American Journal of Pharmaceutical Education 2013; 77 (10) Article 211.

RESEARCH

Reliability of a Minimal Competency Score for an Annual Skills Mastery Assessment

Gregory L. Alston, PharmD, and Wesley R. Haltom, PharmD

School of Pharmacy, Wingate University, Wingate, North Carolina

Submitted April 12, 2013; accepted June 12, 2013; published December 16, 2013.

Objective. To determine whether the modified Angoff process can be used to calculate a reliable minimal competency ("cut") score for the Annual Skills Mastery Assessment (ASMA).

Methods. Three panels of pharmacy faculty members used a modified Angoff method to create a minimal competency score for 60 previously used test items. The panels did not know which items were included. Data were analyzed to determine differences between rating sessions, faculty type, item difficulty, and rater scoring bias.

Results. The cut score generated was not significantly different by session or faculty type. The range of cut scores varied by less than 3% per examination. Faculty panelists correctly predicted student performance on items grouped as easy, medium, and hard.

Conclusion. A properly constructed faculty panel can determine a reliable cut score and accurately rank relative test item difficulty using the modified Angoff process.

Keywords: assessment, progress examination, competency, Angoff method, modified Angoff method, mile marker examination

INTRODUCTION

Many colleges and schools of pharmacy have developed mile marker or progress examinations in response to the Accreditation Council for Pharmacy Education Standards described in Guideline 15.1,[1] which states that the evaluation of student learning should "incorporate periodic, psychometrically sound, comprehensive, knowledge-based, and performance-based formative and summative assessments, including nationally standardized assessments (in addition to graduates' performance on licensure examinations) that allow comparisons and benchmarks with all accredited and college or school-determined peer institutions."[2] Several different methodologies for conducting a psychometrically sound, comprehensive progress assessment have been developed.[3-6] While some may consider it a problem that there is no universally accepted method for performing a progress examination, others would argue that the assessments should be tailored to the curriculum, culture, and resources available to the individual school, and this may necessitate the use of a variety of techniques.

The American Educational Research Association (AERA) developed a policy based on the 1999 Standards for Educational and Psychological Testing that are a professional consensus concerning sound and appropriate test use in education and psychology.[7,8] While each of the AERA's 10 points is a required element of a sound testing program, this study focuses on the issue of setting a defensible passing (cut) score. Downing described a procedure for establishing defensible absolute passing scores on high-stakes examinations in the health professions.[9] All such cut scores reflect the subjective opinions of experts. Five different methods were described: Angoff, Ebel, Hostee, Borderline Group, and Contrasting Group, each of which produced a slightly different cut score and there is no universally preferred method. The key to creating a defensible cut score lies in the choice of credible judges and the use of a sound systematic process of collecting and analyzing their judgments about examination components. One such credible method is the Angoff method.[10]

The process followed in practice is the modified Angoff method. "A slight variation to this procedure is to ask each judge to state the probability that the 'minimally acceptable person' would answer each item correctly. In effect, the judges would think of several minimally acceptable persons, instead of only 1 such person, and would estimate the proportion of minimally acceptable persons who would answer each item correctly." People who fail high-stakes examinations are the ones most likely to file lawsuits. A typical legal strategy is to challenge the fairness of the cut score as a violation of the Federal Civil Rights Act,[11] which places these cases in federal court. All federal courts consider Supreme Court decisions as precedent. It is

Corresponding Author: Gregory L. Alston, PharmD, Wingate University, Wingate, NC. E-mail: galston@wingate.edu

the application of the modified Angoff method to high-stakes examinations that has received acceptance before the United States Supreme Court in employment law court cases.[11] The Angoff method complies with the federal Uniform Guidelines on Employee Selection Procedures which speak to cutoff scores used in developing a legally compliant preemployment testing program.

Biddle suggested the following recommendations to create a reliable examination cut score:[12] (1) use at least 7 to 10 subject matter experts as judges; (2) ask each judge to state the probability for each test item that the minimally acceptable person would answer the item correctly; (3) sum the judges' estimates for each test item, average the score per test item, and then sum the averages for the items on the test to create the cut score; (4) calculate the reliability and standard deviation for the test scores after the test is administered; and (5) consider the standard error of measurement before setting the final cut score.

This study reviewed the ability of the Angoff process used by the Wingate University School of Pharmacy (WUSOP), to reliably produce a cut score for the Annual Skills Mastery Assessment (ASMA). This examination is a key element in the WUSOP Assessment Plan and should be able to withstand legal scrutiny by complying with AERA standards. The only AERA criterion we analyzed was the reliability of the cut score. Future studies will be required to address the validity of the cut score.

Because the ASMA examination purports to assess minimal competency in a variety of terminal ability areas, the process for development of the cut score must be sound. The process developed at WUSOP was designed to adhere to the AERA guidelines and practices. We hypothesized that if the process was reliable, then the cut score created by each independent faculty member panel should be the same and the faculty should be able to predict actual test item difficulty accurately. A secondary objective of this study was to determine potential risks to panel accuracy. This was done by comparing Angoff scoring results by clinical vs nonclinical faculty members and by rater scoring bias. Rater scoring bias was defined as the propensity to be either a below-average or above-average rater. The ASMA examination program is administered in 4 unique versions every year to the first-year (P1), second-year (P2), third-year (P3), and fourth-year pharmacy (P4) students. While we focused on the process for generating the P4 examination cut score, the process used for all other levels of the examination was identical.

METHODS

The ASMA examination assesses student performance on the mastery of terminal ability outcomes as they progress through the curriculum at WUSOP. Since the publication of the original methodology article,[13] the ASMA examination has been refined and lengthened to improve reliability. The essential elements of the ASMA examination remain as described in 2009, but a brief synopsis of the design of the program is included here to provide context.

At WUSOP the curriculum was designed around 195 terminal ability outcomes (TABOs). Each of the TABOs was tagged to the instruction year as a P1, P2, P3, or P4 ability. A sampling of the TABOs was included in the ASMA examination, and a unique multiple-choice, single-best-answer, 4-option examination was created for each class. The test-question bank coded each item to reflect the TABO it intended to test. The Angoff assigned cut score was recorded in a user field in the database. In 2012, the ASMA included 102 test items and 7 TABOs for the P1 students, 124 test items and 12 TABOs for the P2 students, 151 test items and 15 TABOs for the P3 students, and 187 test items and 17 TABOs for the P4 students. All students took the examination on the last Wednesday in March.

Each student received a report with a detailed analysis of their scores at their year-end faculty advisor appointment. This score report compared student performance to the faculty-determined competence level on the total examination and for each TABO. The test items were intentionally written at a level to reflect the ability to be observed rather than simple recall of facts. Students were not told what would be on the examination and were not encouraged to study so that the performance on this examination was reflective of their retention and not their short-term memory skills.

A critical component of this examination was setting the appropriate cut score for the total examination and for each subscore. The WUSOP Angoff process computed a 2-digit score for each test item. This score represented the panel estimation of the percentage of minimally competent students who would answer the item correctly. The Angoff score for each item selected for inclusion in the examination was then used to calculate the Angoff cut score for the entire examination and for each subscore. Testing software generated a grading table that allowed for customized reporting by subscore.

All faculty members with doctor of pharmacy (PharmD) degrees were invited to participate in the test review sessions and voluntarily attended the sessions. These annual sessions were conducted exactly as they had been in previous years. Each session was scheduled for 2 hours to rate between 100 and 120 test items. Faculty members used the Angoff method for scoring the test items and conducted quality control of the items for accuracy. To avoid introducing a different level of attention

to the process than in previous years, panelists were not alerted that data were being collected for this study during the sessions. Session 1 and 2 panels reviewed 102 identical test items per session, 60 of which already had been scored by faculty panels in previous years. The sessions convened with 10 faculty panelist members in the computer laboratory on campus. A proprietary software developed at WUSOP was used to conduct the session. The test items to be reviewed were loaded into the software and appeared on each faculty member's computer terminal 1 question at a time.

The assistant dean for assessment moderated the sessions to provide consistent administration of the process. The session began with a brief description of the process used at WUSOP. Panelists were advised to consider a room full of 100 minimally competent students. Because the minimal competency expectation rating depended on the level of the student being considered, the panelists were instructed to consider the competency of P4 students at the completion of the curriculum as their reference point. The only results included from the historical panels were P4 ratings to provide comparability between years. Experience running multiple Angoff sessions targeted at 4 different levels (P1through P4) suggested that faculty members were most aligned when rating P4 students. After reading each question, the panelists were instructed to predict how many of the 100 minimally competent students would answer the test item correctly. The panelists then entered their prediction into the software. The moderator screen allowed the moderator to view the panelist ratings per item. More importantly, each faculty member panelist could see the other panelists' predictions once they entered in their own prediction. Panelists were free to discuss their reasoning and change their score prior to finalizing the system-generated cut score, but in practice, less than 5% of the item ratings were altered after the initial panelist prediction was entered. The moderator then manually advanced the system to the next test item. The software automatically dropped the highest and lowest prediction to calculate the average of the remaining 8 scores. It computed a system-generated cut score for each test item for each session.

Sixty identical test items were rated by 3 separate, 10-member faculty panels; 10 clinical faculty members comprised the first 2010 session, 7 nonclinical faculty members and 3 clinical faculty members comprised the second 2010 session, and a mix of clinical and nonclinical faculty members comprised a historical panel session using data from 2008 to 2010 retrieved from the test bank software. Nonpharmacist faculty members were excluded from sessions 1 and 2 but were included in the historical panel. Some members of sessions 1 or 2 may have participated in the historic panels but the data were not kept until 2011; thus, the effect cannot be determined with certainty. Sessions 1 and 2 included 42 new test items for which historical data were not available. These panels rated a total of 102 identical test items, only 60 of which had been rated by historical panels. The previously defined modified Angoff process was followed for all sessions and moderated by the assistant dean for assessment.

The raw data for sessions 1 and 2 were further reviewed. The data from all 20 panelists were combined to create a data sheet for analysis using SPSS, version 15, software (SPSS Inc, Chicago, IL). The individual test item predictions made by each panelist were aligned in a single column on the data sheet. In the adjacent columns new values were entered to identify whether the prediction was made during session 1 or 2, the faculty member making the prediction was clinical or nonclinical, the faculty member was a below-average or above-average rater, the faculty panelist was an outlier, and the test item itself was easy, medium, or hard for students to answer correctly based on actual performance on a previous live examination.

Because all panelists possessed a PharmD degree, having a clinical practice site determined clinical vs nonclinical status. Faculty members whose average prediction score value was above the combined average for all 20 panelists were considered above-average raters. Faculty members whose average predictions were below the combined average for all 20 panelists were considered below-average raters. The standard deviation for the group of 20 panelists' predictions was calculated and those faculty members rating outside of ± 1 standard deviation (SD) from the mean of the group were labeled as either $+1$ or -1 outliers. The prediction scores were then analyzed using a 1-way analysis of variance (ANOVA) to compare the means between session 1 and session 2, clinical and nonclinical faculty members, above-average and below-average raters, and ± 1 SD and nonoutliers.

The 60 test items, previously Angoff scored, were selected from the question bank of test items used on an ASMA examination during 2008 or 2009. Actual test item performance history determined whether questions were labeled as easy ($p>0.93$), medium ($p>0.68$ and $p<0.88$), or hard ($p<0.62$). The p value was defined as the percentage of previous examinees who answered the test item correctly on a live examination. The p value range had no significance to the item selection for inclusion other than they were the values required to obtain a collection of 20 items in each category of difficulty.

The proprietary Angoff software system output the scoring data in a spreadsheet. A data table with the system-

generated cut score computed for each of the previously defined sessions was compiled and a new computed value was calculated and added to the data sheet to provide the average of all 3 sessions. Analysis of variance was run to compare session 1, session 2, the historical session, and the average for all sessions. While these Angoff sessions were conducted during our normal ASMA examination development process, the results reported were solely for test items rated at the P4 level. The ASMA examination cut scores for the P1, P2, and P3 examinations were determined by separate faculty panels who had previously rated the number of minimally competent P1, P2, and P3 students who would answer the item correctly.

RESULTS

A comparison of cut scores on a hypothetical 60-item examination using the cut scores created by the Angoff panels showed no significant difference between the 3 panels ($p=0.852$) (Table 1). A comparison was made of the predictions of the 10 panel members of session 1 with the 10 members of session 2 for 102 examination items (Table 1). There was no significant difference between the 2 panels ($p=0.39$). A comparison was also made of the predictions of 13 clinical to 7 nonclinical faculty members for 102 items (Table 1). There was no significant difference between the 2 groups ($p=0.10$). Faculty ability to accurately identify item difficulty on easy, medium, and hard test items was reported in Table 1. The actual performance on test items matched faculty predictions with significance and no overlap between confidence intervals of the difficulties.

There were 9 faculty members who rated below the average of both sessions (mean=52.9) and 11 faculty members who rated above the average of both sessions (mean=58.5), and the comparison of these 2 groups was significant ($p=<0.001$). There were 2 faculty members who were less than 1 standard deviation below the mean (mean=47.8) and 3 faculty members who were greater than 1 standard deviation above the mean (mean=60.8),

Table 1. Comparison of Faculty Panel Determination of Examination Cut Scores

Faculty Panel Determination	N	Mean (SD)	*P*
Angoff system-generated mean cut score for 3 different faculty panel ratings of 60 identical test items			
Historical	60	57.4 (12.9)	0.85
Session 1 – 2010	60	55.1 (12.8)	
Session 2 – 2010	60	56.2 (17.5)	
All sessions' average	60	56.2 (12.4)	
Range of cut scores=2.3 points per 100 test items			
Comparison of the cut score generated by PharmD faculty in session 1 And session 2 after rating 102 identical test items			
Session 1 – 2010	1,020	56.3 (17.1)	0.39
Session 2 – 2010	1,020	55.6 (20.6)	
Both sessions average	2,040	56.0 (18.9)	
Range of cut scores=0.7 points per 100 test items			
Comparison of cut score generated by 13 clinical and 7 nonclinical PharmD faculty members after rating 102 identical test items			
Clinical	1,326	56.5 (17.9)	0.10
Nonclinical	714	55.0 (20.7)	
Combined faculty all types average	2,040	56.0 (18.9)	
Range of cut scores=1.5 points on a 100-item test			
Comparison of the faculty Angoff rating to actual test item difficulty on exam administration 2008-2009			
Easy (93% or more of test takers answered correctly)	60	68.2 (9.3)	<0.001
Medium (68% to 88% of test takers answered correctly)	60	56.1 (12.1)	
Hard (less than 62 % of test takers answered correctly)	60	44.4 (10.9)	
Total items rated and average panel rating	180	56.2 (14.5)	

Abbreviations: PharmD=doctor of pharmacy.

and the comparison of these groups with the mean of the 15 nonoutliers (mean=56.1) was significant ($p=<0.001$). Three groups emerged from the data. The 2 faculty members greater than -1 SD and 3 faculty members greater than +1 SD were significantly different than the 15 who were average. The actual variation between the highest rating panel and the lowest rating panel on the 60-item study sample reviewed by 3 separate faculty panels projected onto a full length 187-item P4 examination was less than 5 correct answers (4.3) or 2.3%. The actual variation between the session 1 and session 2 ratings on the 102-item sample projected on to a full-length, 187-item P4 examination was less than 2 correct answers (1.3) or 0.7%. Both variations were lower than the actual standard error of the P4 examinations which ranged from 4.5 to 5.0 for the 2009 through 2013 P4 examinations.

Three separate panels of 10 faculty members generated identical cut scores. Thirteen clinical faculty members produced an identical cut score to that of 7 nonclinical faculty members. Twenty faculty members also correctly identified easy, medium, and hard questions based on actual examination performance. Five of the 20 faculty panelists created predictions outside of ±1 SD from the combined panel average.

DISCUSSION

Central to using the Angoff standard setting method for rating a college or school of pharmacy's progression examination is the panelist's ability to accurately make item performance predictions for an entry-level, minimally competent pharmacist. If panelists are unable to envision the skills and competencies of pharmacists just beginning their career, the process will not produce a valid cut score.[14] The ability of panelists to make accurate predictions that are replicable from year to year is possibly dependent on well-designed training sessions prior to the actual item-rating session.[9] This could allow for a detailed discussion of specific knowledge, skills, and abilities of minimally competent candidates. Some may argue that the panelists should have access to the previous performance data for a test item in order to adjust their predictions accordingly, but in this case, we chose not to expose the panelists to performance data prior to the session because the point of the study was to determine their ability as a group to make an independent judgment. Exposing each panel to previous performance data could have potentially masked the effect that was being measured.

Because the software that WUSOP uses in the cut scoring process allowed each panelist to see the other 9 Angoff predictions and adjust their score prior to finalizing the result, the process itself may have produced a voluntary leveling effect. There was no significant difference between yearly cut scores generated by the 2 different session panels of clinical vs nonclinical faculty members when compared to a mixed panel used in previous years (Table 1). The difference between the panels was not significant. More importantly, the range of cut scores using the extremes from either group would produce no appreciable difference on an examination the actual length of the 2011 examination. The impact of different system-generated cut score percentages on the actual enforced cut score of a 187-item P4 examination was quite small. All examination items were equally weighted at 1 point and no penalty was subtracted for wrong answers. The actual cut scores generated by these panels would vary by only 4 points on a 187-point examination. This represents a maximum range variation of less than 2.2% from the high to the low scoring panel. Using the session 1 and session 2 scores that were generated by rating a larger sample of 102 test items produced an even more reliable cut score that varied by less than 2 correct answers on a 187-item examination.

Because each examination's actual cut score was calculated from the weighted average of item scores, and the item scores were created using the Angoff process, the ability of the 3 panels to create statistically equivalent minimal competency scores was important in establishing the reliability of the annual ASMA examination. Also, because test items are routinely developed and added each year, these results provide a measure of confidence that the process does not fluctuate widely from panel to panel. The faculty members were adept at correctly identifying which items would be easier, somewhat difficult, or highly difficult for the students to answer correctly. Prior to the items appearing on the ASMA examination, and with no knowledge of item performance, panels correctly rated 60 items without a single error (Table 1).

While no significant differences were found, the clinical faculty members tended to rate the items as less difficult than did the nonclinical faculty members (56.5 vs 55, respectively) (Table 1). This was not unexpected given the clinical nature of most of the questions. Clinical faculty members spend 3 days per week at their practice site providing patient care. Pharmacists who see patients on a daily basis may expect their students to perform better on patient-care questions than nonclinical faculty members. In addition, nonclinical faculty members may be less likely to know the correct answer to clinical questions predisposing them to rate the answer as more difficult. Given the potential for these cohorts to rate differently, it was a welcome finding that the process, as practiced, showed no significant difference.

Study limitations include the possibility that the 3 clinical faculty members in session 2 could have influenced other panelists' scores because the Angoff rating sessions were designed to allow predictions to be viewed by all members. Also, data were not kept from the faculty members of panels from the historical comparator, leading to the possibility that members of historical panels served on sessions 1 or 2. However, the likelihood of panel members remembering the score that they put for a specific test item years earlier intuitively is quite low. Keeping the overarching principles of modified Angoff scoring in mind, panel members should have been able to make a judgment independent of ones from years past.

The greatest potential limitation in the WUSOP Angoff process may have resulted from the composition of the rating panel based on rating style. The spread between below-average raters and above-average raters could have resulted in a potential spread of 6 points on a 100-item examination compared to less than 1 point for the sessions conducted in 2010. Assembling the panel without some indication of rating styles can pose a risk to accuracy. Data collection will continue in order to ensure that deviation between groups and years remains minimal.

CONCLUSION

Properly constructed faculty panels can determine a reasonably reliable cut score that varies by less than 2.3% from panel to panel. In addition, faculty panels can accurately rank relative test item difficulty using the modified Angoff process as practiced by Wingate University School of Pharmacy. The panels correctly identified the difficulty of 60 test items without prior knowledge of their performance. The reliability of panel ratings is likely impacted the most by the rating styles of individual faculty panelists. Outliers have the potential to significantly alter the computed cut scores. Future data collection will establish the validity of the examination, potentially using results to predict students who are at an increased risk of failure on professional licensure examinations.

REFERENCES

1. Plaza CM. Progress examinations in pharmacy education. *Am J Pharm Educ.* 2007;71(4):Article 101.
2. Accreditation Council on Pharmacy Education. Accreditation standards and guidelines for the professional program in pharmacy leading to the doctor of pharmacy degree. https://www.acpe-accredit.org/pdf/FinalS2007Guidelines2.0.pdf. Accessed February 13, 2013.
3. Mészáros K, Barnett MJ, McDonald K, et al. Progress examination for assessing students' readiness for advanced pharmacy practice experiences. *Am J Pharm Educ.* 2009;73(6):Article 109.
4. Austin Z, O'Byrne C, Pugsley J, and Munoz LQ. Development and validation processes for an objective structured clinical examination (osce) for entry-to-practice certification in pharmacy: the Canadian experience. *Am J Pharm Educ.* 2003;67(3):Article 76.
5. Scott DM, Bennett LL, Ferrill MJ, Brown DL. Pharmacy curriculum outcomes assessment for individual student assessment and curricular evaluation. *Am J Pharm Educ.* 2010;74(10): Article 183.
6. Szilagyi JE. Curricular progress assessments: the MileMarker. *Am J Pharm Educ.* 2008;72(5):Article 101.
7. American Psychological Association. The standards for educational and psychological testing. http://www.apa.org/science/programs/testing/standards.aspx. Accessed February 13, 2013.
8. American Educational Research Association. Position statement on high stakes testing in pre-k-12 education. http://www.aera.net/AboutAERA/AERARulesPolicies/AERAPolicyStatements/PositionStatementonHighStakesTesting/tabid/11083/Default.aspx. Accessed February 13, 2013.
9. Downing SM, Tekian A, Yudlowsky R. Procedures for establishing defensible absolute passing scores on performance examinations in health professions education. *Teach Learn Med.* 2006;18(1):50-57.
10. Angoff WH. Scales, norms, and equivalent scores. In: Thorndike RL, ed. *Educational Measurement.* 2nd ed. Washington, DC: American Council on Education; 1971:508-600.
11. Cavanaugh S. Response to a legal challenge: five steps to defensible credentialing examinations. *Eval Health Prof.* 1991; 14(1):13-40.
12. Biddle RE. How to set cutoff scores for knowledge tests used in promotion, training, certification, and licensing. *Public Pers Manage.* 1993;22(1).
13. Alston GL, Love BL. Development of a reliable, valid annual skills mastery assessment examination. *Am J Pharm Educ.* 2010; 74(5):Article 80.
14. Plake BS, Impara JC. Ability of panelists to estimate item performance for a target group of candidates: an issue in judgmental standard setting. *Educ Assess.* 2001;7(2):87-97.