# TR 2008 - 14

# Ohio Graduation Tests
# Standard Setting Report
# for
# Science, Social Studies and Writing

**Michael B. Bunch**
**Elliot Inman**
**Julie Miles**

**Measurement Incorporated ™**

**August 6, 2008**

| Revision | Description | Revised by | Approval | Date |
|---|---|---|---|---|
| Int. Rel. | Edited by T. Moore | TWM | | 8.6.08 |

# TR 2008 – 14

# Ohio Graduation Tests
# Standard Setting Report
# for
# OGT Science, Social Studies and Writing

**Michael B. Bunch**
**Elliot Inman**
**Julie Miles**
**Measurement Incorporated ™**

**June 1, 2005**

# Executive Summary

The Ohio Graduation Tests (OGT) were mandated by Ohio Senate Bill 1 (SB1) of 2001, with modifications per Ohio House Bill 3 (HB3) of 2003. Ohio law calls for tests based on rigorous content and five performance levels: Advanced, Accelerated, Proficient, Basic, and Limited. Students entering ninth grade in 2003 (class of 2007) will be required to obtain a score that places them at Proficient or better in order to graduate. In the spring of 2005, Ohio tenth graders took five components of the graduation tests: Reading, Writing, Mathematics, Science, and Social Studies. Performance standards had been set for the Reading and Mathematics tests in the spring of 2004. On April 25-27, 2005, the Ohio Department of Education convened three groups to review the remaining three tests and recommend performance standards (cut scores) for the five levels. These three groups consisted of 25 panelists in Science, 21 panelists in Social Studies, and 25 panelists in Writing. In each group were classroom teachers, administrators, parents, and business representatives. Classroom teachers included both in-subject and across-subject teachers.

Measurement Incorporated (MI), the contractor responsible for developing the OGT, presented a standard-setting plan to the ODE and its Technical Advisory Committee (TAC) in October 2004. The ODE approved the plan, with minor modifications. Implementing the final, approved version of the plan, MI staff (Dr. Michael Bunch, Dr. Elliot Inman, and Dr. Julie Miles) led the three groups through a three-day process for setting performance standards. The Science and Social Studies groups employed a method commonly referred to as the bookmark procedure, while the Writing group employed a method commonly referred to as the holistic method. The TAC approved a different approach for the Writing (Reading and Mathematics had also employed the bookmark procedure) because of the preponderance of weight in the Writing test attributed to the two writing samples (36 out of 48 points).

The three groups analyzed the tests and recommended for each test a set of cut scores to differentiate among the five groups of students using performance level descriptors (PLDs) that had been developed by the ODE with representative groups of Ohio educators and stakeholders. After three rounds of deliberation, the groups recommended cut scores.

Prior to arriving at these recommended cut scores on April 27, each group first took the Science, Social Studies, or Writing test, discussed at some length the criteria by which the open-ended items on each test were scored, scored their own tests, reviewed and discussed PLDs, received detailed instructions in the application of the bookmark or holistic procedure, and completed a practice exercise. They then worked through three rounds of review and discussion to make their final recommendations. Prior to each round, panelists completed a readiness form indicating that they fully understood the tasks they were to perform.

For Science and Social Studies, panelists placed a bookmark in a specially constructed test booklet (ordered from easiest to most difficult item). Because item difficulty and student achievement level had been plotted on the same scale, the bookmark placements actually identified not only the most difficult item a student at the threshold of a particular category could answer but the achievement level of that student as well. For this Proficient cut score, for

example, we asked panelists to evaluate each item in terms of whether or not a minimally Proficient student would have a 50-percent chance of answering correctly. These individual achievement level estimates were then used to calculate a mean achievement level for each threshold, which corresponded to a unique raw score on the test. These threshold raw scores were the cut scores.

For the writing test, panelists evaluated a set of 150 student work samples. MI staff had selected these samples in advance with the approval of ODE staff. Panelists examined 20-30 samples per round and assigned each sample to one of the five performance categories, using a holistic rating procedure; i.e., given the student's responses to both essays, a short-answer question and ten multiple-choice editing questions, the panelist (without knowing the scores on the short-answer or essay questions) determined whether the whole body of work for that student represented Limited, Basic, Proficient, Accelerated, or Advanced work. MI staff entered these ratings, cross-referenced them with the scores the students had earned, determined median scores for each category, and calculated the midpoints between medians of adjacent categories to determine cut scores.

The cut scores are shown, along with the associated percentages of students at each level in the spring of 2005 (class of 2007) in Table ES-1. The percentages shown in Table ES-1 are preliminary estimates based on a sample of 45,000 students out of approximately 157,000 students who took the test in March 2005.

**Table ES-1**
**Final Cut Scores and Percentages of Students in Each Category**

| | **Science** | | **Social Studies** | | **Writing** | |
|---|---|---|---|---|---|---|
| **Category** | **Cut Score (Out of 48)** | **Percent in Category** | **Cut Score (Out of 48)** | **Percent in Category** | **Cut Score (Out of 48)** | **Percent in Category** |
| **Limited** | -- | 9.5 | -- | 9.4 | -- | 5.1 |
| **Basic** | 14.5 | 17.7 | 15 | 11.1 | 18 | 12.2 |
| **Proficient (Graduation)** | 23.5 | 31.7 | 21.5 | 32.1 | 25.5 | 34.0 |
| **Accelerated** | 32 | 23.4 | 33 | 22.1 | 34 | 41.9 |
| **Advanced** | 37.5 | 17.7 | 39 | 25.3 | 41 | 6.8 |
| **Proficient or Above** | | **72.8** | | **79.5** | | **82.7** |

After the first round of standard setting, each panelist was allowed to see how his or her Round 1 bookmarks or ratings of student work samples compared to those of other panelists. Panelists also received impact information. This information was used to point out how many students would be classified at each level if the first round recommendations held. Panelists examined not only overall score distributions but distributions by race and gender as well and discussed the implications at length.

Panelists were given a second opportunity to place bookmarks (Science and Social Studies) or evaluate student work samples (Writing) dividing the five groups of students. Once again, group facilitators from MI shared with the groups the results and led discussions of the implications. Panelists then went through a third round. For the bookmark procedure, they indicated not only the pages at which the bookmarks should be set, but the corresponding raw cut scores and associated percentages of students scoring at or above that cut score. For the holistic procedure, panelists were given the actual score for each student work sample they rated. These final ratings are reflected in Table ES-1.

## Impact by Group

MI provided impact data to panelists, first for all students tested, and then by subgroup. These final distributions across the five levels are shown in Tables ES-2, ES-3 and ES-4. Because some students failed to indicate race or sex (or either), the numbers do not add up to 45,000. The racial groups included the following: American Indian (AmInd), Asian-Pacific Islander (As-PI), Black-African American (BL-AA), Hispanic (Hisp), Multiracial (Multi), Other (Other), White (White). It should also be noted that the results shown in Tables ES-1, ES-2, ES-3, and ES-4 are based on the sample of 45,000 students for whom complete data were available at the time of the standard setting; i.e., the data that were shared with the panelists.

**Table ES-2**
**Distribution of Students by Race and Sex: Science**
**(Entries are percentages.)**

| | Category | | | | | Proficient or Above |
|---|---|---|---|---|---|---|
| **Group (Tested)** | **Limited** | **Basic** | **Proficient** | **Accelerated** | **Advanced** | |
| AmInd (86) | 12.79 | 22.09 | 34.88 | 18.60 | 11.63 | 65.11 |
| As-PI (545) | 7.16 | 14.68 | 24.22 | 22.94 | 31.01 | 78.17 |
| BL-AA (6323) | 29.81 | 35.52 | 24.77 | 7.37 | 2.53 | 34.67 |
| Hisp (927) | 21.90 | 29.34 | 28.91 | 12.19 | 7.66 | 48.76 |
| Multi (581) | 9.98 | 22.20 | 33.91 | 19.62 | 14.29 | 67.82 |
| Other (140) | 24.29 | 16.43 | 29.29 | 16.43 | 13.57 | 59.29 |
| White (36,419) | 5.62 | 14.29 | 33.07 | 26.53 | 20.49 | 80.09 |
| Total by Race 45,021 | **9.49** | **17.71** | **31.72** | **23.37** | **17.71** | **72.80** |
| | | | | | | |
| Female (22,075) | 9.21 | 19.87 | 32.64 | 22.70 | 15.58 | 70.92 |
| Male (22,891) | 9.71 | 15.61 | 30.82 | 24.05 | 19.80 | 74.67 |
| Total by Sex 44,966 | **9.46** | **17.70** | **31.72** | **23.39** | **17.73** | **72.83** |

Note: Total (percent) by race and Total (percent) by sex may differ due to non-response data. For example, an examinee may fail to report race as one of the categories but did report a gender; or the other way around. The result of such nonreporting is slightly different counts of totals.

**Table ES-3**
**Distribution of Students by Race and Sex: Social Studies**
**(Entries are percentages.)**

| | Category | | | | | Proficient or Above |
|---|---|---|---|---|---|---|
| **Group (Tested)** | **Limited** | **Basic** | **Proficient** | **Accelerated** | **Advanced** | |
| AmInd (73) | 8.22 | 15.07 | 35.62 | 23.29 | 17.81 | 76.72 |
| As-PI (545) | 7.34 | 7.16 | 23.67 | 21.10 | 40.73 | 85.50 |
| BL-AA (6345) | 26.32 | 22.58 | 35.37 | 10.45 | 5.28 | 51.10 |
| Hisp (910) | 22.53 | 20.99 | 30.66 | 14.18 | 11.65 | 56.49 |
| Multi (601) | 9.98 | 11.65 | 37.27 | 19.13 | 21.96 | 78.36 |
| Other (107) | 27.10 | 10.28 | 27.10 | 16.82 | 18.69 | 62.61 |
| White (36,419) | 6.06 | 8.93 | 31.57 | 24.44 | 29.00 | 85.01 |
| Total by Race 45,000 | **9.37** | **11.13** | **32.06** | **22.13** | **25.31** | **79.50** |
| | | | | | | |
| Female (22,132) | 8.77 | 12.29 | 35.18 | 21.28 | 22.48 | 78.94 |
| Male (22,818) | 9.89 | 9.98 | 29.05 | 22.98 | 28.10 | 80.13 |
| Total by Sex 44,950 | **9.34** | **11.12** | **32.07** | **22.14** | **25.33** | **79.54** |

Note: Total (percent) by race and Total (percent) by sex may differ due to non-response data. For example, an examinee may fail to report race as one of the categories but did report a gender; or the other way around. The result of such nonreporting is slightly different counts of totals.

**Table ES-4**
Distribution of Students by Race and Sex: Writing
**(Entries are percentages.)**

| | Category | | | | | Proficient or Above |
|---|---|---|---|---|---|---|
| **Group (Tested)** | **Limited** | **Basic** | **Proficient** | **Accelerated** | **Advanced** | |
| AmInd (86) | 5.81 | 11.63 | 53.49 | 26.74 | 2.33 | 82.56 |
| As-PI (545) | 3.49 | 9.36 | 30.28 | 39.63 | 17.25 | 87.16 |
| BL-AA (6323) | 12.38 | 26.46 | 39.92 | 19.74 | 1.50 | 61.16 |
| Hisp (927) | 13.05 | 23.19 | 35.17 | 25.57 | 3.02 | 63.76 |
| Multi (581) | 5.51 | 11.36 | 38.90 | 37.18 | 7.06 | 83.14 |
| Other (126) | 11.90 | 14.29 | 29.37 | 39.68 | 4.76 | 73.81 |
| White (36,419) | 3.60 | 9.50 | 32.88 | 46.32 | 7.70 | 86.90 |
| Total by Race 45,007 | **5.08** | **12.20** | **33.99** | **41.91** | **6.82** | **82.72** |
| | | | | | | |
| Female (22,158) | 2.73 | 8.91 | 31.55 | 47.58 | 9.23 | 88.36 |
| Male (22,799) | 7.32 | 15.36 | 36.38 | 36.45 | 4.49 | 77.32 |
| Total by Sex 44,957 | **5.06** | **12.18** | **34.01** | **41.93** | **6.82** | **82.76** |

Note: Total (percent) by race and Total (percent) by sex may differ due to non-response data. For example, an examinee may fail to report race as one of the categories but did report a gender; or the other way around. The result of such nonreporting is slightly different counts of totals.

## Results by Rounds

As noted above, panelists engaged in three rounds of setting bookmarks to derive cut scores. The placements, particularly for the Proficient category, were extremely stable across rounds. Results by round are shown in Tables ES-5, 6, and 7. Percentages of students in each group are shown in parentheses. As with the first four tables, the numbers shown in Tables ES-5, 6, and 7 reflect the data that were available on April 25, 2005.

**Table ES-5**
**Cut Scores by Round: Science**
**(Percentages of students in each group are shown in parentheses.)**

| | **Round** | | |
|---|---|---|---|
| **Category** | **1** | **2** | **3** |
| Limited | -- (7.1) | -- (8.5) | -- (9.5) |
| Basic | 13 (17.5) | 14 (18.7) | 14.5 (17.7) |
| Proficient (Graduation) | 22.5 (30.0) | 23.5 (31.7) | 23.5 (31.7) |
| Accelerated | 31 (27.7) | 32 (25.1) | 32 (23.4) |
| Advanced | 37.5 (17.7) | 38 (16.0) | 37.5 (17.7) |
| Percent of Students Proficient or Above | **75.4** | **72.8** | **72.8** |

**Table ES-6**
**Cut Scores by Round: Social Studies**
**(Percentages of students in each group are shown in parentheses.)**

| | **Round** | | |
|---|---|---|---|
| **Category** | **1** | **2** | **3** |
| Limited | -- (8.8) | -- (9.4) | -- (9.4) |
| Basic | 14.5 (12.6) | 15 (11.1) | 15 (11.1) |
| Proficient (Graduation) | 22 (24.3) | 21.5 (26.9) | 21.5 (32.1) |
| Accelerated | 31 (23.2) | 31.5 (25.4) | 33 (22.1) |
| Advanced | 37.5 (31.1) | 38.5 (27.2) | 39 (25.3) |
| Percent of Students Proficient or Above | **78.6** | **79.5** | **79.5** |

**Table ES-7**
**Cut Scores by Round: Writing**
**(Percentages of students in each group are shown in parentheses.)**

| | **Round** | | |
|---|---|---|---|
| **Category** | **1** | **2** | **3** |
| Limited | -- (6.0) | -- (5.1) | -- (5.1) |
| Basic | 19 (14.4) | 18 (12.7) | 18 (12.2) |
| Proficient (Graduation) | 27 (30.9) | 26 (38.7) | 25.5 (34.0) |
| Accelerated | 34 (41.9) | 34.5 (39.7) | 34 (41.9) |
| Advanced | 41 (6.8) | 42 (3.8) | 41 (6.8) |
| Percent of Students Proficient or Above | **79.6** | **82.2** | **82.7** |

As can be seen in Tables ES-5, 6, and 7, there was some movement from Round1 to Round 2, but very little from Round 2 to Round 3. In Science, the cut score for Proficient went up one point from Round 1 to Round 2 (from 22.5 to 23.5), and then did not change in Round 3. In Social Studies, the Proficient cut score dropped half a point from Round 1 to Round 2 (22 to 21.5) and then remained unchanged in Round 3. In Writing, the cut score for Proficient fell one point from Round 1 to Round 2 (27 to 26) and another half a point at Round 3 (to 25.5). Cut scores for the other levels followed a similar pattern. The one odd pattern was in Social Studies; while the cut scores for Proficient were going down, all the others were going up from round to round. The effect was a broadening of the Proficient category.

## Conclusion

Standard setting is a combination of art and science. It combines the democratic process of group interaction and decision making with carefully planned and executed steps based on well-defined mathematical models. The processes by which the standard-setting activities for the Science, Social Studies, and Writing tests of the OGT were carried out, described in detail in the body of this report, were meticulously crafted by experienced psychometricians and reviewed by a national body of experts in the field. The plans were carried out under the supervision of ODE staff and two external reviewers who are also experts in this field. Because there are no "true" cut scores for any test, the recommended cut scores are only as valid as the processes by which they were derived. By all accounts, these standard-setting activities were well planned and executed, and the process was sound.

The 71 panelists in this activity – the Ohio educators and stakeholders invited to review the tests and provide the ratings – voiced overwhelming support for the process and outcomes. A complete analysis of their responses to 14 specific evaluation statements, along with their comments, is included in the final section of this report.

# Overview

On April 25-27, Dr. Michael Bunch, Dr. Elliot Inman, and Dr. Julie Miles of Measurement Incorporated (MI) met with groups of Ohio educators, parents, and community representatives to conduct standard-setting activities for the Ohio Graduation Tests in Science, Social Studies, and Writing. The standard-setting activities were conducted in accordance with a plan developed by Drs. Bunch and Inman and approved by the Ohio Department of Education (ODE) and Ohio's Technical Advisory Committee (TAC). That plan is included as Appendix A to this report.

This report documents the three-day standard-setting process and ensuing cut score recommendations of the panelists. It also provides a record of the fidelity of that process to the plan submitted to the TAC at its October 2004 meeting. Additional details are provided in six appendices:

A Standard Setting Plan
B Standard Setting Materials
C Impact Data
D Evaluations by Panelists
E Technical Details of the Bookmark Procedure
F Secure Documents

## Introductions and Training

**Day 1 (A.M.)**. On Monday, April 25, Dr. Bunch opened the three-day session at 8:30 A.M. and introduced MI staff and observers Dr. Thomas Hirsch and Dr. John Keene of Assessment and Evaluation Services (AES). He then introduced Ms. Judy Feil of the ODE, who introduced other ODE staff. Panelists included teachers as well as administrators, parents, and business and community representatives. A summary of panelist characteristics is included in Table 1.

Dr. Bunch presented an overview of the standard-setting process as well as the test-development process, emphasizing the involvement of teachers and community members at several points in the process. Dr. Bunch also stressed the role of the panelists as advisory to the ODE and the Ohio Board of Education, which will make a final decision on cut scores after considering the recommendations of these groups as well as the TAC and the Test Steering Committee (TSC). His presentation is also included in Appendix B. There were several questions during both presentations, and either MI or ODE staff answered all to the satisfaction of the questioners.

**Table 1**
**Summary of Panelist Characteristics**

| Category | Science | Social Studies | Writing | Total |
|---|---|---|---|---|
| Female | 12 | 9 | 15 | 36 |
| Male | 13 | 12 | 10 | 35 |
| African American | 3 | 5 | 7 | 15 |
| Asian/Pacific Islander | 3 | 0 | 1 | 4 |
| Hispanic | 0 | 1 | 1 | 2 |
| White | 18 | 15 | 15 | 48 |
| Other Race | 1 | 0 | 1 | 2 |
| Teacher | 17 | 13 | 17 | 47 |
| Higher Education | 1 | 1 | 1 | 3 |
| School Board | 0 | 1 | 1 | 2 |
| District/School Administrator | 3 | 2 | 2 | 7 |
| Business/Community Representative/Parent | 4 | 4 | 4 | 12 |
| Total Members | 25 | 21 | 25 | 71 |

At 10:15 A.M., panelists adjourned to their separate meeting rooms and prepared to take the spring test that students had just completed. For the rest of the day, the Science, Social Studies, and Writing groups remained separated from each other, with Dr. Miles leading the Social Studies group, Dr. Inman leading the Science group, and Dr. Bunch leading the Writing group. Overview and directions for test administration took approximately 15 minutes. Testing started at about 10:30 A.M. for both groups and ended between 12 noon and 12:30 P.M.. After a one-hour break for lunch, panelists reconvened to score their tests, using scoring keys and guides prepared by MI. These keys and guides are included in a separate collection of secure documents.

**Day 1 (P.M.)**. The scoring guides consisted of scoring rubrics for each of the open-ended items in the tests, followed by examples of student responses at each score point. Some of the student responses were annotated, and some were not. Drs. Bunch and Inman discussed the rubrics and sample papers for each item and then directed panelists to score their own responses by comparing them to the rubrics and sample responses. Both Dr. Bunch and Dr. Inman stressed the fact that the training they were providing for scoring the open-ended responses was far less detailed than that provided for MI scorers. MI scorer training lasts several days and includes multiple rounds of scoring of sets of student responses, ending with one or more qualifying rounds in which potential scorers must match scores for a range of student responses previously scored by MI senior scoring leaders. Those who fail to match the specified number of scores are retrained or released from the project. The purpose of the training that standard-setting panelists received was not to make them professional scorers but to give them just enough insight into the process to score their own responses.

After panelists had scored their tests, they took a short break and reconvened to examine and discuss achievement level definitions. These definitions, developed by the ODE, describe in general terms the typical skills and proficiencies of students at the Limited, Basic, Proficient, Accelerated, and Advanced levels of achievement. These definitions are included in Appendix B (as Performance Level Descriptors or PLD's). Dr. Bunch led a discussion in the Writing group, while Dr. Inman led a discussion in the Science group, and Dr. Miles led a discussion in the Social Studies group. Participants were encouraged to add clarifying statements to their definitions for use later in the standard-setting process.

Having studied and discussed the definitions for all five achievement levels, panelists in the Science and Social Studies groups were then encouraged to think of the student who would just barely qualify as Basic, Proficient, Accelerated, and Advanced. Dr. Inman emphasized the fact that each achievement level is a broad band with a considerable difference between students at the high end and those at the low end of the band. By focusing on the just barely Basic, Proficient, Accelerated, or Advanced student, panelists would be identifying the starting point for each of these achievement levels; hence, they would be identifying the cut points. These threshold concepts are considered very important to the type of standard-setting activity these groups would conduct. Meanwhile, in the Writing room, Dr. Bunch led a discussion about the performance level definitions that focused on the full range of performance within each category. The discussions of achievement levels ended at approximately 4 P.M. for all three groups. Drs. Bunch, Inman, and Miles collected and accounted for all secure materials and adjourned for the day.

After the first day, which was devoted almost entirely to tasks common to all three content areas, virtually all activities were carried out in group-specific meetings. Thus, the next section of this report is by content area: Science, Social Studies, and Writing.

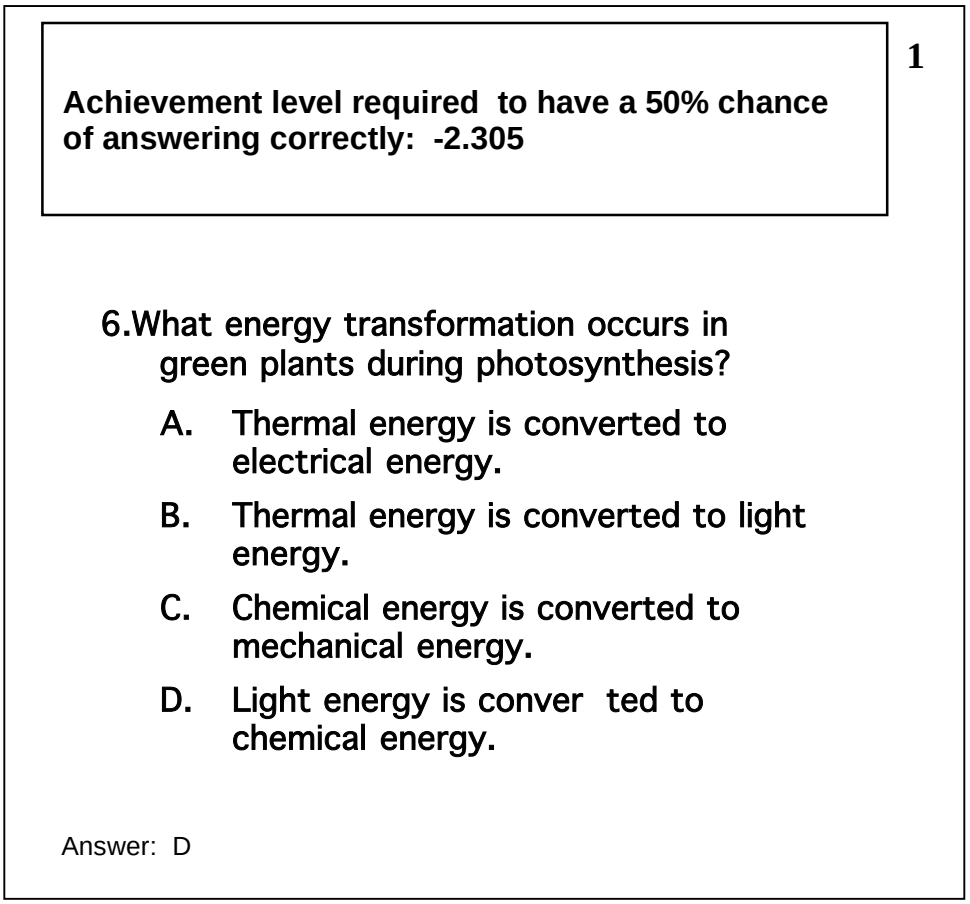
## Standard Setting By Content Area

### Science

**Day 2 (A.M.)**. On Tuesday, April 26, panelists convened at 8:30 A.M. for an orientation to the specific standard-setting procedure they would employ. Dr. Inman led the orientation to the bookmark procedure for the Science and Social Studies panels. A copy of the presentation is included in Appendix B, while technical details are included in the standard setting plan (Appendix A). During the bookmark presentation, Dr. Inman reminded panelists of the discussions the previous day, emphasizing the notion of "just barely" Basic, Proficient, Accelerated, and Advanced.

*Bookmark practice round*. At the end of the bookmark presentation, panelists broke into two groups: Science (led by Dr. Inman) and Social Studies (led by Dr. Miles). Each group received a practice test of six items (representing a total of 10 points), including both multiple-choice and open-ended items. They were instructed to begin on page 1 of the difficulty-ordered test booklet, consider the barely proficient student, and apply the three questions above to each

item. At the point at which they were no longer able to answer "Yes" to the final question, they were to go back to the previous question (i.e., the last one for which they could answer "Yes") and place a bookmark. The practice test booklets are included in a separate document that contains all test booklets and other secure materials. A sample page is shown in Figure 1. The booklet is included in Appendix F.


1

**Achievement level required to have a 50% chance of answering correctly: -2.305**

6.What energy transformation occurs in green plants during photosynthesis?

A. Thermal energy is converted to electrical energy.

B. Thermal energy is converted to light energy.

C. Chemical energy is converted to mechanical energy.

D. Light energy is conver ted to chemical energy.

Answer: D

**Figure 1. Sample page from a difficulty-ordered Science test booklet**

The large number in the upper right corner of the page is the page number in the difficulty-ordered booklet. The item in this example would have been on the first page. The information following the difficulty-ordered page number is the original item number and Rasch-based achievement data showing the student achievement level required to have a 50% chance to answer this question correctly. At the bottom of the page is the correct answer for the item.

Had this been an open-ended item, the question or prompt would have been followed by a sample response at a particular score point. In a full difficulty-ordered test booklet, each such item would appear once for each of its score points, twice for a 2-point item and four times for a 4-point item. Since it is more difficult to receive a 4 than a 1 (i.e., a score of 4 represents a higher level of student achievement than does a score of 1), the page with the "4" response

would appear much later in the test booklet than the page with the "1" response. In the practice booklet, consisting of only six pages, only selected pages were included.

After completing the practice test, panelists called out their bookmarked pages, and Dr. Inman tallied them. Panelists then discussed reasons for setting bookmarks at various pages in the booklet. Finally Dr. Inman, using the Rasch-based student achievement level printed at the top of each page, calculated the mean achievement level associated with the page numbers called out by the panelists and translated this achievement level into a raw score, using the Rasch model. Details of this procedure are included in Round 1.

Panelists then reviewed the procedure, their bookmarks in the practice booklets, and the achievement level definitions. Dr. Inman asked for and responded to questions and called panelists' attention to the Readiness Form (included in Appendix B) in each panelist's packet. Panelists were asked to respond to the following statement:

> I have completed the practice test, and I understand what I need to do to complete Round 1.
> (Circle one): **Yes** **No**

Dr. Inman checked to make sure each panelist had responded positively to the Readiness Forms. All panelists had circled "Yes," so they proceeded to Round 1 after a break for lunch. After collecting and accounting for all secure materials, Dr. Inman dismissed the groups for lunch.

**Round 1**

After lunch, Dr. Inman reminded the group of the task before them and assigned them to teams of four or five such that the subject-matter teachers, non-subject-matter teachers, parents, administrators, and others were evenly divided among teams, giving each team diverse points of view. Panelists remained in these teams for the remainder of the session. Once they had joined their teams, each panelist received a panelist number. Each panelist then recorded this number on each new piece of material.

Dr. Inman distributed the difficulty-ordered test booklets, and a bookmark. The bookmark form is included in Appendix B, and the ordered booklets are included in a separate document with other secure materials. The bookmark has places for panelists to enter three bookmarks for Round 1 and again for Round 2. Panelists were directed to reintroduce themselves to the other members of their teams and to discuss briefly their views of the test contents and the nature of students just barely in the Basic, Proficient, Accelerated, and Advanced categories. They were then directed to work alone in silence to place their bookmark for Proficient. After placing the Proficient bookmark, group members discussed their placements, made any adjustments they thought appropriate, and entered the other three bookmarks.

After all members of a team had placed the remaining three bookmarks, they discussed them. Dr. Inman noted that the purpose of the discussion was to allow all team members to hear

from others on their teams before turning in their Round 1 bookmarks but not to move the team toward a consensus. After small-group discussion ranging from ten to forty minutes, each team member turned in his or her bookmark, test booklet, and other secure materials. After accounting for all materials, Dr. Inman dismissed individual panelists for the day.

After collecting all bookmarks, MI staff entered the page numbers and associated achievement levels into Microsoft Excel spreadsheets which calculated the mean and standard deviation of the achievement levels. These mean achievement levels, along with low, mean minus one standard deviation, mean plus one standard deviation, and high were translated into raw scores using Table 20.1 of the Rasch-based WINSTEPS output. Results are shown in Table 2.

**Table 2**
**Results for Science Round 1**

| Rater | Basic Pg | Basic Ach | Prof Pg | Prof Ach | Acc Pg | Acc Ach | Adv Pg | Adv Ach |
|---|---|---|---|---|---|---|---|---|
| 3 | 9 | -0.96500 | 25 | 0.02100 | 36 | 0.96400 | 43 | 1.46500 |
| 4 | 5 | -1.29500 | 25 | 0.02100 | 30 | 0.57200 | 39 | 1.25100 |
| 5 | 9 | -0.96500 | 27 | 0.50800 | 38 | 1.06100 | 44 | 1.56400 |
| 6 | 12 | -0.75300 | 30 | 0.57200 | 39 | 1.25100 | 46 | 1.91100 |
| 7 | 7 | -1.11200 | 18 | -0.43000 | 33 | 0.65500 | 45 | 1.63900 |
| 8 | 8 | -1.07300 | 24 | -0.05700 | 33 | 0.65500 | 43 | 1.46500 |
| 9 | 11 | -0.81900 | 21 | -0.23700 | 31 | 0.61900 | 38 | 1.06100 |
| 10 | 8 | -1.07300 | 25 | 0.02100 | 35 | 0.73000 | 39 | 1.25100 |
| 11 | 11 | -0.81900 | 21 | -0.23700 | 30 | 0.57200 | 39 | 1.25100 |
| 12 | 8 | -1.07300 | 18 | -0.43000 | 33 | 0.65500 | 43 | 1.46500 |
| 13 | 8 | -1.07300 | 25 | 0.02100 | 34 | 0.67000 | 38 | 1.06100 |
| 14 | 13 | -0.73600 | 31 | 0.61900 | 38 | 1.06100 | 46 | 1.91100 |
| 15 | 12 | -0.75300 | 30 | 0.57200 | 39 | 1.25100 | 46 | 1.91100 |
| 16 | 4 | -1.40000 | 26 | 0.03500 | 38 | 1.06100 | 44 | 1.56400 |
| 17 | 8 | -1.07300 | 18 | -0.43000 | 33 | 0.65500 | 44 | 1.56400 |
| 18 | 9 | -0.96500 | 26 | 0.03500 | 35 | 0.73000 | 42 | 1.42300 |
| 19 | 13 | -0.73600 | 31 | 0.61900 | 38 | 1.06100 | 47 | 2.35700 |
| 20 | 12 | -0.75300 | 18 | -0.43000 | 33 | 0.65500 | 45 | 1.63900 |
| 21 | 9 | -0.96500 | 24 | -0.05700 | 31 | 0.61900 | 43 | 1.46500 |
| 22 | 11 | -0.81900 | 26 | 0.03500 | 37 | 1.01500 | 42 | 1.42300 |
| 23 | 8 | -1.07300 | 18 | -0.43000 | 33 | 0.65500 | 44 | 1.56400 |
| 24 | 18 | -0.43000 | 32 | 0.62000 | 42 | 1.42300 | 47 | 2.35700 |
| 25 | 13 | -0.73600 | 26 | 0.03500 | 34 | 0.67000 | 43 | 1.46500 |
| 26 | 4 | -1.40000 | 16 | -0.50900 | 38 | 1.06100 | 44 | 1.56400 |
| 27 | 8 | -1.07300 | 26 | 0.03500 | 32 | 0.62000 | 42 | 1.42300 |
| Mean | | -0.957 | | 0.021 | | 0.838 | | 1.561 |
| SD | | 0.227 | | 0.374 | | 0.254 | | 0.325 |
| M-1SD | | -1.184 | | -0.353 | | 0.584 | | 1.235 |
| M+1SD | | -0.731 | | 0.395 | | 1.091 | | 1.886 |
| Mean Cut | | 13 | | 22.5 | | 31 | | 37.5 |

Using this table, MI staff prepared figures showing the distributions of bookmarks. These distributions are shown in Figure 2

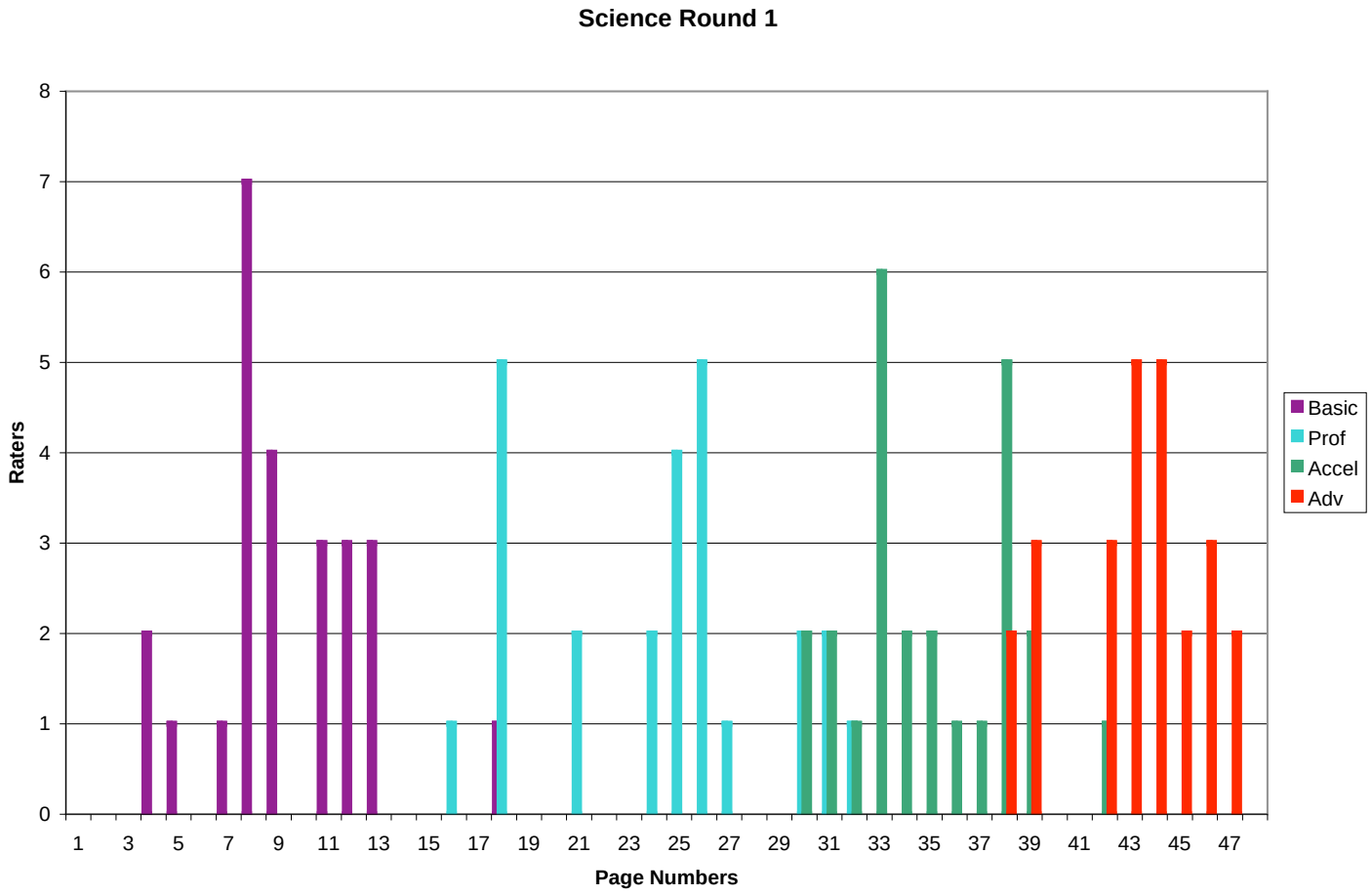



**Figure 2. Round 1 distribution of bookmarks for Science**

MI staff had also previously calculated the raw score distributions based on performances of students who took the tests in March. They rounded the mean cuts for both tests and applied them to the raw score distributions to create impact data for the entire group of students who took the tests. The same process was also performed by race and gender. All impact data are included in Appendix C.

## Round 2

On April 27, Dr. Inman convened the Science panel at 8:30 A.M. and distributed the Round 1 bookmarks, test booklets, and other materials including Table 2 and Figure 2. Large-group discussion of Round 1 began with discussion of Table 2 and Figure 2. Panelists discussed

the reasons for placing individual bookmarks as they had and what those placements represented. There was then further general discussion of the nature of the just barely Proficient student as well as the just barely Basic student and the just barely Accelerated and barely Advanced student, as well as the demands of the tests as a whole and of individual items. Of particular interest were comments regarding the intersection of the cognitive demands of individual items and the achievements and capabilities of students in each of the categories.

Dr. Inman then introduced the impact data in both tabular and graphic form (included in Appendix C) and led discussions of their implications. Results summarized in Table 3 reflect the final tabulations of scores with whole and half-point scores present.

**Table 3**
**Impact Data Summary for Science Round 1**

| Category | Cut Score (Out of 48 Points) | Percent in Category |
|---|---|---|
| Below Basic | -- | 7.1 |
| Basic | 13 | 17.5 |
| Proficient | 22.5 | 30.0 |
| Accelerated | 31 | 27.7 |
| Advanced | 37.5 | 17.7 |

After considerable discussion of the placement of the bookmarks, the relationship between bookmark placement and cut score, and impact data, Dr. Inman asked the panelists if they were ready to move on to Round 2. They agreed that they were ready, and Dr. Inman directed them to complete the appropriate section of the Readiness form:

> I have discussed the results of Round 1, including my ratings, the ratings of others, and the impact data, and I understand what I need to do to complete Round 2.
> (Circle one): **No** **Yes**

Panelists rejoined their teams and received their bookmarks, ordered test booklets, and other materials needed to complete Round 2. Instructions were the same as for Round 1. Panelists were to work alone in silence to place four bookmarks, starting with Proficient. They were to use their first round ratings, the comments they had heard during the morning discussion, the tables and graphs they had received or viewed and the impact data as they found any one or all pieces of information useful. After all members of a team had entered all four bookmarks, the team could discuss their ratings and make a final adjustment before turning in their individual bookmarks. As individual panelists completed their Round 2 tasks, Dr. Inman collected their bookmarks, accounted for secure materials, and dismissed the panel for lunch.

After the last panelist in each group had been dismissed for lunch, MI staff entered the page numbers and associated Rasch-based achievement levels in the same Microsoft Excel workbooks they had used for Round 1. Results are summarized in Table 4 and Figure 3. Table 4 may be interpreted exactly as Table 2 was.

**Table 4**
**Results for Science Round 2**

| Rater | Basic Pg | Basic Ach | Prof Pg | Prof Ach | Acc Pg | Acc Ach | Adv Pg | Adv Ach |
|---|---|---|---|---|---|---|---|---|
| 3 | 9 | -0.96500 | 25 | 0.02100 | 34 | 0.67000 | 41 | 1.31500 |
| 4 | 11 | -0.81900 | 24 | -0.05700 | 34 | 0.67000 | 43 | 1.46500 |
| 5 | 12 | -0.75300 | 26 | 0.03500 | 38 | 1.06100 | 46 | 1.91100 |
| 6 | 13 | -0.73600 | 29 | 0.54500 | 39 | 1.25100 | 46 | 1.91100 |
| 7 | 11 | -0.81900 | 26 | 0.03500 | 39 | 1.25100 | 45 | 1.63900 |
| 8 | 13 | -0.73600 | 26 | 0.03500 | 38 | 1.06100 | 44 | 1.56400 |
| 9 | 11 | -0.81900 | 24 | -0.05700 | 31 | 0.61900 | 43 | 1.46500 |
| 10 | 14 | -0.70400 | 26 | 0.03500 | 38 | 1.06100 | 45 | 1.63900 |
| 11 | 11 | -0.81900 | 29 | 0.54500 | 39 | 1.25100 | 44 | 1.56400 |
| 12 | 12 | -0.75300 | 26 | 0.03500 | 36 | 0.96400 | 43 | 1.46500 |
| 13 | 8 | -1.07300 | 26 | 0.03500 | 34 | 0.67000 | 44 | 1.56400 |
| 14 | 12 | -0.75300 | 28 | 0.53600 | 37 | 1.01500 | 44 | 1.56400 |
| 15 | 12 | -0.75300 | 30 | 0.57200 | 39 | 1.25100 | 46 | 1.91100 |
| 16 | 10 | -0.89400 | 24 | -0.05700 | 34 | 0.67000 | 43 | 1.46500 |
| 17 | 10 | -0.89400 | 24 | -0.05700 | 33 | 0.65500 | 44 | 1.56400 |
| 18 | 9 | -0.96500 | 29 | 0.54500 | 39 | 1.25100 | 43 | 1.46500 |
| 19 | 10 | -0.89400 | 24 | -0.05700 | 34 | 0.67000 | 43 | 1.46500 |
| 20 | 13 | -0.73600 | 25 | 0.02100 | 35 | 0.73000 | 46 | 1.91100 |
| 21 | 10 | -0.89400 | 24 | -0.05700 | 33 | 0.65500 | 43 | 1.46500 |
| 22 | 11 | -0.81900 | 26 | 0.03500 | 34 | 0.67000 | 41 | 1.31500 |
| 23 | 11 | -0.81900 | 25 | 0.02100 | 42 | 1.42300 | 44 | 1.56400 |
| 24 | 18 | -0.43000 | 32 | 0.62000 | 42 | 1.42300 | 47 | 2.35700 |
| 25 | 12 | -0.75300 | 27 | 0.50800 | 34 | 0.67000 | 43 | 1.46500 |
| 26 | 8 | -1.07300 | 18 | -0.43000 | 39 | 1.25100 | 44 | 1.56400 |
| 27 | 7 | -1.11200 | 24 | -0.05700 | 34 | 0.67000 | 45 | 1.63900 |
| Mean | | -0.831 | | 0.134 | | 0.941 | | 1.609 |
| SD | | 0.142 | | 0.283 | | 0.288 | | 0.231 |
| M-1SD | | -0.973 | | -0.149 | | 0.653 | | 1.378 |
| M+1SD | | -0.690 | | 0.417 | | 1.229 | | 1.840 |
| Mean Cut | | 14 | | 23.5 | | 32 | | 38 |

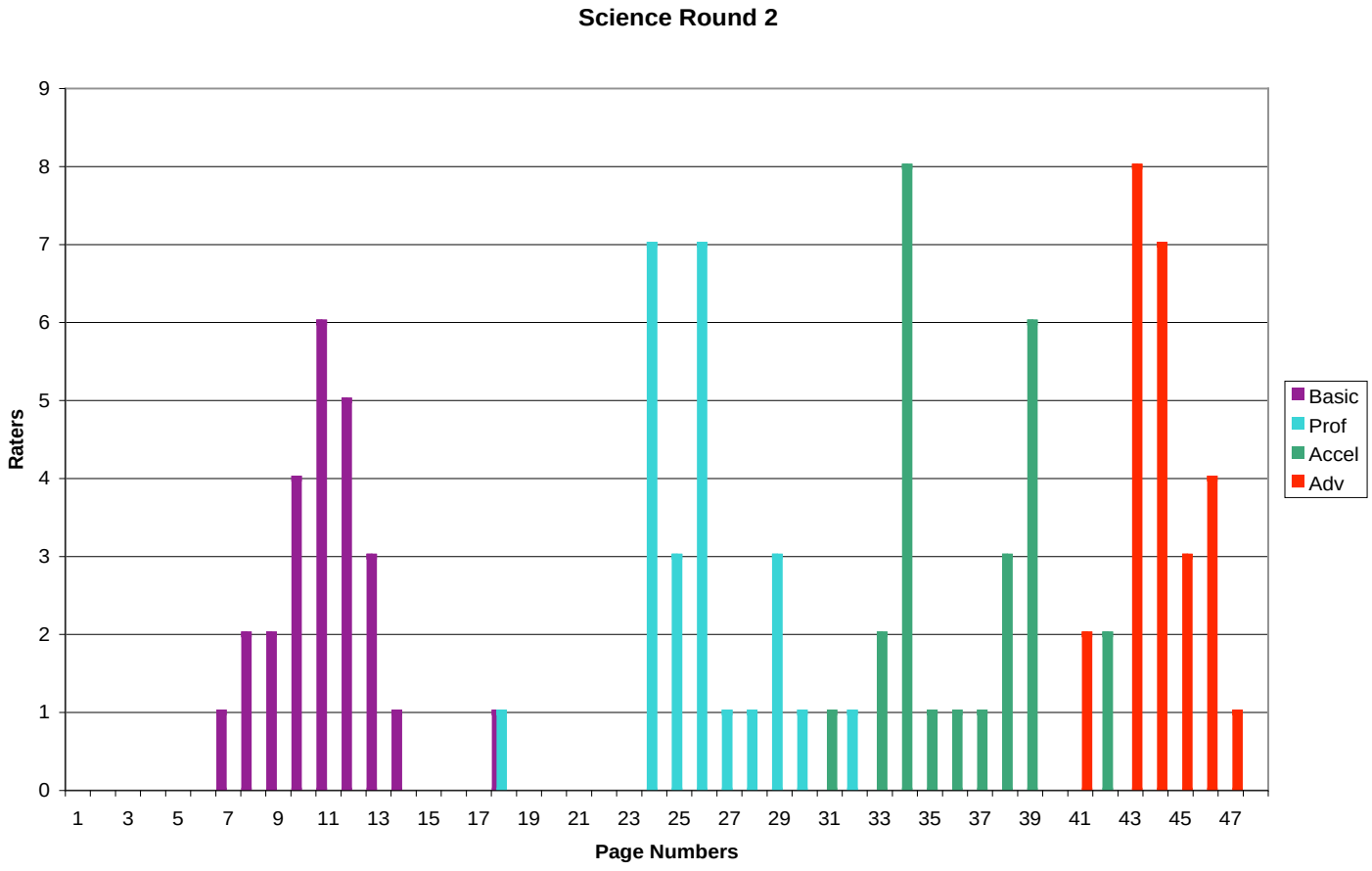



**Figure 3. Round 2 distribution of bookmarks for Science**

## Round 3

Dr. Inman distributed and led discussions of the Round 2 results, pointing out movement in the placement of bookmarks. All cut scores went up by half a point (Advanced) or a whole point (Basic, Proficient, and Accelerated). While there were still some outliers, there were far fewer than in Round 1, indicating that panelists had moved not just upward but toward consensus as well. This movement was reflected in the shrinking of the standard deviations of the Rasch theta values for all cut scores except Accelerated

Once again, panelists discussed their rationales for placing bookmarks as they did, why the percentages of students classified as Below Basic, Basic, Proficient, Accelerated, or Advanced were appropriate or inappropriate, and the relationship between item cognitive demands and the achievement level definitions. After considerable discussion, panelists completed their Readiness Forms, responding to the following statement:

> I have discussed the results of Round 2, including my ratings, the ratings of others, and the impact data, and I understand what I need to do to complete Round 3.
> (Circle one): **No** **Yes**

Dr. Inman inspected the forms and noted that all panelists had answered "Yes." He then directed panelists to begin Round 3 following essentially the same directions they had followed in Rounds 1 and 2, with a slight difference. Round 3 ratings were to include the page number and the actual cut scores and associated percentages of students scoring at or above that cut score, according to the tables distributed at the beginning of Round 2. The purpose of this shift in focus was to allow each panelist to leave the session knowing at least where he or she had set a recommended cut score and the impact that score would have on students.

As panelists completed their Round 3 bookmarks and turned them in, MI staff checked the three sets of numbers to make sure there was no confusion. Every panelist gave completely consistent final ratings and cut scores. Panelists then completed the remaining portion of the Readiness Form and an evaluation form created specifically for this standard-setting activity. The final four statements on the Readiness Form were as follows:

**Round 3:** I have completed my ratings, and I believe that the cut scores I have identified fairly represent minimal performances of students at the Basic, Proficient, Accelerated and Advanced levels
(Circle one): **No** **Yes**

Everyone was encouraged to share his or her ratings and hear those of other panelists**.**
(Circle one): **No** **Yes**

The cut scores we recommended accurately reflect the Basic, Proficient, Accelerated, and Advanced achievement levels.
(Circle one): **No** **Yes**

The process was fair and unbiased.
(Circle one): **No** **Yes**

MI staff checked the Readiness Forms as panelists turned in their materials prior to being dismissed. There was not a single negative response on any of the Readiness Forms collected.

ODE staff thanked the members of each group for their three days of effort and willingness to be a part of the process by which performance standards on the Ohio Graduation Tests are set. Dr. William Batchelor, MI project director also thanked the groups for their participation and distributed expense checks. Dr. Inman collected final ratings and all other secure materials. After each panelist's materials were accounted for, he or she was dismissed with a final word of thanks from Dr. Inman.

## Final Results for Science

Table 5 and Figure 4 summarize the results of Round 3. There was no movement in the cut score for Proficient from Round 2, but there was a further shrinkage of the standard deviation. The cut score for Accelerated also stayed the same as in Round 2, while the cut score for Basic went by half a point (to 14.5), and the cut score for Advanced came down by half a point (to 37.5). Overall, standard deviations shrank from Round 2 to Round 3, indicating greater consensus. In the end, there was only one reversal; one panelist placed the bookmark for Accelerated higher than the lowest bookmark for Advanced.

**Table 5**
**Results for Science Round 3**

| Rater | Basic Pg | Basic Ach | Prof Pg | Prof Ach | Acc Pg | Acc Ach | Adv Pg | Adv Ach |
|---|---|---|---|---|---|---|---|---|
| 3 | 9 | -0.96500 | 22 | -0.06900 | 34 | 0.67000 | 41 | 1.31500 |
| 4 | 11 | -0.81900 | 24 | -0.05700 | 34 | 0.67000 | 43 | 1.46500 |
| 5 | 12 | -0.75300 | 26 | 0.03500 | 38 | 1.06100 | 46 | 1.91100 |
| 6 | 13 | -0.73600 | 27 | 0.50800 | 39 | 1.25100 | 44 | 1.56400 |
| 7 | 11 | -0.81900 | 26 | 0.03500 | 33 | 0.65500 | 44 | 1.56400 |
| 8 | 12 | -0.75300 | 26 | 0.03500 | 38 | 1.06100 | 44 | 1.56400 |
| 9 | 11 | -0.81900 | 21 | -0.23700 | 32 | 0.62000 | 43 | 1.46500 |
| 10 | 14 | -0.70400 | 26 | 0.03500 | 38 | 1.06100 | 45 | 1.63900 |
| 11 | 11 | -0.81900 | 29 | 0.54500 | 39 | 1.25100 | 44 | 1.56400 |
| 12 | 8 | -1.07300 | 26 | 0.03500 | 36 | 0.96400 | 43 | 1.46500 |
| 13 | 11 | -0.81900 | 26 | 0.03500 | 34 | 0.67000 | 44 | 1.56400 |
| 14 | 12 | -0.75300 | 28 | 0.53600 | 37 | 1.01500 | 44 | 1.56400 |
| 15 | 12 | -0.75300 | 29 | 0.54500 | 39 | 1.25100 | 46 | 1.91100 |
| 16 | 10 | -0.89400 | 24 | -0.05700 | 38 | 1.06100 | 44 | 1.56400 |
| 17 | 10 | -0.89400 | 24 | -0.05700 | 33 | 0.65500 | 44 | 1.56400 |
| 18 | 9 | -0.96500 | 26 | 0.03500 | 39 | 1.25100 | 43 | 1.46500 |
| 19 | 10 | -0.89400 | 26 | 0.03500 | 34 | 0.67000 | 43 | 1.46500 |
| 20 | 13 | -0.73600 | 25 | 0.02100 | 35 | 0.73000 | 45 | 1.63900 |
| 21 | 11 | -0.81900 | 24 | -0.05700 | 35 | 0.73000 | 43 | 1.46500 |
| 22 | 11 | -0.81900 | 26 | 0.03500 | 34 | 0.67000 | 41 | 1.31500 |
| 23 | 11 | -0.81900 | 27 | 0.50800 | 41 | 1.31500 | 44 | 1.56400 |
| 24 | 18 | -0.43000 | 32 | 0.62000 | 42 | 1.42300 | 47 | 2.35700 |
| 25 | 12 | -0.75300 | 26 | 0.03500 | 34 | 0.67000 | 43 | 1.46500 |
| 26 | 13 | -0.73600 | 25 | 0.02100 | 39 | 1.25100 | 44 | 1.56400 |
| 27 | 7 | -1.11200 | 24 | -0.05700 | 34 | 0.67000 | 45 | 1.63900 |
| Mean | | -0.818 | | 0.123 | | 0.932 | | 1.585 |
| SD | | 0.132 | | 0.250 | | 0.273 | | 0.211 |
| M-1SD | | -0.950 | | -0.127 | | 0.659 | | 1.374 |
| M+1SD | | -0.686 | | 0.372 | | 1.204 | | 1.796 |
| Mean Cut | | 14.5 | | 23.5 | | 32 | | 37.5 |

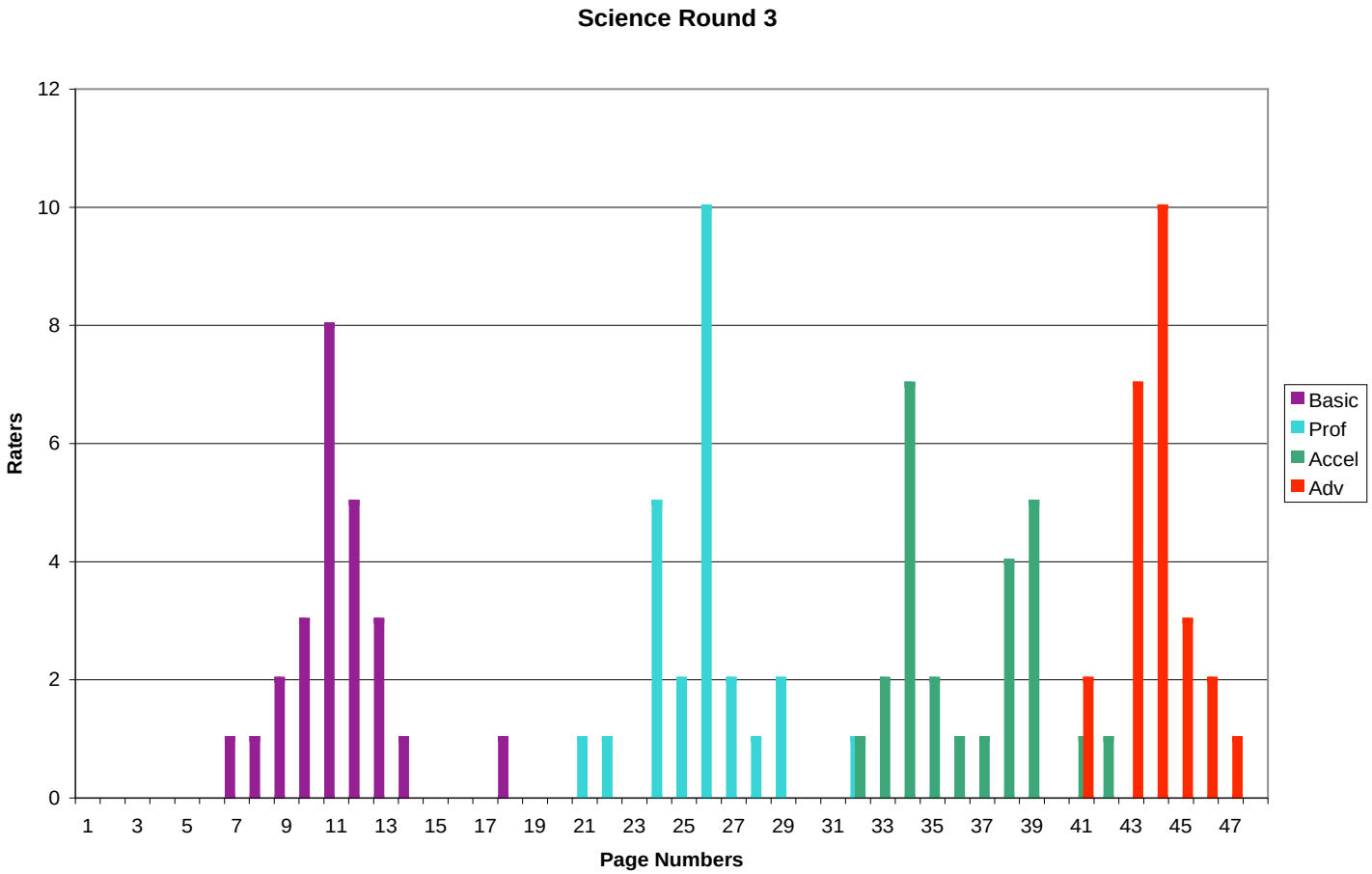



**Figure 4. Round 3 distribution of bookmarks for Science**

Table 6 summarizes the three rounds of ratings and associated raw cut scores over three rounds of ratings. There was minor movement in cut scores from Round 1 to Round 2 and then very little movement from Round 2 to Round 3.

**Table 6**

**Cut Scores by Round: Science**
**(Percentages of students in each group are shown in parentheses.)**

| | **Round** | | |
|---|---|---|---|
| **Category** | 1 | 2 | 3 |
| Limited | -- (7.1) | -- (8.5) | -- (9.5) |
| Basic | 13 (17.5) | 14 (18.7) | 14.5 (17.7) |
| Proficient (Graduation) | 22.5 (30.0) | 23.5 (31.7) | 23.5 (31.7) |
| Accelerated | 31 (27.7) | 32 (25.1) | 32 (23.4) |
| Advanced | 37.5 (17.7) | 38 (16.0) | 37.5 (17.7) |
| Percent of Students Proficient or Above | **75.4** | **72.8** | **72.8** |

**Impact**. Table 6 shows the final (Round 3) cut scores and impact (percentage of students in each category). Figure 5 summarizes the same information graphically.

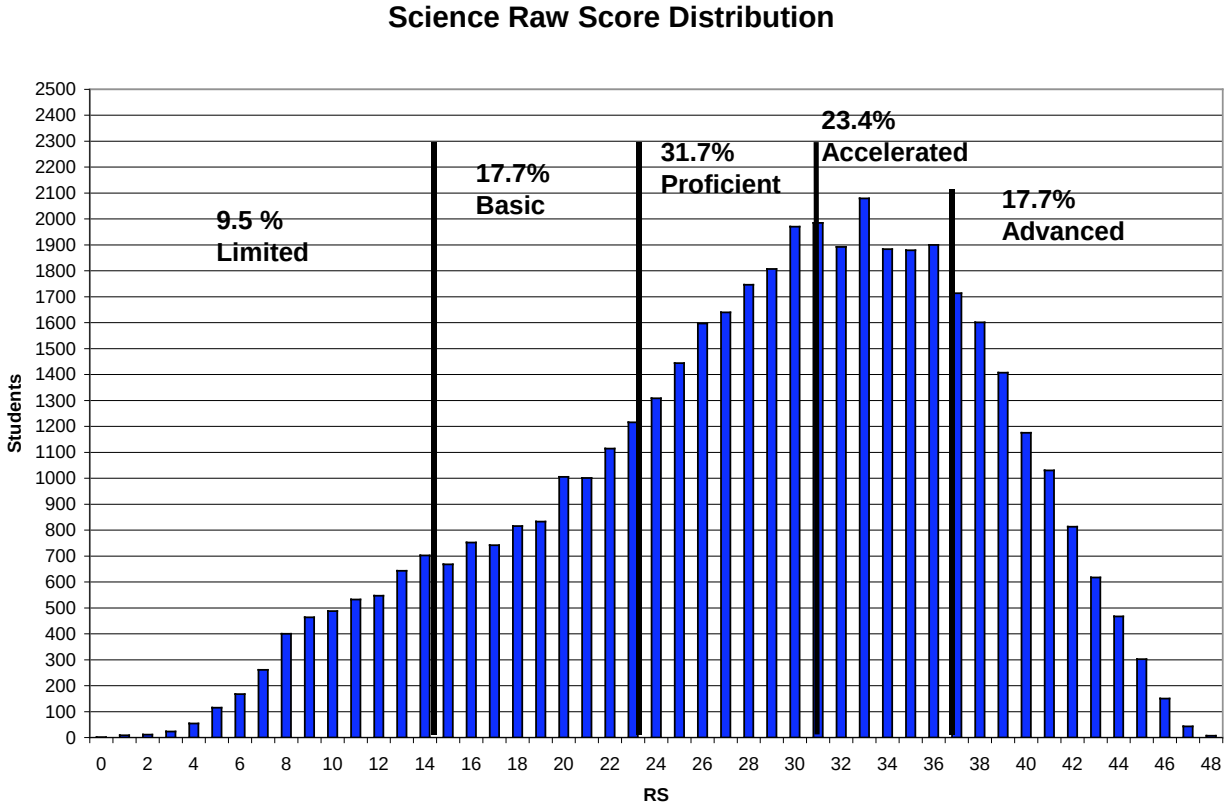



**Figure 5. Raw score distribution for Science**

While Table 5 and figure 5 show the overall distributions of students by achievement level, they do not address possible differences among groups. Table 7 shows these same distributions by race and sex.

**Table 7**
**Distribution of Students by Race and Sex: Science**
**(Entries are percentages.)**

| | Category | | | | | Proficient or Above |
|---|---|---|---|---|---|---|
| **Group (Tested)** | **Limited** | **Basic** | **Proficient** | **Accelerated** | **Advanced** | |
| AmInd (86) | 12.79 | 22.09 | 34.88 | 18.60 | 11.63 | 65.11 |
| As-PI (545) | 7.16 | 14.68 | 24.22 | 22.94 | 31.01 | 78.17 |
| BL-AA (6323) | 29.81 | 35.52 | 24.77 | 7.37 | 2.53 | 34.67 |
| Hisp (927) | 21.90 | 29.34 | 28.91 | 12.19 | 7.66 | 48.76 |
| Multi (581) | 9.98 | 22.20 | 33.91 | 19.62 | 14.29 | 67.82 |
| Other (140) | 24.29 | 16.43 | 29.29 | 16.43 | 13.57 | 59.29 |
| White (36,419) | 5.62 | 14.29 | 33.07 | 26.53 | 20.49 | 80.09 |
| Total by Race 45,021 | **9.49** | **17.71** | **31.72** | **23.37** | **17.71** | **72.80** |
| | | | | | | |
| Female (22,075) | 9.21 | 19.87 | 32.64 | 22.70 | 15.58 | 70.92 |
| Male (22,891) | 9.71 | 15.61 | 30.82 | 24.05 | 19.80 | 74.67 |
| Total by Sex 44,966 | **9.46** | **17.70** | **31.72** | **23.39** | **17.73** | **72.83** |

**Readiness and evaluations**. As noted above, there were no negative responses to any of the statements on the Readiness Form. Results of the evaluation were extremely positive, as shown in Table 8. Evaluations were quite good, with agreement rates ranging from 91 percent (item 7) to 100 percent (six of the nine items). Comments are included in Appendix D.

**Table 8**
**Summary of Evaluations: Science**
**(Entries are percentages.)**

| | Statement | Agree | Disagree |
|---|---|---|---|
| 1 | The workshop leaders clearly explained the purpose of the meeting. | 95% | 5% |
| 2 | The workshop leaders clearly explained my task. | 100% | 0% |
| 3 | The examples and exercises helped me understand how to perform my task. | 100% | 0% |
| 4 | The large and small group discussions helped me understand the process. | 100% | 0% |
| 5 | I was able to follow the instructions and complete the rating sheets accurately. | 100% | 0% |
| 6 | The discussions after the first round of rating were helpful to me. | 100% | 0% |
| 7 | The discussions after the second round of rating were helpful to me | 91% | 9% |
| 8 | The information showing the distribution of student scores was helpful to me. | 100% | 0% |
| 9 | The facilities and food service helped to create a good working environment. | 95% | 5% |

## Social Studies

**Day 2 (A.M.)**. On Tuesday, April 26, panelists convened at 8:30 A.M. for an orientation to the specific standard-setting procedure they would employ. Dr. Inman led the orientation to the bookmark procedure for the Science and Social Studies panels. A copy of the presentation is included in Appendix B, while technical details are included in the standard setting plan (Appendix A). During the bookmark presentation, Dr. Inman reminded panelists of the discussions the previous day, emphasizing the notion of "just barely" Basic, Proficient, Accelerated, and Advanced.

*Bookmark practice round*. At the end of the bookmark presentation, panelists received a practice test of six items (representing a total of 10 points), including both multiple-choice and open-ended items. They were instructed to begin on page 1 of the difficulty-ordered test booklet, consider the barely proficient student, and apply the three questions above to each item. At the point at which they were no longer able to answer "Yes" to the final question, they were to go back to the previous question (i.e., the last one for which they could answer "Yes") and place a bookmark. The practice test booklets are included in a separate document that contains all test booklets and other secure materials. A sample page is shown in Figure 6. The booklet is included in Appendix E.

The large number in the upper right corner of the page is the page number in the difficulty-ordered booklet. The item in this example would have been on the first page. The information following the difficulty-ordered page number is the original item number and Rasch-based achievement data showing the student achievement level required to have a 50% chance to answer this question correctly. At the bottom of the page is the correct answer for the item.

Had this been an open-ended item, the question or prompt would have been followed by a sample response at a particular score point. In a full difficulty-ordered test booklet, each such item would appear once for each of its score points, twice for a 2-point item and four times for a 4-point item. Since it is more difficult to receive a 4 than a 1 (i.e., a score of 4 represents a higher level of student achievement than does a score of 1), the page with the "4" response would appear much later in the test booklet than the page with the "1" response. In the practice booklet, consisting of only six pages, only selected pages were included.

**1**

Achievement level required for a **50%** chance to answer correctly: **-0.960**

7. How did the U.S. Constitution change as a result of the ratification of the 19th Amendment?

A. The right of suffrage was extended to women.
B. Freedom of assembly was restricted.
C. The power of government decreased.
D. Freedom of the press was

Key A

**Figure 6. Sample page from a difficulty-ordered Social Studies test booklet**

After completing the practice test, panelists called out their bookmarked pages, and Dr. Miles tallied them. Panelists then discussed reasons for setting bookmarks at various pages in

the booklet. Finally Dr. Miles, using the Rasch-based student achievement level printed at the top of each page, calculated the mean achievement level associated with the page numbers called out by the panelists and translated this achievement level into a raw score, using the Rasch model. Details of this procedure are included in Round 1.

Panelists then reviewed the procedure, their bookmarks in the practice booklets, and the achievement level definitions. Dr. Miles asked for and responded to questions and called panelists' attention to the Readiness Form (included in Appendix B) in each panelist's packet. Panelists were asked to respond to the following statement:

> I have completed the practice test, and I understand what I need to do to complete Round 1.
> (Circle one): **Yes** **No**

Dr. Miles checked to make sure each panelist had responded positively to the Readiness Forms. All panelists had circled "Yes," so they proceeded to Round 1 after a break for lunch. After collecting and accounting for all secure materials, Dr. Miles dismissed the groups for lunch.

**Round 1**

After lunch, Dr. Miles reminded the group of the task before them and assigned them to teams of four or five such that the subject-matter teachers, non-subject-matter teachers, parents, administrators, and others were evenly divided among teams, giving each team diverse points of view. Panelists remained in these teams for the remainder of the session. Once they had joined their teams, each panelist received a panelist number. Each panelist then recorded this number on each new piece of material.

Dr. Miles distributed the difficulty-ordered test booklets, and a bookmark. The bookmark form is included in Appendix B, and the ordered booklets are included in a separate document with other secure materials. The bookmark has places for panelists to enter three bookmarks for Round 1 and again for Round 2. Panelists were directed to reintroduce themselves to the other members of their teams and to discuss briefly their views of the test contents and the nature of students just barely in the Basic, Proficient, Accelerated, and Advanced categories. They were then directed to work alone in silence to place their bookmark for Proficient. After placing the Proficient bookmark, group members discussed their placements, made any adjustments they thought appropriate, and entered the other three bookmarks.

After all members of a team had placed the remaining three bookmarks, they discussed them. Dr. Miles noted that the purpose of the discussion was to allow all team members to hear from others on their teams before turning in their Round 1 bookmarks but not to move the team toward a consensus. After small-group discussion ranging from ten to forty minutes, each team member turned in his or her bookmark, test booklet, and other secure materials. After accounting for all materials, Dr. Miles dismissed individual panelists for the day.

After collecting all bookmarks, MI staff entered the page numbers and associated achievement levels into Microsoft Excel spreadsheets which calculated the mean and standard deviation of the achievement levels. These mean achievement levels, along with low, mean minus one standard deviation, mean plus one standard deviation, and high were translated into raw scores using Table 20.1 of the Rasch-based WINSTEPS output. Results are shown in Table 9.

**Table 9**
**Results for Social Studies Round 1**

| Rater | Basic Pg | Basic Ach | Prof Pg | Prof Ach | Acc Pg | Acc Ach | Adv Pg | Adv Ach |
|---|---|---|---|---|---|---|---|---|
| 1 | 6 | -0.76794 | 12 | -0.41306 | 26 | 0.16397 | 44 | 1.4007 |
| 2 | 13 | -0.39632 | 32 | 0.589 | 45 | 1.422 | 48 | 1.884 |
| 3 | 2 | -1.49401 | 19 | -0.056 | 28 | 0.32632 | 44 | 1.4007 |
| 4 | 13 | -0.39632 | 29 | 0.333 | 41 | 1.19115 | 46 | 1.511 |
| 5 | 15 | -0.28349 | 29 | 0.333 | 44 | 1.4007 | 47 | 1.755 |
| 6 | 10 | -0.47204 | 23 | 0.06583 | 37 | 0.79873 | 43 | 1.355 |
| 7 | 13 | -0.39632 | 23 | 0.06583 | 41 | 1.19115 | 48 | 1.884 |
| 8 | 10 | -0.47204 | 23 | 0.06583 | 40 | 1.018 | 44 | 1.4007 |
| 9 | 11 | -0.43834 | 19 | -0.056 | 26 | 0.16397 | 45 | 1.422 |
| 10 | 4 | -1.10577 | 13 | -0.39632 | 43 | 1.355 | 46 | 1.511 |
| 11 | 13 | -0.39632 | 29 | 0.333 | 41 | 1.19115 | 45 | 1.422 |
| 12 | 8 | -0.61739 | 29 | 0.333 | 37 | 0.79873 | 43 | 1.355 |
| 13 | 5 | -1.01667 | 23 | 0.06583 | 42 | 1.333 | 47 | 1.755 |
| 14 | 6 | -0.76794 | 19 | -0.056 | 35 | 0.739 | 44 | 1.4007 |
| 15 | 7 | -0.75225 | 19 | -0.056 | 34 | 0.71398 | 45 | 1.422 |
| 16 | 14 | -0.30163 | 23 | 0.06583 | 35 | 0.739 | 48 | 1.884 |
| 17 | 5 | -1.01667 | 29 | 0.333 | 40 | 1.018 | 43 | 1.355 |
| 18 | 13 | -0.39632 | 32 | 0.589 | 41 | 1.19115 | 44 | 1.4007 |
| 19 | 13 | -0.39632 | 33 | 0.69005 | 40 | 1.018 | 43 | 1.355 |
| 20 | 13 | -0.39632 | 23 | 0.06583 | 41 | 1.19115 | 47 | 1.755 |
| 21 | 6 | -0.76794 | 16 | -0.28169 | 29 | 0.333 | 44 | 1.4007 |
| Mean | | -0.621 | | 0.124 | | 0.919 | | 1.525 |
| SD | | 0.319 | | 0.304 | | 0.401 | | 0.198 |
| M-1SD | | -0.941 | | -0.179 | | 0.518 | | 1.327 |
| M+1SD | | -0.302 | | 0.428 | | 1.320 | | 1.723 |
| Mean Cut | | 14.5 | | 22 | | 31 | | 37.5 |

Using this table, MI staff prepared figures showing the distributions of bookmarks. These distributions are shown in Figure 7.

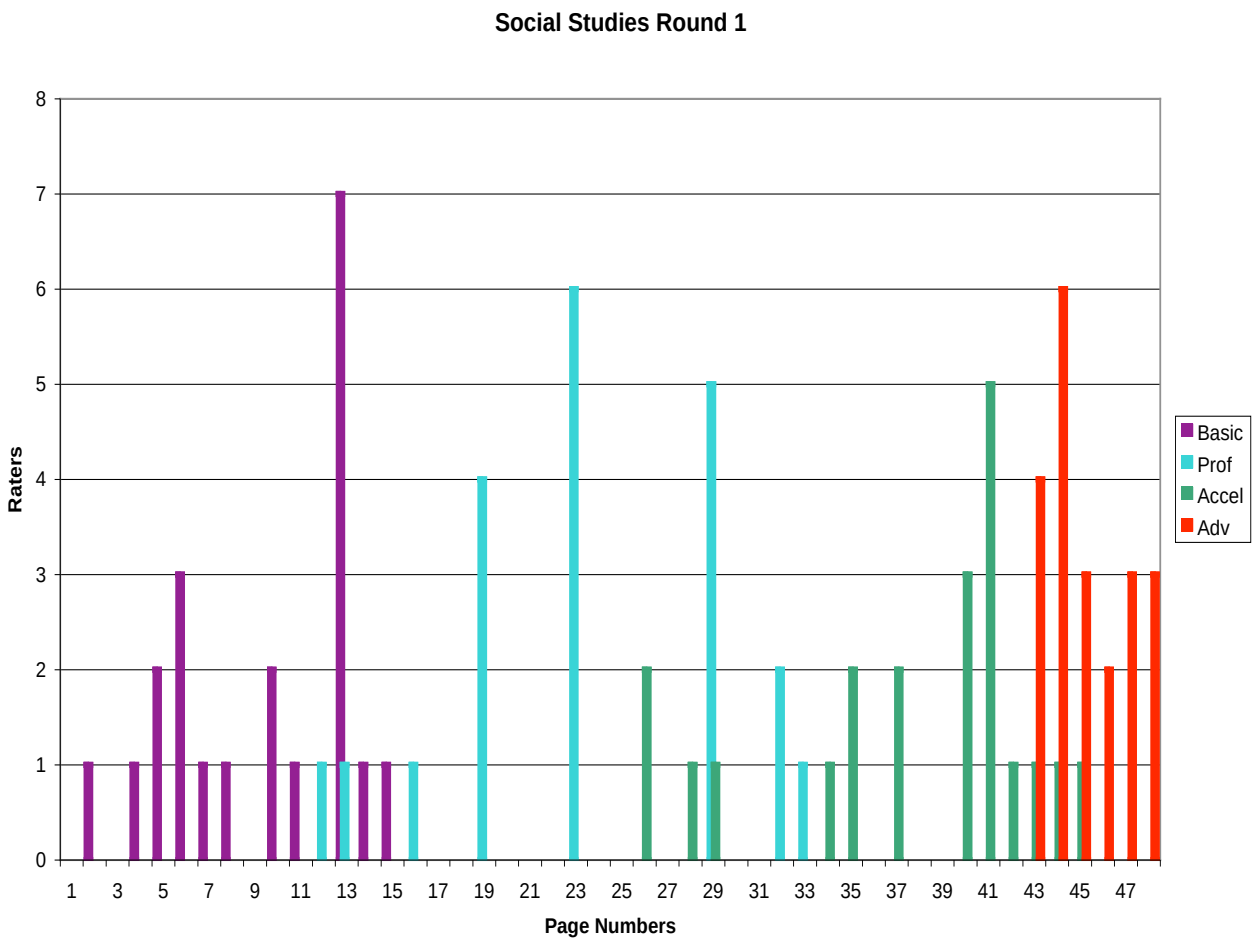



**Figure 7. Round 1 distribution of bookmarks for Social Studies**

MI staff had also previously calculated the raw score distributions based on performances of students who took the tests in March. They rounded the mean cuts for both tests and applied them to the raw score distributions to create impact data for the entire group of students who took the tests. The same process was also performed by race and gender. All impact data are included in Appendix C.

## Round 2

On April 27, Dr. Miles convened the Social Studies panel at 8:30 A.M. and distributed the Round 1 bookmarks, test booklets, and other materials including Table 9 and Figure 7. Large-group discussion of Round 1 began with discussion of Table 9 and Figure 7. Panelists discussed the reasons for placing individual bookmarks as they had and what those placements represented. There was then further general discussion of the nature of the just barely Proficient student as well as the just barely Basic student and the just barely Accelerated and barely Advanced student, as well as the demands of the tests as a whole and of individual items. Of

particular interest were comments regarding the intersection of the cognitive demands of individual items and the achievements and capabilities of students in each of the categories.

Dr. Miles then introduced the impact data in both tabular and graphic form (included in Appendix C) and led discussions of their implications. Results summarized in Table 10 reflect the final tabulations of scores with whole and half-point scores present.

**Table 10**
**Impact Data Summary for Social Studies Round 1**

| Category | Cut Score (Out of 48 Points) | Percent in Category |
|---|---|---|
| Below Basic | -- | 8.8 |
| Basic | 14.5 | 12.6 |
| Proficient | 22 | 24.3 |
| Accelerated | 31 | 23.2 |
| Advanced | 37.5 | 31.1 |

There was considerable discussion of the placements of the bookmarks as well as the impact data. Dr. Miles kept the discussion focused on the PLDs and how different perspectives on the PLDs led to the different cut scores. After it was clear that all panelists had voiced their opinions and asked all the questions they cared to ask, Dr. Miles asked them if they were ready to move on to Round 2 and directed them to the appropriate section of the Readiness Form:

> I have discussed the results of Round 1, including my ratings, the ratings of others, and the impact data, and I understand what I need to do to complete Round 2.
> (Circle one): **No** **Yes**

Panelists rejoined their teams and received their bookmarks, ordered test booklets, and other materials needed to complete Round 2. Instructions were the same as for Round 1. Panelists were to work alone in silence to place four bookmarks, starting with Proficient. They were to use their first round ratings, the comments they had heard during the morning discussion, the tables and graphs they had received or viewed and the impact data as they found any one or all pieces of information useful. After all members of a team had entered all four bookmarks, the team could discuss their ratings and make a final adjustment before turning in their individual bookmarks. As individual panelists completed their Round 2 tasks, Dr. Miles collected their bookmarks, accounted for secure materials, and dismissed the panel for lunch.

After the last panelist in each group had been dismissed for lunch, MI staff entered the page numbers and associated Rasch-based achievement levels in the same Microsoft Excel

workbooks they had used for Round 1. Results are summarized in Table 11 and Figure 8. Table 11 may be interpreted exactly as Table 9 was.

**Table 11**
**Results for Social Studies Round 2**

| Rater | Basic Pg | Basic Ach | Prof Pg | Prof Ach | Acc Pg | Acc Ach | Adv Pg | Adv Ach |
|---|---|---|---|---|---|---|---|---|
| 1 | 9 | -0.614 | 19 | -0.056 | 32 | 0.589 | 44 | 1.4007 |
| 2 | 13 | -0.39632 | 29 | 0.333 | 45 | 1.422 | 48 | 1.884 |
| 3 | 13 | -0.39632 | 29 | 0.333 | 43 | 1.355 | 44 | 1.4007 |
| 4 | 13 | -0.39632 | 27 | 0.215 | 41 | 1.19115 | 46 | 1.511 |
| 5 | 10 | -0.47204 | 23 | 0.06583 | 40 | 1.018 | 48 | 1.884 |
| 6 | 8 | -0.61739 | 19 | -0.056 | 35 | 0.739 | 43 | 1.355 |
| 7 | 13 | -0.39632 | 23 | 0.06583 | 38 | 0.864 | 48 | 1.884 |
| 8 | 7 | -0.75225 | 14 | -0.30163 | 40 | 1.018 | 44 | 1.4007 |
| 9 | 4 | -1.10577 | 17 | -0.27581 | 24 | 0.10441 | 43 | 1.355 |
| 10 | 4 | -1.10577 | 14 | -0.30163 | 33 | 0.69005 | 46 | 1.511 |
| 11 | 10 | -0.47204 | 19 | -0.056 | 41 | 1.19115 | 47 | 1.755 |
| 12 | 9 | -0.614 | 29 | 0.333 | 41 | 1.19115 | 47 | 1.755 |
| 13 | 9 | -0.614 | 27 | 0.215 | 41 | 1.19115 | 46 | 1.511 |
| 14 | 11 | -0.43834 | 29 | 0.333 | 43 | 1.355 | 47 | 1.755 |
| 15 | 13 | -0.39632 | 23 | 0.06583 | 35 | 0.739 | 45 | 1.422 |
| 16 | 10 | -0.47204 | 19 | -0.056 | 35 | 0.739 | 47 | 1.755 |
| 17 | 6 | -0.76794 | 12 | -0.41306 | 30 | 0.43335 | 44 | 1.4007 |
| 18 | 11 | -0.43834 | 32 | 0.589 | 43 | 1.355 | 46 | 1.511 |
| 19 | 13 | -0.39632 | 28 | 0.32632 | 39 | 0.94807 | 48 | 1.884 |
| 20 | 13 | -0.39632 | 23 | 0.06583 | 41 | 1.19115 | 47 | 1.755 |
| 21 | 6 | -0.76794 | 16 | -0.28169 | 28 | 0.32632 | 44 | 1.4007 |
| Mean | | -0.573 | | 0.054 | | 0.936 | | 1.595 |
| SD | | 0.220 | | 0.271 | | 0.369 | | 0.203 |
| M-1SD | | -0.793 | | -0.216 | | 0.567 | | 1.392 |
| M+1SD | | -0.353 | | 0.325 | | 1.305 | | 1.798 |
| Mean Cut | | 15 | | 21.5 | | 31.5 | | 38.5 |

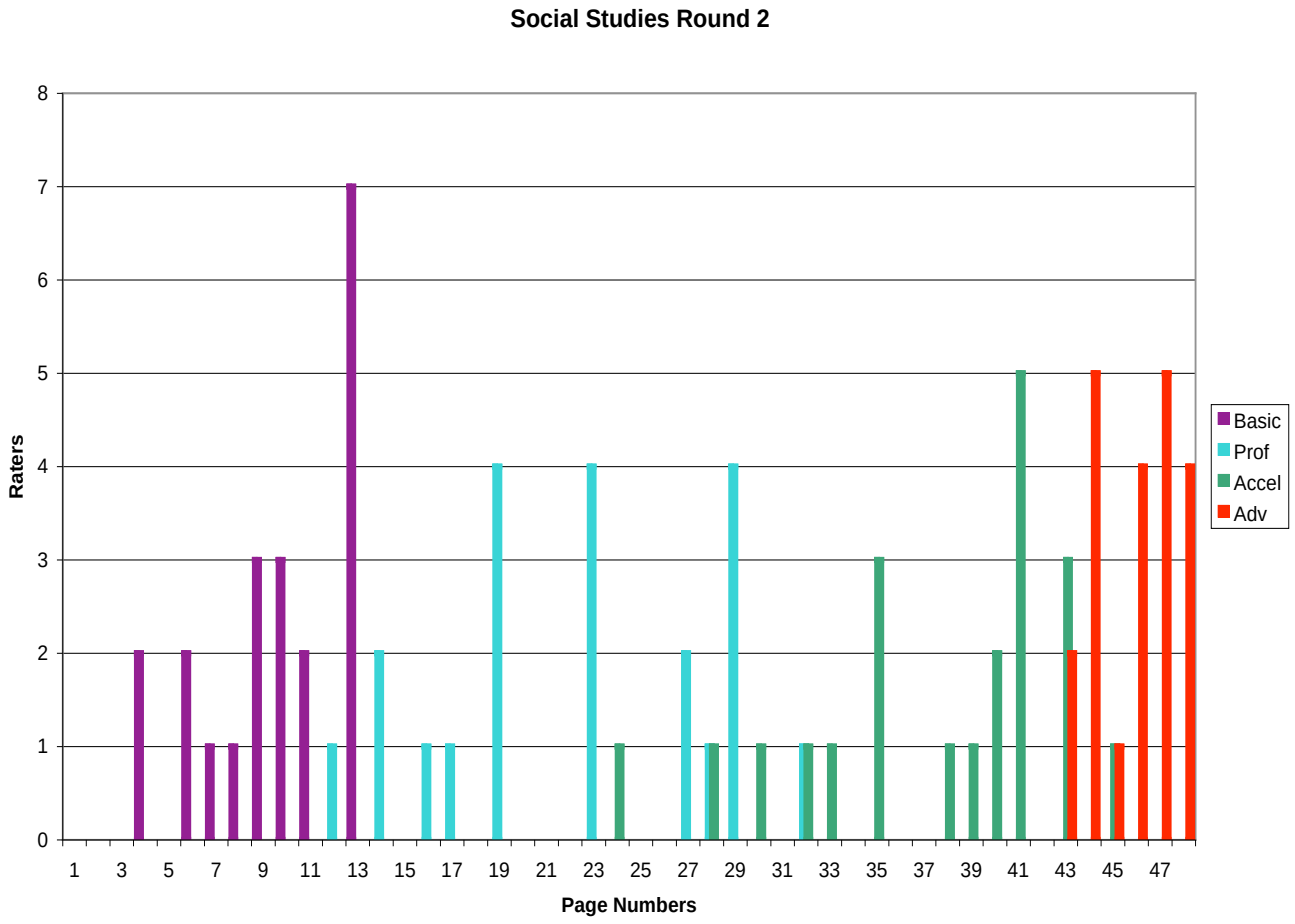



**Figure 8. Round 2 distribution of bookmarks for Social Studies**

From Round 1 to Round 2, the cut scores for Basic and Accelerated went up by half a point, the cut for Proficient went down by half a point, and the cut for Advanced went up by full point. The full-point rise in the Advanced cut score was clearly influenced by the impact data. Panelists were surprised that so many students would be in that category if Round 1 cut scores had held. There were also fewer instances in Round 2 of panelists encroaching on adjacent categories, as shown in Figure 8. In Figure 8, for example, the lowest bookmark placement for Proficient was 12, while the highest bookmark placement for Basic was 13. This situation was less extreme than that shown in Figure 7.

## Round 3

Dr. Miles distributed and led discussions of the Round 2 results, pointing out movement in the placement of bookmarks toward the means. Once again, panelists discussed their rationales for placing bookmarks as they did, why the percentages of students classified as Below Basic, Basic, Proficient, Accelerated, or Advanced were appropriate or inappropriate, and the relationship between item cognitive demands and the achievement level definitions. After considerable discussion, panelists completed their Readiness Forms, responding to the following statement:

I have discussed the results of Round 2, including my ratings, the ratings of others, and the impact data, and I understand what I need to do to complete Round 3.
(Circle one): **No** **Yes**

Dr. Miles inspected the forms and noted that all panelists had answered "Yes." She then directed panelists to begin Round 3 following essentially the same directions they had followed in Rounds 1 and 2, with a slight difference. Round 3 ratings were to include the page number and the actual cut scores and associated percentages of students scoring at or above that cut score, according to the tables distributed at the beginning of Round 2. The purpose of this shift in focus was to allow each panelist to leave the session knowing at least where he or she had set a recommended cut score and the impact that score would have on students.

As panelists completed their Round 3 bookmarks and turned them in, MI staff checked the three sets of numbers to make sure there was no confusion. Every panelist gave completely consistent final ratings and cut scores. Panelists then completed the remaining portion of the Readiness Form and an evaluation form created specifically for this standard-setting activity. The final four statements on the Readiness Form were as follows:

**Round 3:** I have completed my ratings, and I believe that the cut scores I have identified fairly represent minimal performances of students at the Basic, Proficient, Accelerated and Advanced levels
(Circle one): **No** **Yes**

Everyone was encouraged to share his or her ratings and hear those of other panelists**.**
(Circle one): **No** **Yes**

The cut scores we recommended accurately reflect the Basic, Proficient, Accelerated, and Advanced achievement levels.
(Circle one): **No** **Yes**

The process was fair and unbiased.
(Circle one): **No** **Yes**

MI staff checked the Readiness Forms as panelists turned in their materials prior to being dismissed. There was not a single negative response on any of the Readiness Forms collected.

ODE staff thanked the members of each group for their three days of effort and willingness to be a part of the process by which performance standards on the Ohio Graduation Tests are set. Dr. William Batchelor, MI project director also thanked the groups for their participation and distributed expense checks. Dr. Miles collected final ratings and all other

secure materials. After each panelist's materials were accounted for, he or she was dismissed with a final word of thanks from Dr. Miles.

## Final Results for Social Studies

Table 12 and Figure 9 summarize the results of Round 3.

**Table 12**
**Results for Social Studies Round 3**

| Rater | Basic Pg | Basic Ach | Prof Pg | Prof Ach | Acc Pg | Acc Ach | Adv Pg | Adv Ach |
|---|---|---|---|---|---|---|---|---|
| 1 | 13 | -0.39632 | 23 | 0.06583 | 32 | 0.589 | 44 | 1.4007 |
| 2 | 13 | -0.39632 | 29 | 0.333 | 45 | 1.422 | 48 | 1.884 |
| 3 | 13 | -0.39632 | 29 | 0.333 | 43 | 1.355 | 44 | 1.4007 |
| 4 | 13 | -0.39632 | 25 | 0.12973 | 41 | 1.19115 | 46 | 1.511 |
| 5 | 10 | -0.47204 | 23 | 0.06583 | 45 | 1.422 | 48 | 1.884 |
| 6 | 8 | -0.61739 | 18 | -0.13361 | 35 | 0.739 | 43 | 1.355 |
| 7 | 13 | -0.39632 | 23 | 0.06583 | 38 | 0.864 | 48 | 1.884 |
| 8 | 10 | -0.47204 | 19 | -0.056 | 44 | 1.4007 | 46 | 1.511 |
| 9 | 6 | -0.76794 | 19 | -0.056 | 24 | 0.10441 | 43 | 1.355 |
| 10 | 4 | -1.10577 | 14 | -0.30163 | 45 | 1.422 | 47 | 1.755 |
| 11 | 8 | -0.61739 | 19 | -0.056 | 43 | 1.355 | 47 | 1.755 |
| 12 | 10 | -0.47204 | 29 | 0.333 | 43 | 1.355 | 48 | 1.884 |
| 13 | 5 | -1.01667 | 27 | 0.215 | 41 | 1.19115 | 46 | 1.511 |
| 14 | 11 | -0.43834 | 29 | 0.333 | 45 | 1.422 | 48 | 1.884 |
| 15 | 13 | -0.39632 | 23 | 0.06583 | 35 | 0.739 | 46 | 1.511 |
| 16 | 10 | -0.47204 | 19 | -0.056 | 43 | 1.355 | 48 | 1.884 |
| 17 | 5 | -1.01667 | 12 | -0.41306 | 40 | 1.018 | 47 | 1.755 |
| 18 | 11 | -0.43834 | 29 | 0.333 | 43 | 1.355 | 46 | 1.511 |
| 19 | 13 | -0.39632 | 28 | 0.32632 | 39 | 0.94807 | 48 | 1.884 |
| 20 | 13 | -0.39632 | 23 | 0.06583 | 45 | 1.422 | 48 | 1.884 |
| 21 | 6 | -0.76794 | 16 | -0.28169 | 28 | 0.32632 | 45 | 1.422 |
| Mean | | -0.564 | | 0.062 | | 1.095 | | 1.658 |
| SD | | 0.233 | | 0.228 | | 0.397 | | 0.214 |
| M-1SD | | -0.797 | | -0.165 | | 0.698 | | 1.444 |
| M+1SD | | -0.331 | | 0.290 | | 1.492 | | 1.872 |
| Mean Cut | | 15 | | 21.5 | | 33 | | 39 |

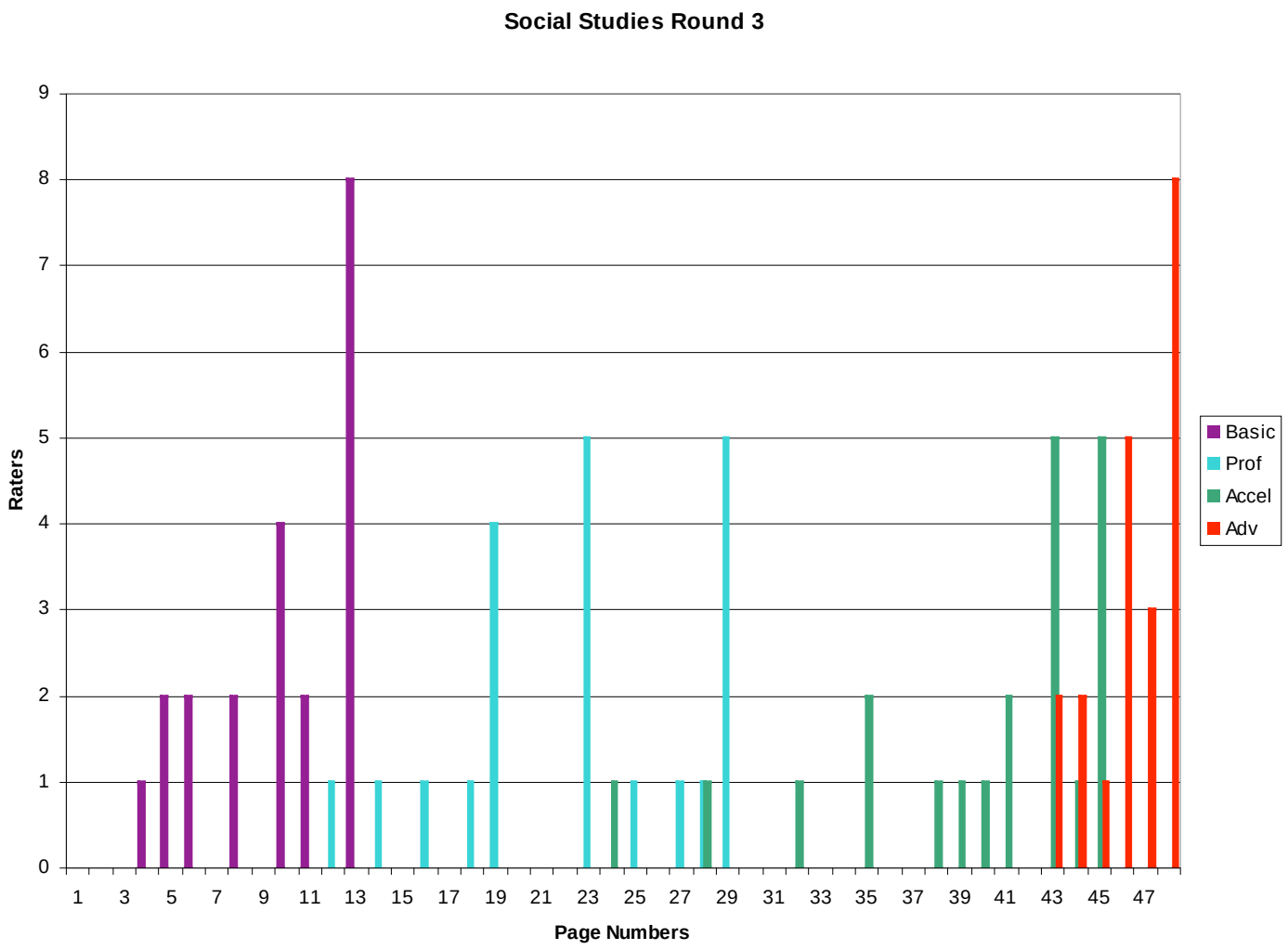



**Figure 9. Round 3 distribution of bookmarks for Social Studies**

From Round 2 to Round 3, the cut scores for Basic and Proficient remained unchanged, but the cut for Accelerated increased by 1.5 (to 33), and the cut for Advanced increased by half a point (to 39). As Figure 9 shows, there was still considerable range in the placements, with overlap at all levels (e.g., some Proficient bookmarks extending down into the region of the Basic bookmarks and some Accelerated bookmarks extending up into the region of the Advanced bookmarks).

Table 13 summarizes the three rounds of ratings and associated raw cut scores over three rounds of ratings. The percentage of students in the top three categories combined changed less than one percent from Round 1 to Round 2 and not at all from Round 2 to Round 3.

**Table 13**

**Cut Scores by Round: Social Studies**
**(Percentages of students in each group are shown in parentheses.)**

| | **Round** | | |
|---|---|---|---|
| **Category** | 1 | 2 | 3 |
| Limited | -- (8.8) | -- (9.4) | -- (9.4) |
| Basic | 14.5 (12.6) | 15 (11.1) | 15 (11.1) |
| Proficient (Graduation) | 22 (24.3) | 21.5 (26.9) | 21.5 (32.1) |
| Accelerated | 31 (23.2) | 31.5 (25.4) | 33 (22.1) |
| Advanced | 37.5 (31.1) | 38.5 (27.2) | 39 (25.3) |
| Percent of Students Proficient or Above | **78.6** | **79.5** | **79.5** |

**Impact**. Table 13 shows the final (Round 3) cut scores and impact (percentage of students in each category). Figure 10 summarizes the same information graphically.

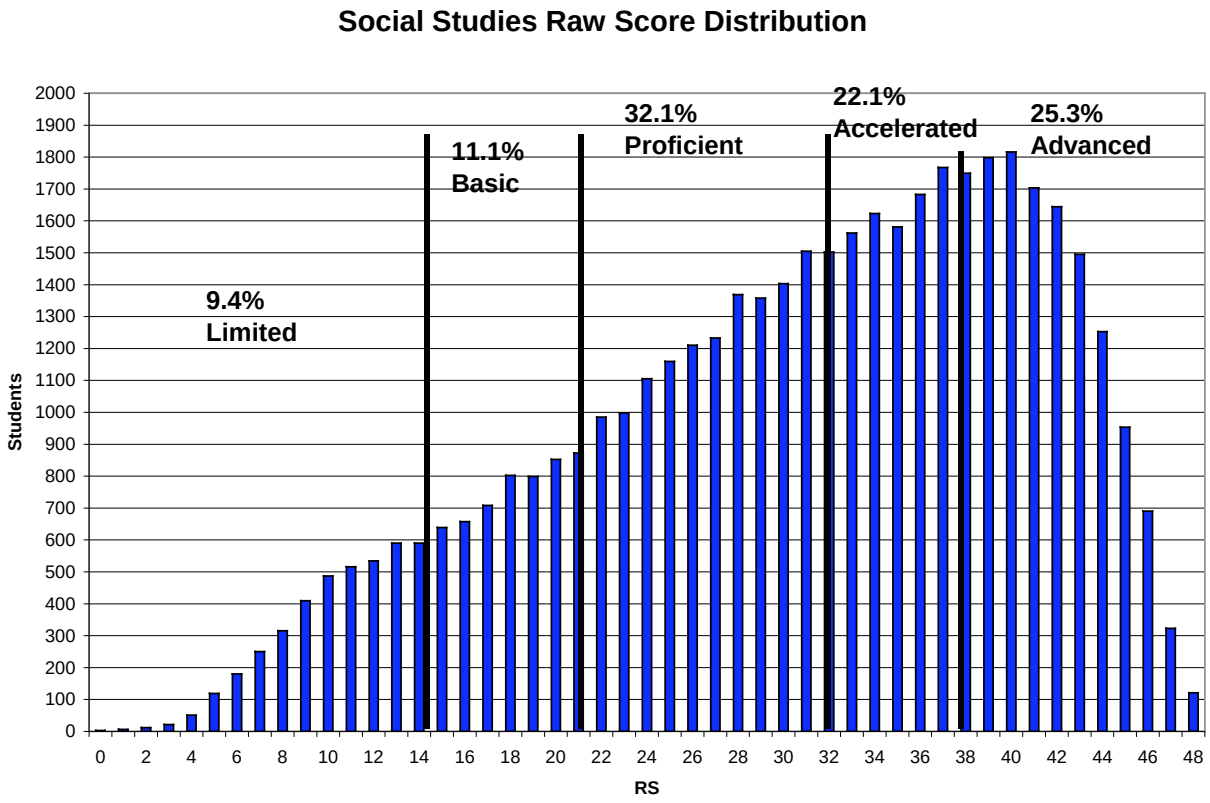



**Figure 10. Raw score distribution for Social Studies**

While Table 12 and figure 10 show the overall distributions of students by achievement level, they do not address possible differences among groups. Table 14 shows these same distributions by race and sex.

**Table 14**
**Distribution of Students by Race and Sex: Social Studies**
**(Entries are percentages.)**

| | Category | | | | | Proficient or Above |
|---|---|---|---|---|---|---|
| **Group (Tested)** | **Limited** | **Basic** | **Proficient** | **Accelerated** | **Advanced** | |
| AmInd (73) | 8.22 | 15.07 | 35.62 | 23.29 | 17.81 | 76.72 |
| As-PI (545) | 7.34 | 7.16 | 23.67 | 21.10 | 40.73 | 85.50 |
| BL-AA (6345) | 26.32 | 22.58 | 35.37 | 10.45 | 5.28 | 51.10 |
| Hisp (910) | 22.53 | 20.99 | 30.66 | 14.18 | 11.65 | 56.49 |
| Multi (601) | 9.98 | 11.65 | 37.27 | 19.13 | 21.96 | 78.36 |
| Other (107) | 27.10 | 10.28 | 27.10 | 16.82 | 18.69 | 62.61 |
| White (36,419) | 6.06 | 8.93 | 31.57 | 24.44 | 29.00 | 85.01 |
| Total by Race 45,000 | **9.37** | **11.13** | **32.06** | **22.13** | **25.31** | **79.50** |
| | | | | | | |
| Female (22,132) | 8.77 | 12.29 | 35.18 | 21.28 | 22.48 | 78.94 |
| Male (22,818) | 9.89 | 9.98 | 29.05 | 22.98 | 28.10 | 80.13 |
| Total by Sex 44,950 | **9.34** | **11.12** | **32.07** | **22.14** | **25.33** | **79.54** |

**Readiness and evaluations**. As noted above, there were no negative responses to any of the statements on the Readiness Form. Results of the evaluation were extremely positive, as shown in Table 15. Evaluations were quite positive. The only item to generate any significant negative response was item 7, concerning the value of the discussion after Round 2; 29 percent of respondents did not find value in that discussion. Otherwise, evaluations were extremely favorable. Details are included in Appendix D.

**Table 15**
**Summary of Evaluations: Social Studies**
**(Entries are percentages.)**

| | **Statement** | **Agree** | **Disagree** |
|---|---|---|---|
| 1 | The workshop leaders clearly explained the purpose of the meeting. | 95% | 5% |
| 2 | The workshop leaders clearly explained my task. | 91% | 9% |
| 3 | The examples and exercises helped me understand how to perform my task. | 95% | 5% |
| 4 | The large and small group discussions helped me understand the process. | 91% | 9% |
| 5 | I was able to follow the instructions and complete the rating sheets accurately. | 95% | 5% |
| 6 | The discussions after the first round of rating were helpful to me. | 95% | 5% |
| 7 | The discussions after the second round of rating were helpful to me | 71% | 29% |
| 8 | The information showing the distribution of student scores was helpful to me. | 95% | 5% |
| 9 | The facilities and food service helped to create a good working environment. | 95% | 5% |

## Writing

**Holistic practice round**. Dr. Bunch presented a set of six work samples to the Writing panelists with the instructions to review each sample and assign it to one of two categories: Basic/Below or Proficient/Above. Panelists examined the samples – each essay, short-answer response, and multiple-choice response for a single student – and entered their ratings on special forms (see Appendix B). Afterwards, Dr. Bunch calculated the cut score – the midpoint between the medians for the two categories – and shared the results with the panelists, who then discussed differences in ratings.

Panelists then reviewed the procedure, their rating sheets, and the achievement level definitions. Dr. Bunch then distributed the Readiness Form. Panelists were asked to respond to the following statement:

> I have completed the practice test, and I understand what I need to do to complete Round 1.
> (Circle one): **Yes** **No**

Dr. Bunch checked to make sure each panelist had responded positively to the Readiness Forms. All panelists had circled "Yes," so they proceeded to Round 1 after a break for lunch. After collecting and accounting for all secure materials, Dr. Bunch dismissed the panel for lunch.

**Round 1**

After lunch, Dr. Bunch reminded the group of the task before them and assigned them to teams of four or five such that the subject-matter teachers, non-subject-matter teachers, parents, administrators, and others were evenly divided among teams, giving each team diverse points of view. Panelists remained in these teams for the remainder of the session. Once they had joined their teams, each panelist received a panelist number.

Student work samples (150 in all) were arranged in numbered folders, with the lowest numbered folder containing the work sample with the lowest score and the highest numbered folder containing the sample with the highest score. This ordered arrangement of work samples was in accordance with advice of the Technical Advisory Committee and was consistent with the difficulty ordering of the test booklets in the bookmark procedure. Panelists entered their ratings for each work sample on special forms, a portion of which is shown in Figure 11. Instead of entering the name of the level, panelists entered numerical ratings: 1 for Limited, 2 for Basic, 3 for Proficient, 4 for Accelerated, and 5 for Advanced.

| Packet | Litho | Level | Packet | Litho | Level | Packet | Litho | Level |
|---|---|---|---|---|---|---|---|---|
| 1 | 101885 | | 26 | 313181 | | 51 | 282230 | |
| 2 | 277570 | | 27 | 278091 | | 52 | 314143 | |
| 3 | 101220 | | 28 | 263359 | | 53 | 262279 | |
| 4 | 282099 | | 29 | 278909 | | 54 | 278097 | |
| 5 | 278075 | | 30 | 281593 | | 55 | 278931 | |
| 6 | 281393 | | 31 | 314138 | | 56 | 314160 | |
| | | | | | | | | |
| 96 | 277503 | | 121 | 281946 | | 146 | 154842 | |
| 97 | 281597 | | 122 | 278969 | | 147 | 230739 | |
| 98 | 281869 | | 123 | 314169 | | 148 | 228316 | |
| 99 | 262261 | | 124 | 282131 | | 149 | 214036 | |
| 100 | 262278 | | 125 | 282282 | | 150 | 157379 | |

**Figure 11. Portion of entry form for Writing work samples**

Panelists worked through the afternoon, taking care to review work samples from the entire range of folders (i.e., some at each end as well as some in the middle). They worked in

this manner until approximately 3:30 P.M., at which time, Dr. Bunch recommended that they discuss their ratings within their small groups, finalize their ratings, and turn in their forms. Dr. Bunch collected all materials and dismissed the panel at 4:00 P.M.

MI staff then entered all ratings and calculated category medians by panelist and then the median of these medians to arrive at overall medians for each category. They then calculated the midpoint between adjacent category medians to determine the cut score for the categories Basic through Advanced. Results for Round 1 are presented in Table 16.

**Table 16**
**Results of Writing Round 1**

| Panelist | Packet | Limited | Basic | Proficient | Accelerated | Advanced |
|---|---|---|---|---|---|---|
| 1 | SUMMARY | 14.25 | 19 | 24 | 34.5 | 47 |
| 2 | SUMMARY | 14.5 | 19 | 25.5 | 33.5 | 37 |
| 3 | SUMMARY | 14 | 22 | 30.5 | 38 | 48 |
| 4 | SUMMARY | | 23 | 32 | 36 | 45 |
| 5 | SUMMARY | | 21.5 | 29 | 39 | 48 |
| 6 | SUMMARY | 17 | 25 | 32.25 | 43 | 48 |
| 7 | SUMMARY | | 25 | 31.5 | 44 | 44.5 |
| 8 | SUMMARY | 14 | 22 | 30 | 35.5 | 39.5 |
| 9 | SUMMARY | 23 | 31 | 34.25 | 41 | 47 |
| 10 | SUMMARY | 15.75 | 22 | 33.25 | 37 | 42.5 |
| 11 | SUMMARY | 18 | 26 | 32 | 34 | 39.5 |
| 12 | SUMMARY | 16 | 22 | 31.5 | 38.5 | 46 |
| 13 | SUMMARY | 16 | 21.5 | 30 | 36 | 42 |
| 14 | SUMMARY | 15 | 18 | 25.5 | 36 | 37.75 |
| 15 | SUMMARY | | 17.5 | 30.5 | 32.75 | 43.5 |
| 16 | SUMMARY | 15 | 23.5 | 32 | 36 | 43.5 |
| 17 | SUMMARY | | 26 | 31.75 | 37 | |
| 18 | SUMMARY | 11.5 | 25 | 31 | 36.5 | 46.5 |
| 19 | SUMMARY | | 26 | 30 | 38.25 | |
| 20 | SUMMARY | 11.5 | | 31 | 38 | |
| 21 | SUMMARY | 17 | 27 | 31.75 | 41.25 | 45 |
| 22 | SUMMARY | 15 | 20 | 31.5 | 38.5 | 42 |
| 23 | SUMMARY | 17 | 27.5 | 31.5 | 38 | 47 |
| 24 | SUMMARY | 14.5 | 17.25 | 28 | 34 | 41.5 |
| 25 | SUMMARY | 18.5 | 26 | 30.75 | 39 | 47 |
| | **MEDIANS** | 15 | 22.5 | 31 | 37 | 44.75 |
| | **MIDPOINTS (Cut Scores)** | | **19** | **27** | **34** | **41** |

The data in Table 16 represent 450 entries, three reviews of each of the 150 student work samples, and an average of 18 reviews per panelist. As expected, the rate of reviews varied from panelist to panelist, ranging from 10 to over 30. This range in count reflected differences in style and familiarity with the task. The two fastest reviewers had also read Advanced Placement

English exams for many years and were very familiar with the process of evaluating student writing. On the other hand, it was also necessary to remind these two panelists not to try to score the samples but simply to place them in a category.

MI staff had also previously calculated the raw score distributions based on performances of students who took the tests in March. They applied these cut scores to the raw score distributions to create impact data for the entire group of students who took the tests. The same process was also performed by race and gender. All impact data are included in Appendix C.

**Round 2**

On April 27, Dr. Bunch convened the Writing panel at 8:30 A.M. and distributed the Round 1 results, including the cut scores and distributions of ratings. Large-group discussion of Round 1 began with discussion of Table 16. Panelists discussed the reasons for placing individual bookmarks as they had and what those placements represented. There was then further general discussion of the nature of the just barely Proficient student as well as the just barely Basic student and the just barely Accelerated and barely Advanced student, as well as the demands of the tests as a whole and of individual items. Of particular interest were comments regarding the intersection of the cognitive demands of individual items and the achievements and capabilities of students in each of the categories.

Dr. Bunch then introduced the impact data in both tabular and graphic form (included in Appendix C) and led discussions of their implications. Results summarized in Table 17 reflect the final tabulations of scores with whole and half-point scores present.

**Table 17**
**Impact Data Summary for Writing Round 1**

| Category | Cut Score (Out of 48 Points) | Percent in Category |
|---|---|---|
| Below Basic | --- | 6.0 |
| Basic | 19 | 14.4 |
| Proficient | 27 | 30.9 |
| Accelerated | 34 | 41.9 |
| Advanced | 41 | 6.8 |

At this point, the discussion turned to the differential impact of the tests themselves on minority students. Some panelists expressed a desire to discuss test bias, and Dr. Bunch pointed out the role of the Fairness and Sensitivity Review Committee in making sure the tests were free of bias. The discussion continued for some time, with every panelist making at least one contribution. Once it was clear that all points of view had been heard, Dr. Bunch asked the group if they were ready to move to Round 2 and directed them to the appropriate portion of the Readiness Form:

| |
|---|
| I have discussed the results of Round 1, including my ratings, the ratings of others, and the impact data, and I understand what I need to do to complete Round 2.<br>(Circle one): **No** **Yes** |

Once again, panelists reviewed student work samples and entered their ratings on the rating forms. The group worked in this manner until around 11:30 A.M., at which time, Dr. Bunch reminded them to review their ratings with team members, finalize their ratings, and turn in their materials. After accounting for all materials, Dr. Bunch dismissed the group for lunch, and MI staff began to enter data for analysis. The results are shown in Table 18. The data in Table 18 are based on 370 reviews (about 2.5 reviews per packet, 15 reviews per panelist). There were fewer reviews in Round 2 than in Round 1, reflecting the fact that there was more discussion in small groups during Round 2.

**Table 18**
**Results of Writing Round 2**

| Panelist | Packet | Limited | Basic | Proficient | Accelerated | Advanced |
|---|---|---|---|---|---|---|
| 1 | SUMMARY | | 18 | 28.25 | 35 | |
| 2 | SUMMARY | 13.75 | 21 | 27.5 | 34 | 42 |
| 3 | SUMMARY | 14.75 | 20 | 31.5 | 40.25 | 44 |
| 4 | SUMMARY | 14 | 23 | 36 | 40 | 47 |
| 5 | SUMMARY | | 21 | 31.5 | | |
| 6 | SUMMARY | 17 | 24.5 | 32.5 | | 45 |
| 7 | SUMMARY | 14 | 16 | 29 | 39.25 | 45 |
| 8 | SUMMARY | 14.5 | 17 | 30.5 | 36 | 46.5 |
| 9 | SUMMARY | 14 | 27 | 31.5 | 38 | 43 |
| 10 | SUMMARY | 15 | 22 | 31 | 36 | 43.5 |
| 11 | SUMMARY | | 26 | 31.5 | 34 | 38.5 |
| 12 | SUMMARY | 14.5 | 23 | 32.5 | | 48 |
| 13 | SUMMARY | 11.5 | 19 | 30 | 37 | 39.5 |
| 14 | SUMMARY | | 20 | 30 | 34 | 47 |
| 15 | SUMMARY | | 15 | 30 | 40 | 48 |
| 16 | SUMMARY | | 21.5 | 30 | | 48 |
| 17 | SUMMARY | 17 | 25 | 30 | 43 | 48 |
| 18 | SUMMARY | 14.5 | 16 | 30.5 | 37.5 | 46 |
| 19 | SUMMARY | 14 | 19 | 31 | 38 | 46 |
| 20 | SUMMARY | | 22.5 | 29.5 | | 48 |
| 21 | SUMMARY | | 28 | 41 | 44 | 46 |
| 22 | SUMMARY | | 25.25 | 33 | 36 | 48 |
| 23 | SUMMARY | 15 | 21 | 33.25 | 39 | 48 |
| 24 | SUMMARY | | | 30 | | |
| 25 | SUMMARY | 18 | 25 | 33 | 42 | 47 |
| | **MEDIANS** | 14.5 | 21.25 | 31 | 38 | 46.25 |
| | **MIDPOINTS** | | 18 | 26 | 34.5 | 42 |

From Round 1 to Round 2, all cut scores moved, but not in the same direction. The cut score for Proficient fell by one point, while all other cuts moved up: one point each for Basic and Advanced, and half a point for Accelerated.

**Round 3**

Dr. Bunch initiated a discussion of the Round 2 results and brought back some of the points panelists had made during the discussion of Round 1. The discussion was less extensive than that after Round 1, and panelists indicated a readiness to move on to Round 3. Dr. Bunch then directed them to the appropriate portion of the Readiness Form:

> I have discussed the results of Round 2, including my ratings, the ratings of others, and the impact data, and I understand what I need to do to complete Round 3.
> (Circle one): **No** **Yes**

Dr. Bunch inspected the forms and noted that all panelists had answered "Yes." He then directed panelists to begin Round 3 following essentially the same directions they had followed in Rounds 1 and 2. Panelists continued to work in their small groups, reviewing and rating student work samples. In all, panelists entered 314 ratings (slightly over two ratings per packet and 12-13 ratings per panelist). At 3:30 P.M., Dr. Bunch directed panelists to begin to wrap up, review and finalize their ratings, and complete their Readiness Forms and evaluation forms. Panelists then completed the remaining portion of the Readiness Form and an evaluation form created specifically for this standard-setting activity. The final four statements on the Readiness Form were as follows:

> **Round 3:** I have completed my ratings, and I believe that the cut scores I have identified fairly represent minimal performances of students at the Basic, Proficient, Accelerated and Advanced levels
> (Circle one): **No** **Yes**
>
> Everyone was encouraged to share his or her ratings and hear those of other panelists**.**
> (Circle one): **No** **Yes**
>
> The cut scores we recommended accurately reflect the Basic, Proficient, Accelerated, and Advanced achievement levels.
> (Circle one): **No** **Yes**
>
> The process was fair and unbiased.
> (Circle one): **No** **Yes**

MI staff checked the Readiness Forms as panelists turned in their materials prior to being dismissed. There was not a single negative response on any of the Readiness Forms collected.

ODE staff thanked the members of each group for their three days of effort and willingness to be a part of the process by which performance standards on the Ohio Graduation Tests are set. Dr. William Batchelor, MI project director also thanked the groups for their participation and distributed expense checks. Dr. Bunch collected final ratings and all other secure materials. After each panelist's materials were accounted for, he or she was dismissed with a final word of thanks from Dr. Bunch.

## Final Results

Table 19 and Figure 12 summarize the results of Round 3.

**Table 19**
**Results of Writing Round 3**

| Rater | COE | Limited | Basic | Proficient | Accelerated | Advanced |
|---|---|---|---|---|---|---|
| 1 | SUMMARY | | 23.5 | 25 | | |
| 2 | SUMMARY | 17 | 22 | 27.5 | 34.75 | 44 |
| 3 | SUMMARY | 17 | 20 | 29 | 37 | 46 |
| 4 | SUMMARY | 14 | 23 | 34 | 43 | |
| 5 | SUMMARY | | 18.5 | 30.25 | 37.5 | 47 |
| 6 | SUMMARY | 19 | 27 | 31.5 | 35 | |
| 7 | SUMMARY | 15 | 20.5 | 32 | 41 | 43 |
| 8 | SUMMARY | 13 | 19 | 29.75 | 38 | 43.75 |
| 9 | SUMMARY | | 22.75 | 31 | 36 | |
| 10 | SUMMARY | 12.75 | 21 | 30.5 | 39 | 44 |
| 11 | SUMMARY | | 22 | 28 | 34 | 39.5 |
| 12 | SUMMARY | | | 29 | 38.5 | |
| 13 | SUMMARY | | 20.25 | 32.5 | | 41 |
| 14 | SUMMARY | | 17 | 28 | 33.5 | 35 |
| 15 | SUMMARY | | 16.25 | 30.5 | 40 | 46 |
| 16 | SUMMARY | | 20 | 32 | 40 | 48 |
| 17 | SUMMARY | | 23.5 | 33.5 | 41 | |
| 18 | SUMMARY | 13.25 | 25 | 34 | 36 | 45 |
| 19 | SUMMARY | 15 | 19 | 23.5 | 41 | 45 |
| 20 | SUMMARY | | | 31 | 35 | 40.5 |
| 21 | SUMMARY | 15 | 21.5 | 29 | 43 | 45 |
| 22 | SUMMARY | 17 | 18 | 27.25 | 38 | 42 |
| 23 | SUMMARY | 14.5 | 17 | 29 | 35.5 | 42 |
| 24 | SUMMARY | | 17 | 31 | 40 | 48 |
| 25 | SUMMARY | 17 | 25 | 34.5 | 42.25 | 47.5 |
| | **MEDIANS** | 15 | 20.5 | 30.5 | 38 | 44 |
| | **MIDPOINTS** | | 17.75 (18) | 25.5 | 34.25 (34) | 41 |

Once again, there was movement of all four cut scores, but this time, the direction was consistent. All four cut scores dropped by anywhere from one-fourth point (Basic, to 17.75, and Accelerated, to 34.25) to a whole point (Advanced, to 41). The cut score for Proficient fell from 26 to 25.5.

Table 20 summarizes the three rounds of ratings and associated raw cut scores over three rounds of ratings. The key numbers in Table 20 are on the final row and are in bold. While there was movement in the raw cut scores, the percentages of students who would score in the Proficient category or above rose by about three percent from Round 1 to Round 2 and another half a percent from Round 2 to Round 3.

**Table 20**

**Cut Scores by Round: Writing**
**(Percentages of students in each group are shown in parentheses.)**

| | **Round** | | |
|---|---|---|---|
| **Category** | 1 | 2 | 3 |
| Limited | --(6.0) | --(5.1) | --(5.1) |
| Basic | 19(14.4) | 18(12.7) | 18(12.2) |
| Proficient (Graduation) | 27(30.9) | 26(38.7) | 25.5(34.0) |
| Accelerated | 34(41.9) | 34.5(39.7) | 34(41.9) |
| Advanced | 41(6.8) | 42(3.8) | 41(6.8) |
| Percent of Students Proficient or Above | **79.6** | **82.2** | **82.7** |

**Impact**. Table 20 shows the final (Round 3) cut scores and impact (percentage of students in each category). Figure 11 summarizes the same information graphically.

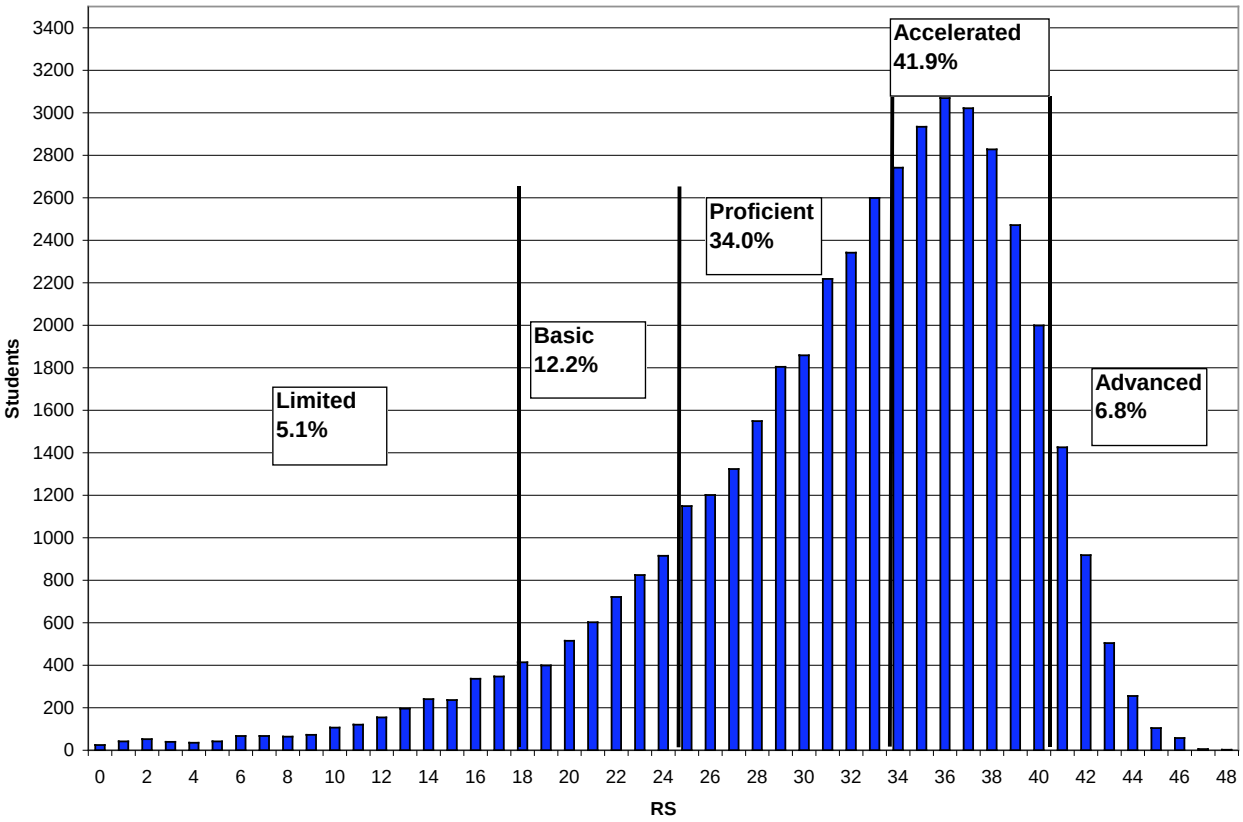



**Figure 12. Raw score distribution for Writing**

While Table 20 and figure 12 show the overall distributions of students by achievement level, they do not address possible differences among groups. Table 21 shows these same distributions by race and sex.

**Table 21**

**Distribution of Students by Race and Sex: Writing**
**(Entries are percentages.)**

| | Category | | | | | Proficient or Above |
|---|---|---|---|---|---|---|
| **Group (Tested)** | **Limited** | **Basic** | **Proficient** | **Accelerated** | **Advanced** | |
| AmInd (86) | 5.81 | 11.63 | 53.49 | 26.74 | 2.33 | 82.56 |
| As-PI (545) | 3.49 | 9.36 | 30.28 | 39.63 | 17.25 | 87.16 |
| BL-AA (6323) | 12.38 | 26.46 | 39.92 | 19.74 | 1.50 | 61.16 |
| Hisp (927) | 13.05 | 23.19 | 35.17 | 25.57 | 3.02 | 63.76 |
| Multi (581) | 5.51 | 11.36 | 38.90 | 37.18 | 7.06 | 83.14 |
| Other (126) | 11.90 | 14.29 | 29.37 | 39.68 | 4.76 | 73.81 |
| White (36,419) | 3.60 | 9.50 | 32.88 | 46.32 | 7.70 | 86.90 |
| Total by Race 45,007 | **5.08** | **12.20** | **33.99** | **41.91** | **6.82** | **82.72** |
| | | | | | | |
| Female (22,158) | 2.73 | 8.91 | 31.55 | 47.58 | 9.23 | 88.36 |
| Male (22,799) | 7.32 | 15.36 | 36.38 | 36.45 | 4.49 | 77.32 |
| Total by Sex 44,957 | **5.06** | **12.18** | **34.01** | **41.93** | **6.82** | **82.76** |

**Readiness and evaluations**. As noted above, there were no negative responses to any of the statements on the Readiness Form. Results of the evaluation were extremely positive, as shown in Table 22. Evaluations were nearly unanimously positive – 100 percent agreement with all statements except # 7, which had 95 percent agreement.

**Table 22**
**Summary of Evaluations: Writing**
**(Entries are percentages.)**

| | **Statement** | **Agree** | **Disagree** |
|---|---|---|---|
| 1 | The workshop leaders clearly explained the purpose of the meeting. | 100% | 0% |
| 2 | The workshop leaders clearly explained my task. | 100% | 0% |
| 3 | The examples and exercises helped me understand how to perform my task. | 100% | 0% |
| 4 | The large and small group discussions helped me understand the process. | 100% | 0% |
| 5 | I was able to follow the instructions and complete the rating sheets accurately. | 100% | 0% |
| 6 | The discussions after the first round of rating were helpful to me. | 100% | 0% |
| 7 | The discussions after the second round of rating were helpful to me | 95% | 5% |
| 8 | The information showing the distribution of student scores was helpful to me. | 100% | 0% |
| 9 | The facilities and food service helped to create a good working environment. | 100% | 0% |

## Conclusions

Standard setting is a combination of art and science. It combines the democratic process of group interaction and decision making with carefully planned and executed steps based on a well-defined mathematical model. The process by which the standard-setting activity for the tests of the OGT was carried out, described in detail in the body of this report, was meticulously crafted by experienced psychometricians and reviewed by a national body of experts in the field. The plan was carried out under the supervision of ODE staff and two external reviewers who are also experts in this field. Because there are no “true” cut scores for any test, the recommended cut scores are only as valid as the process by which they were derived.

Appendix A

# Standard Setting Plan

# Ohio Graduation Tests

# Standard Setting Plan: Spring 2005

**Michael B. Bunch**

**Measurement Incorporated**

In the spring of 2005, all five components of the Ohio Graduation Tests (OGT) will be administered: Reading, Mathematics, Science, Social Studies, and Writing. These tests will be administered with operational consequences; i.e., students will be required to meet established criteria on these tests (or others taken subsequently) in order to graduate from high school.

Staff of Measurement Incorporated will assist the Ohio Department of Education (ODE) in establishing the criteria to determine not only eligibility for high school graduation but to sort students into five performance categories for adequate yearly progress (AYP) reporting purposes. These categories are Limited, Basic, Proficient, Accelerated, and Advanced. In the spring of 2004, we conducted standard setting for Reading and Mathematics, forwarded recommended cut scores to ODE for presentation to the Board of Education. The Board approved all cut scores for these two tests. Thus, in the spring of 2005, it will be necessary to set cut scores for Science, Social Studies, and Writing.

We propose to use the same methods for setting standards in Science and Social Studies that we used for Reading and Mathematics; viz. the bookmark procedure (cf. Bunch, 2004). For the Writing test, we propose to use a holistic procedure that takes advantage of the principal characteristic of the tests; i.e., the fact that most of the items call for extended student responses. This plan addresses all three standard-setting activities in some detail.

## Science and Social Studies

The plan for Science and Social Studies will be essentially identical to one previously approved by the TAC and carried out in 2004. The persons involved, the data required, and the seven steps we will take to complete the task are described in detail in this Plan:

Public Engagement
Preparation
Training
Test Administration
Practice Test
Rating Test Items
Follow-Up and Reporting

Drs. Elliot Inman and Julie Miles will conduct the standard-setting meetings. Dr. Inman conducted the Mathematics standard-setting activity in 2004 with Dr. Miles' assistance. We propose to invite 25 panelists (Ohio educators, parents, and community leaders) per content area to standard-setting sessions to be held in May 2005.

We propose a three-day standard setting session for each test, employing a Rasch-based modified bookmark procedure (Mitzel, Lewis, Patz, & Green, 2001). Prior to the standard-setting meetings, MI will analyze data from the operational tests and calculate Rasch statistics for each item. In preparation for the standard setting meeting, we will rank order the items by IRT difficulty and prepare difficulty-ordered booklets as well as panelist training materials to include the following: overhead transparencies, guides, and forms.

## Public Engagement

We will share final versions of achievement level descriptors with the Test Steering Committee (TSC) and the Fairness and Sensitivity Review Committee (FSRC). We will work directly with ODE staff in the development of achievement level descriptors. The ODE has requested that we share the plans and results of the standard setting activities with the TSC and FSRC prior to the presentation of recommendations to the State Board of Education in June 2005. We will do so and incorporate their comments in our report to the Board.

### Preparation

We propose to use a Rasch-based modified bookmark procedure for standard setting as we did in 2004. This procedure features a difficulty-ordered test booklet which panelists use to indicate points at which students of a specified achievement level would no longer have a reasonable chance of answering items correctly. Panelists mark these points with "bookmarks" to indicate cut points between students of one level of achievement and those of the next higher level.

**Panelists**. While selection of method and facilitators is important to the success of any standard setting activity, the selection and orientation of standard-setting panelists is crucial. The ODE has taken responsibility for identifying, recruiting, and securing the services of panelists to whom we will refer as "panelists" to distinguish their advisory role from the policy-making role of the State Board of Education. MI will provide whatever assistance the ODE requests and will follow up on leads initiated by the ODE.

It is likely that the pool of panelists will include classroom teachers and building-level administrators as well as non-educators (e.g., parents, community leaders). We also anticipate that some teachers will cross subjects; i.e., some science teachers will rate Social Studies tests, and some social studies teachers will rate Science tests. We will maintain a list of panelists and their relevant characteristics (gender, race, educator/non-educator, subject taught) for subsequent analysis of results by group, but we will not directly associate a panelist's name with that panelist's ratings.

**Materials**. MI staff will prepare materials for the spring 2005 OGT standard setting. ODE staff will select panelists. Except for panelist selection, MI will be responsible for all aspects of standard setting. Materials will include the following:

Follow-up letters with instructions
PowerPoint presentation slides
PowerPoint presentation panelist handouts
PowerPoint presenter's notes
Practice tests and scoring guides
Achievement level descriptors
Difficulty-ordered test booklets and scoring guides
Bookmarks
Item data sheets
Impact data sheets
Data entry and processing programs and procedures
Data presentation spreadsheets and graphic displays of rating data
Readiness Form
Evaluation forms

**Achievement levels**. In 2004, ODE staff assumed primary responsibility for the development of the achievement levels and performance level descriptors (PLDs). It is our understanding that they will continue to take this responsibility.

The bookmark method**. We have proposed to use a modified bookmark method (Mitzel, Lewis, Patz, & Green, 2001). A brief overview of the method is provided here.**

The bookmark procedure is so named because panelists identify cut scores by entering markers in a specially designed test booklet. The test booklet consists of a set of items placed in difficulty order, easiest items first and hardest items last. In between, multiple choice (MC) and constructed response (CR) items are intermingled in order of their difficulty. Each CR item appears several times in the booklet, once for each of its score points. For a given CR entry, the item prompt and the rubric for a particular score point appear, along with sample responses illustrating that score point. The method has become quite popular because of its ability to present MC and CR items at the same time and because of its use of item response theory (IRT) analyses.

The difficulty-ordered booklet can be composed of any collection of items spanning the range of content, item types, and difficulty represented in a typical test, and need not consist only of items that have appeared in an intact test. This booklet can have more items or fewer items than a regular test booklet. For the spring 2005 standard-setting activity, we plan to use only the items that appear in the spring 2005 operational test.

With the bookmark procedure, IRT difficulty indices of the MC items and step values of the various score points for the CR items are ordered from least to most difficult. For the Ohio Graduation Test (OGT), we will use a one parameter IRT model, specifically the Rasch model. Each MC item will have one associated Rasch difficulty index, and each CR item will have as many Rasch step (difficulty) functions as it has score points (excluding zero).

Panelists will work in small groups, evaluating the contents of small clusters of items as they appear in the difficulty-ordered test booklet. They will discuss what makes one item or group of items more difficult than those that preceded it and ultimately place a bookmark at a point where they believe the difficulty of the subsequent items exceeds the ability of an identified group of students. Thus, for example, panelists would begin with the first item and ask themselves if a minimally qualified student (or group of students) at a particular achievement level (e.g., just barely *Proficient* or just barely *Advanced* would have a reasonable chance of answering the item correctly (cf. Mitzel, *et al.*, 2001, p. 260). They would then ask themselves the same question for each subsequent item until they reached one where they could not answer affirmatively. The final item yielding an affirmative response would mark the beginning of that performance level, and the panelists would place a bookmark at that point (i.e., after the last attainable item).

**Operationally defining the decision rule**. Crucial to the decision about who does and who does not have a reasonable chance of success is the definition of reasonable chance. Typically, bookmark and other item mapping procedures employ a 2/3 rule; i.e., panelists are asked to consider whether about two-thirds of the students who are just barely *Proficient* (or *Advanced* or *Basic)* would probably answer the item correctly or earn the score point in question. This operational definition of reasonable chance is central to the calculations necessary to determine cut scores. In 2004, the TAC advised ODE to adopt a decision rule that set the threshold at 50 percent; i.e., the point at which a just-barely *Proficient* (or *Advanced* , *Accelerated*, or *Basic*) student would have a 50-percent chance of answering correctly.

We begin with the fundamental equation of the Rasch model for dichotomous items (e.g., Wright & Stone, 1979; equation 1.4.1, p. 15):

$$P(X=1| \quad , \quad _i) = \exp( \quad - \quad _i)/[1 + \exp( \quad - \quad _i)] \qquad (1)$$

where is the Rasch ability (achievement level) of a student at the cut score (SACS), $_i$ is the Rasch difficulty of item *I*, and exp is the natural logarithm raised to the power inside the parentheses. By setting P in equation (1) to .50 and solving for , we find that = $_i$.

The OGT also contains several constructed-response (CR) items, all scored on a two- or four-point scale. For CR items, we employ the Partial-Credit Model (PCM; Wright & Masters, 1982). For CR items, the likelihood ($\pi_{nix}$) of a person with a given ability ($\beta_n$) obtaining any given score (*j*) in any item (*i*) is shown in equation 2, taken from Wright & Masters (1982), equation 3.1.6:

$$\pi_{nix} = \frac{\exp \sum_{j=0}^{x} (\beta_n - \delta_{ij})}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^{k} (\beta_n - \delta_{ij})} \qquad (2)$$

In Wright & Masters' formulation, the difficulties associated with each score point are referred to as step functions ($\delta_{ij}$). A key concept is that the step function for score point 0 is set equal to 0 for equation (2):

$$\delta_{i0} \equiv 0$$

such that $\sum_{j=0}^{0}(\beta_n - \delta_{ij}) = \mathbf{0}$, and $\exp \sum_{j=0}^{0}(\beta_n - \delta_{ij}) = \mathbf{1}$.

The numerator values for the other steps are derived as follows:

**Step 1**. $\sum_{j=0}^{1}(\beta_n - \delta_{ij}) = \sum_{j=0}^{0}(\beta_n - \delta_{i0}) + \beta_n - \delta_{i1} = 0 + \beta_n - \delta_{i1} = \boldsymbol{\beta_n - \delta_{i1}}$. (3)

**Step 2**. By similar logic: $\sum_{j=0}^{2}(\beta_n - \delta_{ij}) = \boldsymbol{2\beta_n - \delta_{i1} - \delta_{i2}}$. (4)

**Step 3**. By similar logic: $\sum_{j=0}^{3}(\beta_n - \delta_{ij}) = \boldsymbol{3\beta_n - \delta_{i1} - \delta_{i2} - \delta_{i3}}$. (5)

**Step 4**. By similar logic: $\sum_{j=0}^{4}(\beta_n - \delta_{ij}) = \boldsymbol{4\beta_n - \delta_{i1} - \delta_{i2} - \delta_{i3} - \delta_{i4}}$ . (6)

The exponential values of these sums are simply the natural logarithm *e* raised to the respective values:

**Step 1**. $\exp(\beta_n - \delta_{i1})$;

**Step 2.** $\exp(2\beta_n - \delta_{i1} - \delta_{i2})$;

**Step 3**. $\exp(3\beta_n - \delta_{i1} - \delta_{i2} - \delta_{i3})$;

**Step 4**. $\exp(4\beta_n - \delta_{i1} - \delta_{i2} - \delta_{i3} - \delta_{i4})$.

The denominator of equation (2) now becomes the simple sum of the four values shown above (for a 4-point item) plus 1 (the exponential value for step 0). The likelihood of obtaining any given score (0 through 4) can be calculated by dividing the numerator associated with that score point by this common denominator.

The derivation of the ability necessary to obtain a given score point is not so direct and neat as for multiple-choice (MC) items. Indeed, it is necessary to calculate a system of probabilities for each CR item (i.e., a probability for each score point) simultaneously. It should also be evident from the development above that the probabilities for each score point are interdependent; i.e., the likelihood of obtaining score point 4 must take into account the likelihood of obtaining score points 3, 2, 1, and 0. Rather than derive a general formula for obtaining these likelihoods, we create item characteristic curves and tables (somewhat similar to Reckase charts) and show the calculated probability of obtaining each score point for a wide range of abilities (e.g., -4 to +4 by increments of 0.1). We then look up the desired value. When

necessary, we interpolate to obtain exact values or enter additional ability estimates in the first column.

In calculating and plotting these probabilities, it becomes very clear that for some score points, there is no ability value associated with a .50 probability of obtaining that score. Indeed, only 0 and 4 (in the present case, or 0 and any perfect score in general) are likely to yield a .50 probability. Our response to this situation has been to consider not just the likelihood of obtaining a given score but of obtaining that score or better. Thus, for example, we concern ourselves not with the likelihood of obtaining a score of 1 or 2 but of obtaining a score of 1 or better and 2 or better. BIGSTEPS/WINSTEPS calculates the expected score point for P =.5 (the Thurstone Expected Score Measures) for score point 1 or better, 2 or better, etc.

**Preparing the difficulty-ordered booklet**. As noted above, MI will prepare a difficulty-ordered test booklet consisting of the items in the spring 2005 test booklets. Table 1 shows sample values for the 2003 OGT Reading test. The CR items are listed two or four times each, once for each non-zero score point. This table shows the Rasch difficulty indices (RASCH) for each item, along with item p values. For CR items, the step values for each score point and corresponding percentages of students scoring at each point or better are shown. For example, on page 10 (15-1), the Rasch Achievement value of .27 indicates that students with a Rasch achievement level ( $_n$ in the formulation above) of .27 would have a 2/3 chance of obtaining a score of 1 or better on item 15.

Translating bookmarks into cut scores is the heart of the procedure. Panelists sometimes have difficulty with the concept, particularly with the mathematics of it. We will use extreme caution and work through this concept carefully with both groups. We will use orientation materials similar to the ones we used successfully in the spring of 2004 with Reading and Mathematics.

## Training the Panelists

The standard setting session will consist of a full day of training, test administration, and scoring, followed by two days of identifying, debating, revising, and setting standards. A sample agenda is provided below.

Day 1

| | |
|---|---|
| 8:00 a.m. | Registration, Materials, Refreshments |
| 8:30 | Introductions; Collection of Security Forms |
| 8:45 | Background and Overview |
| 10:00 | Break |
| 10:15 | Test Administration |
| 12:30 | Lunch |
| 1:30 p.m. | Test Scoring and Discussion |
| 3:00 | Review of Achievement Levels |
| 4:00 | Adjourn |

Day 2

| | |
|---|---|
| 8:00 a.m. | Materials, Refreshments |
| 8:30 | Introduction to the Bookmark Procedure |
| 10:00 | Break |
| 10:15 | Practice Test |
| 11:00 | Questions & Answers |
| Noon | Lunch |
| 1:00 p.m. | Instructions for Round 1 |
| 1:15 | Round 1 |
| 3:45 | Wrap-up |
| 4:00 | Adjourn |

Day 3

| | |
|---|---|
| 8:00 | Materials, refreshments |
| 8:30 | Review of Round 1 |
| 9:45 | Round 2 |
| Noon | Lunch |
| 1:00 p.m. | Discussion of Round 2 |
| 1:30 | Round 3 |
| 3:00 | Final recommendations |
| 3:30 | Closure |
| 4:00 | Adjourn |

Drs. Inman and Miles will train panelists to complete the following tasks:

1. Study and answer all test items in specially ordered test booklets;

2. Identify the knowledge and skills required to answer each item correctly;

3. Determine why later items are more difficult than earlier items in the difficulty-ordered test booklet;

4. Consider whether most of the students just barely performing at the Basic, Proficient, and Advanced levels of achievement would answer the item correctly;

5. Place a mark for that level at each point where levels change;

In the course of the three days of training and rating activity, Drs. Inman and Miles will perform the following tasks:

1. Provide guidance and feedback throughout the three rounds of standard setting with regard to procedure only;

2. Collect all materials at the end of each round, analyze data, and prepare reports for the next round of standard setting;

3. Provide impact and other data and facilitate discussions between rounds;

4. Provide feedback at the end of the third round to let all panelists know what the final recommendations are likely to be.

**Orientation**. Inman will lead the Science group, while Dr. Miles will lead the Social Studies group. By the end of the orientation, panelists will understand the purpose of standard setting, their role in it, the meaning of the achievement level descriptors, the contents of the tests, and the specific procedures they will follow in setting standards.

Orientation basically consists of preparing panelists to answer the following three questions about each test item:

1. What knowledge or skill is required to earn this point?
2. What makes later items more difficult than earlier items?
3. Think of a large group of students at the cut score for this level. Would about 50 percent of them earn this point?

By providing examples and engaging panelists in a discussion of those examples, we will bring all panelists to a good working knowledge of their tasks. With specific regard to the third question above, Drs. Inman and Miles will take great care to explain the importance of the 50 percent rule. As noted previously, we will also make sure that the panelists are fully aware of their advisory role (hence the title "panelist" rather than "standard setter") and the fact that actual performance standards will be set by the State Board of Education, based on recommendations of the panelists.

**Discussion and feedback**. Drs. Inman and Miles will present information from a variety of sources (item difficulty, impact data, results of rating) but will not offer opinions or otherwise attempt to sway panelists. They will, however, use the discussion and feedback sessions to ascertain the level of task comprehension of the panelists. Should they detect confusion or lack of clarity, they will provide additional training to make sure everyone stays on task. All panelists will complete a Readiness Form prior to each round of rating, acknowledging that they have participated in training and that they understand the task they are about to perform.

## Test Administration

All panelists will take the standard form of the OGT in either Science or Social Studies. Drs. Inman and Miles will administer the tests and provide scoring guides (the multiple-choice key plus scoring guides for the constructed-response items). We note that the scoring of CR items is a very involved process requiring days of training for MI scorers. Panelists will not receive the full range of scorer training. Instead, we will present an abbreviated version of scorer training and allow panelists to score their own tests. After they have taken the tests and scored their own papers, panelists will discuss the overall difficulty and contents of the tests.

The scoring guides deserve some attention at this point because they may well influence panelists. As noted above, MI scorers receive extensive guides and days of training before they score student responses to CR items. We will not have several days to train panelists to be fully qualified scorers, nor would it be appropriate to attempt to train them to be scorers. The student responses they will eventually review will have already been scored by qualified scorers. Their own responses (i.e., the ones they will actually "score") will hardly be representative of high school student responses, for whom the scoring guides and sample papers were developed. Instead, our goal is to give each panelist a basic understanding of the requirements of a score point of 1, 2, 3, and so on. During training, we will provide two or three examples of each score point for each CR item (rather than the dozens of sample responses scorers review). Panelists will then determine which exemplars are most like their responses in order to score their own tests. Since we will not take high school diplomas away from teachers who score low on the OGT, the main purpose of the score will be to give panelists a solid understanding of the overall difficulty of the test.

## Practice Test

Our agenda shows a one-hour block of time the morning of the second day for a practice test. The purpose of the practice test will be to give the panelists some experience in setting cut scores before they begin the task with the operational test. The practice test will be a six-item difficulty-ordered booklet with four MC and two CR items. Together, they will span the difficulty of the test; therefore, they will be very distinct from one another. Per previous TAC recommendation, these six items will not come from the actual spring 2005 test, so that there will be no contamination of later ratings of the same items.

Panelists will be asked to review the six-item difficulty-ordered practice test booklet and place a bookmark to separate *Basic* from *Proficient*. We have selected this cut point because it is also the cut point for the graduation standard. We will give panelists an opportunity to work in small groups to review the six items and identify a point that would separate those just barely qualified to graduate from those not qualified to graduate. The pages of the practice test booklet will contain the same information as those of the regular difficulty-ordered test booklet.

After each panelist has placed his or her single bookmark, Drs. Inman and Miles will ask for a show of hands, tally the locations of the bookmarks, and conduct a brief discussion of their distribution. We will discuss the locations of the bookmarks in terms of their relation to the printed achievement level descriptors and dispersion. We will not share impact data at this point because the goal of the exercise will be to make sure panelists fully understand the procedure, not to divide students into categories.

### Rating Test Items

At the end of the discussion, we will ask if there are any questions about the task of rating items. We will then distribute the Readiness Form, a document we produced for the 2004 standard setting which requires all panelists to indicate that they understand the process before

they begin. Panelists will complete the appropriate section of this form and turn it in just before lunch on the second day. We will review the forms during lunch. After lunch, if any panelists have indicated that they are not ready to proceed, we will provide further training. If one or two panelists continue to struggle, we will work individually with them after distributing Round 1 materials and instructions to the rest of the panelists.

**Presenting the difficulty-ordered test booklet**. We will use the spring 2005 operational tests as the basis for the difficulty-ordered test booklets. In preparing the difficulty-ordered booklet, MI will include the page number, the original item number, and the Rasch achievement level associated with a .50 probability of answering a given MC item correctly or obtaining that score point or better for a CR item. Panelists will know at a glance how difficult the item was. They will not know, simply from the information on the page, what the cut score would be or how many students would score above or below the cut score.

Figure 1 shows the layout of a hypothetical difficulty-ordered booklet. The **bold numbers** at the top right of each page indicate the position in the difficulty-ordered booklet. The numbers at the top left indicate the positions in the original booklet. Some of these numbers have hyphens (e.g., 21-1). These numbers refer to the original item number (21, in this case) and the score point represented on that page (1, in this case). Each such item will appear in the booklet once for each of its score points.

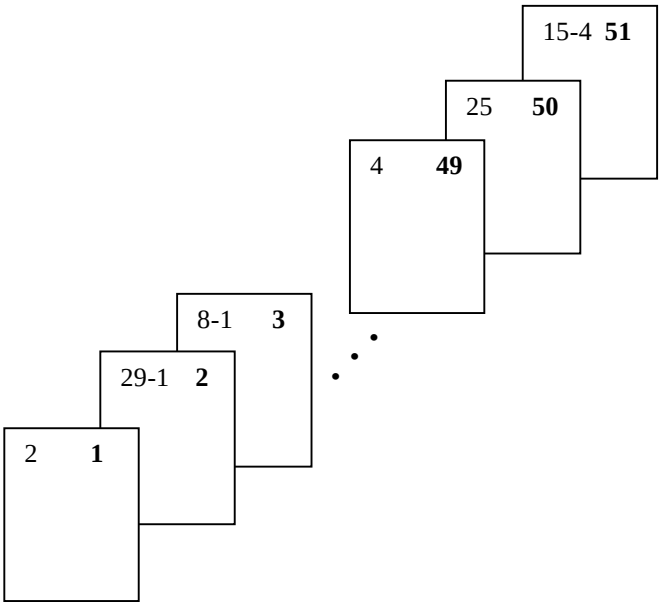



**Figure 1. Hypothetical page arrangement of a difficulty-ordered booklet**

Figure 2 shows an enlargement of a single item page that includes page number, original item number and score point, page number, and the Rasch achievement level required for a 2/3 chance to answer the item correctly. The key (A) is placed at the bottom of the page in a smaller font to serve as a quick check on the panelist's own response to the item without interfering with

the panelist's estimation of the difficulty of the item. In practice, since items associated with a given stimulus are likely to vary widely in difficulty and therefore be scattered throughout the test booklet, all common stimulus materials are placed in a companion booklet. The companion booklet is distributed to panelists along with the difficulty-ordered test booklet. In practice, panelists gain enough familiarity with the stimuli during the first day of test administration and scoring and the first round of item rating that they typically refer to the companion booklet very little during the last two rounds of item rating.

**1**

Item 2

Achievement level required for a 50% chance to answer correctly: **-1.363**

**Which of these best supports the idea that Mary** McLeod Bethune is concerned with helping young people find their way in the world?

A. the legacy she leaves in her will
B. her desire to return and help Essie
C. her zeal for her own place in history
D. the way she inspires Essie to believe

Key = A

**Figure 2. Sample page from a difficulty-ordered booklet**

The page shown in Figure 2 contains all the information a panelist would need to make a decision about the item. This item, on page 1 of the difficulty-ordered booklet, was item 2 in the actual test. The achievement level required for a student to have a 50 percent chance of answering this item correctly is –1.363 (the Rasch difficulty value of the item). All the identifying information is at the top of the page so that it will be easily accessible to panelists.

The page number will be bold and of a larger size to make it clearly distinguishable from the other numbers. We will ask panelists to use the page number as their indicator for a bookmark.

**Round 1**. Each panelist will receive a difficulty-ordered test booklet, stimulus booklet and a set of bookmarks. The difficulty-ordered booklet, constructed in accordance with Table 1, will have one item per page, starting with the easiest item in the test booklet. At the top of each page will be printed the original item number (top left), page number (top right), item Rasch difficulty index (top center), and Rasch achievement level associated with 50 percent chance of answering correctly for MC items or 50 percent chance of obtaining that raw score or higher for CR items. Each CR item will be represented once for each of its score points, as noted above. We will print the item and one sample response per score point. Because there will be several different ways to earn each score point, we will select sample responses to cover the full range of possibilities across the various CR items.

For both tests, we will prepare a separate stimulus booklet. In a difficulty-ordered test booklet, items for a given scenario, map, or other stimulus will be scattered throughout the booklet. Panelists sometimes have some difficulty with this feature of the procedure, and we try to help them in any way we can. We will create a meaningful code for each stimulus and then repeat that code at the beginning of each associated item. Thus, if a panelist needs to refer to a particular stimulus, that panelist would have no difficulty in doing so.

The bookmarks will be printed on one side of a piece of card stock. Each bookmark will be similar to the one shown in Figure 3. In Rounds 1 and 2, panelists will enter the page number for each bookmark. At Round 3, panelists will be familiar with the relationship between page number and cut score. We will ask them to enter page number and associated cut score, as well as the impact data in order to make sure each panelist is fully aware of his or her recommendation. In introducing the bookmark, we will emphasize the fact that the ODE has designated *Proficient* as the graduation standard. Therefore, we will direct panelists to focus on this standard first and then turn their attention to *Basic, Accelerated,* and *Advanced.*

During Round 1, panelists will work in small groups of 3-5 individuals. While they will discuss the item contents among themselves, each panelist will complete his or her own bookmark card. As they complete Round 1, panelists will review their forms to make sure they are complete, return all materials to Dr. Dr. Inman or Dr. Miles, and be dismissed for the day.

**OGT Standard Setting**
**Science**

**Panelist Number _____**

**Bookmarks (Enter Page Number for Rounds 1 and 2.)**

| Round | Basic | Proficient/ Graduation | Accelerated | Advanced |
|---|---|---|---|---|
| 1 | | | | |
| 2 | | | | |

**Round 3**

| | Basic | Proficient/ Graduation | Accelerated | Advanced |
|---|---|---|---|---|
| Page Number | | | | |
| Cut Score | | | | |
| % At or Above | | | | |

**Notes**

Figure 3. Bookmark

**Data analysis and presentation**. At the end of each round, MI staff will collect the bookmark cards and enter the values into a spreadsheet similar to the one shown in Table 2 and a chart similar to the one shown in Figure 4. After tallying the results, we will return the cards to the panelists, along with the results. Table 1 allows panelists to see where their bookmarks fall, relative to those of other panelists. It also gives them a sense of where the group average lies, as well as how far their own bookmarks fall from the group average. For our purposes here, Figure 4 provides a more graphic summary of the bookmarks in the region of the Proficient/Accelerated division, allowing each panelist to see his or her marks, relative to those of other panelists.

**Table 1**
**Sample Output for One Round of Standard Setting**

| Results of Round 1 of Rating | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **Basic** | | **Proficient/Graduation** | | **Accelerated** | | **Advanced** | |
| Panelist | Page | Ach. | Page | Ach. | Page | Ach. | Page | Ach. |
| 1 | 11 | 0.273 | 27 | 0.900 | 36 | 1.200 | 48 | 2.040 |
| 2 | 9 | 0.193 | 23 | 0.616 | 31 | 0.998 | 40 | 1.489 |
| 3 | 13 | 0.420 | 25 | 0.810 | 30 | 0.988 | 49 | 2.080 |
| 4 | 9 | 0.193 | 25 | 0.810 | 40 | 1.489 | 47 | 1.650 |
| 5 | 12 | 0.286 | 26 | 0.891 | 38 | 1.333 | 46 | 1.627 |
| 6 | 7 | -0.176 | 20 | 0.569 | 33 | 1.090 | 44 | 1.589 |
| 7 | 16 | 0.493 | 23 | 0.616 | 41 | 1.510 | 45 | 1.590 |
| 8 | 8 | 0.082 | 27 | 0.900 | 40 | 1.489 | 45 | 1.590 |
| 9 | 10 | 0.270 | 27 | 0.900 | 38 | 1.333 | 45 | 1.590 |
| 10 | 12 | 0.286 | 24 | 0.740 | 38 | 1.333 | 48 | 2.040 |
| 11 | 11 | 0.272 | 21 | 0.579 | 38 | 1.333 | 46 | 1.627 |
| 12 | 11 | 0.272 | 22 | 0.600 | 36 | 1.200 | 44 | 1.589 |
| 13 | 10 | 0.270 | 16 | 0.493 | 35 | 1.191 | 45 | 1.590 |
| 14 | 17 | 0.540 | 23 | 0.616 | 33 | 1.090 | 46 | 1.627 |
| 15 | 14 | 0.440 | 26 | 0.891 | 39 | 1.340 | 48 | 2.040 |
| 16 | 12 | 0.286 | 32 | 1.046 | 39 | 1.340 | 49 | 2.080 |
| 17 | 10 | 0.270 | 29 | 0.945 | 38 | 1.333 | 47 | 1.650 |
| 18 | 9 | 0.193 | 26 | 0.891 | 33 | 1.090 | 43 | 1.586 |
| 19 | 11 | 0.272 | 22 | 0.600 | 39 | 1.340 | 44 | 1.589 |
| 20 | 11 | 0.272 | 25 | 0.810 | 36 | 1.200 | 45 | 1.590 |
| 21 | 12 | 0.286 | 24 | 0.740 | 37 | 1.290 | 50 | 2.120 |
| 22 | 13 | 0.420 | 28 | 0.910 | 35 | 1.191 | 48 | 2.040 |
| 23 | 11 | 0.272 | 23 | 0.616 | 35 | 1.191 | 48 | 2.040 |
| 24 | 10 | 0.270 | 25 | 0.810 | 33 | 1.090 | 47 | 1.650 |
| 25 | 9 | 0.193 | 30 | 0.988 | 38 | 1.333 | 46 | 1.627 |
| **Mean** | | **0.274** | | **0.771** | | **1.253** | | **1.749** |
| SD | | 0.138 | | 0.156 | | 0.145 | | 0.220 |
| M-1SD | | 0.136 | | 0.616 | | 1.108 | | 1.529 |
| M+1SD | | 0.412 | | 0.927 | | 1.397 | | 1.970 |
| **Mean Cut** | | **25.0** | | **32.5** | | **38.5** | | **42.5** |
| Low Cut | | 20.0 | | 28.0 | | 35.0 | | 41.0 |
| Cut-1SD | | 23.5 | | 30.0 | | 36.0 | | 41.0 |
| Cut+1SD | | 27.0 | | 34.0 | | 39.0 | | 42.5 |
| High Cut | | 29.0 | | 35.0 | | 41.0 | | 46.0 |

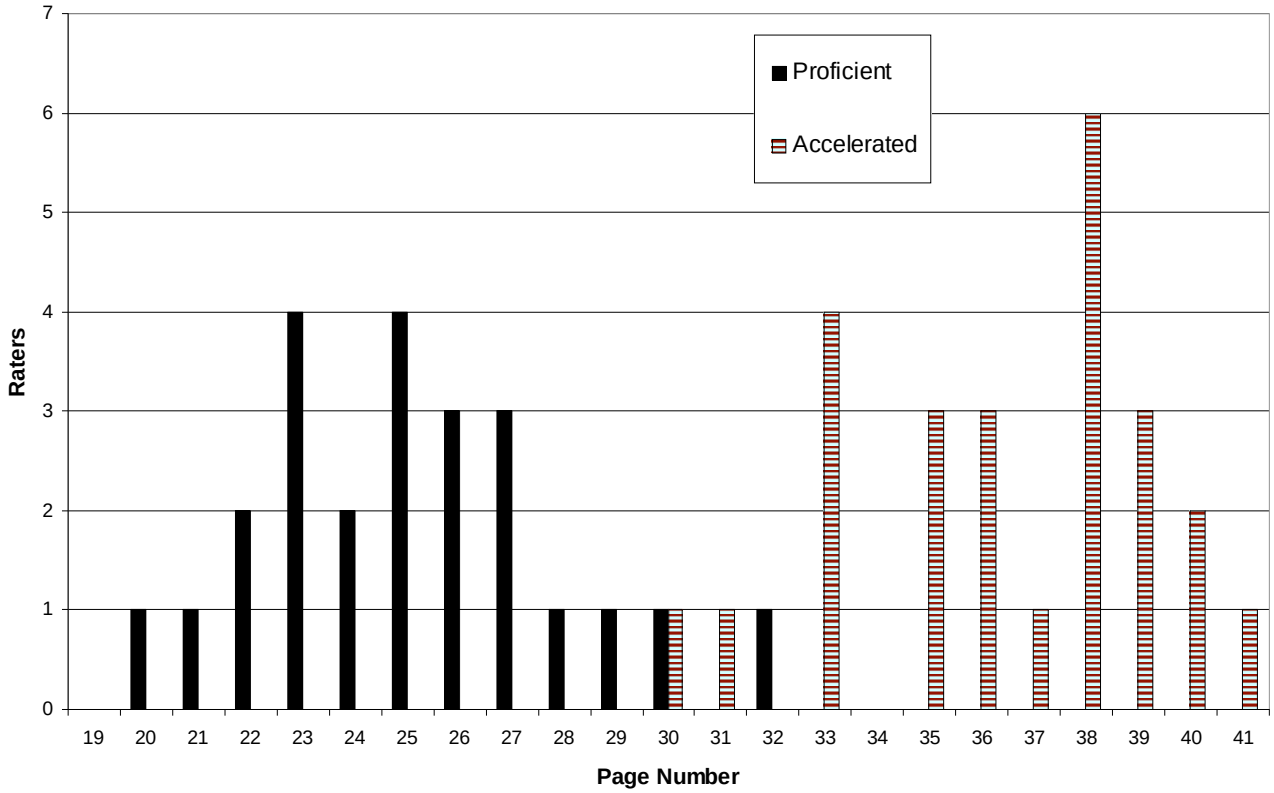



**Figure 4. Data presentation for one round of standard setting showing the Proficient/Accelerated division**

Table 1 provides a summary of bookmark placements, as well as the resulting cut scores. Note that the mean cut score is shown (along with its standard deviation). Also shown are the low cut, high cut, and cuts one standard deviation above and below the mean. Individual cut scores are not shown. The panelists already know those values, having noted them on the pages where they placed their bookmarks. We use Table 1 to draw attention to student achievement levels and the associated mean cut scores. It is also noteworthy that the mean cut score based on the mean Rasch achievement level is not necessarily equal to the arithmetic mean of the individual cut scores. Rather than have to explain why they are not (or even worse, respond to someone's conclusion that we miscalculated the mean cut score), we choose to report bookmarks and associated abilities at the individual panelist level and save cut score for the summary.

Figure 4 offers a different perspective. Here, only the bookmarked page numbers are shown. Panelists get a graphic view of how their bookmarks compare to the bookmarks of the other panelists. Figure 4 also shows where there are gaps; i.e., page ranges no one chose as bookmarks for any cut score. In subsequent rounds, these pages will typically not enter the discussion. This figure also shows where there are overlaps. For example, one panelist chose page 32 as the bookmark for *Proficient*, while another panelist chose 30 and another 31 for *Accelerated.* In effect, one panelist would set the cutoff for *Proficient* higher than at least one panelist would set the cutoff for Accelerated. Visualizations such as Figure 4 are excellent conversation starters in Rounds 2 and 3.

*Impact data*. MI will show impact data to panelists between Rounds 1 and 2. The purpose of the data will be to allow panelists to see how many students would be classified at

each achievement level if the mean cut scores from Round 1 were implemented. Given the timing of the standard-setting activity, we will present impact data for a representative sample of students (N 40,000 – 50,000) who take the test as well as for subgroups based on gender and ethnicity. We plan to present these data in both tabular and graphic form, as we did in 2004. All the impact data for will be in hand prior to the standard-setting session. Calculating numbers and percentages of students in each performance category will be a fairly simple task that can be performed quickly on site.

**Round 2**. On the morning of Day 3, panelists will receive their difficulty-ordered test booklets and other materials from Round 1 plus the data from Round 1 and the impact data. Drs. Inman and Miles will lead a discussion of the Round 1 ratings and impact.

Discussion will focus on range of cut scores, areas of particular disagreement, and concerns about placement of individual items. It is sometimes helpful for panelists to work through the difference between their perceived difficulty of a particular item and the placement of that item relative to others in the test. Once panelists have discussed the results of Round 1 as a total group, we will reassign them to small groups of 3-5 members each, and begin Round 2. The task for Round 2 will be identical to that of Round 1. The primary difference will be the amount of information available to each panelist. At the end of the round, Drs. Inman and Miles will collect all materials and dismiss for lunch.

During lunch, Drs. Inman and Miles will once again tally the cut scores and prepare reports similar to those shown in Table 1 and Figure 4. They will present these to panelists at the beginning of Round 3.

**Round 3**. As they return from lunch, panelists will receive all their Round 2 materials plus a summary of Round 2. We will use a special form of Table 1 which includes actual cut scores associated with the values of Rasch achievement level (see Table 4). Once again, panelists will be able to examine the impact data as part of their discussion. Drs. Inman and Miles will lead a discussion of the impact data and other topics of concern from Round 2. At the end of this discussion, panelists will have a final opportunity to evaluate all their previous ratings and all information at hand and simply enter four bookmarks and the associated cut scores. We will also ask panelists to enter the **% At or Above** each cut score they enter. We have added this value as a check on the accuracy of the cut score entries. Drs. Inman and Miles will check each completed bookmark for accuracy, tally these final ratings and calculate the mean recommended cut score for each achievement level. At the discretion of the ODE, we can report these means to the panelists or not.

## Follow-Up and Reporting

MI staff will summarize the processes and outcomes of the three-day session and present the plans and results to ODE, the TAC, and the TSC and FSRC. It is likely that one or more of the groups will have comments about the results and may even want to alter the proposed cut scores. To preserve the integrity of the overall process, we propose to append the

recommendations of these groups to the original set of recommendations and forward all to the State Board of Education.

Drs. Inman and Miles will work with ODE staff to prepare a report for the Ohio Board of Education, complete with recommendations regarding cut scores. The report will provide a complete description of the process as well as the cut scores recommended by the panelists during the three-day activity as well as any additional recommendations from TAC, TSC, or FSRC. The report will also contain an executive summary for nontechnical audiences. Either Dr. Inman or Dr. Miles will be available for a face-to-face meeting with the Board, if ODE so chooses.

## Writing

The Writing component of the OGT consists of two writing prompts, one short-answer item, and 10 multiple-choice (MC) items. Each MC item counts for one point. The short-answer item counts for four points, and each writing prompt counts for 18 points, for a total of 50 points. Since the bulk of the points on the test (40 out of 50) come from either short or extended student responses, a standard-setting procedure that employs this rich data source is preferable to one that does not. Holistic methods are ideally suited for this purpose. MI used this method in setting standards for the OGT alternate assessments in May 2004 with great success, and we propose to use it in 2005 for the Writing component of the OGT.

Holistic standard setting techniques, (e.g., Kingston, Kahl, Sweeney, & Bay, 2001; Plake & Hambleton, 2001; Jaeger & Mills, 2001), focus on whole collections of student work. Our approach will contain elements of the Analytic Judgment Method (Plake & Hambleton, 2001) and the Body of Work procedure (Kingston, et al., 2001). It will be essentially identical to the procedure we used in May 2004 with the OGT alternate assessment. Our procedure includes the following steps:

Advance preparation
Training
Round 1 and Data Analysis
Round 2 and Data Analysis
Final Cut-Score Setting and Analysis
Follow-Up and Reporting

### Advance Preparation

Prior to standard setting, the client and the contractor create preliminary definitions of the performance levels (Performance Level Descriptors or PLDs). These definitions describe, somewhat generically, what students at each performance level are expected to know or be able to do. Having defined the levels, the client and contractor also work together to select samples of student work representing the full range of scores. Finally, panelists are identified and invited.

**Selection of appropriate committee members.** ODE staff will identify and select committee members just as they did for the April 2004 Reading and Mathematics standard setting and the May 2004 alternate assessment standard setting. ODE will assemble a group of panelists that reflects the diversity of the state and include the stakeholders in the outcome of this assessment: educators, community leaders, teachers, and parents. As some of the work involved in standard setting will require work in small groups, efforts will be made to assure that those small groups include both kinds of panelists, those already familiar with the standards and process and those who are not. As with Science and Social Studies, we recommend 25 panelists.

**Selection of student work or other assessment materials to review.** In the case of a bookmark standard setting approach, the student work presented is usually the actual test booklet items and, for constructed-response items, representative student responses earning that score point. In the case of a holistic standard setting, panelists will review actual student work. The selection of samples of student work is a crucial element in the success of the procedure.

It is expected that approximately 150,000 students will take the Writing test in the spring of 2005. At this time, it is uncertain whether all tests will be scored before standard setting takes place, so the final distribution of scores may not be known. The range of possible scores will be from 0 to 50. We propose to select with a set of 40 work samples to represent the full range of scores. Furthermore, MI will select six additional samples to use in a practice session. These will be chosen to represent the entire spread of the possible point distribution.

**Schedule**. Prior to the standard-setting activity, MI staff will prepare a detailed schedule for the entire event. We propose a three-day session, just as for Science and Social Studies. A preliminary schedule is shown below. We welcome the opportunity to discuss this schedule with ODE and the TAC.

Day 1

| | |
|---|---|
| 8:00 a.m. | Registration, Materials, Refreshments |
| 8:30 | Introductions; Collection of Security Forms |
| 8:45 | Background and Overview |
| 10:00 | Break |
| 10:15 | Test Administration |
| 12:30 | Lunch |
| 1:30 p.m. | Test Scoring and Discussion |
| 3:00 | Review of Achievement Levels |
| 4:00 | Adjourn |

Day 2

| | |
|---|---|
| 8:00 a.m. | Materials, Refreshments |
| 8:30 | Introduction to the Holistic Procedure |
| 10:00 | Break |
| 10:15 | Practice Session |
| 11:00 | Questions & Answers |
| Noon | Lunch |
| 1:00 p.m. | Instructions for Round 1 |

| | |
|---|---|
| 1:15 | Round 1 |
| 3:45 | Wrap-up |
| 4:00 | Adjourn |

Day 3

| | |
|---|---|
| 8:00 | Materials, refreshments |
| 8:30 | Review of Round 1 |
| 9:45 | Round 2 |
| Noon | Lunch |
| 1:00 p.m. | Discussion of Round 2 |
| 1:30 | Final Cut Score Setting |
| 3:30 | Closure |
| 4:00 | Adjourn |

Training

ODE and MI staff will provide an overview of the OGT and the standard-setting activity. Panelists will then have an opportunity to study the PLDs, take the same Writing test that students took, and score their tests. Dr. Bunch will then explain the procedure and provide a short practice exercise to determine the extent to which panelists understand the process. Only after all panelists complete a Readiness Form indicating that they understand the process enough to begin to apply it will they begin the first round of reviewing student work.

**Definition of task for committee members.** For most people who serve on a standard-setting committee, the process is a once-in-a-lifetime experience. Without an appropriate introduction to both the assessment and the task at hand, it is easy for committee members to misconstrue their role. MI staff, working with ODE, will prepare a background document to acquaint panelists with the history of the OGT and the task to be performed. This background document will be integrated into a Powerpoint presentation that will be used during the orientation session for each group. MI staff will prepare drafts of these documents for Department review and approval.

Several critical points will be reiterated throughout the process, including:

The committee is to suggest cut scores to the State Board of Education, but the committee does not make the cut scores law;

The committee is to make a thoughtful judgment about student performance in a standards-based context taking into consideration possible pass rates and retention, but pass rates are not to dictate their judgments about level of achievement;

The committee is to make judgments about the extent to which a student has reached a critical level in mastery of standards, and

Although the group will have the actual rubric used for scoring the tests, the goal is not for panelists to score each student work sample, but to assign a level of performance.

MI staff will cooperate with ODE staff to train the panelists. ODE staff will develop a set of narrative descriptions of the cognitive skills and knowledge that define those levels of achievement. Standard setting panelists will be trained to use these materials. Dr. Bunch will provide a thorough introduction to the Body of Work procedure and provide a practice exercise in which panelists apply the procedure to a small group of student work samples.

**Practice session**. Using the six student work samples selected earlier, Dr. Bunch will lead a practice session in which panelists will evaluate all six work samples and tally their ratings of each sample (from Limited to Advanced). Dr. Bunch will lead a discussion of the similarities and differences in the ratings, noting in particular the features of work samples that were most likely to lead to disagreement and the apparent differences in interpretations of the five PLDs. Panelists will have ample time to discuss their differences in interpretation and points of view they bring to the activity. At the end of the practice session, Dr. Bunch will administer the first segment of the Readiness Form. Only those panelists affirming understanding of the task and readiness to proceed will be allowed to do so.

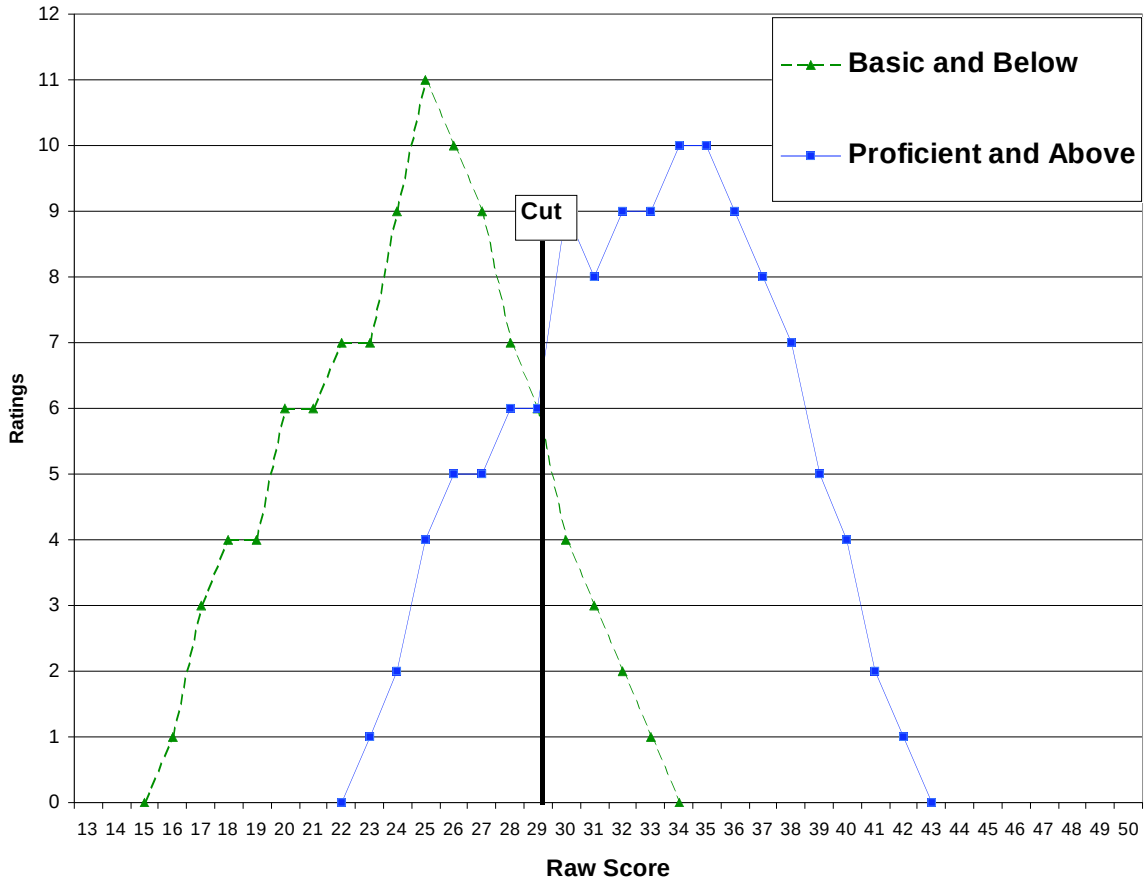
Round 1 and Analysis

During the initial round, panelists will examine a set of carefully selected work samples and assign each sample to one of two categories, using rating forms designed specifically for this task. The two categories will be Basic-and-Below and Proficient-and-Above. We found in May that an initial round in which panelists focus on this simplified task helps them focus on the crucial Basic/Proficient cut score that determines who will earn a diploma. Each panelist will initially make these assignments independently of other panelists and without reference to any scoring information. Later stages of the process can also include discussion of the individual ratings and opportunity for each panelist to change his or her ratings after seeing the ratings of other panelists summarized visually. Each panelist will review as many work samples as possible in approximately 2.5 hours. MI staff will make sure that all 40 work samples receive approximately the same numbers of reviews.

**Data analysis**. Between Rangefinding and Pinpointing (otherwise known as Rounds 1 and 2), MI staff will analyze all panelist responses. Using scores assigned by MI readers, we will calculate means and medians for each performance category, based on panelists' ratings of individual work samples. Kingston, *et al.* (2001) have described a procedure for determining the probability of a given score falling within or above a given performance category using logistic regression. In their model, the cut score ($x$) is a function of the slope and intercept of the regression line describing the relationship between the raw score distribution and classification into a given performance category (or higher). Plake & Hambleton (2001) and Jaeger & Mills (2001) use the same information in a much more direct manner. They simply average the raw scores at the boundaries of each category. We propose to apply the Plake & Hambleton approach to the Body of Work procedure, just as we did for the May 2004 alternate assessment standard setting. In that activity, we calculated the median for each category and then computed the midpoint between adjacent category medians.

Alternatively, we could use category means. A third possibility is to approximate the logistic regression approach by noting the score point at which adjacent score distributions overlap, as shown in Figure 5. The line labeled Cut indicates the cut score based on distribution overlap. In this particular case, the cut score is 29. Given the same data, the cut score based on the midpoints between means would be 28.65. Using the midpoints between adjacent category medians, the cut score would be 29. For reasonably normal distributions, all three approaches yield approximately the same results. We would invite the TAC to make a recommendation as to which approach it would prefer.

**Figure 5. Possible Cut Scores Based on Alternative Approaches**




Round 2

Dr. Bunch will initiate a discussion of the Round 1 results, encouraging panelists to share their opinions with one another but always focusing on the match between PLD and student performance within the context of the content of the test. Panelists will be able to see their own category means or medians, group means and medians, and preliminary cut scores. After panelists have discussed results at this level, Dr. Bunch will introduce impact data; i.e., the total distribution of scores for all students tested in the spring of 2005 (or all tests scored as of the time of standard setting). Panelists will be able to see that their ratings have an impact on the numbers or percentages of students who have qualified to graduate (i.e., those who would be classified Proficient or higher).

After discussion of all Round 1 results and the impact data, Dr. Bunch will administer the second segment of the Readiness Form. Upon affirming understanding of the task and readiness to proceed to Round 2, panelists will be allowed to do so.

As in Round 1, Panelists will evaluate student work samples, this time assigning each sample to one of the five categories: Limited, Basic, Proficient, Accelerated, and Advanced. Panelists will work in small groups and be able to discuss their ratings prior to marking their rating sheets. At the end of Round 2, panelists will turn in all materials for MI staff to analyze.

**Data analysis**. The same procedures that were used to calculate a single cut score for Round 1 will be used to calculate four cut scores for Round 2. Other analyses, such as the distributions of panelists' ratings, will be identical to Round 1.

### Final Cut-Score Setting and Analysis

Dr. Bunch will initiate a discussion of Round 2 results, once again showing the impact of their ratings, this time, however, showing how many students would be in each category, not just the graduate/not-graduate categories. Panelists will have a final opportunity to discuss their points of view about various features or whole work samples and hear the points of view of others. At the end of the discussion, Dr. Bunch will administer the final segment of the Readiness Form and give instructions for the final round which will establish the cut scores to forward to the State Board of Education.

In the final round, panelists will once again work in small groups, assigning work samples to one of the five performance categories. While they will cooperate with others in the review of work samples, each panelist will be free to enter his or her own final judgment on his or her rating sheet. At the end of the round, MI staff will collect all materials and dismiss the panelists.

**Data analysis**. Analysis of the final round of data will be identical to that for Round 2. At the end of the analysis, MI will forward the results to ODE in the form of cut scores, score ranges, and percentages of students in each category.

## Follow-Up and Reporting

MI staff will summarize the processes and outcomes of the three-day session and present the plans and results to ODE, the TAC, and the TSC and FSRC. It is likely that one or more of the groups will have comments about the results and may even want to alter the proposed cut scores. To preserve the integrity of the overall process, we propose to append the recommendations of these groups to the original set of recommendations and forward all to the State Board of Education.

Dr. Bunch will work with ODE staff to prepare a report for the Ohio Board of Education, complete with recommendations regarding cut scores. The report will provide a complete

description of the process as well as the cut scores recommended by the panelists during the three-day activity as well as any additional recommendations from TAC, TSC, or FSRC. The report will also contain an executive summary for nontechnical audiences. Dr. Bunch will be available for a face-to-face meeting with the Board, if ODE so chooses.

**References**


Jaeger, R. M. & Mills, C. N. (2001). An integrated judgment procedure for setting standards on complex, large-scale assessments. In G. J. Cizek (Ed.*), Setting Performance Standards: Concepts, Methods, and Perspectives.* Mahwah, NJ: Lawrence J. Earlbaum Associates.

Kingston, N. M. Kahl, S. R., Sweeney, K. P., & Bay, L (2001). Setting performance standards using the body of work method. In Cizek, G. J. Setting Performance Standards: Concepts, Methods, and Perspectives. Mahwah, NJ: Lawrence Erlbaum Associates.

Mitzel, H. C., Lewis, D. M., Patz, R. J, & Green, D. R. (2001). The bookmark procedure: Psychological perspectives. Cizek, G. J. Setting Performance Standards: Concepts, Methods, and Perspectives. Mahwah, NJ: Lawrence Erlbaum Associates.

Plake, B. S. & Hambleton, R, K. (2001). The analytic judgment method for setting standards on complex performance assessments. Cizek, G. J. Setting Performance Standards: Concepts, Methods, and Perspectives. Mahwah, NJ: Lawrence Erlbaum Associates.

Wright, B. D. & Stone, M. H. (1979). *Best Test Design*. Chicago: MESA Press.

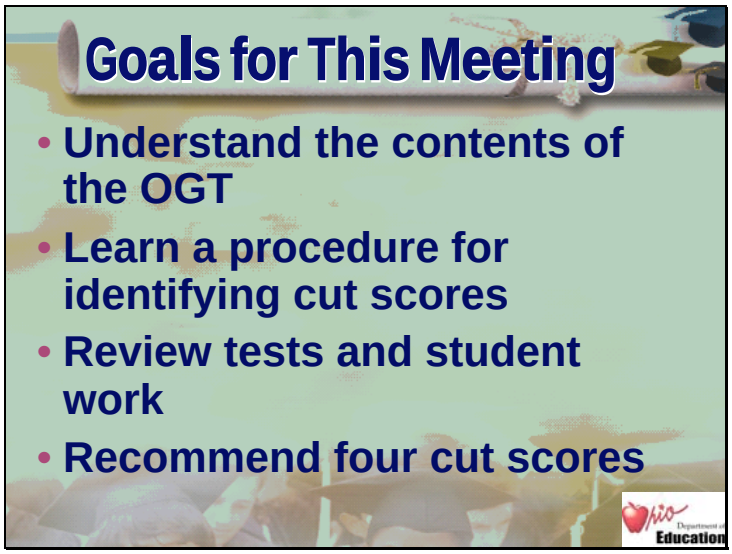
Wright, B. D. & Masters, G. N. (1982). *Rating Scale Analysis*. Chicago: Mesa Press.

# Appendix B

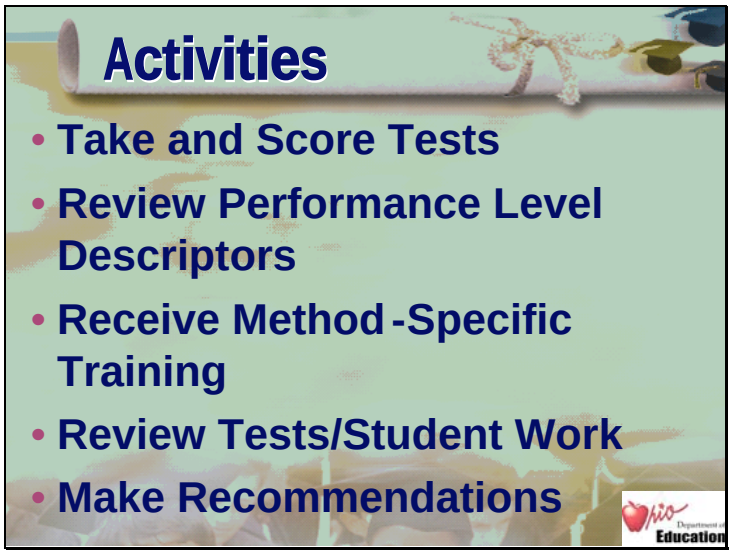
Overview of Standard Setting Process
Final Performance Level Descriptors
Readiness and Evaluation Forms
Bookmark Process
Holistic Rating Form

Slide 1

Slide 2



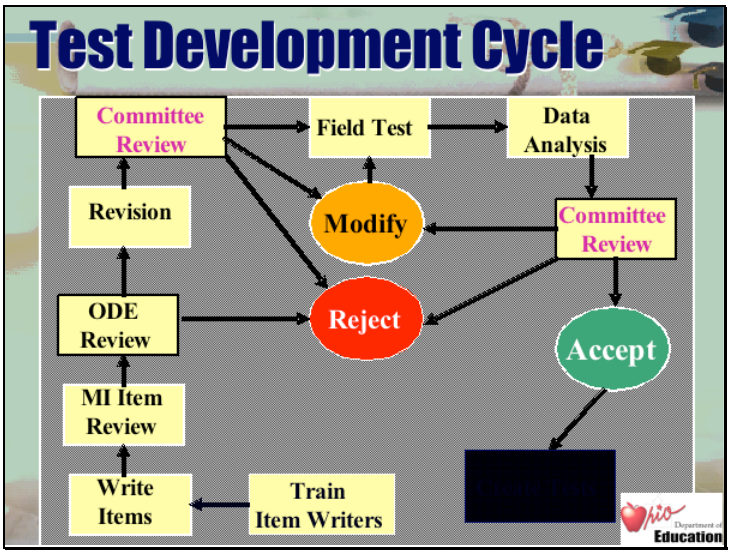
Slide 3



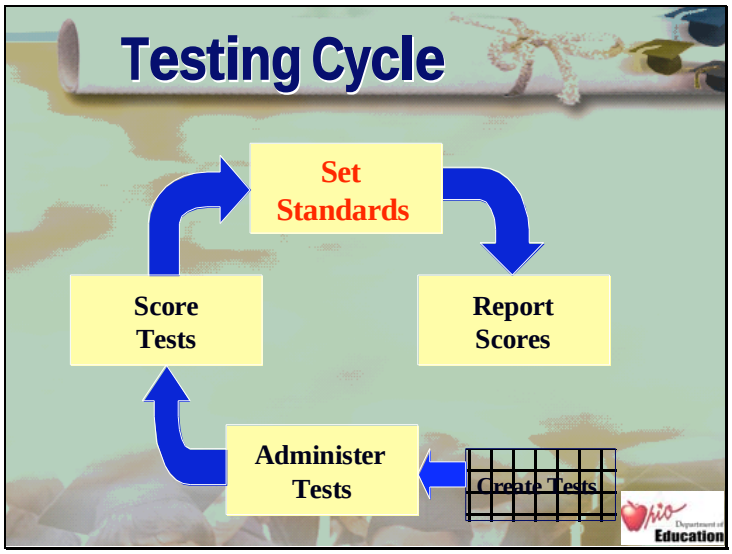

Slide 4

Slide 5

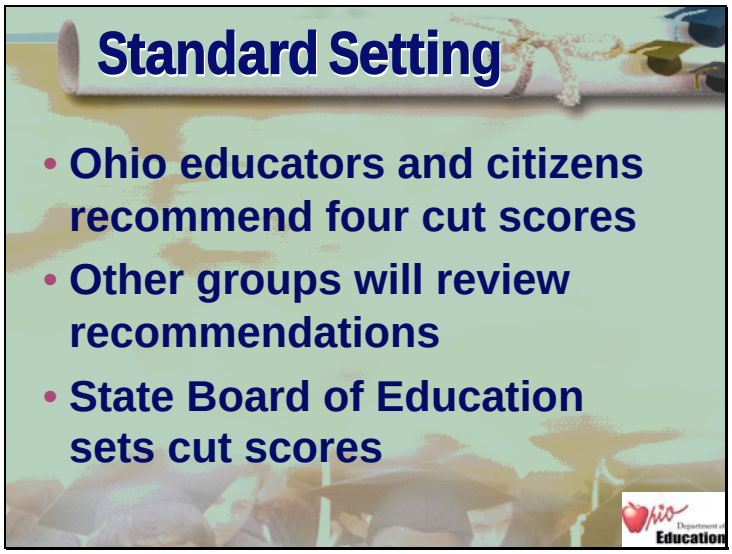
Slide 6



Slide 7

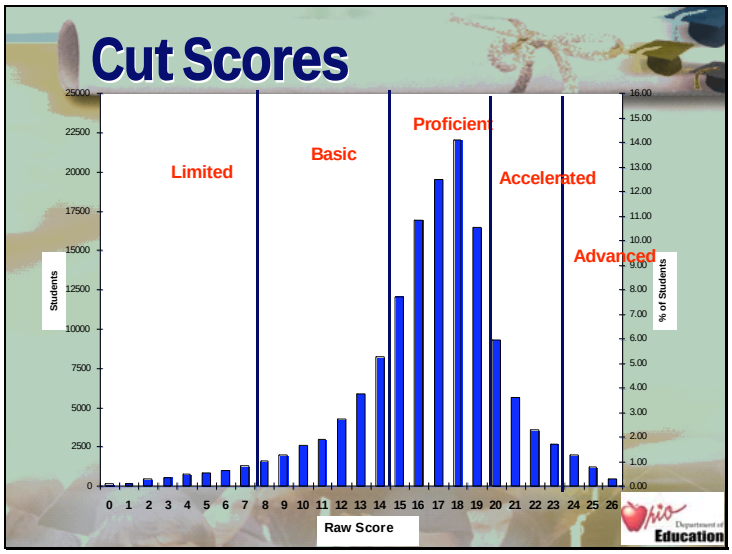



Slide 8

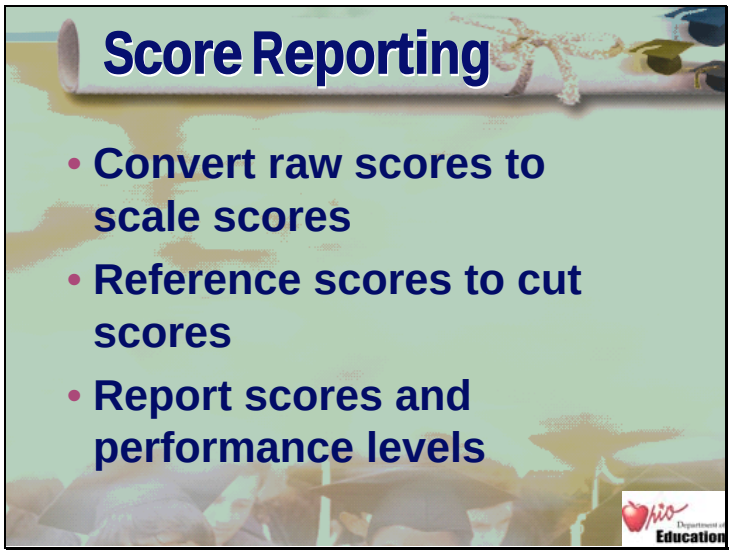


Slide 9

Slide 10



Slide 11




Slide 12

# Ohio Graduation Test
# Performance Level Descriptors – Science
# January 2005

| Limited | **Students demonstrate skills and understanding below the performance required to reach the Basic level.** |
|---|---|
| **Basic** | Students inconsistently identify scientific facts and terms and show a rudimentary understanding of valid scientific concepts, processes and relationships underlying natural phenomena in life, physical, and Earth and space sciences. Given investigative scenarios, they demonstrate an elementary understanding of scientific investigative processes, recognize some laboratory equipment, and outline simple procedures. Given sufficiently rich contexts, they classify based on definitions. They understand basic models and identify some parts of living, physical, and Earth and space systems. They demonstrate some familiarity with technological applications. |
| **Proficient** | Students typically recognize and provide descriptions or explanations showing understanding of scientific concepts and relationships underlying natural phenomena, structures, processes in living, physical and Earth and space systems and cycles (e.g., food webs, electric circuits, water cycle). Given investigative scenarios, they demonstrate a working ability to design scientific investigations. They organize, represent and analyze data in various forms, and detect and summarize data trends. They use information to provide explanations and to draw reasonable conclusions. They demonstrate understanding of physical and conceptual models. They recognize some inputs and outputs, causes and effects, and interactions and relationships within a system. They recognize factors impacting rate of change (e.g., effects of forces on motion). They recognize the practical application of scientific concepts and principles to problems in the real world and show a developmental understanding of technological applications |
| **Accelerated** | Students typically demonstrate solid knowledge and reasoning abilities in the sciences. They design, revise and critique scientific investigations, combining scientific knowledge with information from experience or observation. They use science equations, symbols and chemical formulas to find solutions. They compare and recognize some inherent strengths and limitations of various models. They interpolate, extrapolate or make valid inferences from given information and/or understanding of scientific concepts to describe, explain or draw appropriate conclusions about interactions and relationships within a system. They provide specific, relevant examples to illustrate practical application of scientific concepts and principles to problems in the real world. They design technological solutions for given problems. |
| **Advanced** | Students consistently demonstrate superior knowledge and the ability to integrate understanding of scientific principles. Students use complex reasoning skills to predict and to design investigations that answer questions about real-world situations. They integrate, interpolate, and extrapolate embedded information to draw well-formulated explanations and conclusions. They describe the inherent strengths and limitations of models and revise models based on new information. They recognize relationships within systems and use this knowledge to make reasonable predictions. They describe and explain constant, exponential, or irregular patterns and apply this recognition to make predictions. They evaluate technological solutions for given problems. |

# Ohio Graduation Test
# Performance Level Descriptors – Social Studies
# January 2005

| | |
|---|---|
| **Limited** | Students demonstrate skills and understandings below the performance required to reach the Basic level. |
| **Basic** | Students inconsistently demonstrate the ability to explain issues of social studies content. Their explanations of historical sequence may be incomplete. If prompted, they can view issues from a limited number of social and geographic perspectives. They are able to identify some instances when the government has had a role in economic activities and how applications of the U.S. Constitution have changed. They recognize that rights and responsibilities have to be balanced in a democratic society. They can read source materials and suggest how they would be related to a task. |
| **Proficient** | Students typically demonstrate the ability to explain issues of social studies content. They have a sense of historical sequence and understand that events in history do not exist independently of each other. These students understand that issues can be examined from different social and geographic perspectives. They can explain the roles of government in economic activities and how applications of the U.S. Constitution have changed over time. They can cite examples of balancing rights and responsibilities. They can paraphrase source material and apply it to a task. |
| **Accelerated** | Students typically demonstrate the ability to analyze issues across most areas of social studies content. They can apply concepts of chronology and causation. They draw from different social and geographic perspectives to examine issues. These students can provide detailed explanations about the role of government in economic activities and how applications of the U.S. Constitution have changed over time. They can analyze examples of balancing rights and responsibilities. They are able to organize source material and apply it to a task. |
| **Advanced** | Students consistently demonstrate the ability to analyze and evaluate issues across the entire spectrum of social studies content. Their analysis of causation is generally thorough. They accurately critique information from different social and geographic perspectives. They can distinguish among the roles of government in economic activities and analyze how applications of the U.S. Constitution have changed over time. These students are able to make judgments about balancing rights and responsibilities. They can evaluate the usefulness of source material and its applicability to a task. |

# Ohio Graduation Test
# Performance Level Descriptors – Writing
# January 2005

| | |
|---|---|
| **Limited** | Students performing at the Limited level demonstrate skills and understanding below the performance required to reach the Basic level. |
| **Basic** | Students performing at the Basic level demonstrate a marginal understanding of the writing process and a marginal grasp of the purpose of writing and style. They demonstrate some skills at organizing, revising and editing writing. The students write with some focus and engage a reader through a few developed, unified and coherent ideas. The students use some sentence variety and make effective word choices inconsistently. They also have a marginal understanding of grammar, capitalization, punctuation and spelling conventions. |
| **Proficient** | Students performing at the Proficient level demonstrate an adequate to effective understanding of the writing process and an adequate to effective grasp of the purpose of writing and writing style. They demonstrate developed skills at organizing, revising and editing writing. The students write with a reasonably well developed focus and engage a reader through reasonably well developed, unified and coherent ideas. The students use sentence variety and make effective word choices with some consistency. They also, with some consistency, understand grammar, capitalization, punctuation and spelling conventions. |
| **Accelerated** | Students performing at the Accelerated level demonstrate an excellent understanding of the writing process and an excellent grasp of the purpose of writing and writing style. They demonstrate well developed skills at organizing, revising and editing writing. The students write with a well developed focus and engage a reader through well developed, unified and coherent ideas. The students use sentence variety and make effective word choices with consistency. They also consistently understand grammar, capitalization, punctuation and spelling conventions. |
| **Advanced** | Students performing at the Advanced level demonstrate a superior understanding of the writing process and a superior grasp of the purpose of writing and writing style. They demonstrate exceptional skills at organizing, revising and editing writing. The students write with an exceptional focus and engage a reader through exceptionally well developed, unified and coherent ideas. The students use sentence variety and make effective word choices with a high degree of consistency. They also understand grammar, capitalization, punctuation and spelling conventions at the same high degree of consistency. |

# Ohio Graduation Tests

## Readiness Form

**Rater Number_____**

**Practice Test:** I have completed the practice test, and I understand what I need to do to complete Round 1.
(Circle one): **Yes** **No**

**Round 1:** I have discussed the results of Round 1, including my ratings, the ratings of others, and the impact data, and I understand what I need to do to complete Round 2
(Circle one): **No** **Yes**

**Round 2:** I have discussed the results of Round 2, including my ratings, the ratings of others, and the impact data, and I understand what I need to do to complete Round 3
(Circle one): **No** **Yes**

**Round 3:** I have completed my ratings, and I believe that the cut scores I have identified fairly represent minimal performances of students at the Basic, Proficient, and Advanced levels
(Circle one): **No** **Yes**

Everyone was encouraged to share his or her ratings and hear those of other raters**.**
(Circle one): **No** **Yes**

The cut scores we recommended accurately reflect the basic, proficient, and advanced achievement levels.
(Circle one): **No** **Yes**

The process was fair and unbiased.
(Circle one): **No** **Yes**

# Ohio Graduation Tests

## Standard Setting Workshop Evaluation

Please respond to the statements below by indicating your agreement or disagreement. Check one box for each statement to indicate whether you strongly agree, agree, disagree, or strongly disagree with each statement. Room for comments has been provided at the bottom of the form.

## Subject: ___ Science ___ Social Studies ___ Writing

| | Statement | **Agree** | **Disagree** |
|---|---|---|---|
| 1 | The workshop leaders clearly explained the purpose of the meeting. | | |
| 2 | The workshop leaders clearly explained my task. | | |
| 3 | The examples and exercises helped me understand how to perform my task. | | |
| 4 | The large and small group discussions helped me understand the process. | | |
| 5 | I was able to follow the instructions and complete the rating sheets accurately. | | |
| 6 | The discussions after the first round of rating were helpful to me. | | |
| 7 | The discussions after the second round of rating were helpful to me | | |
| 8 | The information showing the distribution of student scores was helpful to me. | | |
| 9 | The facilities and food service helped to create a good working environment. | | |

Comments

**OGT Standard Setting**
**Science**

**Panelist Number _____**

**Bookmarks (Enter Page Number for Rounds 1 and 2.)**

| Round | Basic | Proficient/ Graduation | Accelerated | Advanced |
|---|---|---|---|---|
| 1 | | | | |
| 2 | | | | |

**Round 3**

| | Basic | Proficient/ Graduation | Accelerated | Advanced |
|---|---|---|---|---|
| Page Number | | | | |
| Cut Score | | | | |
| % At or Above | | | | |

**Notes**

**OGT Standard Setting**
**Social Studies**

**Panelist Number _____**

**Bookmarks (Enter Page Number for Rounds 1 and 2.)**

| Round | Basic | Proficient/ Graduation | Accelerated | Advanced |
|---|---|---|---|---|
| 1 | | | | |
| 2 | | | | |

**Round 3**

| | Basic | Proficient/ Graduation | Accelerated | Advanced |
|---|---|---|---|---|
| Page Number | | | | |
| Cut Score | | | | |
| % At or Above | | | | |

**Notes**

# Ohio Graduation Tests

**Writing**
**Standard Setting Rating Form**

| Packet | Litho | Level |
|---|---|---|
| 1 | 101885 | |
| 2 | 277570 | |
| 3 | 101220 | |
| 4 | 282099 | |
| 5 | 278075 | |
| 6 | 281393 | |
| 7 | 281389 | |
| 8 | 278928 | |

1 = Limited
2 = Basic
3 = Proficient
4 = Accelerated
5 = Advanced

# Appendix C

Impact Data for Round 1 and Round 2
Graphs of Impact Data

**Round 1 Science**

*SCIENCE*
*IMPACT ROUND 1*

*The FREQ Procedure*

| GROUP | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| 1 Limit | 3204 | 7.12 | 3204 | 7.12 |
| 2 Basic | 7873 | 17.49 | 11077 | 24.60 |
| 3 Prof | 13494 | 29.97 | 24571 | 54.58 |
| 4 Accel | 12477 | 27.71 | 37048 | 82.29 |
| 5 Advan | 7973 | 17.71 | 45021 | 100.00 |

*SCIENCE*

*The FREQ Procedure*

*Table of eth by GROUP*

*eth GROUP*

| Row Pct | 1 Limit | 2 Basic | 3 Prof | 4 Accel | 5 Advan | Total |
|---|---|---|---|---|---|---|
| AmInd | 11.63 | 20.93 | 31.40 | 24.42 | 11.63 | |
| As-PI | 4.22 | 14.86 | 22.94 | 26.97 | 31.01 | |
| BL-AA | 22.79 | 39.22 | 25.92 | 9.54 | 2.53 | |
| Hisp | 16.29 | 31.93 | 29.02 | 15.10 | 7.66 | |
| Multi | 7.06 | 22.20 | 31.84 | 24.61 | 14.29 | |
| Other | 21.43 | 17.14 | 29.29 | 18.57 | 13.57 | |
| White | 4.14 | 13.30 | 30.78 | 31.29 | 20.49 | |
| Total | 3204 | 7873 | 13494 | 12477 | 7973 | 45021 |

*SCIENCE*

*The FREQ Procedure*

*Table of Gender by GROUP*

*Gender GROUP*

| Row Pct | 1 Limit | 2 Basic | 3 Prof | 4 Accel | 5 Advan | Total |
|---|---|---|---|---|---|---|
| F | 6.76 | 19.46 | 31.38 | 26.82 | 15.58 | |
| M | 7.41 | 15.56 | 28.61 | 28.62 | 19.80 | |
| Total | 3189 | 7857 | 13475 | 12473 | 7972 | 44966 |

*Frequency Missing = 55*

***Round 2 Science***

*SCIENCE*
*IMPACT ROUND 2*

*The FREQ Procedure*

| GROUP | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| 1 Limit | 3847 | 8.54 | 3847 | 8.54 |
| 2 Basic | 8402 | 18.66 | 12249 | 27.21 |
| 3 Prof | 14279 | 31.72 | 26528 | 58.92 |
| 4 Accel | 11309 | 25.12 | 37837 | 84.04 |
| 5 Advan | 7184 | 15.96 | 45021 | 100.00 |

*SCIENCE*

*The FREQ Procedure*

*Table of eth by GROUP*

*eth GROUP*

*Row Pct ,1 Limit ,2 Basic ,3 Prof ,4 Accel ,5 Advan , Total*

| | | | | | | |
|---|---|---|---|---|---|---|
| Amlnd | 12.79 | 22.09 | 34.88 | 20.93 | 9.30 | |
| As-PI | 5.69 | 16.15 | 24.22 | 24.59 | 29.36 | |
| BL-AA | 26.92 | 38.42 | 24.77 | 7.73 | 2.17 | |
| Hisp | 19.85 | 31.39 | 28.91 | 13.38 | 6.47 | |
| Multi | 8.26 | 23.92 | 33.91 | 21.00 | 12.91 | |
| Other | 22.86 | 17.86 | 29.29 | 17.86 | 12.14 | |
| White | 5.05 | 14.86 | 33.07 | 28.55 | 18.47 | |
| Total | 3847 | 8402 | 14279 | 11309 | 7184 | 45021 |

*SCIENCE*

*The FREQ Procedure*

*Table of Gender by GROUP*

*Gender GROUP*

*Row Pct ,1 Limit ,2 Basic ,3 Prof ,4 Accel ,5 Advan , Total*

| | | | | | | |
|---|---|---|---|---|---|---|
| F | 8.24 | 20.85 | 32.64 | 24.36 | 13.92 | |
| M | 8.79 | 16.53 | 30.82 | 25.90 | 17.96 | |
| Total | 3829 | 8387 | 14262 | 11305 | 7183 | 44966 |

*Frequency Missing = 55*

***Round 1 Social Studies*** *SOCIAL STUDIES*

*IMPACT ROUND 1*

*The FREQ Procedure*

| GROUP | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| 1 Limit | 3970 | 8.82 | 3970 | 8.82 |
| 2 Basic | 5685 | 12.63 | 9655 | 21.46 |
| 3 Prof | 10926 | 24.28 | 20581 | 45.74 |
| 4 Accel | 10419 | 23.15 | 31000 | 68.89 |
| 5 Advan | 14000 | 31.11 | 45000 | 100.00 |

*SOCIAL STUDIES*

*The FREQ Procedure*

*Table of eth by GROUP*

*eth GROUP*

*Row Pct ,1 Limit ,2 Basic ,3 Prof ,4 Accel ,5 Advan , Total*

| | | | | | | |
|---|---|---|---|---|---|---|
| Amlnd | 8.22 | 17.81 | 27.40 | 24.66 | 21.92 | |
| As-PI | 6.79 | 8.26 | 16.70 | 21.47 | 46.79 | |
| BL-AA | 24.93 | 25.53 | 28.07 | 14.23 | 7.23 | |
| Hisp | 21.21 | 23.63 | 24.62 | 15.93 | 14.62 | |
| Multi | 9.48 | 13.98 | 28.62 | 19.13 | 28.79 | |
| Other | 26.17 | 12.15 | 20.56 | 15.89 | 25.23 | |
| White | 5.68 | 10.15 | 23.66 | 25.00 | 35.52 | |
| Total | 3970 | 5685 | 10926 | 10419 | 14000 | 45000 |

*SOCIAL STUDIES*

*The FREQ Procedure*

*Table of Gender by GROUP*

*Gender GROUP*

*Row Pct ,1 Limit ,2 Basic ,3 Prof ,4 Accel ,5 Advan , Total*

| | | | | | | |
|---|---|---|---|---|---|---|
| F | 8.16 | 13.97 | 27.05 | 23.04 | 27.78 | |
| M | 9.40 | 11.31 | 21.59 | 23.29 | 34.40 | |
| Total | 3951 | 5673 | 10913 | 10415 | 13998 | 44950 |

*Frequency Missing = 50*

***Round 2 Social Studies*** *SOCIAL STUDIES*

*IMPACT ROUND 2*

*The FREQ Procedure*

| GROUP | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| 1 Limit | 4217 | 9.37 | 4217 | 9.37 |
| 2 Basic | 5009 | 11.13 | 9226 | 20.50 |
| 3 Prof | 12117 | 26.93 | 21343 | 47.43 |
| 4 Accel | 11420 | 25.38 | 32763 | 72.81 |
| 5 Advan | 12237 | 27.19 | 45000 | 100.00 |

*SOCIAL STUDIES*

*The FREQ Procedure*

*Table of eth by GROUP*

*eth GROUP*

*Row Pct ,1 Limit ,2 Basic ,3 Prof ,4 Accel ,5 Advan , Total*

| | | | | | | |
|---|---|---|---|---|---|---|
| Amlnd | 8.22 | 15.07 | 30.14 | 27.40 | 19.18 | |
| As-PI | 7.34 | 7.16 | 18.53 | 24.77 | 42.20 | |
| BL-AA | 26.32 | 22.58 | 31.24 | 13.90 | 5.96 | |
| Hisp | 22.53 | 20.99 | 27.25 | 16.70 | 12.53 | |
| Multi | 9.98 | 11.65 | 32.28 | 22.63 | 23.46 | |
| Other | 27.10 | 10.28 | 21.50 | 19.63 | 21.50 | |
| White | 6.06 | 8.93 | 26.21 | 27.66 | 31.13 | |
| Total | 4217 | 5009 | 12117 | 11420 | 12237 | 45000 |

*SOCIAL STUDIES*

*The FREQ Procedure*

*Table of Gender by GROUP*

*Gender GROUP*

*Row Pct ,1 Limit ,2 Basic ,3 Prof ,4 Accel ,5 Advan , Total*

| | | | | | | |
|---|---|---|---|---|---|---|
| F | 8.77 | 12.29 | 30.02 | 24.68 | 24.24 | |
| M | 9.89 | 9.98 | 23.92 | 26.09 | 30.11 | |
| Total | 4198 | 4997 | 12104 | 11416 | 12235 | 44950 |

*Frequency Missing = 50*

***Round 1 Writing*** *WRITING*

*The FREQ Procedure*

| GROUP | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| 1 Limit | 2704 | 6.01 | 2704 | 6.01 |
| 2 Basic | 6462 | 14.36 | 9166 | 20.37 |
| 3 Prof | 13912 | 30.91 | 23078 | 51.28 |
| 4 Accel | 18860 | 41.90 | 41938 | 93.18 |
| 5 Advan | 3069 | 6.82 | 45007 | 100.00 |

*WRITING*

*The FREQ Procedure*

*Table of eth by GROUP*

```
eth     GROUP

Row Pct ,1 Limit ,2 Basic ,3 Prof  ,4 Accel ,5 Advan ,  Total
ƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
Amlnd   ,  6.98 , 15.12 , 48.84 , 26.74 ,  2.33 ,
ƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
As-PI   ,  4.22 , 11.19 , 27.71 , 39.63 , 17.25 ,
ƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
BL-AA   , 14.74 , 29.23 , 34.79 , 19.74 ,  1.50 ,
ƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
Hisp    , 14.56 , 26.43 , 30.42 , 25.57 ,  3.02 ,
ƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
Multi   ,  6.20 , 14.11 , 35.46 , 37.18 ,  7.06 ,
ƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
Other   , 11.90 , 15.87 , 27.78 , 39.68 ,  4.76 ,
ƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
White   ,  4.28 , 11.51 , 30.19 , 46.32 ,  7.70 ,
ƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
Total      2704    6462   13912   18860    3069   45007
```

*WRITING*

*The FREQ Procedure*

*Table of Gender by GROUP*

```
Gender    GROUP

Row Pct ,1 Limit ,2 Basic ,3 Prof  ,4 Accel ,5 Advan ,  Total
ƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
F       ,  3.29 , 10.93 , 28.97 , 47.58 ,  9.23 ,
ƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
M       ,  8.60 , 17.66 , 32.80 , 36.45 ,  4.49 ,
ƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
Total      2690    6448   13899   18852    3068   44957
```

*Frequency Missing = 50*

***Round 2 Writing***

*WRITING*
*IMPACT ROUND 2*

*The FREQ Procedure*

| GROUP | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| 1 Limit | 2287 | 5.08 | 2287 | 5.08 |
| 2 Basic | 5705 | 12.68 | 7992 | 17.76 |
| 3 Prof | 17422 | 38.71 | 25414 | 56.47 |
| 4 Accel | 17871 | 39.71 | 43285 | 96.17 |
| 5 Advan | 1722 | 3.83 | 45007 | 100.00 |

*WRITING*

*The FREQ Procedure*

*Table of eth by GROUP*

*eth GROUP*

*Row Pct ,1 Limit ,2 Basic ,3 Prof ,4 Accel ,5 Advan , Total*

| | | | | | | |
|---|---|---|---|---|---|---|
| Amlnd | 5.81 | 12.79 | 54.65 | 25.58 | 1.16 | |
| As-PI | 3.49 | 10.09 | 31.93 | 43.67 | 10.83 | |
| BL-AA | 12.38 | 27.33 | 42.84 | 16.73 | 0.71 | |
| Hisp | 13.05 | 23.73 | 39.37 | 21.79 | 2.05 | |
| Multi | 5.51 | 11.88 | 44.41 | 34.42 | 3.79 | |
| Other | 11.90 | 14.29 | 35.71 | 35.71 | 2.38 | |
| White | 3.60 | 9.90 | 37.96 | 44.22 | 4.32 | |
| Total | 2287 | 5705 | 17422 | 17871 | 1722 | 45007 |

*WRITING*

*The FREQ Procedure*

*Table of Gender by GROUP*

*Gender GROUP*

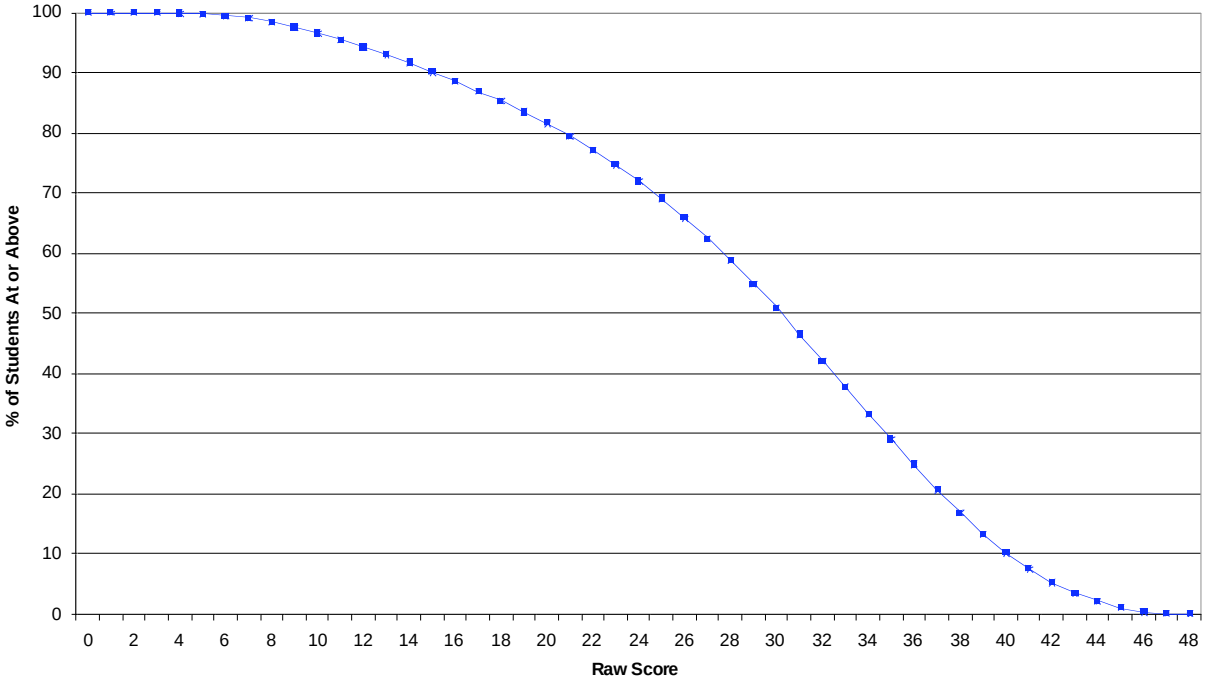
*Row Pct ,1 Limit ,2 Basic ,3 Prof ,4 Accel ,5 Advan , Total*

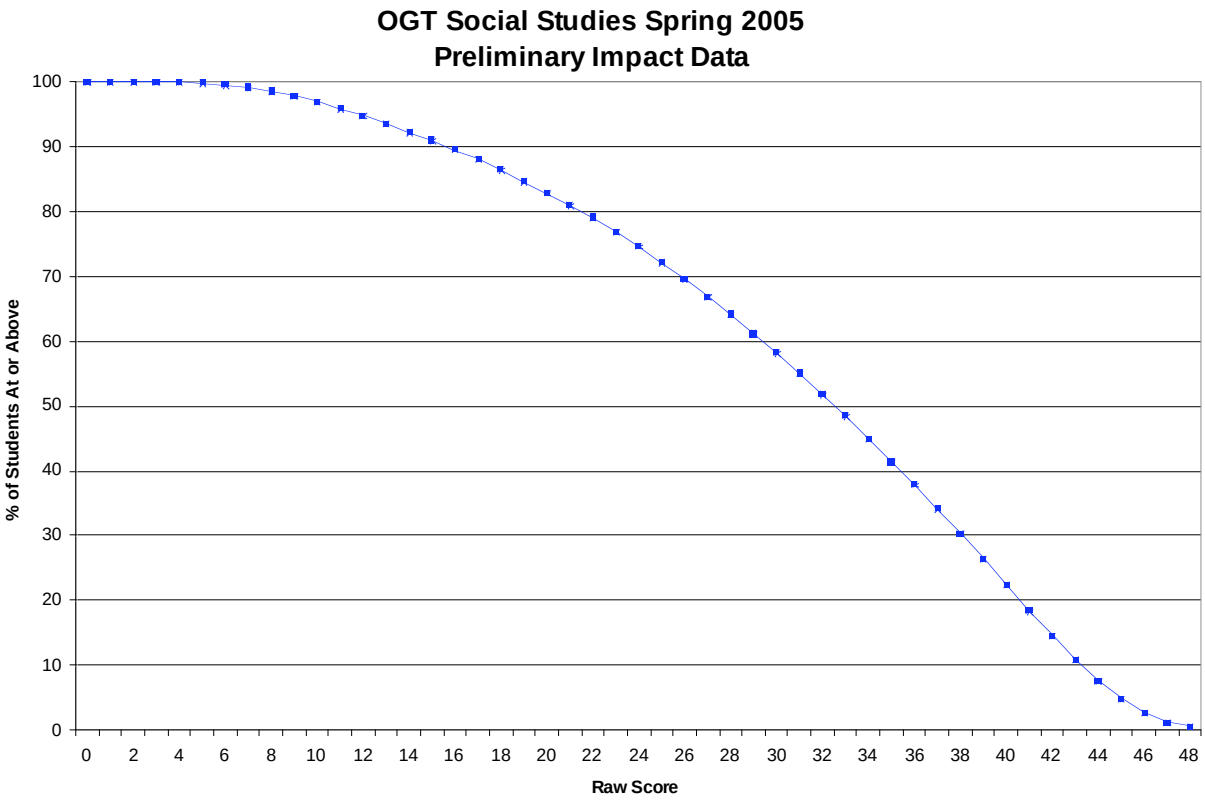
| | | | | | | |
|---|---|---|---|---|---|---|
| F | 2.73 | 9.29 | 36.73 | 45.97 | 5.28 | |
| M | 7.32 | 15.93 | 40.65 | 33.68 | 2.42 | |
| Total | 2275 | 5690 | 17406 | 17865 | 1721 | 44957 |

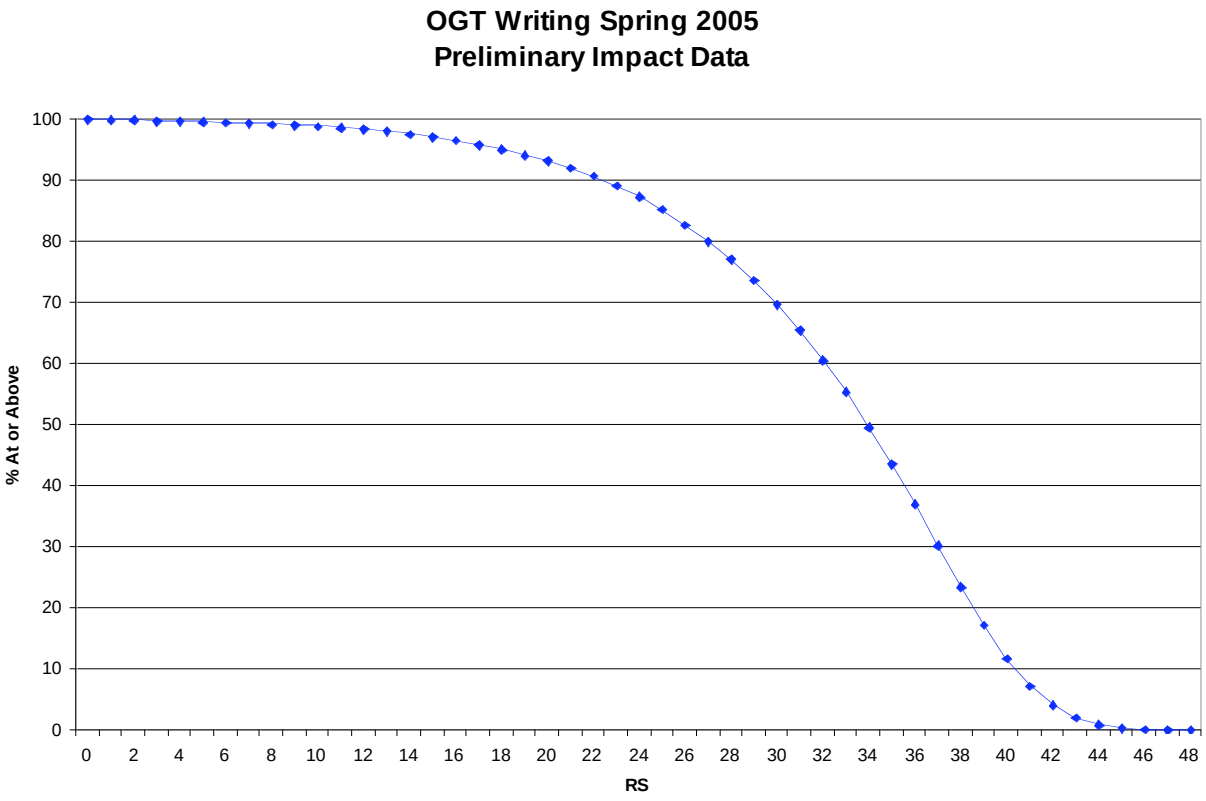
*Frequency Missing = 50*

**Round 3 impact for all subjects is shown in the body of the report.**

**OGT Science Spring 2005**
**Preliminary Impact Data**




OGT Social Studies Spring 2005
Preliminary Impact Data
% of Students At or Above
100
90
80
70
60
50
40
30
20
10
0
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48
Raw Score


OGT Writing Spring 2005
Preliminary Impact Data
% At or Above
100
90
80
70
60
50
40
30
20
10
0
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48
RS

# Appendix D

## Summary of Participants' General Comments

| General Comments | # of Responses |
|---|---|
| Wishes that others were informed about the development of cut scores and the selection process for committee members | 2 |
| Good Professional Development | 3 |
| Interesting and Informative Process | 13 |
| Glad to be a part of the process | 9 |
| Encouraged opinions, but not equally valued | 1 |
| Group facilitator did a good job | 7 |
| ODE should encourage school districts to implement benchmarks for OGT testing | 1 |
| Appreciated great effort to maintain integrity in the process | 1 |
| More discussion was needed in reference to the committee members' purpose | 2 |
| Too much downtime in the meetings | 1 |
| MI did an outstanding job of conducting the meeting | 5 |
| More discussion | 1 |
| ODE representation was positive | 2 |
| Cut scores may have been lowered b/c of pressure to consider particular ethnic groups | 1 |
| Excellent facility | 2 |
| Rooms were not adequate | 1 |
| Need additional information | 1 |
| | |