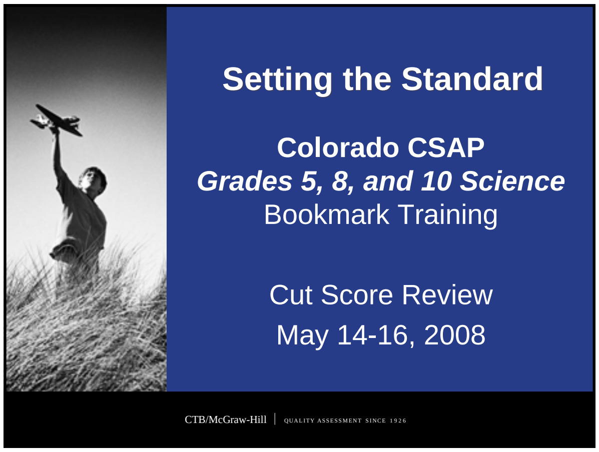
Agenda: Day 3

- Round 2
  - Review Round 1 results in tables
  - Discuss in tables
  - Make new judgments individually
- Round 3
  - Review Round 2 results as a large group
  - Discuss as a large group
  - Make new judgments individually
- Review final recommendations
- Smoothing discussion
- Evaluate the cut score review



Questions?

- Thank you for your participation!

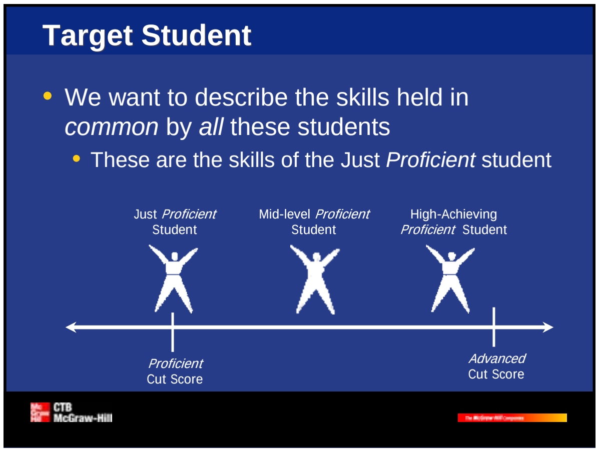


Setting the Standard

Colorado CSAP
Grades 5, 8, and 10 Science
Bookmark Training

Cut Score Review
May 14-16, 2008

CTB/McGraw-Hill | QUALITY ASSESSMENT SINCE 1926



Target Student

- We want to describe the skills held in *common* by *all* these students
  - These are the skills of the Just *Proficient* student

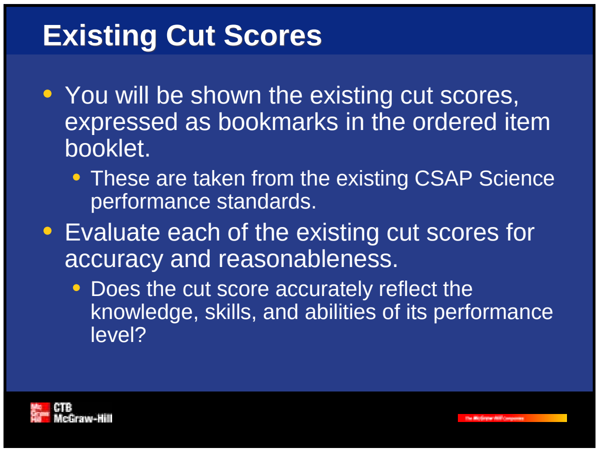
Just *Proficient* Student

Mid-level *Proficient* Student

High-Achieving *Proficient* Student

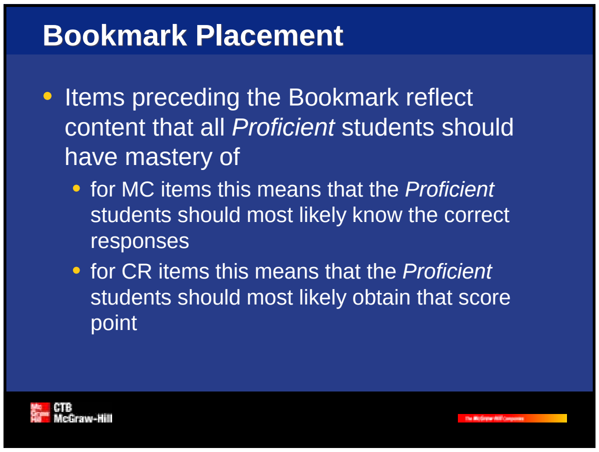
Proficient Cut Score

Advanced Cut Score



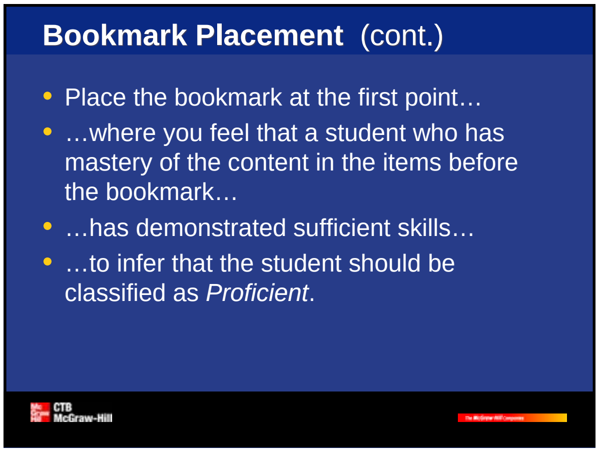
Existing Cut Scores

- You will be shown the existing cut scores, expressed as bookmarks in the ordered item booklet.
  - These are taken from the existing CSAP Science performance standards.
- Evaluate each of the existing cut scores for accuracy and reasonableness.
  - Does the cut score accurately reflect the knowledge, skills, and abilities of its performance level?



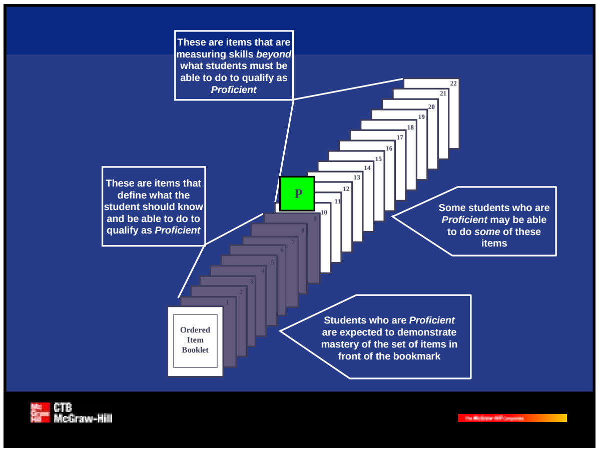
Bookmark Placement

- Items preceding the Bookmark reflect content that all *Proficient* students should have mastery of
  - for MC items this means that the *Proficient* students should most likely know the correct responses
  - for CR items this means that the *Proficient* students should most likely obtain that score point



Bookmark Placement (cont.)

- Place the bookmark at the first point…
- …where you feel that a student who has mastery of the content in the items before the bookmark…
- …has demonstrated sufficient skills…
- …to infer that the student should be classified as *Proficient*.

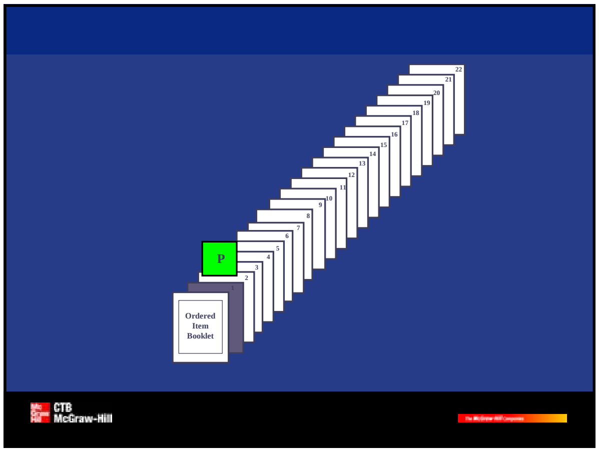


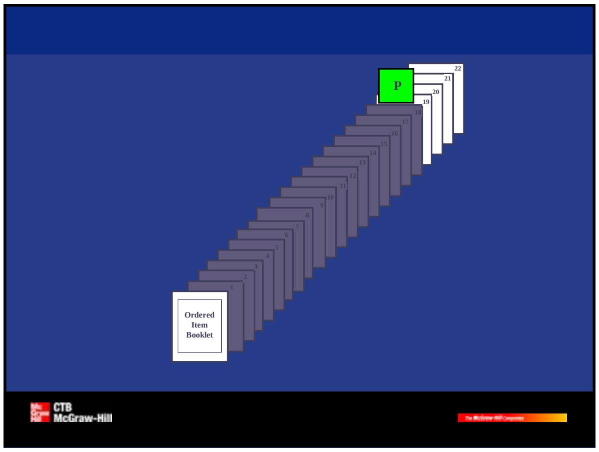
These are items that are measuring skills *beyond* what students must be able to do to qualify as *Proficient*

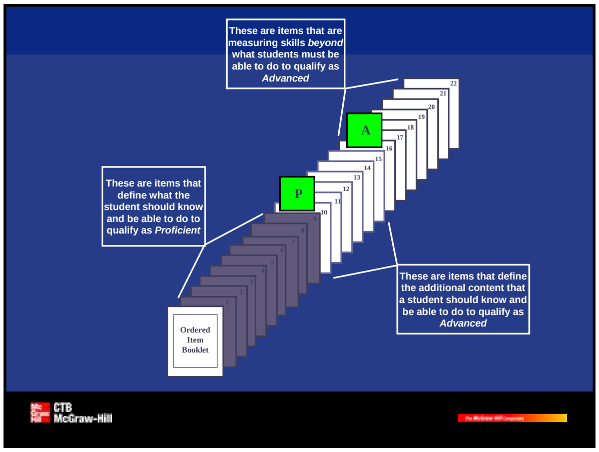
These are items that define what the student should know and be able to do to qualify as *Proficient*

Some students who are *Proficient* may be able to do *some* of these items

Students who are *Proficient* are expected to demonstrate mastery of the set of items in front of the bookmark



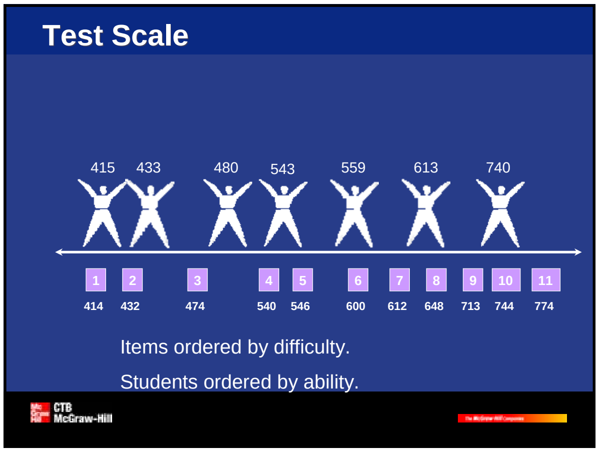




These are items that are measuring skills *beyond* what students must be able to do to qualify as *Advanced*

These are items that define what the student should know and be able to do to qualify as *Proficient*

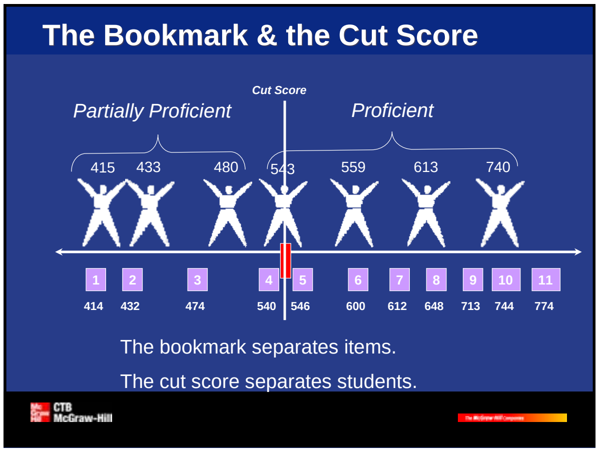
These are items that define the additional content that a student should know and be able to do to qualify as *Advanced*



Test Scale

Items ordered by difficulty.

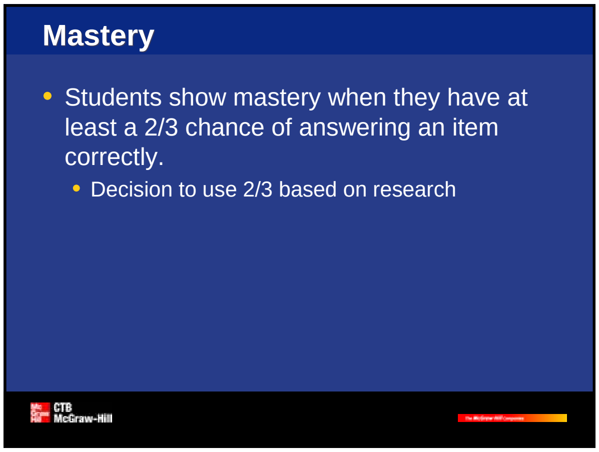
Students ordered by ability.



The Bookmark & the Cut Score

The bookmark separates items.

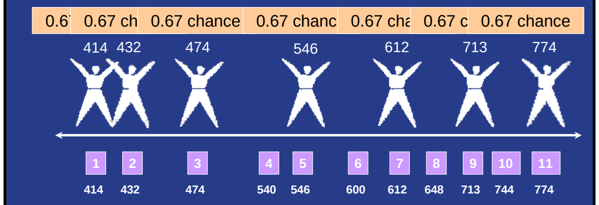
The cut score separates students.



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

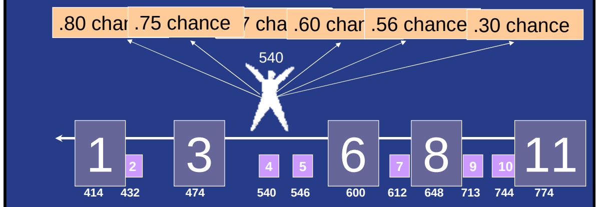
Item Location



Location is an indication of difficulty.

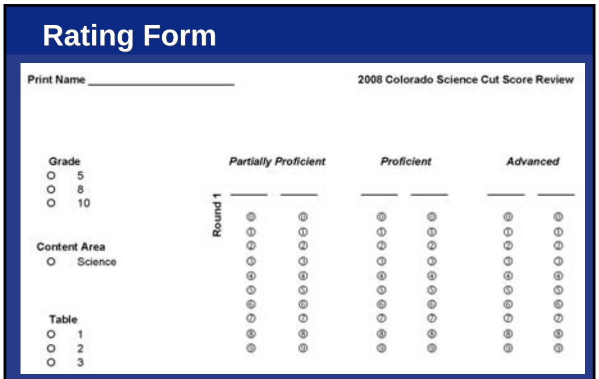
Location represents the ability level necessary to have a .67 chance of answering the item correctly.

Mastery and the Target Student



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

Rating Form



Print Name ____________________

2008 Colorado Science Cut Score Review

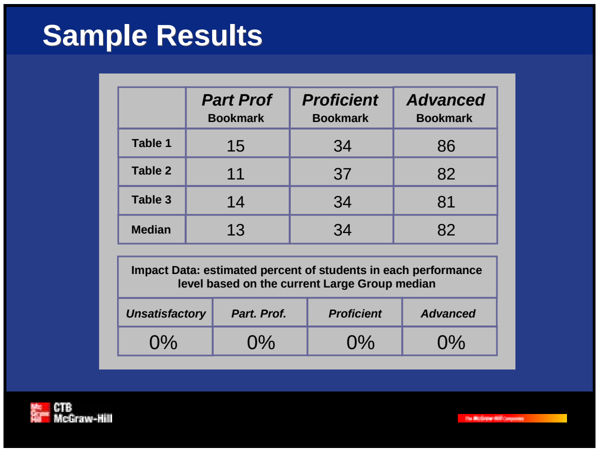
Grade
○ 5
○ 8
○ 10

Content Area
○ Science

Table
○ 1
○ 2
○ 3

Round 1

Partially Proficient		Proficient		Advanced	
____	____	____	____	____	____
⓪	⓪	⓪	⓪	⓪	⓪
①	①	①	①	①	①
②	②	②	②	②	②
③	③	③	③	③	③
④	④	④	④	④	④
⑤	⑤	⑤	⑤	⑤	⑤
⑥	⑥	⑥	⑥	⑥	⑥
⑦	⑦	⑦	⑦	⑦	⑦
⑧	⑧	⑧	⑧	⑧	⑧
⑨	⑨	⑨	⑨	⑨	⑨

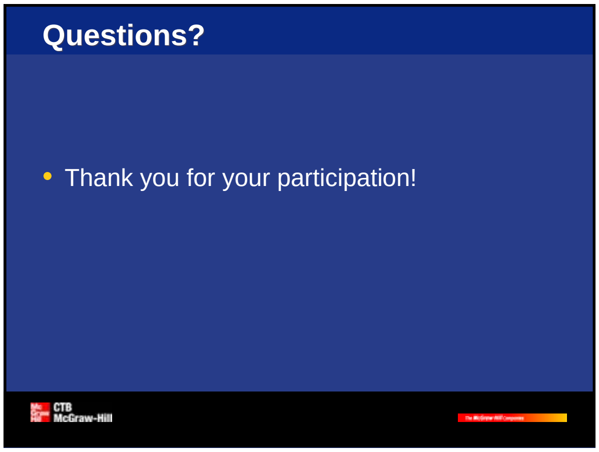


Sample Results

	Part Prof Bookmark	Proficient Bookmark	Advanced Bookmark
Table 1	15	34	86
Table 2	11	37	82
Table 3	14	34	81
Median	13	34	82

Impact Data: estimated percent of students in each performance level based on the current Large Group median

Unsatisfactory	Part. Prof.	Proficient	Advanced
0%	0%	0%	0%



Questions?

- Thank you for your participation!

Section E

Detailed Results of the Cut Score Review

CSAP Grade 5 Science
Round 1 Bookmark Placements

Table	Participant	Partially Proficient	Proficient	Advanced
1	6	20	45	75
1	8	22	46	78
1	9	20	48	73
1	10	21	56	78
2	1	17	51	72
2	2	23	54	74
2	3	27	53	74
2	4	20	57	78

Overall	Median	20.5	52	74.5
	Minimum	17	45	72
	Maximum	27	57	78
	SD	2.92	4.53	2.43

CSAP Grade 5 Science
Round 1 Cut Scores

Table	Participant	Partially Proficient	Proficient	Advanced
1	6	429	500	569
1	8	437	503	572
1	9	429	507	564
1	10	434	517	572
2	1	420	509	562
2	2	443	515	567
2	3	452	513	567
2	4	429	521	572

Overall	Median	429	511	567
	Minimum	420	500	562
	Maximum	452	521	572
	SD	9.89	7.17	3.83

CSAP Grade 5 Science
Round 1 Summary of Bookmark Placements

Statistic	Table	Partially Proficient	Proficient	Advanced
Median	1	20.5	47	76.5
Median	2	21.5	53.5	74
Median	Overall	20.5	52	74.5
Minimum	1	20	45	73
Minimum	2	17	51	72
Minimum	Overall	17	45	72
Maximum	1	22	56	78
Maximum	2	27	57	78
Maximum	Overall	27	57	78
SD	1	0.96	4.99	2.45
SD	2	4.27	2.50	2.52
SD	Overall	2.92	4.53	2.43

Overall	Median	20.5	52	74.5
	Minimum	17	45	72
	Maximum	27	57	78
	SD	2.92	4.53	2.43

CSAP Grade 5 Science
Round 1 Summary of Cut Scores

Statistic	Table	Partially Proficient	Proficient	Advanced
Median	1	432	505	571
Median	2	436	514	567
Median	Overall	429	511	567
Minimum	1	429	500	564
Minimum	2	420	509	562
Minimum	Overall	420	500	562
Maximum	1	437	517	572
Maximum	2	452	521	572
Maximum	Overall	452	521	572
SD	1	3.95	7.41	3.77
SD	2	14.26	5.00	4.08
SD	Overall	9.89	7.17	3.83

Overall	Median	429	511	567
	Minimum	420	500	562
	Maximum	452	521	572
	SD	9.89	7.17	3.83

CSAP Grade 5 Science
Round 1 Median Bookmark Summary

Table	Partially Proficient	Proficient	Advanced
1	20.5	47	76.5
2	21.5	53.5	74
Overall	20.5	52	74.5

Impact Data

	Unsatisfactory	Partially Proficient	Proficient	Advanced
Overall	14.8	42.8	30.8	11.6

CSAP Grade 5 Science
Round 2 Bookmark Placements

Table	Participant	Partially Proficient	Proficient	Advanced
1	6	20	45	75
1	8	20	46	75
1	9	20	46	75
1	10	21	56	78
2	1	17	53	76
2	2	23	54	76
2	3	20	51	74
2	4	20	57	78

Overall	Median	20	52	75.5
	Minimum	17	45	74
	Maximum	23	57	78
	SD	1.64	4.78	1.46

CSAP Grade 5 Science
Round 2 Cut Scores

Table	Participant	Partially Proficient	Proficient	Advanced
1	6	429	500	569
1	8	429	503	569
1	9	429	503	569
1	10	434	517	572
2	1	420	513	571
2	2	443	515	571
2	3	429	509	567
2	4	429	521	572

Overall	Median	429	511	569
	Minimum	420	500	567
	Maximum	443	521	572
	SD	6.43	7.59	1.77

CSAP Grade 5 Science
Round 2 Summary of Bookmark Placements

Statistic	Table	Partially Proficient	Proficient	Advanced
Median	1	20	46	75
Median	2	20	53.5	76
Median	Overall	20	52	75.5
Minimum	1	20	45	75
Minimum	2	17	51	74
Minimum	Overall	17	45	74
Maximum	1	21	56	78
Maximum	2	23	57	78
Maximum	Overall	23	57	78
SD	1	0.50	5.19	1.50
SD	2	2.45	2.50	1.63
SD	Overall	1.64	4.78	1.46

Overall	Median	20	52	75.5
	Minimum	17	45	74
	Maximum	23	57	78
	SD	1.64	4.78	1.46

CSAP Grade 5 Science
Round 2 Summary of Cut Scores

Statistic	Table	Partially Proficient	Proficient	Advanced
Median	1	429	503	569
Median	2	429	514	571
Median	Overall	429	511	569
Minimum	1	429	500	569
Minimum	2	420	509	567
Minimum	Overall	420	500	567
Maximum	1	434	517	572
Maximum	2	443	521	572
Maximum	Overall	443	521	572
SD	1	2.50	7.63	1.50
SD	2	9.50	5.00	2.22
SD	Overall	6.43	7.59	1.77

Overall	Median	429	511	569
	Minimum	420	500	567
	Maximum	443	521	572
	SD	6.43	7.59	1.77

CSAP Grade 5 Science
Round 2 Median Bookmark Summary

Table	Partially Proficient	Proficient	Advanced
1	20	46	75
2	20	53.5	76
Overall	20	52	75.5

Impact Data

	Unsatisfactory	Partially Proficient	Proficient	Advanced
Overall	14.8	42.8	31.6	10.8

CSAP Grade 5 Science
Round 3 Bookmark Placements

Table	Participant	Partially Proficient	Proficient	Advanced
1	6	20	45	75
1	8	20	45	75
1	9	20	46	75
1	10	20	50	75
2	1	17	50	75
2	2	20	50	75
2	3	17	51	74
2	4	20	54	74

Overall	Median	20	50	75
	Minimum	17	45	74
	Maximum	20	54	75
	SD	1.39	3.23	0.46

CSAP Grade 5 Science
Round 3 Cut Scores

Table	Participant	Partially Proficient	Proficient	Advanced
1	6	429	500	569
1	8	429	500	569
1	9	429	503	569
1	10	429	508	569
2	1	420	508	569
2	2	429	508	569
2	3	420	509	567
2	4	429	515	567

Overall	Median	429	508	569
	Minimum	420	500	567
	Maximum	429	515	569
	SD	4.17	5.10	0.93

CSAP Grade 5 Science
Round 3 Summary of Bookmark Placements

Statistic	Table	Partially Proficient	Proficient	Advanced
Median	1	20	45.5	75
Median	2	18.5	50.5	74.5
Median	Overall	20	50	75
Minimum	1	20	45	75
Minimum	2	17	50	74
Minimum	Overall	17	45	74
Maximum	1	20	50	75
Maximum	2	20	54	75
Maximum	Overall	20	54	75
SD	1	0.00	2.38	0.00
SD	2	1.73	1.89	0.58
SD	Overall	1.39	3.23	0.46

Overall	Median	20	50	75
	Minimum	17	45	74
	Maximum	20	54	75
	SD	1.39	3.23	0.46

CSAP Grade 5 Science
Round 3 Summary of Cut Scores

Statistic	Table	Partially Proficient	Proficient	Advanced
Median	1	429	502	569
Median	2	425	509	568
Median	Overall	429	508	569
Minimum	1	429	500	569
Minimum	2	420	508	567
Minimum	Overall	420	500	567
Maximum	1	429	508	569
Maximum	2	429	515	569
Maximum	Overall	429	515	569
SD	1	0.00	3.77	0.00
SD	2	5.20	3.37	1.15
SD	Overall	4.17	5.10	0.93

Overall	Median	429	508	569
	Minimum	420	500	567
	Maximum	429	515	569
	SD	4.17	5.10	0.93

CSAP Grade 5 Science
Round 3 Median Bookmark Summary

Table	Partially Proficient	Proficient	Advanced
1	20	45.5	75
2	18.5	50.5	74.5
Overall	20	50	75

Impact Data

	Unsatisfactory	Partially Proficient	Proficient	Advanced
Overall	14.8	40.8	33.6	10.8

CSAP Grade 8 Science
Round 1 Bookmark Placements

Table	Participant	Partially Proficient	Proficient	Advanced
1	1	3	24	57
1	2	6	23	62
1	3	8	25	66
1	4	5	20	63
1	5	5	23	66
2	6	9	24	60
2	7	6	25	73
2	8	6	20	62
2	9	6	23	66

Overall	Median	6	23	63
	Minimum	3	20	57
	Maximum	9	25	73
	SD	1.73	1.87	4.57

CSAP Grade 8 Science
Round 1 Cut Scores

Table	Participant	Partially Proficient	Proficient	Advanced
1	1	437	517	564
1	2	474	516	570
1	3	479	518	581
1	4	459	499	575
1	5	459	516	581
2	6	481	517	567
2	7	474	518	600
2	8	474	499	570
2	9	474	516	581

Overall	Median	474	516	575
	Minimum	437	499	564
	Maximum	481	518	600
	SD	13.95	7.91	10.88

CSAP Grade 8 Science
Round 1 Summary of Bookmark Placements

Statistic	Table	Partially Proficient	Proficient	Advanced
Median	1	5	23	63
Median	2	6	23.5	64
Median	Overall	6	23	63
Minimum	1	3	20	57
Minimum	2	6	20	60
Minimum	Overall	3	20	57
Maximum	1	8	25	66
Maximum	2	9	25	73
Maximum	Overall	9	25	73
SD	1	1.82	1.87	3.70
SD	2	1.50	2.16	5.74
SD	Overall	1.73	1.87	4.57

Overall	Median	6	23	63
	Minimum	3	20	57
	Maximum	9	25	73
	SD	1.73	1.87	4.57

CSAP Grade 8 Science
Round 1 Summary of Cut Scores

Statistic	Table	Partially Proficient	Proficient	Advanced
Median	1	459	516	575
Median	2	474	517	576
Median	Overall	474	516	575
Minimum	1	437	499	564
Minimum	2	474	499	567
Minimum	Overall	437	499	564
Maximum	1	479	518	581
Maximum	2	481	518	600
Maximum	Overall	481	518	600
SD	1	16.40	7.98	7.33
SD	2	3.50	9.04	14.93
SD	Overall	13.95	7.91	10.88

Overall	Median	474	516	575
	Minimum	437	499	564
	Maximum	481	518	600
	SD	13.95	7.91	10.88

CSAP Grade 8 Science
Round 1 Median Bookmark Summary

Table	Partially Proficient	Proficient	Advanced
1	5	23	63
2	6	23.5	64
Overall	6	23	63

Impact Data

	Unsatisfactory	Partially Proficient	Proficient	Advanced
Overall	31.3	27.2	33.1	8.4

CSAP Grade 8 Science
Round 2 Bookmark Placements

Table	Participant	Partially Proficient	Proficient	Advanced
1	1	6	25	61
1	2	6	23	66
1	3	6	25	66
1	5	6	25	66
2	6	7	24	63
2	7	5	22	71
2	8	5	21	65
2	9	6	22	66

Overall	Median	6	23.5	66
	Minimum	5	21	61
	Maximum	7	25	71
	SD	0.64	1.60	2.88

CSAP Grade 8 Science
Round 2 Cut Scores

Table	Participant	Partially Proficient	Proficient	Advanced
1	1	474	518	568
1	2	474	516	581
1	3	474	518	581
1	5	474	518	581
2	6	478	517	575
2	7	459	512	592
2	8	459	503	579
2	9	474	512	581

Overall	Median	474	516	581
	Minimum	459	503	568
	Maximum	478	518	592
	SD	7.38	5.20	6.73

CSAP Grade 8 Science
Round 2 Summary of Bookmark Placements

Statistic	Table	Partially Proficient	Proficient	Advanced
Median	1	6	25	66
Median	2	5.5	22	65.5
Median	Overall	6	23.5	66
Minimum	1	6	23	61
Minimum	2	5	21	63
Minimum	Overall	5	21	61
Maximum	1	6	25	66
Maximum	2	7	24	71
Maximum	Overall	7	25	71
SD	1	0.00	1.00	2.50
SD	2	0.96	1.26	3.40
SD	Overall	0.64	1.60	2.88

Overall	Median	6	23.5	66
	Minimum	5	21	61
	Maximum	7	25	71
	SD	0.64	1.60	2.88

CSAP Grade 8 Science
Round 2 Summary of Cut Scores

Statistic	**Table**	**Partially Proficient**	**Proficient**	**Advanced**
Median	1	474	518	581
Median	2	467	512	580
Median	Overall	474	516	581
Minimum	1	474	516	568
Minimum	2	459	503	575
Minimum	Overall	459	503	568
Maximum	1	474	518	581
Maximum	2	478	517	592
Maximum	Overall	478	518	592
SD	1	0.00	1.00	6.50
SD	2	9.95	5.83	7.27
SD	Overall	7.38	5.20	6.73

Overall	Median	474	516	581
	Minimum	459	503	568
	Maximum	478	518	592
	SD	7.38	5.20	6.73

CSAP Grade 8 Science
Round 2 Median Bookmark Summary

Table	Partially Proficient	Proficient	Advanced
1	6	25	66
2	5.5	22	65.5
Overall	6	23.5	66

Impact Data

	Unsatisfactory	Partially Proficient	Proficient	Advanced
Overall	31.3	27.2	34.7	6.8

CSAP Grade 8 Science
Round 3 Bookmark Placements

Table	Participant	Partially Proficient	Proficient	Advanced
1	1	5	22	65
1	2	5	21	65
1	3	5	22	65
1	5	6	22	66
2	6	5	24	66
2	7	5	21	71
2	8	5	20	64
2	9	5	22	66

Overall	Median	5	22	65.5
	Minimum	5	20	64
	Maximum	6	24	71
	SD	0.35	1.16	2.14

CSAP Grade 8 Science
Round 3 Cut Scores

Table	Participant	Partially Proficient	Proficient	Advanced
1	1	459	512	579
1	2	459	503	579
1	3	459	512	579
1	5	474	512	581
2	6	459	517	581
2	7	459	503	592
2	8	459	499	578
2	9	459	512	581

Overall	Median	459	512	579
	Minimum	459	499	578
	Maximum	474	517	592
	SD	5.30	6.23	4.50

CSAP Grade 8 Science
Round 3 Summary of Bookmark Placements

Statistic	Table	Partially Proficient	Proficient	Advanced
Median	1	5	22	65
Median	2	5	21.5	66
Median	Overall	5	22	65.5
Minimum	1	5	21	65
Minimum	2	5	20	64
Minimum	Overall	5	20	64
Maximum	1	6	22	66
Maximum	2	5	24	71
Maximum	Overall	6	24	71
SD	1	0.50	0.50	0.50
SD	2	0.00	1.71	2.99
SD	Overall	0.35	1.16	2.14

Overall	Median	5	22	65.5
	Minimum	5	20	64
	Maximum	6	24	71
	SD	0.35	1.16	2.14

CSAP Grade 8 Science
Round 3 Summary of Cut Scores

Statistic	Table	Partially Proficient	Proficient	Advanced
Median	1	459	512	579
Median	2	459	508	581
Median	Overall	459	512	579
Minimum	1	459	503	579
Minimum	2	459	499	578
Minimum	Overall	459	499	578
Maximum	1	474	512	581
Maximum	2	459	517	592
Maximum	Overall	474	517	592
SD	1	7.50	4.50	1.00
SD	2	0.00	8.22	6.16
SD	Overall	5.30	6.23	4.50

Overall	Median	459	512	579
	Minimum	459	499	578
	Maximum	474	517	592
	SD	5.30	6.23	4.50

CSAP Grade 8 Science
Round 3 Median Bookmark Summary

Table	Partially Proficient	Proficient	Advanced
1	5	22	65
2	5	21.5	66
Overall	5	22	65.5

Impact Data

	Unsatisfactory	Partially Proficient	Proficient	Advanced
Overall	23.9	31.7	37.0	7.4

CSAP Grade 10 Science
Round 1 Bookmark Placements

Table	Participant	Partially Proficient	Proficient	Advanced
1	6	12	31	69
1	7	9	30	73
1	8	9	28	68
1	9	12	24	64
2	2	13	33	70
2	3	14	35	77
2	4	13	41	68
2	5	12	31	71

Overall	Median	12	31	69.5
	Minimum	9	24	64
	Maximum	14	41	77
	SD	1.83	5.01	3.85

CSAP Grade 10 Science
Round 1 Cut Scores

Table	Participant	Partially Proficient	Proficient	Advanced
1	6	469	510	583
1	7	457	508	592
1	8	457	504	581
1	9	469	499	572
2	2	471	513	586
2	3	472	515	602
2	4	471	526	581
2	5	469	510	590

Overall	Median	469	510	583
	Minimum	457	499	572
	Maximum	472	526	602
	SD	6.20	8.00	8.97

CSAP Grade 10 Science
Round 1 Summary of Bookmark Placements

Statistic	Table	Partially Proficient	Proficient	Advanced
Median	1	10.5	29	68.5
Median	2	13	34	70.5
Median	Overall	12	31	69.5
Minimum	1	9	24	64
Minimum	2	12	31	68
Minimum	Overall	9	24	64
Maximum	1	12	31	73
Maximum	2	14	41	77
Maximum	Overall	14	41	77
SD	1	1.73	3.10	3.70
SD	2	0.82	4.32	3.87
SD	Overall	1.83	5.01	3.85

Overall	Median	12	31	69.5
	Minimum	9	24	64
	Maximum	14	41	77
	SD	1.83	5.01	3.85

CSAP Grade 10 Science
Round 1 Summary of Cut Scores

Statistic	Table	Partially Proficient	Proficient	Advanced
Median	1	463	506	582
Median	2	471	514	588
Median	Overall	469	510	583
Minimum	1	457	499	572
Minimum	2	469	510	581
Minimum	Overall	457	499	572
Maximum	1	469	510	592
Maximum	2	472	526	602
Maximum	Overall	472	526	602
SD	1	6.93	4.86	8.21
SD	2	1.26	6.98	8.96
SD	Overall	6.20	8.00	8.97

Overall	Median	469	510	583
	Minimum	457	499	572
	Maximum	472	526	602
	SD	6.20	8.00	8.97

CSAP Grade 10 Science
Round 1 Median Bookmark Summary

Table	Partially Proficient	Proficient	Advanced
1	10.5	29	68.5
2	13	34	70.5
Overall	12	31	69.5

Impact Data

	Unsatisfactory	Partially Proficient	Proficient	Advanced
Overall	26.0	26.4	42.1	5.5

CSAP Grade 10 Science
Round 2 Bookmark Placements

Table	Participant	Partially Proficient	Proficient	Advanced
1	6	10	20	66
1	7	9	21	65
1	8	9	27	66
1	9	9	24	64
2	2	13	38	71
2	3	13	38	71
2	4	13	39	71
2	5	13	38	71

Overall	Median	11.5	32.5	68.5
	Minimum	9	20	64
	Maximum	13	39	71
	SD	2.03	8.42	3.14

CSAP Grade 10 Science
Round 2 Cut Scores

Table	Participant	Partially Proficient	Proficient	Advanced
1	6	461	493	576
1	7	457	497	573
1	8	457	501	576
1	9	457	499	572
2	2	471	519	590
2	3	471	519	590
2	4	471	522	590
2	5	471	519	590

Overall	Median	465	512	581
	Minimum	457	493	572
	Maximum	471	522	590
	SD	7.07	12.14	8.53

CSAP Grade 10 Science
Round 2 Summary of Bookmark Placements

Statistic	Table	Partially Proficient	Proficient	Advanced
Median	1	9	22.5	65.5
Median	2	13	38	71
Median	Overall	11.5	32.5	68.5
Minimum	1	9	20	64
Minimum	2	13	38	71
Minimum	Overall	9	20	64
Maximum	1	10	27	66
Maximum	2	13	39	71
Maximum	Overall	13	39	71
SD	1	0.50	3.16	0.96
SD	2	0.00	0.50	0.00
SD	Overall	2.03	8.42	3.14

Overall	Median	11.5	32.5	68.5
	Minimum	9	20	64
	Maximum	13	39	71
	SD	2.03	8.42	3.14

CSAP Grade 10 Science
Round 2 Summary of Cut Scores

Statistic	Table	Partially Proficient	Proficient	Advanced
Median	1	457	498	575
Median	2	471	519	590
Median	Overall	465	512	581
Minimum	1	457	493	572
Minimum	2	471	519	590
Minimum	Overall	457	493	572
Maximum	1	461	501	576
Maximum	2	471	522	590
Maximum	Overall	471	522	590
SD	1	2.00	3.42	2.06
SD	2	0.00	1.50	0.00
SD	Overall	7.07	12.14	8.53

Overall	Median	465	512	581
	Minimum	457	493	572
	Maximum	471	522	590
	SD	7.07	12.14	8.53

CSAP Grade 10 Science
Round 2 Median Bookmark Summary

Table	Partially Proficient	Proficient	Advanced
1	9	22.5	65.5
2	13	38	71
Overall	11.5	32.5	68.5

Impact Data

	Unsatisfactory	Partially Proficient	Proficient	Advanced
Overall	24.1	29.6	40.1	6.2

CSAP Grade 10 Science
Round 3 Bookmark Placements

Table	Participant	Partially Proficient	Proficient	Advanced
1	6	13	28	66
1	7	9	30	65
1	8	11	28	68
1	9	9	24	64
2	2	12	32	69
2	3	12	29	71
2	4	13	36	68
2	5	12	29	71

Overall	Median	12	29	68
	Minimum	9	24	64
	Maximum	13	36	71
	SD	1.60	3.46	2.60

CSAP Grade 10 Science
Round 3 Cut Scores

Table	Participant	Partially Proficient	Proficient	Advanced
1	6	471	504	576
1	7	457	508	573
1	8	465	504	581
1	9	457	499	572
2	2	469	512	583
2	3	469	507	590
2	4	471	516	581
2	5	469	507	590

Overall	Median	469	507	581
	Minimum	457	499	572
	Maximum	471	516	590
	SD	5.86	5.19	6.92

CSAP Grade 10 Science
Round 3 Summary of Bookmark Placements

Statistic	Table	Partially Proficient	Proficient	Advanced
Median	1	10	28	65.5
Median	2	12	30.5	70
Median	Overall	12	29	68
Minimum	1	9	24	64
Minimum	2	12	29	68
Minimum	Overall	9	24	64
Maximum	1	13	30	68
Maximum	2	13	36	71
Maximum	Overall	13	36	71
SD	1	1.91	2.52	1.71
SD	2	0.50	3.32	1.50
SD	Overall	1.60	3.46	2.60

Overall	Median	12	29	68
	Minimum	9	24	64
	Maximum	13	36	71
	SD	1.60	3.46	2.60

CSAP Grade 10 Science
Round 3 Summary of Cut Scores

Statistic	Table	Partially Proficient	Proficient	Advanced
Median	1	461	504	575
Median	2	469	510	587
Median	Overall	469	507	581
Minimum	1	457	499	572
Minimum	2	469	507	581
Minimum	Overall	457	499	572
Maximum	1	471	508	581
Maximum	2	471	516	590
Maximum	Overall	471	516	590
SD	1	6.81	3.69	4.04
SD	2	1.00	4.36	4.69
SD	Overall	5.86	5.19	6.92

Overall	Median	469	507	581
	Minimum	457	499	572
	Maximum	471	516	590
	SD	5.86	5.19	6.92

CSAP Grade 10 Science
Round 3 Median Bookmark Summary

Table	Partially Proficient	Proficient	Advanced
1	10	28	65.5
2	12	30.5	70
Overall	12	29	68

Impact Data

	Unsatisfactory	Partially Proficient	Proficient	Advanced
Overall	26.0	24.0	43.9	6.1

Science

Based on Participants' Round 1 Bookmark Recommendations

Cut score review workshop held May 14-16, 2008

	5	8	10	
Unsatisfactory	14.8%	31.3%	26.0%	**Impact**
Partially Proficient	42.8%	27.2%	26.4%	
Proficient	30.8%	33.1%	42.1%	
Advanced	11.6%	8.4%	5.6%	
Proficient & Above	**42%**	**42%**	**48%**	

Partially Proficient	429	474	469	**Cut Scores**
Proficient	511	516	510	
Advanced	567	575	583	

CSAP Grade 10 Science
Round 3 Median Bookmark Summary

Table	Partially Proficient	Proficient	Advanced
1	10	28	65.5
2	12	30.5	70
Overall	12	29	68

Impact Data

	Unsatisfactory	Partially Proficient	Proficient	Advanced
Overall	26.0	24.0	43.9	6.1

Science

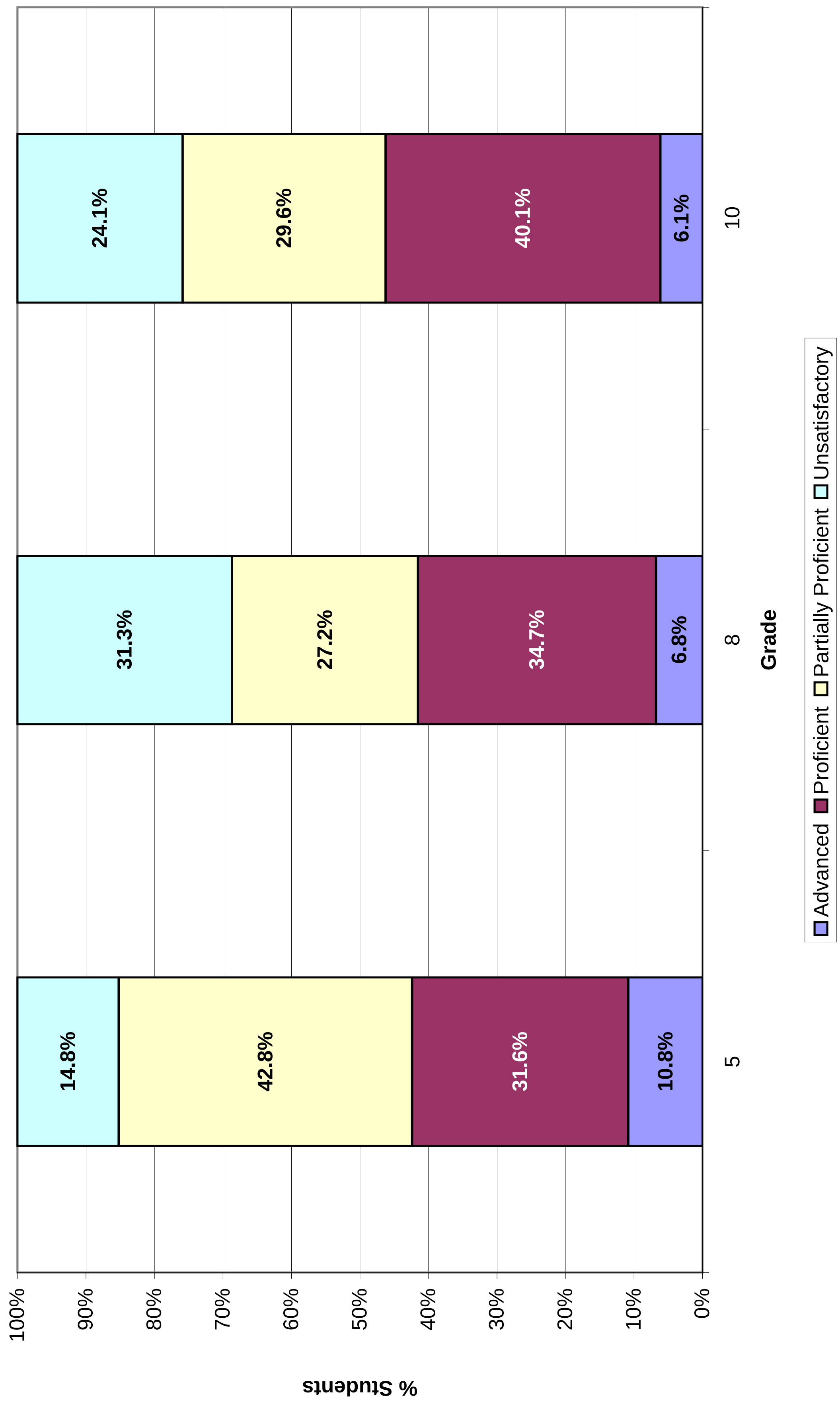
Based on Participants' Round 2 Bookmark Recommendations

Cut score review workshop held May 14-16, 2008

	5	8	10	
Unsatisfactory	14.8%	31.3%	24.1%	**Impact**
Partially Proficient	42.8%	27.2%	29.6%	
Proficient	31.6%	34.7%	40.1%	
Advanced	10.8%	6.8%	6.1%	
Proficient & Above	**42%**	**42%**	**46%**	

Partially Proficient	429	474	465	**Cut Scores**
Proficient	511	516	512	
Advanced	569	581	581	

Colorado Student Assessment Program
Science Round 2 Results: Percent of Students by Performance Level



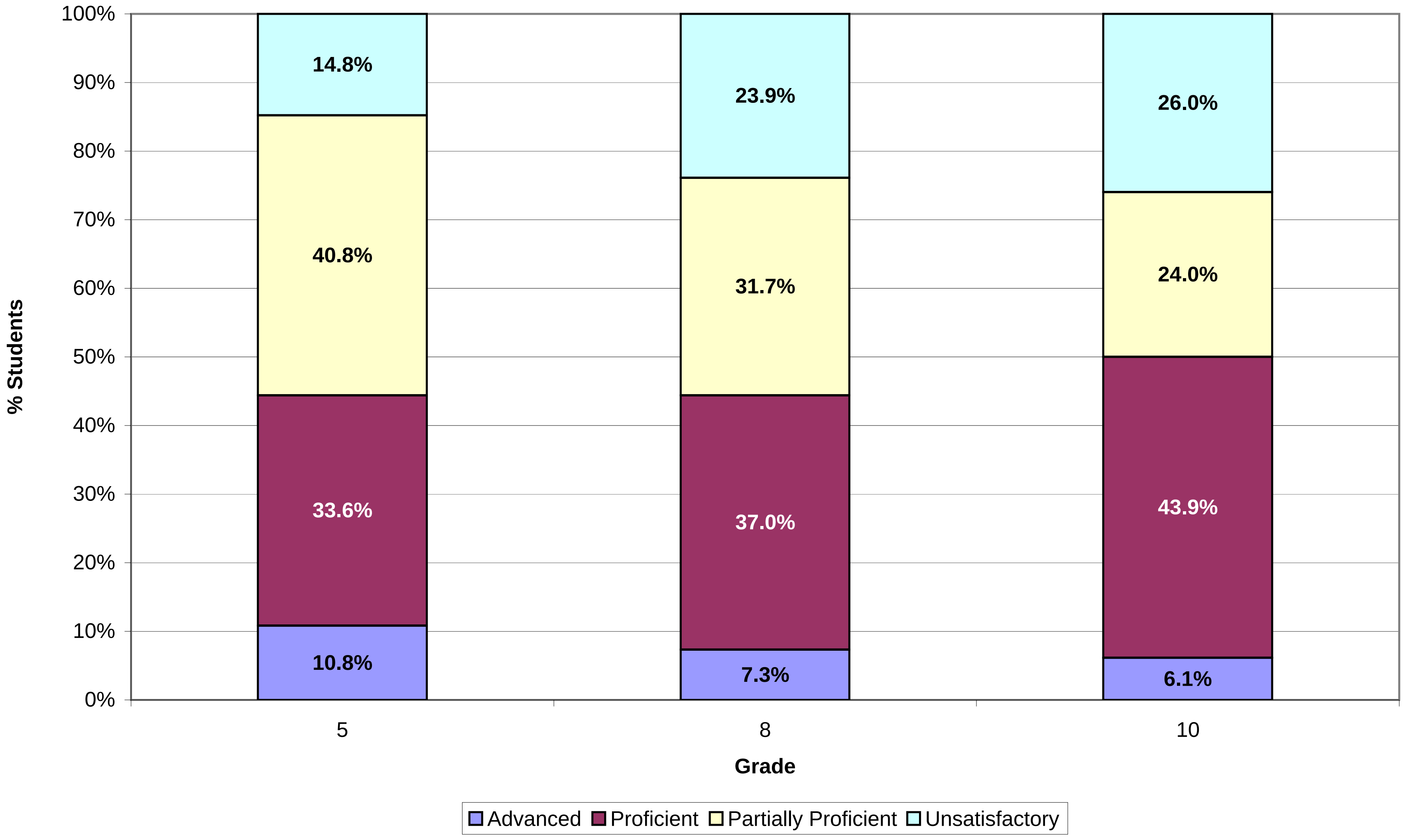
Science **Based on Participants' Round 3 Bookmark Recommendations**

Cut score review workshop held May 14-16, 2008

	5	8	10	
Unsatisfactory	14.8%	23.9%	26.0%	**Impact**
Partially Proficient	40.8%	31.7%	24.0%	
Proficient	33.6%	37.0%	43.9%	
Advanced	10.8%	7.3%	6.1%	
Proficient & Above	**44%**	**44%**	**50%**	

Partially Proficient	429	459	469	**Cut Scores**
Proficient	508	512	507	
Advanced	569	579	581	

Colorado Student Assessment Program
Science Final Round Results: Percent of Students by Performance Level



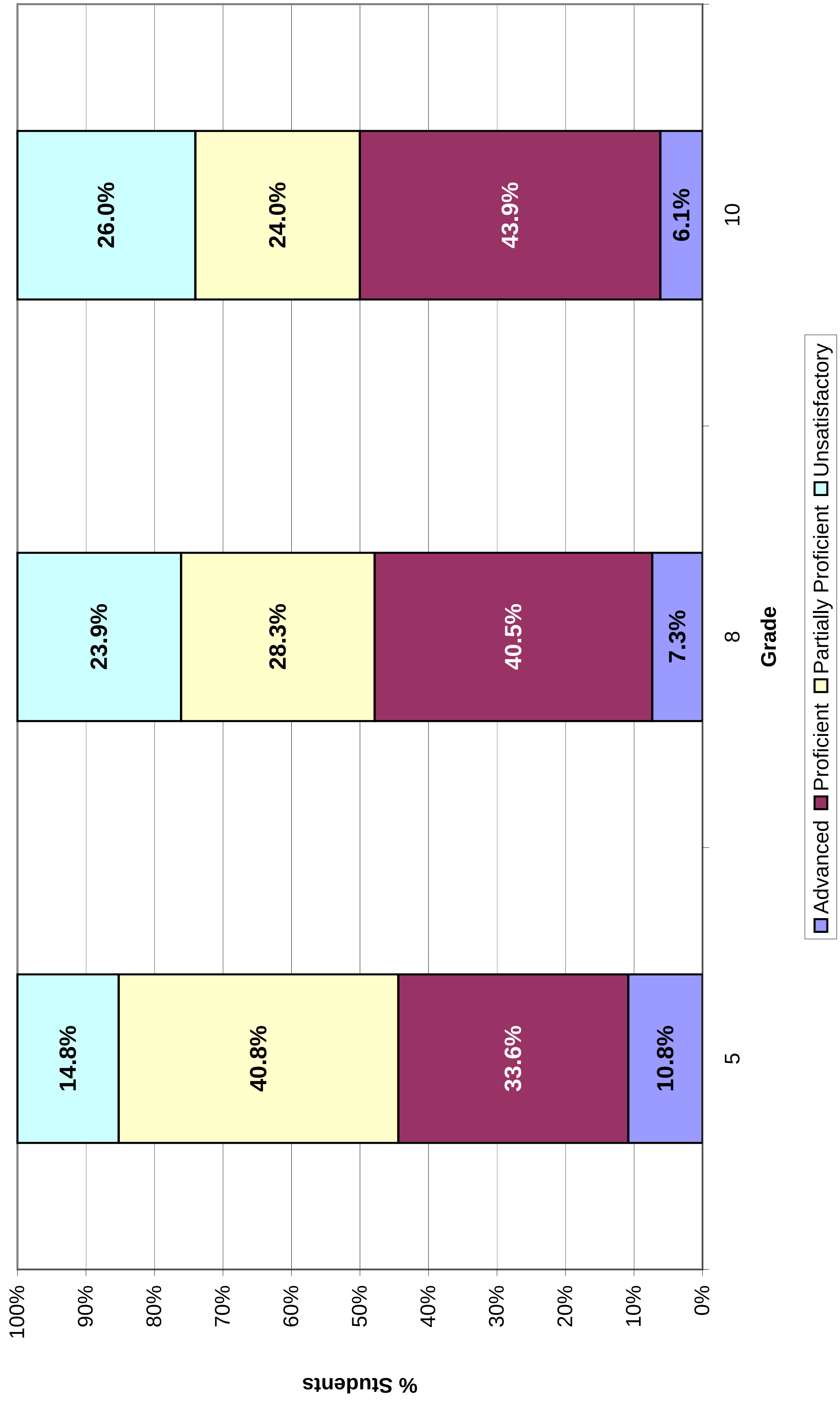
Science

Based on Results from the Smoothing Process
Cut score review workshop held May 14-16, 2008

	5	8	10	
Unsatisfactory	15%	24%	26%	**Impact**
Partially Proficient	41%	28%	24%	
Proficient	34%	41%	44%	
Advanced	11%	7%	6%	
Proficient & Above	**44%**	**48%**	**50%**	

Partially Proficient	429	459	469	**Cut Scores**
Proficient	508	*507*	507	
Advanced	569	579	581	

Colorado Student Assessment Program
Science Smoothing Results: Percent of Students by Performance Level



Science

Based on Results from the Smoothing Process

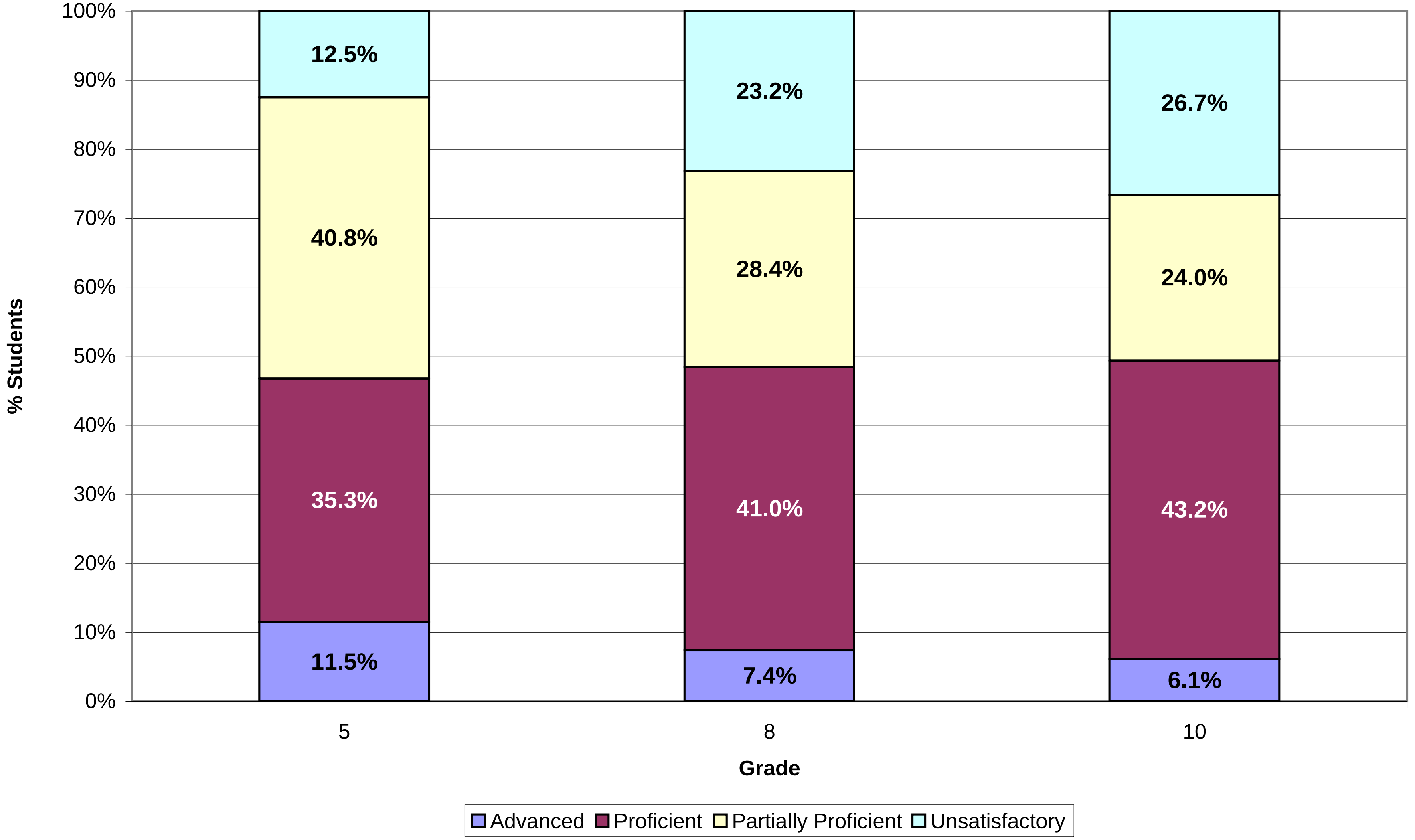
Cut score review workshop held May 14-16, 2008

100% Data

	5	8	10	
Unsatisfactory	12%	23%	27%	**Impact**
Partially Proficient	41%	28%	24%	
Proficient	35%	41%	43%	
Advanced	11%	7%	6%	
Proficient & Above	**47%**	**48%**	**49%**	

Partially Proficient	429	459	469	**Cut Scores**
Proficient	508	*507*	507	
Advanced	569	579	581	

Colorado Student Assessment Program 100% Data
Science Smoothing Results: Percent of Students by Performance Level



Section F

Participants’ Recommended Cut Scores Plus and Minus One, Two, and Three Standard Errors with Associated Impact Data

CSAP Grade 5 Science

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

Performance Level	Unsatisfactory	Partially Proficient	Proficient	Advanced	
SE (cut score)		2.85	5.37	0.83	
Recommended Cut Point* + 3 SE		438	524	571	+ 3 SE
Percent of Students in Each Level	17.8	48.4	23.6	10.2	
Recommended Cut Point* + 2 SE		435	519	571	+ 2 SE
Percent of Students in Each Level	16.8	46.2	26.8	10.2	
Recommended Cut Point* + 1 SE		432	513	570	+ 1 SE
Percent of Students in Each Level	15.8	43.3	30.5	10.4	
Recommended Cut Point*		429	508	569	Recommended Cut Points*
Percent of Students in Each Level	14.8	40.8	33.6	10.8	
Recommended Cut Point* -1 SE		426	503	568	-1 SE
Percent of Students in Each Level	13.8	38.4	36.6	11.2	
Recommended Cut Point* -2 SE		423	497	567	-2 SE
Percent of Students in Each Level	12.9	35.3	40.3	11.5	
Recommended Cut Point* -3 SE		420	492	567	-3 SE
Percent of Students in Each Level	12.1	32.9	43.5	11.5	

* Participants' Large Group Medians

CSAP Grade 5 Science

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

Performance Level	Unsatisfactory	Partially Proficient	Proficient	Advanced	
Standard Error (SE) measurement		16.00	15.00	18.00	
Recommended Cut Point* + 3 SE		477	553	623	+ 3 SE
Percent of Students in Each Level	35.9	46.7	16.3	1.1	
Recommended Cut Point* + 2 SE		461	538	605	+ 2 SE
Percent of Students in Each Level	27.5	47.5	22.3	2.7	
Recommended Cut Point* + 1 SE		445	523	587	+ 1 SE
Percent of Students in Each Level	20.5	45.1	28.8	5.6	
Recommended Cut Point*		429	508	569	Recommended Cut Points*
Percent of Students in Each Level	14.8	40.8	33.6	10.8	
Recommended Cut Point* -1 SE		413	493	551	-1 SE
Percent of Students in Each Level	10.2	35.5	35.9	18.4	
Recommended Cut Point* -2 SE		397	478	533	-2 SE
Percent of Students in Each Level	7.0	29.5	35.4	28.1	
Recommended Cut Point* -3 SE		381	463	515	-3 SE
Percent of Students in Each Level	4.6	23.9	31.8	39.7	

* Participants' Large Group Medians

CSAP Grade 5 Science

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

Performance Level	Unsatisfactory	Partially Proficient	Proficient	Advanced	
Standard Error (SE) measurement + cutscore		16.25	15.93	18.01	
Recommended Cut Point* + 3 SE		478	556	623	+ 3 SE
Percent of Students in Each Level	36.5	47.5	14.8	1.2	
Recommended Cut Point* + 2 SE		462	540	605	+ 2 SE
Percent of Students in Each Level	28.0	48.0	21.2	2.8	
Recommended Cut Point* + 1 SE		445	524	587	+ 1 SE
Percent of Students in Each Level	20.5	45.8	28.1	5.6	
Recommended Cut Point*		429	508	569	Recommended Cut Points*
Percent of Students in Each Level	14.8	40.8	33.6	10.8	
Recommended Cut Point* -1 SE		413	492	551	-1 SE
Percent of Students in Each Level	10.2	34.8	36.6	18.4	
Recommended Cut Point* -2 SE		397	476	533	-2 SE
Percent of Students in Each Level	7.0	28.3	36.6	28.1	
Recommended Cut Point* -3 SE		380	460	515	-3 SE
Percent of Students in Each Level	4.5	22.5	33.3	39.7	

* Participants' Large Group Medians

CSAP Grade 8 Science

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

Performance Level	Unsatisfactory	Partially Proficient	Proficient	Advanced	
SE (cut score)		4.62	3.83	3.65	
Recommended Cut Point* + 3 SE		473	523	590	+ 3 SE
Percent of Students in Each Level	30.8	32.5	32.2	4.5	
Recommended Cut Point* + 2 SE		468	520	586	+ 2 SE
Percent of Students in Each Level	28.2	33.1	33.3	5.4	
Recommended Cut Point* + 1 SE		464	516	583	+ 1 SE
Percent of Students in Each Level	26.2	32.3	35.3	6.2	
Recommended Cut Point*		459	512	579	Recommended Cut Points*
Percent of Students in Each Level	23.9	31.7	37.0	7.4	
Recommended Cut Point* -1 SE		454	508	575	-1 SE
Percent of Students in Each Level	21.7	31.2	38.6	8.5	
Recommended Cut Point* -2 SE		450	504	572	-2 SE
Percent of Students in Each Level	20.1	30.0	40.5	9.4	
Recommended Cut Point* -3 SE		445	501	568	-3 SE
Percent of Students in Each Level	18.2	29.7	41.1	11.0	

* Participants' Large Group Medians

CSAP Grade 8 Science

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

Performance Level	Unsatisfactory	Partially Proficient	Proficient	Advanced	
Standard Error (SE) measurement		15.00	12.00	14.00	
Recommended Cut Point* + 3 SE		504	548	621	+ 3 SE
Percent of Students in Each Level	50.1	29.3	19.9	0.7	
Recommended Cut Point* + 2 SE		489	536	607	+ 2 SE
Percent of Students in Each Level	40.1	32.0	26.0	1.9	
Recommended Cut Point* + 1 SE		474	524	593	+ 1 SE
Percent of Students in Each Level	31.3	32.7	32.1	3.9	
Recommended Cut Point*		459	512	579	Recommended Cut Points*
Percent of Students in Each Level	23.9	31.7	37.0	7.4	
Recommended Cut Point* -1 SE		444	500	565	-1 SE
Percent of Students in Each Level	17.9	29.4	40.6	12.1	
Recommended Cut Point* -2 SE		429	488	551	-2 SE
Percent of Students in Each Level	13.4	26.1	41.5	19.0	
Recommended Cut Point* -3 SE		414	476	537	-3 SE
Percent of Students in Each Level	9.7	22.7	40.3	27.3	

* Participants' Large Group Medians

CSAP Grade 8 Science

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

Performance Level	Unsatisfactory	Partially Proficient	Proficient	Advanced	
Standard Error (SE) measurement + cutscore		15.69	12.59	14.46	
Recommended Cut Point* + 3 SE		506	550	622	+ 3 SE
Percent of Students in Each Level	51.4	29.1	18.8	0.7	
Recommended Cut Point* + 2 SE		490	537	608	+ 2 SE
Percent of Students in Each Level	40.7	32.0	25.6	1.7	
Recommended Cut Point* + 1 SE		475	525	593	+ 1 SE
Percent of Students in Each Level	31.9	33.0	31.3	3.8	
Recommended Cut Point*		459	512	579	Recommended Cut Points*
Percent of Students in Each Level	23.9	31.7	37.0	7.4	
Recommended Cut Point* -1 SE		443	499	565	-1 SE
Percent of Students in Each Level	17.6	29.0	41.3	12.1	
Recommended Cut Point* -2 SE		428	487	550	-2 SE
Percent of Students in Each Level	13.1	25.8	41.6	19.5	
Recommended Cut Point* -3 SE		412	474	536	-3 SE
Percent of Students in Each Level	9.3	22.0	40.8	27.9	

* Participants' Large Group Medians

CSAP Grade 10 Science

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

Performance Level	Unsatisfactory	Partially Proficient	Proficient	Advanced	
SE (cut score)		6.22	10.66	7.52	
Recommended Cut Point* + 3 SE		488	539	604	+ 3 SE
Percent of Students in Each Level	36.9	36.8	24.5	1.8	
Recommended Cut Point* + 2 SE		481	528	596	+ 2 SE
Percent of Students in Each Level	32.4	33.7	31.1	2.8	
Recommended Cut Point* + 1 SE		475	518	589	+ 1 SE
Percent of Students in Each Level	29.1	29.3	37.5	4.1	
Recommended Cut Point*		469	507	581	Recommended Cut Points*
Percent of Students in Each Level	26.0	24.0	43.9	6.1	
Recommended Cut Point* -1 SE		463	496	573	-1 SE
Percent of Students in Each Level	23.1	19.0	49.3	8.6	
Recommended Cut Point* -2 SE		457	486	566	-2 SE
Percent of Students in Each Level	20.6	15.0	53.1	11.3	
Recommended Cut Point* -3 SE		450	475	558	-3 SE
Percent of Students in Each Level	18.0	11.1	56.0	14.9	

* Participants' Large Group Medians

CSAP Grade 10 Science

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

Performance Level	Unsatisfactory	Partially Proficient	Proficient	Advanced	
Standard Error (SE) measurement		14.00	13.00	12.00	
Recommended Cut Point* + 3 SE		511	546	617	+ 3 SE
Percent of Students in Each Level	53.0	25.2	21.0	0.8	
Recommended Cut Point* + 2 SE		497	533	605	+ 2 SE
Percent of Students in Each Level	42.8	26.7	28.8	1.7	
Recommended Cut Point* + 1 SE		483	520	593	+ 1 SE
Percent of Students in Each Level	33.7	26.3	36.7	3.3	
Recommended Cut Point*		469	507	581	Recommended Cut Points*
Percent of Students in Each Level	26.0	24.0	43.9	6.1	
Recommended Cut Point* -1 SE		455	494	569	-1 SE
Percent of Students in Each Level	19.9	21.0	49.1	10.0	
Recommended Cut Point* -2 SE		441	481	557	-2 SE
Percent of Students in Each Level	15.0	17.4	52.2	15.4	
Recommended Cut Point* -3 SE		427	468	545	-3 SE
Percent of Students in Each Level	11.1	14.4	52.1	22.4	

* Participants' Large Group Medians

CSAP Grade 10 Science

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

Performance Level	Unsatisfactory	Partially Proficient	Proficient	Advanced	
Standard Error (SE) measurement + cutscore		15.31	16.81	14.16	
Recommended Cut Point* + 3 SE		515	557	623	+ 3 SE
Percent of Students in Each Level	56.0	28.6	14.9	0.5	
Recommended Cut Point* + 2 SE		500	541	609	+ 2 SE
Percent of Students in Each Level	45.0	30.1	23.6	1.3	
Recommended Cut Point* + 1 SE		484	524	595	+ 1 SE
Percent of Students in Each Level	34.3	28.7	34.0	3.0	
Recommended Cut Point*		469	507	581	Recommended Cut Points*
Percent of Students in Each Level	26.0	24.0	43.9	6.1	
Recommended Cut Point* -1 SE		454	490	567	-1 SE
Percent of Students in Each Level	19.5	18.7	51.0	10.8	
Recommended Cut Point* -2 SE		438	473	553	-2 SE
Percent of Students in Each Level	14.1	14.0	54.4	17.5	
Recommended Cut Point* -3 SE		423	457	539	-3 SE
Percent of Students in Each Level	10.1	10.5	53.0	26.4	

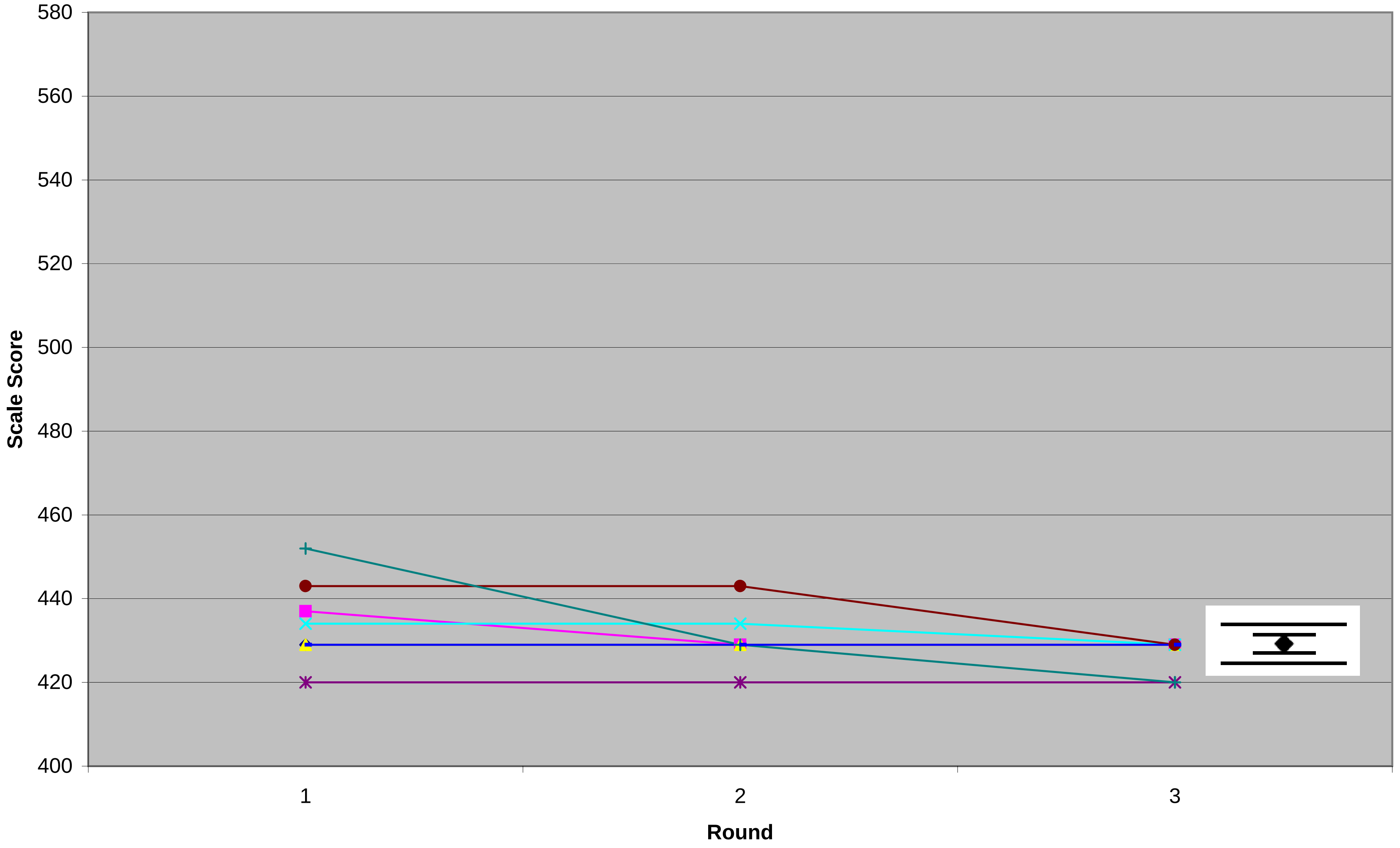
* Participants' Large Group Medians

Section G

Graphical Representations of Participants’ Judgments and Standard Errors

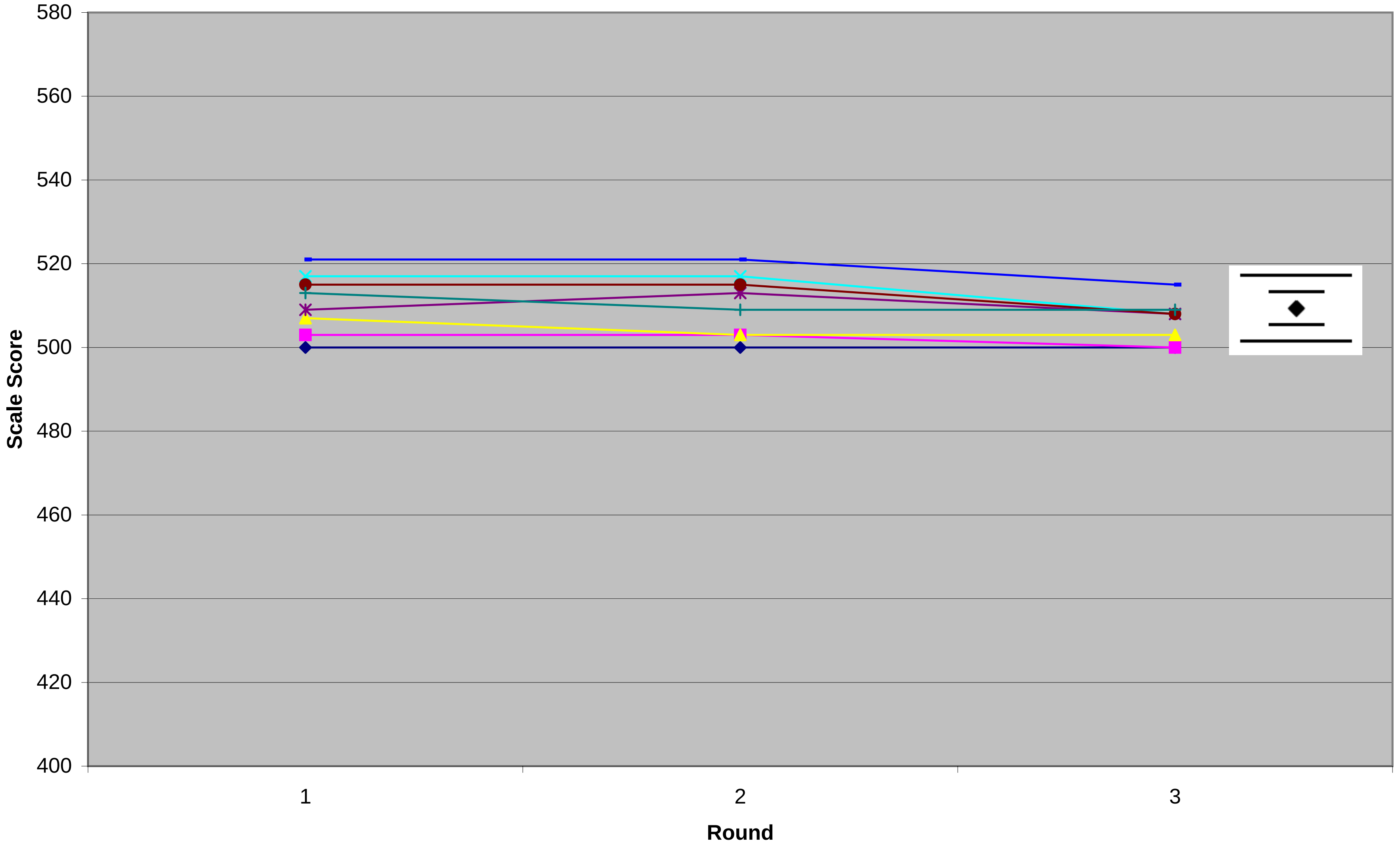
CSAP Grade 5 Science Partially Proficient Cut Point

SEBk = 2.85; r = 0.00



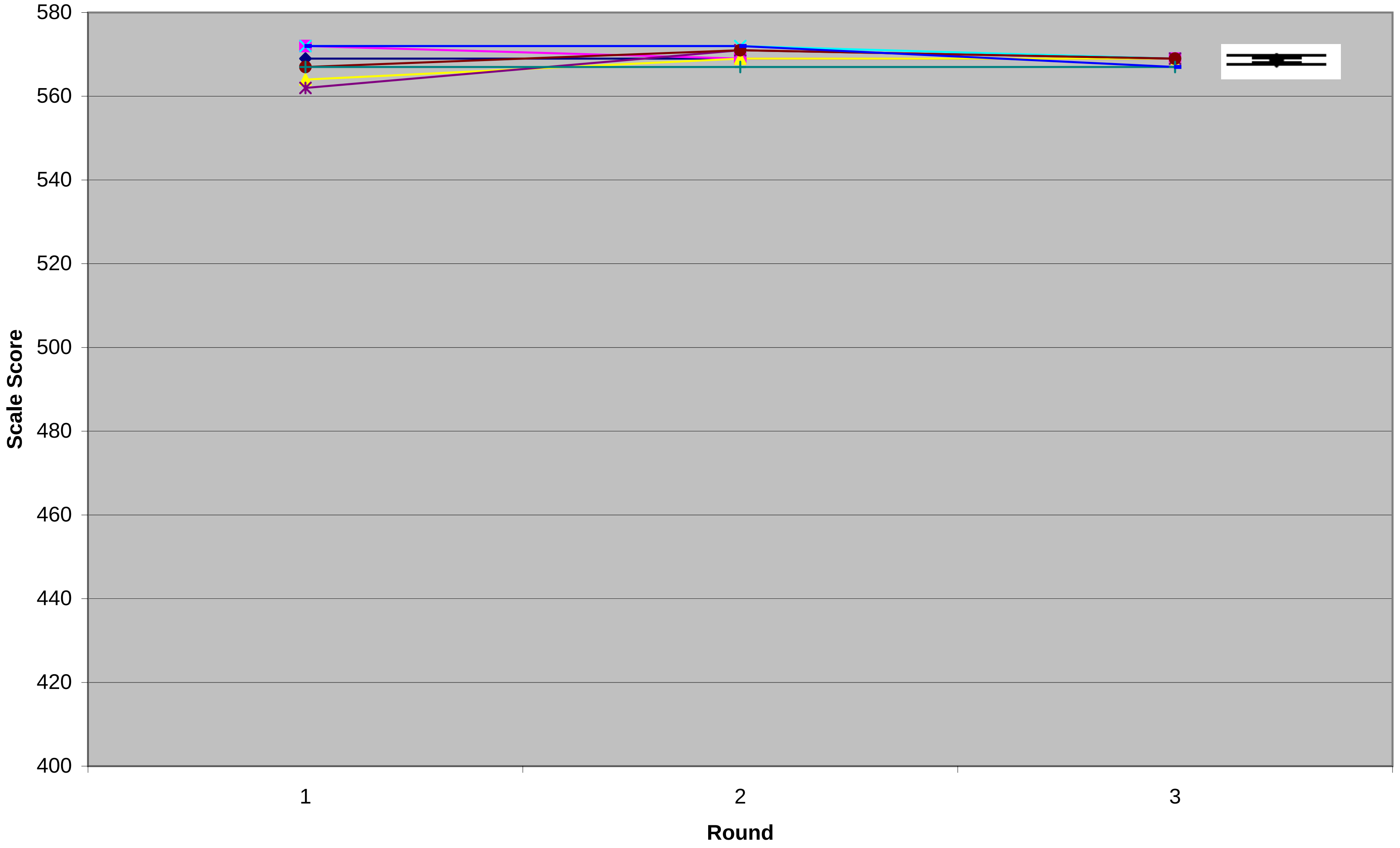
CSAP Grade 5 Science Proficient Cut Point

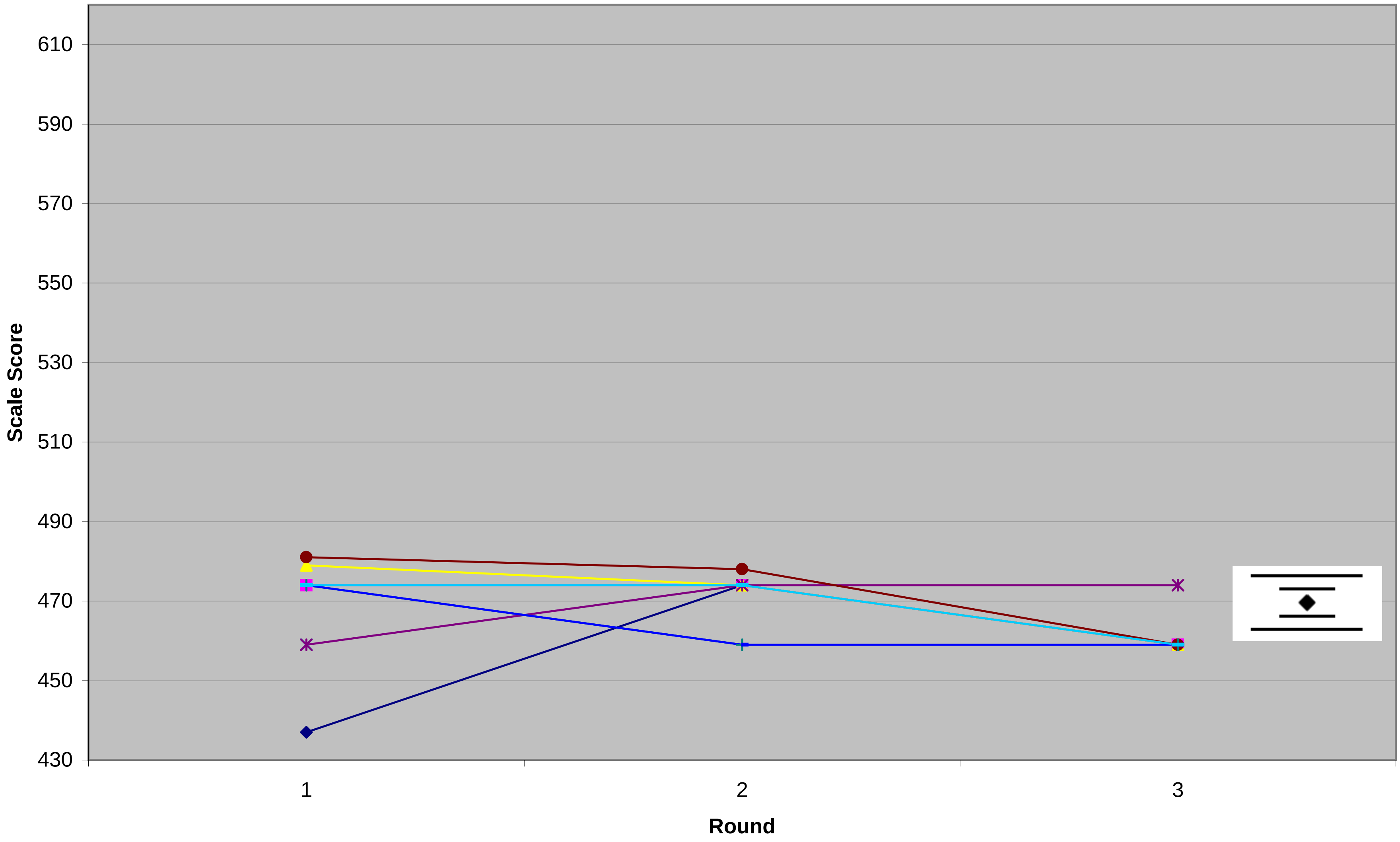
SEBk = 5.37; r = 0.52



CSAP Grade 5 Science Advanced Cut Point

SEBk = 0.83; r = 0.04



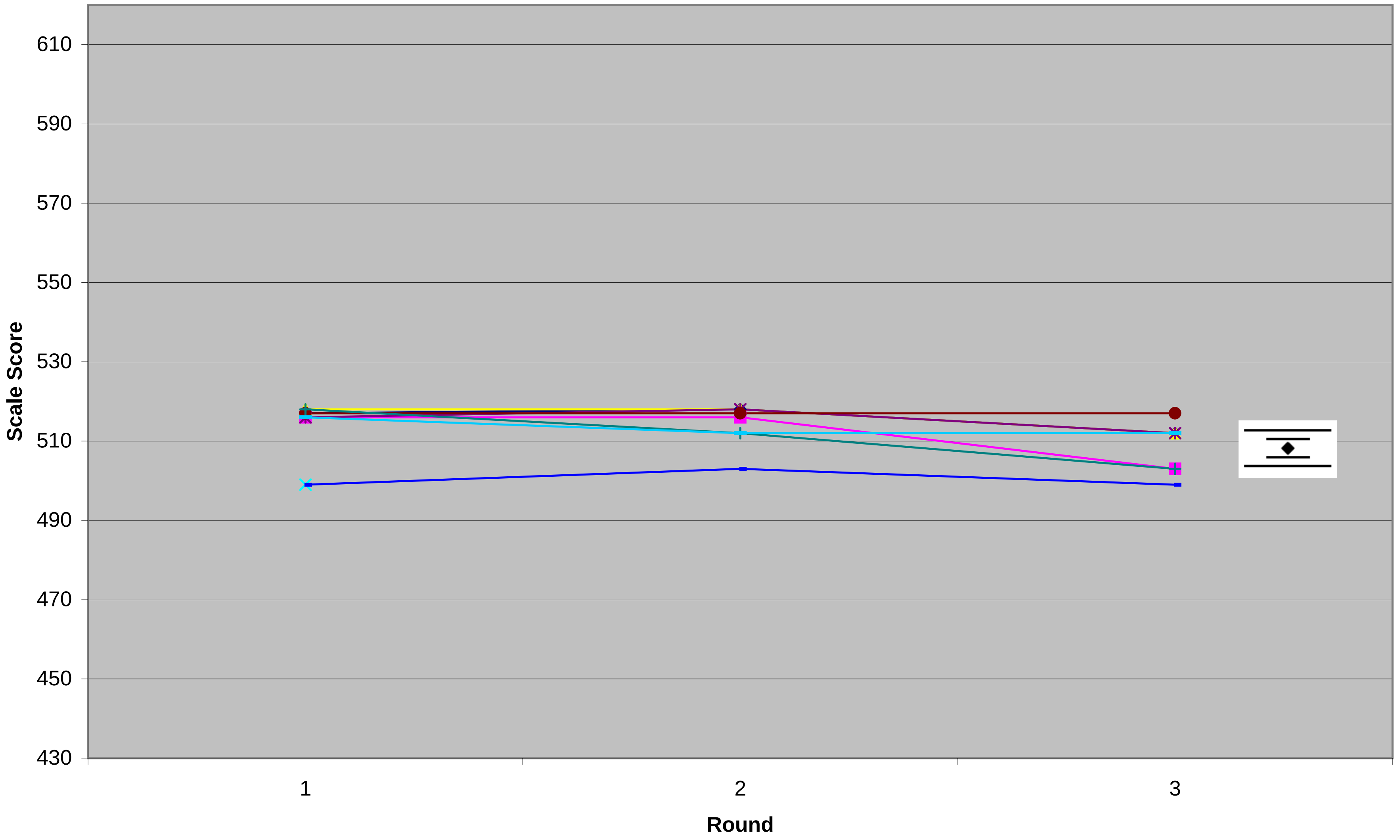


CSAP Grade 8 Science Partially Proficient Cut Point

SEBk = 4.62; r = 0.33

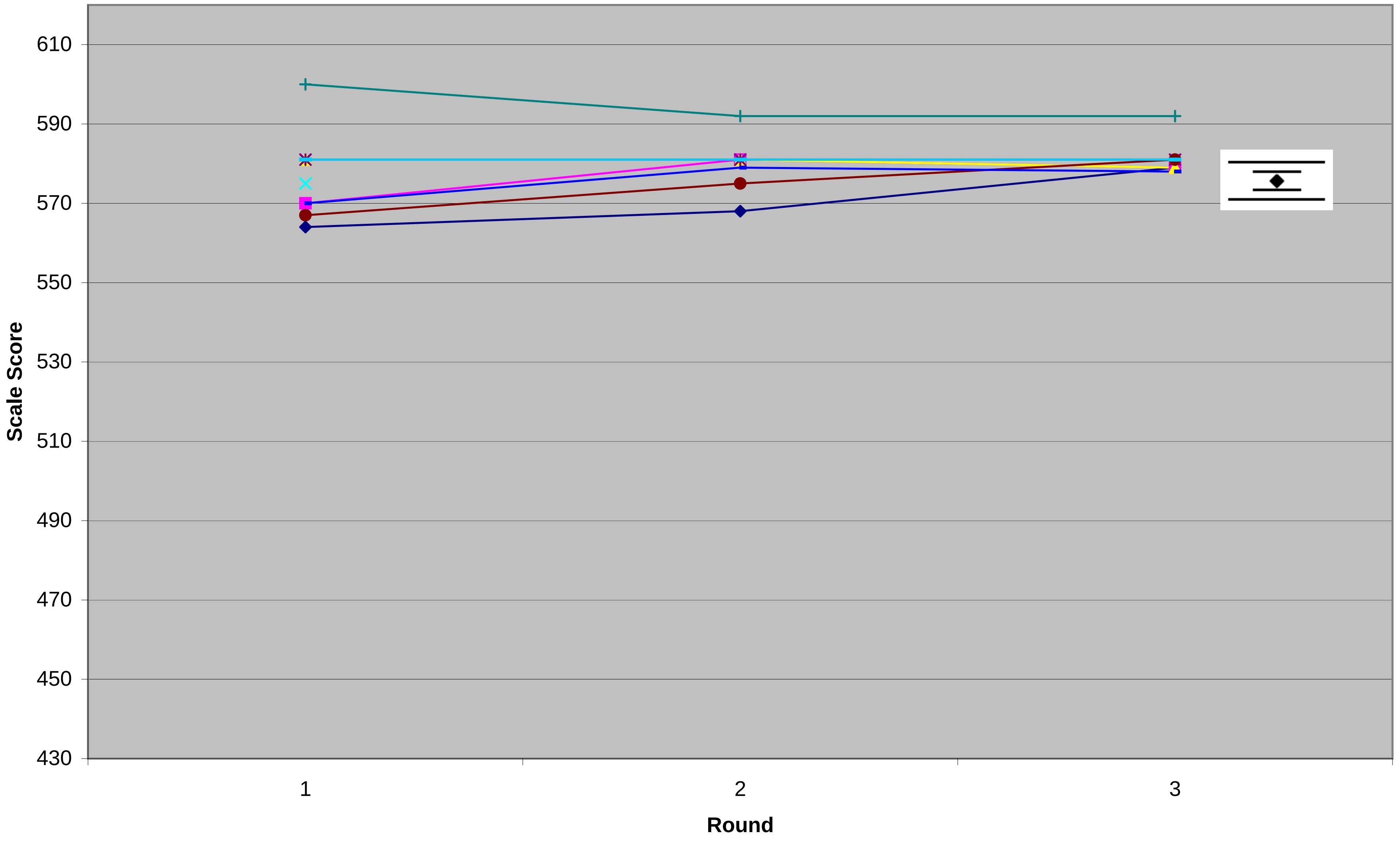
CSAP Grade 8 Science Proficient Cut Point

SEBk = 3.83; r = 0.58



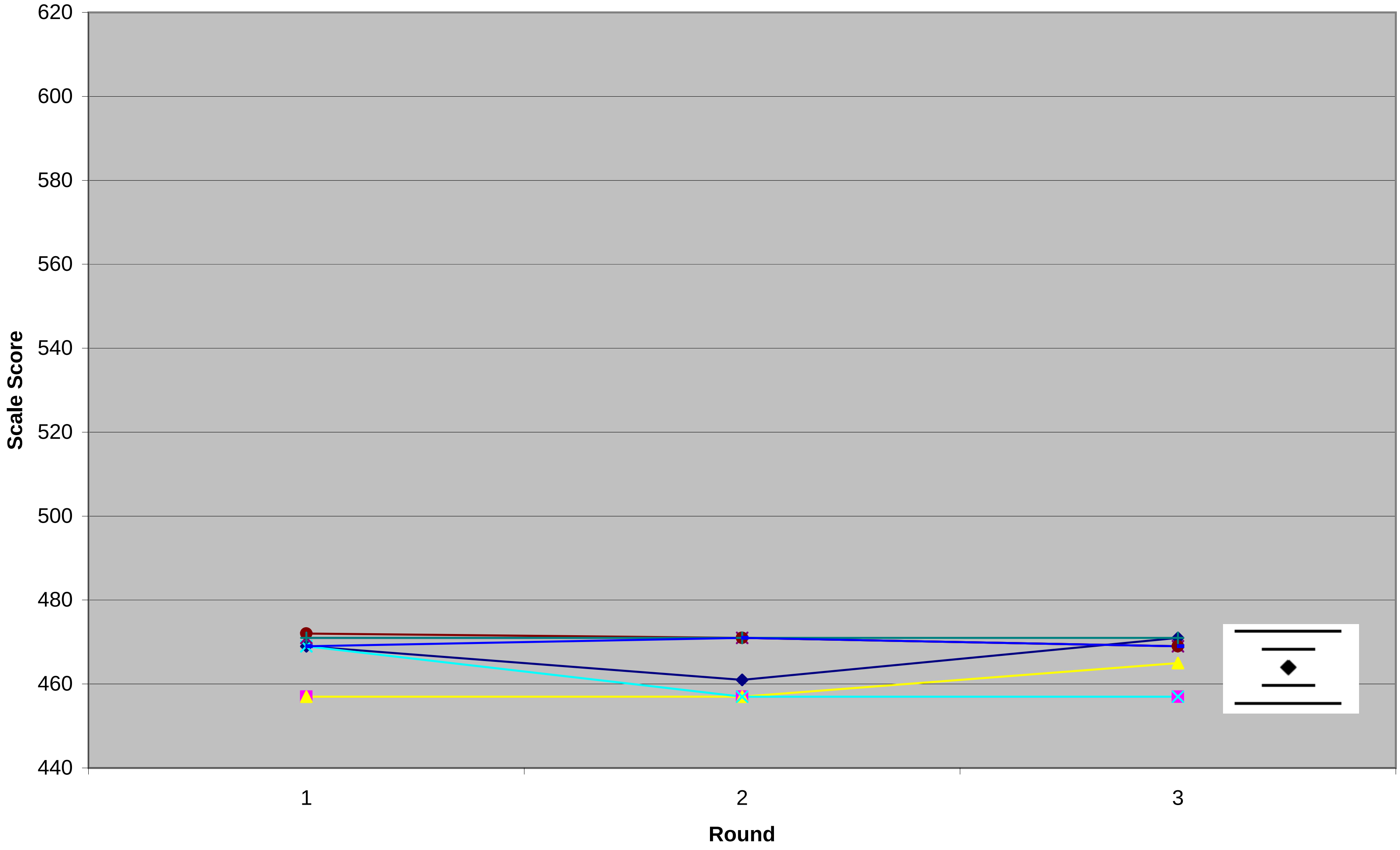
CSAP Grade 8 Science Advanced Cut Point

SEBk = 3.65; r = 0.16



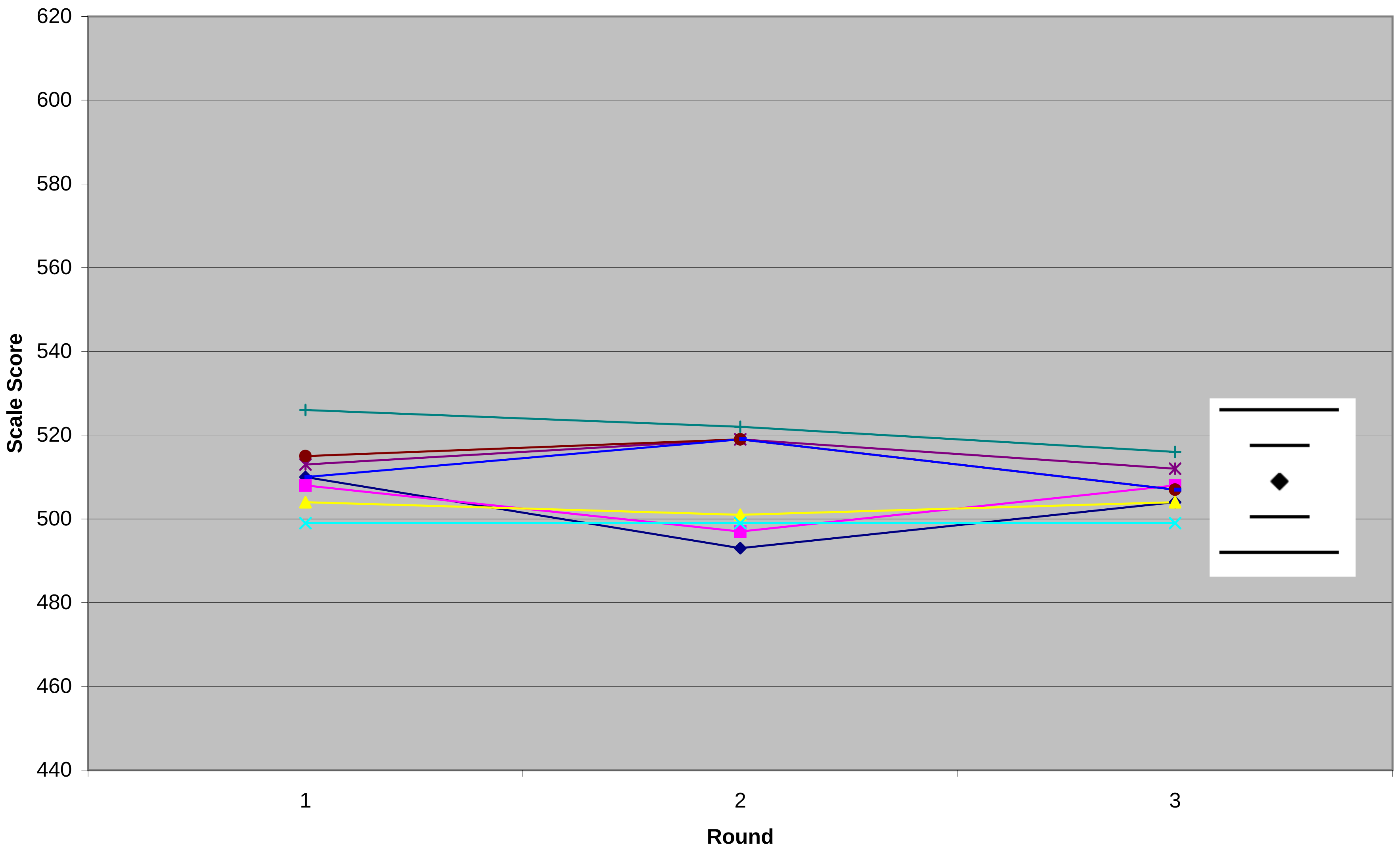
CSAP Grade 10 Science Partially Proficient Cut Point

SEBk = 6.22; r = 0.98



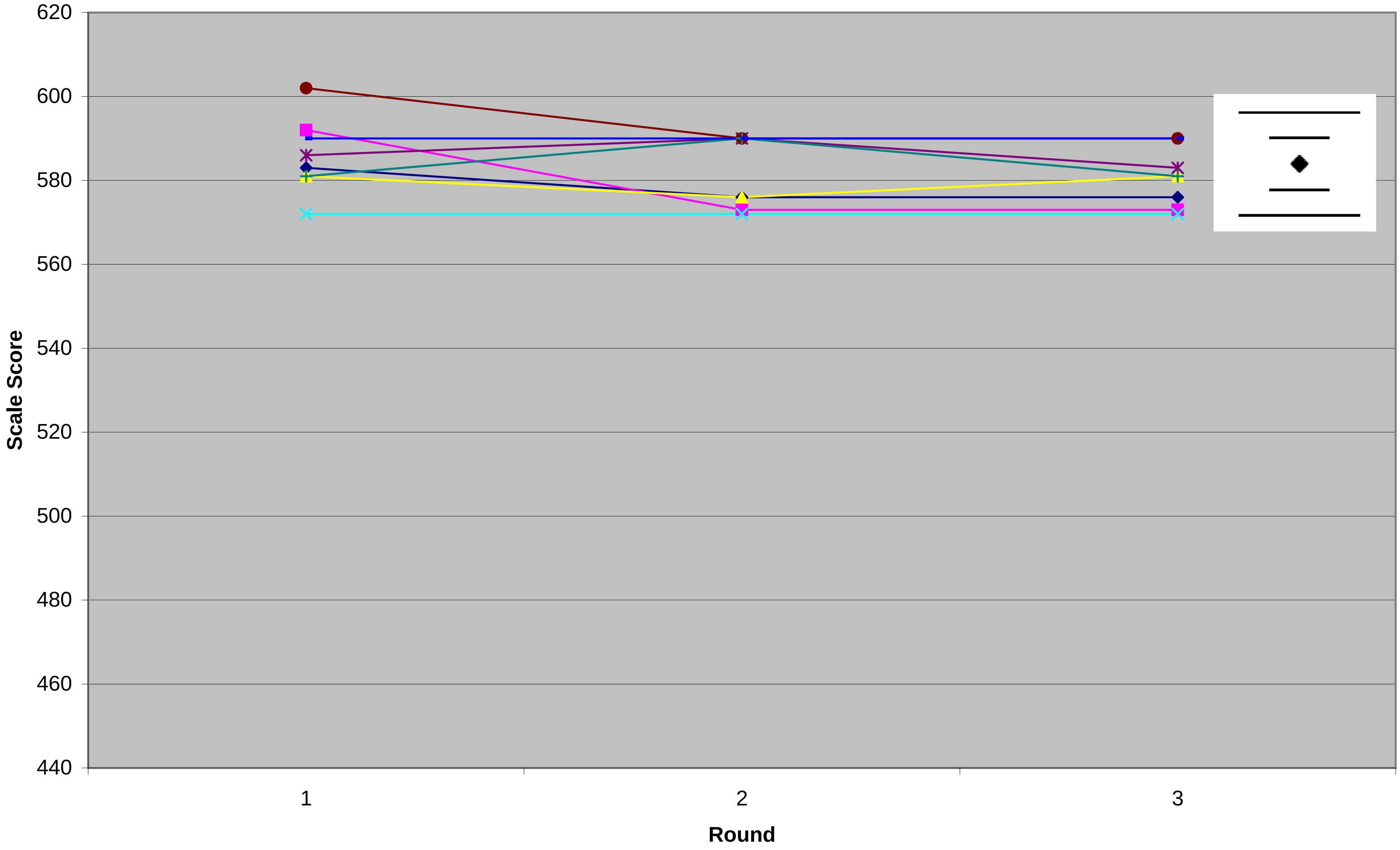
CSAP Grade 10 Science Proficient Cut Point

SEBk = 10.66; r = 0.98



CSAP Grade 10 Science Advanced Cut Point

SEBk = 7.52; r = 0.99



Section H

Participant Training Materials

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/score point?	Why is this item more difficult than the preceding items?
1	220	12	1	MC	2	1		**N/A**
2	225	9	4	MC	3	4		
3	229	9	3	MC	2	5		
4	240	12	2	MC	4	1		
5	241	12	4	MC	2	4		
6	256	12	7	CR	1/2	1		
7	262	9	5	MC	1	1		
8	282	12	7	CR	2/2	1		
9	303	9	6	MC	2	2		
10	321	9	8	MC	2	2		
11	401	9	9	MC	3	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?

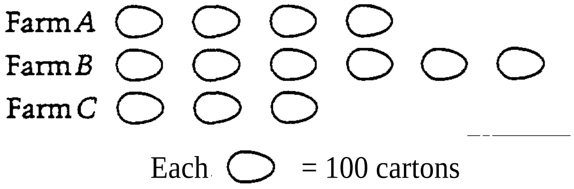
Ⓐ 574 miles

Ⓑ 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?

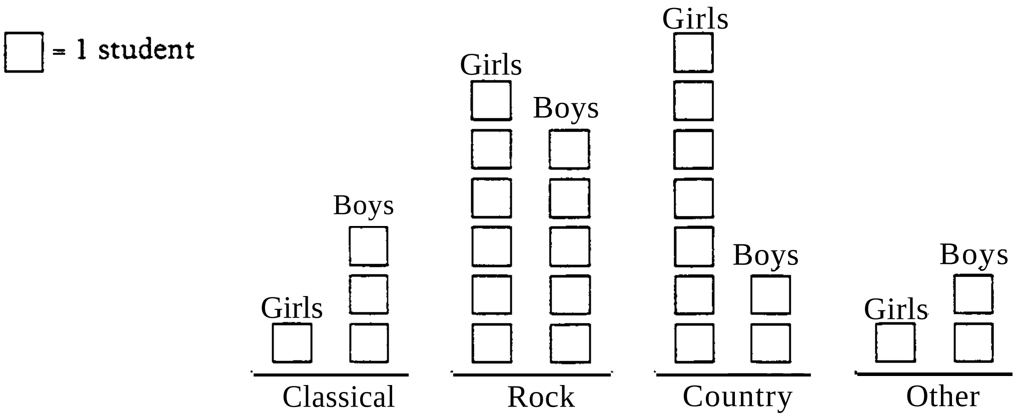
Ⓐ 652

Ⓑ 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

7. Sam can purchase his lunch at school. Each day he wants to have juice that costs 50¢, a sandwich that costs 90¢, and fruit that costs 35¢. His mother has only $1.00 bills. What is the least number of $1.00 bills that his mother should give him so he will have enough money to buy lunch for 5 days?

6 rubric

Item Number: 7 Accession Number: AP000522

Key: None

Classification Codes:
N25M 1 A 04 a PS RECM 02

Open Codes: NA NA NA 3

Rationale Text:
SOLUTION:

For one day the sum is $1.75. For 5 days, the sum is $8.75. Therefore he should ask his mother for nine one-dollar bills (or 1 $5 bill and 4 $1 bills)

Answer may be given pictorially.

Note: No explanation is asked for, so paper could have small error, such as copying a number incorrectly and still get a score of 2, provided method and answer are correct.

Scoring Guide:

0 Incorrect response -- includes $1.75 or $2: also $875 or $875.00

[✓ 1] $8.75 or 875
OR
One day is $1.75 so he needs $2 each day, so $10 for a week
(picture of $10 bill is acceptable)
OR
correct method but rounded down to $8 (this requires work to be shown)
OR
correct method but small error and incorrect response of $7 to $11, inclusive

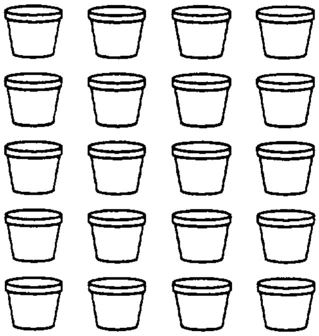
2 Correct response

6 exemplar

Student Sample Response

7. Sam can purchase his lunch at school. Each day he wants to have juice that costs 50¢, a sandwich that costs 90¢, and fruit that costs 35¢. His mother has only $1.00 bills. What is the least number of $1.00 bills that his mother should give him so he will have enough money to buy lunch for 5 days? $8.75

Level:
Partial (1)



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 = □

8
Score Point
2 of 2

7. Sam can purchase his lunch at school. Each day he wants to have juice that costs 50¢, a sandwich that costs 90¢, and fruit that costs 35¢. His mother has only $1.00 bills. What is the least number of $1.00 bills that his mother should give him so he will have enough money to buy lunch for 5 days?

8 rubric

Item Number: 7 Accession Number: AP000522

Key: None

Classification Codes:
N25M 1 A 04 a PS RECM 02

Open Codes: NA NA NA 3

Rationale Text:
SOLUTION:

For one day the sum is $1.75. For 5 days, the sum is $ 8.75 Therefore he should ask his mother for nine one-dollar bills (or 1 $5 bill and 4 $1 bills)

Answer may be given pictorially.

Note: No explanation is asked for, so paper could have small error, such as copying a number incorrectly and still get a score of 2, provided method and answer are correct.

Scoring Guide:

0 Incorrect response -- includes $1.75 or $2: also $875 or $875.00

1 $8.75 or 875
OR
One day is $1.75 so he needs $2 each day, so $10 for a week
(picture of $10 bill is acceptable)
OR
correct method but rounded down to $8 (this requires work to be shown)
OR
correct method but small error and incorrect response of $7 to $11, inclusive

[✓ 2] Correct response

8 exemplar

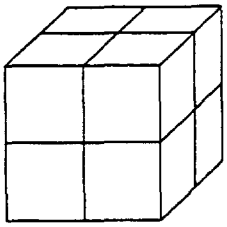
Student Sample Response

7. Sam can purchase his lunch at school. Each day he wants to have juice that costs 50¢, a sandwich that costs 90¢, and fruit that costs 35¢. His mother has only $1.00 bills. What is the least number of $1.00 bills that his mother should give him so he will have enough money to buy lunch for 5 days?

$$\begin{array}{r} \text{¢}50 \\ \text{¢}90 \\ +\text{¢}35 \\ \hline \overset{2}{\$1.75} \\ \times 5 \\ \hline \$8.75 \end{array}$$

9 dollar bills

Level:
Complete (2)



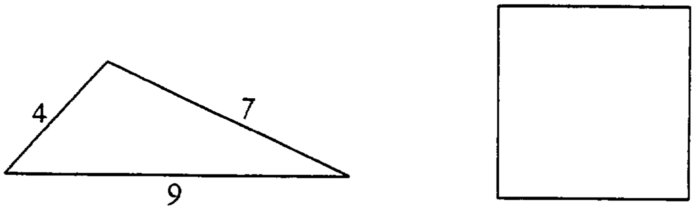
6. In this figure, how many small cubes were put together to form the large cube?

Ⓐ 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

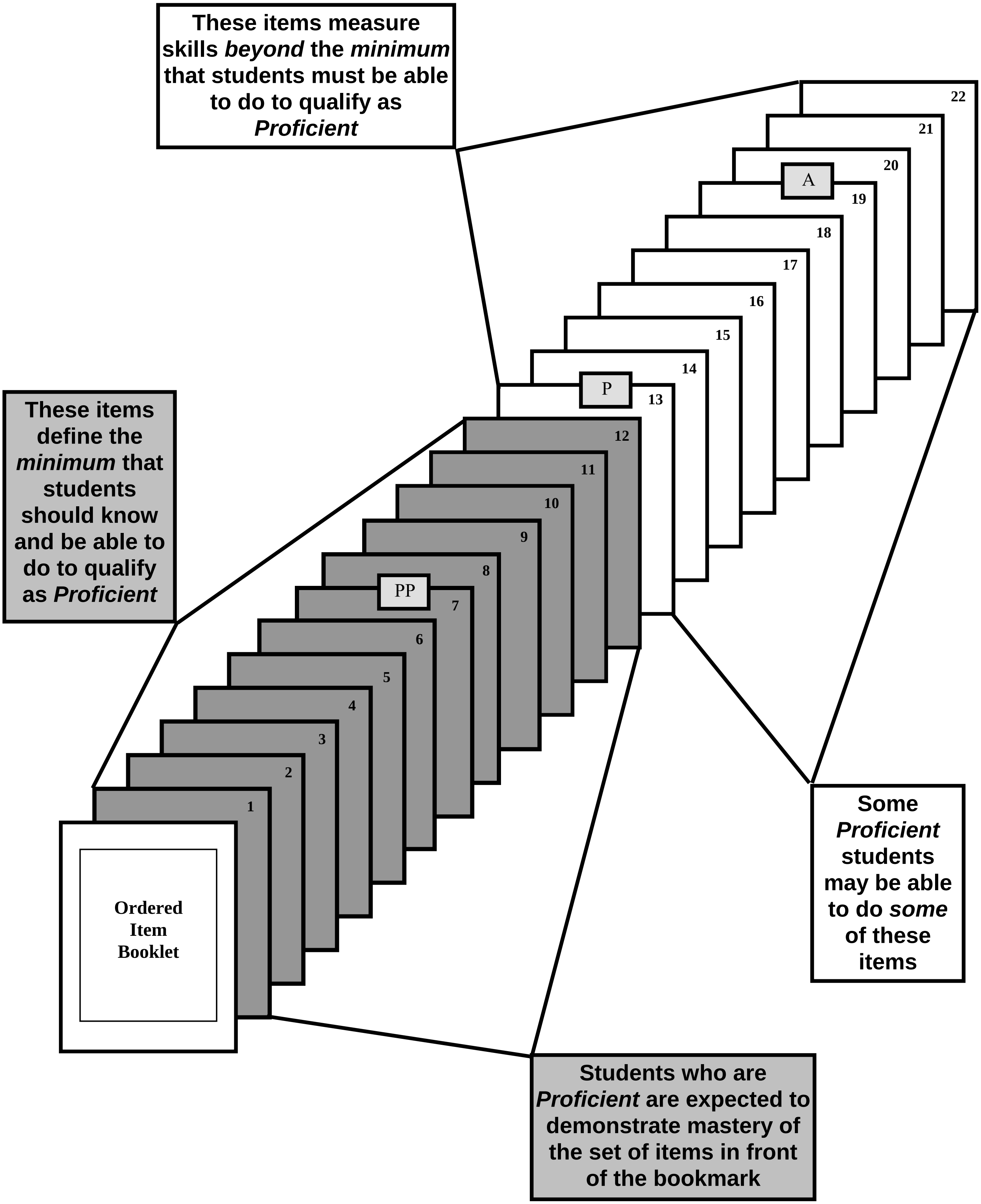
9. There are 3 fifth graders and 2 sixth graders on the swim team. Everyone’s name is put in a hat and the captain is chosen by picking one name. What are the chances that the captain will be a fifth grader?

Ⓐ l out of 5

Ⓑ l out of 3

Ⓒ 3 out of 5

Ⓓ 2 out of 3



These items measure skills *beyond* the *minimum* that students must be able to do to qualify as *Proficient*

These items define the *minimum* that students should know and be able to do to qualify as *Proficient*

Some *Proficient* students may be able to do *some* of these items

Students who are *Proficient* are expected to demonstrate mastery of the set of items in front of the bookmark

Bookmark Placement

These directions are written for placing the *Proficient* bookmark and apply analogously to the *Partially Proficient* and *Advanced* bookmarks.

For whom am I placing this bookmark? The Target Student

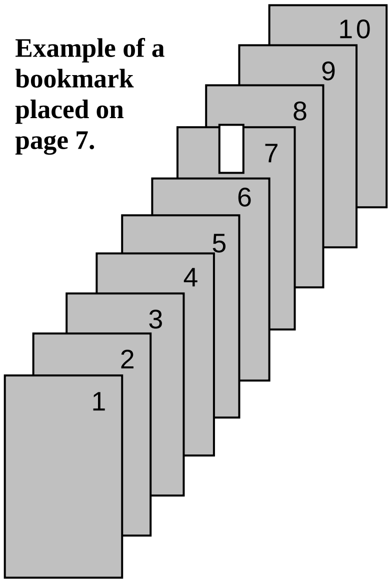
When you place your *Proficient* bookmark, you are separating the highest ability *Partially Proficient* students from the lowest ability *Proficient* students. In other words, you are keeping in mind the Target Student who will just make it into the *Proficient* level.

How do I place my bookmark? The Mechanics

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

To place your bookmark, start at page 1 in the Ordered Item Booklet (OIB). Page through the OIB **looking at the content 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 *Proficient* relative to the content standards. This is the content you are saying a *Proficient* Target Student needs to master to just make it into the *Proficient* level.

Example of a bookmark placed on page 7.



Hold the pages that contain the content 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 Rating 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 *Proficient* Bookmark mean? Some Answers

- You expect *Proficient* students to master the knowledge, skills, and abilities contained in the items *before* your bookmark.
- *Proficient* students should know and be able to do the items *before* the bookmark. For multiple-choice items, *Proficient* students should know the correct response. For constructed-response items, *Proficient* students should most likely achieve the score points before the bookmark.

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 *Proficient* bookmark was placed on page 7. The participant was *not* saying that a student must get six items correct to be classified as *Proficient*. This participant is saying that a barely *Proficient* student must master the content 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 *Proficient* bookmark and apply analogously to the *Partially Proficient* and *Advanced* 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 content you expect students to know and be able to do.

I set my bookmark based on the content I expect students to know and be able to do, that is, the content 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 *Proficient* bookmark and gets some correct after the bookmark, is that student still *Proficient?*

A student does *not* have to get every item before the bookmark correct to be *Proficient*. *Proficient* students 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?

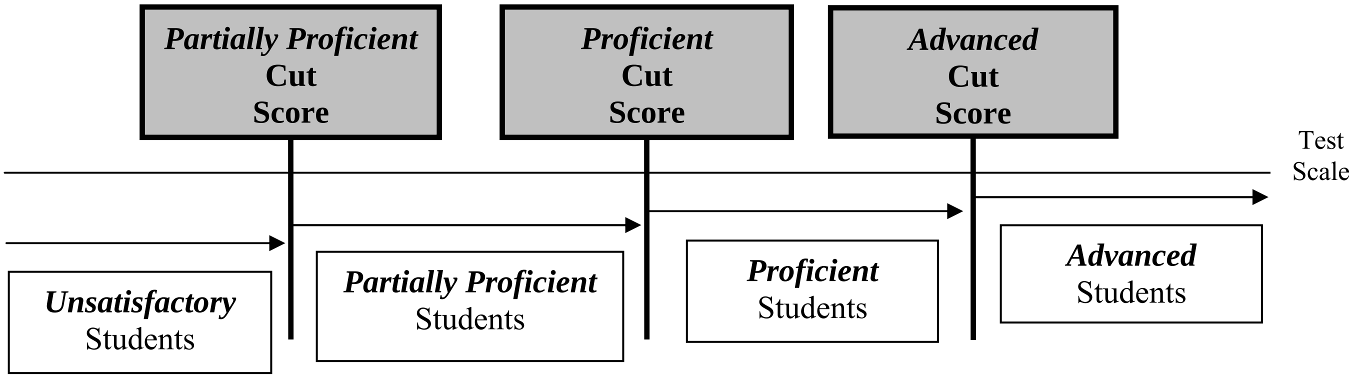
No. Remember, you are placing your bookmark based on the content you expect students to know and be able to do. You are *not* making your decision based on the number of items students must answer correctly. The bookmark is placed on a *page* in the Ordered Item Booklet. This page number corresponds 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 content domain. In some cases, some content 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. In Colorado, you will set three bookmarks to separate students into four performance levels.



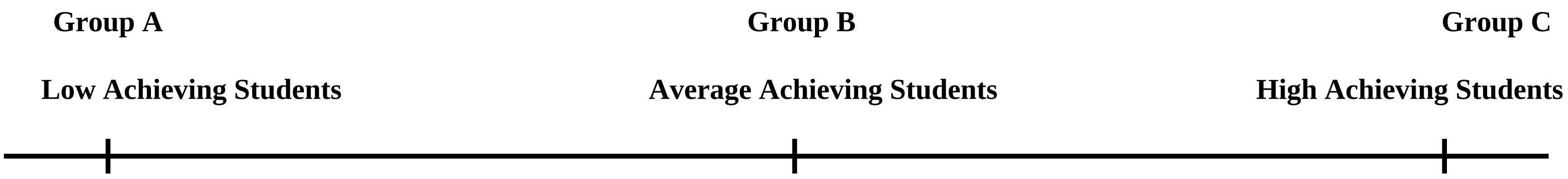
MASTERY

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

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

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 content 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
$\frac{99}{100}$	$\frac{95}{100}$	$\frac{93}{100}$	$\frac{87}{100}$	$\frac{83}{100}$	$\frac{82}{100}$	$\frac{78}{100}$	$\frac{74}{100}$	$\frac{69}{100}$	$\frac{67}{100}$

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
$\frac{63}{100}$	$\frac{60}{100}$	$\frac{59}{100}$	$\frac{58}{100}$	$\frac{57}{100}$	$\frac{53}{100}$	$\frac{52}{100}$	$\frac{50}{100}$	$\frac{50}{100}$	$\frac{49}{100}$	$\frac{49}{100}$	$\frac{48}{100}$	$\frac{47}{100}$	$\frac{43}{100}$	$\frac{41}{100}$	$\frac{39}{100}$	$\frac{37}{100}$	$\frac{35}{100}$	$\frac{34}{100}$	$\frac{31}{100}$

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
$\frac{30}{100}$	$\frac{29}{100}$	$\frac{25}{100}$	$\frac{22}{100}$	$\frac{20}{100}$	$\frac{18}{100}$	$\frac{17}{100}$	$\frac{14}{100}$	$\frac{11}{100}$	$\frac{10}{100}$	$\frac{9}{100}$	$\frac{7}{100}$	$\frac{5}{100}$	$\frac{5}{100}$	$\frac{4}{100}$	$\frac{3}{100}$	$\frac{2}{100}$	$\frac{2}{100}$	$\frac{1}{100}$	$\frac{1}{100}$

Definition of Mastery

We say that a group of like students have demonstrated mastery of the content 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 content reflected in items 1 through 30 of the ordered item booklet, but have not demonstrated mastery of the content 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
$\frac{99}{100}$	$\frac{99}{100}$	$\frac{99}{100}$	$\frac{99}{100}$	$\frac{99}{100}$	$\frac{98}{100}$	$\frac{98}{100}$	$\frac{98}{100}$	$\frac{97}{100}$	$\frac{97}{100}$

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
$\frac{96}{100}$	$\frac{96}{100}$	$\frac{95}{100}$	$\frac{93}{100}$	$\frac{89}{100}$	$\frac{85}{100}$	$\frac{84}{100}$	$\frac{83}{100}$	$\frac{83}{100}$	$\frac{81}{100}$	$\frac{79}{100}$	$\frac{79}{100}$	$\frac{78}{100}$	$\frac{73}{100}$	$\frac{72}{100}$	$\frac{72}{100}$	$\frac{71}{100}$	$\frac{70}{100}$	$\frac{69}{100}$	$\frac{67}{100}$

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
$\frac{65}{100}$	$\frac{63}{100}$	$\frac{63}{100}$	$\frac{61}{100}$	$\frac{58}{100}$	$\frac{57}{100}$	$\frac{57}{100}$	$\frac{55}{100}$	$\frac{55}{100}$	$\frac{54}{100}$	$\frac{53}{100}$	$\frac{53}{100}$	$\frac{52}{100}$	$\frac{51}{100}$	$\frac{44}{100}$	$\frac{41}{100}$	$\frac{39}{100}$	$\frac{37}{100}$	$\frac{35}{100}$	$\frac{33}{100}$

Definition of Mastery

We say that a group of like students have demonstrated mastery of the content 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 content reflected in items 1 through 45 of the ordered item booklet, but have not demonstrated mastery of the content 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
$\frac{99}{100}$	$\frac{99}{100}$	$\frac{99}{100}$	$\frac{99}{100}$	$\frac{99}{100}$	$\frac{99}{100}$	$\frac{99}{100}$	$\frac{99}{100}$	$\frac{97}{100}$	$\frac{97}{100}$

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
$\frac{97}{100}$	$\frac{97}{100}$	$\frac{95}{100}$	$\frac{95}{100}$	$\frac{94}{100}$	$\frac{93}{100}$	$\frac{92}{100}$	$\frac{92}{100}$	$\frac{91}{100}$	$\frac{89}{100}$	$\frac{89}{100}$	$\frac{89}{100}$	$\frac{88}{100}$	$\frac{88}{100}$	$\frac{88}{100}$	$\frac{87}{100}$	$\frac{87}{100}$	$\frac{86}{100}$	$\frac{85}{100}$	$\frac{84}{100}$

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
$\frac{83}{100}$	$\frac{81}{100}$	$\frac{81}{100}$	$\frac{81}{100}$	$\frac{80}{100}$	$\frac{80}{100}$	$\frac{79}{100}$	$\frac{78}{100}$	$\frac{77}{100}$	$\frac{75}{100}$	$\frac{74}{100}$	$\frac{72}{100}$	$\frac{70}{100}$	$\frac{68}{100}$	$\frac{67}{100}$	$\frac{64}{100}$	$\frac{58}{100}$	$\frac{53}{100}$	$\frac{49}{100}$	$\frac{46}{100}$

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

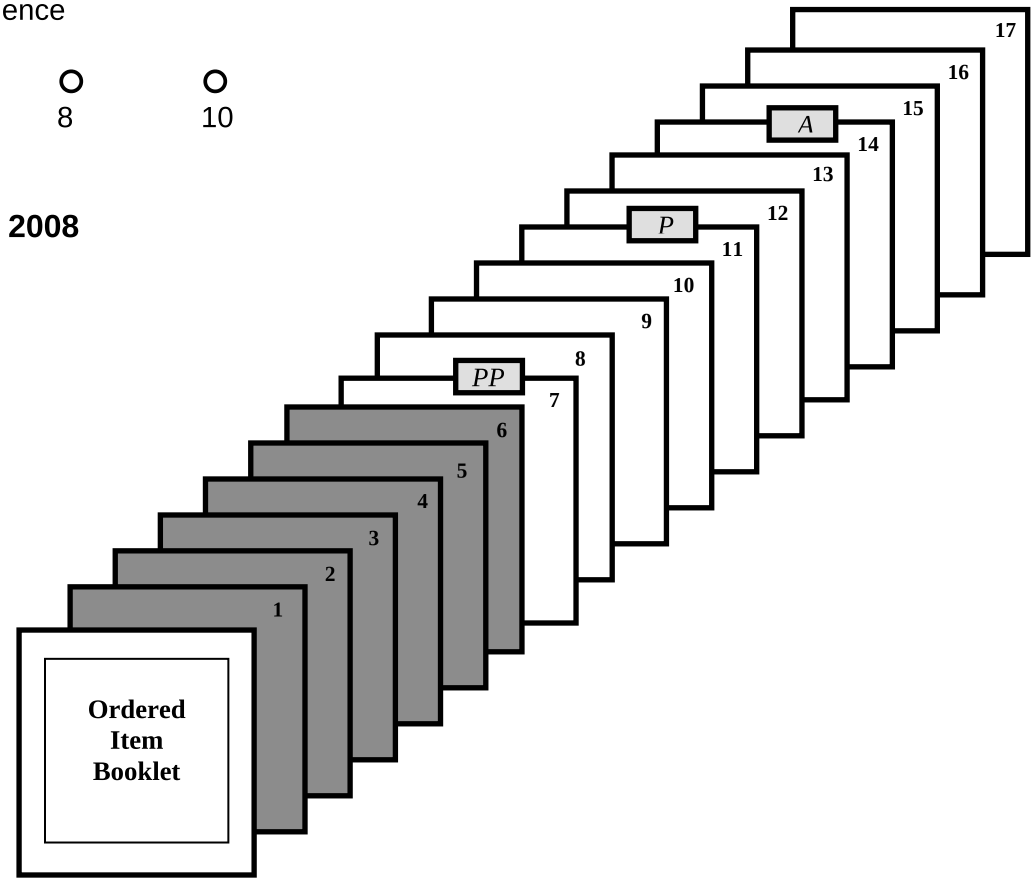
Also, if *you* identify a set of items in the ordered item booklet, the specific point on the test scale at which students have demonstrated mastery of the content you have specified can be determined. This is how the various cut scores are ascertained.

As experts, you will first specify the content in the ordered item booklet that you expect students to master in order to be classified as ***Proficient***. This means that you will identify the items that reflect the knowledge, skills, and abilities you expect all ***Proficient*** 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.

Content Area: ○ Science

Grade: ○ 5 ○ 8 ○ 10

Colorado Science 2008



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

	Partially Proficient **Bookmark on Page #**	***Proficient*** **Bookmark on Page #**	***Advanced*** **Bookmark on Page #**
Round 1	**7**	**11**	**14**

1. Which items does a student need to master to just make it into the *Partially Proficient* performance level?
   ○ 1 to 5 ○ 1 to 6 ○ 1 to 7 ○ 1 to 8

2. If a student mastered only items 1 through 5, in which performance level would this student be?
   ○ *Unsatisfactory* ○ *Partially Proficient* ○ *Proficient* ○ *Advanced*

3. Suppose a student mastered items 1 through 6. Which performance level is this student in?
   ○ *Unsatisfactory* ○ *Partially Proficient* ○ *Proficient* ○ *Advanced*

4. For students who are classified as *Partially Proficient*, with at least what likelihood will they be able to answer item 6?
   ○ 1/3 ○ 1/2 ○ 2/3 ○ 3/4

   Will the items BEFORE the *Partially Proficient* 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 I

Performance Level Descriptors

CSAP Grade 5 – Performance Level Descriptors

Unsatisfactory
Students have a very limited understanding of scientific inquiry processes, as well as life, physical, and earth and space science concepts and vocabulary.

Partially Proficient
Identify appropriate scientific tools used to gather data for an investigation; identify various types of energy and their common sources; recognize that an electrical circuit must be complete to function; describe how animals use food for growth and energy; identify organ systems and major organs; describe the function of various human body systems; sequence the stages of growth of plants/animals; describe ways that plants/animals of the same population and life stage look different; recognize the majority of Earth's surface is covered by salt/fresh water; predict results of experiments when repeated.

Proficient
Identify effects when a variable changes; make conclusions/predictions; identify metric units; show data visually; use tools; explain atoms make up matter; describe states of matter; recognize effects of forces; describe plant/animal structures and effect of ecosystem interactions; identify organisms with similar life stages; classify organisms; describe fossils show change; describe weathering/erosion/deposition; compare weather/climate; identify water cycle parts; describe uses of natural resources/benefit of conserving; describe effects of Earth’s motion; know repetition verifies results; identify model uses.

Advanced
Identify a testable question/state a hypothesis; use data to predict how an event affects the physical property of an object; describe how melting, freezing, evaporation, and condensation change matter from one state to another; describe a force is a push/pull; identify gravity, magnetism, and friction as forces; explain multiple forces may act on an object at the same time; evaluate changes in speed/direction of motion caused by unbalanced forces; predict or infer how fossils are formed from previously living organisms; explain the contribution of volcanic/earthquake activity to changes on Earth's surface.

CSAP Grade 8 – Performance Level Descriptors

Unsatisfactory

Students have a very limited understanding of scientific inquiry processes, as well as life, physical, and earth and space science concepts and vocabulary.

Partially Proficient

Identify independent/dependent variables; record data using tools/units; distinguish physical/chemical changes; describe what makes up white light; identify classifying characteristics; differentiate between animal/plant/single cell organisms; classify communicable/noncommunicable diseases and recognize causes; differentiate between renewable/nonrenewable resources; describe fossil formation; identify causes of weather changes/patterns; explain processes connecting the water cycle; compare Sun/Moon/Earth characteristics; understand Earth's tilt/motion results in days/years/seasons; identify a controlled factor.

Proficient

Design investigations; describe particle arrangement of phases; apply law of conservation of mass to physical changes; describe atoms, elements, molecules; relate distance/time/speed of objects; evaluate acting forces/types of energy; compare circuits; identify organelle function; describe photosynthesis/respiration; analyze flow of energy; describe mitosis/meiosis; infer offspring traits; describe limiting population factors; understand plate boundaries; interpret weather data; identify ocean characteristics; describe effects of Moon location; explain when results are comparable; describe why knowledge changes.

Advanced

Form conclusions/predictions; state if hypotheses are supported; explain patterns using data; describe temperature effect on particles; separate mixtures using density; apply law of conservation of mass to chemical changes; predict gravity effects on mass/weight; describe a compound/mixture; compare wavelengths for colors of light; explain body system interactions; compare gas exchange in organisms; describe how photosynthesis/respiration relate; understand energy pyramids; interpret rock layers; describe the atmosphere; explain theories on Solar System/Earth/Moon formation; describe a model for a given process.

CSAP Grade 10 – Performance Level Descriptors

Unsatisfactory
Students have a very limited understanding of scientific inquiry processes, as well as life, physical, and earth and space science concepts and vocabulary.

Partially Proficient
Record data using appropriate tools/units; describe how technologies are used; identify exothermic/endothermic reactions; describe conduction, convection, radiation; compare total mass/energy of materials; describe different animal structures/behaviors; identify composition of biological molecules; compare energy requirements based on situational needs; predict the niche of an organism; infer that organisms undergo evolution over time; describe uses of renewable/nonrenewable resources; describe cause-effect relationships in a system; identify examples of when new scientific evidence has changed previous views.

Proficient
Design an investigation/identify errors; use the Periodic Table; explain electrons are in orbitals; recognize balanced equations; calculate specific heat; explain energy changes to heat; apply the terms frequency, wavelength, amplitude; apply Newton's Laws; explain community succession; describe DNA structure/replication; construct a classification system; predict how biological variation increases/decreases survival; describe Earth’s internal layers; explain plate tectonics; identify how Earth's movement/ocean currents affect weather; classify stars; identify use for peer review; describe hypothesis/theory/law.

Advanced
Describe a source of unexplained data/how to evaluate it; explain new evidence is used to revise conclusions; describe covalent/ionic bonds; explain frequency and wavelength are inversely related; describe macromolecule functions; describe homeostatic feedback mechanisms; identify immune/endocrine/nervous system functions; relate polarity and properties of water; describe how mutation/natural selection/reproductive isolation/humans affect biodiversity; analyze data on sustainable resource use; explain differential heating/changes in moisture cause weather; analyze technology effects on progression of knowledge.

Section J

Participants' Evaluation of the Cut Score Review

Evaluation of the CSAP Science Cut Score Review Workshop — May 2008

Key: SD=Strongly Disagree D=Disagree N=Neutral A=Agree SA=Strongly Agree	SD	D	N	A	SA
1. The Bookmark Procedure was well described.	O	O	O	O	O
2. The training on bookmark placement made the task clear to me.	O	O	O	O	O
3. The training materials were helpful.	O	O	O	O	O
4. The goals for the Bookmark Procedure were clear.	O	O	O	O	O
5. Taking the test helped me place my bookmarks.	O	O	O	O	O
6. The ordering of the items in the ordered item booklet agreed with my perception of the relative difficulty of the items.	O	O	O	O	O
7. Reviewing the performance level descriptors helped me place my bookmarks.	O	O	O	O	O
8. Reviewing the Target Students helped me place my bookmarks.	O	O	O	O	O
9. I considered the content standards when I placed my bookmarks.	O	O	O	O	O
10. I understood how to place my bookmarks.	O	O	O	O	O
11. I had enough time to consider my Round 1 bookmarks.	O	O	O	O	O
12. During Round 1, I placed my bookmarks without consulting other participants.	O	O	O	O	O
13. I learned how to do the bookmark placement as I went along, so my later ones may not be comparable to my earlier ones.	O	O	O	O	O
14. Overall, my table's discussions were open and honest.	O	O	O	O	O
15. Overall, I believe that my opinions were considered and valued by my group.	O	O	O	O	O
16. The presentation of impact data was helpful to me.	O	O	O	O	O
17. I feel this procedure was fair.	O	O	O	O	O
18. I am confident that the Bookmark Procedure produced valid standards.	O	O	O	O	O
19. Overall, I am satisfied with my group's final bookmarks.	O	O	O	O	O
20. I would defend the *Partially Proficient* cut score against criticism that it is too high.	O	O	O	O	O
21. I would defend the *Partially Proficient* cut score against criticism that it is too low.	O	O	O	O	O
22. I would defend the *Proficient* cut score against criticism that it is too high.	O	O	O	O	O
23. I would defend the *Proficient* cut score against criticism that it is too low.	O	O	O	O	O
24. I would defend the *Advanced* cut score against criticism that it is too high.	O	O	O	O	O
25. I would defend the *Advanced* cut score against criticism that it is too low.	O	O	O	O	O
26. Participating in the standard setting increased my understanding of the test.	O	O	O	O	O
27. This experience will help me target instruction for the students in my classroom.	O	O	O	O	O
28. Overall, I valued the conference as a professional development experience.	O	O	O	O	O
29. The standard setting was well organized.	O	O	O	O	O

30. What is your current profession?

O Teacher
O Administrator
O Other (please specify)

____

31. How many years in your current profession?

O 1–5
O 6–10
O 11–15
O 16–20
O 21+

32. What is your highest education level?

O High school
O Bachelor's
O Master's
O Doctorate

33. What is your race/ethnicity?

O Asian/ Pacific Islander
O Black/ African-American
O American Indian
O White
O Other

34. Are you of Hispanic origin?

O Yes
O No

35. What is your gender?

O Male
O Female

36. Have you taught Special Education in the last 5 years?

O Yes
O No

37. Have you taught Adult Education in the last 5 years?

O Yes
O No

38. Have you taught Alternative Education in the last 5 years?

O Yes
O No

39. Have you taught Vocational Education in the last 5 years?

O Yes
O No

40. Have you taught ELL in the last 5 years?

O Yes
O No

41. Which grade did you work during this standard setting?

O 5
O 8
O 10

Please add your comments on the back of this evaluation.
Thank you!

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."

Question 1

The Bookmark Procedure was well described.

Content Area	Grade Level	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Overall		25	0.0%	0.0%	0.0%	32.0%	68.0%
	Grade 5	9	0.0%	0.0%	0.0%	33.3%	66.7%
	Grade 8	8	0.0%	0.0%	0.0%	37.5%	62.5%
	Grade 10	8	0.0%	0.0%	0.0%	25.0%	75.0%

Question 2

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

Content Area	Grade Level	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Overall		25	0.0%	0.0%	4.0%	28.0%	68.0%
	Grade 5	9	0.0%	0.0%	11.1%	11.1%	77.8%
	Grade 8	8	0.0%	0.0%	0.0%	50.0%	50.0%
	Grade 10	8	0.0%	0.0%	0.0%	25.0%	75.0%

Question 3

The training materials were helpful.

Content Area	Grade Level	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Overall		25	0.0%	0.0%	20.0%	28.0%	52.0%
	Grade 5	9	0.0%	0.0%	11.1%	44.4%	44.4%
	Grade 8	8	0.0%	0.0%	37.5%	25.0%	37.5%
	Grade 10	8	0.0%	0.0%	12.5%	12.5%	75.0%

Question 4

The goals for the Bookmark Procedure were clear.

Content Area	Grade Level	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Overall		25	0.0%	0.0%	4.0%	32.0%	64.0%
	Grade 5	9	0.0%	0.0%	11.1%	22.2%	66.7%
	Grade 8	8	0.0%	0.0%	0.0%	50.0%	50.0%
	Grade 10	8	0.0%	0.0%	0.0%	25.0%	75.0%

Question 5

Taking the test helped me place my bookmarks.

Content Area	Grade Level	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Overall		25	0.0%	4.0%	4.0%	40.0%	52.0%
	Grade 5	9	0.0%	0.0%	0.0%	77.8%	22.2%
	Grade 8	8	0.0%	0.0%	12.5%	25.0%	62.5%
	Grade 10	8	0.0%	12.5%	0.0%	12.5%	75.0%

Question 6

The ordering of the items in the ordered item booklet agreed with my perception of the relative difficulty of the items.

Content Area	Grade Level	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Overall		25	4.0%	28.0%	12.0%	40.0%	16.0%
	Grade 5	9	0.0%	33.3%	11.1%	44.4%	11.1%
	Grade 8	8	12.5%	50.0%	12.5%	25.0%	0.0%
	Grade 10	8	0.0%	0.0%	12.5%	50.0%	37.5%

Question 7

Reviewing the performance level descriptors helped me place my bookmarks.

Content Area	Grade Level	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Overall		25	0.0%	24.0%	20.0%	32.0%	24.0%
	Grade 5	9	0.0%	0.0%	0.0%	55.6%	44.4%
	Grade 8	8	0.0%	62.5%	12.5%	25.0%	0.0%
	Grade 10	8	0.0%	12.5%	50.0%	12.5%	25.0%

Question 8

Reviewing the Target Students helped me place my bookmarks.

Content Area	Grade Level	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Overall		25	0.0%	4.0%	4.0%	60.0%	32.0%
	Grade 5	9	0.0%	0.0%	0.0%	55.6%	44.4%
	Grade 8	8	0.0%	0.0%	12.5%	75.0%	12.5%
	Grade 10	8	0.0%	12.5%	0.0%	50.0%	37.5%

Question 9

I considered the content standards when I placed my bookmarks.

Content Area	Grade Level	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Overall		25	0.0%	4.0%	8.0%	52.0%	36.0%
	Grade 5	9	0.0%	0.0%	11.1%	33.3%	55.6%
	Grade 8	8	0.0%	12.5%	12.5%	62.5%	12.5%
	Grade 10	8	0.0%	0.0%	0.0%	62.5%	37.5%

Question 10

I understood how to place my bookmarks.

Content Area	Grade Level	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Overall		24	0.0%	0.0%	4.2%	33.3%	62.5%
	Grade 5	9	0.0%	0.0%	0.0%	33.3%	66.7%
	Grade 8	8	0.0%	0.0%	12.5%	25.0%	62.5%
	Grade 10	7	0.0%	0.0%	0.0%	42.9%	57.1%

Question 11

I had enough time to consider my Round 1 bookmarks.

Content Area	Grade Level	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Overall		25	0.0%	0.0%	0.0%	48.0%	52.0%
	Grade 5	9	0.0%	0.0%	0.0%	33.3%	66.7%
	Grade 8	8	0.0%	0.0%	0.0%	50.0%	50.0%
	Grade 10	8	0.0%	0.0%	0.0%	62.5%	37.5%

Question 12

During Round 1, I placed my bookmarks without consulting other participants.

Content Area	Grade Level	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Overall		25	0.0%	0.0%	0.0%	32.0%	68.0%
	Grade 5	9	0.0%	0.0%	0.0%	33.3%	66.7%
	Grade 8	8	0.0%	0.0%	0.0%	25.0%	75.0%
	Grade 10	8	0.0%	0.0%	0.0%	37.5%	62.5%

Question 13

I learned how to do the bookmark placement as I went along, so my later ones may not be comparable to my earlier ones.

Content Area	Grade Level	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Overall		25	28.0%	28.0%	12.0%	20.0%	12.0%
	Grade 5	9	44.4%	22.2%	11.1%	22.2%	0.0%
	Grade 8	8	25.0%	37.5%	25.0%	0.0%	12.5%
	Grade 10	8	12.5%	25.0%	0.0%	37.5%	25.0%

Question 14

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

Content Area	Grade Level	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Overall		25	0.0%	4.0%	0.0%	32.0%	64.0%
	Grade 5	9	0.0%	0.0%	0.0%	22.2%	77.8%
	Grade 8	8	0.0%	12.5%	0.0%	62.5%	25.0%
	Grade 10	8	0.0%	0.0%	0.0%	12.5%	87.5%

Question 15

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

Content Area	Grade Level	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Overall		25	0.0%	0.0%	0.0%	40.0%	60.0%
	Grade 5	9	0.0%	0.0%	0.0%	33.3%	66.7%
	Grade 8	8	0.0%	0.0%	0.0%	75.0%	25.0%
	Grade 10	8	0.0%	0.0%	0.0%	12.5%	87.5%

Question 16

The presentation of impact data was helpful to me.

Content Area	Grade Level	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Overall		25	0.0%	0.0%	0.0%	52.0%	48.0%
	Grade 5	9	0.0%	0.0%	0.0%	44.4%	55.6%
	Grade 8	8	0.0%	0.0%	0.0%	75.0%	25.0%
	Grade 10	8	0.0%	0.0%	0.0%	37.5%	62.5%

Question 17

I feel this procedure was fair.

Content Area	Grade Level	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Overall		25	0.0%	0.0%	0.0%	48.0%	52.0%
	Grade 5	9	0.0%	0.0%	0.0%	33.3%	66.7%
	Grade 8	8	0.0%	0.0%	0.0%	75.0%	25.0%
	Grade 10	8	0.0%	0.0%	0.0%	37.5%	62.5%

Question 18

I am confident that the Bookmark Procedure produced valid standards.

Content Area	Grade Level	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Overall		25	0.0%	0.0%	8.0%	44.0%	48.0%
	Grade 5	9	0.0%	0.0%	11.1%	33.3%	55.6%
	Grade 8	8	0.0%	0.0%	12.5%	62.5%	25.0%
	Grade 10	8	0.0%	0.0%	0.0%	37.5%	62.5%

Question 19

Overall, I am satisfied with my group's final bookmarks.

Content Area	Grade Level	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Overall		25	0.0%	0.0%	4.0%	48.0%	48.0%
	Grade 5	9	0.0%	0.0%	0.0%	55.6%	44.4%
	Grade 8	8	0.0%	0.0%	12.5%	62.5%	25.0%
	Grade 10	8	0.0%	0.0%	0.0%	25.0%	75.0%

Question 20

I would defend the Partially Proficient cut score against criticism that it is too high.

Content Area	Grade Level	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Overall		25	0.0%	0.0%	4.0%	44.0%	52.0%
	Grade 5	9	0.0%	0.0%	11.1%	44.4%	44.4%
	Grade 8	8	0.0%	0.0%	0.0%	50.0%	50.0%
	Grade 10	8	0.0%	0.0%	0.0%	37.5%	62.5%

Question 21

I would defend the Partially Proficient cut score against criticism that it is too low.

Content Area	Grade Level	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Overall		25	4.0%	4.0%	0.0%	52.0%	40.0%
	Grade 5	9	11.1%	11.1%	0.0%	44.4%	33.3%
	Grade 8	8	0.0%	0.0%	0.0%	75.0%	25.0%
	Grade 10	8	0.0%	0.0%	0.0%	37.5%	62.5%

Question 22

I would defend the Proficient cut score against criticism that it is too high.

Content Area	Grade Level	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Overall		25	0.0%	0.0%	8.0%	48.0%	44.0%
	Grade 5	9	0.0%	0.0%	11.1%	55.6%	33.3%
	Grade 8	8	0.0%	0.0%	0.0%	50.0%	50.0%
	Grade 10	8	0.0%	0.0%	12.5%	37.5%	50.0%

Question 23

I would defend the Proficient cut score against criticism that it is too low.

Content Area	Grade Level	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Overall		25	4.0%	0.0%	8.0%	44.0%	44.0%
	Grade 5	9	11.1%	0.0%	11.1%	55.6%	22.2%
	Grade 8	8	0.0%	0.0%	0.0%	37.5%	62.5%
	Grade 10	8	0.0%	0.0%	12.5%	37.5%	50.0%

Question 24

I would defend the Advanced cut score against criticism that it is too high.

Content Area	Grade Level	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Overall		25	0.0%	0.0%	0.0%	48.0%	52.0%
	Grade 5	9	0.0%	0.0%	0.0%	55.6%	44.4%
	Grade 8	8	0.0%	0.0%	0.0%	50.0%	50.0%
	Grade 10	8	0.0%	0.0%	0.0%	37.5%	62.5%

Question 25

I would defend the Advanced cut score against criticism that it is too low.

Content Area	Grade Level	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Overall		25	4.0%	4.0%	0.0%	48.0%	44.0%
	Grade 5	9	11.1%	0.0%	0.0%	55.6%	33.3%
	Grade 8	8	0.0%	12.5%	0.0%	50.0%	37.5%
	Grade 10	8	0.0%	0.0%	0.0%	37.5%	62.5%

Question 26

Participating in the standard setting increased my understanding of the test.

Content Area	Grade Level	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Overall		25	0.0%	0.0%	0.0%	12.0%	88.0%
	Grade 5	9	0.0%	0.0%	0.0%	22.2%	77.8%
	Grade 8	8	0.0%	0.0%	0.0%	0.0%	100.0%
	Grade 10	8	0.0%	0.0%	0.0%	12.5%	87.5%

Question 27

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

Content Area	Grade Level	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Overall		25	0.0%	0.0%	12.0%	12.0%	76.0%
	Grade 5	9	0.0%	0.0%	22.2%	0.0%	77.8%
	Grade 8	8	0.0%	0.0%	0.0%	12.5%	87.5%
	Grade 10	8	0.0%	0.0%	12.5%	25.0%	62.5%

Question 28

Overall, I valued the conference as a professional development experience.

Content Area	Grade Level	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Overall		25	0.0%	0.0%	0.0%	12.0%	88.0%
	Grade 5	9	0.0%	0.0%	0.0%	22.2%	77.8%
	Grade 8	8	0.0%	0.0%	0.0%	0.0%	100.0%
	Grade 10	8	0.0%	0.0%	0.0%	12.5%	87.5%

Question 29

The standard setting was well organized.

Content Area	Grade Level	N	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Overall		25	0.0%	0.0%	0.0%	8.0%	92.0%
	Grade 5	9	0.0%	0.0%	0.0%	0.0%	100.0%
	Grade 8	8	0.0%	0.0%	0.0%	12.5%	87.5%
	Grade 10	8	0.0%	0.0%	0.0%	12.5%	87.5%

Question 30

What is your current profession?

Content Area	Grade Level	N	Teacher	Administrator	Other (please specify)
Overall		25	80.0%	8.0%	12.0%
	Grade 5	9	77.8%	11.1%	11.1%
	Grade 8	8	87.5%	0.0%	12.5%
	Grade 10	8	75.0%	12.5%	12.5%

Question 31

How many years in your current profession?

Content Area	Grade Level	N	1-5	6-10	11-15
Overall		25	4.0%	36.0%	16.0%
	Grade 5	9	0.0%	44.4%	22.2%
	Grade 8	8	12.5%	37.5%	25.0%
	Grade 10	8	0.0%	25.0%	0.0%

Content Area	Grade Level	N	16-20	21+
Overall		25	16.0%	28.0%
	Grade 5	9	0.0%	33.3%
	Grade 8	8	12.5%	12.5%
	Grade 10	8	37.5%	37.5%

Question 32

What is your highest education level?

Content Area	Grade Level	N	High school	Bachelor's	Master's
Overall		25	0.0%	8.0%	84.0%
	Grade 5	9	0.0%	11.1%	88.9%
	Grade 8	8	0.0%	12.5%	87.5%
	Grade 10	8	0.0%	0.0%	75.0%

Content Area	Grade Level	N	Doctorate
Overall		25	8.0%
	Grade 5	9	0.0%
	Grade 8	8	0.0%
	Grade 10	8	25.0%

Question 33

What is your race/ethnicity?

Content Area	Grade Level	N	Asian/Pacific Islander	Black/African-American	American Indian
Overall		25	4.0%	0.0%	0.0%
	Grade 5	9	0.0%	0.0%	0.0%
	Grade 8	8	0.0%	0.0%	0.0%
	Grade 10	8	12.5%	0.0%	0.0%

Content Area	Grade Level	N	White	Other
Overall		25	96.0%	0.0%
	Grade 5	9	100.0%	0.0%
	Grade 8	8	100.0%	0.0%
	Grade 10	8	87.5%	0.0%

Question 34

Are you of Hispanic origin?

Content Area	Grade Level	N	Yes	No
Overall		25	0.0%	100.0%
	Grade 5	9	0.0%	100.0%
	Grade 8	8	0.0%	100.0%
	Grade 10	8	0.0%	100.0%

Question 35

What is your gender?

Content Area	Grade Level	N	Male	Female
Overall		25	24.0%	76.0%
	Grade 5	9	11.1%	88.9%
	Grade 8	8	25.0%	75.0%
	Grade 10	8	37.5%	62.5%

Question 36

Have you taught Special Education in the last 5 years?

Content Area	Grade Level	N	Yes	No
Overall		25	8.0%	92.0%
	Grade 5	9	22.2%	77.8%
	Grade 8	8	0.0%	100.0%
	Grade 10	8	0.0%	100.0%

Question 37

Have you taught Adult Education in the last 5 years?

Content Area	Grade Level	N	Yes	No
Overall		25	28.0%	72.0%
	Grade 5	9	22.2%	77.8%
	Grade 8	8	25.0%	75.0%
	Grade 10	8	37.5%	62.5%

Question 38

Have you taught Alternative Education in the last 5 years?

Content Area	Grade Level	N	Yes	No
Overall		25	0.0%	100.0%
	Grade 5	9	0.0%	100.0%
	Grade 8	8	0.0%	100.0%
	Grade 10	8	0.0%	100.0%

Question 39

Have you taught Vocational Education in the last 5 years?

Content Area	Grade Level	N	Yes	No
Overall		25	0.0%	100.0%
	Grade 5	9	0.0%	100.0%
	Grade 8	8	0.0%	100.0%
	Grade 10	8	0.0%	100.0%

Question 40

Have you taught ELL in the last 5 years?

Content Area	Grade Level	N	Yes	No
Overall		25	16.0%	84.0%
	Grade 5	9	22.2%	77.8%
	Grade 8	8	12.5%	87.5%
	Grade 10	8	12.5%	87.5%

Question 41

Which grade did you work during this standard setting?

Content Area	Grade Level	N	Grade 5	Grade 8
Overall		25	36.0%	32.0%
	Grade 5	9	100.0%	0.0%
	Grade 8	8	0.0%	100.0%
	Grade 10	8	0.0%	0.0%

Content Area	Grade Level	N	Grade 10
Overall		25	32.0%
	Grade 5	9	0.0%
	Grade 8	8	0.0%
	Grade 10	8	100.0%

Section K

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

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

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.