RESEARCH ARTICLES

Reliability and Credibility of Progress Test Criteria Developed by Alumni, Faculty, and Mixed Alumni-Faculty Judge Panels

H. Glenn Anderson Jr., PharmD, and Arthur A. Nelson, PhD

Texas Tech University Health Sciences Center, School of Pharmacy

Submitted April 29, 2011; accepted August 4, 2011; published December 15, 2011.

Objective. To compare the reliability and credibility of Angoff-based, absolute criteria derived by faculty, alumni, and a combination of alumni and faculty judge panels.

Methods. Independently, faculty, alumni, and mixed faculty-alumni judge panels developed pass/fail criteria for an 86-item test. Generalizability and decision studies were performed. Root mean square errors (RMSE) and 95% confidence intervals were calculated for reliability and credibility assessment. School graduate performance upon the North American Licensure Examination (NAPLEX) was the comparator for credibility assessment.

Results. RMSEs were 1.06%, 1.42%, and 2.32% for the alumni, faculty, and mixed judge panels respectively. The school's NAPLEX pass rate was 97.5%. This rate triangulated well with the faculty judge panel (pass rate = 93.9%, CI95% = 87.1% - 98.2%), but did not with either mixed judge or alumni judge panels.

Conclusions. Faculty-derived criteria offer superior pass/fail decision defensibility relative to both alumni derived and mixed faculty-alumni derived criteria.

INTRODUCTION

The Accreditation Council for Pharmacy Education (ACPE) standards require colleges and schools to assess student attainment of desired learning outcomes.[1] The student assessment program at the Texas Tech University Health Sciences Center's School of Pharmacy centers on delivery of an annual ability-linked assessment of student knowledge and skills. The assessment domains are based on the expected abilities of a recent pharmacy school graduate. The school requires all fourth-year students to pass the assessment prior to graduation. As such, the school's assessment is a progress test that uses a regional definition of pharmacy practice skills and abilities to determine student readiness to practice pharmacy. There are risks associated with progress tests. As a result of their performance, students are categorized as either passing and ready for program progression or failing and requiring remediation. These pass/fail decisions have the potential to delay or stop student advancement within a program.

Because of the potential for significantly impacting the lives of students, any pass/fail decision resulting from progress testing must be defensible. Establishing the procedural reliability of criterion development and credibility of the pass/fail decision is the cornerstone of claims for defensibility.[2] Process factors, specifically the procedures used for collecting expert judgments, may influence criterion credibility and reliability and thus, have received much attention.[3,4]

Because it is easily adapted to various assessment methods, the Angoff procedure has been extensively studied as a method for establishing absolute assessment criteria.[5,6] The procedure has 5 basic steps: selection of judges, defining "borderline" knowledge and skills, training the judges in use of the method, collecting judgments, and combining judgments to establish a passing score.[3,5,6] Content experts are generally believed to be the most appropriate judges for establishing absolute pass/fail criteria.[3,5] However, selection of judges to include within the procedure can be challenging.

Health profession curricula cover broad subject areas; however, instructors tend to focus on specific areas of expertise and instruction. When establishing criteria for a progress test, selecting judges from among the various instructors within a curriculum may result in overall group expertise, but with the majority of judges having little or no personal knowledge of curricular content beyond the individual courses they teach. For this reason, the composition of "best judges" for use with the Angoff procedure has been questioned.

Corresponding Author: H. Glenn Anderson Jr., PharmD, Associate Dean of Academic and Curricular Affairs, Professor of Pharmacy Practice, Marshall University School of Pharmacy, One John Marshall Drive, Huntington, WV 25755-2950. Tel: 304-696-2305. Fax: 304-696-7309. E-mail: glenn.anderson@marshall.edu

Verhoeven and colleagues argued that the individuals who are most knowledgeable regarding a curriculum's content are the graduates who have successfully completed the curriculum.[7,8] This position was supported by their finding that graduates were able to produce reliable criteria that provided credible pass/fail decisions.[7] Comparisons of graduate criteria to that of faculty experts (the item writers) showed graduate-derived criteria to be more credible (less likely to erroneously identify students as incompetent) than those derived by faculty experts.[8]

The studies by Verhoeven and colleagues suggest that judge panels comprised of program graduates improve reliability and credibility of criteria resulting from the Angoff procedure (ie, reasonable assessment outcomes). The effect on criteria development of using a mixed panel of item writers and graduates as item judges has not been explored previously, but such panels are thought to have the potential to further improve criteria reliability and credibility.

This study investigates the potential effect of using mixed panels of judges on the outcomes of the Angoff procedure. The objective of the study was to compare the reliability and credibility of progress test criteria developed by 3 separate groups of curricular content experts: program graduates, current faculty members, and a group of both faculty members and program graduates.

METHODS

The annual student progress assessment at the Texas Tech University Health Sciences Center School of Pharmacy, a test to determine student readiness to practice, includes both pen-and-paper and objective structured clinical examination subtests.[9,10] Each year, a table of specifications is developed to map the pen-and-paper portion of the assessment to a broad sample of curricular content by domain.

From 2006 to 2008, the pen-and-paper portion of the assessment was comprised of 222 items selected from a test bank written by faculty experts composed of biomedical scientists, pharmaceutical scientists, administrative and behavioral scientists, and practitioner educators. Experts for item writing were defined as individuals practicing, teaching, or performing research within a given curricular content area. All faculty item writers taught within the curriculum. The item sample consisted of 86 recurrent items taken from the 2007 and 2008 progress tests. Prior to being included on the progress test, each item had been tested and, if needed, revised to improve reliability and performance. All items were taken from 3 of the 4 domains assessed within the pen-and-paper portion of each progress test, including basic sciences, dispensing pharmaceuticals, and social and administrative sciences, but excluding pharmaceutical care. Prior to study initiation, the institutional review board granted exempt status for the study. The judges were either volunteers from the school of pharmacy faculty or alumni who had graduated within the past 8 years. Three panels of item judges were compared. The first panel was comprised of pharmacy faculty members who had not received a college degree from the school, including 5 faculty members from the department of pharmaceutical sciences and 5 from the department of pharmacy practice. The faculty judge panel rated the sampled items in October 2007 during criterion development for the 2008 progress test.

The second group was comprised of 6 alumni who graduated from the program between 2001 and 2008. Two alumni panel members were pharmacy faculty members, 3 were adjunct faculty members involved in preceptorship of third-year and fourth-year pharmacy students during experiential training, and 1 was a new graduate and ineligible for preceptor licensure at the time of this study. The alumni panel judged the items included within this analysis in June 2008.

The third panel of judges was comprised of 10 faculty members (equal representation from both school departments) and 3 alumni of the school. This mixed faculty-alumni panel judged items in October 2008. All criteria were estimated using a modified Angoff procedure based on item content and difficulty.[3,10] Judges were asked to imagine a group of 100 borderline students and estimate for each item the number of these examinees who would provide correct answers. Borderline students were defined as students with a 50% chance of passing the progress test. A borderline student was anticipated to spend an average amount of time studying, have knowledge just sufficient to pass the progress test, but frequently have difficulty scoring above 70% on individual course assessments. The 70% score represented the standard for course pass/fail decisions at the school and was familiar to all panel participants.

Judges were provided documents containing all items to be judged (stem, answer, and 3 distracters) and blanks for notation of item judgments. Judges were not provided historical item difficulties or the correct answers to the items reviewed. Judges were instructed not to apply a correction for guessing when rating items. Judgments rendered represented the probability that a borderline student would correctly answer each individual item and could assume a range of 0% to 100%.

Statistical Analysis

All analyses were performed using the SPSS 15.0 (SPSS Inc., Chicago, IL)[11] and GENOVA (The American College Testing Program, Iowa City, IA)[12] statistical packages. To assess how representative the sampled items were,

means and standard deviations for student performance were calculated for all items, the sampled items, and the items not sampled from both the 2007 and 2008 progress tests. Chronbach's alpha reliability coefficients were calculated for each item set with and without correction for item number reduction using the Spearman Brown prophecy formula.[13]

Classical test theory explains observed measurement as the combination of a true score (a measure of actual performance ability) and a single random source of error.[13,14] Examples of error commonly considered during application of classical test theory include occasion of assessment (test-retest reliability) and evaluator (inter-rater reliability). Though classical test theory is a familiar theory, its application is limited by the assumption of a single error source.

Generalizability theory (G-theory) is an alternative to classical test theory defined as a conceptual framework wherein the dependability of behavioral measurements can be considered.[15,16] G-theory is founded on the analysis of variance (ANOVA) statistical model. Because of ANOVA's ability to partition total variance, G-theory uses the ANOVA model to estimate the variance component associated with each source of variation that affects the measurement of interest.[15] Within G-theory, sources of variation are termed *facets* (similar to *factors* in ANOVA) with each facet having one or more *conditions* (comparable to *levels* in ANOVA).

G-theory allows for the development of models wherein the measure of interest (ie, object of measurement), one or more facets, and the interactions of each may be considered simultaneously.[15-17] Variance within the object of measurement can then be broken down into individual variance components for each facet and interaction. Variance components for each facet can then be scrutinized for individual contributions and evaluated to determine whether facet contribution can be expected to increase or decrease when combined with other facets.

Statistical analyses using G-theory are termed *generalizability studies* (G-studies). In a G-study, a researcher would obtain variance components for the object of measurement, for each study facet, and for each interaction. Variance components can be scrutinized for the purpose of explaining measurement outcomes or used to calculate either generalizability coefficients or root mean square errors (RMSE), both of which are indices of measurement reliability.[7,8,15,16]

These indices of measurement dependability are the focus for decision studies, wherein facet conditions are varied within a reasonable range in an attempt to find a point at which the index is maximized. Performance of a decision study is similar to repetitively asking the question, "What if the measurement conditions were changed in this way?"[15,16] The goal of performing a decision study is to identify the set of conditions that allows measurement efficiency to be maximized and measurement error minimized.

In the current study, G-theory was used to investigate criteria reliability.[15,18] A crossed item-by-judge design was used, with the analyses performed separately within each panel.[15] Variance components were estimated and used to calculate RMSE, an estimate of measurement reliability.[8,18,19] After generalizability studies had been completed, decision studies were performed to investigate the effect of varying facet conditions (items, judges) upon RMSE. During these studies, RMSEs were estimated when facet conditions were varied within a reasonable range of values.[8,18,19]

Angoff procedures were considered to be optimized when decision studies identified combinations of facet conditions that would allow attainment of an RMSE goal of 0.5% to 1.0%. This RMSE goal was selected after scrutinizing the 2007-2008 student performance on sampled items. Assuming an approximately normal distribution of student scores, a 1% shift in the criterion would result in a 1% change in failure rate. Using confidence intervals as an approximation of criteria precision, an RMSE of 1.0% or less would limit potential misclassifications of student failures to less than 5%. Criteria were identified as credible when pass/fail decisions triangulated with student performance on examinations assessing similar domains. The 2007-2008 graduate performance on the North American Pharmacy Licensure Examination (NAPLEX) was chosen as the study's credibility comparator.[20,21]

NAPLEX performance data were acquired from 2 sources. The pass rates of school of pharmacy graduate first-time test-takers were acquired from the National Association of Boards of Pharmacy aggregated data.[22] These data provided benchmarks for graduate competency. Disaggregated, individual graduate NAPLEX performance data were then acquired from the Texas State Board of Pharmacy under the Freedom of Information Act for the testing period of May 2007 through May 2009.

All students graduating from the school in 2007 and 2008 ($n = 81$ and $n = 82$, respectively) completed the required progress tests as P4 students prior to graduation. These students' responses to the 86 recurrent items found on the 2007 and 2008 progress tests formed the basis for credibility assessment. Expected passing rates were determined relative to the criteria derived from each of the judge panels (alumni, faculty, and mixed). Individual students were categorized as passing if, on the 86 recurrent

items, they achieved a score that was greater than or equal to the criterion being assessed. Students achieving scores lower than the derived criterion were categorized as having failed and not proving competency. To test pass/fail decision reasonability, the NAPLEX pass rate was compared to the pass rates for each criterion and to the pass rates for the upper and lower limits of each criterion's $CI_{95\%}$.

The RMSE is an estimate of the standard error of the mean (SEM) of Angoff measurements across items and judges,[7,18] which is analogous to the SEM used in the calculation of many common statistical procedures and in confidence intervals. As with the SEM, the RMSE can be used to calculate a confidence interval around a judge panel's criterion, thus identifying a range of values that would likely contain a repeated Angoff procedure criterion at a given level of confidence.

Criteria confidence intervals were calculated after estimating RMSEs for each judge panel. For this test, RMSEs were standardized to panel sizes of 10 judges developing criteria for an 86-item test. Criterion precision (confidence interval of 95% or $CI_{95\%}$) was used as an approximation of worst- and best-case scenarios for repeated criterion development procedures. To assess whether a judge panel criterion was reasonable, worst- and best-case criteria were used to establish an expected range of pass rates with use of each judge panel. The ranges of judge panel pass rate were then compared with the observed NAPLEX pass rate as the first test of credibility.

Although triangulation of failure rates was the primary method of establishing credibility, concerns regarding predictive accuracy of student decisions still existed. To investigate the predictive accuracy of criteria use, student-specific pass/fail decisions arising from use of each criterion were compared to those obtained from NAPLEX performance. Criterion hit rates were calculated after preparation of 2x2 tables.[13,23] The hit rate of the faculty judge panel was considered the base rate for these analyses, as the school's standard operating procedure has been to use faculty members for derivation of all progress test criteria.

RESULTS

Table 1 summarizes the pharmacy students' performance on the overall 2007 and 2008 progress tests, the sampled items, and the nonsampled items. Means, standard deviations, and internal consistency for the sampled items and nonsampled items were comparable. Internal consistency of the standardized progress test ranged from 0.68 to 0.82 (Chronbach's alpha) and was highest for the sampled items. When considering 2007 and 2008 progress tests in combination, scores on the sampled items strongly correlated with overall scores on the progress test ($r = 0.87$, $p < 0.0005$) and moderately with scores on non-sampled items ($r = 0.62$, $p < 0.0005$). In the generalizability study, item judgment rates were similar for the 3 judge panels, with 96.5%, 89.1%, and 92.4% of judgments returned for the alumni, faculty, and mixed judge panels, respectively. Table 2 provides a summary of individual panel member judgments.

Results of the Generalizability Study are summarized in Table 3. Across panels, 46.3% to 66.0% of all variance can be attributed to variance between items or item difficulty. The large degree of variance attributed to items suggests that the progress test includes items with a moderately wide range of difficulty. Although the judge facet contributes only a small amount to overall progress test variance, the mixed judge panel does contain the largest source of judge variance (17.3% versus 7.4% [faculty] and 3.2% [alumni]). The error variance (ij, e) accounts for a moderate amount of the overall variance (range, 30.8% to 45.7%) and may indicate existence of either some degree of item-judge interaction or a systematic unexplained error; however, because of the large item sample size, this source of variability contributes only minimally to the computation of RMSE.[8,19]

Table 1. Mean Progress Test Scores of Fourth-Year Pharmacy Students and Reliabilities of Total Progress Test, Items Not Sampled and Items Sampled for the Angoff Procedure

	Year	Fourth-Year Pharmacy Students, No.	Questions, No.	Correct Score, Mean % (SD)	Bivariate Correlations (r)	Reliability[a]	Standardized Reliability[a,b]
Sampled items	Pooled	163	86	66.4 (6.9)	1.0	0.65	0.82
Total performance test[c]	2007	81	222	65.8 (4.9)	0.84	0.70	0.70
	2008	82	222	70.1 (5.9)	0.90	0.81	0.81
Items not sampled[c]	2007	81	133	66.6 (5.2)	0.53	0.55	0.68
	2008	82	133	71.7 (5.7)	0.68	0.67	0.78

[a] Cronbach's alpha

[b] Estimated using the Spearman Brown prophecy formula; standardized toward 222 items

[c] Not pooled because of dissimilarity of nonsampled items used within 2007 and 2008 progress tests.

Table 2. Group Membership, Demographic Characteristics, and Mean Angoff Estimates for Panel Judges[a,b]

Judge	Judge Panel	Faculty Member	Pharmacist	Prior Angoff Experience	Mean (SD) (%)
1	Faculty	Yes	No	Yes	58.7 (11.7)
2	Faculty	Yes	Yes	Yes	46.9 (17.7)
3	Faculty	Yes	Yes	Yes	48.6 (15.3)
4	Faculty	Yes	Yes	No	49.2 (15.4)
5	Faculty	Yes	Yes	Yes	48.9 (19.4)
6	Faculty	Yes	Yes	Yes	44.5 (17.1)
7	Faculty	Yes	No	No	47.3 (15.9)
8	Faculty	Yes	No	No	44.5 (20.6)
9	Faculty	Yes	No	No	47.9 (7.5)
10	Faculty	Yes	No	No	49.0 (11.9)
11	Alumni	No	Yes	No	37.8 (19.1)
12	Alumni	No	Yes	No	39.1 (16.6)
13	Alumni	No	Yes	No	37.2 (15.2)
14	Alumni	No	Yes	No	41.6 (16.2)
15	Alumni	Yes	Yes	No	42.5 (17.9)
16	Alumni	Yes	Yes	No	44.1 (19.6)
17	Mixed[c]	No	Yes	No	50.6 (17.3)
18	Mixed	Yes	Yes	Yes	48.6 (13.3)
19	Mixed	Yes	Yes	Yes	53.2 (19.0)
20	Mixed	Yes	No	No	59.8 (14.3)
21	Mixed	Yes	No	Yes	61.3 (14.1)
22	Mixed	Yes	No	No	58.9 (13.7)
23	Mixed	Yes	Yes	No	56.7 (18.6)
24	Mixed	Yes	No	No	47.9 (13.60
25	Mixed	Yes	No	Yes	63.8 (14.5)
26	Mixed	No	Yes	No	48.3 (16.8)
27	Mixed	Yes	Yes	No	63.4 (13.6)
28	Mixed	Yes	Yes	No	56.8 (18.9)
29	Mixed	No	Yes	No	46.1 (18.4)

[a] Program graduates are indicated by "No" in the Faculty Member column and a "Yes" in the Pharmacist column

[b] Pharmaceutical and Biological Scientists are indicated by a "Yes" in the Faculty Member column with a "No" in the Pharmacist column

[c] Faculty and alumni mixed-judge panel.

After standardizing panel size to 10 judges, RMSEs for alumni, faculty, and mixed judge panels were 1.06, 1.42, and 2.32, respectively, for the 86 sampled items. Observed RMSE differences can be attributed directly to the relative sizes of the judge variance components. Tables 4, 5, and 6 summarize the results of the decision study. RMSE is displayed as a function of number of items comprising the assessment and the judge panel size. As

Table 3. Analysis of Variance and Estimated Variance Components

Judge Panel	Source of Variability	Sum of Squares	Mean Squares	Estimated Variance Component	Percent of Total Variance
Faculty	Items (i)	114472.48	1300.82	118.51	46.8
	Judges (j)	16118.04	1790.89	18.82	7.4
	Error (ij, e)	91666.36	115.74	115.74	45.7
Graduates	Items (i)	119547.58	1358.50	210.05	66.0
	Judges (j)	5009.87	1001.97	10.15	3.2
	Error (ij, e)	43209.29	98.20	98.20	30.8
Mixed faculty/ graduates	Items (i)	170464.98	1937.10	140.51	46.3
	Judges (j)	57431.11	4785.93	52.53	17.3
	Error (ij, e)	116613.19	110.43	110.43	36.4

Table 4. Root Mean Square Error (RMSE) as a Function of the Number of Judges and the Number of Items for the Faculty and Alumni Mixed Panel

	Number of Judges[b]								
Number of Items	**3**	**5**	**10**	**15**	**20**	**30**	**40**	**50**	**60**
50[a]	4.27	3.31	2.34	1.91	1.65	1.35	1.17	1.05	0.96
100	4.23	3.28	2.32	1.89	1.64	1.34	1.16	1.04	0.95
150	4.21	3.26	2.31	1.88	1.63	1.33	1.15	1.03	0.94
200	4.21	3.26	2.30	1.88	1.63	1.33	1.15	1.03	0.94
225	4.20	3.26	2.30	1.88	1.63	1.33	1.15	1.03	0.94
250	4.20	3.26	2.30	1.88	1.63	1.33	1.15	1.03	0.94
300	4.20	3.25	2.30	1.88	1.63	1.33	1.15	1.03	0.94

[a] RMSE is expressed as percent correct score.
[b] Faculty and Alumni Mixed judge panels are comprised of a 10:3 ratio of faculty:alumni.

expected, RMSE decreases and criterion precision increases with both progress test length and judge panel size; however, changes in judge panel size produced larger RMSE gains.

Tables 4, 5, and 6 identify ratios of panel size to progress test length that would allow achievement of goal RMSEs. The mixed judge panel could attain RMSEs nearing the 0.5% to 1.0% range only when establishing criteria on 50 or more item tests with at least 60 judges. In contrast, the faculty judge panel reached desirable levels of precision with assessments containing 150 or more items and panels of 15 to 20 judges. The precision of the alumni judge panel was greater than the faculty judge panel, attaining desirable precision levels when establishing criteria for assessments of 50 or more items using panels of 10 to 15 individuals.

The school's 2007-2008 mean NAPLEX pass rate was 97.5%. The 3 panels of judges derived criteria of 47.7% (alumni), 57.0% (faculty), and 64.0% (mixed judge panel). Using criteria derived from the alumni, faculty, and mixed judge panels, pass rates would be 100.0%, 93.9%, and 71.8%, respectively. Figure 1 displays the influence of criterion precision on resulting pass/fail conclusions. Use of alumni judge panel criterion would result in stable student outcomes ($CI_{95\%}$ = 46.5% - 50.0%, pass rate = 100.0% to 100.0%). Increasing instability of the pass/fail decision would be expected if either the faculty judge panel ($CI_{95\%}$ = 54.7% to 60.5%, pass rate = 87.1% to 98.2%) or the mixed judge panel ($CI_{95\%}$ = 60.5% to 68.6%, pass rate = 46.6% to 87.1%) were used for criteria development. However, of the 3 panel-derived criteria, only the faculty judge panel criterion resulted in student outcomes that triangulated with the school's 2007-2008 NAPLEX pass rate.

NAPLEX scores were acquired from the Texas State Board of Pharmacy for the 141 (86.5%) 2007/2008 P4 students who underwent examination in the state of Texas. The observed predictive accuracy, or hit rate, between NAPLEX performance and use of each panel-derived criterion is summarized in Table 7. The faculty judge panel, the base rate for these analyses, had a hit rate of 94.3%, with 5.0% of participants expected to be identified as failing the progress test although they had passed the NAPLEX (false positives). What constitutes a "good" hit rate is subjective, but improvement on base rates is a reasonable goal whenever procedural changes are being considered.[23] The mixed judge panel failed to achieve base rate levels of predictive accuracy (hit rate = 73.8%) or to improve on the base misclassification rate (26.2% false positives). The alumni judge panel hit rate was

Table 5. Root Mean Square Error (RMSE) as a Function of the Number of Judges and the Number of Items for the Faculty Panel

	Number of Judges								
Number of Items	**3**	**5**	**10**	**15**	**20**	**30**	**40**	**50**	**60**
50[a]	2.65	2.06	1.45	1.19	1.03	0.84	0.73	0.65	0.59
100	2.58	2.00	1.41	1.15	1.00	0.82	0.71	0.63	0.58
150	2.56	1.98	1.40	1.14	0.99	0.81	0.70	0.63	0.57
200	2.54	1.97	1.39	1.14	0.98	0.80	0.70	0.62	0.57
225	2.54	1.97	1.39	1.14	0.98	0.80	0.70	0.62	0.57
250	2.54	1.96	1.39	1.13	0.98	0.80	0.69	0.62	0.57
300	2.53	1.96	1.39	1.13	0.98	0.80	0.69	0.62	0.57

[a] RMSE is expressed as percent correct score.

Table 6. Root Mean Square Error (RMSE) as a Function of the Number of Judges and the Number of Items for the Alumni Panel

	Number of Judges								
Number of Items	**3**	**5**	**10**	**15**	**20**	**30**	**40**	**50**	**60**
50 RMSE%[a]	2.01	1.56	1.10	0.90	0.78	0.64	0.55	0.49	0.45
100	1.93	1.49	1.06	0.86	0.75	0.61	0.53	0.47	0.43
150	1.90	1.47	1.04	0.85	0.74	0.60	0.52	0.46	0.42
200	1.88	1.46	1.03	0.84	0.73	0.60	0.52	0.46	0.42
225	1.88	1.46	1.03	0.84	0.73	0.59	0.51	0.46	0.42
250	1.88	1.45	1.03	0.84	0.73	0.59	0.51	0.46	0.42
300	1.87	1.45	1.02	0.84	0.72	0.59	0.51	0.46	0.42

[a] RMSE is expressed as percent correct score

97.9%, exceeding the base rate and resulting in 0.0% false positives.

DISCUSSION

As assessments of student competency or readiness for curricular progression, progress tests are significant sources of student stress. Delays in program progression, unanticipated financial burdens, social stigmatization, or loss of career are all possible outcomes of applying progress tests. Thus, progress test decisions must be justifiable to all stakeholders. Increasing the defensibility of progress test decisions requires substantial time and effort. How much time must be committed to this endeavor is difficult to forecast, but a reasonable rule-of-thumb is to increase the rigor of the assessment development process as the severity of assessment consequences increases. Assessment defensibility rests with development of valid assessments that return reliable and credible pass/fail decisions. This study focused on expert judge selection during the criterion development process and how judge selection can affect defensibility on the basis of reliability, credibility, or both.

Prior research suggests that using item writers as judges may not produce criteria that are as reliable as those produced by recent program graduates.[8] In the current study, this conclusion is supported by the RMSE for the alumni judge panel being smaller than the RMSE for the faculty judge panel. By progressing through a curriculum course by course, program graduates, have been hypothesized to have a more global, homogeneous view of the overall curriculum than that of item writers.[7] The current study suggests an expansion of this postulate to faculty members whose experience and expertise are often in focused areas of a pharmacy curriculum.

As both the faculty and alumni curricular viewpoints may have limitations, there may be opportunities for further improvements in criterion reliability with panels comprised of both item writers and alumni.[8] Unfortunately, the results of the current study failed to support this hypothesis, as evidenced by a significantly less-reliable criterion being derived by the mixed judge panel compared with criteria derived by either the faculty or alumni judge panel. Explaining why this may have occurred requires a deeper exploration of the Angoff procedure.

Discussion among panel members is a key component of the Angoff procedure. These discussions center around the highest and lowest judgments rendered. When group variance exceeds a prescribed level, panel members rendering those judgments are required to provide a brief synopsis of the reasoning behind their judgment. As such, the judges have a significant opportunity to influence their peers prior to the rendering of final judgments on the item being considered.

A possible explanation for the mixed judge panel's large judge facet variance and subsequent reliability problems is group-induced polarization.[24] This phenomenon occurs when 2 groups favoring opposite sides of an issue engage in discussion. During discussion, the opinions of each groups' members migrate to a more extreme position than originally held. Such outcomes arise more frequently with subjective decisions, as with judgments made during

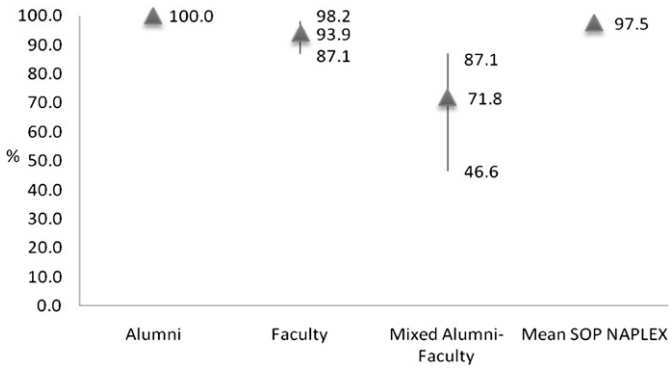

Figure 1. Effect of criterion precision on expected progress test pass rates relative to mean School of Pharmacy (SOP) North American Pharmacy Licensure Examination (NAPLEX) pass rate.

Table 7. NAPLEX vs. Progress Test, Hit Rates by Criteriona

Judge Panel	NAPLEX Performance	Expected Progress Test Performance		Hit Rate (%)[b]
		Failed No.	Passed, No.	
Alumni	Failed	0	3	
	Passed	0	138	97.9
Faculty	Failed	2	1	
	Passed	7	131	94.3
Mixed	Failed	3	0	
	Passed	37	101	73.8

[a] Sample includes 141 P4 students taking NAPLEX in Texas.
[b] Hit Rate = percent of predictions correct = 100 * [(failed and failed) + (passed and passed)]/total.[13]

criterion development, and are more prevalent when discussion exposure is to extremes in opinion rather than the overall distribution of opinions. Our decision to limit discussion to only extreme differences in judgments may have allowed group-induced polarization to occur during development of the mixed judge panel criterion.

This phenomenon could have been reduced or avoided by providing judges with a realistic starting point for their judgments. Providing judges with past item difficulties (item *p* values) would have established a realistic starting point for per-item performance of the overall student body and may have facilitated estimation of borderline student performance.[19,25,26] Revision routinely occurs after items are tested. We chose not to provide item difficulty levels to judges because revision of an item has the potential to change item difficulty, thus rendering past performance estimates invalid.

Criterion reliability may also have been affected by judge panel demographics and mixed judge panel composition. Unfortunately, the current study did not investigate the effects of varying the ratio of faculty members to alumni in the mixed judge panel or the effects of judge demographics. Future investigation into the influences that these factors have upon criterion reliability would provide a clearer picture of the potential benefits of using mixed judge panels.

One method for providing evidence of test credibility is to establish the comparability of outcomes arising from progress tests with similar, validated assessments.[2,27,28] The school's progress test is similar to the NAPLEX both in terms of purpose and in the domains assessed. Both assessments attempt to determine graduate or near-graduate readiness to practice. To evaluate practice readiness, both assessments use ACPE standards as a reference source.[20] In theory, 2 assessments that measure the same behavior, skill, or ability should return similar absolute decisions – in this case, competent or not competent to practice pharmacy.

The student NAPLEX outcomes provided a benchmark for establishing progress test criteria reasonableness. The pass rate obtained by means of the mixed judge panel's criterion triangulated poorly with the NAPLEX pass rate, providing evidence that the mixed judge panel criterion was ill-conceived and likely underestimated student competency. Conversely, use of the alumni judge criterion produced a pass rate that appeared to overestimate student performance and would be unable to identify students who had not acquired competency. The faculty panel was the only group to produce a criterion that resulted in a pass rate that triangulated closely with the NAPLEX pass rate and, thus, could be considered credible.

Achievement of reasonable failure rates provides only part of the evidence required to label a criterion as credible. The criteria used to formulate progress test decisions should also provide valid interpretations of student ability. Calculated hit rates allowed for the assessment of potential competency misclassifications with use of each criterion.

Incorrectly identifying students as progress test failures (false positives) was interpreted as being the least desirable misclassification because of the potential for a student's graduation to be delayed for remediation and reassessment. In comparing hit rates for the 3 panels, we believe that use of the criterion of the mixed judge panel would result in unreasonably high false-positive rates that could have significant, inappropriate student impact. Use of both the faculty and alumni judge criteria would result in reasonable false positive and hit rates.

Identification of a defensible criterion with the desirable characteristics of stability and credibility is the ultimate goal of this study. The criterion of the mixed judge panel failed to achieve either characteristic, and the criterion of the alumni judge panel had desirable stability but poor credibility. Only the criterion of the faculty judge panel met or exceeded both desirable characteristics, rendering it defensible.

This finding differs from the conclusion of prior research that, compared with item writers, alumni produce criteria that are more credible.[8] Differences in Angoff procedure modifications may explain this divergence. Specifically, we chose not to provide judges with the correct answers to the items being evaluated. This decision may have led the judges to rate items as difficult when they themselves did not know the correct answers. Because correct answers are often the subject of group deliberations, the faculty judge panel, which is comprised of item writers, may have been less subject to item-answer uncertainty than were members of the alumni panel.

Along with the decision not to provide judges with past item difficulty data, this decision may have resulted in alumni judges rating borderline student performance on some items inadvertently low. This judging behavior could be one explanation for why use of the alumni judge panel's criterion resulted in passing rates that were significantly higher than the NAPLEX benchmark.

CONCLUSION

Judge selection within Angoff procedures can have significant influence on both criteria stability and student pass rates. Therefore, identifying the best judges for standard setting is paramount to successful implementation of a progress test. The findings of this study suggest that both alumni and mixed faculty-alumni judge panels had difficulty producing credible student outcomes. However, reasonably sized faculty judge panels were able to produce criteria with a balance of reliability and credibility. As such, faculty judge panels should be preferred when establishing progress test criteria.

ACKNOWLEDGMENT

The investigators acknowledge the faculty members and alumni who donated their time and energy to the criterion development processes described herein. The author acknowledges and thanks Summer Balcer, MEd, and Ron G. Hall II, PharmD, MSCS, BCPS, for their review and commentary on earlier versions of this work.

REFERENCES

1. Accreditation Council for Pharmacy Education. Accreditation standards and guidelines for the professional program in pharmacy leading to the Doctor of Pharmacy degree. *Accreditation Council for Pharmacy Education*. http://www.acpe-accredit.org/pdf/FinalS2007Guidelines2.0.pdf. Accessed October 1, 2011.
2. Cizek GJ, Bunch MB. *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests*. Thousand Oaks, CA: Sage Publications, Inc; 2007.
3. Downing SM, Tekian A, Yudkowsky R. Procedures for establishing defensible absolute passing scores on performance examinations in health professions education. *Teach Learn Med*. 2006;18(1):50-57.
4. Norcini JJ, Shea JA. The credibility and comparability of standards. *Appl Meas Educ*. 1997;10(1):39-59.
5. Livingston SA, Ziedy MJ. *Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests*. Princeton, NJ: Educational Testing Service; 1982.
6. Cizek GJ. *Setting Performance Standards: Concepts, Methods, and Perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc; 2001.
7. Verhoeven BH, van der Steeg AF, Scherpbier AJ, Muijtjens AM, Verwijnen GM, van der Vleuten CP. Reliability and credibility of an angoff standard setting procedure in progress testing using recent graduates as judges. *Med Educ*. Nov 1999;33(11):832-837.
8. Verhoeven BH, Verwijnen GM, Muijtjens AM, Scherpbier AJ, van der Vleuten CP. Panel expertise for an Angoff standard setting procedure in progress testing: item writers compared to recently graduated students. *Med Educ*. Sep 2002;36(9):860-867.
9. Mehvar R, Supernaw RB. Outcome assessment in a PharmD program: The Texas Tech experience. *Am J Pharm Educ*. 2002; 66(3):219-223.
10. Supernaw RB, Mehvar R. Methodology for the assessment of competence and definition of deficiencies of students in all levels of the curriculum. *Am J Pharm Educ*. 2002;66(1):1-4.
11. SPSS Inc. *SPSS 15.0 for Windows, Release 15.0.0*. Chicago, IL: SPSS, Inc; 2006.
12. Crick JE, Brennan RL. *A general purpose analysis of variance system [computer program], Version 2.1*. Iowa City, IA: The American College Testing Program; 1983.
13. Crocker L, Algina J. *Introduction to Classical & Modern Test Theory*. Mason, OH: Cengage Learning; 2008.
14. Brennan RL. Generalizability theory and classical test theory. *App Meas Educ*. 2011;24:1-21.
15. Shavelson RJ, Webb NM. *Generalizability Theory: A Primer*. Newbury Park, Ca: Sage Publications, Inc.; 1991.
16. Naizer G. *Basic Concepts in Generalizability Theory: A More Powerful Approach to Evaluating Reliability*. Department of Education; 1992.
17. Di Nocera F, Ferlazzo F, Borghi V. G Theory and the reliability of psychophysiological measures: a tutorial. *Psychophysiology*. 2001;38(5):796-806.
18. Brennan RL. *Elements of Generalizability Theory*. Iowa City, IA: ACT Publications; 1983.
19. Fowell SL, Fewtrell R, Mclaughlin PJ. Estimating the minimum number of judges required for test-centered standard setting on written assessments. Do discussion and iteration have an influence? *Adv Health Sci Educ Theory Pract*. 2008;13(1):11-24.
20. Newton DW, Boyle M, Catizone CA. The NAPLEX: evolution, purpose, scope, and educational implications. *Am J Pharm Educ*. 2008;72(2):Article 33.
21. National Association of Boards of Pharmacy. NAPLEX Blueprint. http://www.nabp.net/programs/examination/naplex/naplex-blueprint/. Accessed October 1, 2011.
22. National Association of Boards of Pharmacy. Statistical Analysis of NAPLEX® Passing Rates for First-time Candidates per Pharmacy School from 2006 to 2010. http://www.nabp.net/programs/assets/NAPLEXpassrates.pdf. Accessed October 1, 2011.
23. Allen MJ, Yen WM, eds. *Introduction to Measurement Theory*. Long Grove, IL: Waveland Press, Inc.; 2002.
24. Fitzpatrick AR. Social influences in standard setting: the effects of social interaction on group judgments. *Rev Educ Res*. 1989; 59(3):315-28.
25. Norcini JJ. Setting standards on educational tests. *Med Educ*. 2003;37(5):464-469.
26. Morrison H, McNally H, Wylie C, McFaul P, Thompson W. The passing score in the objective structured clinical examination. *Med Educ*. 1996;30(5):345-348.
27. American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association; 2004.
28. Downing SM. Validity: on meaningful interpretation of assessment data. *Med Educ*. Sep 2003;37(9):830-837.