RESEARCH

Towards an Operational Definition of Clinical Competency in Pharmacy

L. Douglas Ried, PhD,^a Charles A. Douglas, PhD, MBA^b

^a University of Texas at Tyler, Tyler, Texas

^b Texas A&M University Irma Lerma Pangel College of Pharmacy, Kingsville, Texas

Submitted July 18, 2014; accepted December 7, 2014; published May 25, 2015.

Objective. To estimate the inter-rater reliability and accuracy of ratings of competence in student pharmacist/patient clinical interactions as depicted in videotaped simulations and to compare expert panelist and typical preceptor ratings of those interactions.

Methods. This study used a multifactorial experimental design to estimate inter-rater reliability and accuracy of preceptors' assessment of student performance in clinical simulations. The study protocol used nine 5-10 minute video vignettes portraying different levels of competency in student performance in simulated clinical interactions. Intra-Class Correlation (ICC) was used to calculate inter-rater reliability and Fisher exact test was used to compare differences in distribution of scores between expert and nonexpert assessments.

Results. Preceptors (n=42) across 5 states assessed the simulated performances. Intra-Class Correlation estimates were higher for 3 nonrandomized video simulations compared to the 6 randomized simulations. Preceptors more readily identified high and low student performances compared to satisfactory performances. In nearly two-thirds of the rating opportunities, a higher proportion of expert panelists than preceptors rated the student performance correctly (18 of 27 scenarios).

Conclusion. Valid and reliable assessments are critically important because they affect student grades and formative student feedback. Study results indicate the need for pharmacy preceptor training in performance assessment. The process demonstrated in this study can be used to establish minimum preceptor benchmarks for future national training programs.

Keywords: advanced pharmacy practice experience, continuing education, faculty development, preceptors, simulation, reliability, competency assessment

INTRODUCTION

Competency-based education focuses on integrating competency into all facets of training and assessment.¹ In experiential education, the objectives of competency-based education are to appraise student performance in the clinical setting and to determine if the student is sufficiently competent to enter into professional practice.² However, there is no widely accepted method to evaluate whether these assessment programs actually discriminate between competent and noncompetent students. These objectives are especially elusive for the advanced pharmacy practice experiences (APPEs). The 2008 American Association of Colleges of Pharmacy (AACP) president and the American College of Clinical Pharmacy (ACCP) Educational Affairs Committee called for a standard APPE assessment instrument.^{3,4} Despite best intentions,

several issues must still be addressed to realize valid and reliable assessment of student performance during APPEs across practice settings and preceptors.

First, volunteer preceptors assess student performance during APPEs in a majority of pharmacy programs and are responsible for acting as gatekeepers to practice. These preceptors are expected to assess students' readiness to enter into practice based on a comparison of observed behavior with professional performance standards (eg, the Center for the Advancement of Pharmaceutical Education (CAPE) Educational Outcomes). However, performance levels are not sufficiently operationalized to rate a student as competent or not competent using CAPE or any other educational performance gold standard. In the volunteer preceptor model, the preceptor's practice experience positively affects the quality of student performance ratings, and those preceptors who are judged as better clinicians are usually better at rating the job performance of others.⁵ Additionally, preceptors with little experience or substandard clinical skills have greater idiosyncratic assessment scores that increase score variation.⁶ For example,

Corresponding Author: L. Douglas Ried, PhD, Ben and Maytee Fisch College of Pharmacy, University of Texas at Tyler, 3900 University Blvd, Tyler, TX 75799. Tel: 519-565-5919. E-mail: dried@uttyler.edu

student assessments differ between nursing educators and nonfaculty nursing clinical preceptors.^{7,8} Educators, who often develop assessment instruments, may have a different concept of competency compared to practicing clinicians. In short, experts' assessments are likely to be better than non-experts' assessments, even with standardized tools.

Next, lack of standardization has an impact on the variability of an individual preceptor's assessments. Pharmacy preceptors are urged to assess students' competence based on a comparison of observed behavior with professional performance standards.⁹ However, individual performance standards are as diverse as the number of instruments and preceptors. Preceptors commonly use an intuitive decision-making framework to assess students, commenting that "... 'gut feeling' seems to represent their cognitive integration of those characteristics into a decision about the overall adequacy of performance."¹⁰ Cross and Hicks concluded that preceptors commonly use implicit criteria in the decision making process; preceptors would ask themselves if they would hire this student for their decision-making framework rather than use clinically-based objective measures.¹¹ Alexander's heuristic model found that preceptors assessed whether, in their opinion, the student's performance was representative of the desirable characteristics of an entry-level practitioner.¹² However, these decision-making frameworks are affected by impressions of previous students and, especially, a personal perception of what constitutes an entry-level practitioner.

Finally, pharmacy education and its accreditors (eg, the American Council for Pharmacy Education (ACPE)) provide little guidance on acceptable preceptor inter-rater reliability, and the acceptability may be as diverse as the number of competency assessment instruments. Inter-rater reliability refers to the degree to which assessment scores are consistent and provide clear information useful for differentiating between individual students.¹³ Measures of inter-rater reliability in the clinical setting are often low and estimates of .80 or above are rarely achieved in other health professions.¹⁴ An instrument, even with face validity, is no better or worse than no instrument if it produces inconsistent results between preceptors. For example, if two students' performances are the same at two sites, it is possible that one preceptor may assess their students as noncompetent and require additional training of the student, while the other preceptor may rate their student as competent and the student is deemed ready to practice. While we assume that all preceptors would rate the same interaction similarly, that assumption hasn't been rigorously tested. Thus, guidance on methods to estimate both preceptor accuracy and inter-rater reliability is warranted.

In addition to inter-rater reliability, preceptors' APPE ratings need to accurately discriminate between

students who are competent and students who are not competent. For example, Noel et al measured the accuracy of student assessments made by medical faculty members (n=203), who viewed two clinical case simulations on videotape.¹⁵ Using a validated scoring system, overall accuracy was calculated at 32% for the group using an open-ended evaluation form and 60% for the group using a structured evaluation form. More than half of the participants rated the students' performances in the two scenarios as satisfactory or superior. In both case simulations, however, the designers purposely included enough errors so that all faculty members should have rated the students' performances as less than satisfactory. Thus, accuracy was less than desirable even with the improved inter-rater reliability measures for the group using a structured evaluation form. Noel et al's study illustrates two points. First, a higher inter-rater reliability measure by itself does not support a case for validity. Second, the use of videotaped simulations is a sound method to measure preceptor inter-rater reliability and accuracy.

Given these aforementioned challenges, the goal of our study was to evaluate the inter-rater consistency and accuracy of pharmacists on the Drug Therapy Evaluation and Development competency from the SUCCESS APPE clinical competency assessment tool.¹⁶ Specifically, our objectives were to estimate the inter-rater reliability of preceptors viewing video simulations of student/preceptor clinical interactions and to compare expert and nonexpert ratings of those interactions.

METHODS

In response to preceptors' calls for a single APPE assessment instrument in Florida, SUCCESS was developed to assess student performance during APPEs.¹⁶ The SUCCESS instrument consists of 13 competencies based on 99 subcompetencies and is used by all colleges and schools of pharmacy in Florida for APPE assessment. A multifactorial experimental study design is used to estimate inter-rater reliability and accuracy of preceptors' assessments of student performance in clinical simulations.

A comprehensive description of the vignette and video development can be found elsewhere;¹⁷ however, an abbreviated description is included. The Drug Therapy Evaluation and Development competency (competency #3) from SUCCESS was chosen. Vignette and video development were accomplished in two phases. First, a physician at the local teaching hospital developed a checklist of activities necessary to meet the target competency and subcompetencies. For example, subcompetency statement A reads: "synthesizes complete history, laboratory, and physical examination data to identify problems." The physician-generated checklist required the student to read the history and identify the patient's age, gender, race, symptoms, and

signs of disease and past medical history to demonstrate competency in subcompetency A. The complete checklists for all subcompetencies can be found elsewhere.¹⁷ In addition to the checklist items, the levels of performance on each activity was adapted from the physical therapy assessment instrument.¹⁸ Five domains demonstrate competence and excellence: supervision, quality, complexity, consistency, and efficiency.

The checklist and performance criteria were submitted to a Delphi panel of pharmacy experts (n=22) using online survey methods via Survey Monkey (Survey Monkey, Inc., Palo Alto, CA).^{19,20} We asked pharmacy school experiential directors to nominate preceptors to join our expert panel. The nomination criteria included knowledge of clinical performance assessment principles, high proficiency in assessment of clinical skills, acknowledgement as an exceptional preceptor, and advanced clinical training. Panelists provided comments using a reactive Delphi model, which requires panelists to react to previously prepared material (ie, the expert physician's checklist) rather than generating original material.²¹ The panelists generally agreed on the checklists and performance levels for each of the subcompetencies (eg, >80%). In the end, it was decided that subcompetencies A and B would be used separately, and subcompetencies D and E could be combined as both referred to "design and evaluation of treatment regimens." Consequently, in the next phase, subcompetencies A, B, and D/E were used to illustrate APPE students reporting to their preceptors in a video simulation.

In phase 2, a nonprobabilistic convenience sample (n=12) from the original Delphi panel was recruited based on the quality of their previous comments. Panelists were offered \$400 compensation for their time and effort. Ten preceptors accepted the original invitation and seven (59%) completed phase 2. The expert panelists had two tasks. The first task was to review and edit scripts prior to filming. Scripts were distributed to the expert panel representing each performance level for each subcompetency. The expert panel was asked for input on student behaviors that could be observed and that would illustrate different levels of performance to generate the performance-rating rubric. After panelists' responses to the first round were compiled, new versions of the scripts were returned to the expert panel and finalized after the second review. The second task of the expert panel in phase 2 was to watch the video vignettes and score the performance using the new criteria. After the videos were completed several months later, they were presented to the expert panelists.

Video simulations are generally accepted as a method to evaluate assessment strategies. In a video simulation, preceptors view and assess the exact same performance by actors.²² In theory, the preceptors' assessments should

be the same. In this case, the videotaped scenarios simulated an authentic preceptor encounter with a student during an APPE. Multiple preceptors observed the same scenarios and rated the students' competence. Nine third-year students from a nonparticipating college of pharmacy were recruited and paid \$100 each. Students completed rehearsals and filming in one day. An experienced faculty member played the role of the preceptor in each simulation and consistently presented the scripted prompts to the students.

In terms of participant expertise, a convenience sample of preceptors representing a variety of practice settings, regional locations, and educational institutions was recruited to participate in a 3-hour online continuing education (CE) course on student clinical competency assessment. The course included presentations, video portrayals of APPE performances, and discussion. The CE program was cost-free to the preceptors. The preceptors were offered a \$100 eGift certificate from Amazon.com and were provided with 3.0 contact hours (0.3 continuing education units) of CE credit if they completed the webinar course and submitted the assessments. The Texas Pharmacy Association (TPA) distributed materials that promoted the CE program via one of TPA's regular e-mail lists to pharmacists state-wide. Two Florida-based schools provided mailing lists containing the names and contact information of their preceptors. The promotional materials encouraged the preceptors to participate in the CE program and were distributed directly to the preceptors.

The preceptors watched the video vignettes and evaluated student performance using the TPA CE webinar site.²⁰ A total of 52 participants started the 3-hour webinar. The dropout rate was steady over the duration of the webinar. Complete data was collected from 42 participating preceptors (81%).

Two to three months after giving feedback on the scripts, seven of the initial 10 expert panelists watched the video vignettes of APPE performances and submitted their assessment scores using a website created exclusively for the panel.²⁰ Panelists were offered a \$100 eGift certificate for their assessments. In addition, two members of the expert panel who had already submitted their assessments also attended the preceptor's webinar, but did not submit their assessments a second time. These two expert panelists were invited to the webinar simply to answer questions and to contribute to the discussion if other preceptors asked questions during the webinar. They were also provided with 3.0 contact hours (0.3 continuing education units) CE credit for participating in the webinar.

As stated earlier, subcompetencies A and B were evaluated as separate subcompetencies and D and E were combined because of similar wording. Subcompetency A required students to synthesize complete patient history,

laboratory, and physical examination data to identify problems. Subcompetency B required students to identify and prioritize actual and potential drug-related problems, stating their rationale. Finally, subcompetencies D and E evaluated students' abilities to design and evaluate treatment outcomes using "pharmacokinetic and drug formulation data" (D) and "disease states and previous or current drug therapy . . . including psycho-social, ethical-legal, and financial data" (E). The expert panel was sent 6 case studies. Panelists were asked which case would best illustrate the 3 subcompetencies. The expert panel picked diabetes, heart failure, and anticoagulation for script development. The first 2 cases focus on chronic conditions and the third case offered an inpatient setting. Hypertension, chronic obstructive pulmonary disease, and pain management were not selected. The purpose of the video vignettes was to show different student behaviors for the preceptors to assess. Consequently, each vignette purposefully illustrated different levels of student clinical performance. The preceptors' prompts were the same regardless of the simulated student or the performance level illustrated. Preceptors then watched 9 video vignettes and scored the performances using rubrics with the following performance criteria: Level 1, excellent performance – the student requires minimal clinical supervision (beyond entry level) with simple to highly complex patients and is able to function in unfamiliar or ambiguous situations. The student's clinical reasoning is consistently proficient and shows knowledge and practice competence beyond entry level. The student's performance is timely and efficient. The student willingly assumes a leadership role for managing more difficult cases and is able to serve as a resource for others. The student actively contributes to the enhancement of the pharmacy with an expansive view of the profession; Level 2, entry-level performance – the student requires the expected degree of supervision for an entry-level pharmacist and shows entry-level knowledge and competence. The student consults with others and resolves unfamiliar or ambiguous situations. The student's clinical reasoning is consistently proficient in simple and complex cases. The student performs in a timely and efficient manner; Level 3, deficient (novice) performance – the student requires more than entry-level clinical supervision even with simple patients. Clinical performance is inconsistent and inefficient. Performance reflects little or no experience and is slow and inefficient.

The first 3 vignettes were presented in order from excellent to deficient, or from best to worst performance. However, a preceptor's assessment should be based on the performance criteria (ie, subcompetency description and the checklist) and not influenced by the performance of the previous student or the simulated student's own

performance on other subcompetencies during that interaction. In other words, a student could be competent in one subcompetency, but deficient in another subcompetency during the same interaction. Hence, to assess the potential for a halo effect, both the performance levels and subcompetencies were randomized in the 6 vignettes for the next 2 medical conditions. Both the expert panelists and preceptors were blinded.

The target number of participants was based on the ICC requirements to estimate inter-rater reliability²³ and the Fisher exact test to estimate accuracy.²⁴ Preceptors and video simulations were treated as random factors in a 2-way random effects ICC model. Since the 3 performance levels (eg, excellent, entry-level, and deficient) were hierarchical, and the purpose of the study was to assess preceptor's agreement on assignment to dichotomous categories, the preceptors' ratings were collapsed into two 2x2 tables for comparison. Preceptors' ratings of students as competent were compared to ratings as non-competent in one table and ratings of students' performances as excellent were compared to nonexcellent in the other 2x2 table. Higher ICC values reflected greater inter-rater reliability among preceptor assessments and described as poor (<0.00), slight (0.00–0.20), fair (0.21–0.40), moderate (0.41–0.60), substantial (0.61–0.80), and almost perfect (0.81–1.00).²⁵ Fisher exact test measured preceptors' accuracy and were compared with expert panel members' classification of competent or noncompetent scores with cells having fewer than 5 expected observations.²⁶ Similar to the ICC, the Fisher exact test treated collapsed data as nominal. The sample size was calculated to show if typical preceptors assessed students with the same accuracy ($\pm 10\%$) as the expert panel beyond chance alone. Data were analyzed using SPSS, v19.0 (IBM, Chicago, IL). The study protocol was approved by the Institutional Review Board at the University of Florida and conducted according to the principles of the Declaration of Helsinki.

RESULTS

The expert panelists' ages ranged from 30 to 66 years with a majority (57%) reporting 40 years or younger. Seventy-one percent of the participants were female and all (100%) held a doctor of pharmacy (PharmD) degree. All panelists had earned an advanced degree or a Board of Pharmacy Specialties (BPS) certification. The majority (86%) reported practicing 10 years or less. A majority (86%) reported precepting students for 10 years or less. Eighty-six percent indicated their primary role was preceptor, although one participant reported clinical coordinator. All participants (100%) were affiliated with public institutions. The largest group (71%) worked in hospitals.

Participants represented 5 states, and most respondents were from Alabama (29%) and Texas (29%).

The preceptors' age range was 26 to 61 years. Sixty-seven percent of the preceptors were female. Several participants possessed advanced degrees and certificates. The majority (53%) of them reported practicing pharmacy for 10 years or less. Eighty-four percent reported less than 10 years of experience precepting students. The majority (76%) indicated their primary role with students was preceptor, and 5% were educational coordinators. Sixty-nine percent were affiliated with public institutions, while the remainder were affiliated with private institutions. The largest group (41%) practiced in hospitals. Participants represented 5 states and the majority practiced in Florida (81%) and Texas (12%).

In the diabetes vignettes, each student's script was written so their performance was at the same level across all 3 subcompetencies, and the vignettes were delivered to the assessors in order from excellent (best) to deficient (worst). Student pharmacist Mary was intended to illustrate excellent performance for all 3 subcompetencies (Table 1). Meanwhile, student pharmacist Thomas was intended to illustrate entry-level performance, and student pharmacist Susan was intended to illustrate deficient performance for all 3 subcompetencies. For these nonrandomized vignettes,

the expert panelists' ratings were unanimous and matched the designated levels at the best and worst performance levels. However, the assessments were mixed for Thomas, the student illustrating entry-level performance, with some of the panelists rating his performance as excellent.

While the expert panelists were unanimous in their rating of Mary, just over a third of the preceptors rated her performance as entry-level (24%) or deficient (12%) for subcompetency A (Table 1). Other preceptors rated her performance as deficient for subcompetency B and subcompetency D/E, although the vignette's script was written to match criteria for excellent performance. For the most part, the preceptors rated Susan consistently as deficient (as intended). However, as with the expert panelists, the preceptors' ratings for Thomas were mixed. Thomas was rated as entry-level for the 3 subcompetencies by 45% to 55% of the preceptors. The simulated performance was rated higher than intended by 29% to 43% of the preceptors and lower (ie, deficient) by 26% to 36% of the preceptors.

The designated performance levels for the heart failure vignettes were randomized for each subcompetency. With regard to rating them correctly or incorrectly, the panelists' results were mixed (Table 2). For example, student pharmacist Patricia's video vignette was designed to demonstrate entry-level performance for subcompetency A.

Table 1. Panelists' Ratings of Student Pharmacists' Performances on Simulated Diabetes Clinical Interactions

		Excellent, n (%)		Entry Level, n (%)		Deficient, n (%)	
		Expert Panelists	Preceptors	Expert Panelists	Preceptors	Expert Panelists	Preceptors
Mary	Subcompetency A (Excellent)	7 (100) <i>p</i> value=0.45	27 (64)	0 (0)	10 (24)	0 (0)	5 (12)
	Subcompetency B (Excellent)	7 (100) <i>p</i> value=0.45	28 (67)	0 (0)	9 (21)	0 (0)	5 (12)
	Subcompetency D/E (Excellent)	7 (100) <i>p</i> value=0.53	29 (69)	0 (0)	9 (21)	0 (0)	4 (10)
Thomas	Subcompetency A (Entry)	3 (43) <i>p</i> value=0.15*	9 (21)	4 (57)	22 (52)	0 (0)	11 (26)
	Subcompetency B (Entry)	3 (43) <i>p</i> value=0.15*	8 (64)	4 (57)	23 (55)	0 (0)	11 (26)
	Subcompetency D/E (Entry)	2 (29) <i>p</i> value=0.06**	8 (19)	5 (71)	19 (45)	0 (0)	15 (36)
Susan	Subcompetency A (Deficient)	0 (0) <i>p</i> value=0.86	0 (0)	0 (0)	1 (2)	7 (100)	41 (98)
	Subcompetency B (Deficient)	0 (0) <i>p</i> value=0.73	0 (0)	0 (0)	2 (5)	7 (100)	40 (95)
	Subcompetency D/E (Deficient)	0 (0) <i>p</i> value=NA	0 (0)	0 (0)	0 (0)	7 (100)	42 (100)

N/A=one variable is a constant and association cannot be computed; ** Significant up to 10% difference between two groups, * significant up to 30% difference between 2 groups; *p* value is the probability of obtaining a difference in proportions between the preceptors and the expert panelists as extreme as the one observed beyond chance alone. Since there is no widely accepted criteria for an acceptable difference between the preceptors and panelists, the study notes *p* values at the 10% and 30% levels

While 29% rated her performance as entry level, 57% rated it as deficient and nearly 14% rated it as excellent. In both cases, 5 of the 7 panelists rated her performance incorrectly. Conversely, Patricia's performance on subcompetency D/E was designated as deficient, and all of the expert panelists rated it correctly. The expert panelists' ratings for scenarios for student pharmacists Joseph and Dorothy were also mixed. Notably, Joseph's a priori designations were excellent, deficient, and excellent. His performance was rated as entry-level by the highest proportion of panelists for the 2 subcompetencies designated as excellent, and an equal number of panelists rated his performance as deficient. For the subcompetency designated as deficient, most of the panelists rated it correctly. Dorothy's results showed ratings consistent with the designation for the 2 deficient subcompetencies. However, subcompetency B was designated a priori as excellent and 5 of the 7 panelists rated her performance as deficient, which was consistent with the other 2 performances panelists correctly rated as deficient.

The preceptors' assessments showed the same general pattern (Table 2). As with the expert panelists, most preceptors rated Patricia's performance as deficient on all 3 subcompetencies, even though 2 of the 3 subcompetencies were designated as entry-level a priori. The majority

of preceptors rated Joseph's performances as entry-level or deficient. In the subcompetency written to show Joseph as deficient, half rated him correctly, but the remainder rated him as entry-level or even excellent. In the 2 cases in which Dorothy was designated as deficient, nearly all of the preceptors rated her as deficient or entry-level. However, in the one subcompetency written to show her perform at an excellent level, only 10% of the preceptors rated her correctly and the majority rated her as deficient.

The expert panelists' assessments again showed mixed results for the third vignette on anticoagulation (Table 3). Student pharmacist Linda's scenario was written to demonstrate excellent, entry-level, and excellent performance on the subcompetencies. The largest proportion of panelists rated her performance as excellent for all 3 subcompetencies, including the one subcompetency designated as entry-level. With the exception of subcompetency A, entry-level was the next most frequently occurring rating. Student pharmacist David was designated as entry-level, excellent, and deficient on the 3 subcompetencies. David was rated as deficient most frequently in the subcompetency rated as entry-level, and he was rated most frequently as entry-level on subcompetency B, which was designated as excellent. He was rated correctly in most cases on the subcompetency designated as deficient. David was consistently rated lower than his

Table 2. Panelists' Ratings of Student Pharmacists' Performances on Simulated Heart Failure Clinical Interactions

		Excellent, n (%)		Entry Level, n (%)		Deficient, n (%)	
		Expert Panelists	Preceptors	Expert Panelists	Preceptors	Expert Panelists	Preceptors
Patricia	Subcompetency A (Entry)	1 (14) <i>p</i> value=0.62	4 (10)	2 (29)	15 (36)	4 (57)	23 (55)
	Subcompetency B (Entry)	0 (0) <i>p</i> value=0.51	4 (10)	3 (43)	11 (26)	4 (57)	27 (64)
	Subcompetency D/E (Deficient)	0 (0) <i>p</i> value=0.03**	1 (2)	0 (0)	17 (40)	7 (100)	24 (57)
Joseph	Subcompetency A (Excellent)	2 (29) <i>p</i> value=0.50	9 (21)	3 (43)	24 (57)	2 (29)	9 (21)
	Subcompetency B (Deficient)	0 (0) <i>p</i> value=0.26	6 (14)	2 (29)	15 (36)	5 (71)	21 (50)
	Subcompetency D/E (Excellent)	2 (29) <i>p</i> value=0.66	9 (21)	3 (43)	21 (50)	2 (29)	12 (29)
Dorothy	Subcompetency A (Deficient)	0 (0) <i>p</i> value=0.19*	2 (5)	1 (14)	15 (36)	6 (86)	25 (60)
	Subcompetency B (Excellent)	2 (29) <i>p</i> value=0.44	4 (10)	0 (0)	15 (36)	5 (71)	23 (55)
	Subcompetency D/E (Deficient)	1 (14) <i>p</i> value=0.23**	1 (2)	1 (14)	21 (50)	5 (71)	20 (48)

** Significant up to 10%, * significant up to 30%; N/A = one variable is a constant and association cannot be computed; *p* value is the probability of obtaining a difference in proportions between the preceptors and the expert panelists as extreme as the one observed beyond chance alone. Since there is no widely accepted criteria for an acceptable difference between the preceptors and panelists, the study notes *p* values at the 10% and 30% levels

designation, with the exception of the subcompetency designated as deficient. Finally, student pharmacist Barbara's subcompetencies were written to demonstrate deficient, deficient, and excellent performance. She was generally rated correctly by the panelists on the 2 deficient subcompetencies. However, the majority of panelists also rated her as deficient on the subcompetency, for which her performance was written to demonstrate excellence.

The majority of preceptors rated Linda correctly for subcompetency A, designed to illustrate an excellent performance (Table 3). The next most frequently occurring rating was entry-level; however, nearly 10 percent rated her excellent performance as deficient. In subcompetency B, designated as entry-level, nearly equal proportions of preceptors rated her performance as excellent or entry-level. David's ratings were about evenly distributed across the excellent, entry-level, and deficient ratings for subcompetency A, designed to illustrate entry-level performance and subcompetency B, designated as excellent. As with the expert panelists, preceptors rated him correctly in most cases on the subcompetency designated as deficient. David was consistently rated lower than his designation, with the exception of the subcompetency designated as deficient. For subcompetency D/E, those who did not rate him as deficient were about evenly split

among the excellent and entry-level performance ratings. In Barbara's case, the patterns in the preceptors' ratings were similar to those of the expert panelists. For the 2 subcompetencies designated as deficient, the majority of the preceptors rated her correctly. However, the majority rated her as deficient on the subcompetency, for which her performance was written to demonstrate excellence.

The aggregate ICC point estimates for subcompetencies A, B, and D/E were 0.37, 0.31, and 0.30, respectively, for competent vs noncompetent comparisons (Table 4), indicating fair inter-rater reliability with wide 95% confidence intervals (CI). Individual ICC point estimates for subcompetencies A, B, and D/E in all case scenarios ranged from 0.00 to 0.67 for competent vs noncompetent comparisons. The individual ICC point estimates had similarly wide 95% CI ranges. The aggregate ICC point estimates and 95% CIs for excellent vs entry-level comparisons were generally lower than the competent vs noncompetent comparisons (Table 4). The 0.24, 0.19, and 0.19 aggregate ICC point estimates indicated slight inter-rater reliability for subcompetencies A, B, and D/E, respectively. Individual ICC point estimates for subcompetencies A, B, and D/E in all case scenarios ranged from 0.00 to 0.54, representing poor to substantial reliability. The individual ICC point estimates had similarly wide 95% CI ranges.

Table 3. Panelists' Ratings of Student Pharmacists' Performances on Simulated Anticoagulation Clinical Interactions

		Excellent, n (%)		Entry Level, n (%)		Deficient, n (%)	
		Expert Panelist	Preceptor	Expert Panelist	Preceptor	Expert Panelist	Preceptor
Linda	Subcompetency A (Excellent)	6 (86) <i>p</i> value=0.55	25 (60)	0 (0)	13 (31)	1 (14)	4 (10)
	Subcompetency B (Entry)	3 (43) <i>p</i> value=0.26*	19 (45)	2 (29)	18 (43)	2 (29)	5 (12)
	Subcompetency D/E (Excellent)	3 (43) <i>p</i> value=0.38	17 (40)	2 (29)	18 (43)	2 (29)	7 (17)
David	Subcompetency A (Entry)	1 (14) <i>p</i> value=0.18*	11 (26)	2 (29)	18 (43)	4 (57)	13 (31)
	Subcompetency B (Excellent)	2 (29) <i>p</i> value=0.30*	11 (26)	4 (57)	17 (40)	1 (14)	14 (33)
	Subcompetency D/E (Deficient)	0 (0) <i>p</i> value=0.34	5 (12)	1 (14)	8 (19)	6 (86)	29 (69)
Barbara	Subcompetency A (Deficient)	0 (0) <i>p</i> value=0.62	0 (0)	1 (14)	8 (19)	6 (86)	34 (81)
	Subcompetency B (Deficient)	0 (0) <i>p</i> value=0.18*	2 (5)	0 (0)	8 (19)	7 (100)	32 (76)
	Subcompetency D/E (Excellent)	0 (0) <i>p</i> value=0.54	2 (5)	2 (29)	13 (31)	5 (71)	27 (64)

** Significant up to 10%, * significant up to 30%; N/A=one variable is a constant and association cannot be computed; *p* value is the probability of obtaining a difference in proportions between the preceptors and the expert panelists as extreme as the one observed beyond chance alone. Since there is no widely accepted criteria for an acceptable difference between the preceptors and panelists, the study notes *p* values at the 10% and 30% levels

Table 4. Intraclass Correlation Comparing Competent versus Noncompetent and Excellent vs Entry-Level Simulations

Skill/Scenario	Competent vs Noncompetent			Excellent vs Entry-Level		
	ICC	95% CI	<i>p</i> value	ICC	95% CI	<i>p</i> value
Skill A ^a	0.37	0.20-0.69	0.00*	0.24	0.11-0.55	0.00*
Diabetes	0.66	0.33-0.99	0.00*	0.54	0.22-0.98	0.00*
Heart Failure	0.15	0.03-0.88	0.00*	0.10	0.01-0.84	0.00*
Anticoagulation	0.46	0.17-0.97	0.00*	0.26	0.07-0.93	0.00*
Skill B ^a	0.31	0.16-0.63	0.00*	0.19	0.09-0.49	0.00*
Diabetes	0.63	0.30-0.99	0.00*	0.49	0.19-0.98	0.00*
Heart Failure	0.00	-0.02-0.45	0.37	0.00	-0.02-0.42	0.40
Anticoagulation	0.37	0.12-0.96	0.00*	0.24	0.06-0.93	0.00*
Skill D/E ^a	0.30	0.15-0.62	0.00*	0.19	0.09-0.48	0.00*
Diabetes	0.67	0.34-0.99	0.00*	0.52	0.21-0.98	0.00*
Heart Failure	0.07	0.01-0.78	0.01*	0.02	-0.01-0.61	0.12
Anticoagulation	0.29	0.08-0.94	0.00*	0.18	0.04-0.90	0.00*

ICC=Intraclass Correlation

* *p* value indicates significant ICC results

^a aggregate ICC per subcompetency

In nearly two-thirds of the rating opportunities, a higher proportion of the expert panelists than preceptors rated the student pharmacists' performance correctly (18 of 27 scenarios). Conversely, a higher proportion of preceptors than expert panelists correctly rated the students' performances in only 30 percent of the rating opportunities (8 of 27). The proportion of correct and incorrect ratings was the same for both the expert panelists and the preceptors for only one subcompetency in one scenario. The expert panelists rated student performances correctly more often than the preceptors in every instance on the diabetes scenarios and for 6 of the 9 rating opportunities presented in the heart failure scenarios. The preceptors rated student performances correctly for 5 of the 9 rating opportunities presented in the anticoagulation scenarios. These trends further support concerns about inter-rater inconsistency, even when only classifying a student's performance as competent or noncompetent.

The diabetes simulations, presented first and from the highest to lowest, had the highest inter-rater reliability estimates and smallest 95% CIs for all 3 subcompetencies. The average ICC for the 3 subcompetencies was 0.65. The average ICC for the 3 randomized subcompetencies in the heart failure scenarios was 0.07 and was generally poor. The ICC results increased slightly with the anticoagulation scenarios, although the average was still considered only fair (0.37). This trend was consistent for the excellent vs entry-level comparisons. Both expert panelists and preceptors assessed student performances in the diabetes vignettes, delivered from best to worst, with a greater degree of reliability compared to the heart failure and anticoagulation vignettes, which were delivered in a random order.

DISCUSSION

In general, the inter-rater reliability and rater accuracy were poor in our study, although not different from most other health professions.^{15,16,20} Several factors may have contributed to inter-rater inconsistency or contributed to assessments that differed from the intended performance level(s). These factors include limited number of observations, halo effect, weak manipulations in the "videotaping" and scripts (although they reflected real-life scenarios), inadequate understanding of the performance criteria, and diversity in rater expertise. Several panelists and preceptors commented that they struggled to make accurate and fair assessments based on a single-case presentation. Often during a student's experiential placement, preceptors will have multiple opportunities to observe the student's performance. So, while they might not "get it right" every time in individual observations, preceptors' overall impressions of a student's competence is assumed to be accurate over multiple observations.

Each opportunity to evaluate a student's competence should be considered independent from all other opportunities. However, preceptors do not look at them independently, and their opinions about previous performance(s) roll over to the next one. This phenomenon is known as the halo effect and occurs when a preceptor's general impression (eg, affective like or dislike) influences their ratings of the individual student's performance,²⁷ the preceptor's current assessment is influenced by their previous assessments of the student or their implicit or explicit comparisons of the student with other students' performances,²⁸ or there is a failure by the preceptor to discriminate among distinct facets of the student's competence.²⁹ Therefore, the preceptor can unconsciously or consciously

fail to discriminate between situations in which the student was competent and situations where improvement was needed to meet the minimum standard under each of these 3 circumstances.^{27,30} Anecdotal findings in our study hinted that the halo effect might have been influential among the preceptors' variable ratings. More than one panelist and preceptor openly acknowledged that previous students' performances influenced their rating of other students or the same student's earlier performance influenced subsequent ratings. In other cases, preceptors' comments suggested that a student's performance on one subcompetency in the same vignette influenced their rating of that student on another subcompetency. Accurate and consistent rating requires that each subcompetency is rated based on its own performance criteria and not from a gut feeling or the fact that a student was previously successful. Comments like "giving the student a break" and "[student name] was doing so well" suggest presence of the halo effect and reduces the veracity of the assessments.³¹

Presentation order of the vignettes also supported the halo effect as a plausible reason for inconsistent ratings. The ICCs were highest for the diabetes vignettes, but those findings could largely be a result of performance levels being the same for all 3 subcompetencies in each vignette and being presented in order from highest to lowest competence for all 3 vignettes. Hence, the halo effect may have artificially increased inter-rater reliability because of the order and consistency of the performance within each vignette. However, real life isn't like that. When the performance levels were mixed for the other 2 scenarios, as we might find in authentic clinical assessment situations, the inter-rater reliabilities were much lower. Moreover, when students' performances weren't classified correctly, preceptors were more likely to rate a performance the same as they rated the previous subcompetency vignette. For example, for Barbara, the first 2 subcompetencies were written to demonstrate a deficient or novice performance. However, subcompetency D/E was written to reflect excellent performance; however, the majority of the preceptors and panelists rated that performance as deficient. Preceptor training programs should provide mechanisms to reduce subjective assessments of objectively observable events and improve detection and resolution of halo effects and leniency. For example, school-sponsored programs should train preceptors on the use of performance checklists and provide them training on the checklists' optimal use.³² The best way to avoid leniency and halo effects is to have multiple raters.³³ However, this luxury is often not possible in experiential programs. So, raising preceptor awareness and teaching strategies to detect and reduce halo effect and leniency is warranted.³⁴

When using the halo effect as an explanation for inconsistent ratings, the rater presumably knows the criteria but applies them incorrectly. Another plausible reason for inconsistent assessments is a lack of understanding of the performance criteria²⁹ as a result of little or no training. However, even minimal training may not be sufficient to provide a clear understanding of the assessment instructions.³⁵ For example, we provided preceptors with a 20-minute introduction to assessment principles and to the performance criteria prior to scoring the video vignettes, which was likely insufficient because only about half the students were rated correctly. Preceptors' own anecdotal comments implied they reverted back to previously used heuristics or substituted their own personal standards after this short training program. When preceptors resort to using their own standards of competence rather than the performance criteria, it often results in different preceptors applying different positive or negative assessments for exactly the same performance. For example, one preceptor chastised a student for "poor medication decisions," while another preceptor praised the same student for "... confident recommendations and follow-up plan." Another example of using contradictory personal standards resulting in differing assessments can be seen with this pair of comments: "... the student didn't have all answers and seemed unsure of some things," and "... willing to say, 'I don't know' and check it out and not make false statements." With more comprehensive training, these types of assessment inconsistencies may be reduced by ensuring preceptors consistently use the performance criteria rather than their own practice standards.

The second objective of this study was to compare the assessment results of typical volunteer faculty members with expert opinions. Ideally, the expert panelists' ratings would have been unanimous in every scenario. However, only 12 out of 27 video vignettes produced unanimous agreement or had only a single dissenting panelist. These 12 ratings occurred at the extremes of the performance levels (eg, either excellent or deficient). Even so, the expert panelist ratings were more accurate than preceptor ratings by a nearly 3 to 1 margin. Overall, preceptors' assessments were more lenient than the expert panelists' assessments; specifically, expert panelists only rated 10% of deficient scenarios as competent whereas preceptors rated 27% of deficient scenarios as competent.

Another reason for the rating inconsistencies may be explained by the fact that preceptors and expert panelists may have used different standards of care or clinical databases. Current clinical skills are *sine qua non* for competent clinical assessment. Students in the vignettes who made their recommendations based on current guidelines were sometimes rated as competent; other times, preceptors

rated the same student as deficient. For example, one preceptor reported anecdotally that Mary's performance was excellent; however, the preceptor disagreed with the drug therapy she recommended and rated her as deficient. Preceptors also offered negative comments about Thomas's medication and dose recommendations. In both of these vignettes, the students' recommendations were consistent with nationally accepted treatment guidelines. Poorly performing clinicians are less able to accurately assess others because they may not be aware of their own deficiencies, despite high levels of self-confidence.^{29,36,37} Unfortunately, if administrators and faculty members responsible for experiential programs are focusing their preceptor training on "how to teach" in the clinical environment, they may have only managed to train preceptors to teach the wrong content well if the preceptors' clinical skills are subpar. So, experiential programs need to ensure their preceptors' clinical skills are up-to-date by facilitating preceptor-led clinical learning communities, coordinating peer-to-peer clinical competency evaluations, or inculcating the responsibility for continuing professional development as part of their obligation as professional teachers. While none of these strategies will completely eliminate rater errors, ongoing systematic training reduces inconsistency, improves accuracy, and reduces variability.^{5,6,38,39}

Next, study participants' anecdotal comments suggested that standardized video simulations could be effective tools, although widespread extension of this technology presents challenges. First, each video vignette provided preceptors with a single opportunity to observe a student's performance. However, in actual practice, preceptors routinely make assessments based on multiple observations of a student spanning the entire clinical practice experience. Therefore, video vignettes portraying multiple interactions would make the assessment more authentic.⁴⁰ Paradoxically, the academic clinical assessment literature has little evidence that guides summative assessment of multiple observations.³⁸ Guidelines designed to help preceptors give a single consistent and accurate summative assessment for a single or multiple subcompetencies based on multiple observations are in need of rigorous study. Future work might establish cut-off guidelines for coming to a summary evaluation based on multiple observations.

Next, given the implication of authentically depicted clinical practice in videotaped simulations, their production is an important consideration. Two facets of video production are important to the overall quality: validity of the scripting and the technical production. First, the vignette scripts were scientifically and clinically sound. However, a subset of the preceptor raters still stated they

had difficulty identifying when activities for one subcompetency were completed and activities for another subcompetency started in the videos. Similarly, others opined it was difficult to recognize subtle differences in performance. On the other hand, another subset of preceptors commented that they were able to observe different levels of performance between subcompetencies and that the portrayals of the student-preceptor interactions were more realistic than expected. The point is that these sets of preceptor comments were about the same vignettes. This diversity of comments on the same vignettes raises the question of whether the issue was with the scripts or whether comments were representative of the challenges of authentic clinical competency assessment. We believe the latter was the case for two reasons. First, participating preceptors were most consistent and accurate in their classifications of the best and the worst performances, which, by definition, illustrated the extremes in performance and should have been the easiest to categorize correctly. Across the board, expert panelists and preceptors had the most difficulty with entry-level performance, which, by definition, contained elements of the best and worst performances. Second, with a more homogenous presentation (eg, in best-to-worst order and consistent performance level within a vignette), the preceptors' assessments were far more consistent and accurate. In short, the preceptors' difficulties seemed to be a vocalization of the daily challenges faced by preceptors in consistently and accurately evaluating student performance.

This brings us to a third consideration based on the results of this study. Performance criteria for APPEs based on universal professional standards are core to valid assessment. While this requirement is well known, national policies or guidelines outlining acceptable validation criteria for APPE assessment instruments have yet to be established. For example, while CAPE defines the educational outcomes expected of an entry-level practitioner, the level of a student's performance where a preceptor would categorize a student as competent or incompetent is not described in the document or in any other standard. If no performance standards exist then, by necessity, raters will apply their own professional judgments to student performance. Given the lack of policies and guidance and the number and variety of instruments, preceptors may assess students differently depending on which assessment instrument is used. The net effect is significant variation between preceptors as their assessments are not grounded in professional standards.

In pharmacy education, a nationwide effort similar to that in physical therapy may be needed to improve APPE clinical assessment. However, the development and maintenance of assessment instruments is expensive

and consumes substantial resources, whether done individually or in conjunction with other schools. A national effort to develop and implement a standardized, validated instrument would reduce individual schools' costs by distributing costs across schools. Even though it would be expensive, hidden costs are associated with inconsistent and inaccurate assessments that need to be considered. Moreover, a national assessment tool would encourage development of a standardized national program to train, assess, and improve preceptors' ability to evaluate student competency and readiness to enter into practice, similar to the American Physical Therapy Association (APTA) program. In physical therapy, both students and preceptors are required to complete a standardized online training program developed and sponsored by APTA, which uses video simulations prior to practice experiences. This requirement was based on evidence that showed even experienced preceptors improved with the training.⁴¹

Our study had several limitations. Inter-rater reliability is dependent on the population of preceptors, the number of student-rating opportunities, and the rating instrument. Our convenience sample may not have represented the demographics of the typical US pharmacy school preceptor faculty members. The impact of these discrepancies on the validity and generalizability of the results is uncertain. Future validation studies should recruit participants that represent demographic characteristics generalizable to the schools using the instrument. Next, on average, the 27 video vignettes took over 70 minutes to view. Although panelists and preceptors were able to view and rate the videos at their convenience online and receive CE credit, this is a large number of video vignettes to watch and assess. Therefore, it still may have been a burden for a group of busy professionals and may have influenced the results. Next, we used a college of pharmacy's students, faculty members, and clinical practice laboratory as a matter of convenience to produce the vignettes. It is possible that some of the difficulties mentioned above were a result of amateur production of the videos. Some of the inconsistencies and inaccuracies among the preceptors' ratings could have been improved by professional production.

Another potential limitation of the study's design was that our expert panelists edited the video scripts and designated behaviors and performance levels that determined whether students' performances portrayed in the videos were excellent, entry-level, or deficient. While it is possible that the panelists recalled their own performance classifications from editing specific scenarios (test-retest consistency), individual panelists could not possibly have "recalled" the other panelists' responses and, as a result, increased the inter-rater consistency. Second, the panelists were not classified post hoc as experts in this study simply because they helped shape the scripts. Rather, they

were selected to shape the scripts because they were already acknowledged expert preceptors with considerable clinical expertise. In rating student performance, experts' ratings are more accurate and consistent than nonexperts' ratings; increasing inter-rater consistency.^{7,8} Next, accuracy of recall is likely diminished by both a greater time interval between the first event and the second event^{42,43} and the degree of detail required by the rating.^{44,45} For editing the vignette to actually give the expert panelists an advantage, they would have needed to accurately recall their performance rating from the first event (ie, editing vignette) and apply it to the second event (ie, rating based on the video portrayal). However, several months intervened between the time the panelists' edited the scripts and the time they viewed and assessed students' performances on the videos. Therefore, we posit that panelists were more likely to rely on their clinical expertise to rate student performance on the videos than to rely on memories of their classifications from the script editing. Moreover, the expertise explanation for higher inter-rater consistency is more likely than the recall hypothesis given the complexity of the task and the amount of detail required to rate student performance. In this study, experts were asked to rate 9 scenarios on 3 subcompetencies requiring a total of 27 ratings. While they may have remembered elements of their decisions and the gist of reasons for their ratings during the editing process, accurate recall of all 27 ratings several months later was unlikely and was a potential source of error, lowering test-retest consistency.⁴⁵ So, given the amount of time between the scenario editing and the complexity of evaluating clinical performance, we propose the expert panelists had little real advantage over the nonexpert preceptors as a result of being on the Delphi panel several months before.

On the other hand, this hypothetical study design limitation does provides us with some guidance on the primary question of this study, which asked whether multiple raters viewing the same student-preceptor interaction would all assess the student's performance the same. Under what was arguably ideal conditions, even the experts didn't rate the students' performances the same in every case. Hence, a reality of clinical performance assessment may be that there is a ceiling effect on assessment accuracy and consistency, especially in situations where the students do some things well and other things not so well vs doing things in extreme terms of excellent and deficient. However, while their inter-rater consistency was not perfect, experts' accuracy and consistency were higher than nonexperts' accuracy and consistency, which leads us to believe that training to increase preceptors' expertise would improve student assessments, although perfection may be to be too lofty a goal.

CONCLUSION

Preceptors' assessments of student competencies in APPEs is replete with challenges regarding consistency and accuracy. Preceptor assessments of clinical competency were inconsistent and inaccurate using standardized video vignettes, even with rubric-based response categories. Although consistency and accuracy improved with preceptor expertise, halo effect, leniency, and preceptor knowledge base were potential sources of rater error. Our findings support calls by pharmacy leaders, such as the 2008 AACP president, for a rigorous and validated standardized APPE assessment instrument.^{3,4} Even so, pharmacy education professionals have yet to establish standards for experiential assessment or to outline acceptable validation criteria for APPE assessment instruments. Without rigorous evaluation, there is no assurance that any existing or future proprietary APPE assessment instruments would be capable of assessing student performances meaningfully. Valid clinical competency assessment needs to unequivocally predict whether students are capable of performing in real-world clinical settings.⁴⁶ A national assessment instrument supported by strong evidence of validity would significantly advance the competency-based educational goals of the profession.

REFERENCES

1. Harden RM. Developments in outcome-based education. *Med Teach*. 2002;24(2):117-120.
2. Schuwirth LWT, Southgate L, Page GG et al. When enough is enough: a conceptual basis for fair and defensible practice performance assessment. *Med Educ*. 2002;36(10):925-930.
3. Raehl CL. AACP Pharmacy Education Assessment Services: Outcomes, assessment, accountability. *Am J Pharm Educ*. 2008; 72(1):Article 13.
4. CAPE Advisory Panel on Educational Outcomes. Center for the Advancement of Pharmaceutical Education Educational Outcomes, Revised Version 2004: Report of the 2005 American College of Clinical Pharmacy Educational Affairs Committee. 2006.
5. Landy FJ, Farr JL. Performance rating. *Psychol Bull*. 1980;87(1):72-107.
6. Holmboe ES, Hawkins RE. *Practical Guide to the Evaluation of Clinical Competence*. Philadelphia, PA: Mosby/Elsevier; 2008.
7. Bondy KN. Criterion-referenced definitions for rating-scales in clinical evaluation. *J Nurs Educ*. 1983;22(9):376-382.
8. Lankshear AJ. The use of focus groups in a study of attitudes to student nurse assessment. *J Adv Nurs*. 1993;18(12):1986-1989.
9. Marrelli T, Tondora J, Hoge MA. Strategies for developing competency models. *Adm Policy Ment Health*. 2005;32(5-6):533-561.
10. Jette DU, Berton A, Coots R, Johnson H, McLaughlin C, Weisbach C. Clinical instructors' perceptions of behaviors that comprise entry-level clinical performance in physical therapist students: a qualitative study. *Phys Ther*. 2007;87(7):833-843.
11. Cross V, Hicks C. What do clinical educators look for in physiotherapy students? *Physiotherapy*. 1997;83(5):249-260.
12. Alexander HA. Physiotherapy student clinical education: The influence of subjective judgments on observational assessment. *Assessment & Evaluation in Higher Education*. 1996;21(4):357-366.
13. Streiner DL, Norman GR. *Health Measurement Scales: A practical guide to their development and use*, 2nd ed. Oxford, England: Oxford University Press; 1995.
14. Dauphinee WD. Assessing clinical performance: Where do we stand and what might we expect? *JAMA*. 1995;274(9):741-743.
15. Noel GL, Herbers JE, Caplow MP, Cooper GS, Pangaro LN, Harvey J. How well do internal medicine faculty members evaluate the clinical skills of residents? *Ann Intern Med*. 1992;117(9):757-765.
16. Ried LD, Nemire R, Doty R, et al. An automated competency-based student performance assessment program for advanced pharmacy practice experiential programs. *Am J Pharm Educ*. 2007; 71(6):Article 128.
17. Douglas CA. Towards an operational definition of pharmacy clinical competency. PhD Dissertation. Gainesville, FL: University of Florida, 2011.
18. Roach K, Gandy J, Deusinger SS, et al. The development and testing of APTA clinical performance instruments. *Phys Ther*. 2002;82(4):329-353.
19. Dillman DA, Smyth JD, Christian LM. *Internet, Mail, and Mixed Mode Surveys: The Tailored Design Method*. Hoboken, NY: John Wiley and Sons, Inc. 2009.
20. Survey Monkey [computer program]. Palo Alto, California, USA SurveyMonkey.com, LLC., 2009
21. Powell C. The Delphi technique: myths and realities. *J Adv Nurs*. 2003;41(4):376-382.
22. Winters J, Hauck B, Riggs CJ, Clawson J, Collins J. Use of videotaping to assess competencies and course outcomes. *J Nurs Educ*. 2003;42(10):472-476.
23. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86(2):420-428.
24. Siegel S, Castellan Jr, NJ. *Nonparametric Statistics for the Behavioral Sciences*. 2nd ed. New York, NY: McGraw-Hill, Inc.; 1988.
25. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159-174.
26. Agresti A. *An Introduction to Categorical Data Analysis*. 2nd ed. Hoboken, NJ: John Wiley & Sons, Inc.; 2007.
27. Linn RL, Gronlund NE. *Measurement and Assessment in Teaching*. 8th ed. Upper Saddle River, NJ: Prentice Hall.; 2000.
28. Robbins SP. *Organizational Behavior*. 4th ed. Englewood Cliffs, NJ: Prentice Hall; 1989.
29. Longford NT. Reliability of essay rating and score adjustment. *J Educ Behav Stat*. 1994;19(3):171-200.
30. Saal FE, Downey RG, Lahey MA. Rating the ratings: assessing the psychometric quality of rating data. *Psychol Bull*. 1980; 88(2):413-428.
31. Motycka CA, Rose RL, Ried LD, Brazeau G. Self-assessment in pharmacy and health science education and professional practice. *Am J Pharm Educ*. 2010;74(5):Article 85.
32. Schmutz J, Eppich WJ, Hoffmann F, Heimberg E, Manser T. Five steps to develop checklists for evaluating clinical performance: An integrative approach. *Acad Med*. 2014;89(7):996-1005.
33. Thomas MR, Beckman TJ, Mauck KF, Cha SS, Thomas KG. Group assessments of resident physicians improve reliability and decrease halo error. *J Gen Intern Med*. 2011; (26)7:759-64.
34. Woehr DJ, Huffcutt AI. Rater training for performance appraisal: a quantitative review. *J Occup Organ Psychol*. 1994; 67(3):189-205.
35. Cross V, Hicks C, Barwell F. Exploring the gap between evidence and judgment: using video vignettes for practice-based assessment of physiotherapy undergraduates. *Assessment & Evaluation in Higher Education*. 2001;26(3):189-212.

36. Eva KW, Cunnington JPW, Reiter HI, Keane DR, Norman GR. How can I know what I don't know? Poor self-assessment in a well-defined domain. *Adv Health Sci Educ Theory Pract*. 2004;9(3):211-224.
37. Davis DA, Mazmanian PE, Fordis M, Van Harrison R, Thorpe KE, Perrier L. Accuracy of physician self-assessment compared with observed measures of competence: a systematic review. *JAMA*. 2006;296(9):1094-1102.
38. Beck DE, O'Sullivan PS, Boh LE. Increasing the accuracy of observer ratings by enhancing cognitive processing skills. *Am J Pharm Educ*. 1995;59(3):228-235.
39. Bondy KN. Clinical evaluation of student performance: the effects of criteria on accuracy and reliability. *Res Nurs Health*. 1984;7(1):25-33.
40. Linssen T, van Dalen J, Rethans J-J. Simulating the longitudinal doctor-patient relationship: experiences of simulated patients in successive consultations. *Med Educ*. 2007;41(9):873-878.
41. Vendrely A, Carter R. The influence of training on the rating of physical therapist student performance in the clinical setting. *J Allied Health*. 2004;33(1):62-69.
42. Coughlin SS. Recall bias in epidemiologic studies. *J Clin Epidemiol*. 1990;43(1):87-91.
43. Sudman S, Bradburn NM. Effects of time and memory factors on response in surveys. *J Am Stat Assoc*. 1973;68(344):805-815.
44. Gaugler BB, Thornton GC. Number of assessment center dimensions as a determinant of assessor accuracy. *J Appl Psychol*. 1989; 74(4):611-618.
45. Guerin SA, Robbins CA, Gilmore AW, Schacter DL. Retrieval failure contributes to gist-based false recognition. *J Mem Lang*. 2012;66(1):68-78.
46. Whitcomb ME. Competency-based graduate medical education? Of course! But how should competency be assessed? *Acad Med*. 2002;77(5):359-360.