The Alberta Journal of Educational Research *Vol. 52, No. 1, Spring 2006, 65-79*

Nizam Radwan

and

W. Todd Rogers
University of Alberta

A Critical Analysis of the Body of Work Method for Setting Cut-Scores

The recent increase in the use of constructed-response items in educational assessment and the dissatisfaction with the nature of the decision that the judges must make using traditional standard-setting methods created a need to develop new and effective standard-setting procedures for tests that include both multiple-choice and constructed-response items. The Body of Work (BoW) method is an examinee-centered method for setting cut-scores that applies a holistic approach to student work in order to estimate the cut-scores that differentiate examinees according to their level of performance in situations where both item formats are used. A detailed review of Version 1 and the recent modification, Version 2, are first presented followed by a critical evaluation of the two versions in terms of Berk's (1986) 10 criteria for defensibility. The results reveal that the BoW method appears to be a promising method for setting cut-scores that could be used on a wider scale in Canada. However, as with other methods, the experience gained from using the BoW method in the field will probably lead to further modifications in an attempt to increase efficiency without sacrificing accuracy.

La décision d'augmenter l'emploi des questions à réponse construite dans l'évaluation scolaire a provoqué de l'insatisfaction à l'égard des décisions que doivent prendre les juges qui utilisent des méthodes traditionnelles pour établir les normes. Ce mécontentement a entraîné le besoin de développer de nouvelles procédures efficaces pour établir des normes dans le cas d'examens comportant des questions à choix multiples aussi bien que des questions à réponse construite. La méthode Body of Work (BoW) est une méthode d'établissement de notes de passage qui est centrée sur le candidat et qui repose sur une approche holistique au travail de l'élève pour déterminer les notes de passage qui distinguent les candidats selon leur rendement dans des situations impliquant les deux sortes de questions. Un examen détaillé de la 1re version et de la modification récente qu'on en a faite (la 2e version) est suivi d'une évaluation critique des deux versions d'après les 10 critères de validation établis par Berk (1986). D'après les résultats, la méthode BoW pourrait bien servir dans l'établissement des notes de passage et être mieux diffusée partout au Canada. Comme c'est le cas pour toutes les autres méthodes par contre, l'expérience qu'on retire de l'emploi de la méthode BoW entraînera probablement des modifications visant à la rendre plus efficace sans toutefois en sacrifier la précision.

Introduction

We are witnessing today a greater use of constructed-response items in large-scale, high-stakes tests. This increased use can be traced to concerns about sole reliance on multiple-choice items and a desire to have tests that better reflect

Nizam Radwan is a doctoral graduate and is currently working as a psychometrician in the Education Quality and Accountability Office, Toronto.
Todd Rogers is a professor and Director of the Centre for Research in Applied Measurement and Evaluation. His research interests are in test translation, ethics in testing, and psychometrics.

the learning outcomes students are expected to meet. Shepard (1991) and Wiggins (1993), for example, pointed out that the use of only multiple-choice items, the responses to which are simply scored right or wrong, results in stress being placed on the development of lower- rather than higher-level thinking skills and a consonant narrowing of the curriculum and what is to be learned. This message has been heard, with an attendant increase in the use of constructed-response items that require a higher level of thinking to formulate acceptable responses. Most often the degree of acceptability of responses to constructed-response items or task varies, leading to polytomous and not dichotomous scoring

At the same time, educational officials wish to know if students are performing at acceptable levels described in terms of performance standards (Kane, 2001). Cut-scores that separate adjacent levels of performance (*master/non-master; advanced/proficient; basic/below basic; certified/not certified*) need to be set in the distribution of scores yielded by a test consisting of a set of items that are relevant to and representative of the performance standards. Before the inclusion of constructed-response items, cut-score procedures were developed for dichotomously scored items. The most popular of these is the Angoff (1971) procedure or one of its modifications (Ricker, in press). The Angoff procedures require that a panel of qualified experts responsible for setting the cut-score that separates two adjacent performance levels think of a group of examinees who just qualify for the higher of the two categories (e.g., a group of minimally competent *proficient* students for the *proficient vs. basic* levels). The panelists are asked to estimate the proportion of each group of minimally competent students who would pass each item. However, determining the proportion of a hypothetical group of minimally competent students who will correctly answer an item is not a natural process used by panelists in their daily work. Furthermore, it is difficult and time-consuming to set cut-scores consistently when there is more than one cut-score. Last, procedures like the Angoff procedures are not well suited for polytomously scored items.

In response to these concerns, various procedures for setting cut-scores have been proposed. Two of these newer procedures—the Analytical Judgment Method and the Body of Work Method—use actual student work, be it dichotomously or polytomously scored. The first of these two methods is discussed by Abbott (in press, this issue). The purposes of this article are to describe the Body of Work (BoW) method and to assess its strengths and weaknesses in terms of Berk's (1986) 10 criteria for defensibility.

Like the other procedures used to set cut-scores, the BoW method has evolved over time, beginning with the first version used in the state of Maine in 1993 to the version now used. Two versions of the BoW are described in this article. The first version marks the midpoint of time between 1993 and the present and was used in Massachusetts and New Hampshire. The second version is what is now used (K.P. Sweeney, November 15, 2004, personal communication). These two versions were selected to illustrate the changes made to make the procedure more practical without sacrificing the accuracy of the cut-scores that are set.

The Body of Work Method

A "test" in the BoW method can consist of unrestricted constructed-response items (e.g., essays, performance assessments), restricted constructed-response items (e.g., short-answer items, mathematics problems), and/or multiple-choice items. The student responses to each constructed and multiple-choice item have been previously scored. The constructed-response items are placed first, followed by the multiple-choice items arranged in order of difficulty from easiest to most difficult. Each item is presented in an abbreviated form; panelists can refer as needed to complete wording in supporting reference materials. The score (numerical value for constructed response items and a "+" (correct) and "–" (incorrect) for multiple-choice items) that a student received for each item is reported next to the corresponding item. The test is considered as a whole; item scores are added to obtain the total test score.

The BoW method involves matching the quality of student test responses at each score point in the total score distribution with defined performance standards and then setting cut-scores at points in the total test score distribution that separate adjacent performance levels (e.g., *proficient vs. basic; basic vs. below basic*). These activities are completed in five steps:

- establishing the performance standards,
- creating folders that contain a representative sample of students' tests,
- selecting and training/calibrating members of a cut-score panel,
- matching tests to the performance standards, and
- setting the final cut-scores.

Each of these steps is described below, first for Version 1 and then for Version 2.

Version 1

Establishing the Performance Standards

The performance standards are established in two stages. First, general performance standards are prepared and named. Second, specific student behaviors are specified for each of the general standards to provide a clear indication of what students are expected to learn and be able to do at each performance level. Kingston, Kahl, Sweeney, and Bay (2001) provide the example shown in Table 1. The general descriptions correspond to general objectives for four proficiency levels, and the specific statements related to each general description correspond to instructional objectives.

The general descriptions and the names are developed by policymakers with advice from an educational advisory committee comprising educational stakeholders in a state or provincial department of education. Content specialists, including teachers and university faculty in the subject area, operationalize the general performance standards.

Creating Folders Containing Student Tests

The importance of establishing the performance standards at the level of detail reflected in the second part of Table 1 cannot be overstated (Kane, 2001). In the case of the BoW, members of the panel appointed to set the cut-scores classify a representative sample of students' tests, grouped in folders using the total test score, according to the closest match between the level of student performance on the tests in the folder and the performance standards corresponding

Table 1
Example of a Performance Standard: Grade 8 Mathematics

Proficiency Level	*Description*
Advanced	Students at this level demonstrate a comprehensive and in-depth understanding of rigorous subject matter, and provide sophisticated solutions to complex problems.
Proficient	Students at this level demonstrate a solid understanding of challenging subject matter, and solve a wide variety of problems.
Needs Improvement	Students at this level demonstrate a partial understanding of subject matter, and solve some simple problems.
Failing	Students at this level demonstrate a minimal understanding of subject matter, and do not solve even simple problems.
Specific Statements	
Proficient	Students should be able to • Demonstrate solid understanding of the numeration system • perform most calculations and estimations • define concepts and generate examples and counterexamples of concepts • represent data and mathematical relationships in multiple forms (e.g., equations, graphs) • apply learned procedures and mathematical concepts to solve a variety of problems, including multi-step problems • use a variety of reasoning methods to solve problems • explain steps and procedures • use various forms of representation (e.g., text, graphs, symbols) to illustrate steps to a solution

Kingston et al., 2001, pp. 222-223.

to each of the defined proficiency levels. The tests placed at various proficiency levels serve as an operational definition of that proficiency level. The cut-scores are placed in the total score distribution to separate the tests at the varied proficiency levels defined when establishing the performance standards.

Three types of folders are used to set the cut-scores in Version 1: pinpointing folders, the range-finding folder, and the training/calibration folder. As indicated above, each folder contains sets of student tests. For example, the Massachusetts Comprehensive Assessment System (MCAS) grade 8 mathematics test contained six constructed-response items scored on a 4-point scale, followed by five short-answer and 21 multiple-choice dichotomously ("+" or "–") scored items.

Pinpointing folders. The pinpointing folders contain clusters of students who have like test performances. Each folder contains five student tests at each of four consecutive test score values (e.g., 46, 45, 44, and 43) as shown in Figure 1. Together the pinpointing folders cover the score range from the highest obtained test score to approximately the chance test score (0.25 times the number of multiple-choice items plus the number of constructed response items (Kingston et al., 2001). The pinpointing folders and the tests in them are ordered in terms of increasing test scores. Often the top folder will span more than the top four scores because there will probably be fewer than five tests at

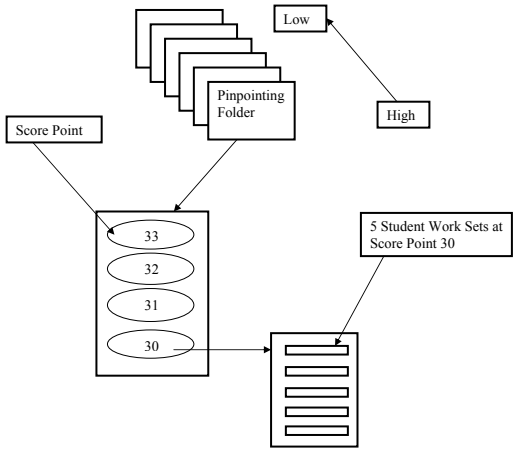

Figure 1. Hypothetical example of preparing pinpointing folders using raw scores.

each of the four highest test score values. In this case, more score points are included to get 20 papers in the top folder. To simplify the process for the panel members, tests of uneven quality (e.g., tests in which scores on the constructed-response and multiple-choice are quite different or students whose scores on the constructed-response items vary) are not selected. More specifically, the criteria used for the selection of student tests might include:

1. Consistency in scores awarded for unrestricted constructed response items: difference between the highest score and the lowest score awarded to these items is not greater than one score point;
2. Consistency between two readers for items involving a writing prompt; scorers assign the same score; and
3. Consistency between multiple-choice items and items that involve a constructed response: regression residual after predicting multiple-choice scores from constructed response scores is not greater than 0.50 (Massachusetts Department of Education, 2002, p. 53).

Range-finding folder. The range-finding folder contains the top and two bottom student tests from each pinpointing folder (Figure 2). To select the top test, the five tests at the highest score in the folder (e.g., score 46 in Figure 1) are examined to identify the test that has the best quality. To select the two bottom tests, the five tests at the lowest score in the file (e.g., score 43 in Figure 1) are examined to identify the two papers with the lowest quality (K.P. Sweeney, November 15, 2004, personal communication). Each pinpointing folder contained 20 student tests (five tests at each of four score points), so the 10 pinpointing folders included 200 student tests. The range-finding folder included 30 tests, ordered by total score (the top and two bottom tests drawn from each of the 10 pinpointing folders).

Training/calibration folder. The training/calibration folder contains the top student test from every other pinpointing folder. For example, the training folder for the MCAS grade 8 mathematics test contained the top student test from pinpointing folders 1, 3, 5, 7, and 9 (Figure 3). Because the top student test from each of these folders was included in the range-finding folder, the student tests in the training folder were also in the range-finding folder.

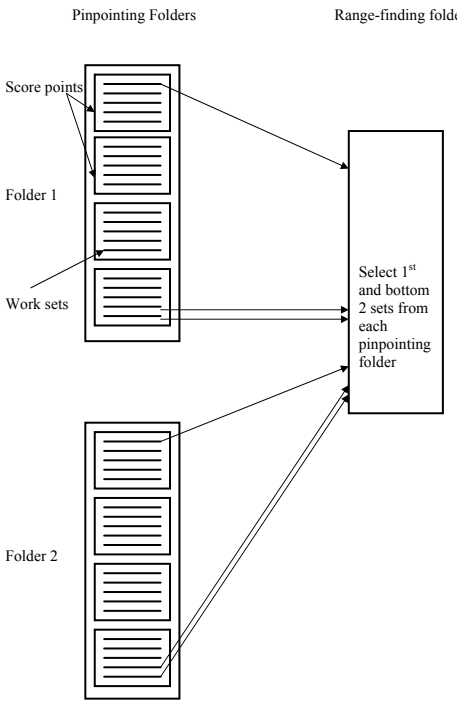

Figure 2. Hypothetical example of preparing a range-finding folder.

In all cases the formation of the folders is governed by the need to get a representative sample of student tests across the test score range from the chance score to the top or maximum score.

Selecting and Training a Cut-Score Panel

Selecting panel members. A sample of panelists representative of the major stakeholders is selected to set the cut-scores. Panel members typically include classroom teachers, administrators, and representatives of the higher education community, business community, local school committees, local or state government committees, parents, and the general public. Often greater representation is given to the groups that work more closely with the students. For example, approximately half a panel might be classroom teachers and a quarter might be school administrators, with the balance made up of representatives of the remaining groups. In the MCAS, a total of 209 panelists set the cut-scores for 12 subject areas. The total group of panelists was composed of 50.7% classroom teachers, 21.5% administrators, 16.7% business community representatives, 7.2% higher education representatives, and 3.8% school committee and local/state government representatives (Kingston et al., 2001). No breakdown was provided for the 12 panels.

Panelist training/calibration. Panel members are first asked to respond to the test questions and then score their own responses so that they become quite familiar with the test content and gain a better appreciation of the difficulty of the items. The panelists are advised not to reveal their scores or volunteer to share their scores with others. The intent of this is to make this part of the process less threatening.

Often panel members are asked to take and score their responses before coming to the location where the cut-scores will be set. However, as with other cut-score setting methods, panel members tended not to complete this task in the case of the MCAS. Hence in agreement with Kingston et al. (2001), we recommend that taking and scoring the test be completed on site as the first training step so as to ensure panelists' compliance.

Next, to further ensure that the panel members have a common understanding of the performance levels and the relationship of these levels to student performance as reflected by their test scores, the following activities, led by a trained facilitator, are needed.

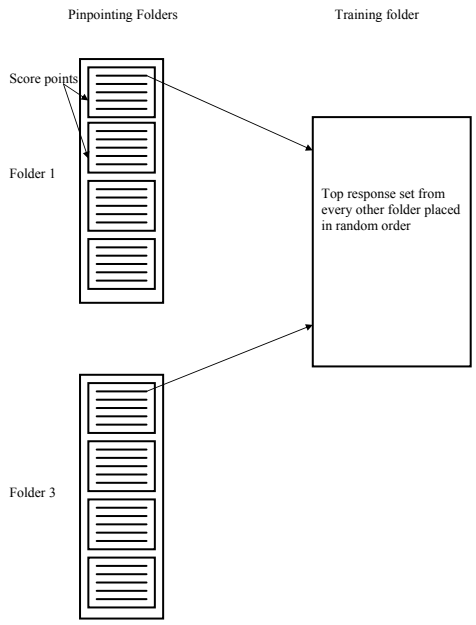

Figure 3. Hypothetical example of preparing a training folder.

- Distribute to and have panel members review and discuss the actual constructed-response items together with their scoring guides and the actual multiple-choice items together with the scoring key.
- Distribute and have the panel members discuss the general and operational performance standard definitions.
- Distribute copies of the training/calibration folder and point out that all items—extended constructed response, short answer, and multiple-choice—must be considered together as a whole.
- Have each panelist independently rank-order the student tests in the training folder based on quality and keeping in mind the general and operational definitions of the performance standards.
- Have the panelists list the students' identification numbers in the same rank order on a separate sheet and then compare their ranking with the actual rankings determined during the creation of the training folder to note the degree of agreement. Discuss discrepancies.
- Have panelists independently assign each rank-ordered test to one of the performance levels (e.g., proficient, basic, below basic).
- Record the performance levels and show on an overhead so that panelists can see the degree of agreement among them. Discuss the findings with the intent of obtaining consensus.

The intent of the last four activities is to familiarize the panel members with the process to be followed while working to calibrate their rankings and proficiency level assignments.

Matching Student Tests to the Performance Standards

The matching of the student tests to the performance standards involves two sequential steps in Version 1. The first is *range-finding* followed by *pinpointing*. The purpose of range finding is to locate the general areas on the total test score scale where the cut-points should be placed. The purpose of pinpointing is to clarify ambiguity in the general area where the cut-score should be set.

Range finding. Following training,

- Distribute the range-finding folder and range-finding rating form to the panelists. The rating form contains the identification numbers of the student tests in the range-finding folder listed in order from high to low quality, and a place for the panel members to enter their names on the form.

- Have the panelists independently enter on the range-finding rating form the ratings they awarded to the student tests included in the training/calibration folder and advise them that if they wish, they may change their ratings of these tests.
- Ask the panelists, to decide independently the performance levels of the remaining tests in the range-finding folder and record their ratings on the range-finding rating form.
- Record and share panelists' assignments of the tests to the performance levels and then discuss the round 1 assignments.
- Ask the panelists to revise independently any assignments they wish in light of the discussion. The round 2 assignments are used to identify the need for pinpointing folders.

Need for and selection of pinpointing folders. As pointed out above, pinpointing is used to clarify ambiguity in the assignment of the student tests in the range-finding folder. If more than two thirds of the panel members agree on the classification of the three student tests that belong to a specific pinpointing folder in range-finding, then the performance level to which these tests are assigned is determined and the corresponding pinpointing folder is not selected for pinpointing. However, if the panel members disagree on the classification of the three student response tests that belong to a specific pinpointing folder, then it is assumed that the panelists are not sure about which of two consecutive performance levels the tests should be assigned. That is, these tests are near the point in the total test score distribution where the cut-score between these two levels should be placed. Kingston et al. (2001) defined the cut-score as the point in the total test score distribution at which the probability of a panelist assigning a paper to one of the two adjacent proficiency levels is one-half (p=0.50). Consequently, the complete pinpointing folder about which there is ambiguity in classifying the three tests used in range finding is selected for pinpointing.

An illustration is provided in Table 2. As shown, more than two-thirds of the 16 panel members who were on the MCAS grade 8 mathematics panel assigned the three student tests from the highest folder, Folder 1, to the advanced performance level and the three student tests from Folder 3 to the proficient performance level during range finding. In contrast, fewer than two-thirds of the panelists assigned the three tests from Folder 2 to the advanced level (62.5%, 37.5%, and 37.5%) or to the proficient level (37.5%, 62.5%, and 62.5%). Although the assignment of the tests from Folders 1 and 3 was clear, there was ambiguity in the assignment of the tests from Folder 2. Seemingly there was close to a 50-50 chance of assigning some of these tests to the advanced performance level and some to the proficient performance level. Hence Folder 2 was selected for pinpointing to determine the cut-score between the advanced and proficient proficiency levels.

Pinpointing. The panelists are divided into subgroups for pinpointing, with the number of subgroups dependent on the number of cut-scores. For example, if there are three cut-scores, then the panelists are divided into three subgroups. The selected pinpointing folders are rotated among the subgroups so that all the panelists examine the folder(s) for each cut-score. The panelists are asked to decide independently if the folder(s) belong to the higher or lower

performance level to be divided by the cut-score. For example, do the tests in Folder 2 belong to the advanced or the proficient performance level? The panelists are given the opportunity to write down their comments in case they decide that the folder does not belong to either of the two suggested performance levels.

Table 3 shows the pinpointing results for Folder 2. The numbers in this case are based on 80 classifications (5 student tests at each of 4 score points and 16 panelists). As shown, there is still some ambiguity at the two lower score points in this folder. Given this, it appears that the cut-score is somewhere in the low 40s on the total score distribution.

Setting the Final Cut-Scores

The final values of the cut-scores on the total test score distribution are set using logistic regression. Logistic regression is used to analyze the relationship between a dichotomous variable such as the probability of being placed in one category (e.g., below basic) as opposed to being placed in another category (e.g., basic and above), and a continuous variable such as performance on a test. The logistic equation for establishing the cut-score between below basic and above basic for the BoW procedure is:

$$\ln \frac{p_{jk}}{1 - p_{jk}} = a + bTS_j ,$$

where ln is the natural logarithm, p_{jk} is the probability of a student with total test score TS_j being assigned to proficiency category k, the basic category or above, and a and b are respectively the slope and intercept of the logistic function.

After estimating a and b, the logistic function equation is solved for the cut-score at each decision point. At each cut-score point the probability of a student being assigned to one of two adjacent proficiency levels is set at 0.50. For example, the logistic function at the cut-score point that separates the below basic level from the basic and above levels, CS_{bB}, is given by:

$$\ln \frac{0.50}{1 - 0.50} = a + b\, CS_{bB} .$$

Table 2
Range-Finding Results for Folders 1, 2, and 3

Folder	Student Test	Panelist Classification: Advanced	Proficient	Needs Improvement
	1	16		
1	19	14	2	
	20	12	4	
	1	10	6	
2	19	6	10	
	20	6	10	
	1	4	12	
3	19		14	2
	20		13	3

Adapted from Kingston et al., 2001, p. 228.

Table 3
Pinpointing Results for Folder 2

Folder	Total Student Test Score	Panelist Classification: Advanced	Proficient	Needs Improvement
2	46	72	8	
	45	58	22	
	44	46	34	
	43	29	34	

Adapted from Kingston et al., 2001, p. 228.

Solving this equation for CS_{bB} yields:

$$CS_{bB} = \frac{-a}{b}.$$

A graphical representation of the logistic function is provided in Figure 4. As shown, the cut-score separating the below basic level from the basic proficiency level is 20.5 test score points, which should be rounded down to 20. Thus students who score below 20 are deemed to be at the below basic performance level.

Returning to the MCAS grade 8 mathematics illustration, the cut-score between the advanced and proficient levels equaled 42.7, which is between 38 and 44. Students with total test scores greater than or equal to 43 were classified as being at the advanced level.

As stated in Standard 4.19 of the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999), a measure of the variability among the panelists' judgments should be provided. In the BoW method, the logistic regression is separately conducted for each panel member at each cut-score. The standard deviation of the separate panelists' estimates is then divided by the square root of the number of panelists to yield the standard error of the mean estimate of the separate panelists' cut-score estimates at each cut-score. Using this method, the standard error due to panelists was approximately half a score point for the MCAS grade 8 mathematics test.

A second estimate of the standard error of the final cut-scores may also be estimated using the full sample of student tests across panelists. Based on a Taylor expansion of $f(a, b) = -\frac{a}{b}$, the variance of the sampling distribution of a cut-score is given approximately by:

$$\text{var}[\, f(a,b)\,] = \frac{1}{b^2}\,\hat{\sigma}_a^2 + \frac{a^2}{b^4}\,\hat{\sigma}_b^2 - \frac{2a}{b^3}\,\hat{\rho}\,\hat{\sigma}_a\,\hat{\sigma}_b ,\sigma$$

where $\hat{\sigma}_a$ and $\hat{\sigma}_b$ are the standard error of estimate of a and b respectively, and $\hat{\rho}$ is the correlation between a and b at each cut-score point on the total score distribution. The square root of the variance is the corresponding standard error of estimate score (Kingston et al., 2001). Using this method, the standard errors were smaller than the standard errors found using the first method and

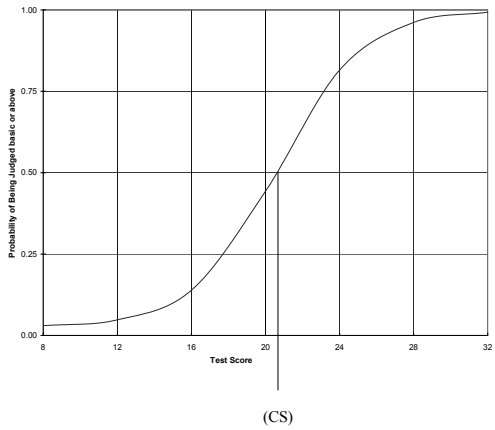

Figure 4. Logistic regression curve for below basic—basic cut-score.

varied between 0.13 and 0.19 score points for the MCAS grade 8 mathematics test.

Version 2

Version 1 of the BoW method is labor-intensive. Considerable time and cost are required to create and print the pinpointing folders. One day is required to complete the pinpointing step. In an attempt to eliminate the pinpointing folders and the pinpointing step, the range-finding results were analyzed using the logistic regression outlined above. The difference between the cut-scores using pinpointing and not using pinpointing was less than a third of a score point (0.1, 0.3, and 0.0 for the advanced/proficient, proficient/basic, and basic/below basic cut-points (Kingston et al., 2001; K.P. Sweeney, personal communication, November 15, 2004).

Consequently, the pinpointing folders and the pinpointing step were dropped. To compensate for the reduction in the coverage of the total test score distribution, the range-finding folder was modified to include two to three student tests at each score point between the chance score and the highest score. Two rounds of range-finding separated by group discussion of the first round assignments are completed. Logistic regression is then used as before to determine the final cut-scores.

Evaluation of the Body of Work Method

Both Version 1 and Version 2 are evaluated in terms of the six technical criteria and four practical criteria. Berk (1986) suggested that these criteria needed to be met to ensure that the cut-scores set were defensible. We mention above that the changes made to Version 1 were intended to make the BoW method more practical without sacrificing the accuracy of the cut-scores that are set. Examining the change between the evaluative ratings for Version 1 and Version 2, the present version, allows an assessment of the tenability of this intent. As with the previous papers, a three-point (3—fully met, 2—partially met, and 1—not met) is used.

Technical Adequacy

1. The method should yield appropriate classification information (Version 1 and 2 Ratings: 3)

General and operational descriptions of the performance levels are established before the creation of the folders in both Version 1 and Version 2. Procedures (discussion, calibration) are in place to help ensure that the cut-score panel members are thoroughly familiar with the performance levels and the differences between the levels. The final cut-scores produced by both versions of the BoW method permit meaningful and appropriate dichotomous classification decisions at each cut-score point. For example, in the 2003 MCAS, the students were appropriately classified at the advanced, proficient, needs improvement, or warning performance levels (J. Nellhaus, personal communication, November 4, 2004). Furthermore, the change to Version 1 does not detract from the appropriateness of the decisions made.

2. The method should be sensitive to examinee performance (Version 1 and 2 Ratings: 3)

The BoW is an examinee-based procedure by which the panel members responsible for setting cut-scores match the quality of student test responses to the quality called for in the statement of the performance standards. For Version 1 the panelists assign a representative sample of student tests in the training/calibration folder and the Version 1 range-finding folder to the appropriate performance levels. Where there is ambiguity in assigning these tests, the remaining student tests in the corresponding pinpointing folders in Version 1 are considered and assigned. For Version 2 the panelists assign the student tests in the training/calibration folder and the Version 2 range-finding folder to the appropriate performance levels.

3. The method should be sensitive to instruction or training (Version 1 and 2 Rating: 3)

The BoW method is instructionally sensitive to the degree that test performance depends on what was taught and learned. However, it appears that provision is not made for the BoW method to consider explicitly information about the opportunity to learn. However, it is reasonable to assume that appropriate instruction has been provided and that the items contained in the test are relevant to and representative of the expected student outcomes based on what is contained in the technical reports (Massachusetts Department of Education, 2002).

4. The method should be statistically sound (Version 1 and 2 Rating: 2)

The decision to be made at each cut-score point is a dichotomous decision; each student is to be assigned to one of two adjacent levels depending on his or her total test score. Thus the outcome variable, the decision made, is a dichotomous variable, and the underlying probability distribution is binomial in form. Consequently, the use of logistic regression in both versions of the BoW to predict the outcome variable from the total score is both appropriate and sound.

As indicated above, a measure of the variability among the panelists' judgments should be provided. Two standard errors are provided, but they are for two separate estimates of the final cut-score. The first is for the mean of the separate logistic regression estimates for each panel member. The other is for

the logistic estimates using all the student tests across all panel members. It is not clear which of the two should be used (Kingston et al., 2001). However, given the observation that they tend not to exceed one score point on the total test score distribution, the issue may be moot. Furthermore, the standard deviation of the panelists' ratings does not appear to be directly presented or discussed in BoW reports.

5. The method should identify the true standard (Version 1 and 2 Rating: 3)
As indicated in the justification of the ratings for the previous criterion, the standard error of estimate is reported for each estimated cut-score. For example, the standard errors for the three cut-scores for the MCAS grade 8 mathematics test were less than half a total score point for both estimation procedures. These low values suggest that the cut-scores are close to the true cut-scores.

6. The method should yield decision validity evidence. (Version 1 and 2 Rating: 1)
As with other cut-score methods, provision is not made in either version of the BoW method about the correctness of the assignment of students to the defined proficiency levels.

Practical Adequacy

7. The method should be easy to implement (Version 1 and 2 Ratings: 1 and 2)
As indicated in the reasons provided for modifying Version 1, Version 1 is labor-intensive and relatively costly. Major time commitments are required to create the pinpointing, range-finding, and training/calibration folders. Copies of these folders must then be made. An extra day is required to work with the pinpointing folders. Given that the cut-scores determined after pinpointing and using the range-finding folder were very close in value, the pinpointing folders were dropped and the range-finding formula was modified. However, the number of student tests in the range-finding folder in Version 2 still appears to be large. Further research is needed to determine whether a reduction in the number of tests, perhaps by first considering every other score point, leads to accurate estimation of the cut-scores. Additional score points could then be considered once the general area of each cut-score is identified.

8. The method should be easy to compute (Version 1 and 2 Rating: 3)
The logistic regression and the standard error of estimates are easy to compute given the availability of fast computers and appropriate statistical programs.

9. The method should be easy to interpret to laypeople (Version 1 and 2 Ratings: 2)
The steps involved in the process used by the panelists were clear and easily understood and followed by the panelists. The panelists were asked to rate the clarity of instructions, level of understanding, and confidence of their ratings. The findings revealed that the instructions were clear, the process was understandable, and the panel members were confident about their ratings. However, it is quite likely that most panel members, if asked, would not understand the statistical analyses used to determine the cuts-scores and their standard errors. Furthermore, no data other than the observation that the BoW has been used in several states over a number of years are provided to see if stakeholders in the educational system (state, province) who were not involved in the cut-score-setting can interpret the final results in terms of the performance

standards and what, if any, actions are taken in response to the results. The evaluation completed as part of the BoW method was rather brief. A more comprehensive and independently conducted evaluation would be much more sound and credible (Joint Committee on Standards for Educational Evaluation, 1994). In fairness, this same concern can be raised with the other methods for setting cut-scores and/or establishing performance standards.

10. The method should be credible to laypeople (Version 1 and 2 Rating: 2)
The BoW seems to be credible to laypeople because it involves the input of a representative sample of stakeholders including parents and members of the general public. However, as pointed out in the discussion of criterion 9, the issue of credibility has not been formally evaluated.

Discussion and Conclusion

The Body of Work appears to be a promising method for setting cut-scores for situations in which constructed-response and multiple-choice items are used. The judgments required in the BoW method are somewhat more familiar to the panelists because they are based on actual student work rather than the judges' concept of a hypothetical group of minimally competent examinees. This feature makes the BoW method very appealing.

Linn (1994) identified four major uses of performance standards: exhortation, exemplification, accountability, and certification. The BoW method as implemented in the MCAS can be used for the first three of these four uses. The performance standards exemplify what students are to learn and what learning is expected at each level of proficiency. This same information coupled with the proportion of students who are classified at each proficiency level can be used to motivate teachers and others who work with students and the students themselves to greater levels of accomplishment and higher levels of proficiency. At the same time, these two sources of information can be used as part of an accountability program in which school officials are held accountable for any discrepancies between desired and actual level of student performance in terms of clearly established performance standards.

Both Version 1 and Version 2 of the BoW method fared reasonably well against Berk's (1986) criteria for evaluating methods for setting cut-scores. The ratings for the technical adequacy for both versions were the same: four of the six ratings were 3, one was 2, and one was 1. The failure to report the standard deviations of the panelists' ratings is easily addressed. The failure to collect decision validity evidence is more problematic, and as suggested, this failure is an endemic problem in the standard-setting regardless of method. Three of the four practical ratings were the same. The ease of implementation of Version 2 was rated 2 whereas the ease of implement of Version 1 was rated 1. The change in procedures had the desired effect of reducing the amount of effort and the cost of the BoW. However, even Version 2 requires more effort and cost than might be needed. The remaining ratings for both versions were 3 for computational ease and 2 for both easy to interpret to laypeople and credible to laypeople. The lower ratings here are principally due to the failure to assess independently and adequately the degree to which the method is easy to interpret and credible to people other than those who were involved in establishing the performance standards and setting the corresponding cut-scores.

Again, this activity is not completed for the other procedures. Taken together, the 10 ratings indicate that the changes made to Version 1 to make it more practical did not adversely influence the accuracy of the cut-scores. Further, as with the other methods, the experience gained from using the BoW in the field will probably lead to further modifications in attempts to increase efficiency without sacrificing accuracy.

References

Abbott, M. (in press). Setting cut-scores for complex performance assessments: A critical examination of the analytic judgment method. *Alberta Journal of Educational Research*.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp.508-600). Washington DC: American Council on Education.

Berk, R.A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research, 56*, 137-172.

Joint Committee on Standards for Educational Evaluation. (1994). *The program evaluation standards: How to assess evaluations of educational programs*. Thousand Oaks, CA: Sage

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* (pp. 53-58). Mahwah, NJ: Erlbaum.

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* (pp. 219-247). Mahwah, NJ: Erlbaum.

Linn, R.L. (1994). *The likely impact of performance standards as a function of uses: From rhetoric to sanctions*. Paper presented at the National Center for Education Statistics and National Assessment Governing Board Joint Conference on Standard-Setting for Large-Scale Assessments, Washington, DC.

Massachusetts Department of Education. (2002). *2001 MCAS technical report*. Malden, MA: Author.

Ricker, K.L. (in press). Setting cut-scores: A critical review of the Angoff and modified Angoff methods. *Alberta Journal of Educational Research.*

Shepard, L.A. (1991). Interview on assessment issues with Lorrie Shepard. Michael W. Kirst. *Educational Researcher, 20*(2), 21-23, 27.

Wiggins, G.P. (1993). *Assessing student performance: Exploring the purpose and limits of testing*. San Francisco, CA: Jossey-Bass.