DOCUMENT RESUME

ED 398 251 TM 025 198

AUTHOR Plake, Barbara S.; Impara, James C.

TITLE Intrajudge Consistency Using the Angoff Standard-Setting Method.

PUB DATE Apr 96

NOTE 13p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New York, NY, April 9-11, 1996).

PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS Cutting Scores; *Interrater Reliability; *Licensing Examinations (Professions); Performance Based Assessment; *Physicians; *Psychometrics; Standards; *Test Items

IDENTIFIERS *Angoff Methods; Experts; *Standard Setting

ABSTRACT

This study investigated the intrajudge consistency of Angoff-based item performance estimates. The examination used was a certification examination in an emergency medicine specialty. Ten expert panelists rated the same 24 items twice during an operational standard setting study. Results indicate that the panelists were highly consistent, in terms of average absolute difference in item performance estimates (mean=0.073, standard deviation=0.066) and in resultant cutscores (16.37 versus 16.25). Features of the standard setting study and psychometric properties of the test were identified as possible contributors to this high level of intrajudge consistency. (Contains one table and eight references.) (Author/SLD)

Intrajudge Consistency Using the Angoff Standard Setting Method

Barbara S. Plake

James C. Impara

University of Nebraska-Lincoln

Running Head: Intrajudge Consistency

Intrajudge Consistency Using the Angoff Standard Setting Method

(Abstract)

This study investigates the intrajudge consistency of Angoff-based item performance estimates. Panelists rated the same 24 items twice during an operational standard setting study. Results indicate that the panelists were highly consistent, in terms of average absolute difference in item performance estimates (Mean = 0.073, SD = 0.066) and in resultant cutscores (16.37 v. 16.25). Features of the standard setting study and psychometric properties of the test were identified as possible contributors to this high level of intrajudge consistency.

Intrajudge Consistency Using the Angoff Standard Setting Method

The purpose of standard setting procedures is to establish the policy or standard to guide decision making. In many applications, the derived standard (or cutscore) determines which candidates or students pass or fail or are licensed or not. When used in such high stakes applications, the psychometric quality of the assessment is especially critical to obtaining valid and reliable decisions.

One of the most prevalent standard setting methods in licensure and certification examinations is the Angoff standard setting method (Sireci & Biskin, 1992). When using the Angoff method (Angoff, 1971) panelists are instructed to estimate independently the probability that a randomly selected minimally competent candidate (MCC) will correctly answer each item in the test. These item performance estimates are then summed across items to yield the each panelist's cutscore. These cutscores are then averaged across panelists to generate the cutscore, or standard, for the test.

The psychometric quality of Angoff-based cutscores has been recently been called into question (Shepard, 1994). In particular, the accuracy of the item performance estimates, especially for items in the difficulty extremes, have been criticized. Shepard argues that the task of making item performance estimates is difficult, if not impossible, for panelists do accomplish accurately.

One approach to investigating the ability of panelists to accomplish the task of accurately estimating item performance would be to compare estimated item performance by the minimally competent candidates to their actual item performance (Kane, 1984). While this approach has been attempted, (see Melican, Mills, & Plake, 1989), the validity of this approach is dependent on accurate identification of the MCC candidates. Typically, true MCC item performance results are not available.

If the task of making item performance estimates is to yield valid values, at minimum, these estimates should not vary dramatically across panelists or across rating occasion. In fact, if the same panelist, when evaluating the same item, provided radically different item performance estimates, very little faith could be placed in the standard devised from aggregates of these unreliable item performance estimates. Thus, another indicator of panelists' ability to produce appropriate item performance estimates would be to analyze the stability or consistency of item performance estimates resulting from repeated rating of the same items.

Very little research has been done to study the question of intrajudge consistency. Plake and Melican (1989) investigated the stability of Nedelsky-based item performance estimates across a year time span. On the 27 item set, the cutscores would have varied only 1 point across the two item ratings. Norcini and Shea (1992) found high levels of agreement between item performance estimates taken two years apart on a 24-item multiple choice examination. Most high stakes examinations consist of several hundred items. On longer examinations, a greater difference in cutscores most likely would have occurred.

The purpose of this study was to investigate the intrajudge consistency of Angoff-based item performance estimates when the items are rated within the same standard setting study. As these items were evaluated by the same judges during the same standard setting exercise, these results should give an optimistic indication of the degree of consistency present in Angoff-based item performance estimates. Lack of consistency in these item performance estimates would bring into question the utility of Angoff-based cutscores for setting performance standards on high-stakes examinations.

Method

Examination. The examination used for this study was a certification examination in an emergency medicine specialty. This 110 item multiple-choice test measures competencies in six content categories deemed by the association's personnel proficiency committee to be critical to practice. Four items were selected from each of the six content categories to serve as "repeaters" for the purpose of this study. In addition, the 22 items that formed the equating block from the previous certification examination were also included in the items to be rated by the panelist. This was done for two reasons: (a) to provide more items to diffuse the memory effect for the repeat items and (b) to allow for comparability of the ratings of the items in the equating block across operational forms. In total, panelists provided item performance estimates for 156 items.

These 156 items were split into two forms. Form A contained 78 items and was parallel in structure to Form B (i.e., both had the same number of items from the six content categories). Further, previous equating block items were positioned across the two forms so that they maintained their same item location from the previous operational form. The repeater items were located in the same item position across forms as well. To counter fatigue and order effects, the panelists were randomly divided into two groups. Group 1 rated items in Form A first, followed by Form B. The order was reversed for Group 2.

Panelists. A total of ten experts in the emergency medical specialty, selected by association's personnel proficiency committee to mimic the geographic and practice characteristics of the organization's national membership, formed the panel. There were seven male and three female panelists.

Procedures. Panelists were convened for a 1.5 day standard setting workshop. Training commenced after dinner on the first day and consisted of an introduction to the standard setting procedure, review of the examination's table of specifications, identification of the characteristics and skills of a hypothetical minimally competent candidate, practice in making item performance estimates, and interpreting feedback about total-group item difficulty values and impact of various cutscores on the proportion of examinees passing who took the 1995 examination. Time was devoted to clarifying procedures and to answering questions as needed.

Subsequent to training, panelists were given the first set of 78 items to evaluate. Panelists were informed that additional non-operational items were included in their packets for the purposes of maintaining a common scale to the previous cutscore. They were also told that some items may look familiar across the total 156 items they would be rating. They were instructed to rate each item independently. After the panelists completed their item performance estimates for their first set of 78 items, they were instructed not to discuss their item ratings among themselves and were dismissed for the night.

The second day commenced with a review of the item performance estimation procedures and a question-and-answer session. Panelists were next given their second set of 78 items and instructed to follow the same procedures as used for the first set the night before. Item performance ratings from Form A and Forms B constituted the Round 1 Angoff ratings for the items. An initial cutscore was calculated using only their ratings for the 110 operational items.

At the conclusion of their Round 1 ratings of the total 156 items (78 from Form A and 78 for Form B), panelists were give two pieces of information: (a) the actual item performance on these items by the total examinee pool, and (b) the initial cutscore value with the proportion of 1995 candidates who would have passed based on their Round 1 cutscore. After discussion, panelists were instructed to review each of the 156 items, in the same order as in Round 1, and revise their item performance estimates, if they so desired, based on the information provided and the earlier discussions. These Round 2 item performance estimates formed the basis for the analyses.

Results

For the 24 repeater items (four from each of the six content categories), the Round 2 item performance estimates were identified. To investigate the stability of the item performance estimates for these items, several analyses were undertaken. First, for each item, the correlation between repeat ratings across the ten panelists was determined. These 24 correlations were then averaged to produce on overall intrajudge consistency index across the repeat items. Next, the absolute difference between repeat item performance estimates across the ten panelists was determined. These values were also averaged to provide another indicator of intrajudge consistency across the repeater items. Finally, the two cutscores for the 24 items were calculated and compared to determine the impact of inconsistency of item performance estimates on the resulting cutscores. Table 1 contains the correlations and average absolute differences for the panelists' ratings of the 24 repeater items.

Insert Table 1 about here

Correlation of repeat items performance estimates. The correlations between repeat item ratings across panelists for the 24 repeat items ranged from a low of -.17 to a high of .91. The average correlation equaled .65 with a standard deviation of .27. The median correlation was .76.

Average Absolute Difference in Item Performance Estimates. Across the 24 repeat items, the average absolute difference in repeat item performance estimates ranged from a low of .05 to a high of .095. The average, across the 24 items, for the absolute mean differences in item performance estimates was .073 (standard deviation = .066).

Impact on Cutscores. Using the first item performance estimate from the ten panelists across the 24 repeat items, a cutscore of 16.37 was determined. A cutscore of 16.25 resulted from using the second item performance.

Discussion

A high level of consistency was observed across the two ratings of these 24 repeat items. The resultant cutscore differed by 0.12, a minimal amount. Even if the test was 4 times longer (96 total items), the expected change in cutscore would be less that .5 of a score point. In practice, high stakes examinations sometimes exceed 200 items. Under those circumstances, the cutscore would still vary by less than one score point.

On average, performance estimates for these items varied less the .10 on a scale from 0 to 1.00. While the panelists had the opportunity to utilize all 100 points in this range, most provided item performance estimates in multiples of .05. Had the panelists used the full scale more fully , it is possible that an even smaller average absolute difference would have resulted.

The correlations of first and second item performance estimates varied dramatically across the 24 items. Correlation values provide an index of rank consistency, not absolute consistency. As was noted in this dataset, even though the actual item performance estimates varied little across occasion, the rank order of these item performance estimates fluctuated substantially across the ratings by the panelists. It is reasonable to speculate that the panelists may have had a common magnitude of item difficulty in mind for the MCC's when they rated each item and the fluctuations observed from first to second ratings are indicative of random error around this latent item performance value. Because the magnitude of the fluctuations in absolute value in item performance estimates across the two rating sessions, on average, was quite small, the amount of random error observed in these item performance estimates was found to be quite low.

The degree of consistency found in these item performance ratings could have resulted from several features of the standard setting study. Several hours were devoted to training, concentrating on the skills and characteristics of minimally competent candidates. Through group discussion during training, over 18 characteristics were identified for the MCCs. Group discussion focused on the types of problems the minimally competent candidate might handle with ease (procedural-type questions) and ones they would likely find more challenging (problem solving in novel situations).

Further, the examination was overall somewhat easy for the 1995 candidates. The median item difficulty, based on the 1995 administration data, was .76. Internal consistency reliability was estimated using KR_{20} to be .86. The results of this study may have been different if the psychometric properties of the items had been substantially different. Additional research is needed to investigate the influence of item and test characteristics on the stability of Angoff-based item performance estimates.

Cutscores derived from standard setting procedures such as Angoff are often used to make critical decisions that affect examinees --- passing a licensure examination, graduating from high school, receiving a meritorious distinction are examples of high stakes decisions that are often based in part on comparing candidates scores to cutscores that were set using standard setting procedures. Therefore, the high levels of psychometric quality for the cutscore are required.

More research is needed to ascertain the boundaries of the conditions that support reliable item performance estimates. Research is also needed to investigate the validity of decisions based on standard setting methods such as Angoff. After all, if the estimates are inaccurate, it is of very little consolation to know that they are highly consistent. Reliability is a necessary but not sufficient condition for validity..

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.

Kane, M.T. (1984, April). Strategies in validating licensure examinations. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Melican, G.J., Mills, C.N. & Plake, B.S. (1987). Accuracy of item and performance predictions based on the Nedelsky standard setting method. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Nedelsky, L. (1954). Absolute grading standards for objective tests. Educational and Psychological Measurement, 14, 2-19.

Norcini, J., & Shea, J. (1992). The reproducibility of standards over groups and occasions. Applied Measurement in Education, 5, 63-72.

Plake, B.S., & Melican, G.J. (1989). Effects of item context on intrajudge consistency of expert judgments via the Nedelsky standard setting method. Educational and Psychological Measurement, 49, 45-51.

Shepard, L. (1994, October). Implications for standard setting of the NAE evaluation of NAEP achievement levels. Paper presented at the Joint Conference on Standard Setting for Large Scale Assessments, National Assessment Governing Board, National Center for Educational Statistics, Washington, DC.

Sireci, S.G., & Biskin, G.H. (1992). Measurement practices in national licensing examination programs: A survey. Clear Exam Review, 3(1), 21-25.

Table 1

Correlations and Average Absolute Differences for the

Panelists' Ratings of the 24 Repeater Items

Item	Average Absolute Difference	Correlation
01	0.030	0.82
02	0.060	0.74
03	0.091	0.17
04	0.075	0.49
05	0.080	0.60
06	0.055	0.80
07	0.055	0.88
98	0.055	0.78
09	0.025	0.85
10	0.050	0.86
11	0.075	0.39
12	0.065	0.88
13	0.050	0.81
14	0.025	0.91
15	0.077	0.26
16	0.095	0.54
17	0.060	0.69
18	0.049	-0.17
19	0.055	0.76
20	0.070	0.34
21	0.035	0.82
22	0.035	0.85
23	0.045	0.73
24	0.040	0.76
Average	0.073	0.65