APPENDIX I—STANDARD SETTING REPORT

MEPA STANDARD SETTING REPORT

The standard setting meetings for MEPA were held February 2–5, 2005 in Waltham, MA. The purpose of the standard setting meetings was to establish cut scores for each of the four grade spans. There were four panels, one for each of the four grade spans (3–4, 5–6, 7–8, and 9–12). Each panel consisted of 20 panelists, with the exception of grade span 5–6, for which there were 18 panelists.

The standard setting method implemented for all grade spans was a modified version of the bookmark method. An overview of this method is described below. All panels followed the same procedures. To help ensure consistency of procedures between panels, each panel was led through the standard setting process by trained facilitators from Measured Progress.

The MEPA standard setting process was divided into the following three stages, each with several constituent tasks.

- Tasks completed prior to the meeting
  - Creation of performance levels and performance level definitions
  - Preparation of materials for panelists
  - Preparation of presentation materials
  - Preparation of instructions for facilitators document
  - Preparation of systems and materials for analysis during the meeting
  - Selection of panelists

- Tasks completed during the meeting
  - Orientation
  - Reviewing assessment materials
  - Filling out item map
  - Round 1 judgments
  - Tabulation of round 1 results
  - Round 2: Comparison of panelist judgments and opportunity for revised judgments
  - Tabulation of round 2 results
  - Round 3: Comparison of panelist results and impact data, and final opportunity to revise judgments
  - Evaluation

- Tasks completed after the meeting
  - Analysis and review of panelists’ feedback
  - Preparation of recommended cut scores
  - Preparation of standard setting report

Tasks Completed Prior to the Standard Setting Meeting

Creation of Performance Levels and Performance Level Definitions

The performance level definitions provided panelists the official description of the knowledge, skills, and abilities students are expected to be able to demonstrate to be classified into each performance level. These performance level definitions were presented to panelists. The definitions are provided in Appendix I–1 of this document.

Preparation of Materials for Panelists

The following materials were assembled into folders for presentation to the panelists at the standard setting meeting:

- Meeting agenda
- Confidentiality agreement
- Performance level definitions
- Assessment booklet
- Scoring rubrics
- Ordered item booklet
- Rating forms
- Evaluation form

Preparation of Presentation Materials

The PowerPoint presentation used in the opening session was prepared prior to the meeting. A copy of the PowerPoint slides is included in Appendix I–2 of this document

Preparation of Instructions for Facilitators Document

A document, "General Instructions for MEPA Standard Setting Group Facilitators," was created for the group facilitators to refer to as they worked through the process. A copy of these instructions is included in Appendix I–3 of this document.

Preparation of Systems and Materials for Analysis During the Meeting

The programming of all analyses to be conducted during the standard setting meeting was completed and thoroughly tested prior to the standard setting meeting.

Selection of Panelists

Panelists were selected prior to the standard setting meeting by the Massachusetts Department of Elementary and Secondary Education. Seventy-eight panelists participated, distributed as follows:

- Grade span 3–4: 20
- Grade span 5–6: 20
- Grade span 7–8: 18
- Grade span 9–12: 20

Of the 78 panelists, there were 58 teachers, 10 administrators, 5 specialists, 4 tutors or coaches, and 1 pathologist. The sample of panelists was chosen to be as geographically representative as possible.

Tasks Completed During the Standard Setting Meeting

Orientation

The standard setting meeting began on Wednesday afternoon with a general orientation session that was attended by all panelists. The purpose of this session was to provide some background information, provide an introduction to the issues of standard setting, and to explain the activities that would occur during the standard setting meeting. At the conclusion of the opening session the floor was opened to questions about the standard setting process. Most of the questions focused on the uses of the standard setting results and other policy-related issues, although some questions addressed the ratings to be made and clarification of the process.

After the large-group session, the panelists assembled into their grade span groups. Each group was in a separate room and each room was further divided into three tables of five to seven panelists each.

Reviewing Assessment Materials

Each panel reviewed the assessment materials, including the MEPA R/W test booklet and the MELA-O materials. The purpose of this step was to make sure the panelists were thoroughly familiar with the assessment and what the students needed to do. Next, the panelists reviewed the performance level descriptors. This step of the process was very important; it was designed to ensure that the panelists thoroughly understood the knowledge, skills, and abilities that students needed to demonstrate in order to be classified as *Early Intermediate, Intermediate,* and *Transitioning*. Discussion of the performance level descriptors occurred both at the tables and among the whole grade-span group, until the panelists felt comfortable that they had come to agreement about what characteristics described a student who was just able enough to be classified into each performance level.

Filling Out Item Map

The purpose of the next step was to ensure that panelists became very familiar with the ordered item booklet and understood the relationships between the ordered items. The ordered item booklet contained one item (or item score category) per page, and was ordered from the easiest item category to the most difficult. The ordered item booklet was created by sorting items by their IRT-based values (b corresponding to $p^{+} = 0.67$ was used). A one-parameter logistic IRT model was used for the dichotomous items, and the partial credit IRT model was used for the polytomous items. The group facilitators explained to the panelists that each MEPA R/W constructed response item and MELA-O indicator would appear multiple times in the ordered item booklet, once for each possible score point.

Each group stepped through the ordered item booklet, item by item, and discussed the knowledge, skills, and abilities students needed to complete each item. Panelists were able to refer to the scoring rubric and the performance level definitions to help them determine this information. Once they were done discussing each item, panelists wrote the knowledge, skills, and abilities onto the item map. The same information was to be filled in each time a particular item appeared in the ordered item booklet.

Round 1 Judgments

In the first round, panelists worked individually to make their initial judgment of where the bookmarks should be placed. For this task, panelists used the performance level definitions, the item map they completed in the previous step, and the ordered item booklet. Starting with the definition of the *Beginning* performance level, panelists considered the skills and abilities students needed to complete each ordered item and asked themselves the question, "Is a student performing at the *Beginning* level likely to have answered this item correctly?" As they read the items in the booklet, each panelist placed a bookmark (representing the cut score between *Beginning* and *Early Intermediate*) before the first ordered item they felt required skills or knowledge beyond those expected of a student performing at the *Beginning* level. The panelists then repeated this process for the *Early Intermediate/Intermediate* and *Intermediate/Transitioning* cut scores. Each panelist used the Rating Form provided to record his/her ratings. Copies of the rating sheets used are provided in Appendix I–4.

Tabulation of Round 1 Results

Each table of panelists received a bar chart for each cut point that showed where each panelist at the table placed his or her bookmarks. This chart was then used to facilitate discussion of the table ratings in round 2.

Round 2: Comparison of Panelist Judgments and Opportunity for Revised Judgments

During round 2, the panelists at each table examined the results from round 1 and discussed their ratings. The panelists shared their rationale for their bookmark placement in terms of the knowledge and skills students need to reach that cut score. After all panelists had an opportunity to discuss their bookmark placement and they completed their discussions, the panelists then had the opportunity to change or revise their round 1 ratings. Each panelist once again used the Rating Form to record his/her ratings.

Tabulation of Round 2 Results

As with round 1, bar charts were provided after round 2; in this case, the graphs showed the bookmark placement of each panelist in the room, rather than by table. In addition for round 2, the average placement for the room as a whole was also provided.

Round 3: Comparison of Panelist Results and Impact Data, and Final Opportunity to Revise Judgments

All the results from round 2 were distributed to panelists prior to the final round of ratings. As a whole room, panelists discussed the round 2 ratings. After the round 3 discussions, each panelist had another opportunity to change his/her ratings, again using the Rating Form.

Evaluation

Upon completion of the rating process, panelists anonymously completed an evaluation form. The results of the evaluations are presented in Appendix I–5.

Tasks Completed After the Standard Setting Meeting

Upon conclusion of the standard setting meeting, several important tasks were completed. These tasks centered on reviewing the standard setting meeting and addressing anomalies that may have occurred in the process or in the results.

Analysis and Review of Panelists' Feedback

Upon completion of the evaluation forms, panelists' responses were reviewed. This review did not reveal any anomalies in the standard setting process or indicate any reason that a particular panelist's data should not be incorporated in obtaining the final results. It appeared that all panelists understood the rating task and attended to it appropriately.

Preparation of Recommended Cut Scores

After the standard setting was completed, the cut points on the ordered item scale and on the theta (θ) scale were calculated based on the panelists' round 3 cuts. In addition, the percentage of students who would be classified into each performance level was determined. These results are presented in Tables I–1 through I–4. In addition, Figure I–1 shows the percentage of students who would fall below each cut point by grade span.

Tables I–1 through I–4 and Figure I–1 show that while the standards set by the panelists were fairly consistent across grade spans for the *Intermediate/Transitioning* (I/T) cut, there were some discrepancies for the other two cut points. In particular, the *Beginning/Early Intermediate* (B/EI) cut for grade span 5–6 and the *Early Intermediate/Intermediate* (EI/I) cuts for grade spans 5–6 and 7–8 were identified as showing fairly substantial differences from the other grade span cuts. In view of these discrepancies, smoothed cut points were also determined. The cuts were smoothed by simply fitting a linear best-fit line to the lines shown in Figure I–1, then determining the theta cut corresponding to the smoothed percent below value. These results, as well as the resulting percents-in-category, are also shown in Tables I–1 through I–4.

The final step in determining the recommended cut points was to convene a group to validate the smoothed cut point values. This group consisted of Department personnel as well as content area experts from the Department and from Measured Progress. Any cut for which the smoothed cut was more than one standard error of measurement different from the original cut was identified for validation. These cuts were the three identified above plus the B/EI cut for grade span 3–4. In addition, all four I/T cuts were validated, since that cut is the most critical in terms of consequences for students. In all, 8 of the 12 cuts were discussed by the validation group. The group felt that the locations of the smoothed cut points were consistent with the performance level descriptors, and the smoothed cuts were adopted as the official cut points to be used for reporting. The final approved cuts are the smoothed values shown in Tables I-1 through I-4.

Table I–1: Summary of MEPA Standard Setting Results—Grade Span 3–4

Performance Level	Standard Setting Round 3			Final Adopted Cuts (Smoothed)		
	Ord Item Cut	Theta Cut	% in Category	Ord Item Cut	Theta Cut	% in Category
Beginning			25.5			20.2
Early Intermediate	22/23	-0.727	20.0	13/14	-0.897	21.0
Intermediate	50/51	-0.269	32.5	49/50	-0.331	34.0
Transitioning	82/83	0.340	22.1	75/76	0.280	24.7

Table I–2: Summary of MEPA Standard Setting Results—Grade Span 5–6

Performance Level	Standard Setting Round 3			Final Adopted Cuts (Smoothed)		
	Ord Item Cut	Theta Cut	% in Category	Ord Item Cut	Theta Cut	% in Category
Beginning			16.4%			24.1
Early Intermediate	14/15	-1.220	14.7	24/25	-0.819	18.2
Intermediate	44/45	-0.580	41.1	65/66	-0.299	32.8
Transitioning	93/94	0.343	27.8	98/99	0.409	24.8

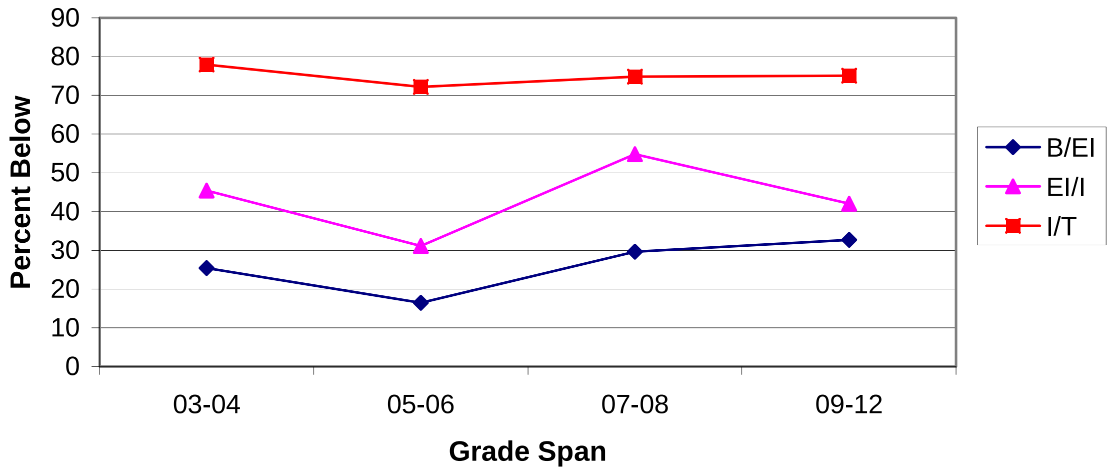
Table I–3: Summary of MEPA Standard Setting Results—Grade Span 7–8

Performance Level	Standard Setting Round 3			Final Adopted Cuts (Smoothed)		
	Ord Item Cut	Theta Cut	% in Category	Ord Item Cut	Theta Cut	% in Category
Beginning			29.6			27.5
Early Intermediate	29/30	-0.582	25.1	24/25	-0.656	15.5
Intermediate	63/64	0.027	20.1	58/59	-0.194	31.2
Transitioning	96/97	0.472	25.2	96/97	0.460	25.8

Table I–4: Summary of MEPA Standard Setting Results—Grade Span 9–12

Performance Level	Standard Setting Round 3			Final Adopted Cuts (Smoothed)		
	Ord Item Cut	Theta Cut	% in Category	Ord Item Cut	Theta Cut	% in Category
Beginning			32.7			31.3
Early Intermediate	29/30	-0.450	9.3	26/27	-0.487	14.0
Intermediate	51/52	-0.205	33.0	58/59	-0.148	27.8
Transitioning	85/86	0.484	25.0	83/84	0.451	26.9

Figure I–1: Percentage of Students Below Each Cut Point by Grade Span



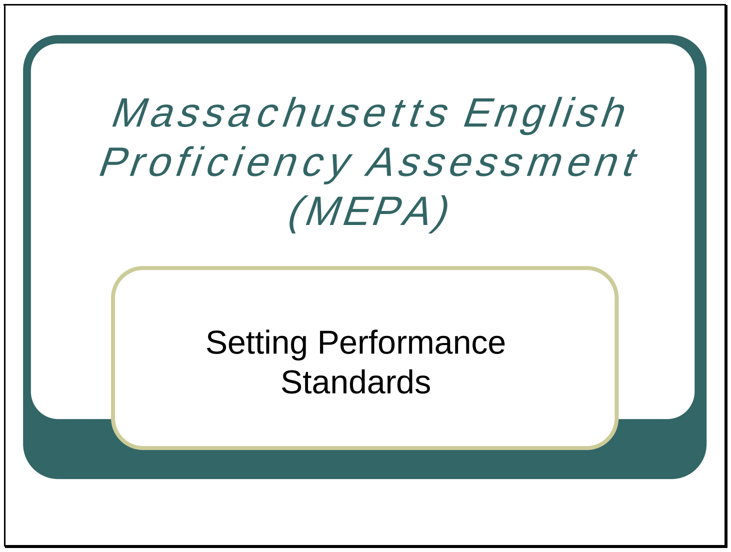
Preparation of Standard Setting Report

This report documents the procedures and results of the standard setting meetings in the establishment of performance standards for the Massachusetts English Proficiency Assessment (MEPA).

APPENDIX I–1
PERFORMANCE LEVEL DESCRIPTORS

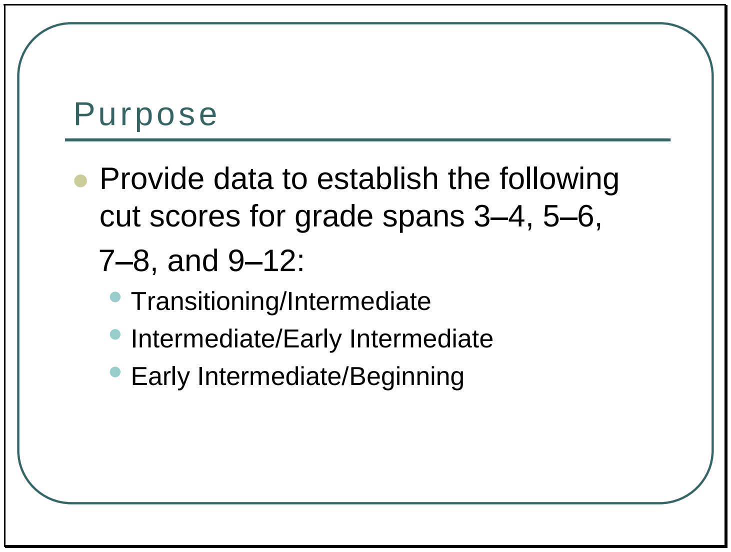
Beginning	*Early Intermediate*	*Intermediate*	*Transitioning*
Reading			
-vocabulary	Recognizes common words	Recognizes common words, and some uncommon and academic words based on context clues	Recognizes most common and academic words based on context clues
-comprehension	Comprehends simple words, phrases and sentences; moderate comprehension of literal meaning (*e.g., identifies facts*) of below grade-level texts	Comprehends literal meaning of simple grade-level texts; limited comprehension of inferential meaning *(e.g., infers meaning based on evidence)* of simple grade-level texts	Comprehends literal meaning and inferential meaning of moderately difficult grade-level texts
-literary elements	Recognizes some elements (e.g., author's purpose, character, and setting) of below grade-level texts	Recognizes elements (e.g., author's purpose, character, and setting) of simple grade-level texts	Recognizes elements (e.g., author's purpose, character, and setting) of moderately difficult grade-level texts
Writing			
-writing	Writes short paragraphs with minimal control of organization, conventions and supporting details	Writes short, simple compositions with limited control of organization, development of ideas, conventions, and supporting details	Writes short compositions with moderate control of organization, development of ideas, conventions and supporting details
-editing	Demonstrates limited knowledge of standard English conventions	Demonstrates partial knowledge of standard English conventions	Demonstrates general knowledge of standard English conventions
Speaking (Production)			
	Uses common words, and simple phrases and expressions in classroom and interpersonal communications; word choice is usually inappropriate or incorrect Uses basic grammar patterns and simple sentence structures in oral communication	Uses words, phrases and expressions with varying appropriateness/accuracy in classroom and interpersonal communications Uses basic grammar patterns and sentence structures in oral communication; uses complex language structures but with frequent errors	Uses appropriate/correct words, phrases, and expressions in classroom and interpersonal communications Uses basic and complex grammar patterns and sentence structures in oral communication; errors do not obscure meaning
Listening (Comprehension)			
	Comprehends basic words, phrases and expressions in oral classroom and interpersonal communications, with frequent need for clarification	Comprehends most oral classroom and interpersonal communications but with some need for clarification	Comprehends extended and sustained oral classroom and interpersonal communication with little or no need for clarification

APPENDIX I–2
OPENING SESSION POWERPOINT PRESENTATION



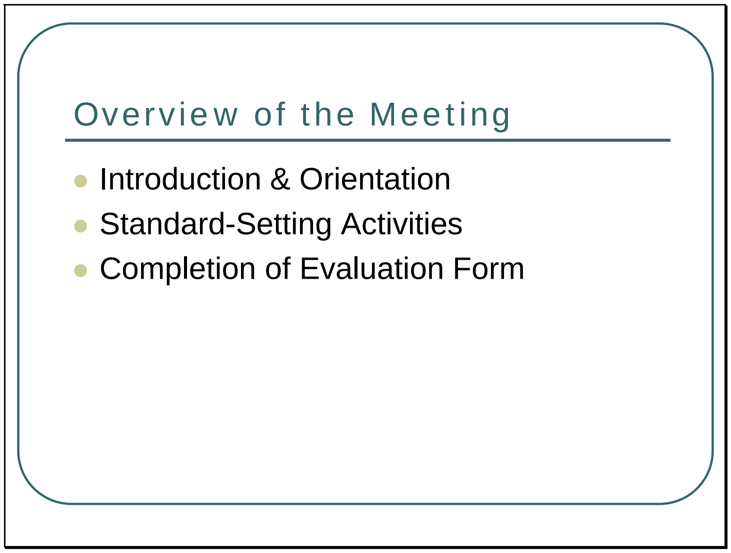
Massachusetts English Proficiency Assessment (MEPA)

Setting Performance Standards



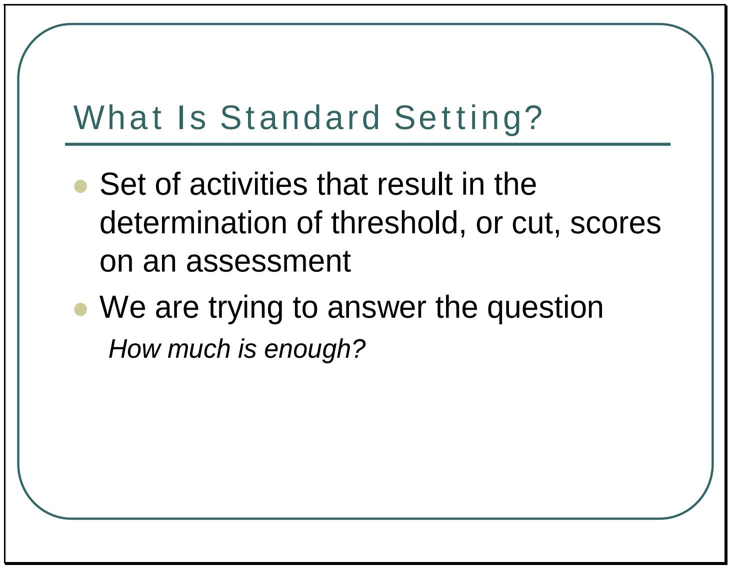
Purpose

- Provide data to establish the following cut scores for grade spans 3–4, 5–6, 7–8, and 9–12:
  - Transitioning/Intermediate
  - Intermediate/Early Intermediate
  - Early Intermediate/Beginning



Overview of the Meeting

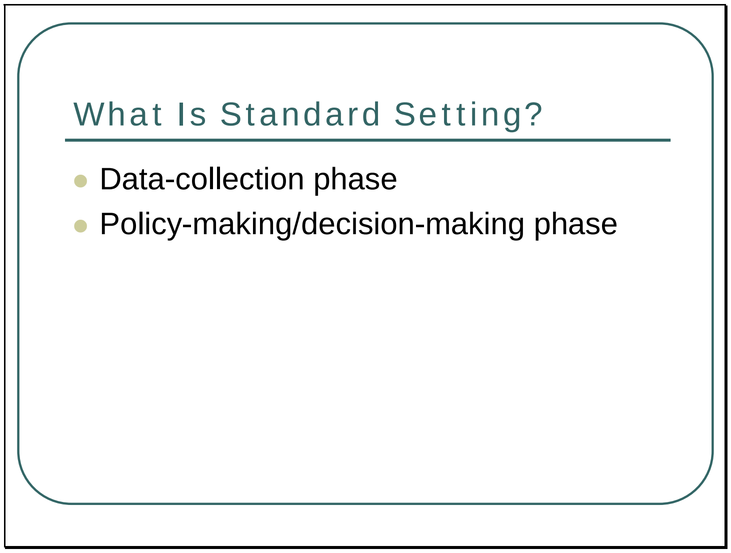
- Introduction & Orientation
- Standard-Setting Activities
- Completion of Evaluation Form



What Is Standard Setting?

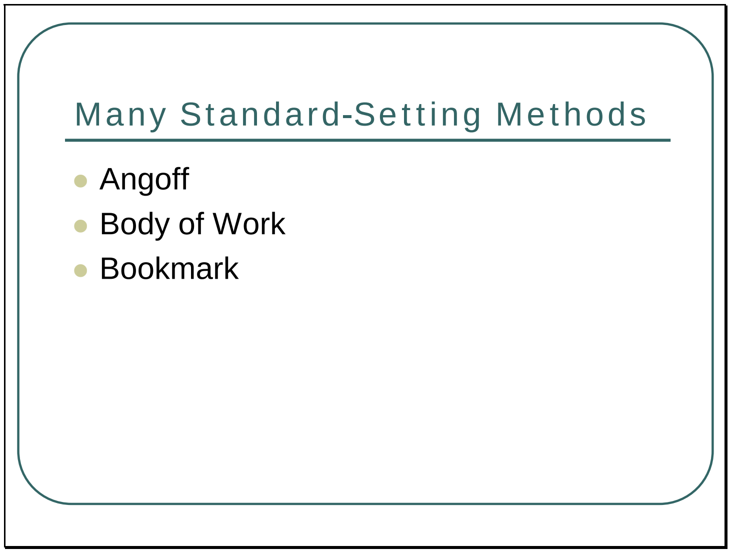
- Set of activities that result in the determination of threshold, or cut, scores on an assessment
- We are trying to answer the question

  How much is enough?



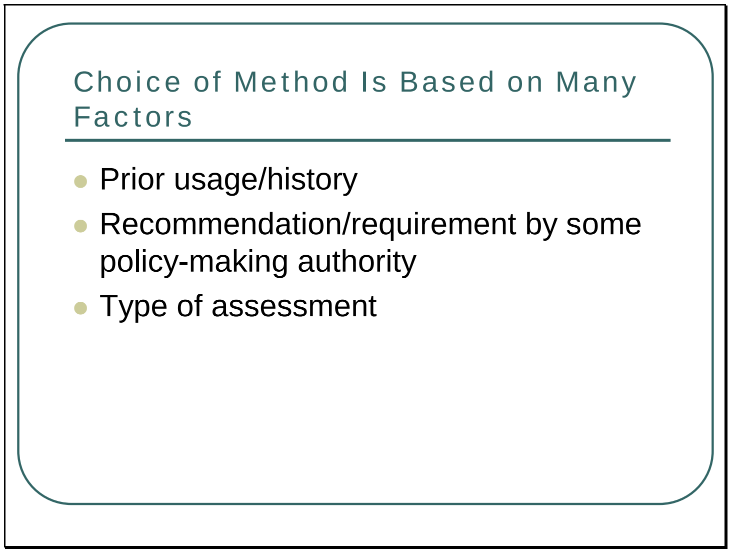
What Is Standard Setting?

- Data-collection phase
- Policy-making/decision-making phase



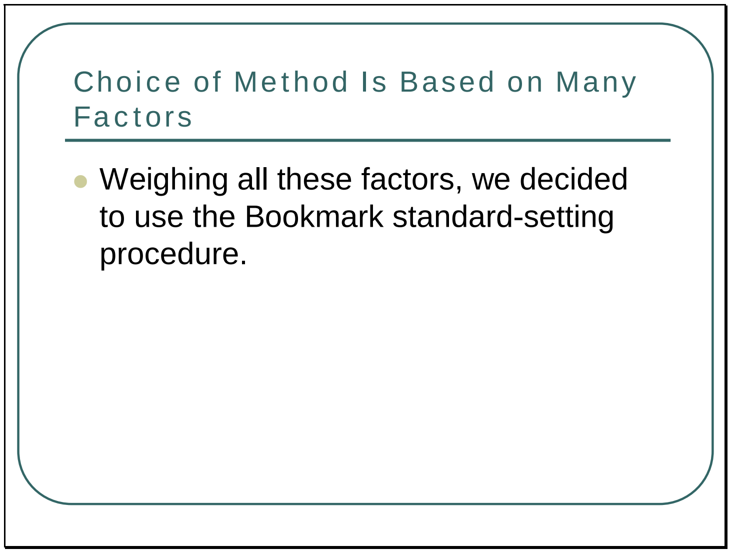
Many Standard-Setting Methods

- Angoff
- Body of Work
- Bookmark



Choice of Method Is Based on Many Factors

- Prior usage/history
- Recommendation/requirement by some policy-making authority
- Type of assessment



Choice of Method Is Based on Many Factors

- Weighing all these factors, we decided to use the Bookmark standard-setting procedure.

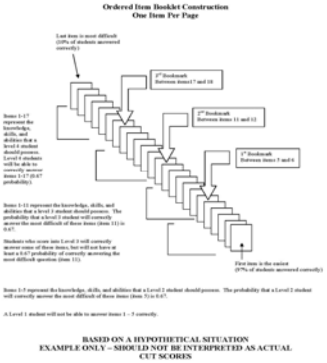
Why Use the Bookmark Procedure Here?

- Well-established procedure that has been successfully used on many assessments
- Used previously in Massachusetts
- Has produced defensible results
- Appropriate for the item types in this assessment

What Is the Bookmark Procedure?

- A standard-setting procedure that uses a book of items (ordered from easiest to hardest)
- Panelists place a bookmark in this book of items

What Is the Bookmark Procedure?



Ordered Item Booklet Construction
One Item Per Page

Last item is most difficult (10% of students answered correctly)

3rd Bookmark Between items17 and 18

2nd Bookmark Between items 11 and 12

1st Bookmark Between items 5 and 6

Items 1-17 represent the knowledge, skills, and abilities that a level 4 student should possess. Level 4 students will be able to correctly answer items 1-17 (0.67 probability).

Items 1-11 represent the knowledge, skills, and abilities that a level 3 student should possess. The probability that a level 3 student will correctly answer the most difficult of these items (item 11) is 0.67.

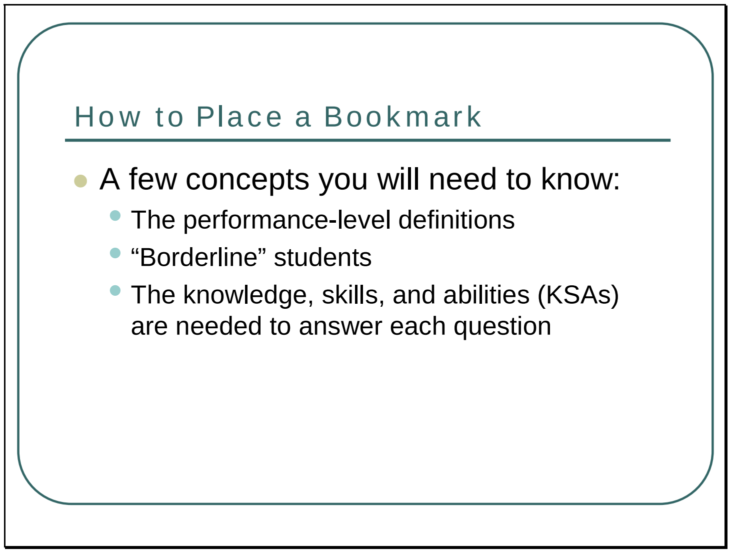
Students who score into Level 3 will correctly answer some of these items, but will not have at least a 0.67 probability of correctly answering the most difficult question (item 11).

First item is the easiest (97% of students answered correctly)

Items 1-5 represent the knowledge, skills, and abilities that a Level 2 student should possess. The probability that a Level 2 student will correctly answer the most difficult of these items (item 5) is 0.67.

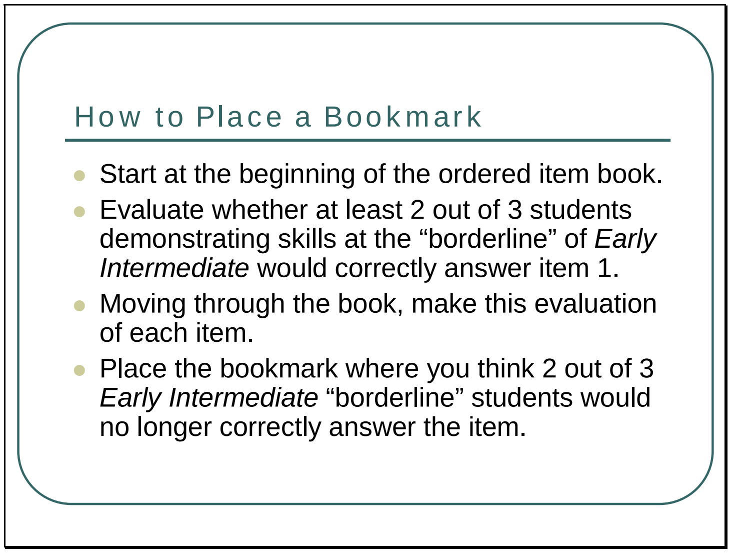
A Level 1 student will not be able to answer items 1 – 5 correctly.

BASED ON A HYPOTHETICAL SITUATION
EXAMPLE ONLY – SHOULD NOT BE INTERPRETED AS ACTUAL CUT SCORES



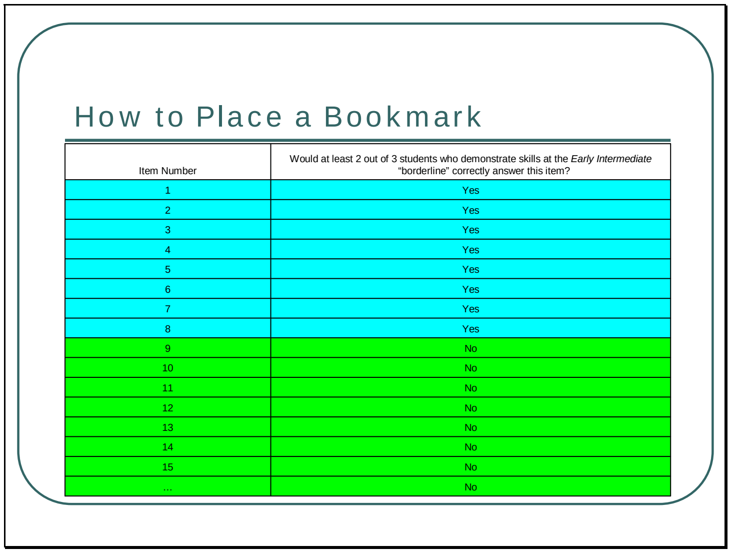
How to Place a Bookmark

- A few concepts you will need to know:
  - The performance-level definitions
  - "Borderline" students
  - The knowledge, skills, and abilities (KSAs) are needed to answer each question



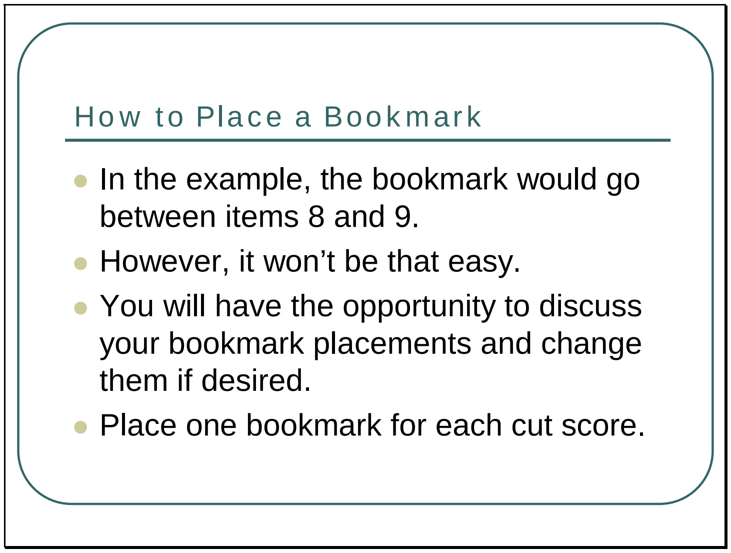
How to Place a Bookmark

- Start at the beginning of the ordered item book.
- Evaluate whether at least 2 out of 3 students demonstrating skills at the "borderline" of *Early Intermediate* would correctly answer item 1.
- Moving through the book, make this evaluation of each item.
- Place the bookmark where you think 2 out of 3 *Early Intermediate* "borderline" students would no longer correctly answer the item.



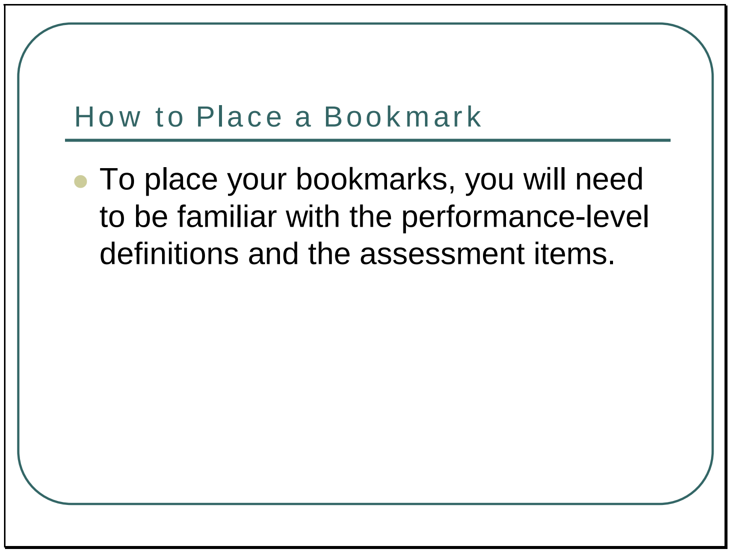
How to Place a Bookmark

Item Number	Would at least 2 out of 3 students who demonstrate skills at the *Early Intermediate* "borderline" correctly answer this item?
1	Yes
2	Yes
3	Yes
4	Yes
5	Yes
6	Yes
7	Yes
8	Yes
9	No
10	No
11	No
12	No
13	No
14	No
15	No
…	No



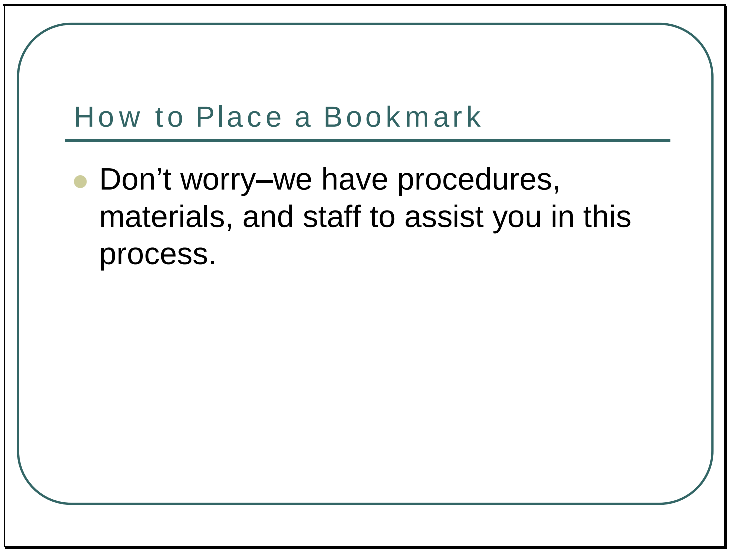
How to Place a Bookmark

- In the example, the bookmark would go between items 8 and 9.
- However, it won’t be that easy.
- You will have the opportunity to discuss your bookmark placements and change them if desired.
- Place one bookmark for each cut score.



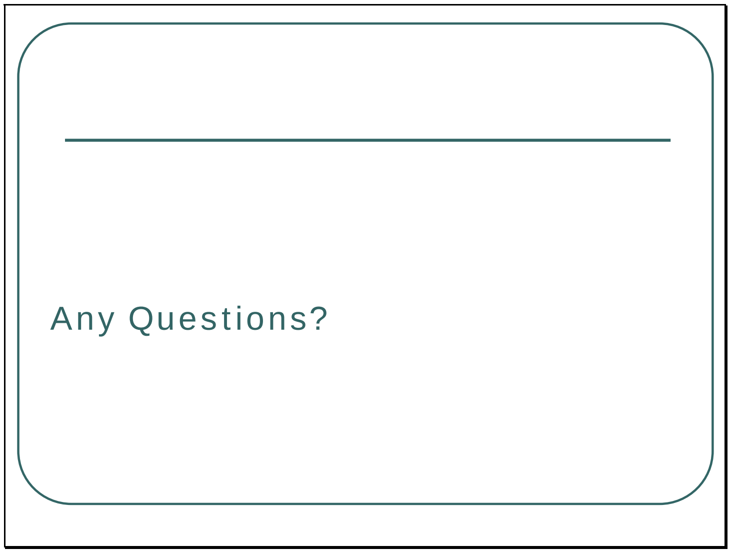
How to Place a Bookmark

- To place your bookmarks, you will need to be familiar with the performance-level definitions and the assessment items.



How to Place a Bookmark

- Don’t worry–we have procedures, materials, and staff to assist you in this process.



Any Questions?

APPENDIX I–3
GROUP FACILITATOR INSTRUCTIONS

GENERAL INSTRUCTIONS FOR MEPA STANDARD SETTING GROUP FACILITATORS PRIOR TO ROUND 1 RATINGS

Introductions

1. Welcome group, introduce yourself (name, affiliation, a little selected background information).
2. Have each participant introduce his/herself.

Review the Test

Overview: In order to establish an understanding of the MEPA test items and for panelists to gain an understanding of the experience of the students who take the test, each participant will review the reading and writing assessments as well as the MELA-O listening and speaking. Panelists may take issue with some of the items in the test. Tell them we will gladly take their feedback to the DOE. However, this is the actual assessment that students took and it is the set of items on which we must set standards.

Activities:

1. Introduce MEPA and convey/do each of the following:
   a. Tell panelists that they are about to review the actual MEPA assessment including the MELA-O.
   b. The purpose of the exercise is to help them establish a good understanding of the test items and to gain an understanding of the experience of the students who take the assessment.
2. Give each panelist a test booklet and the MELA-O.
3. Tell panelists to try to take on the perspective of a student as they review the test.

Fill Out Item Map

Overview: The primary purpose of filling out the item map is for panelists to think about and document the knowledge, skills, and abilities (KSAs) that students need to answer each question. Panelists should have an understanding of what makes one test item harder or easier than another. The notes panelists take here will be useful in helping them place their initial bookmarks and in discussions after the various rounds of ratings.

Activities:

1. Make sure panelists have the following materials:
   a. Item map
   b. Ordered item booklet
2. Review the ordered item booklet and item map with the panelists. Explain what each is, and point out the correspondence of the ordered items between the two.
3. Provide an overview of the task paraphrasing the following:
   a. The primary purpose of this activity is for panelists to think about what makes one question harder or easier than another. For example, it may be that the concept tested is a difficult concept, or that the concept isn't difficult but that

the particular wording of the question makes it a difficult question. Similarly, the concept may be a difficult one, but the wording of the question makes it easier.

   b. Panelists should take notes about their thoughts regarding each question. These will be useful in the rating activities and later discussions.

4. Tell panelists to work individually at first. After they complete the item map they will have the opportunity to discuss with their tables.
5. Each panelist will begin with the first ordered item and compare the second ordered item to it. What makes item #2 harder than item #1? Panelists should not agonize over these decisions. It may be that item #2 is only slightly harder than item #1.
6. Panelists should work their way through the entire item map, making notes as appropriate.
7. Once panelists have completed the item map, they should discuss them as a table.
8. Based on the table discussion, the panelists should modify their own item map (make additional notes, cross things out, etc.).

Discuss Performance Level Definitions & Describe Characteristics of the "Borderline" Student

Overview: In order to establish an understanding of the expected performance of borderline students on the test, panelists must have a clear understanding of:

- The definition of the four Performance Levels (*Transitioning, Intermediate*, *Early Intermediate*, and *Beginning*), and
- Characteristics of students who are "just able enough" to be classified into each performance level (PL). These students will be referred to as borderline students, since they are right on the border between performance levels.

The purpose of this activity is for the panelists to obtain an understanding of the PL definitions with an emphasis on characteristics that describe the borderline student, both what these students can and cannot do.

This activity is critical since the ratings panelists will be making in rounds 1–3 will be based on these understandings.

Activities:

1. Introduce task.
   a. Have panelists individually review the performance level descriptors.
   b. Discuss descriptors in small group.
   c. Generate small group description of borderline Nearing Proficiency, Proficient, and Advanced students.
2. Have panelists individually review the descriptors. They can make notes if they like. The goal here is for the panelists to come to a common understanding of what it means to be in each performance level. It is not unusual for panelists to disagree with the definition they will see; almost certainly there will be some panelists who will want to change the definition. However, the task at hand is for panelists to have a common understanding of what knowledge, skills, and abilities are described by each PL descriptor.
3. After reviewing the definition, have panelists discuss it and provide clarification.

The purpose of this is to have a collegial discussion and bring up/clarify any issues or questions that any individual may have and to reach a consensus on the definition.

4. Once panelists have a solid understanding of the PL descriptors, have them focus their discussion on the knowledge, skills, and abilities of students who are in the *Early Intermediate* category, but just barely. The focus should be on those characteristics and KSAs that best describe the lowest level of performance necessary to warrant a *Early Intermediate* classification.
5. After discussing *Early Intermediate,* have the panelists discuss characteristics of the borderline *Intermediate* student and then characteristics of the borderline *Transitioning* student. Panelists should be made aware of the importance of the *Transitioning* cut.

Round 1

Overview: The primary purpose of Round 1 is to ask each panelist to gauge the level of difficulty of each individual item for those students who barely meet the definition of *Early Intermediate, Intermediate*, and *Transitioning*. The task that panelists are asked to do is to estimate whether a borderline *Early Intermediate* student would answer each question correctly. More specifically, would two out of three borderline students answer the question correctly? This same question is then asked of the borderline *Intermediate* students and the borderline *Transitioning* students.

Activities:

1. Make sure panelists have the following materials:
   a. Round 1 rating form
   b. Ordered item booklet
   c. Item map
   d. Performance level definitions
2. Have panelists write in their ID number and grade span. The ID number is on their name tags.
3. Provide an overview of Round 1. Paraphrase the following:
   a. Primary purpose is for each panelist to individually estimate whether students whose performance is barely *Early Intermediate* would answer each item correctly and to place a bookmark at the location where the answer of 'yes' turns to 'no'. Remind panelists that they should be thinking about two-thirds of the borderline students.
   b. Panelists need to base their judgments on their experience with the content, understanding of students, and the definitions of the borderline students generated previously.
   c. One bookmark will be placed for each cut score.
   d. If panelists are struggling with placing a particular bookmark they should use their best judgment and move on. They will have an opportunity to revise their ratings.
   e. Panelists should feel free to take notes if there are particular points about where they placed their bookmarks that they think are worthy of discussion in future rounds.
4. Tell panelists to work individually.
5. Go over the rating form with panelists.
   a. Lead panelists through a step-by-step demonstration of how to fill in the rating

form.

b. Answer questions the panelists may have about the work in Round 1.

c. Once everyone understands what they are to do in Round 1, tell them to begin.

6. Using the ordered item booklet, each panelist begins at item 1 and decides where to place the bookmark for the *Beginning/Early Intermediate* cut.

7. Continuing through the ordered item booklet they then decide where to place the bookmark for the *Early Intermediate/Intermediate* cut.

8. Continuing through the ordered item booklet, they then decide where to place the bookmark for the *Intermediate/Transitioning* cut.

9. As panelists complete the task, ask panelists to carefully inspect their rating forms to ensure they are filled out properly.
   a. The grade and ID number must be filled in.
   b. The item numbers for each cut score must be adjacent.

Tabulation of Round 1 Results: Tabulation of Round 1 results will be completed over night for discussion the following morning.

Round 2

Overview: The primary purpose of Round 2 is to ask panelists to discuss their initial bookmark placements and to revise their ratings on the basis of that discussion. They will discuss their ratings in the context of the ratings made by other members of the group. The panelists with the highest and lowest ratings should comment on why they gave the ratings they did. The group should get a sense of how much variation there is in the ratings. Panelists should also consider the question, “How tough or easy a panelist are you?” The purpose here is to allow panelists to examine their individual expectations (in terms of their experiences) and to share these expectations and experiences in order to attain a better understanding of how their experiences impact their decision-making. Once panelists have reviewed and discussed their item estimates, they will be given the opportunity to change or revise their Round 1 ratings.

Activities:

1. Make sure panelists have the following materials:
   a. The Rounds 1 and 2 rating form
   b. Ordered item booklets
   c. Item maps
   d. Performance level definitions
2. Panelists should have already filled in their name, ID number, and grade on the rating form.
3. Provide an overview of Round 2. Paraphrase the following:
   a. As in Round 1, the primary purpose is to place bookmarks where you feel the performance levels are best distinguished.
   b. Panelists need to base their judgments on their experience with the content area, understanding of students, the definitions of the borderline students generated previously, discussions with other panelists, and the knowledge, skills, and abilities required to answer each item.
4. Review the feedback information with the panelists.
   a. Show the panelists where they have placed their cuts and where the average for each table is located. This information is useful so that panelists get a sense if they are more stringent or more lenient than other panelists.

5. Panelists should be given a few minutes to review the feedback forms.
6. Beginning with the first cut, panelists should discuss, then revise, their ratings.
   a. On the basis of the discussions and the results presented, panelists should make a second round of ratings.
   b. Panelists should be encouraged to listen to their colleagues as well as express their own points of view.
   c. If the panelists hear a logic/rationale/argument that they did not consider and that they feel is compelling, they should adjust their ratings to incorporate that information.
   d. When making revised ratings, panelists should not feel compelled to change their ratings.
   e. The group does not have to achieve consensus. If panelists honestly disagree, that is fine. We are trying to get the best judgment of each panelist. Panelists should not feel compelled or coerced to making a rating they disagree with.

Encourage the panelists to use the discussion and feedback to assess how stringent or lenient a judge they are. If a panelist is consistently higher or lower than the group they may have a different understanding of the borderline student than the rest of the group, or a different understanding of the performance level definition, or both. **It is OK for panelists to disagree, but that disagreement should be based on a common understanding of the definitions.**

7. When each panel completes its second ratings, collect the rating forms, and carefully inspect them to ensure they are filled out properly.
   a. The name, grade, and ID number must be filled in.
   b. Each cut for Round 2 must have one (and only one) rating.

Tabulation of Round 2 Results: Round 2 results will be tabulated as soon as possible upon receipt of the rating forms.

Round 3

Overview: Round 3 will proceed as Round 2 did, except that the following additional information will be provided to panelists so that they can make more informed ratings:

- The average group rating for each table and the entire room at each cut point
- Each panelist’s cut score
- The p-values of the items

Panelists will compare their results from Round 2 to those of the entire room. Additionally, the panelists will see the actual difficulty of the items to give them a sense of whether the test is as easy or as hard as they suspected. After being presented this information panelists will have an opportunity to again discuss and revise their ratings.

Activities:

1. Make sure panelists have the following materials:
   a. Round 3 rating form
   b. Round 2 results
   c. Ordered item booklet
   d. Item map
   e. Performance level definitions
2. Have panelists fill in their name, ID number, and grade. The ID number is on their name tags.
3. Provide an overview of Round 3. Paraphrase the following:
   a. As in Rounds 1 and 2, the primary purpose is for each panelist to place a bookmark between the performance levels.
   b. Panelists need to base their judgments on their experience with the content area, understanding of students, the definitions of the borderline students, discussions with other panelists, and the knowledge, skills, and abilities required of each items.
   c. Again, a single rating will be made for each cut score.
4. Review the results of Round 2 information with panelists.
   a. Starting at the top of the page, explain what all of the numbers on the page are.
   b. Review the item difficulties with the panelists. Explain what the numbers are. Point out that the session 2 items were easier for those students who took sessions 2 and 3 than for those who took sessions 1 and 2.
5. Panelists should be given a few minutes to review the results. Encourage the panelists to use this information to assess how stringent or lenient a judge they are. If a panelist is consistently higher or lower than the group he/she may have a different understanding of the borderline student than the rest of the group, or a different understanding of the performance level definitions, or both. It is OK for panelists to disagree, but that disagreement should be based on a common understanding of the definitions.
6. Beginning with the first cut, panelists should discuss, then revise, their ratings.
   a. Point out differences between tables. Encourage a discussion between the tables if there appears to be large systematic differences.
   b. On the basis of the discussions and the statistical information presented, panelists should make a third and final round of ratings.
   c. Panelists should be encouraged to listen to their colleagues as well as express their own points of view.
   d. In light of the additional information presented, if panelists hear a logic/rationale/argument that they did not consider and that they feel is compelling, then they should adjust their ratings to incorporate that information.
   e. When making revised ratings, panelists should not feel compelled to change their ratings.
   f. The group does not have to achieve consensus. If panelists honestly disagree, that is fine. We are trying to get the best judgment of each panelist. Panelists should not feel compelled or coerced to making a rating they disagree with.
7. When each panel completes its final ratings, collect the rating forms from each. When you collect the rating forms carefully inspect them to ensure they are filled out properly.

a. The name, grade, and ID number must be filled in.
b. Each cut must have one (and only one) rating.

Tabulation of Round 3 Results: Round 3 results will be tabulated as quickly as possible after the rating sheets have been turned in.

Set Minimum Cut Scores for *Transitioning* Students in Each Sub-Domain

Overview: For a student to be classified as *Transitioning*, two separate criteria must be achieved. First, a student must achieve a passing score on the total MEPA assessment (this score will result from the preceding standard setting activities). Second, students must demonstrate a minimum level of ability in each of the sub-domains (reading, writing, listening, speaking). The purpose of this activity is to recommend those minimum thresholds of performance in each sub-domain.

For this activity there will be only one round of ratings. Panelists will be provided separate ordered item books for each sub-domain. The Massachusetts DOE will provide for each sub-domain an initial threshold score based on recommendations from content experts. These initial threshold scores will be placed as bookmarks in each separate subdomain. Panelists will evaluate the KSAs needed to answer the questions prior to the bookmark and determine if the bookmark is properly placed.

Activities:

1. Make sure panelists have the following materials:
   a. Minimum *Transitioning* rating form
   b. Ordered item booklets
   c. Item map
   d. Performance level definition of *Transitioning*
2. Have panelists fill in their name, ID number, and grade. The ID number is on their name tags.
3. Provide an overview of this activity. Paraphrase the following:
   a. The primary purpose is for each panelist to evaluate the initial placement of each sub-domain bookmark and, if deemed appropriate, to recommend an adjustment to that initial placement.
   b. The judgment the panelists must make is a bit different here than previously. Panelists should focus on what is the **minimum** skill that needs to be displayed in each area for *Transitioning* students. This is NOT the same as setting the *Transitioning* cut in each area. This recognizes the importance of the *Transitioning* cut, and completes the statement that “if a student can’t do X he/she shouldn’t be transitioning,” where X represents the KSAs indicated by the bookmark placement.
4. Panelists need to base their judgments on their experience with the content area, understanding of students, the definition of the *Transitioning* students, discussions with other panelists, and the knowledge, skills, and abilities required of each item. One rating will be made for each sub-domain.
5. Review the placement of the initial bookmark for Reading with the panelists.
   a. Have the panelists discuss the KSAs needed to answer the questions prior to and after the bookmark. Have panelists discuss the appropriateness of the bookmark placement and to revise the placement on the basis of the discussion. Panelists should be encouraged to listen to their colleagues as well

as express their own points of view.

b. The group does not have to achieve consensus. If panelists honestly disagree, that is fine. We are trying to get the best judgment of each panelist. Panelists should not feel compelled or coerced to making a rating they disagree with.

c. Go through the same procedure for Writing, Listening, and Speaking.

6. When each panel completes its final ratings, collect the rating forms from each. When you collect the rating forms carefully inspect them to ensure they are filled out properly.

   a. The name, grade, and ID number must be filled in.

   b. Each cut must have one (and only one) rating.

Complete Evaluation Form

Upon completion of revising the PL descriptors, have panelists fill out the evaluation form. Emphasize that their honest feedback is important.

APPENDIX I–4
RATING FORMS

MASSACHUSETTS ENGLISH PROFICIENCY ASSESSMENT RATING FORM
GRADE SPAN 3–4

Round ________________

Panelist ID # ____________

Table __________________

Beginning	***Early Intermediate***	***Intermediate***	***Transitioning***
(B)	**(EI)**	**(I)**	**(T)**
Item-ordered numbers	Item-ordered numbers	Item-ordered numbers	Item-ordered numbers
First	First	First	First

Directions: Please enter the range of item-ordered numbers that fall into each performance level category according to where you placed your bookmarks.

Important: The ranges MUST be adjacent to each other. For example: Beginning **1–30**, Early Intermediate **31–60**, Intermediate **61–90**, Transitioning **91–117**.

MASSACHUSETTS ENGLISH PROFICIENCY ASSESSMENT RATING FORM
GRADE SPAN 5–6

Round __________________

Panelist ID # _____________

Table ___________________

Beginning **(B)** Item-ordered numbers First	***Early Intermediate*** **(EI)** Item-ordered numbers First	***Intermediate*** **(I)** Item-ordered numbers First	***Transitioning*** **(T)** Item-ordered numbers First

Directions: Please enter the range of item-ordered numbers that fall into each performance level category according to where you placed your bookmarks.

Important: The ranges MUST be adjacent to each other. For example: Beginning **1-30**, Early Intermediate **31-60**, Intermediate **61-90**, Transitioning **91-117**.

MASSACHUSETTS ENGLISH PROFICIENCY ASSESSMENT
RATING FORM
GRADE SPAN 7–8

Round __________________

Panelist ID # _____________

Table ___________________

Beginning **(B)** Item-ordered numbers First	***Early Intermediate*** **(EI)** Item-ordered numbers First	***Intermediate*** **(I)** Item-ordered numbers First	***Transitioning*** **(T)** Item-ordered numbers First

Directions: Please enter the range of item-ordered numbers that fall into each performance level category according to where you placed your bookmarks.

Important: The ranges MUST be adjacent to each other. For example: Beginning **1-30**, Early Intermediate **31-60**, Intermediate **61-90**, Transitioning **91-117**.

MASSACHUSETTS ENGLISH PROFICIENCY ASSESSMENT
RATING FORM
GRADE SPAN 9–12

Round __________________

Panelist ID # _____________

Table ___________________

Beginning **(B)** Item-ordered numbers First	***Early Intermediate*** **(EI)** Item-ordered numbers First	***Intermediate*** **(I)** Item-ordered numbers First	***Transitioning*** **(T)** Item-ordered numbers First

Directions: Please enter the range of item-ordered numbers that fall into each performance level category according to where you placed your bookmarks.

Important: The ranges MUST be adjacent to each other. For example: Beginning **1-30**, Early Intermediate **31-60**, Intermediate **61-90**, Transitioning **91-117**.

APPENDIX I–5
EVALUATION FORM RESULTS

Massachusetts English Proficiency Assessment
2005 Standard Setting Evaluation Form

1. Please mark the grade span for which you set standards.

☒ Grade Span 3-4　□ Grade Span 5-6　□ Grade Span 7-8　□ Grade Span 9-12

2. What was your comfort level with the standard setting process <u>at the beginning</u> of the process?

Extremely Uncomfortable		Somewhat Comfortable		Extremely Comfortable
1	2	3	4	5
0	*11*	*6*	*3*	*0*

3. What was your comfort level with the standard setting process <u>at the end</u> of the process?

Extremely Uncomfortable		Somewhat Comfortable		Extremely Comfortable
1	2	3	4	5
0	*2*	*2*	*7*	*9*

4. To what extent did the training prepare you to complete the task of standard setting?

Not at all		Somewhat Well		Extremely Well
1	2	3	4	5
0	*0*	*4*	*8*	*8*

5. How clear were the performance level descriptors?

Not at all Clear		Somewhat Clear		Very Clear
1	2	3	4	5
0	*1*	*4*	*14*	*1*

6. How clear was the bookmarking task?

Not at all Clear		Somewhat Clear		Very Clear
1	2	3	4	5
0	*0*	*2*	*8*	*10*

7. To what extent was the length of this meeting appropriate for the task of setting performance standards?

Too Little Time		About Right		Too Much Time
1	2	3	4	5
0	*3*	*10*	*7*	*0*

8. How would you characterize the organization of the standard setting session activities?

Disorganized		Somewhat Organized		Extremely Organized
1	2	3	4	5
0	0	1	11	8

9. What is your level of confidence in the bookmarks you placed?

Very Low				Very High
1	2	3	4	5
0	0	1	14	5

10. How influential were the following factors in determining where you set your bookmarks:

	Not at all Influential		Somewhat Influential		Very Influential
A. The performance level descriptors	1	2	3	4	5
	0	0	0	9	11
B. The assessment items	1	2	3	4	5
	0	0	3	9	8
C. Other panelists' comments	1	2	3	4	5
	0	0	5	9	6
D. My professional experience	1	2	3	4	5
	0	0	1	9	9
E. Rater feedback data	1	2	3	4	5
	0	0	7	10	3
F. Item difficulty statistics	1	2	3	4	5
	1	3	4	10	2

11. What confidence do you have in the classification of standards at the Beginning level?

Very Low				Very High
1	2	3	4	5
0	***1***	***6***	***8***	***5***

12. What confidence do you have in the classification of standards at the Early Intermediate level?

Very Low				Very High
1	2	3	4	5
0	***0***	***5***	***10***	***5***

13. What confidence do you have in the classification of standards at the Intermediate level?

Very Low				Very High
1	2	3	4	5
0	***0***	***2***	***14***	***4***

14. What confidence do you have in the classification of standards at the Transitioning level?

Very Low				Very High
1	2	3	4	5
0	***0***	***1***	***10***	***9***

15. What confidence do you have in the classification of standards across all performance levels?

Very Low				Very High
1	2	3	4	5
0	***0***	***2***	***14***	***4***

16. Please use the space below to provide comments about the standard setting process and/or suggestions as to how the process could be improved.

____

____

____

____

____

____

Thank you for your hard work and valuable feedback!

RESPONSES

Extremely fascinating experience and most helpful to discuss assessment items with peers from all over the state.

There was a lot of confusion regarding what part of the question was read aloud, the number of ordered items vs. the number it was on the test. For the first time I think it went very well.

Great experience!! Nice work Liz!!!

It was grueling the first day and better the second. Allow more time for those first few tasks.

The process would have been easier without the MELA-O being intended. The facilitator was very good.

I'd like to see a clearer correlation between the section from the original bracket and the order of difficulties from the binder. Also, I would think it's a lot easier to place the benchmark with out the mixing in MELA-O score.

It was very helpful, hearing the sharing session, very friendly.

Good process. Note: 181738 graphic looks like boy with long shorts, 206706 change chart, 206699 sentence is complete as is. Liz did a great job facilitating.

It was a great learning experience for me! Thank you!

The room setting could have been more conducive to whole group discussion.

The instructions and explanations were very good. I would have liked to have had more access to DOE staff for additional questions concerning MELA-O and MEPA items scored together.

Valuable training. Great opportunity. Quick-paced. Treated well.

Try to keep first evening to orientation – most brain dead at 7:30pm – 8pm. Location numbers were based on a field test – this info given after lunch Fri – would have meant something earlier. Found inclusion of MELA-O items very confusing – and found too much of discussions went off on MELA-O. Distracting.

The room size was a little small and not conducive to whole group discussions.

I felt very confused at end of day one. I became confused. Day two made me feel very comfortable with the process.

I thought we spent too much time at beginning on the PowerPoint overview – time would be better spent familiarizing ourselves with test items and whole process.

I think we needed more time (I did, anyway) to analyze the results. This is largely because I had no personal experience with ELL student performance levels. I also feel that the performance level descriptions need to be more carefully written so there are less assumptions about the skills students can do in each level. The MELA-O obviously caused some trouble.

Directions were not always clear. Liz was helpful, but some of the other people who popped in at times seemed to not be completely sure of their responses.

Massachusetts English Proficiency Assessment
2005 Standard Setting Evaluation Form

1. Please mark the grade span for which you set standards.

□ Grade Span 3–4 ☒ Grade Span 5–6 □ Grade Span 7–8 □ Grade Span 9–12

2. What was your comfort level with the standard setting process at the beginning of the process?	Extremely Uncomfortable		Somewhat Comfortable		Extremely Comfortable
	1	2	3	4	5
	3	2	9	4	2
3. What was your comfort level with the standard setting process at the end of the process?	Extremely Uncomfortable		Somewhat Comfortable		Extremely Comfortable
	1	2	3	4	5
	0	1	1	10	8
4. To what extent did the training prepare you to complete the task of standard setting?	Not at all		Somewhat Well		Extremely Well
	1	2	3	4	5
	0	0	3	13	4
5. How clear were the performance level descriptors?	Not at all Clear		Somewhat Clear		Very Clear
	1	2	3	4	5
	0	2	6	8	4
6. How clear was the bookmarking task?	Not at all Clear		Somewhat Clear		Very Clear
	1	2	3	4	5
	0	0	2	7	11
7. To what extent was the length of this meeting appropriate for the task of setting performance standards?	Too Little Time		About Right		Too Much Time
	1	2	3	4	5
	0	1	15	3	1

8. How would you characterize the organization of the standard setting session activities?

Disorganized		Somewhat Organized		Extremely Organized
1	2	3	4	5
0	*0*	*2*	*10*	*8*

9. What is your level of confidence in the bookmarks you placed?

Very Low				Very High
1	2	3	4	5
0	*0*	*3*	*9*	*8*

10. How influential were the following factors in determining where you set your bookmarks:

	Not at all Influential		Somewhat Influential		Very Influential
A. The performance level descriptors	1	2	3	4	5
	0	*0*	*4*	*8*	*8*
B. The assessment items	1	2	3	4	5
	0	*0*	*3*	*11*	*6*
C. Other panelists' comments	1	2	3	4	5
	1	*2*	*6*	*8*	*3*
D. My professional experience	1	2	3	4	5
	0	*0*	*1*	*7*	*12*
E. Rater feedback data	1	2	3	4	5
	2	*0*	*9*	*7*	*2*
F. Item difficulty statistics	1	2	3	4	5
	1	*3*	*8*	*5*	*3*

11. What confidence do you have in the classification of standards at the Beginning level?

Very Low				Very High
1	2	3	4	5
0	0	2	7	11

12. What confidence do you have in the classification of standards at the Early Intermediate level?

Very Low				Very High
1	2	3	4	5
0	0	3	9	8

13. What confidence do you have in the classification of standards at the Intermediate level?

Very Low				Very High
1	2	3	4	5
0	0	3	12	5

14. What confidence do you have in the classification of standards at the Transitioning level?

Very Low				Very High
1	2	3	4	5
0	0	2	9	8

15. What confidence do you have in the classification of standards across all performance levels?

Very Low				Very High
1	2	3	4	5
0	0	1	14	5

16. Please use the space below to provide comments about the standard setting process and/or suggestions as to how the process could be improved.

Thank you for your hard work and valuable feedback!

RESPONSES

As a reg. ed teacher I was extremely impressed by the level of expertise and dedication shown by the panel members. Their students are well served by their commitment. On a personal level, it forced me to listen and ask many questions. I was forced out of my comfort zone and had to fight through the process. I learned a great deal and appreciate the opportunity to learn from others. Thank you.

I'm a special ed teacher who serves many children about half second (ESL) and 1 or 2 non-English speaking students. At first this was very difficult. I'd no experience. I've learned a tremendous amount these three days and appreciate it very much. Thank you.

This was an extremely interesting experience. It gave the opportunity to analyze the items more closely and to see the bigger picture of students' performance across the state.

The last stage of standard setting procedure was very exhausting, not because of decision making, but because of different item numbers for the same test questions. It took a lot of time looking for the test item on the binder to compare the expected level of difficulty with actual results of students' performance. If I didn't have extra cup of coffee, I may have given up reconsidering my decision one more time. I understand it's a lot of work to prepare this event, but please consider that the last stage of this process requires the most convenience in using given materials. In general, I truly appreciate the amount of work you put into this, I learned a lot! Thank you.

I would have liked more independent thinking time at the beginning of brainstorming / reflecting on proficiency indicators (EI; I; T). Maybe we could have been asked to bring in samples of student work at each level? It would have been beneficial to have more days, but I know this is difficult with teachers and substitutes. Overall, it was a positive, learning experience. Thank you for your hard work and patience.

Students with English as a foreign language instruction in their country of origin might be able to read and write but have low MELA-O levels. Students with gaps in their education might have high

MELA-O levels, but not be readers and writers. The sequencing of MELA-O levels in the binder does not follow the language acquisition process. high levels of comprehension and fluency followed by level 1–2 vocabulary and grammar ratings don't make sense. How are teachers (QMA's) state-wide using the MELA-O matrix? Do they understand what they are rating?

I think it was well run and I learned a lot from the process.

Do not include MELA-O in the reading / writing items. They were distracting from the task at hand. Make sure the test administrators manual is accurately reflected in the binder of ordered items. (When does the administrator read the item to the students?)

Setting norms would have been helpful at the beginning. (there was a lot of sidebar conversing going on). This was a great experience!

Why weren't the standards set first – if this is a standards-based assessment don't we want to set the test standards matched to ELPBO, then give the test. Concerned that data may be flawed. Students who weren't truly LEP, (i.e. hadn't been determined to be LEP by a placement test) were participating. Also, teachers wrongly assigned Int. students to sessions 1 and 2.

MELA-O would have liked to do a separate book marking for these. Having them folded in was like mixing apples and oranges. "Examiner says" I had to redo several "yes" "no" decisions based on this error.

It would have been more helpful to have had the standard number on the Item Map, when using it. Too much time flipping back and forth. Filling in the item map was very long and tedious, especially the comments for why was this item more difficult than the previous one? Having so many items to write about. Otherwise, happy to have been a part of this.

I enjoyed the process of discussing professional issues with qualified colleagues. I found the process of standard setting valuable because it forces you to focus on the different levels of language development. Good Job!!

The process was a learning experience for me. The trainers were clear about the purpose and about the task at hand. I would bring back to my district a positive message about the thoroughness of the process and the dedication of teachers in this field!

Well done! Please re-do the MELA-O training tape!

I think that this process should come before or with the writing of the MEPA test. Standards should be set before writing the test. Why are we doing this backwards? Is there a scope and sequence for ESL? There should be!!

I appreciate being part of this committee. My group was very interested in the process and serious about the task. Sharman was extremely helpful and patient with our group. Thanks for a great couple of days! Very helpful.

Massachusetts English Proficiency Assessment
2005 Standard Setting Evaluation Form

1. Please mark the grade span for which you set standards.

❑ Grade Span 3-4 ❑ Grade Span 5-6 ☒ Grade Span 7-8 ❑ Grade Span 9-12

2. What was your comfort level with the standard setting process at the beginning of the process?

Extremely Uncomfortable		Somewhat Comfortable		Extremely Comfortable
1	2	3	4	5
0	3	10	3	2

3. What was your comfort level with the standard setting process at the end of the process?

Extremely Uncomfortable		Somewhat Comfortable		Extremely Comfortable
1	2	3	4	5
1	1	2	9	4

4. To what extent did the training prepare you to complete the task of standard setting?

Not at all		Somewhat Well		Extremely Well
1	2	3	4	5
0	1	4	10	3

5. How clear were the performance level descriptors?

Not at all Clear		Somewhat Clear		Very Clear
1	2	3	4	5
1	1	7	7	2

6. How clear was the bookmarking task?

Not at all Clear		Somewhat Clear		Very Clear
1	2	3	4	5
0	2	3	3	10

7. To what extent was the length of this meeting appropriate for the task of setting performance standards?

Too Little Time		About Right		Too Much Time
1	2	3	4	5
0	2	12	3	1

8. How would you characterize the organization of the standard setting session activities?

Disorganized		Somewhat Organized		Extremely Organized
1	2	3	4	5
0	*1*	*2*	*7*	*8*

9. What is your level of confidence in the bookmarks you placed?

Very Low				Very High
1	2	3	4	5
0	*2*	*2*	*9*	*4*

10. How influential were the following factors in determining where you set your bookmarks:

	Not at all Influential		Somewhat Influential		Very Influential
A. The performance level descriptors	1	2	3	4	5
	1	*1*	*5*	*7*	*4*
B. The assessment items	1	2	3	4	5
	0	*2*	*2*	*7*	*6*
C. Other panelists' comments	1	2	3	4	5
	0	*1*	*3*	*7*	*7*
D. My professional experience	1	2	3	4	5
	0	*0*	*1*	*9*	*8*
E. Rater feedback data	1	2	3	4	5
	1	*2*	*6*	*8*	*1*
F. Item difficulty statistics	1	2	3	4	5
	0	*0*	*7*	*8*	*3*

11. What confidence do you have in the classification of standards at the Beginning level?

Very Low				Very High
1	2	3	4	5
0	*1*	*4*	*6*	*7*

12. What confidence do you have in the classification of standards at the Early Intermediate level?

Very Low				Very High
1	2	3	4	5
0	*2*	*2*	*8*	*6*

13. What confidence do you have in the classification of standards at the Intermediate level?

Very Low				Very High
1	2	3	4	5
0	*1*	*2*	*10*	*5*

14. What confidence do you have in the classification of standards at the Transitioning level?

Very Low				Very High
1	2	3	4	5
1	*1*	*5*	*8*	*3*

15. What confidence do you have in the classification of standards across all performance levels?

Very Low				Very High
1	2	3	4	5
0	*1*	*5*	*7*	*5*

16. Please use the space below to provide comments about the standard setting process and/or suggestions as to how the process could be improved.

____

____

____

____

____

____

Thank you for your hard work and valuable feedback!

RESPONSES

For the writing we should have seen the actual student's work and placed in categories. Question the validity of book marking considering the domains of reading, writing, listening, and speaking together. They should have been done separately. No clarity on descriptions. I felt rushed through the process of identifying what a student was capable of doing at each level. It should have been flushed out further before this meeting.

More time should have been provided for performance indicators. I don’t think it was necessary to go through all 117 items. The MELA-O related pages should have been treated separately so there wouldn't be so much confusion in placing the book marks.

I thought it a very difficult process. I had done standard setting before and I believe that if the item map had a space to rate each item, beginning, early intermediate, intermediate, transitions. It would have been easier.

Integrating the MELA-O levels without any context was confusing. E.g.. How can an early intermediate student be ranked a 5 in comprehension, able to participate in classroom discussions which at the 7/8 level involves academic content.

MELA-O should not be part of the process. Should be separate as it only confuses the issue. Doing the item map should not be left to the end of the day – too much to do, would be better broken up with break in between or earlier in the day.

The movement from ELBOW to descriptors may have resulted in a discrepancy because the ELBOW were mastery objectives. Descriptors were "beginning borderline" categories. Just keeping that fact in mind was difficult in determining descriptors and consequently book marking. There's a gap in perhaps not so much the understanding of the difference between ELBOW and descriptors, but the determinations of borderline students.

MELA-O items should not have been mixed in with the MEPA items. These made it difficult to compare the items. The two tests should be considered completely different indications of proficiency, considering the wide variation between listening and speaking vs. reading and writing proficiency! The process of flushing out the performance level descriptors seems flawed, since these were not used as much, I think, a the intention was.

Perhaps this was not the place to do it, but there needs to be provision for including classroom teachers in the development of policy regarding these standards. I am very uncomfortable with setting standards when it is so unclear how the standards will be used. This is why I feel so strongly that classroom teachers, who deal with the implementation and result of these scores, should be more involved in the process. I do not think this panel answered this need.

More time was needed to review the descriptors set by the group.

The organization was very good. The facility was excellent. Carolyn was exceptional with her clarification skills. Setting descriptors was a great way to begin as it provided scheme on which to judge the questions. Even though I enjoyed the fast pace and the accomplishment, one more day might have been beneficial.

We could have done a better job with our "room" holistic, entry lever performance level descriptors. The MELA-O caused many panelists to be confused. The primary session provided a great overview, room leader was great facilitator. DOE was accessible and responsive to all queries.

Excellent facilitation. More stress over every task on importance and future use of descriptors in process so group is pleases with it's descriptors. Thank you. Learned so much.

Use of MEPA test taking on the first night was unnecessary. Better examples of expectations could have been given to mark performance levels at the beginning of the conference. These would not have influenced out comes.

The MELA-O ratings were a major distraction because of clustering in ways that can not relate to the acquisition of language. The first round of MEPAs tested everyone, including probable test-outs so that skews data. Inconsistent definition about transitioning, "cusp" of descriptors for book marks, differences between large and small school programs for ELLs etc. all confuse our understanding of skills and definitions. Finally, the transitioning understanding of ELLs and main streaming in the age of NCLB and MCAS makes ESL teachers nervous about these linear markers on a cyclical process of language acquisition and knowledge acquisition. Thanks for the experience.

Everything was great! We were well cared for!!!

Separating the MELA-O component would make it a much easier process. Too many people were confused when trying to put that in with reading and writing. Overall, a great experience.

The MELA-O scores caused a lot of confusion when attempted to place bookmarks. I do not think that filling in the item map for 117 items was helpful.

I think facilitators need to be a bit more forceful to keep folks on task. We were much better the last day. The second day was quite bad due to people going off topic and not understanding the task.

Massachusetts English Proficiency Assessment
2005 Standard Setting Evaluation Form

1. Please mark the grade span for which you set standards.

□ Grade Span 3-4   □ Grade Span 5-6   □ Grade Span 7-8   ☒ Grade Span 9-12

2. What was your comfort level with the standard setting process at the beginning of the process?

Extremely Uncomfortable		Somewhat Comfortable		Extremely Comfortable
1	2	3	4	5
0	4	8	3	5

3. What was your comfort level with the standard setting process at the end of the process?

Extremely Uncomfortable		Somewhat Comfortable		Extremely Comfortable
1	2	3	4	5
0	0	2	4	14

4. To what extent did the training prepare you to complete the task of standard setting?

Not at all		Somewhat Well		Extremely Well
1	2	3	4	5
0	0	3	8	9

5. How clear were the performance level descriptors?

Not at all Clear		Somewhat Clear		Very Clear
1	2	3	4	5
0	1	4	10	5

6. How clear was the bookmarking task?

Not at all Clear		Somewhat Clear		Very Clear
1	2	3	4	5
0	0	0	9	11

7. To what extent was the length of this meeting appropriate for the task of setting performance standards?

Too Little Time		About Right		Too Much Time
1	2	3	4	5
1	2	9	4	4

8. How would you characterize the organization of the standard setting session activities?

Disorganized		Somewhat Organized		Extremely Organized
1	2	3	4	5
0	***0***	***3***	***5***	***12***

9. What is your level of confidence in the bookmarks you placed?

Very Low				Very High
1	2	3	4	5
0	***0***	***1***	***7***	***12***

10. How influential were the following factors in determining where you set your bookmarks:

	Not at all Influential		Somewhat Influential		Very Influential
A. The performance level descriptors	1	2	3	4	5
	0	***1***	***4***	***10***	***5***
B. The assessment items	1	2	3	4	5
	0	***0***	***3***	***8***	***9***
C. Other panelists' comments	1	2	3	4	5
	0	***0***	***3***	***11***	***6***
D. My professional experience	1	2	3	4	5
	0	***0***	***3***	***6***	***11***
E. Rater feedback data	1	2	3	4	5
	0	***1***	***1***	***9***	***8***
F. Item difficulty statistics	1	2	3	4	5
	0	***1***	***5***	***8***	***6***

11. What confidence do you have in the classification of standards at the Beginning level?

Very Low				Very High
1	2	3	4	5
0	1	2	8	9

12. What confidence do you have in the classification of standards at the Early Intermediate level?

Very Low				Very High
1	2	3	4	5
0	1	1	11	7

13. What confidence do you have in the classification of standards at the Intermediate level?

Very Low				Very High
1	2	3	4	5
0	1	1	12	6

14. What confidence do you have in the classification of standards at the Transitioning level?

Very Low				Very High
1	2	3	4	5
0	1	1	7	11

15. What confidence do you have in the classification of standards across all performance levels?

Very Low				Very High
1	2	3	4	5
0	1	0	13	6

16. Please use the space below to provide comments about the standard setting process and/or suggestions as to how the process could be improved.

Thank you for your hard work and valuable feedback!

RESPONSES

Most important: all involved should be familiar with SLA or at least English teachers. More time to go over test. Especially some familiarity with MELA-O.

The process informative: Not real happy with the Measured Progress facilitator. I found her very glib and almost condescending. Although she had a job to do, we are professionals and not children. We should not be treated as a "class" of kids, which is how I often felt when she tried to bring us together as a group. Somewhat insulting, if you will!

I feel that this entire process was informative and quite productive. The span / range of people in the group was wonderful for the task at hand. I feel that we could have used some more time during the item mapping activity. I am very confident in the answers I provided.

It was well done, as far as the process went, but the fact that it was all based on the MEPA, one must assume the MEPA was constructed effectively. Some questions did not test what they were designed to test, however. Work on directions for the test itself.

I think everything was very well organized.

More table space needed for each person.

Some of the descriptions of scoring were more helpful, more descriptive than others.

This was very interesting and I learned so much from this! Thank you.

Less intro material and right to the activity of looking at test. More spaces so participants know what to expect on agenda. Room too small. Table conversation hard to hear.

Shorter session (2 days) could have worked for this exercise. The M.P. people were excellent and most helpful and positive. Kudos to Dona. Kit Viator is the Best!!! Clear, concise, courteous, appreciative, empathetic, well-informed, a great people person.

Teachers tended to confuse the purpose of the test with particular students, underestimating student capacity, I sometimes felt. A little more talk about tests, or the purpose of this one in particular, might have been helpful.

Feedback on forms – Layout: Test booklet items are vertical. Answer bubble are horizontal. Test book add: you may work in test book.

Problems: The "apples and oranges" issue of MELA-O and MEPA R/W. General performance level definitions. Need to include proficiency at test-taking strategies. Need to include reading comprehension strategies other that referencing. I counted eight other strategies that were tested. Language needs to be more specific: "most common words"? What does "recognize" mean here? To Measured Progress: I don't have much confidence that the creators of the test thoroughly grasp all reading comprehension strategies and the various communicative functions of language beyond simple referential meaning.