The Alberta Journal of Educational Research *Vol. 52, No. 1, Spring 2006, 80-83*

W. Todd Rogers
University of Alberta

and

Kathryn L. Ricker
Educational Testing Service

Epilogue

In the foregoing four articles, four cut-score setting methods are reviewed. We deliberately avoid making a recommendation of one procedure over others. Rather, we believe that those responsible for establishing performance standards and setting the cut-scores in a given situation should select what they consider the best procedure to use.

During the course of completing the evaluations using Berk's (1986) two sets of criteria, two main issues were identified that we believe require more research. These issues concern (a) the presentation of normative and impact data, and (b) test dimensionality, and given multidimensionality, the related need to determine which of two scoring models—compensatory or conjunctive—should be used.

Presentation of Normative and Impact Data

The role and effect of presenting student performance data are not clear. First, it is useful to distinguish between two types of student data that may be presented: normative and impact. Normative data are item- or task-level data that reflect the actual performance of relevant groups of examinees. These data are usually used during the cut-score setting procedure. In the case the Angoff (1971) procedure (Ricker, in press) and other like procedures (Nedelsky, 1954) and its modifications (Gross, 1985), normative data are usually presented during the third round to increase agreement among panel members and to provide an empirical check on the cut-scores set so that they are realistic (Hambleton & Powell, 1983; Jaeger, 1989; Linn, 1978). The Bookmark (Lewis, Mitzel, Green, & Patz, 1999), Analytic Judgmental Method (Plake & Hambleton, 2001), and Body of Work (Kingston, Kahl, Sweeny, & Bay, 2001) procedures use actual student data to set the cut-scores. It is argued that in the absence of normative data, unrealistically high or low cut-scores may be identified. However, when normative data are provided, the cut-score(s) set moves toward the level(s) of proficiency reflected by the normative data (Cizek, 2001).

Missing from the discussion of the appropriateness of normative data is consideration of the purposes of establishing performance standards and setting the corresponding cut-scores. If we accept the four major uses of performance standards and cut-scores identified by Linn (1994)—exhortation, exemplification, accountability, and certification—is the use of normative data appropriate for any of these four purposes? It may be that the use of normative data "effectively militate[s] against these uses" (Cizek, 2001, p. 11; Pellegrino, Jones, & Mitchell, 1999).

Turning to the use of impact data: these are data external to the items/tasks used to set the cut-scores that indicate the correctness of classification decisions made about examinees. These data reflect the numbers of students who would pass or who would be placed in one category as opposed to an adjacent category. The question to be answered with impact data is how many false positives and false negatives are there given the expectation that the numbers should be low, if not zero? In contrast to normative data, impact data usually are not available when the cut-scores are set. Furthermore, the collection of impact data is costly because it involves collecting data external to the items/tasks used to set the cut-scores. For example, in a minimum competence situation, it is necessary to wait for an appropriate time to see if those predicted to succeed do indeed succeed and those predicted not to succeed do indeed not succeed. Yet in contrast to normative data, impact data directly contribute to the validity of the performance standards and their associated cut-scores (Kane, 2001; Kane, Crooks, & Cohen, 1999) and, therefore, ought to be collected and used. The question here is how to collect these data in a timely and feasible manner.

The presentation of normative data presupposes that the panel members are unable to make the judgments necessary to identify cut-scores given that performance standards have been established and agreed to. This may well be true, but the absence of studies directed toward determining the processes panel members actually use and the values (Kane, 1998; Zieky, 2001) they hold suggests that such presuppositions may be faulty. Even if the cut-score setting is questionable, questions arise about how the normative data and, if available, impact data should be used to inform the reasonableness of the cut-scores set in light of the purpose for which performance standards were established and cut-scores set. Clearly there is need for research to address these issues.

Test Dimensionality and the Appropriate Scoring Model

An assumption must be made before applying one or more of the four methods reviewed for setting cut-scores that the construct underlying the test has only one dominant dimension. If a test is not strictly unidimensional, then it must be assumed that the distinct dimensions present on the test are *compensatory* in nature and in essence act as a unified construct in order to use the total test score alone. As used, the methods reviewed in the foregoing four articles use compensatory scoring models in that an examinee can meet a performance cut-score either by being minimally competent on all dimensions of a test or by making up for deficiencies on one dimension with strengths in other dimensions. In the case of the Analytic Judgment Method (Plake & Hambleton, 2001), individual questions that correspond to various behaviors are considered. However, the cut-scores for each question are then summed to determine the assessment level of each student in terms of the operational performance standards to which the test and its items are referenced.

However, the presence of more than one distinct, probably related dimension affords an opportunity to use a conjunctive model of scoring. Such an approach would be used when in the case of minimum competence, it was important that examinees be competent along each dimension. Conjunctive scoring is also more diagnostic, allowing the identification of areas of strength and problem areas that need to be addressed. With this model, performance

standards would be set for each dimension, and the corresponding cut-scores would be set. However, the percentage of students who fail to meet the cut-score along each dimension will probably increase. Will this be acceptable to policymakers and parents? Research is needed to ascertain the acceptability of conjunctive scoring to policymakers and parents.

Each of the models reviewed in this issue appears to be amenable to conjunctive scoring. However, the procedures would need to be applied to each dimension. For example, as mentioned above, the Analytic Judgment Method (Plake & Hambleton, 2001) attends to varied questions or sets of questions that arguably measure varied behaviors. The Bookmark (Lewis et al., 1999) could be applied to each dimension or alternatively, multidimensional item response models could be used to obtain the "books" for each dimension. Although the analytic machinery might be in place or within reach, there are still concerns. Will the test and its items and questions reflect simple or complex structure? It seems to us that it probably will be easier to establish performance standards and set cut-scores when there is simple structure than when there is complex structure. However, regardless of the complexity of the test structure, it is not clear how panel members would cope with distinct but related dimensions. Research is needed to determine how panel members would make their judgments given multiple dimensions and perhaps multiple cut-scores along each dimension.

Conclusion

The articles in this set reveal that establishing performance standards and setting cut-scores is complex and controversial. The fact that there is more than one method for setting cut-scores points to the elusive nature of the final outcome. The effect of normative data on the cut-score initially set and how it interacts, perhaps negatively, with the purposes of the performance standards and cut-scores adds to this elusiveness. The high cost of obtaining relevant impact data makes it difficult to know how well the cut-scores are working. And to recognize that many of the behaviors students are to learn and acquire are multidimensional in nature and that perhaps to be most useful, separate scores and, therefore, cut-scores should be set for each dimension leads to even more complexity. Yet the notion of competence and varying levels of competence is held by policymakers and parents and forms a normal part of life. Forcefully stated by Snow and Lohman (1989), psychometricians and cognitive psychologists need to work to ensure that whatever procedures are used for any of the scoring models, the performance standards established and the cut-scores set are reasonable, fair, and defensible given the purpose for which the performance standards and cut-scores are needed. Care needs to be taken to meet the standards for establishing performance standards and setting cut-scores set out in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999).

Note

All work done by K.L. Ricker was conducted while a she was 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, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

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

Berk, R.A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research, 56*(1), 137-172.

Cizek, C.J. (1999). Give us this day our daily bread: Manufacturing crises in education. *Phi Delta Kappan, 80*, 737-743.

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

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., & Powell, S. (1983). A framework for viewing the process of standard-setting. *Evaluation and the Health Professions, 6*, 3-24.

Jaeger, R.M. (1989). Certification of student competence. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485-514). New York: American Council on Education and Macmillan.

Kane, M. (1998). Choosing between examinee-centered and test-centered standard-setting methods. *Educational Assessment, 5*(3), 129-145.

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

Kane, M.T., Crooks, T., & Cohen, A. (1999). The design and evaluation of standard-setting procedures for licensure and certification tests. *Advances in Health Sciences Education, 4*, 195-207.

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

Lewis, D.M., Mitzel, H.C., Green, D.R., & Patz, R.J. (1999). *The bookmark standard setting procedure*. Monterey, CA: McGraw-Hill.

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 Education Statistics and National Assessment Governing Board Joint Conference on Standard-Setting for Large-Scale Assessments, Washington, DC.

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

Pellegrino, J.W., Jones, L.R., & Mitchell, K.J. (Eds.). (1999). *Grading the nation's report card: Evaluating NAEP and transforming the assessment of education progress*. Washington, DC: National Research Council.

Plake, B.S., & Hambleton, R.K. (2001). In C.J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 283-312). Mahwah, NJ: Erlbaum.

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

Snow, R.E., & Lohman, D.F. (1989). Implications of cognitive psychology for educational measurement. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 263-330). New York: American Council on Education and Macmillan.

Zieky, M.J. (2001). So much has changed: How the setting of cutscores has evolved since the 1980s. In C.J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 19-51). Mahwah, NJ: Erlbaum.