APPLYING THE BOOKMARK METHOD TO MEDICAL EDUCATION: STANDARD SETTING FOR AN ASEPTIC TECHNIQUE STATION

MONICA L. LYPSON
STEVEN M. DOWNING
LARRY D. GRUPPEN
RACHEL YUDKOWSKY

Medical Teacher. 2013; 35(7):581-585.

ABSTRACT

Introduction

The purpose was to evaluate the Bookmark standard-setting method for use on a performance-based assessment in medical education.

Methods:

We compared cutscores for Aseptic performance assessment using the modified Angoff, Hofstee and modified Bookmark methods.

Results:

The Angoff produced a cutscore of 62%, SD=18 and a percent passing (pp)=64%. The Hofstee cutscore was 71%, SD=7 and pp= 46%. Bookmark mean cutscores were 65.9% SD=10.7 and pp=42% for advanced beginners; 83.6%, SD=9.2 and pp=17% for competent and the proficient category resulted in a cutscore of 96.4% SD= 3.9 and pp=1%. Faculty judges found the bookmark method to be an easy and acceptable method.

Conclusions:

The Bookmark method was acceptable to faculty, has reasonable quality metrics when compared to other methods and can be a practical tool for establishing standards in performance-based examinations. The Bookmark method could be useful for establishing multiple levels of competency using the Dreyfus criteria.

Practice Points:

1. A modified Bookmark method is suitable for a medical education performance based examination.
2. The Bookmark method is more easily understood and feasible from the judges' perspective.
3. The Bookmark method may produce advantages in a post-graduate medical training setting given the adaptability to level of competence & the focus on milestones.

Declaration of Interest

The authors report no declarations of interest.

Introduction

An understanding of passing standards is essential for establishing competency guidelines and performance milestones, yet standard setting exercises in post-graduate medical education (PGME) have only recently begun to surface in the literature (Wayne et al. 2007, Wayne et al. 2008, Hicks et al. 2010). Standards for performance tests in medical education typically are set by methods such as Angoff, Ebel, Hofstee, and Borderline Group Method (Cusimano 1996; Ben-David 2000; Wass et al. 2001; Zieky 2001; Boulet et al. 2003; Downing et al. 2006; Yudkowsky et al. 2009). The Bookmark method is a relatively new approach developed in 1996 by the CTB/McGraw-Hill research group for use in K-12 education. A key feature of the Bookmark method is the Ordered Item Booklet, which contains the set of items placed on a continuum from simple to difficult and asks judges to determine the placement of the cut score on this continuum (Lewis et al. 1998; Lewis et al. 1999; Cizek 2001; Mitzel et al. 2001; Cizek et al. 2004; Karantonis & Sireci 2006). In this model, item difficulty is determined empirically, typically after calibrating the item using Item Response Theory (IRT) methods (Karantonis and Sireci, 2006). We used the proxy of item difficulty (classical test theory) as defined by trainee performance and not an IRT estimate of item difficulty as calibrated by examinee expertise. While others have used this method as well, the impact on the final standards of using item difficulty instead of IRT are not well understood.

This method is particularly useful for obtaining cut scores that differentiate examinees at different performance levels (e.g. advanced beginners, competent and proficient) (Dreyfus & Dreyfus 1986). While the Bookmark method is currently one of the most popular standard-setting methods in primary education in the United States, there has

been limited usage of this technique in medical education in general and in performance-based examinations specifically (Lewis et al. 1998). In 2007, Ahn and Ahn, used the Bookmark method for the Korean National Medical Licensing Examination and recommended it for further use; they found the Bookmark method to require shorter time commitments compared to the Angoff (Ahn & Ahn 2007).

At the University of Michigan, an intern Objective Standardized Clinical Examination (OSCE) called the Post-graduate Orientation Assessment (POA) is administered during orientation, providing formative feedback on basic competencies for incoming residents as determined by residency program directors (Lypson et al. 2004; Lypson et al. 2010). One of the stations on this OSCE assesses competency at performing aseptic technique, scored by means of a twenty-item checklist (Lypson et al. 2004). In previous work, we found the Aseptic Technique checklist to have satisfactory inter-rater reliability and internal consistency within instrument. It also exhibits validity evidence, making it useful for competency assessment (Lypson et al. 2004). In addition, this station has a unique assessment strategy using expert nurses as graders and includes many Joint Commission patient safety standards–making it a novel tool for hospital accreditation.

In this study, we sought to answer the following research questions: (1) is the Bookmark approach a feasible method for a PGME performance test as represented by the aseptic technique station and (2) will a group of residency educators find this method acceptable for setting standards?

METHODS

The Institutional Review Boards at the University of Michigan and the University of Illinois at Chicago approved this study.

Participants/Data

An expert panel of sixteen judges was purposefully selected based on their expertise in medical education and experience in PGME (Fowell et al. 2008).

This sample included nine physician educators from various specialties (Internal Medicine, Plastic Surgery, Pediatrics, Family Medicine, General Surgery, Obstetrics & Gynecology, and Physical Medicine & Rehabilitation), three PhD medical educators (Surgery and Medical Education), two standardized patient educators (master's level education) and standardized patient involved in residency education initiatives.

Data Collection

Three separate training sessions were held for the judges. 1. Orientation session: judges were provided with detailed information and training on all three standard-setting methods (Modified Angoff, Hofstee and Bookmark methods). 2. Round 1: The judges provided Modified Angoff, Hofstee, and Bookmark method judgments in turn, without access to examinee performance data. 3. Round 2: Two weeks later the panel discussed their judgments and resulting cut score determinations; they were then given item-level performance data from 177 interns who participated in the POA in 2008. The judges then provided a second round of judgments for the Bookmark and Hofstee methods. A second round of the Angoff was not performed due to unanimous agreement among the judges that performance data would not change their previous judgments. In order to prevent sequence effects, we changed the order of the standard setting exercises: in round one, the Hofstee method was conducted first and in round two, the modified Bookmark method was done first.

Modified Angoff Method

The Angoff Method is an absolute standard technique used frequently in medical education for performance-based assessments (Boulet et al., 2003). Panelist/judges are asked to provide an estimate of the percentage of borderline examinees that will answer an item correctly. Often a modified approach is used, in which judges discuss their ratings in subsequent rounds after receiving performance data if available. The original approach was a yes/no approach (will they or won't they) but most people use a modified Angoff that asks for the percent correct or probability of a correct response. This information is then averaged across judges for each item (Boulet et al., 2003).

Hofstee Method

The Hofstee method is often described as an example of the compromise approach that has been developed to utilize the advantages of both relative and absolute standard-setting procedures (Case and Swanson, 1996). This information is plotted with exam performance and the point where the information intersects the performance curve indicated the standard (DeGruijter, 1985). Panel members provided four parameters: (1) minimum acceptable passing score, (2) maximum acceptable passing score, (3) minimum acceptable failure rate, and (4) maximum acceptable failure rate (Hofstee 1983; Downing et al. 2006; Fowell et al. 2008).

Modified Bookmark Method

The Bookmark method was used to categorize performance into three levels based on the Dreyfus model of expertise: advanced beginner, competent and proficient. the Dreyfus categories of novice and expert were not used for this exercise (Dreyfus & Dreyfus 1986). After extensive discussion, judges reached consensus to ensure a common understanding of these categories. The judges and facilitator decided that the definition of "competent" for aseptic technique was the lowest acceptable threshold of performance and was equivalent to the borderline intern in the modified Angoff and Hofstee methods.

Item Response Theory (IRT) parameters were not available to order the items on the checklist. Impara (1997) and Buckendal (2002) used classical test/measurement theory instead of IRT to sequence items in order of difficulty. These scholars argued that using classical test theory allows for better understanding on the part of judges thereby producing an alternative that can be operationalized in almost any setting (Meskauskas 1986; Green et al. 2009b). Following the classical test theory approach our ordered item booklet provided test items listed for the Aseptic Technique station which were arranged according to item difficulty, interpreted as percent correct response on the checklist and based on the performance in the "done" category only (Impara and Plake, 1997; Buckendahl et al., 2002). Judges were asked to indicate within the booklet the point at which a learner would be placed in the Dreyfus "advanced beginner", "competent" and "proficient" categories. According to Bookmark procedures, the judges were instructed that the item preceding the "bookmark" should be interpreted as the point at which at least two-thirds of the examinees at a given level are likely to respond correctly (Lewis et al. 1998; Lewis et al. 1999).

Cut scores and resulting pass rates were calculated for the class of 2008 using each of the three methods. The Angoff and Hofstee cut scores were calculated using means across

judges, and the Bookmark cut score was calculated (a) using medians, per Bookmark guidelines, and (b) using means, to allow for comparisons across methods. The use of the mean cut score should be considered a further modification of the bookmark method.

The feasibility and acceptability of each standard-setting method was evaluated by: (a) quality indices for judges' results as defined below; (b) judges' feedback about the credibility and difficulty of each method, and confidence in the resulting cut scores immediately after the session and 9 months, and (c) the time commitment required for judges to perform the exercise (evaluated immediately afterwards) (Cizek 2001; Boulet et al. 2003; Yudkowsky et al. 2009).

Quality Metrics

Over the past three decades, attention has been given to the characteristics of the cut score in order to provide evidence to stakeholders about the value, quality, credibility, or defensibility of a given cut score (Cizek, 2001). One method of measuring quality is to assess the level of agreement between judges in their determination of the score. This study used two quality metrics: (1) the standard deviation of the Cut Score Judgment (Cusimano 1996; Cusimano & Rothman 2003; Yudkowsky et al. 2009) and (2) the Meskauskas SIS1 index. The Meskauskas SIS1 index is the ratio of the standard deviation of the students' scores to the standard deviation of the judges' cut-scores (Meskauskas 1986). The index was calculated as SD (student scores) / SD (cut-score judgments). The standard deviations of the judgments should ideally be small relative to the SD of student scores, resulting in a Meskauskas index greater than 4.0. In addition, the standard error was calculated for each method. Quality metrics for judges' cut scores typically average all of the judges' decisions

and then look at the mean and standard deviations. The bookmark method however uses a median score to calculate the cut scores at various levels. Thus, for this study, we use both the median cut score and the mean scores to allow for comparisons across methods. The use of the mean cut score should be considered a further modification of the bookmark method.

RESULTS

Table 1 presents the cut scores (percent correct) and resulting passing rates for each standard-setting method. The Bookmark method resulted in the most stringent cut score (84.6%). The Angoff was the most lenient standard at 64%, close to the cut score of 62.5% representing the "advanced beginner" category in the Bookmark method (Table 1). The Hofstee method produced a cut score at an intermediate level of 71%. Using the Angoff method, 64% of beginning interns would pass, compared with 46% using the Hofstee method and only 23% using the Bookmark method.

Table 2 illustrates the application of the Bookmark method results to delineate four levels of performance. Using these cut scores 40% of interns were rated as "novice", 42% as "advanced beginner", 17% as "competent" and only 1% as "proficient" or above.

The quality indices of the judgments are provided in Tables 1 and 2. The Angoff method produced the lowest level of inter-judge agreement. The Bookmark method demonstrated better quality metrics (more agreement among the judges) than did the Angoff. The standard errors varied per round given the variable number of judges involved.

Survey results: Judges were very satisfied with their roles and their cut score determinations. Half (50%) of the judges felt that the Bookmark method offered the easiest approach to standard setting when compared to other methods. Sixty percent of the judges

reported that the Angoff took the longest to complete compared to 20% of the judges in regards to the Bookmark method. While judges found the Bookmark method to be an acceptable and easy method for standard-setting they still considered the Angoff method to be more trustworthy compared to the other methods; they often mentioned the use of Angoff for the national board examinations as the reason for judging it more trustworthy: "If they use it, it must be better".

DISCUSSION

We found the Bookmark method to be a feasible and acceptable means of determining standards for a performance examination. An advantage of the Bookmark Method is that it can provide useful information about multiple levels of performance. The American Board of Internal Medicine (ABIM) recently convened a task force that adopted the Dreyfus model for skill acquisition, detailing milestones of achievement using a five-step progression to competence (Green et al. 2009). The Bookmark method could be used to set standards for achievement at different Dreyfus levels. As the Accreditation Council for Graduate Medical Education (ACGME) and other certification boards develop milestones, the ability to set standards at different levels will be crucial to determining and tracking developmental progression for various skill-based competencies.

The quality indices of the Bookmark method were acceptable, and better than those of the more traditional Angoff method. This provides further validity evidence for the Bookmark method in medical education performance-based assessments (Meskauskas 1986; Zieky 2001).

The Bookmark method resulted in more stringent cut scores than the Angoff or Hofstee, even though the judges considered the definition of "competence" to be the same as that of the "borderline intern." Judges felt that the failure rate 80% of was acceptable given the formative nature of the exam and the risk of iatrogenic infection in cases of poor technique (Green et al. 2009). Judges were willing to defend the standards resulting from the Bookmark method; nonetheless, they still considered the Angoff method to be the most trustworthy, perhaps due to their familiarity with the Angoff method, its use in national-level high stakes assessments, and its detailed item-level judgments.

The Bookmark's "advanced beginner" cut score of 65% was similar to the Angoff cut score of 64%. This suggests that the "borderline" intern may be perceived as an advanced beginner who is not quite ready to perform the procedure on their own. The Bookmark method produced the most stringent standard for competence. If we use this procedure only 18-20% of the interns would be competent at Aseptic technique. This would mean remediation for approximately 80% of interns. This may be acceptable in this formative assessment, but it is not clear how the judges would view this outcome if this was a high stakes examination. Patients, however, might agree that this standard is more than acceptable when dealing with the consequences of poor technique which would mean iatrogenic infection (Green et al., 2009).

There were several limitations to this study. The judgments were based on a group of faculty from a single North American medical school. We used the proxy of item difficulty (classical test theory) as defined by trainee performance and not an IRT estimate of item difficulty. The impact of using item difficulty instead of IRT is not well understood.

CONCLUSION

The goal of this study was to determine the feasibility and acceptability of implementing a modified Bookmark method for a PGME performance test. The Bookmark method was easily learned, acceptable to faculty, and demonstrated acceptable quality indicators. Based on these criteria, the Bookmark method is a reasonable and constructive approach to standard-setting in the post-graduate arena, where competency and the measurement of milestones are an integral part of the educational process.

REFERENCES

Ahn DS, Ahn S. 2007. Reconsidering the cut score of Korean National Medical Licensing Examination. J Educ Eval Health Prof 4:1-6.

Ben-David M. 2000. AMEE Guide No. 18: Standard setting in student assessment. Med Teach 22(2):120-130.

Boulet JR, Dechamplain AF, Mckinley DW. 2003. Setting defensive performance standards on OSCEs and standardized patient examinations. Med Teach 25(3):245-249.

Buckendahl CW., Russell WS, Impara JC, and Plake BS. 2002. A comparison of Angoff and Bookmark standard setting methods. J Educ Measur 39(3): 253-63.

Case SM, Swanson DB. Constructing written test questions for the basic and clinical sciences. Philadelphia: National Board of Medical Examiners, 1996.

Cizek GJ (ed.) 2001. Setting performance standards: Concepts, methods, and perspectives. Mahwah, NJ: Lawrence Erlbaum Associates.

Cizek GJ, Bunch MB, Koons H. 2004. Setting performance standards: Contemporary methods. Educ Measur: Issues Pract 23(4):31-50.

Cusimano MD. 1996. Standard setting in medical education. Acad Med 71(10 Suppl):S112-S120.

Cusimano MD, Rothman I. 2003. The effect of incorporating normative data into a criterion-referenced standard setting in medical education. Acad Med 78(10 Suppl):S88-S90.

DeGruijter DNM. 1985. Compromise models for establishing examination standards. J Educ Measur 22(4):263-69.

Downing SM, Tekaian A, Yudkowsky R. 2006. Procedures for establishing defensible absolute passing scores on performance examinations in health professions education. Teach Learn Med 18(1):50-57.

Dreyfus H, Dreyfus S. 1986. Mind over machine. New York, Free Press.

Fowell SL, Fewtrell R, Mclaughlin PJ. 2008. Estimating the minimum number of judges required for test-centered standard setting on written assessments. Do discussions and iteration have an influence? Adv Health Sci Educ 13(1):11-24.

Green ML, Aagaard EM, Caverzagie KJ, Chick DA, Holmboe E, Kane G, Smith CD, Iobst W. 2009. Charting the road to competence: Developmental milestones for internal medicine residency training. J Grad Med Educ 1(1):15-20.

Hicks PJ, Englander R, Schumacher DJ, Burke A, Benson BJ, Guralnick S, Ludwig, S, Carraccio C. 2010. Pediatrics milestone project: Next steps toward meaningful outcomes assessment. J Grad Med Edu 2(3):577-584.

Hofstee WKB 1983. The case for compromise in educational selection and grading *In:* Anderson, S. & Helmick, J. (eds.) On educational testing. San Francisco: Jossey-Bass Publishers.

Impara JC, Plake BS. 1997. Standard setting: An alternative approach. J Educ Measur 34(1): 353-366.

Karantonis A, Sireci SG. 2006. The bookmark standard-setting method: A literature review. Educ Measur 25:4-12.

Lewis D, Green D, Mitzel H, Baum K, Patz R 1998. The bookmark standard setting procedures: Methodology and recent implementations. *Annual Meeting of the National Council on Measurement in Education.* San Diego, CA.

Lewis D, Mitzel H, Green D, Patz R. 1999. The bookmark standard setting procedures. Monterey, CA, McGraw Hill.

Lypson ML, Frohna JG, Gruppen LD, Woolliscroft JO. 2004. Assessing residents' competencies at baseline: Identifying the gaps. Acad Med 79(6):564-570.

Lypson ML, Hamsta SJ, Ross PT, Gruppen LD, Colletti LM. 2010. An assesment tool for aseptic technique in resident physicians: A journey towards validation in the real word of limited supervision. J Grad Med Educ 2(1):85-89.

Meskauskas JA. 1986. Setting standards for credentialing examinations. Eval Health Prof 9(2):187-203.

Mitzel HC, Lewis DM, Patz RJ, Green DR 2001. The bookmark procedure: Psychological perspectives. *In:* Cizek, G. J. (ed.) Setting performance standards: Concepts, methods and perspectives. Mahwah, NJ: Lawrence Erlbaum Associates.

Wass V, Van der Vleuten C, Shatzer J, Jones R. 2001. Assessment of clinical competence. Lancet 357:S17-S20.

Wayne DB, Barsuk JH, Cohen E, Mcgaghie WC. 2007. Do baseline data influence standard setting for a clinical skills examination? Acad Med 82(10 Suppl):S105-S108.

Wayne DB, Cohen E, Makoul G, Mcgaghie WC. 2008. The impact of judge selection on standard setting for a patient survey of physician communication skills. Acad Med 83 (10 Suppl):S17-S20.

Yudkowsky R**,** Downing S, Tekian A. 2009. Standard Setting. In Downing SM and Yudkowsky R (eds): *Assessment in Health Professions Education*, New York.

Zieky MJ 2001. So much has changed: How the setting of cutscores have evolved since the 1980s. *In:* Cizek, G. J. (ed.) Setting performance standards: Concepts, methods, and perspectives. Mahwah, NJ: Lawrence Erlbaum Associates.

ACKNOWLEDGEMENT

This research was carried out in partial fulfillment of a master's level thesis at the University of Illinois.

TABLE 1
RESULTS OF THE STANDARD SETTING EXERCISE FOR ASEPTIC TECHNIQUE

Judges (n)	Median Minimal Pass Level Cut-Score % Correct	Mean Minimal Pass Level Cut-Score % Correct	% Pass	SIS-1	SD of Judges' Cut-Scores	Standard Error of Judges' Cut-Scores
Angoff – Round 1 (n=14)	n/a	64%	64%	0.7	18	5
Hofstee – Round 1 Without Performance Data (n=12)	n/a	74%	41%	1.7	8	2
Hofstee – Round 2 With Performance Data (n=14)	n/a	71%	46%	2.2	6	2
Bookmark – Round 1 Without Performance Data (n=13)	85% (Competent)	84.6% (Competent)	16%	1.5	8.8	2.4
Bookmark – Round 2 With Performance Data (n=11)	85% (Competent)	83.6 % (Competent)	17%	1.4	9.2	1.9

TABLE 2
THE BOOKMARK METHOD STANDARDS FOR ASEPTIC TECHNIQUE

Calculation	Advanced Beginning Bookmark Cut-score	% of Interns at the Advanced Beginner level	Competent Bookmark Cut-score	% of Interns at the Competent level	Proficient Bookmark Cut-score	%Interns at the Proficient level
Round 1 without Performance Data (n=13)						
MEDIAN Minimal Pass Level Cut Score	55	69%	85	16%	95	4%
MEAN Minimal Pass Level Cut Score (SD)	59.5 (15.9)	59%	84.6 (8.8)	16%	96.2 (3.6)	2%
Meskauskas index SIS1	0.8		1.5		3.7	
Round 2 with Performance Data (n=11)						
MEDIAN Minimal Pass Level Cut Score	65	46%	85	16%	95	2%
MEAN Minimal Pass Level Cut Score(SD)	65.9 (10.7)	42%	83.6 (9.2)	17%	96.4 (3.9)	1%
Meskauskas index SIS1	1.2		1.4		3.4	