MATCHING THE JUDGMENTAL TASK WITH STANDARD SETTING PANELIST EXPERTISE: THE ITEM-DESCRIPTOR (ID) MATCHING METHOD

Steve Ferrara
CTB/McGraw-Hill

Marianne Perie
National Center for the Improvement of Educational Assessment

Eugene Johnson
Independent Consultant

Earlier versions of this paper were presented at the 2002 annual meeting of the National Council on Measurement in Education, New Orleans, and by invitation to the Board on Testing and Assessment of the National Research Council, 2003, Washington, DC. Correspondence regarding this article should be addressed to steve_ferrara@ctb.com.

We appreciate contributions of numerous colleagues to the refinement and description of this method.

MATCHING THE JUDGMENTAL TASK WITH STANDARD SETTING PANELIST EXPERTISE: THE ITEM-DESCRIPTOR (ID) MATCHING METHOD

ABSTRACT

Psychometricians continue to introduce new approaches to setting cut scores for educational assessments in an attempt to improve on current methods. In this paper we describe the Item-Descriptor (ID) Matching method, a method based on IRT item mapping. In ID Matching, test content area experts match items (i.e., their judgments about the knowledge and skills required to respond to an item) to the knowledge and skills described in performance level descriptors that are used for reporting test results. We argue that the cognitive-judgmental task of matching item response requirements to performance level descriptors is aligned closely with the experience and expertise of standard setting panelists, who are typically classroom teachers and other content area experts. Unlike other popular standard setting methods, ID Matching does not require panelists to make error-prone probability judgments, predict student performance, or imagine examinees who are just barely in a performance level. We describe applications of ID Matching in two educational testing programs and provide evidence of the effectiveness of this method. The entire process is described in the first section of the paper. Subsequent sections describe applications of ID Matching for two operational testing programs.

MATCHING THE JUDGMENTAL TASK WITH STANDARD SETTING PANELIST EXPERTISE: THE ITEM-DESCRIPTOR (ID) MATCHING METHOD

INTRODUCTION

Psychometricians have introduced a range of new approaches to standard setting in the 25 years since Glass published his landmark article on setting performance standards for criterion-referenced tests (Glass, 1978). These new approaches include, for example, the Body of Work, Bookmark, the Item-Mapping standard setting method (Wang, 2003), Mapmark, and Policy-Capturing methods (see Cizek and Bunch, 2007, for descriptions and evaluations). This proliferation of methods has at least two explanations: First, there are no true or correct cut scores for a test, only more or less defensible ones; defensibility is based in large measure on the method used to set standards. Second, there is no one best or correct method for setting standards but rather a range of approaches that may be more or less appropriate for a specific situation.

These new approaches can be classified in the same two categories that existed 25 years ago: test-based and examinee-based approaches (e.g., Kane, 2001). Other distinctions have emerged: test-based approaches that focus on individual items (e.g., modified Angoff), test-based approaches that focus on patterns of subscores (e.g., the dominant profile method), examinee-centered approaches that focus on examinees' products (e.g., the Body of Work method), and those that focus on groups of examinees (e.g., the contrasting groups method). (See Hambleton & Pitoniak, 2006, p. 433 and table 12.1 for one taxonomy of standard setting methods and Zieky, Perie, & Livingston, 2007, chap. 4 for another characterization of methods.)

Just as no single approach to standard setting can be considered best for all situations, no single approach is preferred by all measurement specialists. The modified Angoff method[1] often is purported to be the most widely used method for setting cut scores on certification, licensure, and educational achievement tests. Some measurement specialists prefer the modified Angoff method because it does not require the use of IRT machinery or performance data. However, the modified Angoff method is less efficient for standard setting panelists because it requires them to make judgments about every item in a test form for each cut score to be set (e.g., Cizek & Bunch, 2007, p. 159). Other psychometricians prefer the Bookmark method, an approach that is based conceptually and procedurally on item response theory (IRT) item parameters. They may choose this approach so that a test's psychometric framework is entirely IRT-based. The Bookmark method has been used to set cut scores in 31 states (Perie, 2005). Recent research and debate suggests that the Bookmark method may yield scores that are lower than those (a) intended by panelists (Karantonis & Sireci, 2006, p. 8; Reckase, 2006a; see Reckase, 2006b and Schulz, 2006 for commentary and a rejoinder), and (b) produced from other methods (Karantonis & Sireci, 2006, p. 8; Green, Trimble, & Lewis, 2003, p. 26).[2]

Educational measurement specialists continue to adapt and refine existing methods and to spin off new approaches to setting performance standards. In this paper we describe Item-Descriptor Matching (ID Matching), another approach to setting cut scores. Its name describes both the procedure that standard setting panelists follow in recommending cut scores and their cognitive-judgmental task. We provide an overview of the ID Matching method and its origins, compare and contrast it with the modified Angoff and Bookmark methods, give details on the ID Matching method and optional variations, and describe results from applying the ID Matching method in two assessment programs. We compare and contrast the ID Matching method with the modified Angoff and Bookmark methods to highlight relative advantages and disadvantages of all three methods.

[1] All implementations of the Angoff method are, by definition, modified. In the traditional application of a modified Angoff method, standard setting panelists estimate the proportion or percentage of borderline examinees they expect to respond correctly to a multiple-choice item.

[2] Results from both the Bookmark and ID Matching methods may be influenced by the response probability (RP) criterion used to order items in ordered item books; see Karantonis and Sireci (2006, p. 8).

Overview of the Item-Descriptor (ID) Matching Method

In the ID Matching method, test items are arranged in ordered item books, usually starting with the least difficult item and continuing to the most difficult item, where item difficulty is based on the IRT scale location (i.e., difficulty or b parameter). Standard setting panelists examine each item (and the accompanying rubric for constructed-response items); determine the content knowledge, skills, and cognitive processes that each item requires (i.e., item response requirements); and then match those requirements to performance level descriptors. The performance level descriptors define performance standards on the test for which panelists will recommend cut scores. As panelists match items and descriptors, sequences of items emerge in which items in one sequence match more closely one performance level descriptor while items in the next sequence match more closely the next higher adjacent performance level descriptor. Typically, a third sequence of items is identified, between these two sequences, in which items alternate between matching each of the adjacent performance level descriptors. The *threshold region* is defined by this alternating pattern of matches between two sequences of clearly matching items. A cut score typically is placed in the threshold region. In subsequent rounds of matching item response requirements and performance level descriptors, panelists adjust cut scores by determining blocks of items (i.e., as opposed to individual items) that most closely match performance level descriptors.

Unlike other standard setting methods, ID Matching does not require panelists to make judgments about the probability that a student will respond successfully to an item (or item score level). Research over several decades in judgment and decision making is clear: humans can be trained to estimate probabilities “moderately well” (Nickerson, 2004, p. 433), but are susceptible to judgmental biases and are prone to making errors when judging the probability of an occurrence (Nickerson, 2004, chap. 11; Plous, 1993, p. 144). In addition, ID Matching does not require panelists to conceptualize an imaginary student who is just barely Proficient (or just barely in other performance levels). Instead, ID Matching requires standard setting panelists to match knowledge and skill requirements of items (and item score levels) with the descriptions of knowledge and skills in performance level descriptors. In addition, in ID Matching, the response probability criterion is directly relevant only in the item scaling process, not in the instructions to panelists. This simplifies the cognitive complexity of the panelists’ judgmental task, relative to the Bookmark method. In ID Matching, panelists can focus on matching the knowledge and skill requirements of each item to the knowledge and skills articulated in performance level descriptors.

In a typical ID Matching workshop, panelists review the response demands of each item (i.e., the content area knowledge and skills required to respond to items) in an ordered item book[3] and match those demands to the knowledge and skill descriptions in the performance level descriptors. Panelists (a) determine which performance level descriptor most closely matches the response demands of each item or (b) indicate that the item is in the threshold region between two adjacent levels. Using the performance levels Basic, Proficient, and Advanced to illustrate, panelists complete a recording sheet, idealized in Figure 1.

Panelists match items to a performance level descriptor only when they feel that the match is clear; otherwise, they indicate that the item is in the threshold region (described below) between adjacent levels. Panelists may place an item in a threshold region because it does not clearly match one performance level descriptor or because they are not sure which descriptor it matches. Threshold regions also are defined when panelists match a sequence of items to adjacent performance level descriptors in alternating fashion. Either panelists or psychometricians locate cut scores in threshold regions.

[3] In an ordered-item book, items are arranged in ascending order of difficulty.

Figure 1

Illustration of item-descriptor matches.

Item in an Ordered Item Book	Performance Level Descriptor to Which Item Is Matched
1	Basic
2	Basic
3	Basic
4	Basic
5	Threshold region
6	Threshold region
7	Proficient
8	Proficient
9	Proficient
10	Threshold region
11	Threshold region
12	Threshold region
13	Advanced
14	Advanced

ID Matching shares features with other standard setting methods, specifically modified Angoff and Bookmark methods. Typical applications of these methods require panelists to make judgments about items to identify a cut score, involve two or three rounds of judgments, require or employ performance level descriptors as the basis for making judgments about items, and usually provide impact data to panelists. The ID Matching method is distinctive in two ways:

1. It captures information about panelists' thinking. Illuminating sequences of items that clearly match performance level descriptors and sequences of items that represent the threshold region provides a focus for panelist discussions. This focus facilitates common understandings about item response requirements and matches to performance level descriptors and facilitates convergence of judgments.

2. The cognitive task of matching item response requirements to performance level descriptors appears to be aligned closely with the experience and expertise of standard setting panelists, who are typically classroom teachers and other content area experts. Unlike the modified Angoff and Bookmark approaches, ID Matching does not require panelists to make judgments about the probability that examinees will answer an item correctly or about a pair of items between which the probability of a correct response changes. It also does not require panelists to envision imaginary students who are just barely in a performance level.

Origins and Evolution of the ID Matching Method

The ID Matching method evolved from procedures developed jointly by the Maryland State Department of Education and CTB McGraw-Hill to establish a score reporting system for the first administration of the Maryland School Performance Program (MSPAP) in 1991. An early version of the ID Matching method emerged in 1993 when staff of the Maryland State Department of Education and Westat jointly developed procedures for setting additional cut scores for the 1992 administration of MSPAP. Staff of the American Institutes for Research (AIR) further refined the ID Matching method and procedures for setting standards in 2000 for high school end-of-course examinations for the School District of Philadelphia; student achievement and school principal certification tests in Bahia, Brazil, in 2002; and alternate assessments for students with significant cognitive disabilities in New Mexico and South Carolina in 2007. ID Matching has been used in demonstrations and pilot tests for other assessments in Georgia and South Carolina. It has been used in operational testing programs in Chicago and New Jersey. ID Matching also was used outside of K–12 education to establish cut scores for a multistage computer-adaptive test of adult mathematics proficiency in 2007 (Sireci, Baldwin, Martone, & Han, 2007). It is gaining interest among specialists in standard setting and is described elsewhere (see Cizek & Bunch, 2007, chap. 11; Zieky, Perie, & Livingston, in press).

Comparisons of ID Matching With Other Commonly Used Standard Setting Methods

ID Matching is an item mapping standard setting method (see Zwick, Senturk, Wang, & Loomis, 2001) as are the Bookmark, Mapmark, and the Item-Mapping (Wang, 2003) standard setting methods. In this paper, we compare ID Matching with the modified Angoff and Bookmark methods because they are widely used in large-scale and statewide testing programs. Because ID Matching and Bookmark both are item mapping methods, they share similar features. Item mapping methods, in general, share commonalties that distinguish them from item rating methods such as the modified Angoff. Below, we summarize similarities and differences between ID Matching and the Bookmark and modified Angoff methods.

. 2 summarizes the similarities and differences among the IDM, Bookmark, and Angoff methods according to seven key considerations. These considerations include practical factors (e.g., materials, resources, and data needs), applicability to different item formats, cognitive complexity, and relative advantages and concerns.

Figure 2

Comparison of ID Matching, Bookmark, and Angoff methods.

Key Considerations	ID Matching	Bookmark	Modified Angoff
Materials and information required or typically used	Ordered item book from an intact test form or an item bank Performance level descriptors for all performance levels, including the level below the lowest cut score Item map (optional)	Ordered item book from an intact test form or an item bank Performance level descriptors Item map (optional)	Items from a fixed test form (not ordered) Performance level descriptors
Data required	Item difficulty values (IRT or classical) Impact data (optional)	Item difficulty values (IRT or classical) Impact data (optional)	Classical p values (optional) Impact data (optional)
Item and test formats to which method is directly applicable	Dichotomously and polytomously scored items Intact test forms or items sampled from an item bank Computer adaptive tests	Dichotomously and polytomously scored items Intact test forms or items sampled from an item bank Computer adaptive tests	Dichotomously and polytomously scored items Intact test forms
Cognitive judgmental task	Match the knowledge and skills required to respond successfully to each item to the knowledge and skills described in one of the performance level descriptors	Place the bookmark on the page in the ordered item book where students who are just barely in the performance level would be able to respond successfully	Determine the percentage of borderline students at each performance level who would answer this item correctly

(continued)

Figure 2 (continued)

Typical feedback data on panelist judgments and recommended cut scores	Disagreement on item-descriptor matches (after the first round) Pages in threshold regions between performance levels High, low, and median cut score pages at each table (after round 1) High, low, and median cut score pages across all panelists (after round 1 or 2)	High, low, and median bookmark placements at each table (after round 1) High, low, and median bookmark placements across all panelists (after round 1 or 2)	Indicators of differences in panelist percentages across items (e.g., mean and range of ratings) p values (optional)
Relative advantages	Judgmental task similar to judgments that teachers make as part of teaching-learning process Does not rely on a response probability, imagining students who are just barely in a performance level, or estimates of students at each performance level Captures information on items that clearly match performance level descriptors and items that are in thresholds between levels Examining the item map and ordered item book provides useful information to take back to the classroom for panelists who are teachers	Well-known method used in the majority of state assessment programs Combines psychometric information on item difficulty with expert judgments Focusing on pages in OIB enables focused discussion and quicker convergence among panelists Examining the item map and OIB provides useful information to take back to the classroom for panelists who are teachers	Supported by the largest amount of research of all methods for setting cut scores Does not require student data, so it can be applied prior to test administration Many modifications exist so that the essential cognitive judgment can be applied to almost any item type
Concerns	A new method that has been used in a limited numbers of states and other assessment programs No research on panelist thinking and judgments	Little empirical research on this method	Has been criticized in some applications.

Differences between the ID Matching and Modified Angoff Methods

ID Matching differs from modified Angoff in three critical areas: the tools it uses in the standard setting workshop, the judgment required of panelists, and the discussions it encourages through feedback.

Like the Bookmark method, ID Matching requires that test items be re-sorted by item difficulty. Items are presented to panelists in order from easiest to hardest in an item map that provides important information on each item, such as the correct answer, the scale location, and the content strand it measures. Using tools such as an ordered item book (OIB) and an item map has the advantage of providing panelists with information on student performance at the start of the process. In addition, the tools have a side benefit of providing the panelists, usually teachers, with information on which items students tend to answer correctly and which they do not. Panelists are able to determine which content sub-areas appear easier than others and which item formats tend to be more problematic for students. A disadvantage to using these tools is that they require student performance data. The modified Angoff method can be implemented at any point after an operational test form has been assembled. ID Matching requires scaled test data from which the items can be sorted. Standard setting cannot begin until the test has been scored, IRT calibration has been completed, and ordered test books have been assembled. Thus, when a testing program is under tight deadlines and must produce a cut score immediately after the field test, or even before administering a field test, modified Angoff may be the preferred method. If, however, some flexibility in timing is available, ID Matching has other advantages that may make it preferable to modified Angoff.

One possible advantage of ID Matching over the modified Angoff method is the cognitive load placed on the panelists. Even in assessments, such as NAEP, that have removed the cognitive burden of envisioning a borderline student by providing written descriptions of what borderline students can be expected to know and be able to do, the modified Angoff method requires panelists to make judgments in terms of student performance. Panelists predict the probability of a successful response for each item for students on the borderline of a performance level. Some research indicates that panelists have difficulty predicting the performance of borderline students (National Academy of Education, 1997). ID Matching appears to be a simpler task. Panelists decide only whether the knowledge, skills, and cognitive processes required to answer an item successfully match the knowledge, skills, and cognitive processes required at each performance level as detailed in the performance level descriptor. Panelists do not need to predict percentage correct (content experts tend not to judge item difficulty accurately; see Impara & Plake, 1998) or think in terms of probabilities during the process. The use of numbers is minimized in ID Matching, thus simplifying the task for non-psychometricians.

Finally, the feedback data provided in ID Matching workshops focus on the group of items matched to a performance level descriptor, encouraging panelists to focus on exactly what knowledge, skills, and cognitive processes each item requires and what a student must demonstrate to reach each performance level. In modified Angoff procedures, discussions revolve around the predicted percentage correct for each item. Discussions may also center on the average cut score across all panelists, but the discussions rarely converge on the relationship between what students must know and be able to do to reach a performance level and the knowledge and skills elicited by each item. Thus, ID Matching appears to facilitate more focused discussions and may result in greater convergence in judgments on individual items.

Differences between the ID Matching and Bookmark Methods

As discussed earlier, because they are both item mapping methods, ID Matching and Bookmark perhaps are more similar than different. The primary difference is in the cognitive-judgmental task. We argue that ID Matching requires panelists to engage in a simpler task, one that allows them, as content experts, to operate in the realm of their expertise.

ID Matching focuses the task on the performance level descriptor and the response requirements of each item. Using their knowledge of the content area and their familiarity with students, teaching, and learning, panelists determine the item response requirements and match them to the performance level descriptors. Panelists never are asked to envision a hypothetical examinee or to think in terms of a two-thirds probability. All the information they need is provided in the ordered item book, the item map, and the performance level descriptor, and all the understanding they need to complete the task comes from their experiences with the content area.

ID Matching also simplifies the process for panelists because it is a logical step-by-step process, involving first matching the items, looking for areas of transition from one performance level to another, and then drawing a cut score in that area of transition. An immediate advantage to this approach is that it enables panelists to identify outliers more easily. Because items are ordered on the basis of empirical difficulty, they do not always line up from easiest to hardest in the order that a content expert would expect. A content expert may identify one easy item in the midst of several difficult items or vice versa. Unlike the traditional Bookmark method, in which most of the judgmental task is completed in the panelists' minds, ID Matching requires panelists to write the item-descriptor match next to each item. Thus, in the ID Matching process, panelists can see more easily that the one “Basic” item in the midst of several “Proficient” items is most likely an anomaly rather than the place to identify the cut score. In addition, the written item-descriptor matches provide standard setting workshop leaders, individual panelists, and, ultimately, item writers information about panelists' thinking and judgment.

Another way to think about this task of finding the cut score within a threshold region is to compare it with the range-finding and pinpointing tasks in the Body of Work method (Kingston, Kahl, Sweeney, & Bay, 2001). Although the judgmental tasks in ID Matching and Body of Work differ significantly, the idea of using subsequent tasks to narrow the range in which the cut score is located is similar. In ID Matching, the panelists first identify a threshold region, which may contain as few as two items or may include many items. The second step of the process, then, is to examine this range in which the cut score may be found and then narrow the range to find the one page in the ordered item book where the cognitive demands of the items most closely match the next highest performance level descriptor.

One benefit of ID Matching that goes beyond simplifying the task is the robustness of the final cut score. That is, the ID Matching process allows for instability in item parameters. For example, items in new assessment programs often are calibrated on the basis of field-test data. The resulting item parameters may vary somewhat from those derived from calibrations that are based on operational data because of motivation effects on students (i.e., no consequences versus high stakes) and on teachers (i.e., better aligning classroom instruction to test content). ID Matching allows panelists to match items to descriptions without requiring the matches to follow a strict sequential order. In other words, panelists may match one item to the basic level, the next item to the proficient level, and then the following item back to the basic level again. This back-and-forth matching identifies the threshold region. Because the initial cut score is located at the midpoint of the threshold region, drift in the location of individual items may not affect the judgmental process and the location of the cut score as much as it could in the Bookmark method. ID Matching, therefore, may be more robust to instability in item parameter calibrations. This consideration is important for many state assessment programs because time constraints often require using field-test data to set performance standards.[4]

ID MATCHING: MATERIALS AND PROCEDURAL DETAILS

To set standards using ID Matching, panelists:

- Work with an OIB and performance level descriptors.
- Determine the knowledge and skills required to respond to each item by identifying knowledge and skill requirements on the basis of their professional judgment.
- Work with an item map that includes item coding information (e.g., targeted content standards), item scale locations, and other relevant information.
- Make two judgments. They (a) match knowledge and skill requirements of items with knowledge and skill requirements in performance level descriptors and (b) locate cut scores in threshold regions.

Panelists record item-descriptor matches and cut score pages in spaces on the item map or on a separate recording form.

Ordered Item Books

As mentioned previously, panelists use OIBs, in which the items have been ordered from least to most difficult according to their IRT location (i.e., as opposed to the order in which they appeared in examinee test books).[5] OIBs display one item per page. Each page includes the item text and accompanying graphics, the scoring key for a multiple-choice item, the scoring rubric for a constructed-response item, and sample examinee responses for each constructed-response item. More than one

[4] We intend to explore this benefit in two state testing programs, as we set standards on field-test data using ID Matching, and then re-examine the standards one year later when operational data become available.

[5] It also would be reasonable to order items using classical p-values if IRT difficulty locations are not available or not desired. Readers should note that item orderings will differ, depending on whether items are ordered using 3-parameter IRT model locations, 2-parameter IRT model locations, or classical p-values. Rasch model locations and classical p-values order multiple-choice items equivalently.

page may be required to display all the information for a constructed-response item. Each constructed-response item typically includes a separate set of pages for each possible score on the item (except the score of 0). For example, if the item has possible scores of 0, 1, 2, and 3, there would be three separate sets of pages for that item: one set of pages for a score of 1 on the item, another set of pages for a score of 2, and another set of pages for a score of 3. Each set of pages for the different scores on the same constructed-response item will appear at a different place in the ordered item book. Other items are likely to appear in between the pages for a constructed-response item in the ordered item book.

Performance Level Descriptors

The effectiveness of the ID Matching method depends on the quality of the performance level descriptors. In fact, this is true for virtually all standard setting methods. Currently, few studies exist to provide advice on writing performance level descriptors and little research exists on the effects of performance level descriptors on resulting cut scores. Mills and Jaeger (1998) describe a process for writing performance level descriptors and results from a small study comparing standard setting results using (a) general performance level descriptors based on test specifications and (b) more specific “content-grounded” descriptions (p. 82) based on test items. Standard setting panelists in the study reported that the more specific descriptions were better suited to the standard setting task. It seems plausible that this finding could vary for different tests and different standard setting methods. Other writers (e.g., Hambleton, 2001) assert the importance of clear performance level descriptors to the defensibility of the standard setting process and resulting standards and call for research on the role of performance level descriptors. Hambleton and Pitoniak (2006, pp. 452–453) also describe the research and advice on writing performance level descriptions.

Performance level descriptors serve several roles. During standard setting, they provide guidance to panelists in implementing their judgmental tasks and making decisions about locations for cut scores. They can provide a means for guiding educators, families, and the general public in interpreting students’ performances on an assessment. In fact, performance level descriptors are policy statements for an assessment program; they communicate aspirations for the performance and achievement of the program’s examinees in the content area(s) assessed. Thus, the descriptors should be written by content experts and approved by policymakers.

Item Maps

Traditionally, panelists refer to item maps for the content coding information that accompanies items. In an item map, items are ordered from easiest to most difficult and are presented in rows, one row per page in the ordered item book. The following information may be included in each column:

- Item sequence number, corresponding to the ordered item book
- Item ID number, corresponding to the original location of the item in the test book
- Item type (i.e., multiple-choice or constructed-response)
- Item location (i.e., on the IRT scale, often in the form of a transformed scale score)
- Content area and/or strand targeted by the item

Panelists use this information as they match item response requirements to performance level descriptors. Figure 3 displays an illustrative item map with hypothetical item-descriptor matches. The figure illustrates sequences of clear matches between item response demands and a Below Basic (i.e., BB) performance level descriptor (i.e., ordered item book pages 1–8) and Basic, Proficient, and Advanced performance level descriptors (i.e., ordered item book pages 15–19, 22–32, and 39–46). The shaded sequences represent threshold regions, which we describe in detail below. It also illustrates multiple difficulty locations for a constructed-response item (e.g., original item number 16, score levels 1, 2, and 3 on pages 9, 28, and 36 of the ordered item book).

Figure 3

Illustrative item map with hypothetical item-descriptor matches.

OIB Page #	Item type	Item # on Original Test	Scale Location	Content Strand	Item-Descriptor Matches
1	MC	9	462	Data, Statistics, and Probability	BB
2	MC	19	464	Number Systems	BB
3	MC	24	468	Measurement	BB
4	MC	13	482	Geometry	BB
5	CR	35.1	482	Data, Statistics, and Probability	BB
6	MC	27	482	Patterns, Algebra, and Functions	BB
7	MC	20	487	Number Systems	BB
8	MC	34	490	Patterns, Algebra, and Functions	BB
9	CR	16.1	492	Measurement	B
10	MC	10	496	Number Systems	BB
11	MC	32	497	Geometry	B
12	CR	25.1	497	Patterns, Algebra, and Functions	B
13	CR	36.1	499	Data, Statistics, and Probability	BB
14	MC	2	499	Number Systems	BB
15	MC	21	500	Data, Statistics, and Probability	B
16	MC	31	500	Measurement	B
17	MC	4	500	Number Systems	B
18	CR	6.1	501	Patterns, Algebra, and Functions	B
19	CR	15.1	502	Measurement	B
20	MC	14	504	Geometry	P
21	CR	5.2	505	Geometry	B
22	MC	8	507	Number Systems	P
23	MC	1	507	Patterns, Algebra, and Functions	P
24	MC	23	507	Number Systems	P
25	CR	26.1	508	Geometry	P
26	CR	35.2	508	Data, Statistics, and Probability	P
27	MC	30	509	Number Systems	P
28	CR	16.2	509	Measurement	P
29	MC	7	513	Data, Statistics, and Probability	P
30	CR	6.2	513	Patterns, Algebra, and Functions	P
31	CR	15.2	513	Measurement	P
32	CR	5.3	514	Geometry	P
33	MC	3	515	Patterns, Algebra, and Functions	A
34	MC	18	515	Measurement	P

(continued)

Figure 3 (continued)

35	CR	35.3	515	Data, Statistics, and Probability	P
36	CR	16.3	515	Measurement	A
37	MC	11	516	Number Systems	A
38	CR	26.2	517	Geometry	P
39	MC	22	517	Patterns, Algebra, and Functions	A
40	CR	25.3	518	Patterns, Algebra, and Functions	A
41	MC	29	518	Geometry	A
42	MC	12	518	Patterns, Algebra, and Functions	A
43	CR	6.3	518	Patterns, Algebra, and Functions	A
44	CR	26.3	518	Geometry	A
45	MC	17	519	Data, Statistics, and Probability	A
46	CR	15.3	520	Measurement	A
BB = Below the Basic performance level, B = Basic, P = Proficient, A = Advanced. Shaded sequences represent threshold regions; see text for an explanation.					

Threshold Regions

Although items are ordered in item maps by difficulty, panelists do not always match them to performance level descriptors in sequential order. Panelists may find an item that matches the description for the basic level, for example, following an item that matches the description for the higher proficient level. Panelists usually produce item-descriptor matches in an item map that look something like Figure 3 In Figure 3, a run of BBs (Below Basic) is followed by a run of non-systematically alternating BBs and Bs (Basic), followed by a run of Bs, and so on. In this illustrative item map in Figure 3, items 1–8 match the performance level descriptor for the Below Basic level, items 15–19 match the performance level descriptor for Basic, items 22–32 match the performance level descriptor for Proficient, and items 39–46 match the performance level descriptor for Advanced. The areas between these sequences are the threshold regions.

Threshold regions represent sequences of items in which the matches between item knowledge and skill demands and the demands in descriptors are not clear. Several factors can account for this lack of clarity, including panelist judgments (e.g., they may not be sure about a match at round 1), peculiarities in the item response demands, test booklet item ordering effects (e.g., context clues), or vagueness in the performance level descriptors. For example, a distractor or external influences can affect the location of an item in the item map. Moreover, performance level descriptors generally do not provide precisely clear distinctions between levels. Thus, we would expect some overlap or general fuzziness in distinguishing the matches at the borderline between item-descriptor matches. We train panelists to expect that items appear in threshold regions for at least three reasons: (a) a sequence of items has matches between items and descriptors that alternate between two adjacent levels, (b) an item’s demands do not clearly match either of two adjacent performance level descriptors, or (c) a panelist is just not yet sure which descriptor the item most closely matches.

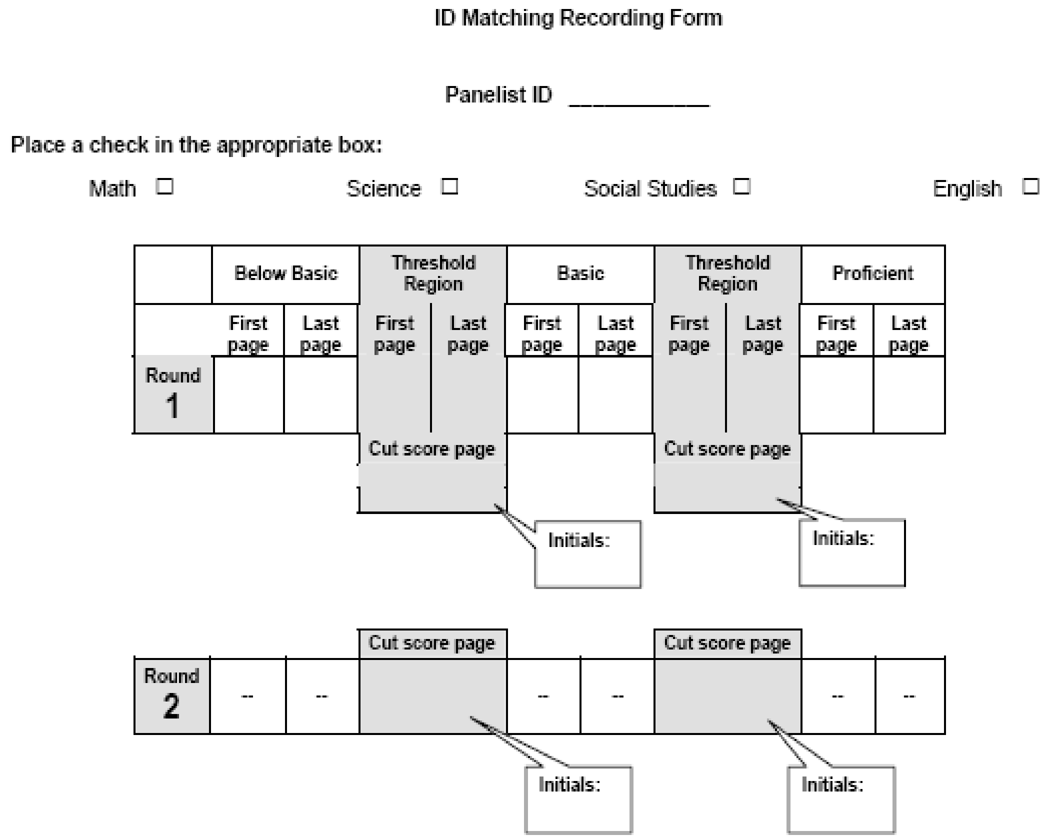
We train panelists to define each threshold region as follows: (a) the first item that matches a higher performance level descriptor, just after a consistent run of matches with a lower performance level descriptor, and (b) the final item just before the first run of three matches to the next higher performance level. In Figure 3, the threshold region between Below Basic and Basic starts with the first B (item sequence number 9) and ends at the item before the first string of three Bs (i.e., item sequence number 14). This rule of using a run of three to define the beginning of a new level is similar to the stopping rule used in individual IQ, achievement, and diagnostic testing.

Recording Item-Descriptor Matches and Identifying Threshold Regions

Panelists record on a recording form the first and last item number (from the OIB) for each sequence of items that matches a performance level descriptor and for each threshold region. Psychometricians enter the cut score item numbers into analysis spreadsheets to calculate median cut scores and table and room reports that are used as feedback at the beginning of subsequent rounds. The recording form is designed to reinforce for panelists the logic of ID Matching: the last ordered item number of a performance level must precede the first ordered item number of the adjacent threshold region, the last item number of the threshold region must precede the item number of the subsequent performance level, and the cut score item number must be in the threshold region (or be the first item number of the subsequent performance level. In a recent application of ID Matching, panelists recorded OIB page numbers on a recording form. Figure 4 is an example of a recording form for a standard setting involving three performance levels set in two rounds for each of four content area assessments. Another option is to have the panelists draw the threshold region directly onto their item map where they have recorded their matches (see Figure 3). They can then draw a second line to indicate where the cut score should fall within the threshold score region. Or, they can do both—that is, first use the item map to record their matches and draw their lines, and then record those line locations onto a recording sheet like the one shown in Figure 4.

Figure 4

Sample ID Matching recording form.



ID Matching Recording Form

Panelist ID ___________

Place a check in the appropriate box:

Math □ Science □ Social Studies □ English □

	Below Basic		Threshold Region		Basic		Threshold Region		Proficient	
	First page	Last page	First page	Last page	First page	Last page	First page	Last page	First page	Last page
Round 1										
			Cut score page				Cut score page			

Initials:

Initials:

			Cut score page			Cut score page		
Round 2	--	--		--	--		--	--

Initials:

Initials:

The Cognitive-Judgmental Matching Task

Panelists use OIBs to determine item response requirements and to match item response requirements to performance level descriptors. They start with the first (easiest) item and determine, using their professional judgment, the item's response requirements. The item's response requirements are the knowledge and skills that examinees must employ to answer correctly each multiple-choice item or to achieve a specific score on a constructed-response item. Next, they examine the description of the knowledge, skills, and cognitive processes that define performance for each performance level. Finally, panelists match the item response requirements to one of the performance level descriptors. They repeat the process for each item. The name "Item-Descriptor Matching" is intended to characterize this cognitive-judgmental task. Panelists are trained to answer the following question for each item-descriptor match:

> Which performance level descriptor most closely matches the knowledge and skills required to respond successfully to this item (or score level for constructed-response items)?
>
> In other words, match knowledge and skill requirements of items to knowledge and skill demands in one of the performance level descriptors.

Because panelists require time to understand and internalize this judgmental task, we display it in PowerPoint slides during training, repeat it often, and ask panelists to state it for their colleagues. Even with regular reinforcement, some panelists may struggle to connect the task statement with the logic of ID Matching. The logic requires that panelists produce sequences of items that match each performance level descriptor and, in most cases, a second set of items between these clearly matched sequences that do not clearly match a performance level descriptors (i.e., threshold regions). We train panelists to understand this logic, using graphical representations (e.g., Figure 3).

Placing Cut Scores in Threshold Regions

In the final step, the cut score is identified within the threshold region. The cut score can be either placed by panelists or calculated by psychometricians. In the example shown in Figure 3, the cut score for the Basic level would fall somewhere between items 9 and 14, the Proficient cut score between items 20 and 21, and the Advanced cut score between items 33 and 38**.**

Several procedures for identifying cut scores have evolved over the course of refining the ID Matching method. The first is to train panelists to use their best judgment to determine where, in the threshold region, the knowledge, skills, and cognitive processes required by the items change and begin more closely to match the description of knowledge, skills, and cognitive processes required just to pass the examination. We train them to place these cut scores by using the following steps: Panelists (a) review the content knowledge and skill demands of all items in their threshold regions and (b) identify the first item in the region whose demands match more closely the demands in the performance level descriptor for the higher of two adjacent levels than the demands in the performance level descriptor for the lower of the two adjacent levels. We recommend this procedure for locating the cut score because panelists (a) have direct judgmental control of the location of cut scores, and (b) can adjust cut scores directly as they reconsider matches between item demands and performance level descriptors in subsequent rounds.

Other options for locating cut scores are available. For example, panelists can be instructed to identify cut scores as the first item above a threshold region. In addition, psychometricians can locate cut scores in threshold regions by (a) calculating the scale value that represents the midpoint between the scale location of the first and last items in a threshold region, or (b) applying logistic regression to items in a threshold region, including or excluding the items in the adjacent performance levels.[6] Deciding whether

[6] For an illustration of using logistic regression to set a cut score in an application of ID Matching, see Sireci et al. (2007). Application of logistic regression was first proposed by Skip Livingston (personal communication, October 24, 2002).

panelists or psychometricians should locate cut scores in threshold regions can be based on the skill and qualifications of the panelists and the availability of time in the standard setting workshop.

A Note about the IRT Response Probability Criterion for Mapping Items

Another difference between the Bookmark and ID Matching methods is the role of the response probability (RP) criterion. The RP criterion is defined by the ability level (i.e., theta) corresponding to a given probability of success on a particular item. For example, a response probability of 0.67 (RP67) corresponds to the location on the theta scale at which the probability of responding successfully to an item is 0.67. Choosing an RP criterion is a policy decision with psychometric implications. The RP criterion affects the location of items in an ordered item book and potentially makes test performance standards easier or harder, depending on the RP criterion chosen. The RP criterion can change the order of the items in an item map, which influences panelists' decisions about placing cut scores. In addition, in the Bookmark method, the RP criterion determines the instructions given to panelists. When RP67 is used in a Bookmark standard setting, panelists are instructed to choose the point at which they expect examinees to answer items successfully with at least a 67% probability of success. We already cited research on how poorly people make probability judgments (see Plous, 1993, p. 144). In ID Matching, the RP criterion is relevant only during the item scaling and mapping process.[7] It is not directly relevant to the judgments that standard setting panelists must make to match items and descriptors and locate cut scores.

Steps in an ID Matching Workshop

Like any other standard setting method, ID Matching works best when the workshop includes sufficient time to train panelists on the content standards and test blueprints or specifications. We recommend having the panelists respond to items from the test on which they will be setting cut scores as part of training. We also recommend giving panelists sufficient time to review the performance level descriptors and talk about the differences between someone who is at the top of one level and someone who is at the bottom of the next level. Understanding these distinctions is particularly important for ID Matching.

We describe the core of the ID Matching workshop in three steps. Typically, an ID Matching workshop can be accomplished over a three-day period. We recommend a three-round process, which we describe below, but it can be conducted in two rounds if there are time constraints. In a two-round workshop, all feedback and data to be provided to panelists must be provided at the start of round 2.

Step 1: Round 1 of standard setting

a. Panelists work collaboratively to answer the two questions about all items in the test or a systematic sample of items.
    i. What do students need to know and be able to do in order to respond successfully to this item?
    ii. What makes this item more difficult than the ones that precede it?

b. Panelists determine item-descriptor matches independently.

c. Panelists locate cut scores in threshold regions.

[7] For a review of recommendations for selecting a response probability criterion, see Karantonis and Sireci (2006, pp. 6–9). Some psychometricians recommend RP 67 for the Bookmark method because panelists appear to understand and accept it more readily than RP 50 (Karantonis & Sireci, 2006) and because it maximizes psychometric information for the correct response for dichotomous items (Huynh, 2006). We believe that policymakers and stakeholders should be involved in the decision of which response probability criterion to use.

Step 2: Round 2 of standard setting

a. Panelists receive information on agreements and disagreements with other panelists on item-descriptor matches and threshold regions.

b. Panelists work individually to review and adjust the locations of their cut scores. In the process of doing so, they review their item-descriptor matches and may adjust them.

Step 3: Round 3 of standard setting

a. Panelists receive information on agreements and disagreements with other panelists on locations of cut scores.

b. Panelists receive impact information in the form of percentages of examinees at and above each performance level based on the median cut score.

c. Panelists work individually to review and adjust the locations of their cut scores and select final locations for cut scores.

Technical reports from ID Matching workshops describe the procedures and synthesize the discussions for each standard setting round; summarize item-descriptor matches, cut scores, threshold regions, and item sequences for each round; tabulate the final cut scores; include revised performance level descriptors (where relevant); and summarize panelists' evaluations of the workshop and results. Evaluation information can be used as evidence of the rigor of the standard setting process and the credibility of the standards.

EXAMPLES

To illustrate the utility of the IDM standard setting method, we provide results from two applications of the method. In the first example, we give a detailed description of workshop procedures and results from the standard setting for high school end-of-course examinations. In the second example, we focus on convergence of panelist judgments in the final round of standard setting and panelist evaluations of the effectiveness of the workshop.

Example 1: School District End-of-Course Examinations

AIR used the ID Matching method to set standards for two high school end-of-course examinations completed in April 2000 for a large, urban school district. For this first year of testing, the school district policymakers wanted only one pass/fail cut score. Thus, standard setting panelists worked with one "passing" description rather than performance level descriptors. This example represents the first operational use of ID Matching, so some of the procedures differ slightly from the steps described earlier.

We used field-test data from the December 1999 administration for two high school end-of-course examinations, in Living Environment I and English 9. OIBs were assembled using item p-values instead of IRT scale locations. Six teachers from each subject area participated in standard setting, along with curriculum specialists from the district central office. We opened training with panelists and content experts from both content areas together. We began by describing the purpose of standard setting and introducing the ID Matching procedure. Next, each panelist took a portion of the test on which the panel was setting standards. The school district content expert then described the content specifications and reviewed and explained the responses. Extra time was spent on the constructed-response items to explain the rubrics and examine the sample responses.

Training in ID Matching

Once the panelists were comfortable with the items and rubrics in each examination, they were trained in ID Matching procedures. The workshop leaders introduced the materials, including the item map and the ordered item book. Some time was spent explaining the p-values and the scale score locations and describing the phrase “knowledge, skills, and cognitive processes.” The next step was to train panelists to determine whether item knowledge, skills, and processes matched the passing description (which they indicated by recording a “yes” on their item maps) or did not match the passing description (indicated by a “no) or whether panelists were not sure whether an item did or did not match the description (indicated by “maybe”). We then trained panelists to expect to see clear and unclear match sequences: that is, a clear sequence of “no” responses, followed by some alternating “yes” and “no” responses and some “maybes,” followed by a clear sequence of “yes” responses. We also explained the threshold region.

Using publicly released grade 8 NAEP history items, we created a practice test for the panelists. We prepared an item map and an OIB from the NAEP items and used the NAEP Proficient achievement level description to define the just-passing level. After showing panelists how to use the two pieces together, we introduced the passing description for NAEP history, highlighting the knowledge, skills, and cognitive processes and the important features of the passing description. Then we modeled matching the knowledge, skills, and cognitive processes of the items to the knowledge, skills, and cognitive processes of the passing description. The panelists needed additional time to work with the constructed-response items to understand how the same item could be matched to different categories depending on the point value of the score.

The panelists then matched items on their own. The workshop leaders asked them to explain their responses when they disagreed on item-descriptor matches. Then panelists identified the threshold region. (At this stage of the development of ID Matching, panelists were taught to identify the threshold regions themselves.) We instructed panelists on how to determine the cut score. For this session, panelists were instructed as follows:

> Use your best judgment to determine where, in the threshold region, the knowledge, skills, and cognitive processes required by the items change and begin to match more closely the description of knowledge, skills, and cognitive processes required just to pass the examination.

Again, the group was given the chance to practice, using the NAEP items with the matches they had determined in the previous training session.

The large group then broke into two groups so that each group could receive subject-specific training on the passing description. Training panelists to understand all components of the description and to internalize the details is one of the most important components of the standard setting process, so some time was spent ensuring that all panelists understood and were comfortable with the passing description.

In the remainder of this section, we describe procedures and deliberations of the science panel so that we can provide specific details of an application of ID Matching.

Round 1—Matching Items and Descriptors

Panelists first worked individually to match the items to the passing descriptions, using the field-test data from the December 1999 administration of the science examination. In their item map, they marked “N” for no, if the knowledge, skills, and cognitive processes of the item did not match the passing description, and “Y” for yes, if they did. After panelists completed the item maps on their own, the facilitator integrated and summarized the maps and led a discussion. Initially, the panelists identified the threshold regions based on their item-descriptor matches as shown in Table 1.

Table 1

Threshold Regions after Rounds 1 and 2

Panelist	Ordered Items	Corresponding Scale Scores[8]
1	20–42	189–203
2	21–30	189–194
3	13–28	186–193
4	24–32	191–196
5	17–47	187–207
Panel	13–47	186–207

The panel threshold region fell between ordered items 13 and 47 in the ordered item book (scale scores 186–207); that is, the items corresponding to the highest and lowest scale scores across all panelists. The panelists agreed that this range was too large, so they discussed each of the items at both ends of the threshold to see whether they could narrow the threshold region. After the discussion, panelists seemed satisfied with defining the threshold region as ordered items 20–41 (i.e., scale scores 189–201). At the individual level, Panelists 2 and 4 maintained their threshold regions where they had them because their original decisions were within the group threshold region. The other panelists modified their ranges slightly to match the group range.

Round 2—Determining a Cut Score

The next step was to place the cut score by working through the threshold region and finding the point where the knowledge, skills, and cognitive processes of the items begin to match more closely the knowledge, skills, and cognitive processes of the passing description. This required about 20 minutes. The initial results appear in Table 2.

Table 2

Cut Scores on a Science Examination: Round 2

Panelist	Ordered Item Number on Each Side of the Cut Score	Cut Score Location on the Scale Score Scale
1	37, 38	200
2	24, 25	191
3	28, 29	193
4	23, 24	190
5	37, 38	200

The midpoint line (i.e., median) for all the panelists was located at the scale score 193, and the average was 195. The panelists engaged in discussion at this point, but none was willing to adjust his or her cut scores.

Round 3—Impact Data and Convergence

At this point, we gave the panelists the impact data as shown in Table 3.

[8] In this workshop, we used RP 50 to calculate the corresponding scale scores for each item.

Table 3

Impact Data, Science Exam: Round 3

Cut Scores	Scale Score	Percentage of Students Who Would Pass
Lowest	191	67%
Median	193	64%
Mean	195	61%
Highest	200	55%

After giving them time to review and discuss the impact data, we told the panelists that we wanted them to reach some degree of convergence but that we did not expect all of them to agree on a single cut score. To provide another way of thinking about the task, we asked them to consider the following: Which is the first item (i.e., in the item difficulty order) that students must answer correctly to demonstrate that they have sufficient knowledge of science to pass this test? Thinking about the task this way, panelists agreed that the turning point was between items 29 and 30. Thus, students scoring at or below 193 would fail the test, and those scoring at or above 194 would pass it. With this cut score, 64% of the students in the field test would have passed the test.

Debriefing

Panelists from both subject groups came together for the debriefing. Each panelist completed a written evaluation in about 15 minutes. Then, we asked them questions about their understanding of the terms and tasks. Panelists showed fairly consistent understanding of the process, and they indicated that it had become clearer after they practiced with the NAEP items. They also indicated that they had confidence that the process resulted in fair distinctions between passing and failing and that they would be willing to defend the process and the final cut score to their peers.

Example 2: High School End-of-Course Examinations

In summer 2007, panelists in three separate workshops followed ID Matching procedures in two rounds, as described in the Steps in an ID Matching Workshop section above, to set two cut scores for high school end-of-course examinations. Student performance results were used to evaluate and refine the impact of innovative curriculum programs in three content areas. Table 4 contains median cut score pages (calculated across all panelists) for the language arts assessment for rounds 1 and 2 plus associated standard errors.

Table 4

Descriptive Statistics for an End-of-Course Examinations Standard Setting

	Round 1				Round 2			
	Median	High	Low	Range	Median	High	Low	Range
				Biology				
Proficient	44.5	48	31	17	41	45	33	12
Basic	18.5	28	11	17	18	20	18	2

(continued)

Table 4 (continued)

Environmental/Earth Science								
Proficient	44	50	28	22	43.5	48	37	11
Basic	14.5	28	9	19	17.5	28	16	12
English 9								
Proficient	33	37	23	14	32.5	34	27	13
Basic	18	20	11	9	17	19	11	8

Note. Each multiple choice item and constructed-response item score level appears on only one page in the ordered item book.

It is clear in Table 4 that even though the cut score pages change little from round 1 to round 2, in most cases the cut score ranges decrease, sometimes substantially. For example, the range of panelists' cut score pages for Proficient for Biology decreased from 17 pages in round 1 to 12 pages at the end of round 2. These results demonstrate that panelist judgments about cut scores do converge, even after only two rounds, and provide evidence of the effectiveness of the ID Matching method and workshop procedures. And we would have expected their judgments to converge even more, if we had been able to conduct a third round.

DISCUSSION AND CONCLUSION

ID Matching is offered as another approach to standard setting with its own set of advantages and disadvantages. Although similar to the Bookmark method, it is unique in several ways. ID Matching:

- Capitalizes on panelists' content area expertise, including identifying what students need to know and be able to do in learning and assessment situations.
- Does not require panelists to consider probabilities of successful responses.
- Does not require panelists to consider imaginary students who are just barely in a performance level.
- Is more robust to minor fluctuations in item parameters.
- Provides detailed information about panelist thinking relevant to placing cut scores.

But what evidence do we have that ID Matching produces credible, defensible, and valid results? How do we know it works?

Evaluating the Validity of Results from the ID Matching Method

Hambleton and Pitoniak (2006, pp. 457–463) propose numerous types of evidence for documenting and evaluating standard setting studies. They include questions about the standard setting panel, method used, implementation of procedures, documentation of the process, communication of the final standards, and support for interpretation. One evaluation question refers to panelists' qualifications to "make the required ratings" (question 2, p. 109). Earlier we discussed that panelists apply what they know about curriculum, instruction, and the students they teach to make item-descriptor matching judgments. A second evaluation question refers to the robustness of the method. ID Matching can be implemented successfully by testing program staff who have no graduate training in measurement (e.g., in Bahia, Brazil) and by other psychometricians (e.g., see Sireci et al., 2007).

Further, the evidence from the two school district end-of-course examinations and the state alternate assessment show convergence of panelists' judgments: the standard deviation and range of panelists' recommended cut scores decreased over standard setting rounds. Finally, panelists' evaluations of the training, the standard setting process, and the reasonableness of the recommended cut scores have been positive in the three ID Matching workshops described in this paper.

We propose an additional question related to the notion of the construct validity of performance standards: Is the cognitive-judgmental task that panelists use to set recommended cut scores consistent with the intended meaning of the performance standards? Performance level descriptors define what students who perform at a level are likely to know and be able to do. The cognitive-judgmental task in the Angoff and Bookmark methods requires panelists to estimate probabilities. Estimating probabilities seems to be a step removed from the intended meaning of performance level descriptors. In contrast, the ID Matching method appears to align the cognitive-judgmental task closely to the intended meaning of performance level descriptors by requiring panelists to match item response requirements with performance level descriptors. We advocate studies that examine our hypotheses about the cognitive-judgmental task required for various standard setting methods.

CLOSING

ID Matching provides another promising method for test developers and psychometricians to use to set cut scores. The cognitive-judgmental task seems well suited to the experience and expertise of typical standard setting panelists. Panelists report that they are comfortable with the item-descriptor matching task and confident in the results it produces. Standard deviations of individual panelists' cut scores tend to be relatively small in relation to average panelists' cut scores, and we have strong evidence that panelists' item-descriptor matches and recommended cut scores converge by the end of four rounds. Future applications of ID Matching will provide more information on the validity of decisions based on ID Matching cut scores, the robustness of cut scores, and comparisons with other standard setting methods.

REFERENCES

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.

Ferrara, S., & DeMauro, G. E. (2006). Standardized assessment of individual achievement in K–12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 579–621). Westport, CT: American Council on Education/Praeger.

Ferrara, S., Phillips, G., Williams, P., Leinwand, S., Mahoney, S., & Ahadi, S. (in press). Vertically articulated performance standards: An exploratory study of inferences about achievement and growth. In R. Lissitz (Ed.), *Assessing and modeling cognitive development in school: Intellectual growth and standard setting*.

Glass, G. V. (1978). Standards and criteria. *Journal of Educational Measurement, 15*, 237–261.

Green, D. R., Trimble, C. S., & Lewis, D. M. (2003). Interpreting the results of three standard-setting procedures. *Educational Measurement: Issues and Practice*, *22*(1), 22–32.

Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.). Westport, CT: American Council on Education/Praeger Publishers.

Huynh, H. (2006). A clarification on the response probability criterion RP67 for standard settings based on Bookmark and item mapping. *Educational Measurement: Issues and Practice*, *25*(2), 19–20.

Impara, J. C., & Plake, B. S. (1998). Teachers’ ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement, 35*(1), 69–81.

Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates.

Karantonis, A., & Sireci, S. G. (2006). The Bookmark standard-setting method: A literature review. *Educational Measurement: Issues and Practice*, *25*(1), 4–12.

Kingston, N. M., Kahl, S. R., Sweeney, K. P., & Bay, L. (2001). Setting performance standards using the Body of Work method. In G.J. Cizek (Ed.), Setting performance standards: Concepts, methods, and perspectives. Mahwah, NJ: Lawrence Erlbaum Associates.

Mills, C. N., & Jaeger, R. M. (1998). Creating descriptions of desired student achievement when setting performance standards. In L. Hansche (Ed.), *Handbook for the development of performance standards*. Washington, DC: U.S. Department of Education and Council of Chief State School Officers.

National Academy of Education. (1997). *Assessment in transition: Monitoring the nation’s educational progress*. Palo Alto, CA: Stanford University.

Nickerson, R. S. (2004). *Cognition and chance: The psychology of probabilistic reasoning*. Mahwah, NJ: Lawrence Erlbaum Associates.

Perie, M. (2005). *Angoff and Bookmark methods*. Workshop presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Plous, S. (1993). *The psychology of judgment and decision making*. New York: McGraw-Hill.

Reckase, M. D. (2006a). A conceptual framework for a psychometric theory for standard setting with examples of its use for evaluating the functioning of two standard setting methods. *Educational Measurement: Issues and Practice*, *25*(2), 4–18.

Reckase, M. D. (2006b). Rejoinder: Evaluating standard setting methods using error models proposed by Schulz. *Educational Measurement: Issues and Practice*, *25*(3), 14–17.

Schulz, E. M. (2006). Commentary: A response to Reckase’s conceptual framework and examples for evaluating standard setting methods. *Educational Measurement: Issues and Practice*, *25*(3), 4–13.

Sireci, S. G., Baldwin, P., Martone, D., & Han, K. T. (2007). Establishing achievement levels on a multi-stage computerized-adaptive test: An application of the Item Descriptor Matching method. In B. Plake (Chair), *Innovations in Standard Setting*, a symposium conducted at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Wang, N. (2003). Use of the Rasch IRT model in standard setting: An item mapping method. *Journal of Educational Measurement, 40*(3), 231–253.

Zieky, M., Perie, M., & Livingston, S. (in press). *Cutscores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.

Zwick, R., Senturk, D., Wang, J., & Loomis, S. C. (2001). An investigation of alternative methods for item mapping in the National Assessment of Educational Progress. *Educational Measurement: Issues and Practice, 20*(2), 15–25.