The Alberta Journal of Educational Research *Vol. 52, No. 1, Spring 2006, 16-24*

W. Todd Rogers
University of Alberta

and

Kathryn L. Ricker
Educational Testing Service, Princeton, NJ

Establishing Performance Standards and Setting Cut-Scores

This article serves as an introduction to the following four articles in which four methods for establishing standards and setting cut-scores are presented and evaluated. The purposes, nature, and characteristics of performance standards are first reviewed. This is followed by a brief discussion of the methods for setting cut-scores. Berk's (1986) criteria for evaluating four of these procedures are then presented and described.

Cet article sert d'introduction aux quatre articles suivants dans lesquels sont présentées et évaluées quatre méthodes pour établir des normes et des seuils de passage. Dans un premier temps, les objectifs, la nature et les caractéristiques des normes de rendement sont examinés. Ensuite, on offre un survol des méthodes employées pour déterminer les seuils de passage. En dernier lieu, on présente et décrit les critères de Berk (1986) pour évaluer quatre de ces procédures.

The establishment of performance standards and the setting of cut-scores has become one of the critical steps in the design of high-stakes testing programs (e.g., high school graduation tests, licensure, certification, and recertification tests). For example, in the province of Ontario students must pass the Ontario Secondary School Literacy Test (OSSLT) prior to graduation. In Alberta high school students wishing to graduate must write and achieve at least an "acceptable standard" on grade 12 exit examinations (www.learning.gov.ab.ca/k-12/testing). University applicants whose prior education was delivered in a language other than English need to provide evidence that their command of English is acceptable before their applications are considered. Most often the evidence is provided by performance on a test such as the Test of English as a Foreign Language (TOEFL: Educational Testing Service, 2004). Persons wishing to immigrate to Canada as skilled workers need to exhibit at least basic proficiency in English and/or French before making their applications. Evidence of their language proficiency is assessed using tests referenced to the *Canadian Language Benchmarks 2000* (www.language.ca).

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. Kathryn Ricker is an associate measurement statistician in the Center for Statistical Analysis. At the time of writing she was a doctoral student in the Centre for Research in Applied Measurement in Evaluation at the University of Alberta.

Each of these decisions involves either a *go-no go* decision or a placement decision. Students who pass the OSSLT are eligible to graduate whereas those who do not pass must complete a language proficiency course to graduate. Only the applications of university applicants who do achieve a high enough score on the TOEFL are considered. Depending on their level of language proficiency (www.cic.gc.ca/english/skilled/), skilled workers can gain up to 24 of the 67 points necessary to be considered for immigration. The potential for false positives and false negatives makes these decisions controversial. Some applicants are passed when in fact they should not have been, and some applicants are failed when in fact they should not have been. Where the cut-score is set for each test can have enormous effects on the rate of false positives and negatives.

The establishment of performance standards and the setting of cut-scores also play an important role in monitoring educational systems. Often the monitoring is done by examining the proportions of students at various levels of performance such as *advanced, proficient, basic*, and *below basic* used in the National Assessment of Educational Progress and many state assessments in the United States. School performance in Alberta is monitored in part by examining the proportions of students who achieve at an acceptable level and at an excellent level. The intent of this type of reporting is to make the results more meaningful to legislators and the people they represent, namely, the public. Armed with this knowledge, they hold schools accountable for their performances. Again, where the cut-scores are set to separate students at adjacent levels can have a tremendous effect on the percentages and consequently the policy decisions that are made about school performance.

It is appropriate at this point to examine the uses of performance standards and the decisions to be made or conclusions to be drawn about the performances of the examinees and/or schools. Linn (1994) identified four major uses:

1. exhortation: motivate students and their teachers to achieve more and higher levels of the knowledge, skills, and behaviors identified in the performance standards;
2. certification: decide to which level examinees have acquired or possess the knowledge, skills, and behaviors called for in the performance standards;
3. exemplification: provide clarity to what students are expected to learn and acquire; and
4. accountability: reward teachers and schools that perform well as reflected by their students' performance, and sanction those who do not.

Determining the appropriateness of the conclusions reached about an individual or a unit comprising individuals based on their standing relative to a cut-score involves collecting validity evidence. The aim of this is to provide a convincing indication that examinees or candidates who score above the cut-score possess the behaviors called for in the performance standards, and that the goals of the decision process are well served. When collecting validity evidence, the purposes and uses of establishing performance standards and setting cut-scores need to be considered. For example, many procedures for setting cut-scores involve iterations or rounds. Group discussions among the panel members established to set the cut-scores and showing impact data between rounds are often used in an attempt to increase agreement among

panel members about cut-score locations. A major side effect of the additions is to regress *what should be* (exhortation) toward *what is* (certification). Further, to be well served, the rates of false positives and negatives should be as close to zero as possible in the case of certification, licensure, recertification, and graduation decisions. In the case of policy decisions regarding the performance of a school and/or curriculum, the results should be meaningful to legislators, education officials, parents, and the public. To achieve this aim, there must be a clearly demonstrable difference between adjacent performance levels.

Performance Standards

A *performance standard* describes the level of performance in terms of what examinees at that level *know* and *can do*. They provide *qualitative descriptions* of the knowledge, skills, and behaviors individuals at each level of distinction possess (e.g., *master, non-master; below basic, basic, proficient, advanced*). They must be stated in such a way that it is clear what are the differences in knowledge, skills, and behaviors between adjacent levels. These statements should be used to evaluate the relevance and representativeness of the items included in the assessment instruments and procedures used to find the degree to which the examinees acquired the knowledge, skills, and behaviors identified in the performance standards so that valid interpretations about performance can be made (Messick, 1989).

Cut-Scores

Cut-scores are used to delineate levels of performance as identified in the performance standards. They are specified on a score scale. Most often the score scale is a continuous rather than a discrete scale. Of concern is that the cut-scores mark points where the performance, as set out in the performance standards, changes. Put another way, the standard embodied in the cut-score needs to be reasonable or rational, given the purpose of the decision process. That is, examinees and applicants below the cut-score should not be able to answer correctly most, if any, of the items/tasks associated with performance above the cut-score.

The Arbitrariness of All of This

The establishment of performance standards and the setting of cut-scores are seen by many as an arbitrary set of activities. Glass (1978) was the first off the mark and implied that the whole exercise was folly. In response, Popham (1978), Block (1978), Hambleton (1978), Linn (1978), and Jaeger (1991) suggested that although establishing performance standards and setting cut-scores is arbitrary in the sense that they are based on judgments, they do not have to be arbitrary in the sense of being capricious. Indeed, many other aspects of measurement involve professional judgment such as developing test items and tasks, scoring constructed responses, interpreting test scores and information, and determining the validity of the interpretations made. Moss and Schutz (2001) summarized this situation as follows:

> Educational measurement professionals have brought considerable technical resources to bear on many of the challenges involved in creating defensible standards-based assessments. Nearly all of their activity, however, has focused on the aspects of the development process that take place after a set of

> standards has been created, appearing to rely on implicit and potentially problematic assumptions about what the standards "are."
>
> In their most basic sense, standards-based assessments are supposed to derive their warrant from a community consensus. It is fundamentally because the community has agreed on a set of standards that it is reasonable to use them to orient assessment instruments. Yet we really do not know much about the actual process of achieving such a consensus. We do not know what sort of agreement is reasonable to expect or what the implications of this lack of agreement might be for justifying the interpretation, use, and ultimate impact of standards-based assessment. (p. 38)

The distinction between establishing performance standards and setting cut-scores (Kane, 1994, 2001; Rogers & Dawber, 2002) is not universally made in the standard-setting literature (Cizek, 2001). Instead the process has often been simply described as standard-setting. However, the distinction helps to clarify the full process and to generate the validity questions and gather the needed validity evidence (Hambleton, 2001). Consequently, the distinction between the two processes is maintained in this article and the four articles that follow.

"Standard-Setting Methods"

Several different methods for setting cut-scores have been developed. More often than not, these procedures are referred to as "standard-setting procedures." However, as suggested above, in this set of articles these methods are referred to as setting cut-scores methods or cut-score setting methods.

There are four main approaches to setting cut-scores: relative cut-scores, contrasting or borderline group cut-scores, absolute cut-scores, compromise cut-scores, and item mapping. Relative cut-scores are set in relation to relative examinee performance and take into account quotas and/or space budget availability. For example, the number of seats available in classrooms and budgetary considerations often dictate the minimum grade point average students must have to gain admission to a university or a faculty in a university. Setting cut-scores using contrasting groups (Livingstone & Zieky, 1982) requires that there be, a priori, substantial evidence about which examinees are *masters* and which are *non-masters* in the sample of students used to set the cut-scores. The cut-score is set at a point that separates the groups. Absolute cut-scores involve making judgments about student performance on an item/task. The Nedelsky (1954) procedure and its modifications (Gross, 1985; Maguire, Skakun, & Harley, 1992) and the Angoff procedure (1971) and its modifications (Ricker, in press) are prominent procedures used to set absolute cut-scores. Setting compromise cut-scores involves aspects of both the contrasting groups and absolute procedures (Fielding et al., 1996; Hofstee, 1983). Item mapping is a general approach by which items or tests are mapped onto locations on a score scale such that students with scores at or above the location of the specific items or tests can be inferred to hold the knowledge, skills, and abilities to respond successfully to those items. The Bookmark (Lewis, Mitzel, & Green, 1996), Body of Work and its modifications (Kingston, Kahl, Sweeny, & Bay, 2001), and Analytic Judgment Method (Plake & Hambelton, 2000, 2001) are item-matching approaches used to set the locations.

Criteria for Evaluating a Performance Standard Setting Study

Various sets of criteria (Berk, 1986; Fitzpatrick, 1984; Hambleton, 2001) have been published to guide the conduct of a performance standard-setting study in which performance standards are established and the corresponding cut-score(s) are set in an attempt to reduce the degree of arbitrariness. Berk developed 10 criteria organized into two sets. The first set deals with the technical adequacy of the method used, and the remaining four with the practicality of method. The technical criteria were based on: (a) current expert opinion of standard-setting researchers; pertinent standards and guidelines from the *Standards for Educational and Psychological Testing* (American Educational Research Association, AERA, American Psychological Association, APA, and the National Council of Measurement in Education, NCME, Joint Committee, 1985); (b) then current expert opinion of "standard" setting; and (c) legal decisions that dealt with issues related to performance standards and their associated cut-scores. The technical criteria reflected the state of measurement research, as well as the political climate, that surrounded establishing performance standards and setting cut-scores at the time. The four remaining criteria were used to assess what Berk termed the *practicability*, or how feasible a method was to use and how easy it was for students, parents, and teachers to accept and interpret the results. Each of the criteria, organized in the two sets, is briefly described below.

Technical Adequacy

1. The method should yield appropriate classification information. The cut-scores set for each boundary condition (e.g., *below basic/basic; basic/proficient; proficient/advanced*) should allow for the classification of the examinees into the two adjacent categories that reflect demonstrable performance differences. Berk (1986) is implying with this first criterion that cut-scores must yield meaningful categories or classes. This criterion and the next require that the items/tasks in the assessment instrument be relevant to and representative of the performance standards.

2. The method should be sensitive to examinee performance. The method should be sensitive to the performance of the students as reflected by their performance on the items/tasks or test that reflect the performance standards. These characteristics include item difficulty and item discrimination at the item/task and the mean, standard deviation, and internal consistency (reliability) at the test level. For example, if the item difficulties for a minimum competence test range between 0.30 and 0.50, but the panel members responsible for setting the cut-score for minimum competence do not know this, then an unrealistically high cut-score could be set.

3. The method should be sensitive to instruction or training. This is the "opportunity to learn" criterion. Panelists need to be told and to take into account the actual instruction or training the applicants or examinees were exposed to before taking the test. For example, if decisions are to be made about minimum competence or high school graduation, then the students should be taught the knowledge and skills necessary to do well on the test. If the evidence suggests that this did not happen, then the cut-scores need to be adjusted accordingly to yield valid decisions.

4. The method should be statistically sound. Correct statistics should be used to describe test performance and summarize judgments. Care should be taken to ensure that the statistics used are appropriate for the measurement scale used (nominal, interval, and maybe ratio).

5. The method should identify the true cut-score. The true cut-score referred to here is a true score and not an observed score. If true scores are not available, measurement errors need to be taken into account. Berk (1986) quotes the *Standards for Educational and Psychological Testing* (AERA, APA, NCME Joint Committee, 1985): "the reliability … of decisions and inferences based on cut scores from educational certification test needs to be studied carefully" (p. 50) and "the standard errors of measurement should be reported for score levels at or near the cut score" (p. 22). The value of the standard errors should be close to zero. Care must be taken to use the error of measurement for a criterion-referenced interpretation and not to use the error of measurement of a norm-referenced situation (Brennan, 2001).

6. The method should yield decision validity evidence. Knowledge of the probabilities of correct and incorrect classification decisions are essential to determine the validity of inferences and decisions made based on cut scores. This concept is shown in Figure 1. The distribution of scores for the non-masters appears on the left; the corresponding distribution for the masters appears on the right. The cut-score was set to control the two types of error that can occur in criterion-reference test score interpretation: (a) "true" masters fail; (b) "true" non-masters pass.

Both types of error were considered equally serious in Figure 1. Consequently, the cut-score was set to equate the two rates of misclassification. This need not always be the case, though. Where the cut-off is established depends on which error, if either, is considered to have more deleterious consequences.

Practicability

1. The method should be easy to implement. The steps taken to set the cut-scores should be clearly understood by the panelists who will set them. The process should take a reasonable amount of time (2 days seems to be the maximum

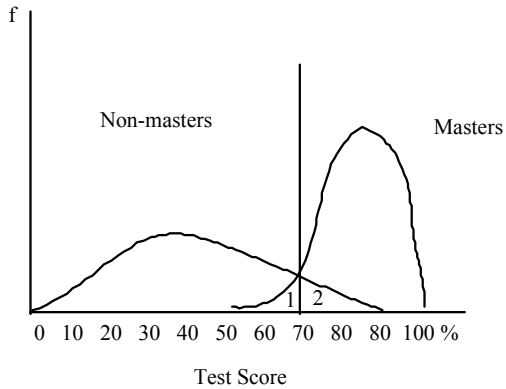

Figure 1. Distribution of scores for masters and non-masters.

amount of time). The selection of judges, the procedures they will employ, the training they receive, and the analysis of the agreement among judges should be clearly set out and systematically followed.

2. The method should be easy to compute. This criterion calls for calculations that can easily be done on a hand-held calculator or with available computer programs. This is needed where iterations or rounds are used, with discussion and sharing of information between consecutive rounds.

3. The method should be easy to interpret to laypeople. The cut-scores, together with a description of the cut-setting method used and validity evidence, should "be reported promptly to all appropriate parties, including students, parents, and teachers" (AERA, APA, & NCME Joint Committee, 1985, p. 53). The descriptions of the method should be clear and understandable, and presentation of results should be easily interpretable for the intended audiences and for those who may have to justify and defend subsequent decisions.

4. The method should be credible to laypeople. A "statistically magical" method is typically not credible to students, parents, and teachers. Similarly, neither is a method that is conceptually confusing and intuitively unsound. To ensure greater credibility, provision should be made to obtain and consider input from representative samples of people—students, parents, teachers, principals and central district staff, school board members—who might be affected or otherwise influenced by the results.

It is necessary to return to the sixth technical criterion. This criterion speaks to the consequences of using cut-scores to classify students or make decisions about what they know and can do. As pointed out and shown in Figure 1, both true masters and true non-masters can be incorrectly classified. However the validation of cut-scores cannot be done in isolation from the validation of performance standards for which the cut-scores were set. That is, performance data and the cut-score, as well as panels and panel reviews, should be used to assess the performance standards. Given the general purpose of the decision process (e.g., classify examinees according to the behaviors they possess), we need not only to conduct reviews of the constitutive definitions or descriptions of the performance standards, but also to evaluate the correspondence between the performance standards and the cut-scores. The aim of collecting the validity evidence is to provide convincing evidence that examinees or candidates who score above a cut-score possess the knowledge, skills, and behaviors called for by the corresponding performance standard, and that the purposes of the decision process are well served. To be well served, the purposes of the performance standards and cut-scores should be clear and accepted, the items/tasks and tests used to obtain the performance data should be relevant to and representative of the set of performance standards, and the rates of false positives and negatives should be as close to zero as possible in the case of certification, licensure, recertification, and graduation decisions. In the case of policy decisions regarding the performance of a school and/or curriculum, the results should be meaningful to legislators, education officials, and the public. To achieve this aim, clearly demonstrable differences between the performance levels must be shown, and the cut-scores must clearly delineate one level from the next.

The following four articles review and evaluate the Angoff (1971) procedure and its modifications, and the three item-matching approaches to setting cut-scores. Of the four procedures, the Angoff procedures are the oldest. When initially developed, most decisions were dichotomous such as deciding whether a student is or is not minimally competent. Despite their age, the Angoff procedures are still frequently used today. The Bookmark, in terms of its use, appears to be close to that of the Angoff procedures. The Body of Work and Analytic Judgment Method are two of the most recently introduced procedures. Unlike the previous two, these two procedures involve working with actual student products. Berk's (1986) 10 criteria are used as the base for each evaluation. No attempt is made to rank order the four procedures. Rather the intent of this set of articles is to provide information to professionals responsible for establishing performance standards and setting the corresponding cut-scores so that fair and equitable decisions about examinees are made. An Epilogue in which issues that need to be addressed when establishing performance standards and setting cut-scores follows the fourth article to provide direction for future research.

Note

All work done by K.L. Ricker was conducted while she was a graduate student in the Centre for Research in Applied Measurement and Evaluation, University of Alberta. The opinions presented here are solely those of the authors.

References

American Educational Research Association (AERA), American Psychological Association (APA), and National Council of Measurement in Education (NCME) Joint Committee. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Society.

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*(1), 137-172.

Block, J.H. (1978). Standards and criteria: A response. *Journal of Educational Measurement, 15*, 291-295.

Brennan, R.L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Cizek, G.J. (Ed.). (2001). *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Erlbaum.

Educational Testing Service. (2004). *Test of English as a foreign language*. Princeton, NJ: Author.

Fielding, D.W., Page, G.G., Rogers, W.T., O'Byrne, C.C., Schulzer, M., & Moody, K. (1996). Standard setting for a Test of Pharmacy Practice Knowledge: Application in high-stakes testing. *American Journal of Pharmaceutical Education, 60*, 20-29.

Fitzpatrick, A.R. (1984, April). *Social influences in standard setting: The effect of group interaction on individuals' judgments*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Glass, G.V (1978). Standards and criteria. *Journal of Educational Measurement, 15*, 237-261.

Gross, L.J. (1985). Setting cutoff scores on credentialing examinations: A refinement of the Nedelsky procedure. *Evaluation and the Health Professions, 8*, 469-493.

Hambleton. R.K. (1978). On the use of cutoff scores with criterion-referenced tests in instructional settings. *Journal of Educational Measurement, 15*, 277-290.

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

Hofstee, W.K.B. (1983). The case for compromise in educational selection and grading. In S.B. Anderson & J.S. Helmick (Eds.), *On educational testing* (pp. 109-127). San Francisco, CA: Jossey-Bass.

Jaeger, R.M. (1991). Selection of judges for standard setting. *Educational Measurement: Issues and Practice, 10*(2), 3-6, 10, 14.

Kane, M. (1994). Validating performance standards associated with passing scores. *Review of Educational Research, 64*, 425-461.

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

Kingston, N.M., Kahl, S.R., Sweeny, K.P., & Bay, L. (2001). In C.J. Cizek (Ed.), *Setting performance standards* (pp. 219-248). Mahwah, NJ: Erlbaum.

Lewis, D.M., Mitzel, H.C., & Green, D.H. (1996, June). *Standard setting: A bookmark approach*. Paper presented at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Boulder, CO.

Linn, R.L. (1978). Demands, cautions, and suggestions for setting standards. *Journal of Educational Measurement, 15*, 301-308.

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

Livingstone, A.A., & Zieky, M.J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.

Maguire, T., Skakun, E., & Harley, C. (1992). Setting standards for multiple-choice items in clinical reasoning. *Evaluation and the Health Professions, 15*, 434-452.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education and Macmillan.

Moss, P.A., & Schutz, A. (2001). Educational standards, assessment, and the search of consensus. *American Educational Research Journal, 38*, 37-71.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 14*, 3-19.

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

Plake, B.S., & Hambleton, R.K. (2001). The analytic judgment method for setting standards on complex performance assessments. In C.J. Cizek (Ed.), *Setting performance standards* (pp. 283-312). Mahwah, NJ: Erlbaum.

Popham, W.J. (1978). As always provocative. *Journal of Educational Measurement, 15*, 297-300.

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

Rogers, W.T., & Dawber, T.B. (2002). Book review of *Setting performance standards: Concepts, methods, and perspectives*. *International Journal of Testing, 2*, 161-168.