ED 442 828 TM 031 248

AUTHOR	Buckendahl, Chad; Impara, James C.; Giraud, Gerald; Irwin, Patrick M.
TITLE	The Consequences of Judges Making Advanced Estimates of Impact on a Cut Score.
PUB DATE	2000-04-15
NOTE	11p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Attitudes; Certification; *Cutting Scores; Elementary Secondary Education; *Expectation; *Judges; *Scoring; *Teachers
IDENTIFIERS	Standard Setting

ABSTRACT

School districts and credentialing agencies use information gathered in standard setting studies to establish minimum passing scores (MPS) for a variety of purposes. These scores may be used to make decisions ranging from subject remediation to licensure. Multiple standard setting methods may be used to provide a range of scores to the policy-making entity. The independence of these methods is important to the validity of the score recommendations. This paper examines the potential for introducing bias into the standard setting process by asking panelists for their expectations of impact prior to their making item performance estimates as might be done if using methods recommended by G. Dillon (1996) or Angoff (W. Angoff) "corrections" as recommended by D. de Gruiter (1985) or W. Hofstee (1983). Mixed results were found from five standard setting applications conducted in a variety of arenas. Three studies were conducted in school districts where teachers served as panelists, and two studies were conducted in the context of certification examinations so that the tests ranged from low-stakes to very high-stakes. A total of 70 panelists participated in all 5 studies. (Contains 10 references.) (SLD)

The Consequences of Judges Making Advanced Estimates of Impact

On a Cut Score

Chad Buckendahl
University of Nebraska – Lincoln

James C. Impara
University of Nebraska – Lincoln

Gerald Giraud
Methodist College

Patrick M. Irwin
Millard Public Schools

Paper presented at the annual meeting of the American Educational

Research Association, New Orleans, Louisiana.

April 25, 2000

Abstract

School districts and credentialing agencies use information gathered in standard setting studies to establish Minimum Passing Scores (MPS) for a variety of purposes. These scores may be used to make decisions ranging from subject remediation to licensure. Multiple standard setting methods may be used to provide a range of scores to the policy-making entity. The independence of these methods is important to the validity of the score recommendations. This paper examines the potential for introducing bias into the standard setting process by asking panelists for their expectations of impact prior to their making item performance estimates as might be done if using methods recommended by Dillon (1996) or Angoff "corrections" as recommended by de Gruiter (1985) or Hofstee (1983). Mixed results were found from five standard setting applications conducted in a variety of arenas.

Introduction

It has been suggested that when a formal process, such as that proposed by Angoff (1971), is undertaken to set cut scores on a test, that the final cut score may be unacceptable. Specifically, the cut score may have an impact (pass/fail rate) that is inconsistent with the expectations of the panelists and perhaps the policy makers who determine the final performance standard (Shepard, 1995; Dillon, 1996). Corrections for adjusting the cut score have been suggested by de Gruijter (1985) and Hofstee (1983) that entail asking panelists to estimate the percent of test items that will be answered correctly by the target candidate and also ask for the panelists' expectations regarding the percentage of target candidates who will fail. Dillon (1996) proposed using these data to form a "window of expectation" (p. 24) to determine the score region where the cut score might fall.

The purpose of this study is to examine the potential for introducing bias into the standard setting process by asking panelists for their expectations of impact prior to their making item performance estimates. It is likely that each panelist has some expectation in mind prior to undertaking the task of making performance estimates on each item. However, this estimate may be rather fluid and could be adjusted by the panelist based on any discussion during the training process or after impact data were provided (a common procedure in Angoff studies). However, if each panelist is asked to formalize their expectation by writing it down prior to making performance estimates, it may be more salient to them and therefore become a "target" value that they feel expected to hit when setting their individual cut score during the operational rounds of judgments.

Since cut scores represent policy decisions, multiple methods are often times used to recommend a range of possible cut scores to a policy making body (Livingston, 1995; Jaeger, 1989). If a method such as the one examined unduly influences another method, it may be necessary to avoid using them in combination. With an increase in the number of assessments that are employing standard setting methods to set cut scores, the importance of validity evidence becomes a greater concern for standard setting research.

Methods

We investigated this question in a variety of settings in which variations of the Angoff method were used to set a cut score. In each setting panelists were asked, prior to making their initial performance estimates, to provide their group performance expectation of the percent of candidates who would "fail" the test. After making initial item performance estimates, panelists were given feedback on actual item performance and impact data based on the panelists' initial performance estimates. If panelists were influenced by their own group performance expectations, it would be expected that their second round of item performance estimates would result in a shift in their individual cut score in the direction of their initial expectations. For example, if panelists' expectation was that 20% of the candidates would fail and the impact of the first round estimates (for all panelists) was that 25% of the candidates would fail, then panelists who were influenced by their group performance expectation would adjust their item performance estimates to lower the cut score.

Three studies were conducted in school district settings where teachers served as panelists and the cut score was to be used to identify students who needed "extra" help in a particular subject area (2 studies) or where the cut score represented the minimum score required for graduation (1 study). In addition to the school settings two studies were conducted in the context of setting the cut score for a certification examination. Thus, the five studies ranged from fairly low stakes tests to very high stakes tests.

Procedures

The same basic methods and procedures were followed in each of the studies, so only a general overview of the process is provided. For some tests all items were multiple choice, whereas others included both multiple choice and constructed response questions. When there was a mixture of both item types the feedback in the form of actual item performance and impact data were provided at several intervals instead of just once. A detailed discussion of the procedures for tests with a mix of item types is given in Buckendahl, Plake, and Impara (1999). Because this situation was most prevalent, it is the mixed model that is described.

The meeting at which the cut score was to be set opened with introductions of the participants and facilitators and was followed by a brief orientation to the standard setting process. The table of specifications for the test was described and participants engaged in a discussion to define the target candidate. This discussion (consistent with the procedures recommended by Mills, Melican, and Ahluwalia, 1991) provided panelists an opportunity to arrive at a common understanding of the behavioral characteristics of the target candidate. This was followed by a practice exercise in which panelists made item performance estimates for multiple choice items similar to those on the operational test. In all but one study, the performance estimates for multiple choice items were dichotomous, that is, panelists estimated whether or not the target candidate would answer correctly or not (as described in Impara and Plake, 1997). These item performance estimates were discussed and feedback on actual item performance was provided for each item. Moreover, a cumulative frequency distribution was provided and the "impact" of the cut score for the practice items was shown. It was explained that these items were not representative of the total test, nor was the distribution of scores necessarily similar to the total test. Panelists were engaged in discussion about the feedback data to insure they understood it and then were permitted to make a second estimate of item performance, just as they would be permitted to do when rating the operational items.

This selected response practice exercise was followed by practicing on one constructed response item where the panelists were provided benchmark responses (used to train scorers as to the definition of a response at that point on the score scale) and asked to identify the responses that would most closely represent the response of a target candidate. After reading the benchmark papers and making their selections, the papers were discussed in terms of the behavioral characteristics of the target candidate. The average score for all candidates was provided along with a cumulative frequency distribution. The impact of the panelists' cut score was then shown. Panelists were engaged in discussion about the feedback data to insure they understood it and then were permitted to review the benchmark papers and make a second selection, as they would be permitted to do in the operational test.

Following these practice ratings, panelists were provided a form on which the following question appeared:

"What percent of examinees (this word varied depending on the context of the study, in school settings it was 'students in the district', in the certification examination the word used was 'candidates') do you think will be classified as not proficient?" We did not use the term fail because that had a negative connotation in the school district setting. Panelists were told their group performance expectation would be averaged with the group performance expectations from all the other panelists and given to the policy making body (school board, board of trustees) as another element of data for consideration in setting the final cut score.

Because there were often several constructed response questions and thus several times when feedback data were provided, only the initial provision of feedback data is examined in the analysis. In most settings, the initial set of performance estimates was made on multiple choice items. Panelists were never told specifically what their first round cut score was, so they did not know if their cut score was consistent with their expectation or not. They did, however, know how their cut score was calculated (the sum of the items they said the target candidate would answer correctly).

If providing the group performance expectation did not bias the second round item performance estimates, then any changes in cut scores after panelists were provided with data would be random. If the panelists were influenced, then the second estimates of item performance would be systematic. Specifically, those panelists whose group performance expectation of the percent passing was below the percentage passing based on the initial cut score would be expected to lower their item performance estimates to result in a lower cut score. For example, a panelist whose initial group performance expectation was that 20 % would fail, but the first round cut score resulted in a 25% failure rate (indicating a higher cut score than expected), would make item performance estimates that would lead to a lower cut score in the second round, thus resulting in a lower failure rate.

A sign test was used to assess the possible influence of making the initial estimate of impact. Panelists in each study were sorted such that their group performance expectation was classified as either below or above the impact value after the first round. Each panelist's cut score was computed for the first round and for the second round. The direction of change was noted as moving toward their group performance expectation or away from it and the direction of these changes were tested in each study. For panelists whose group performance expectation was below the impact based on the total panelists' cut score, their second round cut score was expected to be lower, and vice versa for panelists whose group performance expectations were higher than the impact based on the total groups' first round cut score.

Results

Analyses included data from five standard setting studies conducted in various arenas during the past two years. The analyses employed simple sign tests. Panelists in each study were sorted such that their group performance expectation was classified as either below or above the impact value (percent of candidates that would fail the examination) after the first round. Each panelist's cut score was computed for the first round and for the second round. The direction of change was noted as moving toward their group performance expectation or away from it and the direction of these changes were tested in each study. The results of these sign tests are shown in Table 1.

TABLE 1. Summary of sign tests conducted on five studies.

Study	Valid N	# toward (+)	# away (-)	# no change	p (two-tailed)
A	22	17	5	0	.008
B	21	5	16	1	.013
C	11	10	1	2	.006
D	9	5	4	3	.500
E	7	6	1	4	.062

For a panelist's cut score to be included in the analysis, there had to be a change in that cut score between rounds. In Study A, changes were observed in the cut score for all 22 panelists. Seventeen panelists' second round cut score moved in a direction that would more closely align to their group performance expectation. Conversely, only five panelists' second round cut score moved away from their group performance expectation. The sign test yielded a statistically significant result (p = .008) meaning that there is a low probability the observed data occurred by chance. Practically, this means that panelists generally moved toward their initial group performance expectation.

For Study B, changes were observed in the cut score for 21 of 22 panelists. Five panelists' second round cut score moved in a direction that would more closely align to their group performance expectation. Conversely, sixteen panelists' second round cut score moved away from their group performance expectation. Using the smaller number (5) as the test statistic, the sign test yielded a statistically significant result (p = .013) meaning that there is a low probability the observed data occurred by chance. For this study, this means that panelists generally moved away from their initial group performance expectation.

For Study C, changes were observed in the cut score for 11 of 13 panelists. Ten panelists' second round cut scores moved in a direction that would more closely align to their initial group performance expectation. Conversely, only one panelist's second round cut score moved away from their group performance expectation. The sign test yielded a statistically significant result (p = .006) meaning that there is a low probability that the observed data occurred by chance. Practically, this means that panelists generally moved toward their initial group performance estimation.

For Study D, changes were observed in the cut score for 9 of 12 panelists. Five panelists' second round cut score moved in a direction that would more closely align to their group performance expectation. Conversely, only four panelists' second round cut score moved away from their group performance expectation. The sign test did not yield a statistically significant result (p = .500) meaning that the observed data could have occurred by chance.

For Study E, changes were observed in the cut score for 7 of 11 panelists. Six panelists' second round cut score moved in a direction that would more closely align to their group performance expectation. Conversely, only one panelist's second round cut score moved away from their group performance expectation. The sign test did not yield a statistically significant result (p = .062) meaning that the observed data could have occurred by chance.

Two of the five studies (A and C) yielded statistically significant results indicating panelists generally move toward their initial group performance expectations when conducting a sign test on the change in cut scores between rounds one and two. This provides some evidence that providing group performance expectations may influence the change in cut score judgments. One of the studies (B) yielded a statistically significant result indicating panelists generally moved away from their initial group performance expectations. This also provides some evidence that providing group performance expectations may influence the change in cut score judgments, however, in the opposite direction. Finally, two of the five studies (D and E) did not yield statistically significant results suggesting that the change in cut scores between rounds may be a random occurrence.

Conclusion

There is some limited evidence that having panelists provide group performance estimates prior to making item performance may influence the change in individual cut scores and the direction of that change. However, because it is expected that some change will occur between rounds one and two due to the influence of the feedback data, some of the statistically significant findings may be random results. Further study of this question using more powerful experimental designs is warranted because of the potential implications to cut scores set for high stakes examinations. As additional methods for recommending cut scores are employed, it is important to determine whether these methods result in truly convergent results or that one method influences another.

References

Angoff, W. H. (1971). Scales, Norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational Measurement, 2nd Edition, p. 508-600. Washington, DC: American Council on Education.

Buckendahl, C. W., Plake, B. S., & Impara, J. C. (April, 1999). Setting Minimum Passing Scores on High-Stakes Assessments that Combine Selected and Constructed Response Formats. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec, Canada. (ERIC TM029986)

de Gruiter, D. N. M. (1985). Compromise models for establishing examination standards. Journal of Educational Measurement, 22(4), 263-269.

Dillon, G. F. (1996). The expectations of standard setting judges. Clear Exam Review, Summer, 22-26.

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

Impara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. Journal of Educational Measurement, 34(4), 355-368.

Jaeger, R. M. (1989). Certification of student competence. In R. Linn (Ed.), Educational Measurement (3rd ed.) (pp. 485-514). Washington, DC.

Livingston, S. A. (1995). Standards for reporting the educational achievement of groups. Joint Conference on Standard Setting for Large Scale Assessments, Proceedings Volume II. Washington, DC: National Assessment Governing Board, National Center for Educational Statistics.

Mills, C. N., Melican, G. J., & Ahluwalia, N. T. (1991). Defining minimal competence. Educational Measurement: Issues and Practice, 10 (2), 7 – 10.

Shepard, L. A. (1995). Implications for Standard Setting of the NAE Evaluation of NAEP Achievement Levels. Joint Conference on Standard Setting for Large Scale Assessments, Proceedings Volume II. Washington, DC: National Assessment Governing Board, National Center for Educational Statistics.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC

TM031248

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: The Consequences of Judges Making Advance Estimates of Impact on a Cut Score

Author(s): Chad Buckendahl, James Impara, Gerald Giraud, Patrick Irwin

Corporate Source: Buros Center for Testing

Publication Date: April, 2000

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☒

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) and paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

□

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

□

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here,→ please

Signature: Chad W. Buckendahl

Printed Name/Position/Title: Chad Buckendahl, Research Associate

Organization/Address: 135 Bancroft Hall, UNL
Lincoln, NE 68588-0352

Telephone: (402) 472-6244

FAX: (402) 472-6207

E-Mail Address: biaco@unl.edu

Date: May 24, 2000

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20772
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com