Abstract

The Mapmark standard setting method was adapted to a higher education setting in which faculty leaders were highly involved. Eighteen university faculty members participated in a day-long standard setting for a general education communications test. In Round 1, faculty set initial cut-scores for each of four student learning objectives. In Rounds 2 and 3, participants used a Mapmark item map to consider information from four student learning objectives at one glance and to integrate this information into a single cut-score. Participants and faculty leaders reported that the process was intuitive, and there was support for a defensible cut-score from the majority of participants and faculty leaders. Practical suggestions and implications are discussed.

AUTHORS

S. Jeanne Horst, Ph.D.
James Madison University

Christine E. DeMars, Ph.D.
James Madison University

CORRESPONDENCE

Email
hortstsj@jmu.edu

Higher Education Faculty Engagement in a Modified Mapmark Standard Setting

In higher education, setting a standard on an assessment can assist faculty and administrators to distinguish between students who are or are not meeting learning objectives. Standard-setting is the process of selecting cut-scores on a test that will separate examinees' scores into achievement categories (Cizek, 2001; Cizek, Bunch, & Koons, 2004). This facilitates the interpretation of scores in a criterion-referenced fashion, because each category is accompanied by a description of what examinees in that category typically know or can do. For example, on certification exams, the cut-score may be used to indicate whether an examinee has at least adequate knowledge or skills to perform in a job or profession.

Standard-setting has long played a role in primary and secondary education, from the minimal competency or graduation tests common in the 1970s-1990s, to the many statewide tests created in response to the No Child Left Behind (NCLB) legislation, and now to tests under development for the Common Core standards (e.g., Borque & Hambleton, 1993; Tong, Patterson, Swerdzewski, & Shyer, 2014). Even when cut-scores are not used for purposes of passing a test, proficiency categories help students and instructors understand what a score means (AERA, APA, & NCME, 2014, Chapter 5). Although less common, standard-setting is also helpful in higher education. Although higher education scores are typically reported as percent-correct, depending on the difficulty and content-coverage of a test, the percent-correct score may have different meanings. For example, on a test designed to measure a wide range of difficulty spanning four years of education in a major, first-year students scoring 60% may have exceeded the expectations faculty set based on the first-year curriculum. However, if the test only covers foundational concepts students should know before entering the program, this same 60% is likely below the faculty's standard. Proficiency

categories help clarify what a score of 60% means in each of these contexts. In this paper, we will describe a standard-setting workshop for university faculty to set a cut-score on a required communications test. We will discuss the ways the procedure was adapted to meet the needs of the faculty and highlight unique features of the higher-education context.

Standard Setting Procedures

For example, in the current study education professionals requested that participants examine separate ordered item booklets for each of four objectives, rather than one comprehensive ordered item booklet. For this reason, the Mapmark standard setting procedure offered an appealing alternative.

Many methods have been developed for setting standards. Common to most are (a) the development of performance standards (i.e., qualitative descriptions of performance levels, or what students should know and be able to do at the particular level) and (b) the setting of cut-scores (i.e., the score at which an examinee is said to have met the standard; Kane, 1998, 2001). In this study, following the development of performance standards by faculty experts, we used a modification of the Mapmark method, which is closely related to the bookmark method. Mapmark has been used at the national level for setting standards related to the National Assessment of Educational Progress (ACT, Inc., 2007). For purposes of contrast, it is important to briefly introduce one of the most commonly used standard setting methods, the Angoff standard setting method.

Angoff Standard Setting Procedure. Although there are several variants, the Angoff standard setting procedure typically requires standard setting participants (i.e., experts or judges) to conceptualize a "hypothetical minimally acceptable person" (Cizek et al., 2004, p. 40). During the standard setting, participants view test items and make judgments about whether they believe the hypothetical examinee could correctly answer each item. Often participants indicate the proportion of minimally acceptable students who would correctly answer each item. Alternatively, in one common variant of the Angoff procedure, participants respond yes (1) or no (0) regarding whether the hypothetical examinee could correctly answer each item (Impara & Plake, 1997). The cut-score is determined from the average across the items and participants. For example, if the average rating across items and participants is .58, then the cut-score would be 58% correct (Cizek et al., 2004). Other common modifications of the Angoff procedure include multiple rounds (typically two or three) of judgements. Between rounds, workshop leaders facilitate discussions about differences in cut-score judgements. Before the final round of judgements, participants generally receive feedback about their own and others' cut-scores, as well as information about student performance relative to the cut-score, termed *impact* because this information can be used to assess the impact of the cut-score on students.

Inherent within the Angoff method is the assumption that participants are able to adequately conceptualize the knowledge, skills, and abilities of the hypothetical minimally-acceptable examinee, and are able to predict how well that examinee would be able to perform on each item (Impara & Plake, 1998). Moreover, as may be expected, participants *do not* always accurately conceptualize the abilities of the minimally-acceptable examinee (Impara & Plake, 1997, 1998). The bookmark standard setting method attempted to simplify the cognitive task required of Angoff participants by providing booklets of items ordered by empirical difficulty.

Bookmark Standard Setting Procedure. The bookmark standard setting procedure was developed for purposes of minimizing the cognitive tasks and number of judgments required of standard setting participants (Mitzel, Lewis, Patz, & Green, 2001). The central feature of the bookmark method is the ordered item booklet, which consists of test items presented in order of item difficulty. Additionally, participants are provided an item map, which is a table that summarizes the item location information (Mitzel et al., 2001). Standard setting participants place a bookmark at the page at which a minimally-competent examinee would have *mastered* the items prior to the bookmark and would have *not mastered* the items following the bookmark. To "master" an item refers to the point at which the *just-competent* examinee would answer the item correctly, roughly 67% of the time (70-75% with guessing).[1]

Bookmark standard settings typically involve three rounds, similar to many Angoff standard settings. Following orientation, participants review each item in small groups. Participants attempt to identify the knowledge, skills, and abilities required of each item, and the features of each item that make it more difficult than previous items (Mitzel et al., 2001). Following Round 1, participants individually place bookmarks. During Round 2, small group

participants discuss the group's bookmarks in light of the characteristics of the items that fall within the group's range, as well as what students should know at the various proficiency levels. Based on small group discussions, participants again place a bookmark. Following Round 2, the median for each small group and the total group is presented, along with impact data (the percentage of students who would have achieved each performance level). Round 3 involves a discussion among the entire group of participants, following which participants again place individual bookmarks; the final cut-score is the median of these bookmarks. The final cut-score and impact data are presented.

One benefit of the bookmark method over other methods is that item difficulties have been empirically computed, allowing panelists to focus on the *content* of the items (Shulz & Mitzel, 2011). However, one quandary is how to manage the ordered item booklets when test developers desire close attention to items by objectives or domains. For example, in the current study education professionals requested that participants examine separate ordered item booklets for each of four objectives, rather than one comprehensive ordered item booklet. For this reason, the Mapmark standard setting procedure offered an appealing alternative.

Mapmark Standard Setting Procedure. The Mapmark method enhances the bookmark standard setting procedure by assigning the item map a central role in the process (Schulz & Mitzel, 2011). However, unlike the item map provided in the bookmark method, which is simply a list of empirical information about each item in the item booklet, the item map in the Mapmark method presents the information visuo-spatially. By providing spatial information for panelists to judge the distance between the difficulty of the items (see Figure 1), the Mapmark method offers "holistic feedback" on the entire test (Schulz & Mitzel, 2011, p. 168). Round 1 bookmarks are placed in ordered item booklets, as in the bookmark method, but in successive rounds the bookmarks are placed on the item map. Sometimes there are large score gaps between items in the item booklet. In the bookmark procedure, participants must choose a specific item for the cut-score, but in the Mapmark procedure participants can choose to place the cut-score anywhere on the scale, even at scores to which no item difficulties are mapped. As seen in Figure 1, in one glance, panelists are able to focus on the spread of difficulty across domains or objectives. This particular feature of the Mapmark standard-setting procedure was of interest to the current study, in which we were interested in simultaneously presenting information on four separate communication learning objectives.

Context for the Current Study

At a mid-sized public university in the Mid-Atlantic region all students are required to take a basic communications course that covers four learning objectives: (a) Construct messages consistent with the diversity of communication purpose, audience, context, and ethics; (b) Respond to messages consistent with the diversity of communication purpose, audience, context, and ethics; (c) Explain the fundamental processes that significantly influence communication; (d) Utilize information literacy skills expected of ethical communicators. The course is part of the General Education program, which is divided into five components called Clusters. The communications course is part of Cluster 1: Foundations, which includes critical thinking, writing, communication, and information literacy. The current Cluster 1 coordinator is also a Speech Communications professor and the former course director.

The proficiency classifications are used specifically for assessment purposes, to help faculty to judge whether curricular/ instructional changes are needed, and for external accountability reporting.

All basic communication students take a common 100-item course-embedded final exam, which includes 25 items mapped to each of the four learning objectives. The exam is administered in a proctored computer lab. There are approximately 70-80 sections of the course each semester, with 4,000-4,500 students per year. Each instructor can choose the specific

1 Selecting the appropriate response probability (RP) value can be controversial and can influence the order of items in the ordered item booklet. The RP plays a role in determining the location of items when an item response theory model other than Rasch is employed, and influences the description of the standard setting procedure to workshop participants. Participants seem able to adjust the bookmark to partly but not fully compensate for changes in the RP (National Academies of Sciences, 2005, Ch. 5). Traditionally, the bookmark procedure included .67 RP (Mitzel et al., 2001); however, other response probabilities have been investigated (Karantonis & Sireci, 2006). For example, a practitioner may choose to select .50 RP, in which to "master" an item the just-competent examinee would answer the item correctly roughly 50% of the time. However, it is argued that because .67 is above .50, it is more consistent with arguing that a just-competent examinee has mastered an item than .50 RP (Karantonis & Sireci, 2006).

Proportion Correct at Scale Score					Items near Scale Score, by page #			
Scale Score	Obj. 1	Obj. 2	Obj. 3	Obj. 4	Obj. 1	Obj. 2	Obj. 3	Obj. 4
≤200	48%	53%	48%	65%	1,2,3,4	1,2,3,4,5,6	1	1,2,3,4,5,6,7,8,9
210	49%	54%	49%	65%	.	.	.	.
220	50%	55%	50%	66%	.	.	.	.
230	52%	56%	51%	67%	5	7	.	.
240	53%	57%	52%	68%	.	.	.	.
250	54%	58%	53%	69%	6	8	.	10
260	55%	59%	54%	69%	.	9	.	11
270	56%	60%	55%	70%	.	.	2	12
280	58%	61%	56%	71%	.	.	.	13,14,15
290	59%	62%	57%	71%	.	.	.	.
300	60%	63%	58%	72%	7	10	.	.
310	61%	64%	59%	73%	.	.	3	.
320	63%	65%	60%	73%	.	11	4,5	.
330	64%	66%	61%	74%	.	.	6	16
340	65%	67%	63%	75%	.	.	7	.
350	67%	68%	64%	75%	.	12	8	.
360	68%	69%	65%	76%	8,9,10	.	9,10	.
370	69%	70%	66%	76%	.	.	.	.
380	70%	71%	67%	77%	11,12	13,14	.	.
390	72%	72%	68%	77%	13	15,16	.	.
400	73%	73%	69%	78%	14	17	.	.
410	74%	74%	70%	78%	15,16	.	11	.
420	75%	75%	71%	79%	17	.	.	17
430	76%	76%	72%	79%	.	.	12,13	.
440	77%	76%	73%	80%	18	18	14,15	.
450	78%	77%	74%	80%	.	.	16	.

Figure 1. *Mapmark item map. The complete item map extended to a score of 800.*

learning activities, but all sections use the same textbook and cover the same objectives. The basic course director, a Speech Communications professor, facilitates consistency across the many instructors, and oversaw faculty who wrote the test items.

Rationale for Standard Setting

The cut-score corresponding to the proficiency standard is not used to determine whether students pass or fail the course. The continuous score on the final exam, not the dichotomous proficiency classification, is incorporated as one part of each student's course grade, along with presentations and other in-class assignments. The proficiency classifications are used specifically for assessment purposes, to help faculty to judge whether curricular/instructional changes are needed, and for external accountability reporting.

Thus, faculty leaders wanted the procedure modified to separate the learning objectives, yet yield a single cut-score. Therefore, the Mapmark standard setting procedure was chosen as a viable standard setting method.

Because the context of the current study differs from the traditional K-12 standard setting, it is important to carefully define three roles: faculty leaders, workshop leaders, and participants. For the purpose of program evaluation and accountability reporting, the course director and Cluster 1 coordinator requested the assistance of faculty at the university's assessment office in setting a proficiency standard on the final exam. The term *faculty leaders* will be used to refer to the Cluster Coordinator and the course director. The term *workshop leaders* will be used to refer to the personnel who did the psychometric work, prepared materials, and helped facilitate the workshop. These labels are arbitrary because both groups are faculty and both groups participated in leading the workshop, but short labels are needed for description. The workshop leaders played the role typically fulfilled by testing company staff when setting standards for statewide K-12 tests or certification tests. The faculty leaders, on the other hand, have no direct parallel. Because of the scale of statewide K-12 tests, curriculum leaders are generally not personally known by the standard setting participants the way the faculty leaders were in this context. Because the standard-setting took place in a single university, and most of the participants taught General Education courses, the faculty leaders were viewed as colleagues. Finally, the term *participants* will be used to refer to faculty members who served as content experts throughout the workshop.

The faculty leaders had participated in other standard setting workshops at the university, using a modified bookmark procedure (for example, DeMars, Sundre, & Wise, 2002). In previous standard settings, all items were included in one ordered item booklet,

regardless of the objective to which the item aligned. Faculty leaders felt it was confusing combining items mapped to four separate learning objectives into one ordered item booklet, making it difficult to discuss what each item was measuring and why it might be harder than the item before it. Thus, faculty leaders wanted the procedure modified to separate the learning objectives, yet yield a single cut-score. Therefore, the Mapmark standard setting procedure was chosen as a viable standard setting method.

Purpose

The purpose of this study was to illustrate a variation on the Mapmark standard setting procedure designed to highlight multiple learning objectives assessed by one test. A secondary purpose was to illustrate standard setting within a higher-education context. The context of the current study was unique, relative to traditional standard settings, given that faculty leaders were highly involved in the process. Moreover, faculty leaders felt strongly that items should be considered by learning objective. Also unique to the higher education setting was the length of the standard setting workshop. Rather than several days, the current standard setting was conducted in one day, to minimize demands on faculty time. The current study summarizes this adaptation of the Mapmark standard setting procedure in a higher education context.

Method

Modification of the Mapmark Procedure

Because the faculty were dissatisfied with previous standard-settings, in which items with different learning objectives were interspersed within the ordered item booklets, workshop leaders and faculty leaders discussed ways of separating the task by learning objective. Faculty agreed that they wanted a single cut-score on the test as a whole, not four separate standards. One option was for the participants to set four separate standards using the bookmark procedure and combine them at the end of the workshop. One concern was that, with the shorter ordered-item-booklets resulting from dividing the items by objective, there would be many score gaps within each booklet. Imagine that the just-proficient student envisioned by a particular participant has the skills corresponding to a scaled score of 328. The standard-setting participant does not know the value 328, but can, hypothetically, envision skills and knowledge at this level. But there may not be any items close to this level; perhaps there is a large gap between an item located at 280 and another located at 362. Another problem is that if each standard were set in isolation, the standards for each learning objective would likely end up at very different points on the proficiency continuum and the mean would not represent the desired proficiencies well. This might be hidden from the participants by using a method that sets the standard on the percent-correct metric, such as the Angoff method; participants would assume that objectives where they set the percent-correct cut-score high were easier than objectives where they set the percent-correct cut-score low. Of course, hiding the incongruity from the participants does not make it go away. Setting the cut-score on the percent-correct metric could also be problematic when the test forms changed; the cut-score might correspond to a different percent-correct when the new form was equated. The faculty leaders also were comfortable with the Bookmark method and did not want to replace it.

The context of the current study was unique, relative to traditional standard settings, given that faculty leaders were highly involved in the process.

The Mapmark procedure provided a way to incorporate the learning objectives because it displays the expected percent-correct by objective or content area. Although participants using the Mapmark procedure generally use a single item booklet in Round 1, with items from different objectives interspersed, we modified Round 1 to include four separate ordered-item booklets, and participants set four separate bookmarks. During Round 2, participants received feedback on where their bookmarks for the different objectives fell relative to the scale scores and to bookmarks set by others. Each participant then set a single bookmark directly on the overall scale in successive rounds.

Preparation

Performance-level descriptors were written by the faculty leaders. Detailed descriptors are important for helping standard setting participants envision students who just meet the criteria for each performance level (Kane, 1998, p. 134; 2001, p. 59). Without written descriptors, participants will implicitly define the performance levels for themselves, which

can lead to wide variation in interpreting the performance levels. Perie (2008) provided practical suggestions on developing performance-level descriptions.

The faculty wanted a single cut-score on the test as a whole, which implies a unidimensional scoring model. It seems somewhat cognitively inconsistent to emphasize the uniqueness of the learning objectives yet score the test using a unidimensional model. To make sure that a single score on the test was meaningful, we ran a multidimensional 3-parameter-logistic (3PL) confirmatory factor model. The latent (disattenuated) correlations among the first three factors were estimated to be 1. The factor tapping Objective 4 was estimated to be correlated .83 with the other three factors. The RMSEA[2] was .01 for both the 4-dimensional model and the 1-dimensional model, suggesting both models fit acceptably. Thus, it seemed reasonable to follow the faculty desire for a single score (unidimensional model).

Materials were prepared for Round 1 following the usual bookmark procedures. Based on the unidimensional 3PL[3] calibration, the item location was calculated. The item location was defined as the ability at which an examinee would have a 2/3 probability of correct response, not counting correct guessing (Lewis, Green, Mitzel, Baum, & Patz, 1998), also referred to as .67 RP. Recognizing that the choice of RP can be controversial, we chose the .67 RP (i.e., 2/3 probability of correct response), which aligns with the original description of the Bookmark method (Mitzel et al., 2001) and is consistent with findings suggesting that participants more easily conceptualize .67 as examinee mastery of items (Karantonis & Sireci, 2006). The item locations were linearly transformed to the scaled scores used in score reporting, ranging from 200 to 800. In a typical bookmark or Mapmark standard setting, items are ordered by location. In this modification, items were separated by objective and ordered within each objective. Each item was printed on a separate page, along with information about the proportion of students in the upper and lower thirds of the score distribution who chose each option.

In a typical bookmark or Mapmark standard setting, items are ordered by location. In this modification, items were separated by objective and ordered within each objective.

For Rounds 2 and 3, an item map was assembled showing scaled scores in increments of ten. At each scaled score, the expected proportion correct was displayed for each objective, followed by the page numbers of items that mapped to that scaled score after rounding. An example of the first part of the scale range is shown in Figure 1—the complete scale range was printed out on a single sheet of 11 by 17 paper for each participant. Figure 1 illustrates, for example, that students who scored 300 would have average raw scores of 60% on Objective 1, and 63%, 58%, and 72% on Objectives 2, 3, and 4, respectively. About 2/3 of the students at score 300 have mastered the 7th item in Objective 1, plus a few more would get it right by guessing. Higher proportions of the students at score 300 have mastered the first 6 items in Objective 1, and lower proportions have mastered the harder items ordered after item 7. This item map helps the participants put the separate learning objectives back into the context of the test as a whole. Score gaps are also evident in Figure 1. For example, using the Mapmark item map, participants could place the cut-score at a score of 370, which would not be possible using the bookmark procedure because there are no items located near that score.

Workshop Activities

The 18 participants completed the test prior to the workshop so that the entire standard-setting could take place in a single day. After providing an overview of the day's activities, faculty leaders provided a context for the test's use within the general education program and discussed the development of the test. Workshop leaders discussed item writing, the way in which distractors contribute to an item's difficulty, and introduced activities that would occur throughout the day. Prior to the beginning of the session, the entire group discussed performance level descriptors. Given that the task was to set one cut-score, there were two performance-level descriptors written by the faculty leaders.The Developing student was described as:

2 The RMSEA used here is based on marginalizing estimations from full-information methods down to bivariate moments so that fit indices developed for limited-information methods can be estimated (Maydeu-Olivares & Joe, 2014).

3 More precisely, a bifactor model was used with secondary factors to account for dependence between some pairs of items, with the parameter estimates projected onto the primary factor (Kahraman & Thompson, 2011) to produce a unidimensional scale.

> "Students below the proficient category have not demonstrated the skills necessary to be able to recognize the fundamental processes that significantly influence communication. Students at this level have not demonstrated an ability to ethically construct and respond to messages consistent with the diversity of communication purposes, audiences, and contexts. They may be unable to utilize information literacy skills or to construct and/or respond to messages effectively or ethically. This category denotes partial but insufficient mastery."

The Proficient student was described as:

> "Students meeting this standard are able to explain the fundamental processes that significantly influence communication. Students at this level demonstrate an ability to ethically construct and respond to messages consistent with the diversity of communication purposes, audiences, and contexts. Students who achieve this standard are able to utilize information literacy skills expected of ethical communicators. Although further development is expected, students achieving this level or higher have the knowledge necessary to communicate effectively within the [institution] academic community."

Participants were each provided a notebook that included: agenda, background context, performance level descriptions, and the four ordered item booklets, one per learning objective.

Round One. Participants divided into four table groups. Starting with Objective 1, participants followed the usual bookmark procedure for Round 1. Each group discussed what each item measured and why it was more difficult than the previous item. A separate item map was provided for each objective, so participants could see when the locations of adjacent items were similar and not spend time trying to discern nonexistent or small differences in item difficulty. Table leaders encouraged full participation from everyone at their table. After all tables discussed Objective 1 items, the bookmark process was explained. After placing bookmarks for Objective 1, table groups discussed Objective 2 items, placed bookmarks, and proceeded through the remaining Objectives. Workshop leaders calculated scale scores for (a) each participant's four bookmarks, (b) mean ratings across each participant's four bookmarks, (c) each table's median rating, and (d) each table's lowest and highest average bookmark scale score.

Round Two. After a lunch break, table group results and Mapmark item maps were explained. Once participants demonstrated that they understood the Mapmark item map, they were encouraged to flag the place on the scale next to the bookmark they selected for each objective and their table's lowest and highest bookmark. Table leaders directed participants' attention to the items between the table's lowest and highest bookmarks. Participants discussed the knowledge, skills, and abilities they believed the items were measuring and whether just-Proficient students should be expected to master the content represented by the items. After small group discussion, each participant individually placed *one* Round 2 bookmark, indicating the scale score appropriate for a just-Proficient student. Workshop leaders tabulated each participant's response and provided the median small group scale score.

Round Three. Following a break, the entire group resumed for discussion. Workshop leaders presented a summary of each table's median scale score as well as impact data for the entire group's median cut-score. The impact data were based upon data from the previous year's administration of the test, and indicated the percent of examinees scoring at or above Proficient level based on the Round 2 median bookmark. Following discussion, participants were instructed to place their third and final bookmarks. Workshop leaders tabulated the data and presented the final cut-score and impact data. Faculty leaders and workshop leaders led discussion with participants about their satisfaction with the final cut-score and the day's experiences. Participants completed an evaluation prior to leaving.

Because we were adapting the Mapmark method to our context, it was crucial to evaluate the appropriateness of the method.

Results and Validation

Scores are on a scale from 200-800, with a mean of 500 and standard deviation of 100. The recommended cut-score following Round 3 was 480. Impact data computed from the previous year's administration of the test indicated that with this cut-score, 58% of students

taking the basic communications course would have been classified as Proficient. Although 58% Proficiency may seem stringent, faculty leaders and participants expressed strong support for the score.

Other distinctive features of the process were that faculty leaders were highly involved throughout the standard setting, and that, with the exception that we required participants to complete the test prior to the standard setting, the standard setting occurred in only one day.

In the context of describing the choice of an appropriate standard setting method, Kane (1998) noted, "it is not easy to evaluate how well a standard-setting procedure is working" (p. 130). That is, standard settings are fraught with subjectivity and arbitrary decisions (Kane, 1994). Cut-scores are representative of the value judgments of the standard setting participants (AERA, APA, & NCME, 2014, p. 101). At best, evaluation of the effectiveness of a standard setting method involves consideration of the appropriateness within the context and purpose for the standard setting, and evaluation of the validity of inferences drawn from application of the standard. The current context was an educational setting, in which faculty leaders were highly involved in the process and would use the information for improvement of their program, rather than high-stakes student pass/no-pass decisions. As such, we felt the strongest evidence would be to adopt the validity argument approach to evaluating the appropriateness of the adaptation of the Mapmark to the current context. At least three forms of validity evidence are recommended: procedural, internal consistency, and external evidence (Kane, 1994, 2001).

Procedural Evidence

Kane (1998) stressed that cut-scores are set, not estimated. There is no "true" cut-score. Thus, procedural evidence often plays a large role in validating the cut-score (Kane, 1994, 1998, 2001). Because we were adapting the Mapmark method to our context, it was crucial to evaluate the appropriateness of the method. We attempted to stay true to the traditional bookmark and Mapmark procedures, as well as general best practices described within the standard setting literature (e.g., Hambleton, 2001; Plake, 2008). And, although anecdotal, standard setting participants seemed to easily grasp the concept of the Mapmark item map. For purposes of assessing procedural validity, we administered a paper-pencil questionnaire immediately following the standard setting.

Table 1

Responses to Satisfaction Questions (Procedural Validity)

Satisfaction with final cut-scores
100.0% (18) Satisfied/Very Satisfied
0.0% (0) Neither satisfied nor dissatisfied
0.0% (0) Dissatisfied/Very Dissatisfied
Satisfaction with standards-referenced nature of cut-scores
94.5% (17) Satisfied/Very Satisfied
5.6% (1) Neither satisfied nor dissatisfied
0.0% (0) Dissatisfied/Very Dissatisfied
Satisfaction with consideration of values/opinions
88.9% (16) Satisfied/Very Satisfied
5.6% (1) Neither satisfied nor dissatisfied
5.6% (1) Dissatisfied
0.0% (0) Very Dissatisfied
Defending the cut-point
83.3% (15) would defend the cut-point
16.7% (3) would not defend the cut-point
Round 3 bookmark changes
38.9% (7) changed bookmark but not as a result of the impact data
38.9% (7) changed bookmark based on the impact data or others' reactions to it
22.2% (4) did not change bookmark
Confidence in Bookmark Procedure for setting valid standards
72.2% (13) Confident/Very Confident
27.8% (5) Neutral
0.0% (0) Not Confident /Not at all Confident
Agreement with item ordering in booklets
88.9% (16) Generally/Somewhat Agreed
5.6% (1) Neither Agreed nor Disagreed
5.6% (1) Somewhat Disagreed
0.0% (0) Generally Disagreed

Table 2

Response to Workshop Setting (Procedural Validity)
Organization of workshop
94.4% (17) Very Organized/Organized
0.0% (0) Neither Organized nor Disorganized
5.6% (1) Disorganized
0.0% (0) Very Disorganized
Quality of general Bookmark training
44.4% (8) Excellent
38.9% (7) Good
16.7% (3) Fair
0.0% (0) Poor
0.0% (0) Fail
Quality of workshop leaders
50.0% (9) Excellent
38.9% (7) Good
11.1% (2) Fair
0.0% (0) Poor
0.0% (0) Fail
Overall Value of Workshop as Professional Development Experience
66.7% (12) Excellent
27.8% (5) Good
5.6% (1) Fair
0.0% (0) Poor
0.0% (0) Fail
Value of Interacting with peers in the group
83.3% (15) Excellent
11.1% (2) Good
5.6% (1) Fair
0.0% (0) Poor
0.0% (0) Fail
Value of constructing better classroom tests (1 missing)
52.9% (9) Excellent
29.4% (5) Good
17.6% (3) Fair
0.0% (0) Poor
0.0% (0) Fail
Value of targeting instruction (2 missing)
37.5% (6) Excellent
31.3% (5) Good
31.3% (5) Fair
0.0% (0) Poor
0.0% (0) Fail

Overall, program participants were satisfied with the workshop. Only one person expressed dissatisfaction with the extent to which participant opinions were considered and valued. A majority of participants (83.3%) stated that they would defend the final cut-score. Three participants who indicated they would not defend the cut-score also indicated that they had changed their cut-score in Round 3; two reported changing their cut-scores as a result of something other than the impact data, and one participant reported changing his/her cut-score based on impact data. All who elected not to change their bookmark at Round 3 were among those who indicated that they would defend the cut-score if asked. See Tables 1 and 2 for a summary of responses.

Most participants expressed confidence in the validity of the standard setting process. A large majority of participants generally or somewhat agreed with the item ordering found in the booklets. Although not indicated in the numeric data, one respondent reported feeling that Round 1 evaluation of Objective 1 was a training session, resulting in less valid Objective 1 bookmarks than subsequent objectives' bookmarks. However, individual objective bookmarks were simply used as a starting point for the exam's cut-score and no single objective in Round 1 should have a large influence on the final cut-score. Although most participants expressed satisfaction with the process, confidence in the cut-score, and appreciation for the workshop as a form of professional development, it

was clear that the process was not perfect. Three participants stated they would not defend the cut-score. However, the proportion of participants who defended the cut-score was similar to the proportion who indicated the same during a prior year's bookmark standard setting. Moreover, confidence in the order of the ordered item booklets increased in the current Mapmark standard setting (88.9%) relative to the prior year's bookmark standard setting (68%), in which items were combined across objectives. However, given that the prior year's standard setting involved a different test and different participants, comparisons across years were made cautiously. The majority of participants indicated they would use the information gained through the standard setting process to enhance their pedagogy.

Both faculty leaders had also been involved in the prior year's bookmark standard setting and noted that the Mapmark was an improvement over the bookmark method. In particular, the Mapmark allowed participants to consider each of the four Objectives individually, while at the same time setting one cut-score. In their estimation, the Mapmark method was a success. However, it was also important to evaluate other forms of evidence.

Internal-Consistency Evidence

In addition to procedural evidence, evaluation of internal consistency of participants' ratings is also a component of a sound validity argument (Kane, 1994; 2001). Figure 2 portrays individual participants' cut-scores across the three rounds. Note that variation in Group 1 participants' cut-scores decreased across the three rounds (e.g., cut-scores converged). In contrast, the remaining groups' ratings converged at Round 2, following table discussions. However, following Round 3 discussions, some participants changed their cut-scores. Although there was still variation in participants' final cut-scores, the least variability was following the Round 3 discussion.

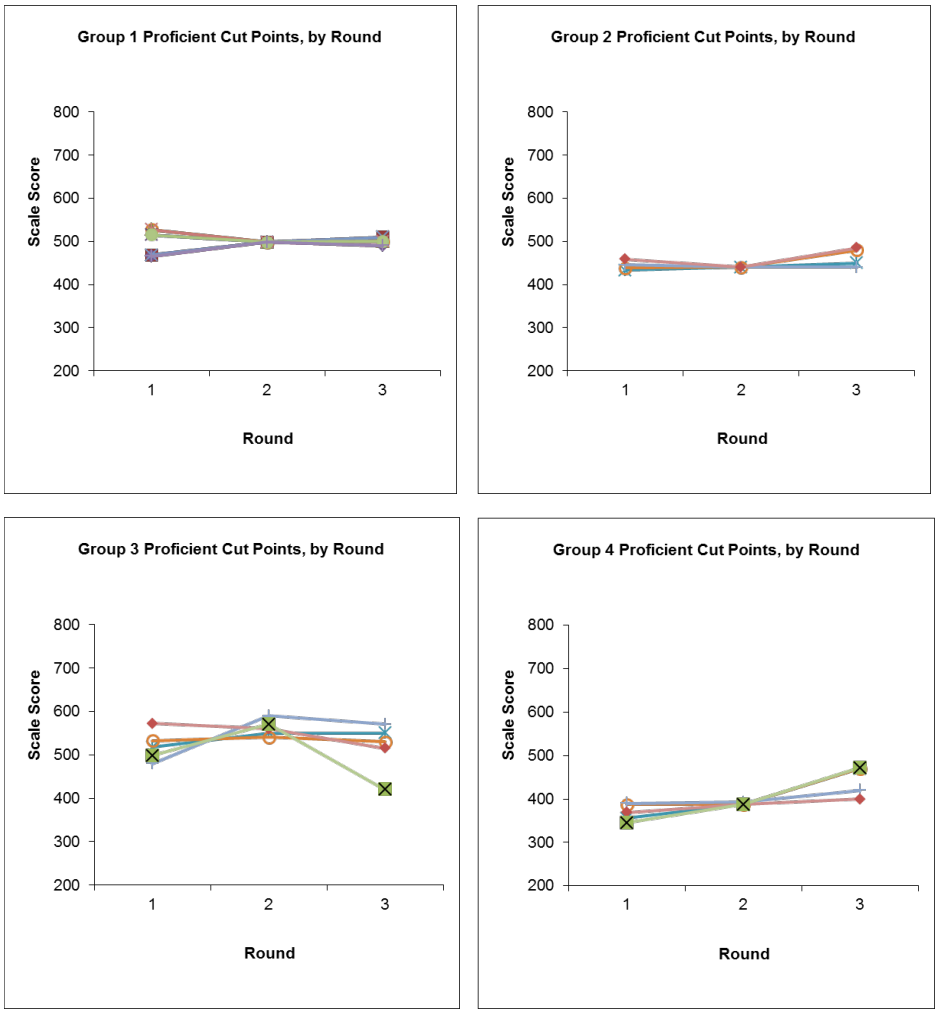

Figure 2: *Variance in Bookmarks*

Hypothetically, the standard error of the cut-score would be the standard deviation of the final cut-score set at each of an infinite number of workshops, with different participants at each workshop. Because the table groups are relatively independent at Round 2, data from Round 2 are typically used in the estimate of the standard error (Lewis et al., 1998; Mitzel et al., 2001), calculated as:

$$SE = \sqrt{\frac{s^2}{N}\left[1 + (n-1)r\right]},$$

where s^2 is the variance of the cut-scores, N is the total number of participants, n is the number of groups, and r is the intraclass correlation, which adjusts the SE to take into account dependency within group. If the median Round 3 cut-scores from different workshops were more alike than the median Round 2 cut-scores from different tables within the same workshop, this would be an overestimate of the standard error (or conversely, it would be an underestimate if groups were more alike within workshops than between workshops). In Figure 2, it is evident that the variance within groups is much smaller than the variance across groups; the intraclass correlation is 0.92. Thus, the estimated standard error of the cut-score was 32.9; it would have been 16.6 simply using the unadjusted standard error of the mean.

External Evidence

Finally, the collection of external validity evidence contributes to a strong validity argument (Kane, 1994, 2001). One form of external validity evidence for the current test is whether the cut-score can aid in identifying groups of students that may need extra support. Anecdotal and empirical evidence (i.e., average percent correct) at the university in which the current study was conducted identified several groups that seem to struggle with passing the test. For the purpose of understanding student performance on the test, examination of Developing/Proficient rates using the cut-score were computed, identifying groups who are still in the Developing category. Analysis of the previous year's data indicated that there was a large group of international students (70.4% Fall 2013; 77.8% Spring 2014) identified as Developing (not yet Proficient). Across both semesters, male students, on average, scored below the cut-score; whereas female students' average was above the cut-score. In sum, the external evidence that was available pointed to meaningful interpretations when applying the cut-score.

In sum, recognizing that further study and direct comparisons with other standard setting methods should be conducted, we cautiously recommend the modified Mapmark process for use by higher education practitioners when evaluation of items by objective or domain is desired.

Discussion and Conclusions

The current study presents an application of the Mapmark standard setting procedure to a higher education setting, during which a standard was set for a test mapping to multiple learning objectives. Other distinctive features of the process were that faculty leaders were highly involved throughout the standard setting, and that, with the exception that we required participants to complete the test prior to the standard setting, the standard setting occurred in only one day. In general, faculty leaders and participants expressed appreciation for the process and most supported the standard that was set. Nonetheless, the process was not perfect and the subjectivity and arbitrariness inherent within any standard setting was evident in the procedural validity feedback from participants.

The cut-score adopted in the current study is used for program assessment purposes. However, there are other reasons that higher education assessment practitioners may want to create a cut-score. For example, unlike the current study, in a previous standard setting we set a cut-score for our university's information literacy assessment test, in which the cut-score is used for pass/fail determinations. Students who do not meet the cut-score are required to repeat the test, until they have mastered the test at a proficient level of competency. Another use for cut-scores within higher education is for university placement. For example, performance on foreign language or mathematics tests frequently determine placement into the appropriate level of language or mathematics course. The procedures described in this study are applicable across these standard-setting contexts.

Future Study and Limitations

As mentioned by Kane (1994), "There is no gold standard. There is not even a silver standard" (p. 448). Comparing cut-score classification with a direct behavioral assessment would provide validity evidence for the performance descriptor and the cut-score. Conducting a standard setting for the communications test using another standard setting method (e.g., Angoff) and comparing results would provide further external validity evidence (Kane, 1994). However, doing so in an applied context where participant time is costly would be prohibitive and outside the mission of practitioners at the university. Continued application of the method and ongoing evaluation of validity evidence for resulting cut-scores is warranted.

Practical Suggestions

In sum, recognizing that further study and direct comparisons with other standard setting methods should be conducted, we cautiously recommend the modified Mapmark process for use by higher education practitioners when evaluation of items by objective or domain is desired. The concept of the holistic item map was easily grasped by workshop participants and the process resulted in a cut-score that was approved by most participants. The following are some practical suggestions that one may want to consider if planning a similar standard setting.

Detailed performance level descriptors should be reviewed at the beginning of the standard setting and be provided for participants to consult throughout the session. Without detailed descriptors, participants may rely on their own personal definitions of competence, resulting in greater variation in cut-scores than desired (Kane, 1998; 2001). Flexibility in the schedule is also recommended. Given that Round 1 involves the careful identification of the knowledge, skills, and abilities required to correctly answer each item, it is important to allow participants enough time to fully complete this step. Allowing some flexibility within the schedule permits organizers the opportunity to lengthen the time allotted to the various rounds, as needed.

Finally, assessment practitioners who conduct standard settings within higher education may want to consider involving faculty leaders throughout the process. The faculty leaders' involvement lent credibility—they were curricular leaders and colleagues to the participants. Faculty leaders provided a perspective that resonated with participants, they supported and defended the assessment process, and they were able to provide an educational perspective to the discussion. Consequently, Round 3 discussions were lively and collegial. Faculty members who teach downstream from the communications course counted the experience as professional development and expressed appreciation for knowing what to expect of students' communication knowledge, skills, and abilities. Nonetheless, when including faculty leaders it is important to consider whether unwanted influence on ratings is introduced through their participation. In the current study, we felt that course director participation enhanced the process and outweighed any potential sources of bias. However, there may be situations in which this is not the case, and assessment practitioners would want to take sole responsibility for the workshop.

Conclusion

The current study offers support for an adaptation of the Mapmark standard setting method to a higher educational setting. Inclusion of the Mapmark item map in Rounds 2 and 3 of the bookmark standard setting allowed participants to consider information from all four objectives at one glance. Participants and faculty leaders reported that the process was intuitive, and there was support for a defensible cut-score from the majority of participants and the faculty leaders.

References

ACT, Inc. (2007). *Developing achievement levels on the 2006 national assessment of educational progress in grade twelve economics: Progress report.* Iowa City, IA: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Borque, M. L., & Hambleton, R. K. (1993). Setting performance standards on the national assessment of educational progress. *Measurement & Evaluation in Counseling & Development, 26*, 41–47.

Cizek, G. J. (2001). Conjectures on the rise and call of standard setting: An introduction to context and practice. In G. J. Cizek (Ed.), *Setting performance standards* (pp. 3–51). Mahwah, NJ: Lawrence Erlbaum.

Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice, 23*(4), 31–31.

DeMars, C. E., Sundre, D. L, & Wise, S. L. (2002). Standard setting: A systematic approach to interpreting student learning. *Journal of General Education, 51*, 1-20.

Hambleton, R.K. (2001) Setting performance standards on educational assessments and criteria for evaluating the process. In G.J. Cizek (Ed.), *Setting performance standards* (pp. 89–116). Mahwah, NJ: Lawrence Erlbaum.

Impara, J.C., & Plake, B.S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement, 34*, 353–366.

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*, 69–81.

Kahraman, N., & Thompson, T. (2011). Relating unidimensional IRT parameters to a multidimensional response space: A review of two alternative projection IRT models for scoring subscales. *Journal of Educational Measurement, 48*, 581–601.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 62*, 425–461.

Kane, M. (1998). Choosing between examinee-centered and test-centered standard-setting methods. *Educational Assessment, 5*, 129–145.

Kane, M. (2001). So much remains the same: Conception and status of validation in setting standards. In G.J. Cizek (Ed.), *Setting performance standards* (pp. 53–88). Mahwah, NJ: Lawrence Erlbaum.

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

Lewis, D. M., Green, D. R., Mitzel, H. C., Baum, K., & Patz, R. J. (April, 1998). *The bookmark standard setting procedure: Methodology and recent implementations*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA.

Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research, 49*, 305–328.

Mitzel, H.C., Lewis, D.M., Patz, R.J., & Green, D.R. (2001). The bookmark procedure: Psychological perspectives. In G.J. Cizek (Ed.), *Setting performance standards* (pp. 249–281). Mahwah, NJ: Lawrence Erlbaum.

National Academies of Sciences (2005). *Measuring literacy: Performance levels for adults, interim report* Available from http://www.nap.edu/catalog/11267/measuring-literacy-performance-levels-for-adults

Perie, M. (2008). A guide to understanding and developing performance-level descriptors. *Educational Measurement: Issues and Practice, 27*(4), 15–29.

Plake, B.S. (2008). Standard setters: Stand up and take a stand! *Educational Measurement: Issues and Practice, 27*(1), 3–9.

Schulz, E.M., & Mitzel, H.C. (2011). A Mapmark method of standard setting as implemented for the National Assessment Governing Board. *Journal of Applied Measurement, 12*, 165–193.

Tong, Y., Patterson, B., Swerdzewski, P., & Shyer, C. (2014, April). *Standard setting for a Common Core aligned assessment*. Paper presented at the annual meeting of the National Council on Measurement in Education, Philadelphia, PA.