

Using the Angoff Method to Set Defensible Cutoff Scores for Standardized Patient Performance Evaluations in PA Education

Jim Carlson, MS, PA-C; John Tomkowiak, MD, MOL;
Curt Stilp, MS, PA-C

Purpose: This study explored the reliability and credibility of a standardized patient (SP)-based performance exam in physician assistant (PA) education with passing standards set using the Angoff method. **Methods:** PA faculty were asked to serve as judges, using the Angoff method to set a passing cut-off score for a four-case SP exam. Fifty-eight clinical-year PA students were evaluated using the exam. Passing cut-off levels and passing rates were reported for the Angoff method, for a fixed percentage cut-off score of 70%, and for a norm-referenced standard. Reliability of test items was assessed using Cronbach's alpha. Judge agreement was evaluated using kappa statistics. Credibility of judge's ratings was assessed using Pearson correlation, comparing student performance on test items with item difficulty as defined by the Angoff method. **Results:** The passing score was 62% (100% pass rate) for the Angoff method, 70% (88% pass rate) for the fixed percentage, and 72% (81% pass rate) for the relative standard. Cronbach's alpha for the test items was 0.75. Judge agreement was substantial (kappa coefficient = 0.71). Pearson correlation between actual item difficulty and difficulty as predicted by the Angoff method showed significant positive correlations (+0.44, $p < 0.05$). **Conclusions:** The Angoff method proved to be a reliable and credible method for setting a passing cut-off score for the exam. This study demonstrates that different standard-setting methods yield different results and care should be taken to engage in a defensible process when making grading decisions for SP examinations.

J Physician Assist Educ 2009;20(1):15-23

Jim Carlson, MS, PA-C, is an assistant professor in the Physician Assistant Program and director of the Education and Evaluation Center at Rosalind Franklin University of Medicine and Science. **John Tomkowiak, MD, MOL**, is associate dean for curriculum, Chicago Medical School, and director of interprofessional clinical education and simulation, Rosalind Franklin University of Medicine and Science. **Curt Stilp, MS, PA-C**, is an assistant professor in the Physician Assistant Program, Oregon Health & Science University.

Correspondence should be addressed to:

Jim Carlson, MS, PA-C
Director, Education & Evaluation Center
Rosalind Franklin University of
Medicine and Science
3333 Green Bay Road
North Chicago, IL 60064
Voice: (847) 578-8464
E-mail: james.carlson@rosalindfranklin.edu

INTRODUCTION

The literature suggests that the majority of physician assistant (PA) training programs use standardized patient (SP)-based performance exams to train students and inform grading decisions regarding clinical competence.^{1,2} SPs are people who do not actually have the specific medical problem in question, but are trained to consistently portray a scripted case focused on that medical problem.³ Typically, students engage the SP by performing clinical skills germane to the case presentation, such as taking a focused history and physical, and are then evaluated on their competency by the SP, faculty, or both.⁴ While there is a growing body of literature documenting the frequent use of SP methodology to assess PA student competency, virtually no literature exists in the PA community documenting the methods used to create defensible passing standards for SP-

based examinations.⁵ Considering that student performance on these exams influences grading decisions, the lack of clearly defined standards for SP exams in PA education should be somewhat alarming to educators.

Guidelines to follow when developing reliable SP cases and assessment tools are documented in the medical literature.^{6,7} The accuracy of SP case portrayal, rater training, and the number of items on a case checklist are but a few of the many items that can affect exam reliability.^{8,9} However, these methods only help to develop reliable measurement processes; they do not set a passing level for the SP case or exam. In fact, determining credible cut-off scores for SP-based examinations can be quite challenging as there is no one gold standard to use.^{10,11} Simply put, different standard-setting procedures will produce different passing cut-off scores. Ultimately, the method cho-

sen to determine a passing cut-off score for an SP case is as essential to the credibility of an examination as the case development process itself.¹⁰⁻¹³

Passing cut-off scores can be determined by either relative or absolute standards.¹² Relative standards (normative) are widely used and define a cut-off score by identifying passing and failing groups relative to another group or the group as a whole. Using this method, cut-off scores could be set at one or more standard deviations below the mean, for example. The fact that test difficulty is automatically corrected for is a significant advantage; however, there are inherent problems in using a relative standard to make pass-fail decisions that determine competency.¹⁴ Perhaps the most significant issue is that students demonstrating competent behavior could fail the exam or that students who demonstrate poor competence could pass the exam. Thus, using a relative standard may not serve to accurately assess student skill during SP exams.

In contrast, absolute standards define a cut-off score that is based on explicit criteria that defines a competent student. For example, if a cut-off score of 70% were used, falling below 70% would be defined as failing regardless of the class mean or standard deviation. Using an absolute standard allows for the possibility that all students could pass and be defined as competent or that all students could fail if competence is not demonstrated. Especially when documenting clinical competence, which all students should demonstrate before graduating, an absolute standard is usually most appropriate.^{10,12,15}

The Angoff Method

A variety of methods for setting defensible absolute cut-off scores for student performance assessment have been reported in the literature.^{10,15,16}

Specifically, the Angoff method has a well-established history of determining credible passing standards for multiple-choice examinations and is easily adapted for use SP examination.^{10,12,17} The method involves three basic elements: conceptualizing the borderline examinee, identification of specific test items, and using expert judges to estimate whether a borderline examinee will appropriately perform each of the test items. Borderline examinees are students who demonstrate behaviors that are sometimes correct, but often not. They have a 50:50 probability of passing or failing the exam, which places them just at the cut-off score for a given competency exam.

Within SP exams, the test items are usually defined on a case-specific checklist. Case-specific checklists are typically composed of behavioral items that students are expected to demonstrate during a given case.¹⁸ For example, if the SP case involves chest pain, the checklist items may document whether the student asked historical items about the nature of the chest pain (onset, duration, location, etc.) or included heart auscultation as part of her interaction with the SP. Typically, most case checklists should include enough items to appropriately measure competence, but not so many elements that reliability is decreased due to poor rater recall.^{8,9} The number of items accurately performed vs. the number of items on the checklist produces a score for the case.

When the Angoff method is used to set a standard for SP examinations, expert judges (usually core or clinical faculty) define the characteristics of a borderline student and then try to estimate if a borderline candidate is likely to correctly perform each of the items on the case-specific checklist. In the modified method most often applied to SP

examinations, and used within this study, a panel of expert judges make predictions for each case checklist item, guided by pilot data.^{10,12,15} The average of the judges' ratings on the case checklist sets the passing standard for the case. The average of the passing levels of all case checklists sets the passing standard for an examination with multiple stations.

Advantages of the Angoff method are that it is fairly easy to employ because it does not require faculty to directly observe every student's performance, a process that is very time-consuming during multi-station SP exams.¹⁰ In fact, it does not require that a judge actually watch any student-SP encounters at all. While there is a robust body of research to support its use, a disadvantage of the method is that judges occasionally report feeling that there is no firm basis for the standard that is set, since they are predicting performance as opposed to directly observing examinee performance.¹² Regardless, due to its potential to provide an efficient and feasible method of setting defensible cut scores during SP examinations in PA education, adaptation of the method using PA faculty as expert judges should be explored.

Purpose and Research Questions

The purpose of this study was to engage a national group of PA faculty to serve as expert judges in using the Angoff method to set passing cut-off levels for four SP cases used in a multi-station SP exam in PA education. Specifically, this study addresses the following questions:

1. What cut-off scores should be set for the individual SP cases and the combined four-case multi-station SP exam, using the Angoff method?
2. What is the degree of reliability and credibility of the passing stan-

dards set by a group of PA faculty serving as expert judges?

3. When used for a multi-station SP examination with an actual student cohort:

- What passing rates are observed using the cut-off levels set by the Angoff method, an arbitrary absolute passing standard of 70%, and a relative standard of one standard deviation below the mean?
- What is the reliability of the measurement tools when assessing student clinical ability?

METHODS

Twenty-five PA faculty participated in a standard-setting workshop during the 2007 PAEA Annual Education Forum. Participants were given an overview of the Angoff method and its application when applied to standardized patient evaluation and asked to serve as expert judges when using the method to set a performance standard for one of four standardized patient cases. The cases and the standards set were used to assess PA student performance during a mandatory clinical year SP-based clinical skills examination. The Rosalind Franklin University of Medicine and Science (RFUMS) IRB granted approval for this study.

Standardized Patient Case and Checklist Development

The four SP cases and their corresponding assessment checklists were selected from a case bank at our institution's SP program. The cases and checklists were developed by consensus of RFUMS MD and PA faculty and had been successfully piloted with clinical-year PA students. The case topics selected for this study are noted below.

- Case #1: Pre-Operative H+P — 50 y/o male, outpatient presentation.
- Case #2: Episodic Shortness of Breath (not acute) — 30 y/o male or female, outpatient presentation.
- Case #3: Acute Abdominal Pain — 40-something y/o female, ER presentation.
- Case #4: Low Back Pain/Radicular Symptoms (subacute) — 40-50 y/o male or female, outpatient presentation.

These specific case scenarios were selected based on the feeling of core PA program faculty that they represent common problems in which clinical-year students should be able to demonstrate competency prior to graduation. Specifically, eight RFUMS PA faculty discussed the cases and felt the topics were appropriate for measuring clinical competency midway through clinical training. While various clinical elements can be evaluated using these cases, this study chose to focus specifically on the case-specific checklist elements used to assess clinical data gathering, eg, student ability to elicit essential history, and physical exam elements defined as important to the work-up of the case topic.

It is worth mentioning that communication elements are frequently evaluated during SP examinations but were deliberately not explored within the context of this study. The dichotomous ratings (done/not done) for the case history and physical exam elements lend themselves well to the Angoff method since they allow judges to rate whether a borderline candidate will perform a specific item as either “yes” or “no.” Communication skills items are often global in nature and are scored on a Likert scale.¹⁹ While these items may

also be relevant to demonstrating clinical competence in these cases, employing the Angoff method with the additional communication items was speculated to be too time consuming for the time-limited PAEA workshop. Instead, the authors chose to focus only on case history and physical elements.

Using the Angoff Method to Set Passing Cut-off Scores for Each Case

The short 50-minute workshop held at the 2007 PAEA Annual Education Forum was designed to introduce PA faculty to the Angoff method and have them serve as expert judges to set a passing cut-off score for an SP case and checklist. Judges were divided into groups of six or seven. This is consistent with recommendations for engaging 5-10 judges when using the Angoff method.²⁰ Unfortunately, no background information was collected regarding judges' experience with the Angoff method or years as a PA educator. Judges were recruited as a convenience sample solely on the basis of their attendance at the PAEA workshop.

Following introduction of the method, judges conceptualized the characteristics of borderline clinical-level PA trainees (students who have completed some clinical training, but who at times demonstrate clinically incompetent behavior (Figure 1)). A modified Angoff approach was used in which the difficulty of checklist items, determined from prior piloting each of the four cases, was used to guide rater judgment.^{10,12} As is customary with the modified Angoff method, the difficulty of each test item was reported to judges as the percentage of examinees receiving credit for the item during pilot study. This percentage is known as a “p-value” (this is different from and not to be confused with a traditional *p*

Figure 1. Sample Characteristics of Borderline Examinees Listed by Expert Judges During the Standard-Setting Process

- Know how to take a medical history but are unable to identify what is pertinent.
- Difficulty focusing physical exam – will do the “hint” full physical on opposed to a physical exam that provides meaningful data for the problem at hand.
- Poor interpersonal communication with the patient.
- Will forget basic things like existing medical.
- Heavily unprepared – will wonder if medical data is recalled.
- Inappropriate responses to patient questions.
- Not being able to connect up with a little medical diagnosis.
- Difficulty getting relevant information – always searching for more information.
- Overconfident, doesn't do enough before the path of least resistance.
- Disorganized approach to collecting patient information.
- Weak knowledge of what history to ask and what PE to perform.
- Superficial physical exam skills – not enough detail to provide meaningful information.
- Unable to self-correct/re-direct actions when “stuck.”
- Poor technique on exam – techniques will not regard to gathering information going through the motions.
- Unable to justify using medical paper and examinee's response.

value that defines statistical significance). Figure 2 shows a sample Angoff judge's rating sheet that includes the case checklist items, specific item p-value, and the judges' options for rating each item. Judges individually determined whether a borderline candidate would be likely to perform each of the checklist items for a specific case (rated as either “yes” or “no” on the Angoff judge's rating sheet). Judges were instructed to discuss their decisions for each item and share their rationale, but ultimately each judge was free to submit a final checklist for which his or her judgment on each item was final; coming to a consensus opinion for every item was not required.

Other Standard-Setting Procedures

As a frame of reference, two other passing cut-off levels were set using more commonly recognized methods. First, an alternative absolute passing standard was set using the traditionally recognized passing cut-off level of 70%. This was selected because the PA program running this study com-

monly uses 70% as a passing cut-off for other measures (multiple choice tests, etc.). Additionally, a relative (norm-referenced) standard was set at one standard deviation below the mean for the student cohort taking the exam. While the reason for selecting either of these two cut-off levels was generally arbitrary, these methods are commonly used to make grading decisions.^{5,20}

Exam Administration and Collection of Student Data

Fifty-eight clinical-year PA students participated in a multistation standardized patient exam composed of the four cases. Students saw all four cases in succession and were told that they would be assessed on their ability to take a focused history and physical exam for each of the problems presented during the case scenarios. Standardized patients were recruited from the institution's SP pool. All had prior experience evaluating clinical-level PA and MD students. Each SP received at least 4 hours of training in how to accurately portray one of the selected cases and use of the

corresponding case checklist. Checklists were completed by SPs immediately after each encounter to document the degree to which students correctly performed case history and physical exam elements. Checklist scores were averaged (total number of items performed correctly/total number of possible items) to determine a score for each case and overall score for all four cases combined for each student.

Statistical Analysis

Judge Reliability. Inter-rater agreement among the expert judges was determined using Cohen's kappa coefficient. A kappa score was determined for each case checklist and the combined four-case exam. Using Landis and Koch's rating of kappa reliability, $k > 0.60$ was considered acceptable agreement.²¹

Exam/Test Item Reliability. Internal consistency of the case and combined four-case exam was measured using Cronbach's alpha.²² An alpha coefficient of 0.70-0.90 was considered an appropriate level of homogeneity. This is an indicator of whether or not the checklist items provide a reliable measurement of the construct being assessed (clinical data collection).

Credibility. Credibility of the standard set by the judges for each case and exam was evaluated using the Pearson correlation coefficient.¹⁴ Since the Angoff standard-setting process asks judges to predict exam and item difficulty, a significant, positive correlation between the level of difficulty that judges assigned to checklist items (the likelihood that a borderline candidate will perform an item) and the actual item difficulty (the percentage of students who appropriately performed the checklist item during the exam itself) lends credibility to the accuracy of the judges ratings.

Figure 2. Sample Angoff Standard-Setting Judges' Worksheet, Case #1 Pre-Op Assessment

Item	Difficulty estimated by setting judge (0-1.0)	% of students predicted to pass	Passing rate (0-1.0)
HISTORY CHECKS AND VITALS			
1. Why was your scheduled to have surgery?	0.10	97%	98%
2. Do you have any current medical conditions/medical conditions you are treated for?	0.08	98%	98%
3. Are you on any current medications? (Do not include over the counter medications)	0.08	98%	98%
4. Do you have any current pain (please state where, how, how long, and how often)? (Do not include over the counter medications)	0.11	97%	98%
5. I asked the status of the chest pain (onset, duration, onset, radiation, quality, severity, associated and relieving factors - related onset and/or last time)	0.15	97%	98%
6. I have not had any shortness of breath (do you have any conditions associated)?	0.10	97%	98%
7. Have you ever had surgery (where)?	0.10	97%	98%
8. Do you smoke?	0.10	97%	98%
9. Do you use alcohol?	0.10	97%	98%
10. Do you have any past prescriptions (and surgery)?	0.10	97%	98%
11. Allergy (and you describe)	0.08	98%	98%
12. Do you have any allergies to medication?	0.08	98%	98%
PHYSICAL EXAMINATION			
13. I asked the patient to stand without leaning on his primary support (if any) to look at his feet.	0.08	98%	98%
14. Auscultated (over the heart) lung fields - 4 fields - bilaterally and symmetrically	0.14	97%	98%
15. All Auscultated (over the heart) lung fields	0.08	98%	98%
16. Palpated the chest wall to feel for any (gross) crepitations of pain (not over the sternum or ribs)	0.10	97%	98%
17. Auscultated (over the heart) the heart	0.08	98%	98%
18. Auscultated (over the heart) the heart	0.08	98%	98%
19. Auscultated (over the heart) the heart	0.08	98%	98%
20. Auscultated (over the heart) the heart	0.08	98%	98%
21. Auscultated (over the heart) the heart	0.08	98%	98%
22. Auscultated (over the heart) the heart	0.08	98%	98%
23. Palpated in abdomen (for stool) if found necessary for the examination (do not touch abdomen if not necessary)	0.10	97%	98%

Four case exam =

0.07500

Descriptive Statistics. Statistics for student exam scores were explored in the context of each of three standard-setting procedures. Specifically, the passing cut-off level and passing rate were explored for two absolute standards (Angoff-generated cut-off score and an arbitrary assignment of 70% as an absolute cut-off score) and a relative standard (one standard deviation below the mean). Passing cut-off levels and passing rates were reported for each case and the combined four-case exam.

RESULTS

Figure 1 highlights examples of borderline characteristics of clinical PA students articulated by the judges during the Angoff procedure. These characteristics were used to guide judges' perspectives when setting passing standards for each case and checklist item. Table 1 highlights inter-rater reliability among judges, as defined by Cohen's kappa (k), Cronbach's alpha, and Pearson correlation between item difficulty, as

demonstrated by students during the full exam and judges' predicted item difficulty for each case and the combined four-case exam. All four cases and the combined exam demonstrated substantial judge inter-rater reliability when determining a passing level using the Angoff procedure. The overall exam, as well as the shortness of breath, back pain, and abdominal pain case checklist item difficulties predicted by the expert judges showed significant ($p < 0.05$)

Table 1. Inter-Judge Agreement for Test Item Difficulty (Reliability) and Correlation of Item Difficulties Set by Angoff Method and Item Difficulty Demonstrated by Actual Student Data (Credibility)

Case	Inter-Judge Agreement (Kappa Coefficient)	Pearson correlation coefficient between judge prediction of item difficulty and actual item difficulties during student examination (* $p < 0.05$)
Pre-operative H+P	0.79	+0.27
Shortness of breath	0.71	+0.55*
Acute abdomen	0.67	+0.48*
Low back pain	0.65	+0.57*
Full exam	0.71	+0.44*

*Standards of strength of kappa coefficient (Landis and Koch 1977)³³

< 0 = poor

.01-.20 = slight agreement

.21-.40 = fair agreement

.41-.60 = moderate agreement

.61-.80 = substantial agreement

.81-1.0 = almost perfect agreement

Table 2. Cronbach's α Value for Case Checklist Items and Full Exam

Case	Checklist Item Reliability (Cronbach's α)
Pre-operative H+P	0.48
Shortness of breath	0.76*
Acute abdomen	0.35
Low back pain	0.47
Full exam	0.75*

*acceptable item reliability

positive correlations with the item difficulties determined by student performance during the actual exam. Table 2 highlights Cronbach's alpha values for the case checklist and the exam as a whole. Only the shortness of breath case checklist demonstrated an acceptable Cronbach's alpha value (0.76) as an individual case; however, the full four-case combined exam yielded acceptable internal consistency ($\alpha = 0.75$). This indicates that as a whole, the exam items measured a similar construct: clinical data gathering.

Table 3 demonstrates the mean and standard deviation of each case and the full exam per student per-

formance data as well as the passing cut-off value and passing rate for each case and the combined four-case exam. The passing standard set by the Angoff procedure (62%) yielded a 100% pass rate for the full examination, while the other standards used generally yielded higher passing cut-off levels and lower passing rates.

DISCUSSION

In this study, a convenience sample of PA faculty who served as expert judges during a one-hour workshop demonstrated substantial agreement on the passing standards set for an SP exam, using the Angoff method. Additionally, the expert judges were

able to use the method to accurately predict test item difficulty for the exam as a whole and for three of the four cases. While there is room for improvement, the fact that such agreement and predictive value could be reached in such a short training workshop suggests that the method is a potentially feasible and defensible method for setting standards in SP exams in PA education. The method offers reliable standards, is simple to employ, and is not overly time consuming for busy PA faculty.

In addition to documenting credible standards in terms of judge accuracy and agreement, this study also highlights several key observations about integrating reliable SP examinations within a PA curriculum. First, it should be recognized that the Angoff method itself is only capable of setting a defensible passing cut-off score and does not itself guarantee a reliable evaluation process. As noted earlier, many factors affect the reliability of SP examinations, including rater training, checklist length, and the number of case stations used in a multi-station exam.^{8,9} Care should be taken to develop appropriate assessment tools and to ensure that raters, whether SPs or faculty, are appropriately trained in the rating process and demonstrate appropriate accuracy and recall when assessing student performance.

Particularly highlighted in this study is the need for multiple stations to achieve a reliable assessment of student performance. This is consistent with previous studies that have advocated the need for multiple stations when assessing students during high-stakes SP-based exams.^{6,7} Cronbach's alpha has been used to demonstrate the reliability of SP evaluations by measuring the internal consistency of checklist items and the degree to which checklist items measure a defined construct.¹⁶ In this

Table 3. Case and Exam Passing Levels and Passing Rates for Two Absolute Standards (including Angoff) and One Relative Standard

Overall Student Exam Data (N=58 examinees)			Standard-Setting Method					
			Angoff (absolute)		Arbitrary 70% Cut-off (absolute)		Norm Referenced (relative)	
Case	Mean	Standard Deviation	Cutoff	Passing Rate	Cutoff	Passing Rate	Cutoff	Passing Rate
Pre-op H&P	86%	9.0	63%	100%	70%	89%	77%	82%
Shortness of breath	73%	11.8	70%	69%	70%	69%	61%	86%
Acute abdomen	86%	8.8	62%	100%	70%	93%	77%	77%
Low back pain	72%	13.0	52%	91%	70%	58%	59%	79%
Full exam	79%	7.1	62%	100%	70%	88%	72%	81%

exam, the construct measured was clinical data gathering in terms of history and physical examination relevant to the case topics. While the shortness of breath case checklist demonstrated an acceptable alpha, the other three cases did not demonstrate this same degree of internal consistency unless combined in the four-case multi-station evaluation. While the purpose of the Angoff method is designed to set a passing cut-off level, the data collected on student performance during this study reinforce the need to use multiple cases in order to reliably measure student performance in high stakes SP examinations.

The advantages to the Angoff method in terms of defensibility of the assessment process become clearer when compared to arbitrarily assigning an absolute passing standard of 70% or applying a relative passing standard of one standard deviation below the mean. For the students participating in the exam generating data for this study, the Angoff method yielded a 100% passing rate (percentage of students who scored above the passing cut-off)

compared to lower passing rates for both other standards set. This may imply that the cut-off scores produced by the Angoff method were too lenient; however, it should be noted that all students within this exam cohort went on to graduate a few months after the exam and were in good academic standing. Using either cut-off score other than the Angoff-generated passing level would have meant an exam failure for students with good academic standing. This reinforces the need for programs using SP examinations for high stakes assessment to employ a defensible absolute standard when making pass/fail decisions as opposed to arbitrarily assigning passing standards.

Limitations

This study does have several limitations. First, a 50-minute training workshop at the 2007 PAEA Annual Education Forum was likely not an optimal situation in terms of providing sufficient time to use the Angoff method, for item discussion, and for establishing the most acceptable consensus among the judges.^{10,12} While the ideal discussion time necessary

for Angoff judges to set a reliable standard for an SP case is not established, the process is generally considered complete when judges have had time to thoroughly discuss each item, share their rationale for each rating, and make rating adjustments if necessary.¹⁷ Since the primary purpose of the conference workshop was to expose participants to the Angoff method, rating decisions had to be made more quickly than they would have under highly controlled use of the method, in which experienced judges had the time they felt necessary to review each checklist item thoroughly. It is encouraging that the judge's ratings demonstrated significant agreement and predictive value for examinee performance despite these limitations; however, more substantial agreement between judges might have been attained with greater time to employ the method.

It would have been helpful to collect data regarding each judge's experience as a PA educator and prior use of the Angoff method. While this was not officially determined, the authors suspect that none of the workshop participants had used the Angoff

method to set passing standards in SP examinations prior to the workshop. Also, novice faculty might lack the frame of reference necessary to accurately conceptualize a borderline candidate or may have a very different perspective about what borderline means in terms of student performance. Further study using the method should account for these variables as it is very possible that greater experience with the Angoff method and greater years as a PA educator would result in the most credible passing standard. Additionally, determining whether different cohorts of judges set a similar passing standard for the same cases would be helpful in establishing whether the method is easily reproducible when used by different judges.

Another limitation is that the student data evaluated in this study are all from one program. The judges' ratings derived from a national sample of PA educators, but it will be helpful to explore whether PA students from different training programs demonstrate similar competence when being assessed by the cases and measurement tools in this study. Further study of these cases should be done across multiple PA training programs at different institutions to explore the reproducibility of the process and findings. Additionally, using the cases and instruments to assess the performance of other learners at a similar level of training (eg, clinical-level medical students) would help to establish exam validity.²³

Future Directions and Opportunities

As noted above, the purpose of the Angoff method is to define a passing standard for a case or exam. It is not a process for developing a reliable case or checklist. Faculty still need to engage in credible methods of case

and checklist development, including proper rater training, determining an acceptable number of case items, verifying inter-rater reliability with actual student exam data, and attempting to explore the validity of the measurement tools. However, the Angoff method does offer an opportunity to promote a healthy discussion among faculty about expectations for student performance and defining clinical competence at a program and national level. Through the debate, faculty may come to better understand and more clearly articulate their expectations of students and understanding of student clinical behavior. Additionally, the resulting consensus-based passing standard will likely reflect a more appropriate standard of care for a case rather than the individual practice preferences of a case author.

If enough PA faculty become familiar with the process, it is even possible that this method is simple enough for expert judges across multiple institutions to use within an online discussion setting. This possibility offers the opportunity to efficiently engage large numbers of expert judges in determining passing cut-off scores for many SP cases that could be used to grow a SP case library for use within PA education. The creation of such a case library linked with credible performance standards would not only provide training programs a potentially more reliable means of assessing student clinical performance, it would also provide a more standardized set of assessment tools through which to research curricular impact and PA student ability across multiple programs. Since physician training also makes heavy use of SP methods, a bank of cases with credible passing standards also offers the opportunity for comparison of clinical skill across professions.

Given the frequency with which PA programs report using SP-based evaluation to influence student grading decisions, it is important that PA educators take steps to use this resource wisely. Also, SP-based assessment is often costly to implement in terms of faculty time, student time, and money. Without engaging in a credible standard-setting procedure, PA faculty risk squandering the resource and making inaccurate decisions about student ability. Using the Angoff method appears to offer PA faculty a simple and defensible method for making pass/fail decisions during SP examinations and its impact should be explored further.

REFERENCES

1. Calhoun B, Vrbin C, Grzybicki D. The use of standardized patients in the training and evaluation of physician assistant students. *J Physician Assist Educ.* 2008;19(1):18-23.
2. Calhoun B, Chambers D. Standardized patients and simulated patient encounters in the evaluation of students. *Perspective on Physician Assistant Education.* 2004;15(2):99-101.
3. Barrows HS. *Simulated Patients (Programmed Patients): The Development and Use of a New Technique in Medical Education.* Springfield, IL: Charles C Thomas; 1971.
4. Epstein RM. Assessment in medical education. *N Engl J Med.* 2007;356:387-396.
5. Anderson K, Johnston A, Knudson J, Tournell C. *The Use of Case-Specific Checklists in Evaluating Physician Assistant Student Performance During Standardized Patient Encounters.* [Master's Thesis]. Rosalind Franklin University of Medicine and Science; May 2006.
6. Williams R. Have standardized patient examinations stood the test of time and experience? *Teach Learn Med.* 2004;16(2):215-222.
7. Barman A. Critiques on the objective structured clinical examination. *Annals Academy of Medicine Singapore.* 2005;(34):478-82.

8. Huber P, Baroffio A, Chamot E, Herrmann F, Nendaz MR, Vu NV. Effects of item and rater characteristics on checklist recording: what should we look for? *Med Educ.* 2005; 39(8):852-8.
9. Vu NV, Marcy MM, Colliver JA, Verhulst SJ, Travis TA & Barrows HS. Standardized (simulated) patients' accuracy in recording clinical performance check-list items. *Med Educ.* 1992;26(2):99-104.
10. Downing S, Tekian A, Yudkowski R. Procedures for establishing defensible absolute passing scores on performance examinations in health professions education. *Teach Learn Med.* 2005;18(1):50-57.
11. Resse A, Chung E, Gardiner M, Williams S. Competency domains in an undergraduate objective structured clinical examination: their impact on compensatory standard setting. *Med Educ.* 2008;42:600-606.
12. Noricini J. Setting standards on educational tests. *Med Educ.* 2003;37: 464-469.
13. Boulet JR, De Champlain AF, McKinley DW. Setting defensible performance standards on OSCEs and standardized patient examinations. *Medical Teacher.* 2003;25(3): 245-9.
14. Verhoeven B, Van der Steeg A, Scherpbier A, Muijtens A, Verwijnen G, van der Vleuten C. Reliability and credibility of an Angoff standard setting procedure in progress testing using recent graduates as judges. *Med Educ.* 1999;33: 832-837.
15. Kaufman D, Mann K, Muijtens A, van der Vleuten C. A comparison of standard-setting procedures for an OSCE in undergraduate medical education. *Acad Med.* 2000;75(3): 267-71.
16. Kramer A, Muijtens A, Jansen K, Dusman H, Tan L, Vleuten C. Comparison of a rational and empirical standard setting procedure for an OSCE. *Med Educ.* 2003;37: 132-139.
17. Ricker K. *Setting Cut Scores: Critical Review of Angoff and Modified-Angoff Methods.* Center for Applied Measurement and Evaluation. University of Alberta, Canada. <http://www.education.ualberta.ca/educ/psych/crame/files/RickerCSSE2003.pdf>. Accessed March 1, 2008.
18. Gorter S, Rethans JJ, Scherpbier A, van der Heijde D, Houben H, van der Vleuten C, van der Linden S. Developing case-specific checklists for standardized patient-based assessments in internal medicine: a review of the literature. *Acad Med.* 2000;75:1130-7.
19. Schirmer J, Mauksch L, Lang F, Marvel K, Aoppi K, Epstein R, Brock D, Pryzbylski M. Assessing communication competence: a review of current tools. *Fam Med.* 2005;37(3):184-92.
20. Sanju G, Haque M & Oyebode F. Standard setting: comparison of two methods. *BMC Med Educ.* 2006;6:46.
21. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33:159-74.
22. Downing SM. Validity on the meaningful interpretation of assessment data. *Med Educ.* 2003;37:830-837.
23. Asprey D, Hegmann T, Bergus G. Comparison of medical student and physician assistant student performance on standardized-patient assessments. *J Physician Assist Educ.* 2007;18 (4):16-19.