MASSACHUSETTS DEPARTMENT OF EDUCATION

MCAS

MASSACHUSETTS
COMPREHENSIVE
ASSESSMENT
SYSTEM

2007 MCAS
Standard Setting Report

Mathematics Grade 3

High School Science and Technology/Engineering

Table of Contents

Overview

Introduction

The purpose of this report is to document the technical details and procedures used to accomplish the standard-setting and validation tasks for 2007, including the context and rationale for setting standards, the specific methods used, the results obtained from the standard-setting panelists, and the actions taken in response to those results. The goal of the report is to provide sufficient detail to enable an accurate evaluation of the validity of the final cut-points adopted for use by the Massachusetts Department of Education.

History of MCAS Standard Setting

The MCAS tests were first administered to students in Massachusetts in 1998. At that time, mathematics, science and technology, and English language arts (ELA) were the subjects assessed. In subsequent years, additional grades and subjects were added. After the initial administration of each new test, performance standards were set. Table 1 displays the years through 2003 in which standards were set for various subjects in grades 3–10, the standard-setting method used, and the responsible contractor. Table 2 displays the grades and subjects for which standards were set in 2006. No standard setting activity occurred for MCAS in 2004 and 2005.

Table 1
Standard Setting History, 1998-2003

Grade	Content	Year	Method	Contractor
3	Reading	2001	Bookmark	HEM, BETA
4	Mathematics	1998	Body of Work	ASME
4	ELA	1998/2001	Body of Work	ASME/HEM, BETA
5	Science and Technology/Engineering	2003	Body of Work	HEM,BETA
6	Mathematics	2001	Body of Work	HEM, BETA
7	ELA	2001	Body of Work	HEM, BETA
8	Mathematics	1998	Body of Work	ASME
8	Science and Technology/Engineering	2003	Body of Work	HEM, BETA
8	ELA	1998	Body of Work	ASME
10	ELA	1998	Body of Work	ASME
10	Mathematics	1998	Body of Work	ASME

The No Child Left Behind law (NCLB) required that students be tested in grades 3 through 8 and 10 by 2006. In order to meet these requirements, it was necessary to add tests for the grades and subjects that were not previously included (see Table 2). NCLB also required that four performance levels be reported. Therefore, in 2006, it was necessary to set standards on the new assessments, and to also establish a fourth performance level (*Above Proficient*) for grade 3 reading. Table 2 displays the grades and subjects for which standards were set in 2006, as well as the standard-setting method utilized.

Table 2
Standard Setting History, 2006

Grade	Content	Method	Notes
3	Reading	Body of Work	Establish only *Proficient*/*Above Proficient* cut point
3	Mathematics	Body of Work	
5	Language & Literature	Body of Work	
5	Mathematics	Body of Work	
6	Language & Literature	Body of Work	
7	Mathematics	Body of Work	
8	Language & Literature	Body of Work	

The new grades and content areas assessed in 2006 “filled in the gaps” so that grades 3–8 are now assessed and reported in the areas of mathematics and ELA.

Current Context

The high school Science and Technology/Engineering (STE) subjects were operationally assessed for the first time in 2007. A study of the technical quality of these assessments was conducted by researchers at the University of Massachusetts at Amherst (Hambleton et al., 2007), and the results of the investigation supported the use of the assessments for measuring end-of-course performance in the four subjects and for reporting student performance according to Massachusetts’s four performance levels. To establish the cut-points necessary for reporting, standard setting was required for each of the assessments: Biology, Chemistry, Introductory Physics, and Technology/Engineering.

Existing MCAS cut points for grade 3 mathematics were set in summer 2006; however, because the test was designed to reflect only three performance levels (*Warning*, *Needs Improvement*, and *Proficient*), the panelists were unable to establish a meaningful third cut point to distinguish between *Proficient* and *Above Proficient*. To address this issue, the test was constructed in 2007 to include more cognitively demanding items so that the test had the measurement characteristics necessary to accurately identify students who performed at the *Above Proficient* performance level. Therefore, panelists were asked to validate starting cuts for the *Warning/Needs Improvement* and *Needs Improvement/Proficient* cut points and to establish the cut point for *Proficient/Above Proficient*.

2007 Standard-Setting Activities

The standard-setting meeting to establish the *Proficient/Above Proficient* cut point for the MCAS grade 3 mathematics test was held Wednesday and Thursday, August 15 and 16; standard setting for high school STE tests was held Tuesday through Thursday, August 14–16. In addition, based on the results a validation meeting was conducted for introductory physics on Monday and Tuesday, September 17 and 18. Standard setting followed the procedures specified in the proposal submitted to the Massachusetts Department of Education (Department) by Measured Progress in April 2007. A copy of the proposal is included as Appendix A.

As mentioned above, in grade 3 mathematics, the panel used a combined standard-setting/standards-validation process. Panelists were given starting cut points for the *Warning/Needs Improvement* and *Needs Improvement/Proficient* cut points, and either validated those starting cuts or recommended changes to them. For the *Proficient/Above Proficient* cut, no starting cuts were provided and the panelists followed a standard-setting process. The starting cut points for the lower two cuts were those established during the standard setting that occurred in August 2006. For the high school STE tests, with the exception of the September validation meeting for introductory physics, panelists recommended cut scores using a standard-setting process in which starting cuts were not provided. For the introductory physics validation meeting, panelists were not given starting cuts, but were asked to consider only two of the cut points (*Failing/Needs Improvement* and *Needs Improvement/Proficient*). (See “Tasks Completed After the Standard-Setting Event” on pages 9-11 for details about the introductory physics validation meeting.)

The standard-setting method implemented for all tests is a modified version of the Body of Work (BoW) method (Kingston, et al., 2001). In the original BoW method, panelists review different folders of student work in each round; for MCAS, the primary modification is the presentation of the same student folders during multiple rounds of ratings. A second modification is the inclusion of an item-mapping process for multiple-choice items. Details of the BoW method and the item-mapping activity are provided in the "Tasks Completed During the Standard-Setting Event" section on pages 5-9. To help ensure the consistency of procedures between panels, each panel was led through the standard-setting process by trained and experienced facilitators from Measured Progress.

Tasks Completed Prior to the Standard-Setting Event

Creation of Performance Level Descriptors

The Performance Level Descriptors provided panelists with descriptions of the knowledge, skills, and abilities students are expected to be able to demonstrate at the *Advanced/Above Proficient*, *Proficient*, and *Needs Improvement* performance levels. These Performance Level Descriptors are provided in Appendix C of this document.

The use of well-defined and vetted performance level descriptors was critical to ensuring the validity of the body of work standard-setting method. To ensure the quality of the content-specific performance level descriptors, a group of STE educators was convened in May 2007 to examine them. The educators who participated in this meeting were all current Assessment Development Committee (ADC) members and therefore familiar with both the content standards and the MCAS high school STE items. The specific goal of the meeting was to verify the appropriateness and clarity of the descriptors in each content area, as well as the comparability of the descriptors for each performance level across the four subjects. Educators in each content area reviewed the descriptors in their subject in terms of the appropriateness of the abilities for each performance level and the clarity and logic of the progression across the performance levels. Educators then examined the rigor of the performance level descriptors for each performance level across the content areas. Educators compared the skills and knowledge required for each performance level in each content area to determine whether comparable rigor was evident in the descriptors. They concluded that the list of skills, knowledge, and cognitive demand generated by the educators for each performance level was a valid characterization for all four content areas. The consensus of the educators was that they were confident that the descriptors for each performance level had comparable rigor across all four content areas.

Preparation of Materials for Panelists

The following materials were assembled into folders for presentation to panelists and are described in greater detail in the "Tasks Completed During the Standard-Setting Event" section of this report:

- Meeting agendas
- Confidentiality agreement
- Student test booklet
- Answer key/scoring rubrics
- Item map

- Performance level descriptors
- Training set of student folders
- 51 student folders for standard setting[1]
- Rating forms
- Evaluation form

The meeting agendas, performance level descriptors, sample rating forms, and results of the evaluations are provided in the appendices to this report.

Preparation of Presentation Materials

The PowerPoint presentations used in the plenary sessions were prepared jointly by Measured Progress and Department staff prior to the meetings. Copies of the PowerPoint slides are included in Appendix D of this document.

Preparation of Instructions for Facilitators

Two versions of the document "General Instructions for MCAS Standard Setting Group Facilitators" were created for group facilitators' reference during the meetings: one for grade 3 mathematics, and another for high school STE. Copies of these instructions are included in Appendix E of this document.

Preparation of Systems and Materials for Analysis During the Meeting

The programming necessary to conduct all analyses during the standard-setting meetings was completed, and systems were thoroughly tested ahead of time. The on-site analyses consisted of calculating the group average cut scores after each round of ratings using logistic regression. Specifically, for a given cut, each panelist's rating for each student folder was dichotomized (i.e., given a score of zero or one, where zero indicates that the panelist rated the folder as being below the cut). A logistic function was fit to the data for that cut, and the point of inflection on this curve was used to establish each panelist's cut point on the raw score scale. The cuts were then averaged across the panelists to come up with the overall group average cut score.

Determination of Starting Cut Points (grade 3 mathematics)

As described above, the purpose of the 2007 standard-setting activity for grade 3 mathematics was to establish the score for the *Proficient/Above Proficient* cut point and to validate the cuts established in August 2006 for the lower two cut points. Therefore, the August 2006 cuts were presented to the panelists as starting cuts. The 2007 starting cut points were calculated by equating the 2007 test onto the 2006 test-score scale and finding the raw scores on the 2007 form equivalent to the lower two raw cut scores established in 2006.

Recruitment and Selection of Panelists

Panelists were selected prior to the standard-setting meeting by the Department. A listing of these

[1] One folder for each possible score point from chance level (i.e., the score a student would be expected to obtain by guessing) to the highest possible score of 60. For technology/engineering, for which no students received scores of 59 or 60, only 49 student folders were prepared.

panelists by content area may be found in Appendix G. The total number of panelists who participated was 116, distributed as follows:

- Grade 3 Mathematics: 20
  - Teachers: 13
  - Administrators: 6
  - Community Representative: 1
- Biology: 22
  - Teachers: 16
  - Administrators: 5
  - Consultant: 1
- Chemistry: 23
  - Teachers: 18
  - Administrators: 3
  - Consultant: 1
  - Scientist: 1
- Introductory Physics (initial standard-setting meeting): 17
  - Teachers: 13
  - Administrators: 3
  - Industry: 1
- Introductory Physics (validation meeting): 18
  - Teachers: 13
  - Administrators: 4
  - Industry: 1
- Technology/Engineering: 16
  - Teachers: 11
  - Administrators: 3
  - Industry: 2

Of the 18 panelists who participated in the validation meeting for introductory physics, 14 had also participated in the initial standard-setting meeting, and four were new to the process.

The sample of panelists was chosen to be as geographically representative as possible of the diverse, educationally oriented, concerned citizen population of Massachusetts. Copies of the application forms that potential panelists were asked to submit are included in Appendix F, and a list of the panelists and their affiliations is included in Appendix G.

As noted in Appendix G, three panelists had to leave the standard-setting event before its completion, due to family illness/emergencies.

Tasks Completed During the Standard-Setting Event

General Orientation

Each standard-setting meeting began with a plenary session on the first morning, attended by all panelists. This session, which was presented jointly by Measured Progress and Department staff, provided a general orientation, including review of the meeting agenda (see Appendix B), background information, and an introduction to the issues of standard setting. The activities that would occur during standard setting were also explained. The Plenary Session PowerPoint Presentations are found in Appendix D. At the conclusion of the plenary session, questions about the standard-setting process were answered.

After the general orientation, the panelists assembled into their content panels. Each panel met in a separate room with a trained room facilitator from Measured Progress. The remaining standard-setting tasks were accomplished in these content panel groups.

Orientation to Assessment

Once the panelists assembled into their content panels, they took the same test the students took. This gave panelists the opportunity to become familiar with the assessment and with what students needed to do to score well. Panelists were asked to try to take on the perspective of students as they took the test. Once panelists had completed the test, the test's answer key and scoring rubrics were distributed, and panelists were allowed to self-score their tests. The panelists then discussed any issues or questions that arose about the test items or scoring rubrics.

Completing the Item Map for Multiple-Choice Items

Prior to starting the item-mapping activity, the room facilitator led a review of the multiple-choice item summary and the item-map documents with the panelists.

In each student folder was a multiple-choice (MC) item summary sheet that listed the test's MC items in order from the easiest to the most difficult based on each item's *p*-value, or percentage of students who got the item correct[2]. The summary provided the following information for each MC item: 1) the item's rank order, where item #1 was the easiest; 2) the item's position in the test booklet; 3) a brief summary of the item's text; 4) the student's answer to the item (either a plus sign if the student got the item right, or the incorrect option he or she chose); and 5) the item's *p*-value. The items were organized by the content strand with which each item was associated.

The first three of the four columns of the item map were the same as those of the MC item summary, indicating each item's rank order and position in the test booklet, along with a summary of the item's text. The final column of the item map was left blank for panelists' notes.

Each panel reviewed the MC item summary for its test, item by item, discussing the knowledge, skills, and abilities students needed to complete each item, and referencing the scoring rubrics and performance level descriptors. Panelists also discussed why each item was more difficult than the previous item. Panelists wrote the knowledge, skills, and abilities the item measured onto their item maps. They were also advised to include any other information on the item map that might help them as they rated items.

Discussion of the MCAS Performance Level Descriptors

Next, the panelists reviewed the performance level descriptors to ensure that they thoroughly understood the knowledge, skills, and abilities that students needed to demonstrate in order to be classified as *Needs Improvement*, *Proficient*, or *Advanced* (or *Above Proficient* for grade 3 mathematics). The panelists began by individually reviewing the general and content- and grade-

[2] For biology and introductory physics, grade 9 students were used in the calculation of p-values. For chemistry, grade 10 students were used, as there were virtually no grade 9 students who took the test. For technology/engineering, grade 9 and 10 students were used due to the small number of students who took the test.

specific descriptors. They then participated in a group discussion led by the facilitator to further clarify what knowledge, skills, and abilities were specified by each performance level descriptor. Bulleted lists of characteristics for each level were generated based on the whole group discussion and were posted in the room for panelists to refer to throughout the standard-setting process.

Training Round

Before beginning the individual rating process, each panel completed a training round that consisted of classifying a set of five training folders into the four performance levels. The purpose of the training round was to ensure that all panelists had a complete understanding of the rating task before they began their actual review of the full set of student folders.

To begin the training round, the facilitator briefly reviewed the performance level descriptors and the set of five training folders, which were selected to represent performance across the range of possible raw scores. The facilitator first reviewed each open-response item and then reviewed the multiple-choice item summaries. The facilitator emphasized that multiple-choice items should be considered carefully by the panelists in making their ratings since the majority of points on each test come from multiple-choice items.

The panelists then individually reviewed all five of the training sets—which were presented in random order—and placed them in order from lowest to highest. Once this was completed, the facilitator tallied the extent to which the panelists agreed about the order of the folders. The facilitator then led a group discussion of the characteristics of the folders, starting with the lowest scoring and pointing out why it belonged in the *Warning* or *Failing* performance level. For the discussion of each folder, the facilitator pointed out the connections between the knowledge, skills, and abilities demonstrated and the performance level descriptors for its performance level.

Individual Ratings

In the first step of the actual rating process, each panelist made an initial judgment of how each student folder should be categorized. Panelists used the performance level descriptors, their completed item maps, and the student test booklet to rate the student folders for their content area. Fifty-one student folders were assembled and presented to panelists for all tests. Starting with the first folder (corresponding to the lowest overall raw score), the panelists individually considered the knowledge, skills, and abilities demonstrated by the student. Panelists for high school STE then decided into which performance level each folder should be placed. For grade 3 mathematics, the process was basically the same, except that, for those folders that were precategorized according to the starting cuts, panelists were instructed to consider whether they were placed appropriately. Panelists used this same process to rate all folders and recorded their initial classification for each folder on the Round 1 Rating Form. For grade 3 mathematics, which included only two official rounds of ratings, these classifications were indicated in the Individual Rating section of the Round 1 Rating Form. Samples of the rating forms are provided in Appendix H.

Revised Ratings after Group Discussion (grade 3 mathematics only)

Once the grade 3 mathematics panelists completed their individual reviews and initial classifications of all of the student folders, they discussed each folder as a group, starting with the first folder. Panelists discussed the knowledge, skills, and abilities demonstrated in each folder and how they corresponded to the Performance Level Descriptors. The facilitator focused the discussion on any student folders for which there was disagreement among the panelists about categorization or disagreement with categorizations according to the starting cut points. The facilitator emphasized that while panelists did not need to come to a consensus about how to categorize the folders, they should express their own opinions while listening to the opinions of the other panelists. As the panelists completed their group discussion of each folder, each panelist entered a rating for that folder in the "Revised Rating After Discussion" section of the Round 1 Rating Form. Facilitators made it clear that each panelist, in each round of rating, should indicate her or his *individual* judgment on the rating form. These ratings were the grade 3 mathematics panelists' official Round 1 judgments.

Tabulation of Round 1 Results (all panels)

After the panelists had recorded their Round 1 ratings, the rating forms were returned to the Research and Analysis staff and the results were analyzed. Prior to beginning Round 2, panelists were given feedback on the Round 1 ratings. The information consisted of the group's average cut scores based on the Round 1 ratings, which were determined using logistic regression. Specifically, for a given cut, each panelist's rating for each student folder was dichotomized (i.e., given a score of zero or one, where zero indicates that the panelist rated the folder as being below the cut). A logistic function was fit to the data for that cut, and the point of inflection on this curve was used to establish each panelist's cut point on the raw score scale. The cuts were then averaged across the panelists to come up with the overall group average cut score.

Round 2 Ratings (all panels)

During Round 2, the panelists in each panel examined the results from Round 1 and discussed their ratings. Focusing on student folders near the cut points, panelists discussed any folders for which there was disagreement about classification; for grade 3 mathematics, they also discussed folders for which the Round 1 classification differed from the initial classification based on the given starting cut points. The panelists were encouraged to share their classification rationales in terms of the knowledge, skills, and abilities students must be able to demonstrate. Again, panelists were told that they did not need to come to a consensus, but that they should participate in the discussion and listen to other panelists' points of view.

After all discussions had been completed, panelists recorded their ratings on the Round 2 Rating Forms. For grade 3 mathematics, the Round 2 ratings represented the panelists' final judgments, while for high school STE, there was a third round of ratings.

Tabulation of Round 2 Ratings (high school STE only)

After the panelists had recorded their Round 2 ratings, the rating forms were returned to the Research and Analysis staff for analysis as described for Round 1.

Round 3 Ratings (high school STE only)

Round 3 proceeded similarly to Round 2: Panelists examined the Round 2 results (i.e., the group average cut points) and discussed their ratings, focusing on student folders near the cut points and folders for which there was disagreement as to how they should be categorized. Again, panelists were encouraged to share their rationales but told that they did not need to come to a consensus. After all discussions were completed, panelists recorded their final judgments on the Round 3 Rating Form.

Evaluation

Upon completion of the rating process, panelists anonymously completed an evaluation form. The evaluation forms and a tabulation of the results for each panel are included in Appendix I. The next section contains analysis of the evaluations.

Tasks Completed After the Standard-Setting Event

Upon the conclusion of the standard-setting event, the process and results, including the final cuts recommended by each panel, were reviewed and analyzed for anomalies by both Measured Progress and Department staff. The outcome of this review was that the results for all groups were found to be reasonable, with the exception of the *Failing/Needs Improvement* and *Needs Improvement/Proficient* cuts for introductory physics.

Review of Introductory Physics Results

A summary of the raw score point ranges that resulted from the cuts recommended by the panelists at the end of the standard-setting process is presented in Table 3 below. Note that the raw scores presented in Table 3 and throughout this report are specific to the 2007 tests. In future years, test results will be equated back to the 2007 test and equated raw scores will be determined that yield equivalent scaled scores and performance level designations.

As shown in Table 3, the number of points needed to achieve both *Needs Improvement* and *Proficient* was noticeably lower for Introductory Physics than for the other three subjects, which were all within a few points of each other.

Table 3
Point Ranges in Each Performance Level
Based on Panelists Final Ratings, August 2007

Performance Level	Biology	Chemistry	Introductory Physics	Technology/ Engineering
Failing	0-23	0-26	0-17	0-25
Needs Improvement	24-34	27-35	18-28	26-36
Proficient	35-50	36-47	29-49	37-51
Advanced	51-60	48-60	50-60	52-58

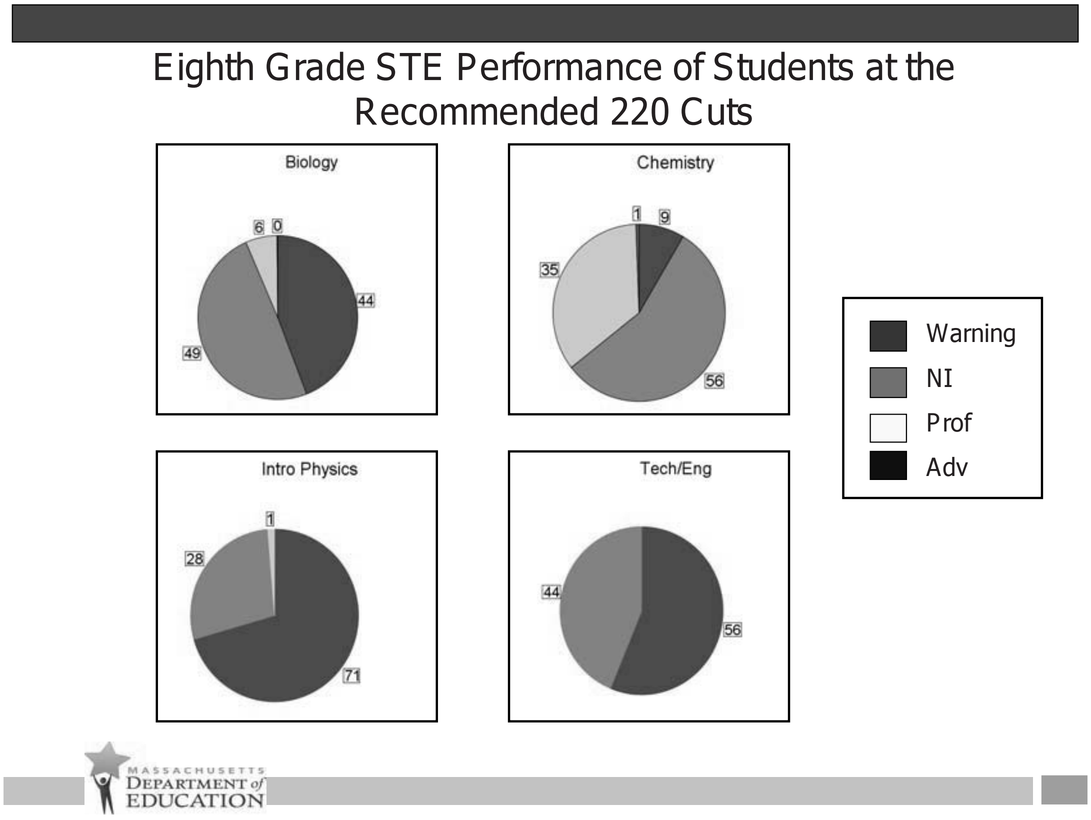
The *Failing/Needs Improvement* and the *Needs Improvement/Proficient* cuts for introductory physics appeared to be out of range (in terms of the number of points) when compared to the other STE subjects. This pattern was not seen for the *Proficient/Advanced* cut.

IRT and classical test analyses indicated that the different tests were not performing substantially differently, psychometrically. The level of difficulty did not appear to vary radically from subject to subject. Test Characteristic Curves (TCCs) were constructed for each test, and the raw score cut points were mapped onto the underlying theta metric. The same pattern of discrepancies in the theta cuts was observed as for the raw score cuts. That is, the theta values associated with the two lower cuts for introductory physics appeared to be below the range of the same cuts for the other

STE disciplines, and the top cut was consistent with the other areas.

Because the tests were on different scales, and different populations of students took each test, comparing the relative cut scores in either the raw score or the theta metrics did not by themselves provide compelling evidence that there was an actual discrepancy in the standard-setting results. It was, however, an indicator that further investigation was prudent.

To this end, the relationship between students' scores on the high school STE tests and the same students' scores on the grade 8 STE test was examined. Specifically, the subset of students who earned the minimum score necessary to be classified in the *Needs Improvement* level (i.e., a scaled score of 220) was selected and their performance on the Grade 8 test was examined. As shown below[3], the percentage of this subgroup of students who had performed at *Warning* on the grade 8 STE test (71%) was substantially higher for introductory physics than for the other three subjects. These results supported the hypothesis that the cut score between *Failing* and *Needs Improvement* represented a different level of achievement in introductory physics than in the other areas.



Eighth Grade STE Performance of Students at the Recommended 220 Cuts

Again, these analyses did not provide conclusive evidence, yet they did provide additional support to the idea that the lower introductory physics cut scores were outside the range of the others. Consequently, these data prompted further review of the bodies of work that were used in the standard-setting process and of the proposed cuts of all four content areas to determine whether the recommendations across subjects were in line with each other. To accomplish this review the Department engaged in the following two-tiered review process:

[3] Graphic provided by the Massachusetts Department of Education.

1) The first stage was a blind review of bodies of work around each cut point across the four STE subjects conducted by content area experts. STE staff at the Department and Measured Progress reviewed three student folders just below the proposed cuts and three student folders just above the proposed cuts in each of the four subjects. The conclusion of the reviewers was that the level of skill and knowledge in a large, broad sense did not seem as high for introductory physics as for the other three subjects, and that the other three subjects better matched the abilities expected at each performance level as defined by the general STE descriptors.

2) The second stage involved a critical review of the papers around each cut point, using the test items and specific performance level descriptors to evaluate whether the cut seemed appropriately placed. STE content staff at the Department and Measured Progress read the student folders around the proposed cuts, starting with three folders just below and three just above, to evaluate whether the cuts seemed appropriate based on the evidence in the folders. If the proposed placements of the cuts were questionable, more papers along the raw score continuum were read to determine where the actual cut perhaps should have been placed. The conclusion of the reviewers was that for introductory physics, the folders did not show sufficient partial understanding at the proposed *Failing/Needs Improvement* cut, nor did they generally demonstrate sufficient solid understanding at the proposed *Needs Improvement/Proficient* cut. In both cases, the reviewers felt that the cuts should have been higher: somewhere between 20 and 27 for the lower cut, and between 32 and 39 for the middle cut.

Based on the above research and reviews, the Department decided neither to accept the proposed lower two introductory physics cut points nor to make a statistical adjustment, but rather to reconvene the panel to focus on the lower two introductory physics cut points as they compared to the corresponding cut points in the other three content areas. Measured Progress contacted the panel members to notify each of them about this decision and invite them to the September introductory physics validation meeting. Following this initial contact, the Department held a conference call to explain to the panel members the reasoning for this decision (described above) and to address any questions or concerns from the panel about this decision. In addition, members of the biology, chemistry, and technology/engineering panels were invited to join the original standard-setting panel to ensure that similar interpretations of the broad standards were used.

The process for the September introductory physics validation meeting was basically the same as described above for the August standard-setting meeting, except that 1) panelists were given only 32 student folders to review, two folders at each of the score points in the ranges identified above; 2) ratings were done in two rounds instead of three; and 3) panelists were given a brief questionnaire after the calibration step, asking whether they felt able to distinguish among the performance levels and ready to begin the validation task. All 18 panelists answered in the affirmative to the two questions, suggesting that they felt prepared for the validation task. The results of the September meeting are presented in Tables 14 and 15.

Analysis and Review of Panelists' Feedback – August Standard Setting

After completing the standard-setting activities, panelists' evaluation feedback was reviewed[4]. The review of the evaluation forms from all the August 2007 standard-setting meetings did not reveal any anomalies in the standard-setting process or indicate any reason that a particular panelist's data should not be incorporated in the final results. It appeared that all panelists

[4] A summary of the panelists' evaluations can be found in Appendix I.

understood the rating task and attended to it appropriately. However, at least 2 introductory physics panelists expressed concern over the uniformity with which all panelists were applying the intended standard-setting protocols. Table 4 below provides demographic information for the groups of panelists who completed the evaluation for each content area.

Table 4
Demographic Information on Panelists Completing Evaluations

	Grade 3 Mathematics	Biology	Chemistry	Intro. Physics	Tech./Eng.
Panelist is a:					
Classroom teacher	10	15	20	13	10
K-12 administrator	4	3	0	1	2
University-level educator	1	1	1	0	0
Business/community representative	2	1	2	1	1
Other	3	2	0	0	3
Gender:					
Male	4	4	7	4	11
Female	16	18	16	11	5
Total:	20	22	23	15	16

Question 3 on the grade 3 mathematics evaluation form and question 4 on the high school STE evaluation asked the panelists whether they relied primarily on the open-response or multiple-choice items, or on both equally, in determining their ratings. The majority of the panelists indicated that they relied on both equally; the percentage ranged from 87% for the initial introductory physics standard setting to 100% for both grade 3 mathematics and the introductory physics validation. Ninety-one percent relied on both equally for biology and chemistry, and 94% relied on both equally for technology/engineering. Of the remaining seven panelists, four relied primarily on the open-response questions (one for biology, two for chemistry, and one for technology/engineering), and three relied primarily on the multiple-choice questions (one for biology and two for the initial introductory physics standard setting).

The next nine questions (4–12 for grade 3 mathematics and 5–13 for high school STE) asked panelists their opinions about the organization and process of the standard setting, as well as the final results. The response options were "Strongly Disagree," "Disagree," "Agree," and "Strongly Agree." Table 5 below shows the percentage of panelists who indicated that they either agreed or strongly agreed with each statement by content area.

Table 5
Summary of Responses to Evaluation by Content Area – Questions 5–13 (4–12 for Mathematics)

	Percent Responding Agree or Strongly Agree				
	Grade 3 Mathematics	Biology	Chemistry	Intro. Physics	Tech./Eng.
5 (4). Overall environment and accommodations were comfortable and appropriate	94%	100%	100%	100%	100%
6 (5). Background information provided improved my ability to set standards	100%	100%	96%	93%	100%
7 (6). Taking and discussing the exam helped me understand the purpose and process	100%	100%	96%	100%	94%
8 (7). By the end of the calibration training, I could distinguish among Performance Level Descriptors	100%	100%	91%	100%	100%
9 (8). Overall, I was provided with clear instructions	100%	95%	91%	100%	100%
10 (9). The group discussions after the first round improved my ability to set standards	100%	100%	96%	100%	100%
11 (10). I am confident that the ratings I provided were consistent with the Performance Level Descriptors	100%	100%	96%	100%	100%
12 (11). The standard-setting process provided for a reliable classification of student work	100%	91%	91%	100%	100%
13 (12). The facilitator was effective	100%	91%	96%	100%	100%

As seen in Table 5, a large majority of panelists who participated in the initial standard-setting meeting said they either agreed or strongly agreed with each statement. In every case, the percentage was greater than 90%. Looking at the full results (see Appendix I), there was only one instance where a panelist said he or she strongly disagreed with one of the statements: One panelist in the biology group strongly disagreed with the statement that the standard-setting process provided for a reliable classification of student work (question 12). For chemistry, either one or two panelists disagreed with every statement except the first, that the environment and accommodations were comfortable and appropriate (question 5). Finally, in biology, two panelists disagreed that the facilitator was effective (question 13). There were only five additional responses of "disagree," and they were scattered among the panels and questions.

The remaining questions (13–20 for grade 3 mathematics and 14–21 for high school STE) asked the panelists to indicate whether they felt the amount of time allotted to each of the various standard-setting steps was appropriate. The response options were "Far too short," "Too short," "Approximately right," "Too long," and "Far too long." Table 6 below indicates the percentage of panelists who felt the time allowed for each step was "Approximately right," by content area.

Table 6
Summary of Responses to Evaluation by Content Area – Questions 14–21 (13–20 for Mathematics)

	Percent Responding Approximately Right				
	Grade 3 Mathematics	Biology	Chemistry	Intro. Physics	Tech./Eng.
14 (13). Initial background information	85%	64%	61%	73%	63%
15 (14). Taking and discussing the exam	95%	95%	96%	87%	94%
16 (15). Learning about and discussing Performance Level Descriptors	70%	73%	91%	93%	81%
17 (16). Ranking, discussing, and classifying student work (calibration)	90%	86%	87%	87%	94%
18 (17). Initial individual classification of student work	95%	55%	87%	87%	88%
19 (18). Group discussion regarding initial ratings	80%	59%	91%	73%	88%
20 (19). Rating student work for the second time	90%	82%	78%	93%	88%
21 (20). Final rating of student work	95%	100%	91%	93%	94%

Overall, as can be seen in Table 6, the percentage of panelists indicating that the time allowed was "Approximately right" ranged from 55% to 100%. Among responses other than "Approximately right," responses of either "Too long" or "Far too long" outnumbered responses of "Too short" or "Far too short" by about three to one overall. However, there was quite a bit of variability across content areas, ranging from two to one for chemistry to nine to one for grade 3 mathematics (see Appendix I).

Across all panels, there were only five instances in which fewer than two-thirds of the panelists answered "Approximately right"; for biology, chemistry, and technology/engineering, over a third of the panelists felt that the time allotted for the initial background information was "Too long" or "Far too long" (question 14).

For biology, only 55% of the panelists felt that the time allotted to the initial individual classification was "Approximately right" (question 18); of the remaining panelists, one thought the time allotted was "Far too short," but the remaining felt it was either "Too long" (six panelists) or "Far too long" (three panelists). Only 59% of the biology panelists felt that the time allotted to the group discussion of the initial ratings was "Approximately right" (question 19); of the remaining panelists, two thought the time was "Too short," six thought it was "Too long," and one thought it was “Far too long.”

In general, responses of “Too short” or “Far too short” are of greater concern than responses of “Too long” or “Far too long,” since having too little time is presumably more likely to have a negative effect on the panelists’ ability to perform the task. As mentioned above, there were approximately a third as many responses indicating not enough time than responses indicating too much time. There was only one response of “Far too short”: question 18 for biology (as described above). In addition, in general, responses indicating too little time tended to be scattered around rather than concentrated for particular panels or questions. The main exception was question 16 (question 15 for grade 3 mathematics), where 12 panelists overall said the time allotted was “Too short”; for all the other questions, five or fewer panelists overall indicated the time was “Too short.”

Analysis and Review of Panelists’ Feedback – September Introductory Physics Validation Meeting

Table 7 below provides demographic information about the group of panelists who participated in the introductory physics validation meeting.

Table 7
Demographic Information of Panelists Completing Evaluations – Introductory Physics Validation

Panelist is a:	
Classroom teacher	12
K–12 administrator	5
University-level educator	0
Business/community representative	0
Other	1
Gender:	
Male	5
Female	13
Total:	18

Question 3 on the evaluation asked the panelists whether they relied primarily on the open-response or multiple-choice items, or on both equally, in determining their ratings. All 18 of the panelists (100%) indicated that they relied on both equally.

Questions 4–10 asked panelists their opinions about the organization and process of the standard setting, as well as the final results. The response options were “Strongly Disagree,” “Disagree,” “Agree,” and “Strongly Agree.” Table 8 below shows the percent of panelists who indicated they either agreed or strongly agreed with each statement by content area.

Table 8
Summary of Responses to Evaluation by Content Area — Questions 4–10 — Introductory Physics Validation

	Percent Responding Agree or Strongly Agree
4. Overall environment and accommodations were comfortable and appropriate	83%
5. Explanation of purpose of standards validation improved my ability to set standards	72%
6. Taking and discussing the exam helped me understand the purpose and process	61%
7. The group discussions after the first round improved my ability to set standards	78%
8. I am confident that the ratings I provided were consistent with the Performance Level Descriptors	94%
9. The standards-validation process provided for a reliable classification of student work	83%
10. The facilitator was effective	89%

As can be seen in Table 8, the percentage of panelists who agreed or strongly agreed with each statement was considerably lower than those reported in Table 5 for the panelists who participated in the initial standard setting. In addition, there were either one or two panelists for each statement who indicated that they strongly disagreed. However, the questions with the lowest ratings (questions 5, 6, and 7) asked about the usefulness of various parts of the process (the explanation in the plenary session, taking the exam, and the group discussions after the first round of ratings); many of the panelists had participated in the initial standard setting as well, so they were repeating the process for the second time, which may explain at least some of the "Disagree" and "Strongly disagree" responses.

The remaining questions (11–17) asked the panelists to indicate whether they felt the amount of time allotted to each of the various standard setting steps was appropriate. The response options were "Far too short," "Too short," "Approximately right," "Too long," and "Far too long." Table 9 below indicates the percentage of panelists who felt the time allowed for each step was "Approximately right," by content area.

Table 9
Summary of Responses to Evaluation by Content Area – Questions 11–17 — Introductory Physics Validation

	Percent Responding Approximately Right
11. Information provided during the orientation session	56%
12. Taking and discussing the exam	61%
13. Learning about and discussing Performance Level Descriptors	56%
14. Ranking, discussing, and classifying students' work (calibration)	78%
15. Initial individual classification of student work	89%
16. Group discussion regarding initial ratings	72%
17. Final rating of student work	72%

Overall, as can be seen in Table 9, the percentage of panelists indicating that the time allowed was "Approximately right" ranged from 56% to 89%. Among responses other than "Approximately right," responses of either "Too long" or "Far too long" outnumbered responses of "Too short" or "Far too short" by almost four to one. There was only one response of "Far too short," for question 17. Again, the responses to these questions may reflect the fact that some of the panelists were completing these steps for the second time.

Preparation of Recommended Cut Scores

After each round of ratings, the raw score cut points were calculated based on the average across all panelists' cuts, where the cuts for each panelist were calculated as described in the Tabulation of Round 1 Results section on page 8 of this document. Both the mean and the median were calculated, and the mean values were used; in most cases, the mean and median yield the same cut-points. In addition, the percentage of students who would be classified into each performance level was determined. This calculated impact data was not provided to panelists as part of the standard-setting process. The results from the initial standard setting for all content areas are presented in Tables 10 through 13 below.

Table 10
2007 MCAS Standard-Setting Results: Grade 3 Mathematics

Performance Level	Starting Cuts		Panelist Cuts: Round 1		Panelist Cuts: Round 2		Final Adopted Cuts	
	Raw Score Range	% in Level	Raw Score Range	% in Level	Raw Score Range	% in Level	Raw Score Range	% in Level
Warning	0-21	15.0	0-21	15.0	0-21	15.0	0-21	15.0
Needs Improvement	22-29	24.5	22-29	24.5	22-28	20.2	22-29	24.5
Proficient			30-36	41.5	29-36	45.8	30-36	41.5
Above Proficient			37-40	19.0	37-40	19.0	37-40	19.0

Table 11
Cut-point Statistics by Round: Grade 3 Mathematics

	Round 1			Round 2		
Cut	Mean	SD	Median	Mean	SD	Median
W/NI	21.3	0.43	21.5	21.3	0.43	21.5
NI/P	29.0	0.84	29.5	28.9	0.86	29.5
P/A	36.6	0.29	36.5	36.6	0.21	36.5

Table 12
2007 MCAS Standard-Setting Results: High School Science and Technology/Engineering

Test	Performance Level	Panelist Cuts: Round 1		Panelist Cuts: Round 2		Panelist Cuts: Round 3		Final Adopted Cuts	
		Raw Score Range	% in Level	Raw Score Range	% in Level	Raw Score Range	% in Level	Raw Score Range	% in Level
Biology	*Failing*	0-22	18.8	0-23	20.5	0-23	20.5	Same as Round 3	
	Needs Improvement	23-33	20.9	24-34	21.7	24-34	21.7		
	Proficient	34-49	44.1	35-50	44.6	35-50	44.6		
	Advanced	50-60	16.2	51-60	13.2	51-60	13.2		
Chemistry	*Failing*	0-26	44.9	0-26	44.9	0-26	44.9	Same as Round 3	
	Needs Improvement	27-35	19.3	27-35	19.3	27-35	19.3		
	Proficient	36-47	24.8	36-47	24.8	36-47	24.8		
	Advanced	48-60	11.0	48-60	11.0	48-60	11.0		
Intro. Physics	*Failing*	0-19	21.3	0-17	16.8	0-17	16.8	See Table 14 on page 17	
	Needs Improvement	20-29	21.4	18-28	23.9	18-28	23.9		
	Proficient	30-48	43.1	29-48	45.2	29-49	47.1		
	Advanced	49-60	14.2	49-60	14.2	50-60	12.2		
Tech./ Eng.	*Failing*	0-24	24.6	0-25	27.5	0-25	27.5	Same as Round 3	
	Needs Improvement	25-36	38.7	26-37	39.0	26-36	35.8		
	Proficient	37-51	34.3	38-51	31.1	37-51	34.3		
	Advanced	52-58	2.5	52-58	2.5	52-58	2.5		

Table 13
Cut-point Statistics by Round: High School Science and Technology/Engineering

Test	Cut	Round 1			Round 2			Round 3		
		Mean	SD	Median	Mean	SD	Median	Mean	SD	Median
Biology	*F/NI*	22.6	3.24	22.5	23.2	1.99	22.5	23.1	1.43	22.5
	NI/P	33.1	3.83	32.5	34.4	2.49	34.5	34.6	1.91	34.5
	P/A	49.6	2.93	49.5	50.2	2.27	50.5	50.2	1.88	50.5
Chemistry	*F/NI*	26.1	3.85	26.5	26.3	2.68	26.5	26.5	1.78	26.5
	NI/P	35.2	3.94	34.5	36.0	1.83	36.5	35.8	1.42	36.5
	P/A	47.9	4.56	46.5	47.9	2.17	47.5	47.4	1.68	46.5
Intro. Physics	*F/NI*	19.0	2.48	18.5	18.0	1.06	17.5	18.0	0.99	17.5
	NI/P	29.3	3.94	28.5	28.6	1.92	28.5	28.9	1.24	28.5
	P/A	48.0	3.25	48.5	48.3	1.57	48.5	49.2	1.33	48.5
Tech./ Eng.	*F/NI*	24.5	4.44	23.5	25.9	1.82	26.0	25.9	1.54	25.5
	NI/P	36.9	3.29	37.5	37.6	2.57	37.5	37.0	2.28	37.5
	P/A	51.3	2.48	52.0	51.4	2.32	51.5	51.7	1.42	52.5

As mentioned on pages 10-12, a review of the Round 3 results for introductory physics raised some concerns about the recommended cut scores. As a result, a standards-validation meeting was conducted in September to revisit the lower two cutpoints, *Failing/Needs Improvement* and *Needs Improvement/Proficient*. Tables 14 and 15 below show the results of the introductory physics validation meeting.

Table 14
2007 MCAS Standard-Setting Results: Introductory Physics Validation

Performance Level	Panelist Cuts: Round 1		Panelist Cuts: Round 2		Final Adopted Cuts	
	Raw Score Range	% in Level	Raw Score Range	% in Level	Raw Score Range	% in Level
Failing	0-22	27.7	0-21	25.7	Same as Round 2	
Needs Improvement	23-33	23.9	22-33	25.9		
Proficient	34-49	36.2	34-49	36.2		
*Advanced***	50-60	12.2	50-60	12.2		

**Panelists were not given the option of changing the *Proficient/Advanced* cut

Table 15
Cut-point Statistics by Round: Introductory Physics Validation

Cut	Round 1			Round 2		
	Mean	SD	Median	Mean	SD	Median
F/NI	22.0	1.49	22.0	21.8	2.00	22.0
NI/P	33.9	0.88	33.5	33.9	0.89	33.5
*P/A***						

**Panelists were not given the option of changing the *Proficient/Advanced* cut

Finally, because of the method used for establishing scaled scores on the MCAS tests (excepting the grade 3 tests, for which scaled scores are not reported), it was necessary to translate the final adopted raw score cuts shown in Tables 12 and 14 to effective raw score cuts. The effective cuts are sometimes lower than the cuts presented in Tables 12 and 14 because of the rounding rules used in the scaling process of MCAS. A given raw score might fall below the adopted raw cut, but if it scales to a value of 219.89, for example, then the scaled score would be rounded to 220. The raw score ranges associated with the effective cuts are shown in Table 16 below.

Table 16
2007 MCAS Standard-Setting Results: Effective Raw Score Ranges

Performance Level	Biology	Chemistry	Intro. Physics	Tech./ Eng.
Failing	0-20	0-23	0-19	0-23
Needs Improvement	21-34	24-35	20-33	24-36
Proficient	35-49	36-47	34-48	37-51
Advanced	50-60	48-60	49-60	52-60

Summary

This report summarizes the rationale, methods, and results of the standard-setting process that was used in August and September 2007 to establish performance level cut scores for the new MCAS assessments in high school STE and the process used to establish a single cut score for the grade 3 mathematics cut between *Proficient* and *Above Proficient*.

For grade 3 mathematics, the Department accepted the recommendations of the standard-setting panelists. The *Proficient/Above Proficient* cut determined at the standard-setting meeting was adopted and used for reporting the 2007 test results.

For high school biology, chemistry, and technology/engineering, the Department also accepted the recommendations of the standard-setting panelists. However, based on the review of the results conducted by the Department and Measured Progress content experts, as well as some subsequent analyses, a judgment was made that the two lower introductory physics cut scores needed to be revisited in a validation meeting. This decision reflected the fact that introductory physics functions in tandem with the other three high school STE tests as part of a comprehensive Science and Technology/Engineering assessment system. For this reason, the cut-points used for reporting must be coherent across the four STE tests. The results of the analyses and review done subsequent to standard setting indicated that the lower two cuts for introductory physics were not consistent with the cuts recommended for the other tests. The cuts were revisited in the September validation meeting for introductory physics, and the recommendations made by the validation panelists were accepted by the department and used for reporting the results of the 2007 high school introductory physics test.

References

Hambleton, Ronald, Yue Zhao, Zachary Smith, Wendy Lam, and Nina Deng. 2007. Psychometric Analyses of the 2006 MCAS High School Science Tests. Massachusetts Department of Education Technical report.

Kingston, Neal M., Stuart R. Kahl, Kevin P. Sweeney, and Luz Bay. 2001. Setting performance standards using the body of work method. In G.J. Cizek (Ed.), *Setting performance standards: concepts, methods, and perspectives*, pp. 219-248. Mahwah, NJ: Lawrence Erlbaum.