DOCUMENT RESUME

ED 465 778 TM 034 163

AUTHOR Wiley, Andrew; Guille, Robin
TITLE The Occasion Effect for "At-Home" Angoff Ratings.
PUB DATE 2002-04-00
NOTE 8p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 2-4, 2002). Supported by the American Board of Internal Medicine.
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Certification; Interaction; *Judges; *Licensing Examinations (Professions); Medical Education; *Standard Setting; Statistical Bias
IDENTIFIERS *Angoff Methods

ABSTRACT

This exploratory study extends the work done by B. Plake and others (2000) and R. Guille and others (2001) by investigating whether a negligible occasion facet would still be found when ratings for licensure and certification examinations were completed in isolation. A set of items was sent to a standard-setting committee to be reviewed at home, completely independently of all other members of the committee. Seven to nine raters reviewed each item. The examination was a medical certification examination administered once a year. Approximately half the 200 items had Angoff ratings already assigned from use on a previous examination, so only half of the items needed Angoff values assigned to them. For the study, a set of items needing Angoff ratings was sent to committee members at home, along with a set of 13 anchor items that had previously been rated. For the at-home ratings, the mean and standard deviation for the items were compared with the values obtained at the most recent standard-setting meeting. Results provide consistent evidence of a slight item-occasion interaction that appears to be the result of the increased variability in the ratings during the at-home session as compared to the normal session. Findings do provide preliminary support for the idea that Angoff ratings obtained at home may not be significantly different from ratings provided during traditional standard setting meetings. (Contains 1 table and 14 references.) (SLD)

TM



# The Occasion Effect for "at-home" Angoff ratings

Andrew Wiley, Robin Guille
American Board of Internal Medicine

This research was supported by the ABIM but does not necessarily reflect its opinions.

Paper to be presented at the Annual Meeting of the National Council on Measurement in Education, 2002, New Orleans, LA.

## Introduction

All examinations used for licensure or certification must set a passing score, above which a candidate is deemed suitably qualified for the license or certificate. However, by their nature, standards are somewhat arbitrary, resulting in a dichotomous classification of examinees even though the underlying variable is really continuous (Glass, 1978; Kane, 1994; Popham, 1978). Despite the arbitrary nature of standards, evidence can still be collected to support the use of a passing score on an exam (Norcini & Shea, 1997).

The 1999 *Standards for Educational and Psychological Testing* state that the "rationale and procedures used for establishing cut scores should be clearly documented" (pg. 59). Other efforts have outlined evidence needed to support the use of a passing score with an examination (Cizek, 1996; Norcini & Shea, 1997). One important feature continuously mentioned is that the rating of judges should be consistent with the difficulty of the test items (Reid, 1991). In addition, the judges' ratings should show stability over time (Norcini & Shea, 1992; Plake, Melican, and Mills, 1991).

Clauser, Clyman, and Swanson (1999) used generalizability theory to investigate the reliability of scores assigned to a performance assessment and looked for an "occasion facet" in the ratings. Evidence of this facet would indicate that each occasion of ratings being gathered had a unique component that influenced the judgments of the experts. While they did not find a consistent facet, they did recommend that researchers ensure that an occasion facet did not exist in their ratings.

Plake, Impara, and Irwin (2000), using standard setting ratings obtained from Angoff standard setting meetings, looked at the occasion effect for 24 items over 2 years. They found very consistent ratings across the years; the average difference in mean ratings was less than .01. As part of their discussion, Plake et al suggested that future research should focus on the generalizability of their results.

Guille, Wiley and Norcini (2001) reviewed a series of examinations to ascertain the consistency of standard setting ratings. Guille et al used generalizability theory to investigate if an occasion facet similar to the one reviewed by Clauser et al could be found when looking at standard setting ratings. They found that the difficulty of the items was the most prominent factor and explained the greatest amount of variance in the standard setting ratings. They found little evidence of an occasion facet.

This exploratory study extends the work done by Plake et al (2000) and Guille et al (2001) by investigating if a negligible occasion facet would still be found when ratings where completed in isolation. A set of items was sent to a standard setting committee to be reviewed at home, completely independent of all other members of the committee.

## Methods

The exam chosen for this study is a medical certification exam administered once a year. The paper and pencil examination consists of 200 multiple-choice items designed to evaluate the capacity of candidates to synthesize information, make medical judgments and to recall factual knowledge.

The exam has a passing score determined by a group of medical experts through the use of a modified Angoff procedure. Before each annual standard setting meeting, the 200 items to appear on the upcoming examination are selected. Of the 200 items, approximately 100 will have Angoff ratings already assigned from use on a previous exam. Because of this "Angoff item bank," normally only one-half of the items need Angoff values assigned to them.

At each Angoff session, all 200 items for an upcoming examination are reviewed. Items that have an Angoff rating are reviewed to ensure that the assigned value is still valid. The committee reviews all items without values and each member assigns a value to these. Before committee members determine an Angoff value, they are provided the item difficulty, item discrimination, and item difficulty for the top and bottom third of candidates. In addition to reviewing all items without a value, a subset of items that had Angoff values from previous years are rated again. The items in the subset are textually unchanged between years. These "anchor" items are evaluated in order to provide data on occasion and rater effects.

Raters were committee members who had worked on the design and development of the examinations. All committee members had previously been certified in their respective specialties and were recognized as experts. Committee members serve five-year terms and rotate on and off the committee, two each year. During each meeting, seven to nine committee members reviewed each item.

For this study, a set of items needing Angoff values was sent to committee members at home. All committee members completing the "at-home" Angoff ratings had participated in at least one full Angoff Standard setting and all had participated in the most recent Angoff meeting. Committee members were sent instructions reminding them how to assign their Angoff rating. The instructions discussed how to define the borderline group of candidates as well as how the ratings would be used to arrive at a passing score. In addition to the text of the items, committee members were provided the same statistical information that is provided during a regular meeting. For this rating, committee members were required to assign Angoff values for each of the items without consulting other members of the committee. Members completed the exercise and mailed the Angoff values back to staff, who compiled and aggregated the Angoff values across all committee members.

In addition to sending a set of items that had no Angoff values, a set of 13 "anchor" items was also sent to provide information regarding occasion and rater effects. All anchor items had been rated in the most recent meeting, 6 months prior to the time the "at-home" Angoff rating was completed. The 13 anchor items came from all content areas within the examination and, on average, were somewhat easier than the overall exams. In 1999, the average difficulty for the 13 items was 0.77 in comparison to a difficulty of 0.69 for the entire examination. In 2000, the average difficulty for the 13 items was 0.75 in comparison to a difficulty level of 0.67 for the exam.

Once the data had been compiled, means and standard deviations for the ratings were calculated. For the "at-home" ratings, the mean and standard deviation for the items were compared with the values obtained at the most recent standard setting meeting. Also, a generalizability analysis was completed using the GENOVA program (Crick & Brennan, 1983) for a random Item x Occasion x Rater design.

## Results

The mean "at-home" Angoff values obtained for the 13 anchor items can be seen in Table 1. The mean Angoff values from the previous standard setting meeting are also included. In addition, the differences in mean ratings across the two occasions are provided. For all but three of the 13 items, the difference between the two mean ratings was less than 10 points. Across all 13 items, the average difference between the two sets of ratings was 1.20 points. The average absolute value of the difference between the two means is slightly larger (6.88).

The standard deviations obtained for the anchor items, across all raters, for both sets of meetings, are also specified in Table 1. For all but 1 item, the standard deviation for the "at-home" ratings is larger than the

standard deviation for the ratings from the recent Angoff meeting. The mean difference in standard deviation across the two meetings was -4.20.

The results of the generalizability analysis can be seen in Table 2. While the item facet accounted for a significant amount of the variance, the occasion and rater facets did not account for any of the variance observed. The item-occasion interaction also accounted for a significant amount of variance.

## Discussion

This study was designed to evaluate the consistency of Angoff ratings obtained outside of the typical Angoff standard setting meeting. Committee members with experience with the Angoff standard setting evaluated a set of test items for a medical certification examination at home and provided ratings independent of all other committee members. If items obtained from this "at-home" standard setting were consistent with ratings obtained during a normal standard setting meeting, an attractive alternative to the normal standard setting meeting may be viable.

Reviewing the means and standard deviation of the Angoff values obtained, there did not appear to be drastic difference in the mean ratings. The mean difference for the items between the two sets of ratings was only 1.20; with an absolute value difference of 6.88. Given that the vast majority of raters used ratings divisible by 5, a change in rating of 5 points, is not that great. The "at-home" ratings did appear to create more variance among the raters. The average standard deviation for the "at-home" Angoffs was 7.80 as compared to an average of 3.60 obtained during a regular standard setting meeting.

In the generalizability analysis, the authors were most interested in determining if an occasion facet would be present. If this facet were present, it would indicate that something unique occurred at either the "at-home" Angoff sessions or the regular Angoff standard setting meeting. Reviewing the results, no evidence was found for a consistent occasion effect.

There does appear to be consistent evidence of a slight Item x Occasion interaction. This interaction appears to be the result of the increased variability in the ratings during the at-home session as compared to the normal session. This is not an unexpected finding. During a normal session, the presence of other committee members may create a pressure to conform to other committee member's ratings. When the ratings are completed at home, this pressure does not exist. However, the exact interpretation of this interaction is difficult to ascertain. In addition to differences in the method of standard setting, a significant period of time elapsed between the collection of the two sets of ratings. Because of this, any interpretation of the "occasion" facet, is confounded with a time factor. A further study eliminating or reducing the time period between the collection of ratings could begin to address this issue.

This paper presents findings that provide preliminary support for the idea that Angoff ratings obtained at home may not be significantly different than ratings provided during the traditional standard setting meeting. However, a number of limitations of this study need to be considered before drawing any conclusions. The first and most obvious example is the small number of items reviewed. In addition, this study was completed using only one medical subspecialty examination, how well results from such an examination can generalize to others is still unknown. Finally, this study used a set of raters who were involved in the test development process and who were experienced in the Angoff standard setting process. How well this research could generalize to raters not involved with the test development process or not as experienced remains to be seen. Given the limitations mentioned, and the increased variability seen with these items, it is advisable that further research be completed before making any decision on the feasibility of this new standard setting procedure.

Table 1: Mean and Standard deviation for 13 anchor items from the normal standard setting meeting and the "at-home" standard setting.

| | Mean Values | | | | Standard Deviation values | | | |
|---|---|---|---|---|---|---|---|---|
| | Normal session | At-home session | Difference | Absolute value (Difference) | Normal session | At-home session | Difference | Absolute value (difference) |
| 1 | 53.75 | 48.75 | 5.00 | 5.00 | 4.43 | 8.76 | -4.33 | 4.33 |
| 2 | 74.38 | 70.63 | 3.75 | 3.75 | 4.17 | 10.84 | -6.66 | 6.66 |
| 3 | 60.63 | 44.38 | 16.25 | 16.25 | 3.20 | 9.80 | -6.59 | 6.59 |
| 4 | 78.75 | 71.25 | 7.50 | 7.50 | 5.82 | 9.91 | -4.09 | 4.09 |
| 5 | 60.63 | 60.00 | 0.63 | 0.63 | 3.20 | 8.02 | -4.81 | 4.81 |
| 6 | 58.13 | 63.75 | -5.63 | 5.63 | 2.59 | 7.44 | -4.85 | 4.85 |
| 7 | 66.88 | 71.88 | -5.00 | 5.00 | 4.58 | 6.51 | -1.93 | 1.93 |
| 8 | 75.00 | 80.00 | -5.00 | 5.00 | 3.78 | 2.67 | 1.11 | 1.11 |
| 9 | 46.88 | 58.13 | -11.25 | 11.25 | 2.59 | 8.84 | -6.25 | 6.25 |
| 10 | 58.13 | 63.13 | -5.00 | 5.00 | 4.58 | 8.84 | -4.26 | 4.26 |
| 11 | 61.25 | 66.25 | -5.00 | 5.00 | 2.31 | 6.41 | -4.09 | 4.09 |
| 12 | 76.88 | 70.63 | 6.25 | 6.25 | 3.72 | 8.21 | -4.49 | 4.49 |
| 13 | 69.38 | 56.25 | 13.13 | 13.13 | 1.77 | 5.18 | -3.41 | 3.41 |
| Average | 64.66 | 63.46 | 1.20 | 6.88 | 3.60 | 7.80 | -4.20 | 4.38 |

Table 2: Estimates for Variance Components from Item x Occasion x Rater Design

| | | | | Sources of Variance | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Exam | Items | Occasion | Raters | Item (I) | Occasion(O) | Rater (R) | I * O | I * R | O * R | I * O * R |
| Test A | 13 | 2 | 8 | 63.07 | 0.00 | 0.00 | 21.72 | 0.00 | 3.18 | 44.76 |

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Brennan, R.L (1995). Standard setting from the perspective of generalizability theory. In *Proceedings of the joint conference on standard setting for large-scale assessments* (Vol. II, pp. 269-287). Washington, DC: National Center for Education Statistics and National Assessment Governing Board.

Cizek, G.J. (1996). Standard setting guidelines. *Educational Measurement: Issues and Practices*, 15(1) 13-21, 12.

Clauser, B.E., Clyman, S.G., & Swanson, D.B. (1999). Components of rater error in a complex performance assessment. *Journal of Educational Measurement*, 36, 29-45.

Crick, J. & Brennan, R.L. (1983). Manual for GENOVA: A generalized analysis of variance system [ACT Technical Bulletin #43]. Iowa City, Iowa: ACT.

Glass, G.V. (1978). Standards and criteria. *Journal of Educational Measurement*, 15, 237-261.

Guille, R., Wiley, A. & Norcini, J. (2000, April). The occasion effect in standard setting. Paper presented at the National Council on Measurement in Education. Seattle, WA.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, 64, 425-461.

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

Norcini, J.J. & Shea, J.A. (1997). The credibility and comparability of standards. *Applied Measurement in Education*, 10, 39-59.

Plake, B.S., Melican, G.J., & Mills, C.N. (1991). Factors influencing intrajudge consistency during standard-setting. *Educational Measurement: Issues and Practices*, 10 (2), 15-16.

Plake, B.S., Impara, J.C. & Irwin, P.M. (2000). Consistency of Angoff-based predictions of item performance: Evidence of technical quality of results from the Angoff standard setting method. *Journal of Educational Measurement*, 37 (4), 347-355.

Popham, W. J. (1978). As always provocative. *Journal of Educational Measurement*, 15, 297-300.

Reid, J.B. (1991). Training judges to generate standard setting data. *Educational Measurement: Issues and Practices*, 10 (2), 11-14.

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

ERIC

TM034163

# REPRODUCTION RELEASE

(Specific Document)

## I. DOCUMENT IDENTIFICATION:

Title: The Occasion Effect for "at-home" Angoff ratings

Author(s): Andrew Wiley, Robin Guille

Corporate Source: American Board of Internal Medicine

Publication Date: April, 2002

## 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 ↑ [X]

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: | Printed Name/Position/Title: Andrew Wiley / Psychometrician |
| Organization/Address: ABIM, 510 Walnut Street Suite 1700, Philadelphia, PA 19106 | Telephone: 215 446-3492 FAX: 215-446-3475 |
| | E-Mail Address: awiley@abim.org Date: April 10, 2002 |

(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 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**
**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**

EFF-088 (Rev. 2/2000)