

Running head: THE ANALYTIC JUDGMENT METHOD

Standard Setting For Complex Performance Assessments: A Critical Examination of the  
Analytic Judgment Method

Marilyn Abbott

University of Alberta

Paper presented at the Annual Congress of the Canadian Society for the Study of Education,  
Halifax, Nova Scotia, May 2003.

### Abstract

The purpose of this paper is to (a) describe the analytic judgment method (AJM) for setting cutscores, and (b) critically examine one investigation where the AJM was used for setting standards on the Pennsylvania Grade 8 Mathematics Achievement Test. Although Plake and Hambleton's (1998, 2000, 2001) AJM studies demonstrate that the AJM is an attractive iterative procedure which utilizes independent item or component judgments of actual student work, more research is necessary to replicate the results and determine whether the AJM would produce high inter-rater reliability with more traditionally sized panels of 20 or more judges.

## Standard Setting For Complex Performance Assessments: A Critical Examination of the Analytic Judgment Method

In an educational context, complex performance assessments involve making judgments about the students' knowledge, skills, and abilities based on behavioural 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 setting more rigorous academic standards, which results in higher cutscores or passing marks on mandated assessments, will promote more effective educational practices (Lockwood, 1998). The process employed to establish standards and set cutscore(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, cutscore setting activities are then used to determine the cutscores or points on the score scale that separate one performance level or standard from another. 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 are purported to assess higher order thinking skills better than multiple-choice items (Madaus, 1993). Not surprisingly, these two notions have increased the popularity of complex performance assessments, thereby creating a

need for validated methods for setting multiple cutscores on relevant and representative performance assessments.

Although numerous methods have been proposed for setting cutscores on multiple-choice assessments, the problem of setting multiple cutscores on complex performance assessments composed of both constructed-response and multiple-choice items is not well-addressed in the literature. In an attempt to fill this gap, the main objectives of this paper are to (a) describe the analytic judgment method (AJM) for setting cutscores given the performance levels have been established, and (b) critically examine one investigation where the AJM was used for setting standards on the Pennsylvania Grade 8 Mathematics Achievement Test.

*Influential Procedures on the Development of the Analytic Judgment Method*

The AJM grew out of the need for a cutscore setting method that would appropriately separate examinees into ordered performance categories on either constructed-response items exclusively or in some combination with multiple-choice items (Plake & Hambleton, 2000). The AJM is a general procedure that has judges rate actual student work that has been selected to represent the full score continuum on the assessment. Initially, judges classify anonymous student papers into one of several performance categories defined to capture levels of performance as expressed by the standards. Each of the test components (usually sections or items) is considered independently by the judges. The component parts are naturally occurring sections of the assessment or some other division of the test, such as individual items, sections with common content, or sections with common formats. The judges' ratings for the first component are discussed by panels of judges and then reconsidered before moving on to the student papers for the next component. To prevent an order effect, student papers are not presented in the same order across the components. Once the judges have rated, discussed, and

re-rated all of the assessment components, the relationship between the examinees' assessment scores and the judges' classifications is used to calculate the cutscores for the associated performance standards.

Several variants of the AJM have been investigated by Plake and Hambleton (1998, 2000, 2001). This paper will only focus on the first variant (AJM1) as most achievement tests tend to have larger numbers of multiple-choice items than constructed response items, and the AJM1 is appropriate for use with exams that have large numbers of multiple-choice questions. The AJM1 employs a variant of the yes/no Angoff (Nassif, 1978) for setting cutscores 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 Cutscores with Multiple-Choice Items*

In a review of the early approaches to standard setting, Berk (1986) identified 23 different procedures that were predominantly designed for setting standards or cutscores (given that the standards have been set) on multiple-choice exams. Using Jaeger's (1989) classification schema, these approaches may be categorized as test-centered, examinee-centered, or a combination of these two approaches. Test-centered procedures do not use any empirical evidence in determining the cutscore(s). Teams of expert judges, usually teachers, examine test items or tasks and decide on the level(s) of competence required to meet the performance standard(s). In contrast, the judges in examinee-centered methods actually categorize examinee performance directly into performance levels using definitions of adequate performance for each level and information about the level of achievement of each examinee. While the number of methods for establishing standards and setting cutscores continues to grow, the most frequently

used standard setting procedure for high stakes tests has been the test-centered method known as the Angoff procedure or some variant thereof (Kane, 1995).

The original Angoff method requires the judges to independently evaluate each of the test items in order to assign a number between 0 and 1 that corresponds to the probability that a group of minimally knowledgeable persons would be able to answer the item correctly. Each of the judge's estimated probabilities is then summed to calculate the minimally acceptable standard for that judge. The minimum passing score, known as the cutscore, is the mean of all of the judges' item probabilities.

Despite the fact that the Angoff method is easy to “implement, understand and compute” (Berk, 1986, p. 148), the procedure has been highly criticized. One frequently noted shortcoming of the method is that even if the cutscores appear reasonable, the cutscores set using this procedure might actually be impossible to achieve because they are often set too high. This is why the Angoff method is frequently supplemented with normative and / or empirical data. When the judges have the opportunity to examine such data, the method is generally perceived to be considerably more objective. The Angoff procedure has also been criticized because the validity of the cutscores rest on the ability of the judges to conceptualize and make accurate performance estimates for minimally competent candidates (Shepard, 1995). Although several measurement professionals (e.g., Shepard, Glaser, Linn & Bohrnstedt, 1993) have criticized the Angoff procedure due to its *nearly impossible cognitive task* (p. xxiv) of estimating the proportion of a group of minimally competent people that would answer each item correctly, others (e.g., Cizek, 1993; Kane, 1995; Mehrens, 1995), defend the Angoff method due to its ease of use, technical soundness, and the widespread satisfaction with the numerous cutscores that have been set using this method.

Many of the desirable features attributed to the Angoff procedure are reflected in several newer standard setting methods. Methods that are closely related to the original Angoff are referred to as variants, derivatives, and modified or extended Angoff procedures (see Ricker, 2003). The Angoff derivative that has influenced the development of the AJM1 is known as the two-choice or yes/no Angoff (Nassif, 1978). In this version, the judges are asked whether a minimally competent person should be able to answer each item correctly. After adjusting for measurement error, the mean number of the *yes* items is used to determine the cutscore.

In contrast, the AJM1 version of the yes/no Angoff has the panelists predict whether a typical student within each of ten performance categories (i.e., novice-medium, novice-high, apprentice-low, apprentice-medium, apprentice-high, proficient-low, proficient-medium, proficient-high, advanced-low, advanced-medium) would be able to answer each multiple-choice item correctly (Plake & Hambleton, 2001). The number of *yes* items for each hypothetical examinee determines the judge's estimated test score for each category. The hypothetical test scores are averaged across judges and used to determine the cutscores for the boundary categories (i.e., novice-high and apprentice-low, apprentice-high and proficient-low, and proficient-high and advanced-low), which are then averaged and used as the final multiple-choice cutscores. Consequently, only the scores from the student papers that the judges classify in the borderline categories are used in the calculation of the cutscores. Thus the main differences between the two-choice Angoff and the AJM1 yes/no version are reflected in the number of cutscores produced for each item and the type of candidate the judges are asked to envision: The two-choice Angoff has the judges determine whether a minimally competent candidate would answer the item correctly (producing one cutscore), whereas the AJM1 has the

judges predict whether typical examinees in each of the different performance categories would answer each item correctly (producing three cutscores – one for each boundary category).

### *Setting Cutscores with Constructed-Response Items*

Plake and Hambleton (2001) divide the methods for setting performance standards on assessments composed of constructed-response items into two main approaches: (1) analytic component approaches, and (2) holistic full test approaches. In the analytic approaches, panelists examine the test questions or sections one at a time in order to identify minimum passing levels for each question or section. The minimum passing levels are set by either 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, full-test approaches are based on score profiles or item response theory where the score distributions or ability estimates calculated from the constructed-response scores are collectively used to set the performance standards. The AJM1 is an examinee-centered, analytic component approach for setting cutscores for the constructed-response items within the test or assessment instrument.

The escalating discontent with test-centered standard setting methods, which require panelists to estimate the difficulty of items for minimally competent examinees has fostered the development of examinee-centered standard setting methods such as the body of work (BoW) method, which are based on the examination of empirical data (i.e., student responses). The BoW method asks judges to

examine complete response sets and match each student response set to performance level categories based on previously agreed on descriptions of what students at the different levels should know and be able to do. (Kingston et al., 2001, p. 221)

The main similarity between the BoW method and the AJM is that both methods have the judges match student responses to performance categories. However, the main difference between these two methods is that while the BoW method is a holistic method, the AJM is an analytic method: In the BoW method, the judges are asked to examine and classify student responses on the whole test, whereas the AJM asks the judges to categorize each item or section of the test.

In comparison with test-centered methods such as the Angoff for setting cutscores, the BoW and the AJM approaches are much more reasonable in that they require the judges (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 BoW and AJM procedures appear to be more congruent with teacher behaviour than the test-centered methods.

#### *Setting Standards on Complex Performance Assessments Using the AJM*

From 1995 to 1999, Hambleton, Jaeger, Mills, and Plake worked on a research project funded by the National Science Foundation that focused on the development of new procedures for setting standards on complex performance assessments with multiple performance levels. One of the methods that grew out of their research was the AJM. Plake and Hambleton (1998, 2000, 2001) developed and investigated three variations of the AJM. The present paper deals with the first of these variations.

Table 1 provides an overview of the research design used to investigate the AJM1. The AJM1 focused on the utility of two procedures: one for rating multiple-choice items and the other for rating constructed-response items. The study also examined whether the judges were satisfied with the procedures and standards produced by the method.

*AJM1 Procedures*

The first application of the AJM was used to set multiple cutscores on the Pennsylvania Grade 8 Mathematics Achievement Test (see Plake & Hambleton, 2000, 2001). Although this assessment was comprised of 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 the study. In conjunction with a variant of the two-choice Angoff for the dichotomously scored multiple-choice items (described above), the AJM1 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 judges, (b) provided with a review of the performance standards and the assessment's scoring rubrics, (c) trained to use the AJM, and (d) required to participate in an AJM practice session. After the training and practice session, each group was given a different 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 judges. Working independently, the judges first predicted whether a typical student within each of the performance categories would be able to answer each multiple-choice item correctly. Then they classified each constructed response into one of several performance categories using the classification scale shown in Figure 1. Thus, each constructed-response item for each student was categorized as novice-low, novice-medium, novice-high, apprentice-low, apprentice-medium, apprentice-high, proficient-low, proficient-medium, proficient-high, advanced-low, advanced-medium or advanced-high. Once the judges completed their initial classifications, each

group was asked to compare their decisions. The discrepant classifications were discussed and the judges were asked to make any final independent adjustments to their ratings.

For the constructed-response items, the relationships between actual examinee scores and the judges' boundary category ratings were used to calculate the performance cutscores that separated one performance standard from another. The overall cutscores 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 cutscore produced by the yes/no Angoff variant for the multiple-choice items. Upon completion of the standard setting activities, the judges 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 cutscores for each of the constructed response questions are presented in Table 2. The results for the two common questions show that the cutscores 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 cutscores for the boundary categories of apprentice, proficient, and advanced were 1.28, 2.44, and 3.54 respectively. For common question 2, the weighted mean cutscores were 1.54, 2.52, and 3.56, respectively. Although the cutscores produced by each group for the two matrix sampled questions were not comparable because each group rated different questions, a similar trend observed across the boundary categories for the common questions occurred with the unique questions; that is, the cutscores generally increased across the categories (see Table 2).

The results for the 75 multiple-choice questions are presented in Table 3. The cut points based on the judges' performance estimates for the boundary categories were 24.21, 46.11, and 65.57. After averaging the cutscores for the boundary categories on the unique questions and adding them to the boundary category means for the common questions, the overall performance standard for each category was calculated by adding the boundary cutscores for the four constructed-response questions to the boundary cutscores 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.

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

Overall, the results from the AJM1 study suggest that the AJM is a feasible standard setting method for two main reasons: first, the judges were generally quite satisfied with the procedures and cutscores; and second, the AJM1 produced essentially replicable cutscores across the four subgroups of judges.

#### *Strengths and Weaknesses of the AJM1*

The strengths and weaknesses of the AJM are assessed in terms of two sets of criteria: Berk's (1986) ten criteria for evaluating standard setting methods, and Hambleton's (2001) 20 questions for evaluating a performance standard setting study. A four-point Likert scale (1 = strongly disagree, 2 = disagree, 3 = agree, 4 = strongly agree) was used to indicate the degree to which each criterion was satisfied.

Berk (1986) defines his criteria for evaluating a method in terms of the method's technical adequacy and practicability: A method that is technically adequate yields appropriate classification information, is sensitive to examinee performance, is sensitive to instruction or training, is 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.

The ratings in Table 4 indicate that the AJM1 satisfies four of the six criteria for technical adequacy and three of the four criteria for practicability. With respect to technical adequacy, the AJM1:

- yields appropriate classification information, which can be used to make appropriate inferences from the standards (e.g., the standards can be readily used to identify students who need remediation);
- is sensitive to examinee performance as the constructed response cutscores reflect actual student performance (however, this criteria is not met by the multiple-choice cutscores as they are produced using a variant of the yes/no Angoff procedure);
- is sensitive to the instruction or training that the examinees receive as the judges were mainly middle math school teachers who had taught Grade 8 Mathematics;
- is statistically sound in that the statistics used to summarize the judgements and describe test performance were appropriate and interpreted correctly;
- does not identify the true standard as measurement error was not taken into account; and
- does not yield decision validity evidence in that it did not provide estimates of the probabilities of correct and incorrect classification decisions.

With respect to practicality, the AJM1 is generally credible, and easy to compute and interpret to laypeople. However, it is not so easy to satisfy the physical demands of selecting, photocopying, and managing the large number of legible student papers required by the method.

The ratings in Table 5 indicate that 11 of Hambleton's (2001) 20 criteria for evaluating standard setting studies (see Table 5) were satisfied by the AJM1 study. The ratings indicate that the study presents little information about the representation and proportion of panelists selected. The panels were extremely small (only three or four members). Although four subpanels were formed, the generalizability of the constructed response results was limited as only two of the four items were categorized by all four subpanels. The small number of judges indicates that the study's resources were insufficient. The AJM1 was not field-tested. The method was appropriate and described in detail. At the beginning of the standard-setting meeting, the panelists received adequate information on the purposes, scoring and uses of the assessment. Some information was provided about the qualifications of the panelists (i.e., that they were high school, middle school and school administrators). However, other relevant demographic data about the panelists was not specified. The panelists were not administered any portion of the assessment prior to classifying the items. The panelists received excellent training on the method. The performance descriptors were clear and effectively used by the panelists. An iterative process was used for discussing and reconciling rating differences. However, no feedback was provided to the panelists. The process was conducted efficiently as the materials, forms, codes, etc. were clear. Panelists did not receive any performance data or consequential data. The approach for calculating the cutscores was relatively clear and understandable. Judges' evaluations of the process indicated that they had confidence in the training, standards and method. No validity evidence was compiled. The standard setting process was clearly

documented with the exception of specifying how the panelists were selected. The final criterion was considered not applicable because the standards were not used in practice as only a subsample of the items was examined in this study. Therefore, no steps were taken to communicate the performance standards to the stakeholders.

### *Summary of the Strengths and Weaknesses of the AJM1*

The AJM1 has several strengths. First, 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 standard setting because it does not require the selection of illustrative papers to serve as benchmarks for the performance levels in advance of the standard setting activities (Plake & Hambleton, 2001). However, it does require the selection of student papers that represent the full continuum on the assessment.

In general, the major strength of the AJM1 is that it is one of only a few examinee-centered approaches to setting standards on complex performance assessments that have the judges complete a realistic task (i.e., judges review actual students' constructed-responses and make judgments about student performance levels). Such procedures can help the judges avoid setting cutscores that are unrealistically high or low. However, while the AJM1 has the judges complete a realistic rating task when examining the constructed-response items, it requires the judges to complete an unrealistic, difficult task when rating the multiple-choice items (i.e., judges estimate whether or not minimally competent examinees in ten different categories could answer each question).

The main weakness of the AJM1, therefore, concerns the use of the yes/no Angoff 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, 1986). In addition, while the judges reported that they were generally satisfied with the AJM1 procedures and cutscores, research comparing the yes/no method with the Angoff estimated percentages method indicates that judges’ 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 the yes/no method to produce lower and higher cutscores than the panelists intend. Therefore, in addition to the problem of speculating about the performance of minimally competent 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 judges rate all of the mathematics achievement test items. Therefore, the AJM1’s cutscores could not be compared with the actual Pennsylvania state cutscores. This is unfortunate as such comparisons would have had the potential to support the reasonableness of the cutscores produced by the AJM1. As Kane suggests (2001), if two methods produce similar results, “we have more confidence in the resulting cutscores 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. Satisfying the physical demands of selecting,

photocopying, and managing the large number of legible student papers required by the AJM1 can be challenging (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.

The main disadvantage of the AJM1's boundary paper method for calculating the cutscores is that it only uses the scores from the student papers that the judges classify in the borderline categories to calculate the cutscores. Thus, not all of the information is used in determining the cutscores.

An additional weakness of the AJM1 study is that Plake and Hambleton (2001) neither specified how the judges were selected nor whether they reflected a balance of geographic distribution, ethnicity, and knowledge of the Grade 8 mathematics curriculum. To promote credible standards judges 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 authors do not report estimates of intra-rater consistency, inter-rater reliabilities, standard errors of measurement or indices of dependability. These estimates could provide evidence that (a) the judges used the achievement levels consistently, (b) the cutscores were reproducible and dependable, and (c) the procedures were credible enough to support reasonable interpretations about the meaning of the achievement levels. In the AJM1 study, however, the estimates of internal consistency were not likely reported because the stability of the results was affected by the small numbers of judges that participated in each of the panels.

## Conclusion

One goal of this paper was to promote an increased awareness of the processes for setting standards on complex performance assessments. A second intent was that the above discussion of the issues surrounding the topic of standard setting on complex performance assessments would extend current practitioners' knowledge of standard setting methods and inform existing standard setting practices. Although the AJM1 has several limitations, at this point in time, there is still no one best method for standard setting on complex performance assessments that are comprised of both selected and constructed response items.

Given that no standard setting procedure is ideal in and of itself, positive elements from several different methods have influenced the AJM1 procedures. For example, Angoff-based and BoW methods that provide a reasonable basis for setting standards have played a role in the development of the AJM1. Although the AJM1 is an attractive iterative procedure which utilizes independent item or component judgments in addition to actual student work, more research is necessary to replicate the results and to determine whether the AJM1 would produce high inter-rater reliability with more traditionally sized panels of 20 or more judges (Plake & Hambleton, 2000, p. 214).

Despite the limitations of the frequently used iterative modified Angoff method, when compared with the AJM1, the modified Angoff appears to offer a better balance between technical adequacy and practicability than the AJM1. However, if achievement tests continue to make greater use of constructed response items, examinee-centered methods may prove to be more appropriate methods for setting cutscores on these types of questions. Therefore, more research will be required to find a method for setting cutscores that is more consistent with more complex performance assessments.

## References

- 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.
- 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.
- Hambleton, R. (1998). Setting performance standards on achievement tests: Meeting the requirements of Title I. In L. Hansche (Ed.), *Handbook for the development of performance standards* (pp. 87-114). Washington, D.C.: U.S. Department of Education and the Council of Chief State School Officers.
- Hambleton, R. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 89-116). Mahwah, NJ: Erlbaum.
- Jaeger, R. (1989). Certification of student competence. In R. Linn (Ed.), *Educational Measurement* (3<sup>rd</sup> ed., pp. 485-514). New York: American Council on Education and Macmillan.
- Kane, M. (1994). Validating performance standards associated with passing scores. *Review of Educational Research*, 64, 425-461.
- Kane, M. (1995). Examinee-centered versus task-centered standard setting. In the *Proceedings*

- of the Joint Conference on Standard Setting for Large-Scale Assessment*, (Vol. 2, pp. 119-141). Washington, D.C.: National Assessment Governing Board (NAGB) and National Center for Education Statistics (NCES).
- 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.
- Mehrens, W. (1995). Methodological issues in standard setting for educational exams. In the

- Proceedings of the Joint Conference on standard setting for Large-Scale Assessment*, (Vol. 2, pp. 221-263). Washington, D.C.: National Assessment Governing Board (NAGB) and National Center for Education Statistics (NCES).
- Mehrens, W., & Popham, W. (1992). How to evaluate the legal defensibility of high-stakes tests. *Applied Measurement in Education*, 5, 255-283.
- Nassif, P. (1978, March). *Standard setting for criterion referenced teacher licensing tests*. Paper presented at the annual meeting of the National Council of Measurement in Education, Toronto.
- Plake, B. (1998). Setting performance standards for professional licensure and certification. *Applied Measurement in Education*, 11, 65-80.
- 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.
- 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. (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).
- 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. (2003, May). *Angoff method: Still a lot to do?* Paper presented at the annual meeting of the Canadian Society for Studies in Education, Halifax, NS.
- Shepard, L. (1995). Implications for standard setting of the National Academy of Education Evaluation of the National Assessment of Educational Progress Achievement Levels. In the *Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessment*, (Vol. 2, pp. 143-160). Washington, D.C.: National Assessment Governing Board (NAGB) and National Center for Education Statistics (NCES).
- Shepard, L., Glaser, R., Linn, R., & Bohrnstedt, G. (1993). *Setting performance standards for student achievement*. Stanford, CA: Stanford University National Academy of Education.
- Zieky, M. (2001). So much has changed. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 19-88). Mahwah, NJ: Erlbaum.

Table 1

*An Overview of the AJM1 Research Design*

## AJM1 – Pennsylvania

## Assessment

Pennsylvania Grade 8 Mathematics Achievement Test

## Panelists

14 (11 mathematics teachers and 3 school administrators)

## Classification method

Constructed-response items: Direct (question-by-question)

Multiple-choice items: Yes/no Angoff Variant

## Rating Scale

12-point (see Figure 1)

## Performance Descriptors

Novice, Apprentice, Proficient, or Advanced

---

*Note.* NAEP = National Assessment of Educational Progress.  
(Adapted from Plake & Hambleton, 2000, p. 205)

Table 2  
*Boundary Cutpoints from Groups Based on Constructed-response Questions – AJM1*

| Common Question 1 |            | Boundary Categories |           |  |
|-------------------|------------|---------------------|-----------|--|
| Group             | Apprentice | Proficient          | Advanced  |  |
| A                 | 1.27 (22)  | 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     | 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.* Number of papers in brackets (Adapted from Plake & Hambleton, 2001, pp. 294-295)

Table 3

*Boundary Cutpoints From Groups Based on Multiple-choice Component – AJMI*

| Multiple-Choice<br>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)

Table 4  
*A Comparison of the Strengths and Weaknesses of the AJM Using Berk's (1986) Criteria for Evaluating Standard Setting Procedures*

| Berk's Criteria for Evaluating a Method          | AJM1 |
|--|------|
| <i>Technical Adequacy</i>                        |      |
| 1) Yields appropriate classification information | 4    |
| 2) Sensitive to examinee performance             | 3    |
| 3) Sensitive to instruction or training          | 3    |
| 4) Statistically sound                           | 3    |
| 5) Identifies true standard                      | 1    |
| 6) Yields decision validity evidence             | 1    |
| <i>Practicability</i>                            |      |
| 7) Easy to implement                             | 2    |
| 8) Easy to compute                               | 3    |
| 9) Easy to interpret to laypeople                | 3    |
| 10) Credible to laypeople                        | 3    |

*Note.* 1 = strongly disagree; 2 = disagree; 3 = agree; 4 = strongly agree.

Table 5

*Evaluating the Strengths and Weaknesses of the AJM Using Hambleton's (2001) Standard Setting Evaluation Criteria*

| Hambleton's Criteria for Evaluating a Study  | AJM1 |
|--|------|
| 1) Representation and proportion of judges were considered                                       | 2    |
| 2) Panel was large enough and representative of stakeholders                                     | 1    |
| 3) Subpanels were formed to check generalizability of standards                                  | 3    |
| 4) Resources were sufficient   | 2    |
| 5) Method was field-tested and revised accordingly   | 1    |
| 6) Method matched the assessment and was described in detail                                     | 4    |
| 7) Purposes, scoring and uses of assessment were explained to panelists                          | 4    |
| 8) Data was collected on judges' qualifications and demographics                                 | 3    |
| 9) Judges were administered the assessment or a portion of it                                    | 1    |
| 10) Judges were trained to use the method  | 4    |
| 11) Performance descriptors were clear and effectively used                                      | 3    |
| 12) Appropriate discussion of ratings and use of feedback  | 3    |
| 13) Process was efficient (i.e., clear materials, forms, codes, etc.)                            | 3    |
| 14) Performance data was used effectively  | 1    |
| 15) Consequential data was used effectively  | 1    |
| 16) Procedure for calculating the standards was clear and understandable                         | 3    |
| 17) Judges' evaluations of the process demonstrated confidence in training, standards and method | 3    |
| 18) Validity evidence was compiled   | 2    |
| 19) Entire process was clearly documented  | 3    |
| 20) Performance standards were effectively communicated  | NA   |

*Note.* 1 = strongly disagree; 2 = disagree; 3 = agree; 4 = strongly agree.

|  |          |          |  |          |          |  |          |          |   |           |           |
|--|----------|----------|--|----------|----------|--|----------|----------|---|-----------|-----------|
| <b>Novice</b>  |          |          | <b>Apprentice</b>  |          |          | <b>Proficient</b>  |          |          | <b>Advanced</b>   |           |           |
| <ul style="list-style-type: none"> <li>minimal understanding of rudimentary basic concepts and skills</li> </ul> |          |          | <ul style="list-style-type: none"> <li>partial understanding of basic concepts and skills</li> </ul> |          |          | <ul style="list-style-type: none"> <li>general understanding of basic concepts and skills</li> </ul> |          |          | <ul style="list-style-type: none"> <li>broad and in-depth understanding of complex concepts and skills</li> </ul> |           |           |
| <b>1</b>   | <b>2</b> | <b>3</b> | <b>4</b>   | <b>5</b> | <b>6</b> | <b>7</b>   | <b>8</b> | <b>9</b> | <b>10</b>   | <b>11</b> | <b>12</b> |
| low  | medium   | high     | low  | medium   | high     | low  | medium   | high     | low   | medium    | high      |

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