DOCUMENT RESUME

ED 430 004 TM 029 717

AUTHOR	Plake, Barbara S.; Impara, James C.; Irwin, Patrick
TITLE	Validation of Angoff-based Predictions of Item Performance.
PUB DATE	1999-04-00
NOTE	35p.; Paper presented at the Annual Meeting of the American Educational Research Association (Montreal, Quebec, Canada, April 19-23, 1999).
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Cutting Scores; *Estimation (Mathematics); *Judges; Performance Factors; Reliability; Standards; *Test Items; *Validity
IDENTIFIERS	Accuracy; *Angoff Methods; *Standard Setting

ABSTRACT

Judgmental standard setting methods, such as the Angoff method (W. Angoff, 1971), use item performance estimates as the basis for determining the minimum passing score (MPS). Therefore the accuracy of these item performance estimates is crucial to the validity of the resulting MPS. Recent researchers (L. Shepard, 1994; J. Impara, 1997) have called into question the ability of the judges to make accurate item performance estimates for target subgroups of candidates, such as minimally competent candidates. The purpose of this study was to examine both the reliability and accuracy of item performance estimates from an Angoff standard setting application using as a model the framework suggested by M. Kane (1994). Results from the standard setting example involving 29 judges for an international certification program in financial management provide evidence that item performance estimates were both valid and reliable. Factors that might have influenced this high degree of reliability and validity in the item performance estimates in a standard setting study are discussed. (Contains five tables and nine references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 430 004

Validation of Angoff-based Predictions of Item Performance

Barbara S. Plake

James C. Impara

Patrick Irwin

University of Nebraska-Lincoln

Running Head: Validation of Angoff-based Predictions

Paper presented at the 1999 annual meeting of the American Educational Research Association, Montreal, CANADA

TM029717

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

□ Minor changes have been made to improve reproduction quality.

● Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Barbara Plake

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

BEST COPY AVAILABLE

Validation of Angoff-based Predictions of Item Performance

(Abstract)

Judgmental standard setting methods, such as the Angoff (1971) method, use item performance estimates as the basis for determining the minimum passing score (MPS). Therefore the accuracy of these item performance estimates is crucial to the validity of the resulting MPS. Recent researchers (Shepard, 1994; Impara, 1997) have called into question the ability of judges to make accurate item performance estimates for target subgroups of candidates, such as minimally competent candidates. The purpose of this study was to examine both the reliability and accuracy of item performance estimates from an Angoff standard setting application using as a model the framework suggested by Kane (1994). Results provide evidence that item performance estimates were both valid and reliable. Factors that might have influenced this high degree of reliability and validity in the item performance estimates in a standard setting study are discussed.

Validation of Angoff-based Predictions of Item Performance

Minimum passing scores (MPSs) are frequently used to make critically important decisions about individuals. Based in part on their performance on the test, for example, candidates may be granted a license to practice in their chosen profession or students may graduate from high school. The procedures used to set these MPSs, or cutscores, therefore must stand up to close scrutiny for validity and reliability.

Most licensure and certification programs use a judgmental standard setting method. The most prevalent judgmental approach is the Angoff (1971) method, with minor variations (Plake, 1998). The Angoff standard setting method relies on expert panelists making item performance estimates for minimally competent candidates (MCCs). These item performance estimates are aggregated across items and averaged across panelists to yield the recommended cutscores. Therefore, the accuracy of these item performance estimates is central to the validity of the resultant cutscore.

The Angoff method has recently come under criticism for being potentially invalid due to concerns about the ability of panelists to make accurate item performance estimates, especially for difficult or easy items (Shepard, 1994; Impara, 1997; Impara & Plake, 1998). Shepard (1994) argues that the task presented to the panelists (who set cutscores that classify students into several proficiency categories on the National Assessment of Educational Progress) was cognitively complex and exceeded the capacity of human raters. Impara and Plake (1998) used data from teachers in a public school setting who were very familiar both with their students' capabilities and the content of the test. They argue that if teachers (who have in-depth knowledge of their students' capabilities by virtue of working with them for a full academic year) have difficulty making accurate performance estimates for their students on a test that the teachers have used for several years, then it is improbable that panelists (who have little or no experience with candidates and little knowledge of the test) will be able to make accurate estimates of candidates' performance on items on the test.

The purpose of this study is to use strategies recommended by Kane (1994) to investigate the consistency and accuracy of item performance estimates from the Angoff standard setting method. Data came from three consecutive years of operational standard setting for a certification program in financial management. This certification program consists of three development levels with examinations for each of the levels.

Intra- and inter-rater consistency was examined. Intra-rater consistency was evaluated within the same standard setting occasion and inter-rater consistency was assessed across years. Accuracy was examined two ways. First, panelists' item performance estimates for the target group of candidates were compared with the actual item performance by a subgroup of panelists whose test scores fell within one standard error of measurement of the MPS. Second, the performance of passing and failing candidates on the same or next developmental level of the assessment program was considered. Retake success levels of failing candidates, with scores within one SEM of the cutscore, was contrasted with that of candidates who scored more than one SEM below the cutscore. Further, candidates who passed but who had scores within one SEM below the cutscore were identified as were candidates who scored more than one

SEM above the cutscore. When these four groups of candidates (two groups of failing candidates and 2 groups of passing candidates) took either the same level test (failing candidates) or the next level examination (passing candidates) in the assessment program, their passing rates were compared. Taken together, these analyses provide a comprehensive investigation of the consistency and accuracy of item performance estimates from an Angoff standard setting study.

This study addressed the following four research questions:

1. Do panelists make consistent item performance estimates within a standard setting study (intra-rater consistency)? Does their level of consistency vary as a function of item difficulty?
2. Do the same items, across years, receive consistent item performance estimates (inter-rater consistency across years)?
3. How close are the item performance estimates made by the panelists for MCCs to the actual item performance of candidates who score within 1 standard error or measurement of the MPS?
4. Do candidates who pass or fail the examination make predicted progress on the same or next developmental level on the examination program?

Method

The study was conducted as part of several operational standard setting workshops for an international certification program in financial management. Data were gathered across three years of standard setting: 1996, 1997, and 1998. This program assesses candidates at three sequential developmental levels: Stage I, Stage II, and Stage III. The Stage I examination consists of 230 multiple choice questions designed to measure entry level knowledge and applications for the content areas identified by the Stage I Examination table of specifications. Both the Stage II and III examinations consist of 15 – 20 open-ended, constructed response questions with 360 total points; each scored using an analytical scoring key. Point values for the questions vary as a function of the questions' importance within the table of specifications. Trained scorers using prescribed scoring guides score candidates' responses to the questions.

In setting the MPS for the Stage I Examination, panelists made two rounds of item performance estimates, with candidate performance information (total group) shared between rounds. For the Stage II and III examinations, panelists used a paper selection method for Round 1 to identify, for each item in the test, the work of the "Just Qualified Candidate" (JQC; this phrase was used in this application as a substitute for "Minimally Competent Candidate", or MCC). Candidate performance data were shared between rounds; at Round 2 panelists identified their anticipated level of performance on the question for the JQC candidate. For both selected-response (Stage I) and constructed-response (Stages II and III) questions, panelists made item performance estimates for the JQC, making the strategy consistent with the Angoff standard setting method.

MPSs point estimates, coupled with information about panelists' variability, are communicated to the organization's policy-making board. The board considers the MPS information from the standard setting studies, in addition to other relevant information. Consistent with Kane's (1994) advocacy that determining the performance standard is a policy issue, it is the responsibility of the policy board to set the final MPSs for the examinations. Table 1 shows descriptive statistics from three developmental levels for each of the three developmental levels, including overall candidate sample size, overall test means and standard deviations, internal consistency reliability estimates, MPS values set by the standard setting panels, and the policy board's decision for the MPSs.

For reporting purposes, each of the research questions is treated separately. The specific procedures are identified for each of the research questions. Within each study, results and limited conclusions are presented. A more comprehensive discussion and conclusions follow.

Research Question #1: Do panelists make consistent item performance estimates within a standard setting study (intra-rater consistency)? Does their level of consistency vary as a function of item difficulty?

Data from the 1996 operational standard setting for the Stage I examination were used to answer this question.

<u>Subjects</u>. A total of 29 panelists were selected by the certifying organization to be representative of a financial management field. The panelists were split into two groups (A and B) in such a way as to maintain this representation.

<u>Instrument</u>. The test was developed by the organization to measure the 8 content domains from the table of specifications. From the 230 test questions, two 115 item, psychometrically equivalent forms were prepared (Forms A & B). These forms were designed to be as equivalent as possible in terms of coverage of the table of specifications for the Stage I certification examination, average difficulty within content domain, overall average performance, and internal consistency reliability. From each of these two 115 item tests, 24 items from each were selected for this study, resulting in a total of 48 items. These items were

selected from the 8 content categories so that difficult (difficulty less than .40), moderate (difficulty between .40 and .70), and easy (difficulty greater than .70) items were selected, resulting in the 24 items per form for a total of 48 items. This 48-item form was labeled Form C. Thus, when rating the items in Form C, panelists re-rated 24 items that they had rated earlier either in Form A or B when they rated these items for the first time.

Procedure. Panelists met for the two-day standard setting workshop. Following an orientation that described the purposes and procedures of the standard setting workshop, a discussion that was designed to elicit the knowledges, skills, and abilities (KSAs) of the Just Qualified Candidate (JQC), and a practice session in which they made item performance estimates for selected items from a previous year's examination, the panelist groups were convened in separate rooms. Panelists were given a copy of the test form assigned to their group (either form A or B) and asked to make independent item performance estimates for each item. Panelists' Round 1 estimates and materials were collected.

Following completion of their Round 1 estimates for their respective form, all panelists were given the 48-item Form and asked to make item performance estimates for these items. Panelists did not have access to their Round 1 ratings when making their item performance estimates for Form C. This concluded the first day of the standard setting workshop.

At the start of day #2 panelists were given item performance data from the most recent administration of the examination, including information about the proportion of candidates who would pass using the cutscore based on their

Round 1 item performance estimates on their 115 item test form. Item performance data were also provided for Form C, but without the impact data.

Next, panelists completed Round 2 where they were given the opportunity to make adjustments to their item performance estimates, first for their respective form and then for Form C. As was the case for the Round 1 item performance estimates, panelists did not have access to the Round 2 ratings for their respective form when they completed their Form C item performance estimates. The Round 2 item performance estimates for the 24 repeater items within each form were analyzed for consistency across ratings. That is, for items that appeared both on Forms A and C the panelists' Round 2 ratings of these items were compared. Likewise, Round 2 item ratings were compared for those items common to both Form B and Form C.

Results

Items in Form C were classified by the results from their operational administration into easy, moderate, and difficult categories. The difficult items, on average, had a difficulty value (proportion correct) of 0.36 (sd = 0.06), the moderate items had an average difficulty of 0.56 (sd = 0.08), and the easy items had an average difficulty of 0.80 (sd = 0.06). The absolute value of the differences in item performance estimates for the Round 2 estimates (when the item was imbedded in the full 115-item form and when the item was rated as part of Form C) were averaged across the 8 items within each difficulty category.

For Form A, across the 15 panelists assigned to Group A, this average difference in absolute value of the item performance estimates for the 8 difficult items was 0.067 (sd = 0.039). For the moderately difficult items the mean difference was 0.067 (sd =0.046) and the average value was 0.083 (sd = 0.029) for the easy items. For Form B, the counterpart averages are as follows: difficult items, average absolute difference, 0.040 (sd = 0.023); moderate items, 0.055 (sd = 0.030); easy items, 0.059 (sd = 0.028).

Conclusions from Study 1. Panelists were systematically more consistent in their item performance for the hardest items and the moderately difficult items than for the easiest items. However, the magnitude of difference was small; on average this difference never exceeded 0.09 on a scale going from 0 to 1.00.

Research Question #2: Do the same items, across years, receive consistent item performance estimates?

Data from the 1996 and 1997 standard setting workshops were used to address this question.

Subjects. Thirty panelists attended the 1997 Stage I standard setting workshop, 12 of whom had participated in the 1996 workshop and 18 who had no prior experience as a panelists in a standard setting workshop.

Instrument. The 1997 Stage I Examination was built to the exact same specifications as the Stage I Examination in 1996. Twenty-four items were repeated in both the 1996 and 1997 examinations. This analysis is based on the average item performance estimates provided on these 24 repeater items across the 1996 and 1997 examinations. The results are based on two indices: (1) the average difference in item performance estimation for these 24 items from 1996 - 1997 and (2) the average absolute value of the difference in these item performance estimates from 1996 - 1997. A summary of the results from these analyses is shown in Table 2.

Average difference in item performance estimates. The average overall difference between the 1996 and 1997 ratings of these 24 items was -0.0001 (standard deviation = 0.056). This implies there was virtually no difference in the overall ratings of these items by both the 1996 and 1997 panelists. It is important to note that the average of the item rating differences were, for all practical purposes, zero. The average rating is used in the calculation of the MPS. Therefore, if these 24 items comprised the Stage I examinations in 1996 and 1997, these panelists' item performance estimates would have yielded essentially identical recommended MPS values for the 1996 and 1997 Stage I examinations. This provides strong evidence for the replicability of the MPS values across years for the Stage I examination.

Average absolute difference in item performance estimates. Because the difference in item performance estimates could be either positive or negative, their average value could be near zero, even though the magnitude of their differences was large. In order to address this issue, the absolute value of the differences in average item performance estimates for the 1996 and 1997 panelists were determined and then averaged across the 24 repeated items. The average of the absolute values of the was calculated to be 0.05 (standard deviation = 0.032). This is remarkable because these item performance estimates are on a scale that ranges from 0.00 to 1.00. The smallest absolute difference in item performance estimates across years was 0.006 and the largest was 0.130.

Because 12 of the 30 panelists participated in the Stage I standard setting in both 1996 and 1997, there is the remote possibility that they remembered their 1996 item performance estimates and simply repeated them in 1997. This would have the effect of inflating the agreement between the 1996 and 1997 item

ERIC

performance estimates artificially. To check for this possibility, the analyses were rerun, separating out the repeat panelists from the nonrepeat panelists. The mean difference in item performance estimates for the nonrepeat panelists was -0.009; average absolute difference equaled 0.05. For the repeat panelists, their average difference in item performance estimates equaled -0.008 with an average absolute difference of 0.05. Therefore, the results do not appear to be an artifact of repeat panelists from 1996 – 1997.

Conclusions for Study 2. These results indicate that the item performance estimates that form the basis for calculating the MPS from the standard setting workshops are consistent across different panels and the resulting item performance estimates appear to be trustworthy. Moreover, the inter-rater consistency from one year to the next was as high as the intra-rater consistency for panelists within years. Based on these analyses, it appears that the MPSs generated from the Stage I standard setting studies are highly likely to be consistent and repeatable across panels.

Research Question #3: How close are the item performance estimates made by the panelists for JQCs to the actual item performance of candidates who score within 1 standard error or measurement of the MPS?

To answer this question data from the Stage I standard setting workshops in 1996 and 1997[1] were used and data from the Stages II, and III standard setting workshops for 1996, 1997, and 1998 were used.

Panelists. In 1996 and 1997, respectively, approximately 30 panelists were convened to participate in each of the standard setting workshops. About half of the panelists were repeated across years. The others (18 in 1996 and 15 in 1997) were unique to that standard setting workshop. In 1998, a total of 28 panelists (10 panelists who participated in the 1997, but not the 1996, standard setting workshop and 8 panelists who participated in both the 1996 and 1997 workshops and ten panelists who had no prior experience as participants in a standard setting study) were convened. In each of the years the panelists were selected by the organization to be representative of the demographics of the organization. Within each year and standard setting panel, two groups were formed, maintaining to the degree possible the representation on each panel of the panelists to the organization's demographics.

Candidates. Many thousand candidates take the examination each year (see Table 1 for candidate numbers for each year). For each of the years 1996, 1997, and 1998 candidates whose total test score fell within one standard error of measurement of the cutscore set that year by the Board formed the "Empirical Just Qualified Candidate" (EJQC) group. Performance of these candidates on the test questions was compared to the item performance estimates made by the panelists for each year.

Instruments. The Stage I examinations in 1996 and 1997 were built to the same table of specifications and each contained 230 operational items. They were divided into two nearly psychometrically equivalent forms as described earlier. Likewise the 1996, 1997, and 1998 Stage II and III examinations were divided into two sets of items, designed to meet the following criteria: a) consists of nearly the same number of total points, and b) be nearly equivalent in cognitive demand and complexity. In all cases, within year both panelist groups considered a set of common questions (48 for Stage I; 2 for Stages II and III).

Procedure. As described above, for each Stage, panelists met for the two-day standard setting workshops. The same format was followed for each of the standard setting workshops, across years and levels.

General method for addressing the validity of panelists' anticipation of performance for the JQC. Data from the Stage I, II, and III examinations were considered for the standard setting workshops from 1996, 1997, and 1998. Using the reliability coefficient from the full examination, the standard error of measurement was calculated and the 67% confidence interval around the MPS was determined (MPS ± 1 SEM). All candidates whose test scores for that year fell within this interval were identified as the "Empirical Just Qualified Candidates" (EJQC). Selecting only the EJQCs, their performance on the questions that comprised the test was determined and compared with the standard setting panelists' expectations for how JQCs would perform on each question on the test. The difference in EJQC performance and the predicted JQC performance was calculated for each question. These differences were averaged as were their absolute values to ascertain how close, on average, the panelists' JQC estimates were to the performance of EJQC. These results are summarized in Table 3.

Stage I Examination. A total of 2,340 candidates in 1996 had a Stage I test score that fell within one SEM of the board-determined MPS. For these candidates, the proportion who answered correctly each of the 230 test questions was calculated and compared to the standard setting panels' average estimation of the proportion of JQCs who would correctly answer these items. On average the difference between the EJQC proportion and the panelists' estimate of the proportion of JQCs getting these items correct was -0.01 (sd = 0.09). In absolute value, these estimates differed on average 0.07 (sd = 0.05). Very similar results were found for the 2,937 EJQCs for the 1997 Stage I examination: average difference in actual and estimated performance = -0.001 (sd = 0.09); average absolute difference = 0.07 (sd = 0.06).

Stage II Examination. In 1996, a total of 2,666 candidates had test scores that fell within 1 SEM of the MPS set by the Board of Trustees. The average difference between the EJQC performance on the constructed response components of the Stage II examination and the panelists' estimation of JQC performance was -1.18 (sd = 1.92); in absolute value terms, the average was 1.78 (sd = 1.34). In 1997, there were 2,815 EJQCs. The average difference in actual and anticipated performance equaled -0.28 (sd = 1.83); absolute value differences averaged to 1.29 (sd = 1.30). In 1998, there were 3,340 candidates whose test scores fell within 1 SEM of the MPS set by the Board of Trustees. The mean difference in actual (EJQC) and anticipated (JQC) performance equaled 0.03 (sd = 1.34); absolute difference values averaged to 0.96 (sd = 0.92).

Stage III Examination. The results for the Stage III examination, across 1996, 1997, and 1998 are strikingly similar to those reported for the Stage II examinations. In 1996, a total of 1,175 candidates had scores that fell between the MPS and plus or minus 1 SEM. The average difference in their actual performance on the 1996 Stage III examination and that anticipated by the standard setters for the 1996 Stage III examination was -1.40 (sd = 2.34); the absolute value of these differences averaged 1.44 (sd = 2.31). For 1997, there were 1,956 Stage III candidates whose test scores qualified them for EJQC status for this analysis. The average difference between actual and anticipated performance on the 1997 Stage III questions equaled -0.31 (sd = 1.37); average absolute value difference was 1.16 (sd = 0.77). On the 1998 Stage III examination, 3,004 candidates had scores between the MPS and 1 SEM; their average performance difference from that expected by the 1998 standard setting panelists was -0.26 (sd = 0.77); average absolute difference was 0.64 (sd = 0.49).

Conclusions for Study 3. Regardless of whether the analysis was based on the Stage I, II, or III examination in 1996, 1997, or 1998, the conclusions that can be drawn from these results are fairly consistent. There is substantial agreement between the panelists' estimation of the performance of the JQC and the actual performance of candidates whose scores are close to the MPS. For the multiple-choice questions on Stage I, these differences, on average, were .010 or smaller and in absolute terms were routinely less than a tenth of a point. For the Stage II and III examinations, these difference rarely exceeded a full point, and most often were less than half a point on the total test scale of 360 possible points. There appears to be a very high degree of accuracy in the panelists' anticipated performances of the JQC.

Research Question #4: Do candidates who pass or fail the examination make predicted progress on the next developmental level on the examination program?

Data from the Stage I, II, and III examinations for 1996, 1997, and 1998 were used in this analysis. This analysis took an in-depth look at candidate performance during the years 1996 – 1998. The analyses focus on two categories of candidates. The first category consisted of those candidates who were successful in their first assessment experience during 1996 or 1997 (those who either passed Stage I or Stage II in 1996 or 1997). The analyses concentrated on how they performed when (and if) they took the next level during the 1997 – 1998 time period. The second category of candidates are those who were not successful in their initial assessment experience during 1996 or 1997 (those who did not pass the Stage I or II or III examinations in 1996 or 1997). The analyses track how these candidates faired when (and if) they retook the examination in subsequent years during the 1997-1998 time period.

Category 1: Candidates who passed either Stage I or Stage II in 1996 or 1997. There were 5,946 candidates who took and passed Stage I in 1996 and then took Stage II in 1997; 2,739 who took (and passed) Stage I in 1996 and then took the Stage II in 1998, and 6,864 candidates who passed Stage I in 1997 and then took Stage II in 1998. The correlation of their scores on Stages I and II indicates the degree to which scores on the Stage I examination are consistent with scores on the Stage II examination. These correlations are as follows: 1996 – 1997: 0.61; 1996-1998: 0.34; 1997-1998: 0.65[2]. All of these correlations are significantly greater than zero. Therefore it appears that candidates who pass Stage I tend to score in a consistent manner across Stages I and II.

A similar analysis was conducted for the candidates who passed Stage II in either 1996 or 1997 and then took the Stage III examination, either in 1997 or 1998. Correlations for scores across Stages II and III for these candidates are as follows: 1996 – 1997 (n = 4,308) 0.47; 1996 – 1998 (n = 2,241) 0.18; 1997-1998 (n = 4,670) 0.56. Again, these correlations are all significantly greater than zero.

The next set of analyses, still focusing on the candidates who passed either Stage I or Stage II in the 1996-1997 time period, examined how candidates performed on the subsequent level examination as a function of how well they performed on their initial examination. Thus, for those candidates who took

Stage I in 1996 or 1997, but whose scores were close to the cutpoint, the analyses examined their performance when (and if) they took the Stage II examination in the years 1997 or 1998. A similar analysis was conducted for those candidates with scores close to the cutscore, but who passed the Stage II examination in either 1996 or 1997.

Candidates who took the Stage I examination in 1996 had to score 146 or higher to pass the examination. Of the candidates who took the Stage II examination in 1997, 930 had scores on the 1996 Stage I examination between 146 and 152 (a standard error of measurement above the cutscore). Of these 930 candidates, only 298 (32%) passed the Stage II examination in 1997. In 1998, there were 426 candidates who took the Stage II examination for the first time who had scored between 146 and 152 when they took the Stage I examination in 1996. Of these 426 candidates, 162 (38%) passed when they took the Stage II examination in 1998. For candidates who took the Stage I examination in 1997, a score of 148 was needed to pass the examination. In 1998, there were 1,275 candidates who took the Stage II examination who had scored between the cutscore one standard error of measurement above the cutscore on the 1997 Stage I examination; 420 (33%) of these candidates passed the Stage II examination in 1998.

For Stage II, of the 1,403 candidates who scored between the cutscore and 1 SEM above the 1996 cutscore, 41% of them passed the Stage III examination in 1997. Of the 759 candidates who took the Stage III examination and who had scored between the cutscore and 1 SEM above the Stage II cutscore in 1996, 38% passed. For the candidates who scored between the cutscore and 1 SEM above the 1997 Stage II cutscore, 1,462 attempted Stage III in 1998. They had a pass rate of 41%.

Therefore, regardless of whether the candidates took Stages I, II, or III, if they passed, but scored close to the cutscore, their chance of passing the next level examination was less than 50%. The next set of analyses singled out those candidates who did exceptionally well on their first assessment experience (passing either the Stage I or II examinations with scores more than 2 standard errors of measurement above the cutscore) and examining how they did when (and if) they took the next level examination. As before, these analyses focus on the cohort of candidates during the 1996 – 1998 time period.

There were 4,197 candidates who took the Stage II examination in 1997 who had scored more than 2 SEMs above the cutscore when they took the Stage I examination in 1996. These candidates tended to have a high pass rate on the Stage II examination in 1997; 76% passed the Stage II examination in 1997. There were 965 candidates who scored more than 2 SEMs above the Stage I cutscore on the 1996 Stage I examination who took the Stage II examination for the first time in 1998. Of these 965 candidates, approximately 68% passed the Stage II examination. In 1998 there were 4,625 candidates with Stage I scores more than 2 SEMs above the Stage I cutscore in 1997 who took the 1998 Stage II examination. Of these candidates, over 83% of them passed the Stage II examination in 1998.

For the Stage II examination, similar results were found when high scoring candidates took the Stage III examination. For 1997, there were 1,597 Stage III candidates who scored more than 2 SEMs above the cutscore of the 1996 Stage II. Nearly 80% of these Stage III candidates passed the examination in

1997. In 1998, 313 candidates who scored more than 2 SEMs above the Stage II cutscore took the Stage III examination for the first time. Of this small group of Stage III candidates, 68% passed the Stage III examination in 1998. There were 236 candidates taking the Stage III examination in 1998 who had scored more than 2 SEMs above the Stage II cutscore in 1997. Of this group of candidates in 1998, 82% passed the Stage III examination. These results are summarized in Table 4.

Therefore, it appears that candidates who passed the Stage I and II examinations during 1996 – 1997 tended to also pass when (and if) they first attempted the next Stage examination in the following years. Candidates who scored extremely well on their Stage I or II examinations had a much higher probability of passing the next level examination than the general candidate groups. These results lend credence to the integrity of MPS values for the assessment program, as it is to be expected that high performing candidates on one level would be high level performers on subsequent levels of the examination program.

Category 2: Candidates who failed either the Stage I or II or Stage III examination in 1996 or 1997. The next set of analyses focused on the performance of candidates who did not pass when they took the Stage I, II, or III examinations in either 1996 or 1997. These analyses investigated whether candidates who scored close to the cutscore when they failed their Stage I, II, or III examinations were more likely to pass when they retook the examination than those candidates who scored more than one SEM below the cutscore when they failed their examination. For the Stage I examination, of the 406 candidates who scored close to the cutscore (within one SEM below), but failed the Stage I examination in 1996 and retook it in 1997, 63% passed in 1997. Of the 1,695 Stage I candidates who scored more than 1 SEM below the cutscore when they took the Stage I examination in 1996, only 29% passed the Stage I examination in 1997. There were 108 candidates who scored within one SEM below the 1996 cutscore who retook the Stage I examination in 1998. Of these candidates, 73% passed the Stage I examination on their initial retake in 1998. Of those candidates who scored more than 1 SEM below the cutscore when they took the Stage I examination in 1996, only 38% passed the Stage I examination in 1998 on their initial retake. Of the 586 1997 Stage I candidates who failed by less than 1 SEM who retook the Stage I examination in 1998, 84% passed. Of those 2,474 candidates who scored more than 1 SEM below the 1997 Stage I cutscore who retook the Stage I examination in 1998, 44% passed. Therefore, for the Stage I examination, it appears that candidates who fail, but are close to the cutscore, have a higher probability of passing the examination when they retake it within 2 years than do candidates who score much lower. These results are presented in Table 5.

Similar results were found for the Stage II examination. Those candidates repeating the examination who had scored close to the cutscore had a higher probability of passing the examination on retake than did candidates who scored lower. For the 1996 – 1997 Stage II repeat candidates, of the 688 who scored within 1 SEM below the cutscore, 63% passed in 1997. Only 31% of the 782 candidates who scored more than 1 SEM below the Stage II cutscore in 1996 passed in 1997. For the 221 candidates who retook the Stage II examination in 1998, but had scored within 1 SEM of the cutscore in 1996, 65% passed; of the 431 who retook the Stage II examination in 1998 who had scored more than 1 SEM below the cutscore when they failed the examination in 1996, 43% passed. Likewise for the candidates who failed the Stage II examination by less than 1 SEM in 1997, 66% passed in 1998 whereas of those candidates with scores more than 1 SEM below the cutscore in 1997, 44% passed in 1998. Exactly the same pattern is present when the Stage III candidate results are compared for those who fail, but score close to the cutscore with those who score lower. Again these results suggest that overall there is consistency in performance on these certification examinations. It is expected that candidates who fail, but score close to the cutscore, would be more likely to pass upon repeat testing, all things being equal.

Conclusions Study 4. These analyses examine the consistency of candidate performance across the examination program by addressing how successful and unsuccessful candidates progress through the program or when they retake failed examinations, respectively. When looking at the successful candidates, those who score high on lower level examinations tend to score high on subsequent levels (and vice versa). This is particular true for those candidates who score very well on the examinations (more than 2 SEMs above the cutscore). Similar results were found when examining the performance of unsuccessful candidates. Those candidates who scored closer to the cutscore were far more likely, in general, to pass upon retake than were candidates who scored lower.

Discussion and Conclusions

These studies address two fundamental precepts of testing: reliability and validity. By demonstrating (as was shown in Studies 1 and 2) that the item performance estimates are virtually equivalent across panels and within panels, the consistency of these estimates have been shown. Because the recommended MPS is based on the aggregation of item performance estimates, it is critical to demonstrate that they are consistent across replications. In Study 3, the central issue of validity of the panelists' estimates was considered. It was shown that, across years and levels of the examination program, the ratings provided by the standard setting panelists were in close agreement to performance of candidates who score close to the MPS values. Finally, Study 4 showed expected consistency in score patterns for successful and unsuccessful candidates as they progress through the examination program.

The high degree of reliability and validity found in the MPSs from this assessment program has not been found in some other investigations that also focused on the psychometric quality of results from an Angoff-based standard setting method. For example, Impara and Plake (1998) found that when teachers, who were very familiar both with their students and the examination, were asked to make item performance predictions for their "barely passing students", their estimates were not a close match to these students' actual performance on the test. However, the procedure used by Impara and Plake was to have the rating forms delivered to the teachers with only written instructions. The teachers did not meet as a group and did not benefit from a discussion of the characteristics of the barely passing students. The teachers received no training, nor did they get an opportunity to practice the item estimation process or to ask questions. Further, they did not estimate performance more than once and did not receive any student performance data. This latter point may not be as critical because the examination had been used in the school system for several years and the teachers were familiar with how their past students had performed on the test overall. However, it is unlikely that they had any knowledge of the item difficulty values (proportion correct, or p-values).

In the standard setting program used for this study, training took a central role. For each level in each of the years, the panelists participated in an in-depth training session lasting approximately 4 hours. Over one hour was devoted to a discussion designed to elicit from the panelists the knowledge, skills, and abilities of the JQC, focusing specifically on the components of the table of specification for the examination. At the completion of the discussion, panelists were given copies of these KSAs listed by component of the table of specifications. As part of the evaluation process, panelists were asked to rate the quality of the training in general, and the specific components, in particular. High rating were uniformly provided by the panelists on the "Discussion of the JQC", as well as the other training components.

Cizek, 1996, specifies that training is one of the features of a standard setting workshop that should be documented. Reid, 1991, states that there is little standardization in what constitutes appropriate levels of training for a standard setting study although he contends that training is an essential component of such a study. The results of this study suggest that training may be a key component to the psychometric quality of the resulting MPS.

Reports that the Angoff standard setting method is fundamentally flawed (NAE, 1993) may be overstated for programs such as the one we studied. The results of this study indicate, at least in this assessment program, MPSs based on item performance estimates (which is basically the data used in an Angoff standard setting method) are consistent, repeatable, congruent with actual candidate performance, and lead to reasonable conclusions about candidate future performance. These are all indicators of high levels of psychometric quality in the MPS values derived from the standard setting process.

Future research should focus on the generalizability of these results and the conditions that supported the high degree of technical quality of the results. Training has been highlighted as one possible link to the quality of these results. Other features, such as the content area, the use of analytical scoring, the quantitative nature of the discipline, and the relative location of the MPS to the central location of most of the scores in the score distribution all may have also contributed to the fact that the results from this standard setting program yielded MPS scores that showed excellent psychometric properties.

Footnotes

[1] In 1998, the was no standard setting workshop for Stage I as an equating strategy was used to put the MPS from the 1997 Stage I examination on the 1998 scale.

[2] The magnitude of these correlations is attenuated due to restriction of range. Because only scores from candidates who passed the Stage I examination were included, the full range of Stage I scores was not considered in the analyses. These correlations, then, should be considered as underestimates of the correlation of performance between scores on the Stage I and II examinations.

References

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

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

Impara, J. C. (1997, October). Setting Standards Using Angoff's Method: Does the Method Meet the Standard? Invited address to Division D of the Midwestern Educational Research Association, Chicago.

Impara, J.C., & Plake, B.S. (1998) Teachers' ability to estimate item difficulty: A test of the assumptions of the Angoff standard setting method. Journal of Educational Measurement, 35, 69-81.

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

National Academy of Education (1993). Setting performance standards for student achievement, Stanford, CA: Author.

Plake, B.S. (1998). Setting performance standards for professional licensure and certification. Applied Measurement in Education, 11, 65-80.

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

Shepard, L.A. (1995, 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.

Table 1. Descriptive Statistics for Examinations by Level and Year

	Stage I			Stage II			Stage III		
	1996	1997	1998[a]	1996	1997	1998	1996	1997	1998
N	14,381	16,832	21,737	7.098	8,493	10,295	3,121	5,316	6,651
Mean	144.35	146.37	152.99	206.79	211.97	230.89	199.12	200.96	194.81
SD	30.36	32.03	31.27	34.34	37.73	39.54	29.11	30.00	27.95
Rxx'[b]	0.9544	0.9574	0.9580	0.8043	0.8574	0.8329	0.7458	0.7672	0.6844
MPS_{panel}	138.46	148.54	NA	227.16	215.25	225.08	207.00	204.33	198.46
Sdpanel	13.21	6.70	NA	10.33	8.97	11.16	8.32	8.94	13.54
MPS_{Board}	146.00	148.00	152.00	200.00	208.00	225.00	190.00	197.00	192.00

[a] An equating strategy was used to determine the equivalent score on the 1998 Stage I examination scale to the MPS set by the Board in 1997.

[b] Internal consistency reliability; estimated by KR_{20} for Stage I; Coefficient alpha for Stages II and III.

Table 2. Inter-item consistency of panelists' item performance estimates on Stage I Examination, 1996 - 1997.

Item Location for Standard Setting		Panelists' Estimate			
1996	1997	1996	1997	Difference	Abs (Diff)
B-43	220	0.5776	0.5577	-0.0199	0.0199
B-45	222	0.5300	0.5362	0.0062	0.0062
B-49	228	0.4433	0.4100	-0.0333	0.0333
B-60	123	0.7840	0.7485	-0.0355	0.0355
B-7	180	0.6507	0.6400	-0.0107	0.0107
B-6	182	0.5000	0.4523	-0.0477	0.0477
B-3	189	0.5667	0.5446	-0.0221	0.0221
B-74	142	0.7180	0.6846	-0.0334	0.0334
A-49	227	0.5967	0.5269	-0.0698	0.0698
A-7	181	0.5067	0.4531	-0.0536	0.0536
A-29	197	0.5547	0.5938	0.0391	0.0391
A-75	139	0.6453	0.6192	-0.0261	0.0261
A-75	140	0.6933	0.7077	0.0144	0.0144
A-74	143	0.6200	0.6346	0.0146	0.0146
A-85	144	0.6007	0.6462	0.0455	0.0455
B-91	37	0.7060	0.7000	-0.0060	0.0060
B-99	41	0.6360	0.5917	-0.0443	0.0443
B-61	1	0.7887	0.7375	-0.0512	0.0512
B-30	80	0.4900	0.6167	0.1267	0.1267
B-106	114	0.7100	0.7567	0.0467	0.0467
B-73	19	0.7007	0.7708	0.0701	0.0701
A-92	48	0.5100	0.6362	0.1262	0.1262
A-103	116	0.5907	0.6383	0.0476	0.0476
A-79	26	0.6233	0.5375	-0.0858	0.0858
AVERAGE				-0.0001	0.0449
STANDARD DEVIATION				0.056	0.032

Table 3: Relationship between actual and anticipated performance on the Stage I, II, and III examinations

		Sample size	Average	Average
Stage	Year	EJOC	(actual – anticip)	|(actual - anticip)|
I	1996	2340	-0.010 (0.089)	0.074 (0.050)
I	1997	2937	-0.005 (0.093)	0.073 (0.057)
II	1996	2666	-1.180 (1.92)	1.780 (1.34)
II	1997	2815	-0.280 (1.83)	1.290 (1.30)
II	1998	3340	0.030 (1.34)	0.960 (0.92)
III	1996	1175	-1.400 (2.34)	1.440 (2.31)
III	1997	1956	-0.310 (1.37)	1.160 (0.77)
III	1998	3004	-0.260 (0.77)	0.640 (0.49)

Table 4. Percent of Candidates Passing Next Level Examination as a Function Performance Status on Previous Examination

Stage I – Stage II

		Within +1 SEM		Greater than +2 SEM	
Stage IStage II		n	% Pass	n	% Pass
96	97	930	32%	4197	76%
96	98	426	38%	965	68%
97	98	1275	33%	4625	83%

Stage II – Stage III

		Within +1 SEM		Greater than +2 SEM	
Stage II	Stage III	n	% Pass	n	% Pass
96	97	1403	41%	1597	80%
96	98	759	38%	313	68%
97	98	1462	41%	236	82%

Table 5. Percent of Repeat Candidates Passing Retake of Same Examination as a Function of Failing Performance on Previous Examination

Stage I – Stage I

		Within -1 SEM		More than -1 SEM	
Stage IStage In		% Pass	n	% Pass	
96	97	406	63%	1695	29%
96	98	108	73%	762	38%
97	98	586	84%	2474	44%

Stage II – Stage II

		Within -1 SEM		More than -1 SEM	
Stage II	Stage II	n	% Pass	n	% Pass
96	97	688	63%	782	31%
96	98	221	65%	431	43%
97	98	763	66%	1378	44%

Stage III – Stage III

		Within -1 SEM		More than -1 SEM	
Stage III	Stage III	n	% Pass	n	% Pass
96	97	391	74%	413	44%
96	98	108	56%	187	30%
97	98	650	63%	1032	37%

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

ERIC®

TM029717

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Validation of Angoff-based Predictions of Item Performance

Author(s): Barbara S. Plake, James C. Impara, Patrick Irwin

Corporate Source: (?) AERA presentation	Publication Date: 4/99

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: 135 Bancroft Hall, University of Nebraska Lincoln, NE 68588-0348	Telephone: 402-472-3280	FAX: 402-472-6207
	E-Mail Address: bplake@unl.edu	Date: 4/14/99

(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

ERIC Full Text Provided by ERIC