DOCUMENT RESUME

ED 417 231 TM 032 178

AUTHOR	Plake, Barbara S.; Impara, James C.; Hertzog, Melody; Giraud, Gerald; Spies, Robert
TITLE	Utility of a Concept-Focusing Strategy on Judgmental Standard Setting Results.
PUB DATE	1997-00-00
NOTE	12p.; Paper presented at the Annual Meeting of the Midwestern Educational Research Association (Chicago, IL, 1997).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Concept Formation; High School Students; High Schools; *Judges; Mathematics Tests; Performance Factors; Probability; *Standards; Test Construction; Validity
IDENTIFIERS	Experts; *Standard Setting

ABSTRACT

Judgmental standard setting approaches rely on the perceptions of experts about examinee performance on a test. Traditional standard setting methods ask panelists to predict the probability that a randomly selected, hypothetical minimally competent candidate (MCC) will correctly answer test questions. Item performance predictions are difficult for panelists to do accurately; however, the validity of these performance standards rests on the ability of panelists to conceptualize the skills and abilities of the MCC accurately and to make accurate performance predictions. This study investigated the utility of a strategy to aid in the conceptualization of the MCC. Panelists were asked to envision the typical candidate and the MCC, and then to make item performance predictions for the typical student first and then the MCC. Results with 24 panelists predicting performance of students on a high school mathematics examination showed that the strategy resulted in significantly lower performance standards than did the traditional approach. Validity data were used to evaluate the accuracy of judgments resulting from the experimental and traditional approaches. More research is needed to clarify the utility of a concept focusing strategy on the judgmental performance standards. (Contains six references.) (Author/SLD)

Utility of a Concept-Focusing Strategy on Judgmental Standard Setting Results

Barbara S. Plake
James C. Impara
Melody Hertzog
Gerald Giraud
Robert Spies

University of Nebraska-Lincoln

Running head: Concept-focusing strategy

Paper presented at the 1997 Annual Meeting of the Midwestern Educational Research Association, Chicago, Illinois.

Utility of a Concept-Focusing Strategy on Judgmental Standard Setting Results

Abstract

Judgmental standard setting approaches rely on the perceptions of experts about examinee performance on a test. Traditional standard setting methods ask panelists to predict the probability that a randomly selected, hypothetical minimally competent candidate (MCC) will correctly answer test questions. Item performance predictions are difficult for panelists to do accurately; however the validity of these performance standards rests on the ability of the panelists to accurately conceptualize the skills and abilities of the MCC and make accurate performance predictions. This study investigated the utility of a strategy to aid in the conceptualization of the MCC. Results showed that the strategy resulted in significantly lower performance standards than did the traditional approach. Validity data was used to evaluate the accuracy of the judgments resulting from the experimental and traditional approaches. More research is needed to clarify the utility of a concept-focusing strategy on the judgmental performance standards.

Utility of a Concept-Focusing Strategy on Judgmental Standard Setting Results

Introduction

Typically, the purpose of standard setting procedures is to establish a minimum passing score on a test. This score is then used as the criterion to determine whether individual examinees pass the test. Examinees whose test scores fall below the minimum passing score do not pass the test. Consequences of not passing the test can be severe: licenses or certifications may be denied or students may be denied graduation or promotion in school settings. Therefore, the accuracy of these performance standards can be critically important to individual examinees in terms of their future admission in professions or other educational programs.

Several methods exist for establishing the minimum passing score on a test. Most of these methods were designed for use with multiple choice examinations. Some of the methods rely on empirical methods, such as regression modeling or contrasting groups, to set the passing scores. Other methods, called Judgmental Standard Setting Methods, depend on expert panelists' perceptions of the skills and abilities of examinees who are just barely competent in the area being assessed.

In Judgmental Standard Setting Methods, panelists are typically asked to predict the performance on the test questions by examinees who are minimally competent. Referred to as the Minimally Competent Candidates (MCCs), these examinees represent the group of candidates whose knowledge, skills, and achievements are just barely sufficient to warrant a passing score. The Angoff Standard Setting Method (1971), for example, requires panelists to predict the performance on each item in the test by a randomly selected, hypothetical MCC. Often this task is operationalized as predicting the proportion of 100 hypothetical MCCs who would correctly answer each question in the test.

The item performance predictions form the basis for determining the recommended performance standard. Most often, the item performance predictions are summed across the items in the test to set the individual panel member's estimate of the minimum passing score. These minimum passing scores are then averaged across panelists to determine the recommended minimum passing score, or performance standard, for the test.

Therefore, the resulting performance standards from the judgmental standard setting methods rely directly on the accuracy of the item performance predictions made by the panelists. These item performance estimates are actually conditional probability estimates, i.e., they represent the panelists' prediction of correct performance conditioned on the ability level of the MCC. The perception of the skills and competencies of the MCC, then, are central to the accuracy of these item performance estimates.

Research has shown that item performance estimates are difficult for panelists to make accurately. Even when the target group of candidates is the total group, panelists have difficulty providing accurate performance estimates (Lorge & Kruglov, 1953, Thorndike, 1980; Bejar, 1983). In a study focusing on the just competent student in sixth grade science, Impara and Plake (1995) found that teachers, who were familiar not only with their students but also with the assessment, systematically underestimated the performance of their "D/F" students (which served as the operational definition of the minimally competent student in their classes). Therefore, a strategy that improved the accuracy of item performance estimates would make an important contribution to the methodology of judgmental standard setting approaches.

The purpose of this study was to investigate the utility of a concept-focusing strategy on the performance standards set using a judgmental standard setting approach. Formation of a clear concept of the minimally competent candidate is critical to the accuracy of the judgments. The concept is generally developed through group discussion of the characteristics of this examinee, focused on the MCC's performance on the objectives measured by the assessment. It may be possible to improve the clarity of the concept in the minds of the panelists by contrasting it with a familiar, related concept. In this case, the panelists were instructed to envision two examinee groups: the MCC and the typical student. Panelists made item performance predictions for the typical student first, then made performance predictions for the MCC. The impact of this concept-focusing strategy was investigated in this study.

Method

A large midwestern school district served as the setting for this study. This school system has adopted a Gateway Assessment Program aimed at identifying students from differing developmental levels in a variety of content areas who are in need of additional educational programming. For each content area, Essential Learner Outcomes (ELOs) have been established. The district uses a series of assessments to measure student achievement tied to these ELOs; minimum passing scores are determined, using a judgmental standard setting approach, to identify students who qualify for additional services. The High School Mathematics Gateway Assessment was used for this investigation.

Instrument. The High School Mathematics Gateway Assessment consists of 62 questions, 30 items selected from the Metropolitan Achievement Test (MAT) and 32 items (called the Supplemental Test) developed specifically for this assessment to cover the remaining ELOs for high school mathematics. Items from the MAT were exclusively multiple choice, while items on the Supplemental Tests were primarily short answer and problem solving. The Supplemental Test items are scored using a pre-established rubric by trained assessors; the key provided by the MAT was used to score the MAT items. The assessment was administered to all students in tenth grade during the previous semester. Student scores had not been reported at the time of the standard setting workshop.

Panelists. A total of 24 panelists participated in the study, which was administered as part of the operational standard setting workshop for high school mathematics. Panelists represented a variety of content areas within mathematics and taught students at varying levels of mathematics, from remedial mathematics to pre-calculus. All of the high schools in the district were represented on the panel. The panelists were divided into two groups in such a way that there was a balance in the schools represented.

Procedure. Prior to participating in the standard setting workshop, most of the panelists were asked to make "global ratings" of their students on their projected performance on the High School Mathematics Gateway Assessment. Using only the table of specification as the frame of reference, these teachers were asked to classify their students into one of 4 performance categories: NO: have not mastered sufficient skills to be deemed masters; SOME: have mastered some of the skills to be deemed masters; YES: have mastered sufficient skills to be considered masters, and BORDERLINE: have just the minimum level of skills to be at the SOME level.

Panelists were kept together during the majority of the orientation and training. The purpose of the standard setting workshop was described. The panelists were all familiar with the High School Mathematics Gateway Assessment, but the table of specifications was reviewed nonetheless. After a discussion of the assessment, the panelists engaged in a concept-formation exercise where they were first asked to think of a specific student they felt was "just barely competent" in the ELOs measured by the test. By describing the knowledge, skills, and achievements of those students, the group arrived at a definition of the Just Competent Student (JCS). This discussion was directed specifically at the content components of the High School Mathematics Gateway Assessment. At the conclusion of this discussion, the panelists were divided into two groups, meeting in separate rooms.

Group 1, the control group, made their item performance predictions only for the JCS. For each item in the test, panelists were asked to make independent judgments of whether the JCS they had in mind during the earlier discussion would be able to answer this question correctly. This is a slight modification of the traditional Angoff Standard Setting Method where panelists are asked to make item performance estimates using the full probability scale. This variation, called the Yes/No Method (Impara & Plake, 1997) has been shown to provide comparable results to the Angoff approach in less time. After completing their initial item performance decisions, panelists were shown data on how students in the school system performed on the test during the most recent administration. Data included the proportion of the total group of students who answered each item correctly (p-values for each item) and the impact of employing the panelists' Round 1 cutscore on the proportion of students system-wide who would qualify for additional educational services. Panelists were given item performance information for each of the four performance groups (NO, BORDERLINE, SOME, YES) of students as well. Following discussion of this data, panelists were given the opportunity to revise their initial item performance decisions. Panelists engaged in a practice session involving these steps prior to embarking on the standard setting process with the operational assessment. After Round 1 in the practice session, panelists discussed their reasons for believing the JCS would answer each practice question in a particular way. This tended to further clarify the definition of the JCS and to help panelists connect the more abstract characteristics of a JCS with performance on a specific test item. There was no discussion of this type during the operational standard setting.

Group 2, the experimental group, followed the steps outlined above for Group 1 with one exception. Panelists in the experimental group were asked to make item performance decisions for two levels of student performance: the "typical" student followed by the JCS. Panelists were told to consider the typical student for the district as a whole, which may differ from the typical student in the courses they currently teach. As with the control group, panelists had an opportunity to practice making their judgments and interpreting empirical data. However, their discussion during the practice session involved reasons for the judgments they had made about the typical student as well as those made for the JCS.

Results

The performance standard derived from the control group's estimates equaled 36.08 (sd = 6.46), compared to 28.25 (sd = 6.40) for the experimental group. These values differ significantly, (t(22) = 2.92; p< .01). In order to assess the accuracy of the panelists' predictions, their predictions were compared to the p-values for the students categorized as BORDERLINE by the teachers making global ratings. The RMSE of the predictions was computed for each group across all 62 items. Neither group was particularly accurate; the control group's RMSE equaled 0.20 while the RMSE for the experimental group's RMSE value was 0.19. These values, though, represented systematic differences in direction of the error: The control group systematically overestimated the performance compared to the BORDERLINE group while the experimental group's predictions were systematically lower than the actual performance of the BORDERLINE group.

Conclusions and Discussion

The control and experimental groups produced performance standards that differed significantly. The validity data suggest that neither performance estimate was superior in accuracy; they differed nearly equally from the target group's performance but in opposite directions. However, the global ratings are themselves fallible as validity measures. For example, it is possible that global ratings are influenced by a halo effect to a greater extent than are item- level ratings. This might take the form of a systematic negative bias in global ratings of students who have not mastered several of the test's objectives, a tendency to generalize low expectations of their performance. If this were occurring, then some students who belong in the SOME category would probably be erroneously classified as BORDERLINE. This could inflate the p-values for the BORDERLINE group, which would change the conclusion drawn concerning the accuracy of the panelists' estimates. In such a case, the estimates of the experimental group would be more accurate than those produced by the control group. The foregoing scenario is purely speculative, but illustrates the need for further investigation of the accuracy of the global ratings and identification of additional sources of validity information.

One limitation of the current study was the absence of a clear definition of the typical student. After Round 1, the variance of individual panelists' estimates of the minimum passing score was approximately four times as large for the experimental group as for the control group. The variances converged during Round 2, after feedback about actual student performance had been given. One plausible explanation for this finding is that initially the experimental group was not uniform in its interpretation of "typical," but that the empirical data helped to create a common definition. In future studies, a definition could be developed explicitly as part of the training of panelists, in the same way that a definition of the JCS is developed.

Strategies to improve judgmental standard setting would benefit from a deeper understanding of the mental process a panelist goes through as he or she conceptualizes the JCS and makes a performance estimate. To some extent, "just competent" only has meaning if we have a concept of "competent." To what extent do panelists implicitly contrast the concept of the JCS to a typical student? Alternatively, do they apply some other standard of comparison as they define the JCS? Qualitative investigations of these sorts of questions are needed.

Despite these qualifications, the concept-focusing strategy used in this study appeared to have a considerable effect on the judgments of the panelists and warrants further study.

References

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.

Bejar, I. I. (1983). Subject matter experts' assessment of item statistics. Applied Psychological Measurement, 7, 303-310.

Impara, J. C., & Plake, B. S. (April, 1995). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Impara, J. C., & Plake, B. S. (March, 1997). An alternative standard setting approach: Variations on a Theme by Angoff. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Lorge, I., & Kruglov, L. K. (1953). The improvement of estimates of test difficulty. Educational and Psychological Measurement, 13, 34-46.

Thorndike, R. L. (1980). Item and score conversion by pooled judgment. Educational Testing Service Conference on Test Equating. Princeton, NJ.

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

ERIC®

TM028178

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Utility of a Concept-Focusing Strategy on Judgmental Standard Setting Results

Author(s): Plake, B., Impara, J., Hertzog, M., Giraud, G., + Spies, R.

Corporate Source: University of Nebraska - Lincoln

Publication Date: Oct 1997

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/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. 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 two options and sign at the bottom of the page.

☒ Check here **For Level 1 Release:** Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical) *and* paper copy.

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)

Level 1

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

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 2

□ Check here **For Level 2 Release:** Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical), but *not* in paper copy.

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither 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/optical 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: Melody Hertzog

Printed Name/Position/Title: MELODY HERTZOG/University-Wide Assessment Coordinator

Organization/Address:
University of Nebraska
208 Canfield Admin. Blg.
Lincoln, NE 68588

Telephone: 402-472-3899

FAX:

E-Mail Address: mhertzog@unlinfo.unl.edu

Date: 12/3/97

(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 sourc please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria a 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 reproduction release is held by someone other than the addressee, please provide the appropriate name and addres

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

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

ERIC Processing and Reference Facility
1301 Piccard Drive, Suite 100
Rockville, Maryland 20850-4305

Telephone: 301-258-5500
FAX: 301-948-3695
Toll Free: 800-799-3742
e-mail: ericfac@inet.ed.gov

(Rev. 3/96/96)