DOCUMENT RESUME

ED 422 354 TM 028 906

AUTHOR	Plake, Barbara S.; Giraud, Gerald
TITLE	Effect of a Modified Angoff Strategy for Obtaining Item Performance Estimates in a Standard Setting Study.
PUB DATE	1998-04-00
NOTE	13p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Estimation (Mathematics); *Evaluators; Performance Factors; Standards; *Test Items
IDENTIFIERS	*Angoff Methods; *Standard Setting

ABSTRACT

In the traditional Angoff Standard Setting Method, experts are instructed to predict the possibility that a randomly selected, hypothetical minimally competent candidate will be able to answer each multiple choice question in the test correctly. These item performance estimates are averaged across panelists and aggregated to determine the minimum passing score for the test. Some applications have used a modification of this method where panelists are instructed to provide their item performance estimates in deciles, with each decile representing a 10-point probability range. The purpose of the study was to investigate the validity of this approach, in terms of comparability of results to that which would occur from the traditional, open-ended administration procedures. Differences were found between the minimum passing scores across the two methods. A variation that gathered restricted item performance estimates for the initial round and reverted to the full probability scale for round 2 was shown to reduce these differences. Discussion focuses on situations where this variation to the modified Angoff method may be particularly attractive. (Contains one table and four references.) (Author/SLD)

Effect of a Modified Angoff Strategy for Obtaining Item Performance Estimates in a Standard Setting Study

Barbara S. Plake

Gerald Giraud

University of Nebraska-Lincoln

Running Head: Modified Angoff Estimates

Paper presented at the 1998 annual meeting of the American Educational Research Association, San Diego, CA

Effect of a Modified Angoff Strategy for Obtaining Item Performance Estimates in a Standard Setting Study

(Abstract)

In the traditional Angoff Standard Setting Method, experts are instructed to predict the probability that a randomly selected, hypothetical minimally competent candidate will be able to correctly answer each multiple choice question in the test. These item performance estimates are averaged across panelists and aggregated to determine the minimum passing score for the test. Some applications have used a modification of this method where panelists are instructed to provide their item performance estimates in deciles, with each decile representing a ten-point probability range. The purpose of this study was to investigate the validity of this approach, in terms of comparability of results to that which would occur from the traditional, open-ended administration procedures. Differences were found between the minimum passing scores across the two methods. A variation which gathered restricted item performance estimates for the initial round and reverted to the full probability scale for round 2 was shown to reduce these differences. Discussion focuses on situations where this variation to the modified Angoff method may be particularly attractive.

Effect of a Modified Angoff Strategy for Obtaining Item Performance Estimates in a Standard Setting Study

Angoff (1971) recommended using estimates of item performance of the minimally competent candidate (MCC) as a means for establishing the minimum passing score, or cutscore, on a multiple choice test. Over the years, the Angoff method has become the dominant approach for setting performance standards on licensure and certification tests (Sireci & Biskin, 1992). The Angoff method requires that panelists estimate the probability, from 0 - 1.00, that a hypothetical, randomly selected MCC will be able to correctly answer each item in the test. These item performance estimates are aggregated across items and averaged across panelists to determine the minimum passing score for the test.

Several practitioners have used a modification of the traditional Angoff standard setting method (see for example, Cross, Impara, Frary, & Jaeger, 1984). In this modification, panelists are asked to make their item performance estimates using deciles, rather than the full probability range. This approach has been endorsed because often panelists restrict their item performance probabilities to deciles, or half-deciles, instead of utilizing the full 100 point scale.

The use of this method allows for the easy application of machine scorable answer sheets, facilitating quick turnaround in results. This is especially attractive when using an iterative approach (Jaeger, 1989) where panelists are given examinee performance data between rounds of item performance estimates. In order to provide impact data on the proportion of examinees who would pass or fail with the imposition of the Round 1 cutscore, the Round 1 results must be analyzed to determine the initial, Round 1, cutscore. With large groups and long tests, data entry is very time consuming especially if the 2 digit probability values are required. For example, it is not uncommon for a licensure test to consist of over 200 items and for a standard setting panel to be comprised of 20 or more panelists. In order to process the Round 1 item performance estimates using a microcomputer, for example, over 8000 keystrokes would be required. By limiting the item performance estimates to a single digit, data entry is streamlined and the use of scanable answer sheet is facilitated.

Many low budget licensure programs are strapped for resources and may not have the funds or technical capacity to include scanners or data entry personnel to be on hand to enter the data. Other strategies have been employed to speed up the processing of Round 1 data, including the use of multiple rating forms for the panelists' use. This allows for part of the data to be entered while the panelists are working on latter items in the test.

Because panelists often work at different rates, it is common for some panelists to have to wait for the rest of their colleagues to complete their item performance estimates. Panelists, especially those who are leaving professional practices to participate in the standard setting workshop, sometimes express aggravation over the amount of wait time that occurs between Rounds 1 and 2. Streamlining the coding decision to a single digit would likely speed up the process of making item performance estimates. For the agency funding the standard setting workshop, efficiencies in gathering and entering item performance estimates can result in less resources being devoted to the standard setting activity. A strategy that would yield valid results in less time would be an attractive alternative, from the perspective of the agency and most panelists.

The need for rapid data entry and quick turnaround of results is typically only crucial for the Round 1 results. Most standard setting panelists are not informed of the Round 2 results at the conclusion of the standard setting activity. This information is typically withheld from the panelists as the final cutscore decision is a policy decision left up to the Board of Directors. If preliminary information about the final cutscore were to be communicated by the panelists to practitioners in the field, this could have detrimental effects on the policy-making decisions. Therefore, the need for imposing a restriction on the estimated probability values is really only warranted at Round 1. For this reason, it was decided to investigate a variation of this modification of the Angoff method that would gather decile ratings at Round 1 and full scale probability values at Round 2. By obtaining item performance estimates using decile values, a quicker turnaround in computing the Round 1 cutscore would result without sacrificing much precision in the Round 2 cutscore.

The design of this study allowed for the investigation of the use of a modification of the Angoff Method in two ways. First, at Round 1, the design permitted a comparison of the results between the traditional and modified Angoff methods. By returning to the full probability scale for Round 2, the design also allowed for a direct comparison on the Round 2 cutscores across the two methods. Therefore, the impact of the modification strategy could be investigated at the conclusion of Round 1 (when the item performance estimates were not gathered on the same scale) and at the conclusion of Round 2 (when the results should be directly comparable).

Procedure

Data for this study were gathered during an operational standard setting study using a state level licensing examination in a health profession. Performance on this 100 item test determines, in part, who will be licensed in this profession. There are seven subcomponents of the test, with the items balanced proportionally across these subcomponents to align with the test's table of specifications.

A total of 14 experts were identified by the state licensing board for participation on the standard setting panel. These panelists represented the profession in terms of tenure, geography, and credentials. For the purpose of this study, these 14 experts were split into 2 equal sized groups matched to be as close as possible on these background characteristics.

The panelists met as a group for the initial orientation and training during which they were informed of the purpose of the standard setting study and given an overview of the procedures. In addition, the knowledge, skills, and abilities of the "just competent" practitioner in this health specialty were identified.

The two groups of panelists met in separate rooms for the practice exercises and operational tasks. During practice, each panel member made item performance predictions on a subset of items from a previous version of the examination, with items selected for the practice exercise to represent the seven content areas and a range of item difficulties. Panelists in Group A recorded their item performance estimates on a form that permitted use of the full probability range of values. Panelists in Group B were giving a coding system for reporting their item performance estimates. This coding system utilized decile ranges, with possible codes ranging from 0 (probability ranges .00-.10) to 9 (probability ranges from .90-1.00). During practice, panelists in both groups experienced making item performance estimates. They were given actual item performance data, including the impact of the cutscore on the examinees based on their item performance estimates on test consisting of only the practice items.

For the operational test, panelists in both groups were instructed to make item performance estimates using the method consistent with the one they used in practice. Following their Round 1 results, panelists were informed of their Round 1 cutscore and the impact of employing this cutscore on the proportion of examinees passing the test. In addition, the panelists were given the actual proportion of the total group of examinees who correctly answered each item. The feedback was identical in format to what was provided during practice.

After reviewing this information, the panelists were asked to make a Round 2 estimate for each item. For Group A, this entailed making revisions, as they deemed appropriate, to their Round 1 performance estimates. A second column of the rating form was provided for the panelists to record their Round 2 item performance estimates adjacent to their Round 1 values. Panelists were instructed to enter a value in the second column, even if that value was unchanged from their Round 1 estimate. This was done so that it would be clear that the panelists reconsidered their performance estimates at Round 2 for every item in the test.

Panelists in Group B were given the same feedback information as was provided to Group A. Instead of having them continue using the decile scale for their Round 2 ratings, the panelists were now instructed to use the full probability scale to make their Round 2 estimates. This allowed for a direct comparison of the Round 2 results from Groups A and B.

Results

Round 1 Results. Results are displayed in Table 1. The Round 1 cutscore derived from the Group A panelists' item performance estimates equaled 79.54 with a standard deviation across panelists' individual Round 1 cutscores equaling 3.68. The cutscore based on the results from Group B's Round 1 performance estimates equaled 72, with a standard deviation of 6.4. These values differ significantly ($t(12) = 2.73$, $p<.01$; effect size = 1.49). Therefore, the panelists in Group B provided a significantly lower cutscore than their counterparts evaluating the same items but using the traditional Angoff approach. The fact that the standard deviation for this group was larger than that for Group A is not surprising, given the restriction in score points available for their selection. However, the difference in magnitude between the Group A and B cutscores may be, in part, due to the scale used to record the item performance ratings in Group B. For Group B, panelists were instructed to use the lower boundary digit to each ten point probability range (e.g., the rating of 2 was used for probability values ranging from .20 - .29). On average, then, these values would be .05 lower than their probability rating counterparts. Across a 100 item test, this would add an additional 5 points to the average score, resulting in an average for Group B of 77, which is not significantly different than the value reported for Group A ($\underline{t}$ (12) = 0.75, p>.05). This adjustment, while reasonable and logical, it is not typically done when this modification of the Angoff method (see for example, Cross, Impara, Frary, & Jaeger, 1984). Therefore, for consistency purposes, the comparisons to the Round 2 results will be based, initially, on the unadjusted Group B Round 1 results.

Round 2 Results. The group results were much closer at Round 2. Group A results yielded a cutscore of 80.0 with a standard deviation of 4.9. Still more variable than their Group A counterparts, the Group B Round 2 cutscore was 78.0, with a standard deviation of 7.6. This difference in cutscores is not statistically significant ($\underline{t}$(12) =0.58 , p>.05). However, it is interesting to note that the standard deviation of the panelists' cutscores increased from Round 1 to Round 2, for both groups. Typically, the outcome of providing performance information and group discussion between rounds is a cutscore closer to the actual group performance (which occurred for both groups) and a reduced standard deviation. It is not clear why these two groups showed a higher standard deviation of panelists' cutscores following their second round of ratings, but this does not appear to be related to the experimental design. The standard deviation for Group A changed from 3.68 at Round 1 to 4.90 at Round

2, a change of 1.22. The standard deviation for Group 2 changed from 6.4 at Round 1 to 7.6 at Round 2, a gain of 1.20. Therefore, the amount of increase did not appear to be a function of group assignment. The increase could simply be an result of the small sample sizes used in this study.

Discussion

The results of this study suggest that different cutscores would result at Round 1 from the application of a modification Angoff method as opposed to the traditional Angoff method. However, using the strategy advocated in this study of utilizing the full probability range for the panelists' Round 2 ratings, resulted in minimal cutscore differences across groups.

Some practitioners select this modification of the Angoff approach as a means to facilitate the data analysis between Rounds 1 and 2 of the standard setting study. In that situation, the need for rapid data entry is only present for the Round 1 results. The results of this study suggest that the strategy of using a restricted set of probability ranges for the Round 1 item performance estimates, followed by the utilization of the full range of probability values for their Round 2 item performance estimates, yields results comparable to those from a traditional Angoff standard setting method.

Therefore, it appears that this modification of the Angoff standard setting method has promise for streamlining the standard setting study by making the data entry more efficient while keeping the results from Round 2 comparable to those derived from the traditional Angoff method.

However, the results when adjusting the Round 1 results for the unbalanced rating scale indicates strong agreement between the traditional Angoff and modified Angoff results, even at Round 1. Practitioners should be careful to verify the scale used for gathering ratings when this modification is employed as the scale along could be responsible for a bias in the results, both at Rounds 1 and 2.

No studies have been done on the comparability of the Round 2 results from a traditional Angoff standard setting method and the modifications employed in some earlier studies. Unless it can be shown that the large difference present in this study at Round 1 are reduced by Round 2, use of the modified Angoff method for the both Rounds 1 and 2 should be considered with caution. In this study, the Round 1 cutscores differed substantially between the traditional method and the modified approach. Until research can confirm the comparability of the Round 2 results across these standard setting approaches, the practitioner would be well advised to return to the traditional approach, at least for Round 2.

Further research is needed to verify the stability of the results found in this study. However, the two stage approach appears to have promise for meeting the psychometric needs of comparability of results to the Angoff standard setting approach while also meeting the needs of the practitioner who is interested in streamlining the analyses from the Round 1 results.

References

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

Cross, L. H, Impara, J.C., Frary, R.B., & Jaeger, R.M. (1984). A comparison of three methods for establishing minimum standards on the National Teacher Examination, Journal of Educational Measurement, 21, 113-130.

Jaeger, R.M. (1989). Certification of student competence. In R.L. Linn (Ed.), Educational Measurement, (3rd ed., 485-514). New York: Macmillan.

Sireci, S.G., & Biskin, B.H. (Winter, 1992). Measurement practices in national licensing examination programs: A survey. CLEAR Exam Review.

Table 1: Comparison of Round 1 and Round 2 results using the traditional and modified Angoff methods

Method	Round 1	Round 2
Traditional		
Mean	79.54	80.00
SD	3.68	4.90
Modified		
Mean	72.00	78.00
SD	6.40	7.60
Adjusted		
Mean	77.00	---
SD	6.40	---

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

ERIC®

TM028906

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Effect of a Modified Angoff Strategy for Obtaining Item Performance Estimates in a Standard Setting Study	
Author(s): Barbara S. Plake, Gerald Giraud	
Corporate Source:	Publication Date:

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	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY ____ Sample ____ TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	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	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY ____ Sample ____ TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ↑ [✓]	Level 2A ↑ [ ]	Level 2B ↑ [ ]
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	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: Barbara S Plake	Printed Name/Position/Title: Barbara S. Plake	
Organization/Address: University of Nebraska-Lincoln 208 Canfield Admin Bldg. Lincoln, NE 68588	Telephone: 402 472-3280	FAX: 402 472-6207
	E-Mail Address: bplake@unl.edu	Date: 4/27/98

(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:

THE UNIVERSITY OF MARYLAND
ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
1129 SHRIVER LAB, CAMPUS DRIVE
COLLEGE PARK, MD 20742-5701
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
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com