The Alberta Journal of Educational Research *Vol. 52, No. 1, Spring 2006, 25-35*

Marilyn L. Abbott
Alberta Education

Setting Cut-Scores for Complex Performance Assessments: A Critical Examination of the Analytic Judgment Method

The purpose of this article is to promote an increased awareness of the processes for setting cut-scores for complex performance assessments by (a) describing the Analytic Judgment Method (AJM) for setting cut-scores, and (b) critically evaluating the technical adequacy and practicability of the AJM by focusing on one investigation where the AJM was used by Plake and Hambleton (2001) for setting standards on the Pennsylvania Grade 8 Mathematics Achievement Test. Although Plake and Hambleton (1998, 2001) demonstrate that the AJM is an attractive iterative procedure that uses independent judgments of actual student work, more research is necessary to replicate the results and determine whether the AJM would produce high interrater reliability with more traditionally sized panels of 20 or more representatives.

L'objectif de cet article est de conscientiser quant aux processus employés pour déterminer les seuils de passage dans les évaluations de tâches complexes. D'une part, nous décrivons la méthode basée sur le choix raisonné (Analytic Judgment Method, AJM) pour établir les seuils de passage et, d'autre part, nous évaluons la suffisance technique et la possibilité de mise en oeuvre de la AJM en étudiant l'emploi qu'en ont fait Plake and Hambleton (2001) dans l'établissement de normes pour un test de rendement en mathématiques qu'on fait passer aux élèves en 8e année en Pennsylvanie. Bien que Plake and Hambleton (1998, 2000, 2001) ait démontré que la AJM constitue une procédure itérative prometteuse pour évaluer de façon indépendante les travaux des élèves, il faudrait entreprendre davantage de recherche pour répéter les résultats et déterminer si la AJM entraînerait un indice plus élevé de cohérence inter-juges chez un groupe d'experts de taille plus traditionnelle où il y irait au moins 20 membres représentatifs.

In an educational context, complex performance assessments involve making judgments about the students' knowledge, skills, and abilities based on behavioral observations and/or inspections of their work (Gitomer, 1993). The current interest in performance assessment is directly related to two notions that are deeply rooted in the educational reform movement. First, many proponents of educational reform believe that the practice of establishing more rigorous academic standards, which results in higher cut-scores or passing marks on mandated assessments, will promote more effective educational practices (Lockwood, 1998). A second commonly held notion of educational reform is that all ailments in student assessment will be cured by replacing multiple-choice tests with performance assessments consisting of constructed-response items or other types of so-called authentic tasks, which purport to assess higher-order thinking skills better than multiple-choice items (Madaus, 1993). Not surprisingly, these two notions have increased the popularity of

Marilyn L. Abbott is a psychometrician. Her research interests include test development, differential item functioning, and fairness in assessment.

complex performance assessments, thereby creating a need for validated methods for setting multiple cut-scores on relevant and representative performance assessments.

The process employed to establish standards and set cut-score(s) on a test that is relevant to and representative of the standards is referred to as the standard-setting procedure or method. Standard-setting procedures are used to derive levels of performance on educational assessments by which decisions or classifications of persons and corresponding inferences will be made (Cizek, 1993). For example, the level of student performance demonstrated on an assessment could be classified as basic, proficient, or advanced in relation to the content specified in the curriculum (i.e., what students are expected to know and be able to do). Once the performance standards are established, cut-score setting activities are then used to determine the cut-scores or points on the score scale that separate one performance level or standard from another.

Although numerous methods have been proposed for setting cut-scores on multiple-choice assessments, the problem of setting multiple cut-scores on complex performance assessments composed of both constructed-response and multiple-choice items is not well addressed in the literature. Two methods have been proposed: the Body of Work (see Radwan & Rogers, in press, this issue) and the Analytic Judgment Method (AJM). The main objective of this article is to promote an increased awareness of the processes for setting cut-scores on complex performance assessments by (a) describing the AJM for setting cut-scores given that the performance levels have already been established, and (b) critically evaluating the technical adequacy and practicability of the AJM by focusing on one investigation where the AJM was used by Plake and Hambleton (2001) for setting cut-scores on the Pennsylvania Grade 8 Mathematics Achievement Test.

Several variants of the AJM have been investigated by Plake and Hambleton (1998, 2000, 2001). In this article I focus on only one variant (referred to below as the AJM1) because most achievement tests tend to have larger numbers of multiple-choice items than constructed-response items, and this variant is appropriate for use with exams that have large numbers of multiple-choice questions whereas the other AJM variants are not. The AJM1 employs a variant of the yes/no Angoff (Angoff, 1971; Ricker, in press, this issue) for setting cut-scores with multiple-choice items, and a variant of the Body of Work (BoW) method (Kingston, Kahl, Sweeney, & Bay, 2001) for the constructed-response items.

Setting Cut-Scores with Constructed-Response Items

Plake and Hambleton (2001) divide the methods for setting cut-scores for assessments composed of constructed-response items into two main approaches: analytic component and holistic full-test. In the analytic approach, panelists examine the test questions or sections one at a time in order to identify the cut-scores for each question or section. The cut-scores are set either by identifying the expected scores for minimally competent candidates (which is a test-centered approach) or by selecting student responses that represent the work of minimally competent candidates (which is an examinee-centered approach). In contrast, the holistic approach sets cut-scores in the total score

distribution, where the total score is calculated from the constructed-response scores either as a simple or weighted sum. In these terms, the AJM1 may be defined as an examinee-centered, analytic component approach for setting cut-scores with constructed-response items.

The AJM1 has the panelists rate student work that has been selected to represent the full-score continuum on the assessment. Initially panel members classify anonymous student papers into one of 12 performance categories defined to capture levels of performance as expressed by the standards (see Figure 1). Each of the items is considered independently by the panelists. The panelists' ratings of the student papers for the first constructed-response item are discussed by panel members and then reconsidered before they move on to the student papers for the next item. To prevent an order effect, the student papers are not presented in the same order across the items. Once the panelists have rated, discussed, and rerated all the items, the relationship between the examinees' assessment scores and the panelists' classifications is used to calculate the cut-scores for the associated performance standards. These procedures are explained in more detail below.

In comparison with test-centered methods such as the Angoff procedure and its modifications for setting cut-scores, the AJM1 is much more reasonable in that it requires the panel members (who are usually teachers) to evaluate student work in relation to specifically defined performance levels. Clearly this is a realistic task that teachers in particular are accustomed to performing, whereas speculating about item difficulty is not. Therefore, the AJM1 procedure for establishing cut-scores with constructed-response items appears to be more congruent with teacher behavior than the test-centered methods.

Setting Cut-Scores With Multiple-Choice Items

Many of the desirable features attributed to the Angoff procedure are reflected in the AJM1 procedures for setting cut-scores with multiple-choice items. The AJM1 version of the yes/no Angoff procedure has the panelists predict whether a typical student in each of the performance categories would be able to answer each multiple-choice item correctly (Plake & Hambleton, 2001). The number of *yes* items for each hypothetical examinee determines the panelists' estimated test score for each category. However, only the means of the panelists' estimated test score boundary categories for the novice-high and apprentice-low categories, the apprentice-high and proficient-low categories, and the proficient-high and advanced-low categories are used as the final multiple-choice cut-scores. Consequently, only the scores from the student papers that the panel members classify in these three borderline categories are

Novice	Apprentice	Proficient	Advanced
• minimal understanding of rudimentary basic concepts and skills	• partial understanding of basic concepts and skills	• general understanding of basic concepts and skills	• broad and in-depth understanding of complex concepts and skills
1 2 3	4 5 6	7 8 9	10 11 12
low medium high	low medium high	low medium high	low medium high

Figure 1. AJM1 multipoint classification scale used by the panelists to rate examinee performance on constructed-response and multiple-choice items.

used to calculate the cut-scores. Thus the main differences between the most common Angoff procedure and the AJM1 yes/no procedure are reflected in the number of cut-scores produced for each item and the type of candidate the panelists are asked to envisage: The two-choice Angoff has the panelists determine whether a *minimally competent* candidate would answer the item correctly (producing one cut-score), whereas the AJM1 has the panelists predict whether typical examinees in each of the various performance categories would answer each item correctly, and only the boundary scores are used in the final calculations to produce three cut-scores.

Setting Cut-Scores for Complex Performance Assessments Using the AJM1

Plake and Hambleton's (2000, 2001) investigation of the AJM1 focused on the utility of two procedures: one for rating multiple-choice items and the other for rating constructed-response items. In addition, the study examined whether the panelists were satisfied with the procedures and cut-scores produced by the method.

AJM1 Procedures

Plake and Hambleton (2000, 2001) used the AJM1 to set multiple cut-scores on the Pennsylvania Grade 8 Mathematics Achievement Test. Although this assessment comprised 105 multiple-choice questions (75 common and 30 matrix-sampled) and four constructed-response items (two common and two matrix-sampled), only the common multiple-choice items and the four constructed-response questions were examined in their study. In conjunction with a variant of the two-choice Angoff for the dichotomously scored multiple-choice items (described above), Plake and Hambleton used an examinee-centered, question-by-question approach to rate the four polytomously scored constructed-response items.

Fourteen panelists (11 mathematics teachers and 3 school administrators) were (a) divided into four small groups consisting of three or four panelists, (b) provided with a review of the performance standards and the assessment's scoring rubrics, (c) trained to use the AJM1, and (d) participated in an AJM1 practice session. After the training and practice session, each group was given a sample of 50 student papers selected to represent the full-score continuum on the assessment. Neither the students' scores nor their identities were revealed to the panel members. Working independently, the panelists in each small group first predicted whether a typical student in each of the performance categories would be able to answer each multiple-choice item correctly. They then classified each constructed-response into one of 12 performance categories using the classification scale shown in Figure 1. Once they had completed their initial classifications, the panelists in each group compared their decisions. Discrepant classifications were discussed, and the panelists were asked to make any final independent adjustments to their ratings.

For the constructed-response items, the relationships between actual examinee scores and the panelists' boundary category ratings were used to calculate the cut-scores that separated one performance standard from another. The final cut-scores for the performance standards were calculated by taking the mean of the actual constructed-response item scores received by the student papers in the respective boundary categories (i.e., novice-high and apprentice-

low, apprentice-high and proficient-low, and proficient-high and advanced-low), and adding the mean boundary category cut-score produced by the yes/no Angoff variant for the multiple-choice items. On completion of the cut-score setting activities, the panelists were asked to fill out an evaluation form designed to explore how satisfied they were with the method and the standards produced using the AJM1.

AJM1 Results

The four groups' boundary cut-scores for each of the constructed-response questions are presented in Table 1. The results for the two common questions show that the cut-scores across the four groups increased monotonically across the three levels. For the first common question (scored on a scale of 1-4), the weighted mean cut-scores for the boundary categories of apprentice, proficient, and advanced were 1.28, 2.44, and 3.54 respectively. Weighted means were calculated because two of the groups had three panelists and two of the groups had four. For common question 2, the weighted mean cut-scores were 1.54, 2.52, and 3.56 respectively. Although the cut-scores produced by each small group for the two matrix sampled questions were not comparable because each group rated different questions, a similar trend was observed across the boundary categories for the unique questions; that is, the cut-scores generally increased across the categories (see Table 1). However, on Common Item 1 (Advanced cut-score), the difference between Group A's cut-score (3.13) and Group C's cut-score (3.91) is 0.78, approximately 20% on a 4-point scale. Similarly, the Group C (3.95) to Group A (3.19) difference on Common Item 2 is 0.76. These differences are substantial enough to yield markedly different cut-scores.

The results for the 75 multiple-choice questions are presented in Table 2. The cut-scores based on the panelists' performance estimates for the boundary categories were 24.21, 46.11, and 65.57. After averaging the cut-scores for the boundary categories on the unique questions and adding them to the boundary category means for the common questions, the final cut-score for each pair of adjacent categories was calculated by adding the boundary cut-scores for the four constructed-response questions to the boundary cut-scores for the multiple-choice questions. This resulted in the following performance standards for the apprentice, proficient, and advanced categories: 30.03, 54.91, and 78.03 respectively. However, because the group differences on the multiple-choice item cut-scores are between 5% and 10%, these differences are enough to affect the overall cut-scores.

The results of the evaluation form, reported on a scale of 1 to 4 (1=not successful to 4=very successful), revealed that the panelists were generally satisfied with the procedures of AJM1 and confident in the cut-scores. The panelists' mean ratings for overall satisfaction and confidence in the cut-scores produced using the AJM1 were 3.2 and 3.4 respectively.

Ten Criteria for Evaluating Cut-score Setting Methods

Berk's (1986) 10 criteria for evaluating standard-setting methods are defined in terms of the technical adequacy and practicability of the method. A method that is technically adequate yields appropriate classification information, is sensitive to examinee performance, is sensitive to instruction or training, is

Table 1
Boundary Cut-scores from Groups Based on Constructed-Response Questions—AJM1

Common Question 1	*Boundary Categories*		
Group	*Apprentice*	*Proficient*	*Advanced*
A	1.27 (22[a])	2.27 (37)	3.13 (16)
B	1.28 (46)	2.53 (32)	3.69 (18)
C	1.27 (55)	2.68 (28)	3.91 (24)
D	1.30 (10)	2.30 (20)	3.29 (21)
Mean	1.28 (133)	2.45 (117)	3.51 (79)
SD	0.01	0.20	0.36
Weighted mean[b]	1.28	2.44	3.54
Common Question 2	*Boundary Categories*		
Group	*Apprentice*	*Proficient*	*Advanced*
A	1.27 (22)	2.27 (37)	3.19 (16)
B	1.63 (19)	2.33 (66)	3.41 (34)
C	1.67 (34)	2.73 (55)	3.95 (38)
D	1.55 (38)	3.00 (21)	3.40 (25)
Mean	1.53 (113)	2.58 (179)	3.49 (113)
SD	0.18	0.35	0.32
Weighted mean	1.54	2.52	3.56
Unique Question 1	*Boundary Categories*		
Group	*Apprentice*	*Proficient*	*Advanced*
A	1.00 (21)	1.00 (3)	2.00 (49)
B	1.00 (13)	1.09 (85)	1.87 (48)
C	1.00 (13)	1.43 (35)	2.00 (39)
D	1.00 (9)	1.35 (17)	1.87 (24)
Unique Question 2	*Boundary Categories*		
Group	*Apprentice*	*Proficient*	*Advanced*
A	2.47 (32)	3.11 (19)	3.64 (22)
B	1.46 (13)	2.76 (74)	3.86 (35)
C	1.20 (10)	3.19 (86)	3.75 (20)
D	2.00 (3)	1.72 (21)	3.37 (19)

Note. [a]Number of papers in brackets (adapted from Plake & Hambleton, 2001, pp. 294-295). [b]Weighted means were calculated because two of the groups had three panelists and two of the groups had four panelists.

statistically sound, identifies the true standard (i.e., takes measurement errors into account), and yields decision validity evidence (i.e., provides estimates of decision consistency). A method is practical if it is credible and easy to implement, compute, and interpret to laypeople. Based on a comprehensive analysis of the information provided in Plake and Hambleton's (2000, 2001) study of the AJM1, I employed a three-point Likert-type scale (1=not met, 2=partially met,

Table 2
Boundary Cutpoints From Groups Based on Multiple-Choice Component—AJM1

Multiple-Choice Group	Boundary Categories Apprentice	Proficient	Advanced
A	19.50	46.00	67.00
B	28.13	47.25	63.63
C	26.63	48.50	66.50
D	20.50	41.50	65.50
Mean	24.21	46.11	65.57
SD	4.68	3.09	1.66

(Plake & Hambleton, 2001, p. 296).

3=fully met) to indicate the degree to which each criterion was satisfied by the method.

Evaluating the AJM1 Using Berk's (1986) Ten Criteria

Technical adequacy

1. The method should yield appropriate classification information (Rating: 3). The AJM1 produces three fairly unambiguous cut-scores, which can be used to make appropriate inferences from the standards. For example, the cut-scores can be readily used to determine whether students have met the performance criteria outlined in the standards or to identify students who need remediation.

2. The method should be sensitive to examinee performance (Rating: 2). The AJM1 method for setting cut-scores on the constructed-response items is sensitive to examinee performance because these cut-scores reflect actual student performance. However, this criterion is not met by the AJM1's approach to setting the multiple-choice cut-scores because they are produced using a variant of the yes/no Angoff procedure, which reflects panelists' expectations of student performance.

3. The method should be sensitive to instruction or training (Rating: 3). The AJM1 procedures are sensitive to the instruction or training that the examinees receive because the panelists were mainly middle-school math teachers who had experience teaching grade 8 mathematics. Thus the teachers' knowledge of the examinee population would have allowed them to take into account whether the examinees would have had the opportunity to learn the assessment material.

4. The method should be statistically sound (Rating: 3). The AJM1 is statistically sound in that the statistics (i.e., the means) used to summarize the judgments and describe the students' test performance were appropriate, easy to calculate, understandable, and interpreted correctly.

5. The method should identify the true standard (Rating: 1). The AJM1 does not identify the true standard because measurement error was not taken into account in the calculation of the cut-scores.

6. The method should yield decision validity evidence (Rating: 1). The AJM1 did not yield decision validity evidence because estimates of the probabilities of correct and incorrect classification decisions were not provided.

Practicability.

7. The method should be easy to implement (Rating: 2). Although the AJM1's rating procedures are systematic and easily understood, the physical demands of selecting, photocopying, and managing the large number of legible student papers required by the AJM1 are not easy to satisfy.

8. The method should be easy to compute (Rating: 3). The AJM1 cut-scores are easily computed using a hand-held calculator.

9. The method should be easy to interpret to laypeople (Rating: 3). The AJM1 procedures are clear and conceptually understandable. Therefore, the stakeholders should not have difficulty understanding the methods used in determining the cut-scores.

10. The method should be credible to laypeople (Rating: 2). The explanations of the AJM1 procedures are clear, conceptually simple, and comprehensible. In this respect the AJM1 should be credible to laypeople. However, because the panelists selected were not representative of all the stakeholders, the cut-scores may tend to possess less credibility than if the panelists had been selected as a representative sample of all the stakeholders.

Overall, then, the AJM1 fully satisfies three of Berk's (1986) six criteria for technical adequacy and two of the four criteria for practicability.

Strengths and Weaknesses of the AJM1

The AJM1 has several strengths. First, as discussed above, the AJM1 procedures are conceptually clear and easy to explain. Second, the constructed-response rating methods match the assessment method, as examinee-centered approaches (which evaluate student work) tend to function well with constructed-response items (Plake, 1998). Third, when compared with other examinee-centered methods such as the BoW, the AJM1 has the potential to reduce preparation costs and the time between scoring and cut-score setting because it does not require the selection of illustrative papers to serve as benchmarks for the performance levels in advance of the cut-score setting activities (Plake & Hambleton, 2001). However, it does require the selection of student papers that represent the full score continuum on the assessment.

In general, the major strength of the AJM1 approach to setting cut-scores with constructed-response items is that it is one of only a few examinee-centered approaches to setting standards on complex performance assessments that have the panel members complete a realistic task (i.e., panelists review students' actual constructed-responses and make judgments about their performance levels). Such procedures can help the panelists avoid setting cut-scores that are unrealistically high or low.

However, although the AJM1 has the panelists complete a realistic rating task when examining the constructed-response items, it requires the panelists to complete an unrealistic, difficult task when rating the multiple-choice items (i.e., the panelists predict whether typical examinees in each of the performance categories would answer each item correctly). The main weakness of the AJM1, therefore, concerns the use of the yes/no Angoff procedure for rating the

multiple-choice items. As Berk (1986) suggests, the "yes-no format limits item probabilities to 0% and 100%" (p. 148). Because an individual's performance on most multiple-choice items usually relies on partial knowledge relating to multiple components of the question, a continuum of probabilities is more appropriate for most test items (Berk). In addition, although the panelists reported that they were generally satisfied with the AJM1 procedures and cut-scores, research comparing the yes/no method with the Angoff estimated percentages method indicates that panelists' responses are less positive for the yes/no method than the estimated percentages method (Loomis, Bay, Yang, & Hanick, 1999; Loomis, Hanick, Bay, & Crouse, 2000a, 2000b). Furthermore, for the lowest and highest performance categories, Reckase (1998) and Reckase and Bay (1999) found that the yes/no method produced lower and higher cut-scores than the panelists intended. Therefore, in addition to the problem of speculating about the performance of typical candidates, the limited item probabilities (0% and 100%) and unintended results detract from the procedural validity of the AJM1.

Another weakness of the AJM1 is related to the fact that Plake and Hambleton (2001) did not have the panelists rate all the mathematics achievement test items. Therefore, the AJM1's cut-scores could not be compared with the actual Pennsylvania State cut-scores set for this test. This is unfortunate because such comparisons would have had the potential to support the reasonableness of the cut-scores produced by the AJM1. As Kane (2001) suggests, if two methods produce similar results, "we have more confidence in the resulting cut-scores than we would have if either method were used alone" (p. 75).

A further disadvantage of the AJM1 involves the logistic requirements for implementing the student work classification process. As mentioned in the above section, satisfying the physical demands of selecting, photocopying, and managing the large number of legible student papers required by the AJM1 can be taxing (Plake & Hambleton, 2001). These logistic challenges also tend to cause the administrative costs of such examinee-centered approaches to be higher than those of the test-centered methods.

An additional weakness of the AJM1 study is that Plake and Hambleton (2001) specified neither how the panelists were selected nor whether they reflected a balance of geographic distribution, ethnicity, and knowledge of the grade 8 mathematics curriculum. To promote credible standards panelists must be both broadly representative of the relevant stakeholders and qualified to make judgments about what students should be able to do at each performance level (Kane, 2001). Furthermore, although Plake and Hambleton (2001) report "there was a high level of agreement in the ratings, both within group and across group" (p. 297), the differences between the small-group cut-scores revealed that they were not completely replicable across the four subgroups of panelists for either the constructed-response or the multiple-response items. In addition, Plake and Hambleton do not report estimates of intrarater consistency, interrater reliabilities, standard errors of measurement, or indices of dependability. These estimates could provide evidence that (a) the panelists used the achievement levels consistently, (b) the cut-scores were reproducible and dependable, and (c) the procedures were credible enough to support reasonable interpretations about the meaning of the achievement levels.

One final concern worth mentioning is that cut-score-setting methods such as the AJM1 that use group discussion and normative information change the focus of the procedure from expected examinee behavior to acceptable examinee behavior (Cizek, 2001). Thus in Cizek's terms, the AJM1 practices militate against the motives for setting standards and cut-scores by regressing what might result from the AJM1 procedures toward what is. This suggests that rather than motivating students and teachers to higher levels of performance and clarifying improved achievement expectations, the AJM1 procedures effectively maintain the status quo.

Conclusion

The primary goal of this article is to promote increased awareness of the processes for setting cut-scores on complex performance assessments by describing one variant of the AJM and critically evaluating the technical adequacy and practicability of the method using Berk's (1986) 10 criteria for evaluating standard-setting methods. It was assumed that a detailed discussion of the issues surrounding the topic of cut-score-setting on complex performance assessments would extend current practitioners' knowledge of cut-score-setting methods and inform existing cut-score-setting practices.

Although the AJM1 has several limitations, at this point there is still no one best method for setting cut-scores on complex performance assessments that comprise both selected and constructed-response items. If achievement tests continue to make greater use of constructed-response items, examinee-centered methods such as the AJM1 may prove more appropriate methods for setting cut-scores on these types of questions. Although the AJM1 is an attractive iterative procedure that uses independent item judgments in addition to actual student work, more research is necessary to replicate the results and to determine whether the AJM1 would produce high interrater reliability with more traditionally sized panels of 20 or more representative members.

Acknowledgment

I thank Todd Rogers and the three anonymous reviewers for their helpful comments on this article.

References

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

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

Cizek, G. (1993). Reconsidering standards and criteria. *Journal of Educational Measurement, 30*, 93-106.

Cizek, G. (2001). Conjectures on the rise and call of standard-setting: An introduction to context and practice. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 3-17). Mahwah, NJ: Erlbaum.

Gitomer, D. (1993). Performance assessment and educational measurement. In R. Bennett & W. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response* (pp. 241-263). Hillsdale, NJ: Erlbaum.

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

Kingston, N., Kahl, S., Sweeney, K., & Bay, L. (2001). Setting performance standards using the body of work method. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 219-248). Mahwah, NJ: Erlbaum.

Lockwood, A. (1998). *Standards: From policy to practice*. Thousand Oaks, CA: Sage.

Loomis, S., Bay, L., Yang, W., & Hanick, P. (1999, April). *Field trials to determine which rating method to use in the 1998 NAEP achievement levels-setting process for civics and writing*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal.

Loomis, S., Hanick, P., Bay, L., & Crouse, J. (2000a). *Setting achievement levels on the 1998 National Assessment of Educational Progress in civics interim report: Field trials*. Iowa, City, IA: ACT.

Loomis, S., Hanick, P., Bay, L., & Crouse, J. (2000b). *Setting achievement levels on the 1998 National Assessment of Educational Progress in civics interim report: Pilot study*. Iowa, City, IA: ACT.

Madaus, G. (1993). A national testing system: Manna from above? An historical/technological perspective. *Educational Assessment, 1*, 9-26.

Plake, B. (1998). Setting performance standards for professional licensure and certification. *Applied Measurement in Education, 11*, 65-80.

Plake, B., & Hambleton, R. (1998, April). *A standard-setting method designed for complex performance assessments with multiple performance categories: Categorical assignments of student work*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA. (ERIC Document Reproduction Service No. ED 422 371)

Plake, B., & Hambleton, R. (2000). A standard-setting method designed for complex performance assessments: Categorical assignments of student work. *Educational Assessment, 6*, 197-215.

Plake, B., & Hambleton, R. (2001). The analytic judgment method for setting standards on complex performance assessments. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 283-312). Mahwah, NJ: Erlbaum.

Radwan, N., & Rogers, T. (in press). A critical analysis of the body of work method for setting performance standards. *Alberta Journal of Educational Research*.

Reckase, M. (1998). *Setting standards to be consistent with an IRT item calibration*. Iowa City, IA: ACT.

Reckase, M., & Bay, L. (1999, April). *Comparing two methods for collecting test-based judgments*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal.

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