GPHE 102517—8/12/2004——129410

Pharmacy Education, Month 2005; 00(0): 1–12

Reliability, validity, and generalizability of an objective structured clinical examination (OSCE) for assessment of entry-to-practice in pharmacy

LILA QUERO MUNOZ[1], CAROL O'BYRNE[1], JOHN PUGSLEY[1], & ZUBIN AUSTIN[1,2]

[1]*Pharmacy Examining Board of Canada, Toronto, Ont., Canada M5B 2E7, and* [2]*Leslie Dan Faculty of Pharmacy, University of Toronto, 19 Russell Street, Toronto, Ontario, Canada M5S 1A1*

Abstract

This paper describes the evaluation of an objective structured clinical examination (OSCE) and the assessment outcomes for reliability, validity and generalizability for the entry-to-practice context in pharmacy in Canada. A total of 190 participants were involved: 153 entry-to-practice candidates and 37 pharmacists who were already licensed. Two balanced forms of an OSCE were developed, consisting of 26 stations (18 interactive and 8 non-interactive stations). Descriptive analysis for all data was undertaken, and detailed analysis of data from Form I of the OSCE (including generalizability and dependability studies) are reported. Based on findings of this study, conclusions were made regarding OSCEs for entry-to-practice assessment in pharmacy. A key finding of this study was that a 15-station OSCE, using one pharmacist-assessor per station, yielded consistent and dependable scores when holistic scoring was used to assess both qualifying candidates and practising pharmacists.

Keywords: *OSCE, assessment, pharmacy education, clinical skills evaluation, performance assessment*

Background

The objective structured clinical examination (OSCE) is a recognized, valid and reliable method for assessment of integration of knowledge, skills and abilities (Sloan, Donnelly, Schwartz and Strodel, 1995, Colliver and Swartz, 1997). In general, OSCEs consist of a series of stations through which all candidates rotate on a timed basis. Each station depicts a situation in practice that is either commonly encountered or critical in nature. The candidate is required to perform specific functions to complete the task or address the problem depicted in the simulation. An assessor is present in most stations to provide real-time, direct observation and assessment of each candidate's performance. By virtue of its performance-based nature, the OSCE is well-suited to complement traditional paper-based, computer-based, or other testing methods (Robb and Rothman, 1985) and, together with these other methods, provides a more robust vehicle for assessing competency. OSCEs are now an accepted part of the education, testing and certification of students and candidates in a wide variety of health professions, including medicine (Barrows, 1993, Dupras and Li, 1995, Singer, Robb, Cohen, Norman and Turnbull, 1996, Grand'Maison, Brailovsky, Lescop and Rainsberry, 1997), pharmacy (Fielding Page, Fevang and Thomas, 1997), and other health professions (Woodburn and Sutcliffe, 1996)

OSCEs demonstrate particular advantages over traditional forms of testing (such as multiple choice tests), in assessing communication and interpersonal skills, professional judgment and moral/ethical reasoning. As a result, momentum has been building within several health professions to expand the role of the OSCE from beyond education to assessment of professional competency at the entry-to-practice level and for assessment of continuing competency for practitioners (Benson, 1991, Fielding Page, Fevang and Thomas, 1992, Campbell, Parboosingh and Slotnick, 1999, Fielding Page, Fevang and Thomas, 2001).

Correspondence: Z. Austin, Leslie Dan Faculty of Pharmacy, University of Toronto, 19 Russell Street, Toronto, Ont. M5S 1A1. Tel: 1 416 978 0186. E-mail: zubin.austin@utoronto.ca

ISSN 1560-2214 print/ISSN 1477-2701 online © 2005 Taylor & Francis Ltd
DOI: 10.1080/15602210400025347

Within the health professions in Canada, medicine, pharmacy and physiotherapy have been pioneers in the use of entry-to-practice OSCEs. In these disciplines, the OSCE has been used (in conjunction with other testing formats), to assess application and integration of professional knowledge and skills. Entry-to-practice OSCEs typically consist of a series of timed (usually between five and ten minute) stations through which each candidate must rotate sequentially. Individual OSCE stations may consist of both interactive and non-interactive stations. The former, typically involving "simulated clients" (actors who are specifically trained to portray patients with medical conditions or needs, or other health professionals), and the latter typically involving written responses to tasks or problems.

The inclusion of an OSCE component in the entry-to-practice examination for pharmacists in Canada has been evolving for many years. In the 1970s, the College of Pharmacists of British Columbia (CPBC, the regulatory body for pharmacists in Canada's third largest province) first introduced the OSCE format as a method of evaluating entry-to-practice and continuing competency of its members (Fielding Page, Fevang and Thomas, 1981). In 1996, the Ontario College of Pharmacists (OCP, the regulatory body for pharmacists in Canada's largest province) instituted an OSCE as part of its compulsory Quality Assurance and Practice Review program (Austin, Croteau, Marinia and Violato, 2003).

The Pharmacy Examining Board of Canada (PEBC), works on behalf of participating provincial regulatory authorities to evaluate qualifications, knowledge, skills and competencies of candidates seeking licensure as pharmacists. The qualifying examination (QE) is a pre-registration requirement in all Canadian provinces (except Quebec). Traditionally, the QE has been a multiple-choice assessment of applied knowledge and practice-based problem solving skills. In response to changes occurring in professional practice and the requirement for pharmacists to demonstrate patient-care competencies, momentum developed for the inclusion of a performance-based direct assessment of candidates for licensure. In 1995, based on the experience of using OSCEs in other professions and other contexts within pharmacy, PEBC began to explore the value and feasibility of incorporating an OSCE in the PEBC qualifying examination.

Pharmacists' roles and scope of practice considerations

Within the context of regulated health professions, performance-based assessment of skills and applied knowledge is often seen as a critical requirement for ensuring public protection and the licensure of safe and effective practitioners. Performance-based assessment is intimately linked to scope of professional practice; the performance being assessed must reflect and represent the practice of the profession. By definition, regulated health professions are limited in their scopes of practice by a variety of regulations. Such limitations ensure that professional scopes of practice correlate with societal expectations, and serve to protect the public from unskilled health care workers.

Unlike other health professions (including medicine, physiotherapy and nursing), pharmacists (in general) are neither required nor expected to demonstrate a broad repertoire of psycho-motor skills related to patient assessment and therapy. Though an emerging part of professional practice, the use of physical assessment skills (such as inspection, palpation, percussion or auscultation) is generally limited and in some circumstances prohibited. Instead, performance within the profession of pharmacy is more highly focused on verbal and non-verbal communication, observation, and review and management of patient information and medical information databases, for the purpose of identifying and resolving clients' drug related problems or other health care needs. Thus, much of a pharmacist's practice involves questioning, listening, observing and problem-solving.

Pharmacists are also increasingly being relied upon to work collaboratively with patients and other practitioners to ensure appropriate prescribing and use of medications by patients. Use of structured patient interviewing techniques to gather health and medication histories, make assessments, and solve problems, is a central task for pharmacists. Communication techniques such as consultation, teaching and knowledge translation to make scientific drug information relevant for patients are among the most important tools pharmacists use in their professional practice.

In Canada, the National Association of Pharmacy Regulatory Authorities (NAPRA, an umbrella organization representing participating provincial regulatory bodies) has described competencies expected of pharmacists at entry-to-practice in a document entitled "Professional Competencies for Canadian Pharmacists at Entry to Practice".

While drug distribution, dispensing, supervision and management of technical support staff are described, substantial emphasis is placed on communicative competencies, integrated knowledge and skills necessary to support patient care and interprofessional collaboration. Specific competencies related to effective use of verbal and non-verbal communications are described in the context of the pharmacist's primary role of assisting patients and other health care providers to identify and resolve potential or actual medication-related problems and promoting health and wellness.

By their very nature, assessment of communicative and cognitive skills may be somewhat more subjective and culturally-bound, and the reliability of direct assessment of these skills may be questioned (Abedi, 2004). In developing assessment tools, the psychometric stability, reliability and generalizability of instruments are of significant concern. Of equal concern is the need to ensure that tasks and activities within an OSCE are reflective of professional standards, expectations and practice.

OSCE development and field testing

Building upon the tradition of performance-based assessment already in place in Canadian pharmacy practice, the Pharmacy Examining Board of Canada, and the College of Pharmacists of British Columbia collaborated in the development and field testing of an entry-to-practice OSCE for pharmacy in Canada.

Given the importance of ensuring that entry-level pharmacists are capable of safe, effective, independent practice, priority was first given to development of a process for determining and weighting critical competencies for direct assessment through an OSCE. Various processes were considered, including statistical sampling techniques, top–down decision making, or majority-voting, and in turn each was rejected as being insufficiently inclusive or representative. Since the prioritization and weighting of competencies would *de facto* define the essence of professional practice, it would be inappropriate for any single group of regulators, academics, or practitioners in isolation to make this decision. Instead, a steering committee, composed of representatives from various constituencies within the profession of pharmacy was convened and worked through a consensus building exercise to define the key functions and activities to be assessed. Through this process, agreement was reached and the decision was made to assess communication skills in general, and the following five pharmacy-specific functions and activities specifically:

- Gather, assess, and interpret information from clients (patients, physicians, health care providers), and other sources.
- Select and recommend appropriate therapeutic options
- Communicate and educate effectively
- Prepare and distribute medications and drug-delivery devices
- Exercise professional and ethical judgments

While these functions and activities provided a general framework for OSCE stations, additional content specifications were required to ensure that specific stations comprising the OSCE were representative of professional practice and adequately sampled the range of competencies expected of entry-level (or qualifying) candidates. These content specifications were defined to assist station-writers in developing cases that, when considered together in an OSCE, would be broadly representative of the profession's expectations and standards of practice at entry-level.

Content specifications were developed that included:

- Critical and commonly encountered disease states (those important medical conditions most frequently seen by pharmacists.)
- Type of practice setting (community pharmacy, hospital or long-term care facility.)
- Drug related problems (such as incorrect dosages, allergy concerns, drug–drug interactions, etc.)
- Patient populations (such as the elderly, ambulatory adults or children)
- Types of interventions (such as management of a patient's drug therapy, or referral to a physician or other health care professional.)
- Complexity (i.e. complicating factors such as the existence of multiple disease states, psycho–social issues, etc.)
- Level of risk if problems were not resolved.

These content specifications, together with the functions and abilities identified previously, formed the template by which case writers could work in developing meaningful OSCE stations, (either interactive or non–interactive), for the assessment of professional competencies and communication abilities in general.

To ensure broad representation from professional practice and to optimize buy-in from the profession at large, case writers were selected from across the country, representing different constituencies of the profession (e.g. community practice or hospital practice). Case writers and standardized patient-trainers, (SP Trainers), were invited to attend a two-day workshop wherein principles of case development were presented and discussed and development of assessment instruments was reviewed. Case writers with subject matter expertise worked in pairs to develop OSCE stations, along with the support of a SP Trainer. At a subsequent two-day workshop, each case was reviewed by a different group of three pharmacists, along with an SP Trainer.

Though a costly and time-intensive exercise, the development of stations and assessment tools is pivotal to the success of an OSCE and essential in ensuring validity of the outcome. Moreover, the involvement of practising pharmacists in the case writing and review process sent an important signal to the profession that the content of these stations was representative of practice and aligned with pharmacists' real-life experiences with patients, not simply academic exercises.

Method for assessment of performance within client interaction stations

Assessment of performance within a station is potentially complicated by the fact that "performance" is an integration of communicative competency with clinical problem-solving. This may introduce interactivity or subjectivity into the assessment with a corresponding negative impact on reliability. Cognizant of this potential limitation, the decision was made to use both analytical (checklist) and global (holistic) assessments and evaluate the impact on scoring and candidates' results.

In the pharmacy OSCE context, the analytical checklist consists of performance-based, behaviorally specific observations, such as interview questions, interpretive or informative questions, etc. Outcomes of these behaviors are observable; the analytical checklist provides assessors with a tool for documenting and evaluating these outcomes to assess application of professional knowledge and problem solving abilities in the context of each station. It is important to note that significant interpretation is still required on the part of the assessor. For example, a typical analytical checklist item for a pharmacy OSCE may be:

- Asks about side effects patient has experienced while taking morphine (e.g. nausea, constipation.)

A candidate in this station may not be so direct, and may instead ask the question, "Have you had any problems recently, things like feeling drowsy during the day?" Strictly speaking, and directly from the words of the analytical checklist item cited above, the candidate has not asked about side effects, but "problems", nor have they cited the two most common side effects associated with the medication in question. Nonetheless, in general parlance, the word "problem" can be seen to be a proxy for "side effect" in the context of a pharmacist's interview with a patient, and the "problem" discussed (drowsiness), though not cited as an example in this case, may be a reasonable example to use as a prompt for a discussion with a patient and thus be given credit as a "unique response" or proxy for the checklist item.

In order to optimize reliability of assessment, contingencies such as this were considered in design of an analytical checklist. Different assessors (all of whom would be pharmacists) may offer different views of the situation cited above, and could justify why they would choose to give the candidate the point or not. Despite its seeming objectivity, the analytical checklist may be open to interpretation and debate. While assessors are asked to observe specific reproducible activities, the nature of analytical checklists, reflective of pharmacy practice, focus on reasonable approximations rather than exact reproduction of an expected behaviour.

Conversely, analytical checklists may be too reductionist in some cases, unfairly disadvantaging exceptional candidates who do not require a sequential interview technique to arrive at the correct endpoint. Since the patient interview is a complex human dynamic, different interviewing styles and processes may be effective in eliciting the required information and may not correspond to the process described in a checklist. If the expected endpoint is reached in a safe, effective (and often-times more efficient) manner, candidates should not be penalized for this by scores based exclusively on the analytical checklist. Global (or holistic) scales were designed to capture a candidate's performance in communication skills and integrative problem solving, considering the more critical behaviours listed in the analytical checklist along with the accuracy and logical ordering of the candidate's responses. Therefore, analytical checklists were used by assessors to guide and support holistic ratings.

Previous studies found that global rating scales had better reliability and construct validity than checklist scores (subsequently published in Regeher, MacRae, Reznick and Szalay, 1998) and that they "appear to be able to capture aspects of student performance that are difficult to capture in specific behavioral ratings (checklists)", (subsequently published in Solomon, Szauter, Rosebraugh and Callaway, 1999). The OSCE scores were, therefore, based on global ratings supported by scoring rubrics that incorporated "critical items" from the checklists.

Recognizing the dimensions of communications, problem solving and other performance characteristics in pharmacy practice, three holistic rating scales were developed: communications, outcome, and performance. For each category, four unique levels were developed with accompanying descriptors, each describing a specific level of attainment. A four point scoring system was developed (1 = unacceptable; 2 = marginally unacceptable; 3 = marginally acceptable; 4 = acceptable). Binary descriptors, (Yes/No), were provided for two additional performance categories: misinformation (i.e. incorrect information provided) and risk (i.e. probability of need for further intervention as a result of an error or omission, resulting in compromised patient care). The Performance rating captures different elements, including communications, outcome, misinformation and risk, and the overall quality of the candidate's performance.

Field testing the OSCE

Pre-testing of each station in an OSCE format was undertaken using a panel of practising pharmacists, (including some recent graduates who were newly licensed). During this field test, it was possible to observe each station in action to ensure functionality of the station and scoring criteria, and the feasibility

and acceptability of the OSCE format. A total of 51 cases, (and their corresponding analytical checklists and holistic rating scales), were developed, reviewed, revised and finalized using the process outlined above. Based on positive field test results, minor modifications were made to stations and processes, and the decision was made to pilot the OSCE to determine its utility for entry-to-practice assessment in pharmacy.

Research objectives

While previous experience with OSCEs in pharmacy had demonstrated their value in a variety of educational, certification/licensing, and continuing competency settings, no systematic study of OSCEs at entry-to-practice had been formally undertaken. Consequently the primary objective of this study was to evaluate the OSCE and the assessment outcomes for reliability, validity, and generalizability in the national entry-to-practice context in pharmacy in Canada. Ancillary objectives included evaluation of the use of simulated patients as assessors in the OSCE, and exploration of the feasibility of the OSCE to assess continuing competency of pharmacists already licensed and in practice.

Methods

Two balanced forms of the OSCE were created for the pilot, based on the results of development and field testing described previously. Each form consisted of 18 interactive stations and 8 non-interactive stations, drawn from the bank of cases developed and described previously. The pilot was administered over two separate occasions in June 1998. For logistical and cost reasons, this study was undertaken in one center (Vancouver, Canada) on two separate occasions. This center was selected because of its previous experience in operating pharmacy OSCEs, its partnership with PEBC in developing the new OSCE stations and exam administration processes and the opportunity to pilot the assessment in the context of a high stakes (licensing) examination.

A total of 190 participants were involved: 37 were practising (i.e. already licensed and experienced) pharmacists from the province of British Columbia, and 153 participants were either new graduates, (i.e. entry-level pharmacists from the University of British Columbia), or international pharmacy graduates seeking licensure in Canada. Of the 37 practising pharmacists, all had selected participation in the pilot as a method of continuing competency assessment. Of the 153 entry-to-practice participants, all were required to take and pass this examination as a condition for licensure in British Columbia. Due to the large number of participants, it was not possible to run the pilot on one day. As a result, two different forms of the OSCE were assembled, each of which was administered on a separate date. Form I of the OSCE was administered and scored for 75 entry-to-practice participants and 21 practising pharmacists; Form II was administered to and scored for 78 entry-to-practice participants and 16 practising pharmacists.

A variety of different assessor allocations were used in the interactive stations in the pilot, in order to determine optimal assessor configuration for the pharmacy OSCE. In some stations, two pharmacist-assessors were present—one full-time assessor and one "roving" assessor (i.e. a pharmacist-assessor who moved from station to station, assessing candidates in multiple stations over the course of the OSCE). In other stations, one pharmacist-assessor, or one pharmacist-assessor and one simulated patient-assessor were utilized.

Scoring and standard setting methods

Candidates' overall scores were computed by aggregating raw scores from all three holistic scales across all stations. Standard setting to establish a passing score in performance-based assessment is complex. For the pilot, standard setting judges were selected to be broadly representative of pharmacy practice in Canada. Six judges were selected from British Columbia (where experience with the OSCE format was more long-standing than in other parts of Canada) and six judges were selected from other parts of the country, all with experience in pharmacy practice. All standard setters had experience with OSCEs, and as a group of twelve judges, adequately represented both hospital and community pharmacy practice.

Standard setters were first charged with defining and identifying the distinguishing characteristics of the "borderline qualified entry-to-practice candidate". As individuals, each standard-setter estimated expected scores for the communication, outcome, and performance scales for this borderline candidate for each OSCE station. A minimum performance level (MPL) for the holistic ratings for each station was calculated as the mean of the standard-setters final estimates. Overall the passing standard for the whole OSCE was calculated by summing mean MPLs for all stations used in the OSCE.

Generalizability (G) studies were conducted to analyze results of the pilot study.

These studies are undertaken to discern true scores from errors. G studies are more sophisticated than traditional reliability tests, (such as Cronbach alpha), insofar as they may be used to analyze multiple sources of measurement errors, (variance components).

Psychometrically, OSCEs are complex phenomena, producing scores with potential errors from multiple sources, including assessors, stations (tasks), forms, scoring methods, and so forth. G studies pinpoint sources of error in scores, estimate the magnitude of

Table I. Comparative results of OSCE pilot (Form I): Analytical scores.

Function	Weighted *p*-value (QC)*	Weighted *p*-value (Ph)
Gather, assess, and interpret information from clients (patients, physicians, etc.) and other sources	0.67	0.59
Make appropriate recommendations to clients	0.62	0.60
Communicate and educate effectively	0.51	0.42
Prepare and distribute medications and drug-delivery devices	0.70	0.73
Exercise professional and ethical judgments	0.58	0.57
Weighted analytical mean	0.59	0.58

QC = Qualifying Candidate ($n = 75$); Ph = Practising Pharmacist ($n = 21$).
* Weighted *p*-value for each of the five functions (analytical subscores) is the frequency of performance of checklist items related to that function within station, averaged across all stations in which the function was assessed.

these errors under different conditions so that cost-efficient measurements can be built, and provide summary reliability coefficients (generalizability coefficients and dependability coefficients) (Brennan, 1995). The G coefficient expresses the consistency of scores relative to the performance of other candidates. It is a relative comparison to the performance of the group usually used with norm-referenced examination interpretations. When using criterion-referenced interpretations, the D coefficient expresses the consistency of the absolute scores of candidates on an examination, or the true level of performance.

When used in conjunction with G studies, D studies can be valuable tools for interpreting the extent to which major sources of variance have affected measurement within a performance-based examination and for making improvements to the examination; thus, these studies are essential in ensuring fairness of high-stakes test outcomes.

Results

Results of the OSCE pilot are presented in Tables I–IX.

Statistical analysis of entry-to-practice participants' performance was undertaken. Reliability analyses using Cronbach's alpha coefficient were complemented by generalizability, (G coefficient), and dependability, (D coefficient), studies to address issues most frequently representing threats to the consistency, validity, and generalizability of passing scores and dependability of results in a performance-based examination. To identify and estimate potential and actual sources of variance in the analytical (checklist) and global (holistic) scores, the following facets and their interactions were investigated: candidates, stations (or tasks), assessors, (including pharmacist-assessors and simulated patient-assessors), and occasions. The number of stations, number of assessors and type of assessor required to produce scores that are generalizable and dependable pass–fail decisions, were estimated for both analytical and holistic scoring methods. Comparative results for qualifying candidates and for practising pharmacists on analytical and holistic scores is presented in Tables I and II. Descriptive analysis of all data was undertaken. For cost and logistics reasons, detailed analysis of data (including generalizability and dependability studies) were undertaken for Form I only, to study sources of measurement error.

Discussion

Key findings from this study addressed major questions with respect to design of an entry-to-practice OSCE for pharmacy:

What number and what type of assessors are required to obtain consistent and dependable candidates' scores? How does use of simulated patient-assessors affect reliability?

The logistics of co-ordinating a multi-station OSCE can be daunting and the costs prohibitive. One major

Table II. Comparative results of OSCE pilot (Form I): Holistic scores.

Variable	Mean score (QC) (out of 4)	Weighted *p*-value (QC)	Mean score (Ph) (out of 4)	Weighted *p*-value (Ph)
Communication*	3.73	0.93	3.77	0.94
Outcome	2.96	0.74	2.95	0.74
Performance	2.99	0.75	2.96	0.74
Weighted holistic mean	3.14	0.78	3.18	0.79

QC = qualifying candidate ($n = 75$); Ph = practising pharmacist ($n = 21$).
Scoring rubric: 1–4: (1) unacceptable, (2) marginally unacceptable, (3) marginally acceptable, (4) acceptable.
Weighted *p*-value for each of the holistic scores is the proportion of the maximum rating achieved by the candidate on each scale within station, averaged across all stations in which the parameter was assessed.
* Communication scores applied to 15 (client-interactive) stations only; Outcome and performance scores applied to 20 (client-interactive and client non-interactive) stations.

Table III. Comparison of scaled scores in the OSCE (Form I and II) and the PEBC written (MCQ) examination for qualifying candidates.

	OSCE	MCQ	Correlation (r)
Form I ($n = 65$)	586	506	0.40
Form II ($n = 62$)	632	542	0.47
Form I + II ($n = 127$)	608.5	523.6	0.44

Note: Form I and II of the OSCE were administered on two separate occasions.

cost driver for the OSCE is exam personnel. Key questions for OSCE administrators include: (a) how many assessors are required in each station, and whether these assessors must all be health care professionals; (b) can simulated patients act as assessors and (c) can the same person both simulate the client role in the station and assess the candidate's performance simultaneously?

Roving pharmacist-assessors' scoring was compared to pharmacist-assessors' scoring, with data presented in Tables IV and V. Reliability (Crohnbach's alpha) coefficients ranged from 0.97 to 0.99 for analytical and holistic grading, respectively, Generalizability (G). Studies also confirmed the high consistency between pharmacist-assessors when using both analytical and holistic scoring. The implication of this finding is that one pharmacist-assessor is sufficient, and little or no additional benefit is accrued by having a second pharmacist-assessor involved. This finding is of significance in implementing a cost-effective OSCE (Carpenter, 1995).

Comparison of analytical and holistic scores between pharmacist-assessors (who observed and scored), and simulated patient-assessors (who simultaneously simulated roles and assessed), showed much greater variation: alpha coefficients ranged from $r = 0.61$ (for the Communication component of the holistic scoring) to $r = 0.98$ (for the analytical (checklist) scoring). The rater effect, although small, was larger for the simulated patient-assessor and pharmacist-assessor combination than for the assessor and roving assessor combination for both the analytical and the holistic scoring as illustrated in Tables VI and VII.

It is important to note that simulated patient-assessors were trained in portrayal of the role, but received only limited training in use of holistic scales for this study. Furthermore, only the Communication scale was used by the simulated patient-assessors, and it was a slightly modified and paraphrased version of the scale used by the pharmacists. Consequently, it may be reasonable to suggest that training and use of the same tool by pharmacists and simulated patients may improve these indices.

How many stations are required to maintain consistency and generalizability of the candidates' scores?

The OSCE relies heavily upon adequate and representative sampling of professional practice. The overall design of the OSCE is based on the notion that one performance in one station cannot adequately capture a candidate's abilities. Multiple direct observations across multiple stations provides a more defensible outcome. This, however, must be balanced against practical and logistical constraints which dictate the number of stations that is feasible to develop and administer within one examination sitting.

In this study, a 26-station OSCE was administered in both Form I and II because the CPBC candidates were taking the OSCE for licensing purposes, and the OSCE had to match the CPBC specifications. However, for the purpose of the joint project the 15 SP stations were mainly used for these generalizability analyses. Results suggest there is no significant or important difference in stations (T) variance components when adding stations beyond 15 stations (shown in Tables VI and VII). In addition, data analyzed illustrates no significant or important differences in D (dependability), or G (generalizability) coefficients between a 15-station and

Table IV. Analytical scoring G studies (σ^2) and D studies (P^2) with estimated variance components for qualifying candidates (Form I): Assessors and roving assessors.

G study σ^2			D study estimated variances components				
		Percentage	$n\tau = 3$ $n\tau = 1$	$n\tau = 5$ $n\tau = 2$	$n\tau = 10$ $n\tau = 2$	$n\tau = 15$ $n\tau = 2$	$n\tau = 15$ $n\tau = 1$
P	0.0022	4.28	0.0018	0.0018	0.0018	0.0018	0.0018
R	0.0000	0.00	0.0000	0.0000	0.0000	0.0000	
T	0.0110	21.44	0.0035	0.0021	0.0016	0.0007	0.0007
PR	0.0000	0.00	0.0000	0.0000	0.0000	0.0000	
PT	0.0320	62.37	0.0163	0.0064	0.0032	0.0021	0.0021
RT	0.0007	1.36	0.0002	0.0001	0.0000	0.0000	
PRT	0.0054	10.52	0.0018	0.0005	0.0003	0.0002	0.0004
Generalizability coefficients relative decisions (P^2)			.12	.20	.34	.43	.42
Absolute decisions (ϕ)			.10	.16	.28	.37	.35

$P = 13$ Candidates; $R = (2)$ roving assessors and assessors; $T = 3$ Stations (22, 48, P01).

Table V. Holistic scoring G studies (σ^2) and D studies (P^2) with estimated variance components for qualifying candidates (Form I): Assessors and roving assessors.

G study (∂^2)			D study estimated variances components				
		Percentage	$n\tau$=3 $n\tau$=1 $n\tau = 1$	$n\tau$=5 $n\tau$=2 $n\tau = 2$	$n\tau$=10 $n\tau$=2 $n\tau = 2$	$n\tau$=15 $n\tau$=2 $n\tau = 2$	$n\tau$=15 $n\tau$=1 $n\tau = 1$
P	0.28300	20.02	0.28300	0.28300	0.28300	0.28300	0.28300
R	0.00000	0.00	0.00000	0.00000	0.00000	0.00000	0.00000
T	0.00000	0.00	0.00000	0.00000	0.00000	0.00000	0.00000
PR	0.00053	0.03	0.00053	0.00027	0.00027	0.00027	0.00053
PT	0.86050	60.89	0.28680	0.17210	0.08610	0.05740	0.05740
RT	0.11530	8.150	0.03850	0.01154	0.00560	0.00390	0.00770
PRTe	0.15380	10.88	0.05130	0.01540	0.00770	0.00510	0.01020
Generalizability coefficients relative decisions (P^2)			.45	.60	.75	.82	.81
Absolute decisions (ϕ)			.43	.59	.74	.81	.79

$P = 13$ Candidates; $R = (2)$ roving and assessor; $T = 3$ stations (22, 48, P01).

20-station OSCE. There was no apparent improvement in the dependability and generalizability of results when more than 15-stations were included in an entry-to-practice pharmacy OSCE, when comparing SPs and pharmacist assessors' scores. Fifteen stations with one pharmacist assessor produced G and D coefficients as high as 0.81 and 0.79, respectively. Much lower dependability of results was achieved when using only 10 stations. Thus, the addition of more stations beyond 15 did not result in any meaningful improvement in assessment, and is thus deemed to be not cost-effective.

An examination of sources of candidates' score variance was undertaken. This analysis suggests that, when using analytical scoring, the stations themselves (*T*) contributed most to the error variance in three of the four studies (Tables IV to VII), suggesting that stations varied in difficulty and/or checklists varied in comprehensiveness and context. Although to a lesser degree, stations also contributed most to error variance when scored by SPs using an holistic (Communications) scale. This was not the case when using scores by pharmacists using holistic scales. This might not be considered a surprising result as stations were intended to measure different skills and abilities and candidates were expected to possess higher or lower competency across different contexts in pharmacy practice.

How do different scoring methods (i.e. analytical (checklist) vs. global (holistic)) affect candidates' scores?

Correlations between holistic and analytical scoring ranged from $r = 0.39$ (for 15 and 20 OSCE stations) to $r = 0.43$ (for all 26 stations). These relatively low correlations suggest that holistic and analytical scoring may not be interchangeable, and if used in isolation, may yield different end results, particularly for borderline candidates. There was a notable difference in mean performance between holistic and analytical scores (weighted *P* value = 0.81 for holistic and 0.59 for analytical). These differences may be due to the all-inclusive nature of the checklists, including items that are inconsequential and missed by many candidates, (although these candidates would have provided the responses deemed essential to be judged as adequate). Overall, the study suggests that holistic scoring may be somewhat more forgiving and stable across tasks and assessors than analytical scoring.

Table VI. Analytical scoring for qualifying candidates (Form I): Standardized patients and assessors.

G studies (σ^2)			D study estimated variance components			
	Generalizability	Percentage	$n\tau$=5 $n\Gamma$=2	$n\tau$=10 $n\Gamma$=1	$N\tau$=15 $n\Gamma$=1	$n\tau$=20 $n\Gamma$=1
P	0.001400	2.63	0.0014	0.014	0.014	0.014
T	0.018300	34.43	0.004	0.002	0.002	0.0009
R	0.003600	6.77	0.002	0.004	0.004	0.004
$P \times T$	0.01800	33.87	0.004	0.002	0.0012	0.0009
$P \times R$	0.000037	0.06	0.00002	0.00004	0.00004	0.00004
$T \times R$	0.001800	3.38	0.0002	0.0002	0.00012	0.00009
$P \times T \times R$	0.10000	18.81	0.00096	0.00096	0.00064	0.00050
P^2			.23	.33	.43	.50
ϕ			.12	.14	.17	.19

P = qualifying candidates (= 75); T = number of stations/tasks (= 15); R = number of raters (= 2); τ = stations/tasks; Γ = raters; p^2 = generalizability coefficients relative decision; ϕ = absolute decision.

Table VII. Holistic (communication) scoring for qualifying candidates (Form I): Standardized patients and assessors.

	G studies (σ^2)		D study estimated variance components			
	Generalizability	Percentage	$n\tau$=5 $n\Gamma$=2	$n\tau$=10 $n\Gamma$=1	$n\tau$=15 $n\Gamma$=1	$n\tau$=20 $n\Gamma$=1
P	0.0240	2.96	0.024	0.0240	0.0240	0.0240
T	0.0670	8.28	0.013	0.0067	0.0045	0.0034
R	0.0960	11.86	0.048	0.0960	0.0960	0.0960
$P\times T$	0.0950	11.74	0.019	0.0095	0.0060	0.0050
$P\times R$	0.0063	0.77	0.003	0.0063	0.0063	0.0063
$T\times R$	0.2450	30.28	0.025	0.0250	0.0164	0.0123
$P\times T\times R$	0.2756	34.07	0.028	0.0280	0.0184	0.0140
P^2			.33	.35	.44	.49
ϕ			.15	.12	.14	.15

P = qualifying candidates (= 75); T = number of stations/tasks (= 15); R = number of raters (= 2); τ = stations/tasks; Γ = raters; p^2 = generalizability coefficients relative decision; ϕ = absolute decision.

What is the validity and defensibility of standard-setting procedures and pass/fail decisions?

A competence-based standard setting procedure was used. Sources of variance in standard-setting were examined, including standard-setters (judges), stations, and occasions. Particular emphasis was placed on establishing the number of standard-setters and stations necessary to produce defensible pass–fail decisions. The standard setters were given the task to score hypothetical borderline-qualified candidates' performance for each of the 26 stations for both Form I and II.

In all cases, mean scale scores were close to 3.0 for each of the 4-point holistic scales. In some cases,(depending upon individual standard-setters' perception of station difficulty), individual standard-setters estimated that "borderline-qualified" candidates would perform at a higher (i.e. acceptable (= 4)) or lower (i.e. marginally unacceptable (= 2)) level.

Results (presented in Tables VIII and IX) indicate that the most important contributor of variance was due to different perceptions of station difficulty, and the least important contributor was the rater effect. On a percentage basis, the largest component of variance was the interaction between standard-setters and stations, suggesting that judges' perception of borderline-qualified candidates' performance across stations was variable.

In part, this may be due to the nature of the standard-setters themselves, since they were deliberately selected to represent different aspects of pharmacy practice (i.e. hospital and community, where professional practice and expectations may differ significantly). Since the entry-to-practice and licensure system in pharmacy does not differentiate between hospital and community pharmacists such variance may be inevitable and unavoidable. Other, smaller variance components included the interaction of standard-setters and occasions. Overall, results show that a 15-station examination with six standard-setters provides passing scores with adequate dependability.

How valid are scores and pass–fail decisions?

The Pharmacy Examining Board of Canada administers a written multiple-choice Qualifying Examination for entry-to-practice candidates seeking licensure in Canada. The written examination was based on a blueprint testing the pharmacy practice competencies similar to those defined for the OSCE Pilot. To support the validity of scores from the OSCE pilot, candidates' scores on the written examination were matched and compared with their performance on the OSCE. As shown in Table III, a moderate

Table VIII. Estimated variance components and generalizability coefficients for standard-setting processes (Forms I and II).

	P	T	$P\times T$ interaction	p^2	ϕ
T = 15 P = 12	0.022	0.044	0.130	0.71	0.65
T = 20 P = 12	0.023	0.070	0.163	0.74	0.66
T = 26 P = 12	0.019	0.061	0.166	0.74	0.68
T = 15 P = 6(BC)	0.017	0.059	0.104	0.71	0.61
T = 20 P = 6(BC)	0.018	0.062	0.201	0.64	0.57
T = 26 P = 6(BC)	0.025	0.041	0.185	0.78	0.75
T = 15 P = 6(PEBC)	0.033	0.043	0.141	0.78	0.73
T = 20 P = 6(PEBC)	0.031	0.081	0.122	0.84	0.76
T = 26 P = 6(PEBC)	0.013	0.010	0.127	0.73	0.61

T = stations/tasks; P = raters; PEBC = raters from across Canada with Pharmacy Examining Board of Canada; BC = raters from College of Pharmacists of British Columbia only; p^2 = generalizability coefficients relative decision; ϕ = absolute decisions.

Table IX. Estimated variance components, generalizability (Rho) coefficients, dependability (Phi) coefficients for standard setting (Forms I and II).

	26-station Form	20-station Form	15-station Form
No. of standard setters	12	12	12
Standard setters (*P*)	0.019	0.023	0.022
Stations (*T*)	0.061	0.070	0.040
Interaction ($P \times T$)	0.166	0.163	0.130
Generalizability coefficients (Rho)	0.74	0.74	0.71
Absolute decisions (Phi)	0.68	0.66	0.65

correlation between these scores was observed for both Form I ($r = 0.40$) and Form II ($r = 0.47$) of the OSCE. Mean scores for the OSCE were higher, (and statistically significant (at $p = 0.001$)) than for the MCQ test. These findings suggest that the OSCE complements, but does not replace, the MCQ test. It also suggests the OSCE measures different competencies not currently or adequately captured through the written test.

Although not statistically defensible (due to small numbers of participants and methods of selection into this study), it is interesting to compare the performance of the 21 practising pharmacists to the 127 entry-to-practice candidates. Although similar, mean holistic scores for pharmacists tended to be slightly lower for entry-to-practice candidates (mean (pharmacists) = 2.96 vs. mean (entry-to-practice) = 3.08). Similar trends were noted in the analytical scores. Caution should be exercised in interpreting these results since the small number of volunteers who self-selected for involvement in this research were not necessarily representative of all practising pharmacists. Furthermore, the examination was not constructed to reliably rank order candidates on the score scale, but to produce valid pass–fail decisions. (It should also be noted that all of these pharmacists' scores exceeded the passing score set for the examination, whereas some of the entry-to-practice candidates did not.)

Conclusions

The process of defining competencies and an examination blueprint, and developing and pre-testing individual stations was laborious and time-consuming, but necessary to ensure buy-in from the profession. As a result of this study, the following conclusions and recommendations were made for the design and deployment of an entry-to-practice OSCE for pharmacy in Canada:

- A 15-station OSCE, using one pharmacist assessor per station, yielded consistent and generalizable scores when holistic scoring was used to assess qualifying candidates and practising pharmacists.
- At this time, simulated patients should not be relied upon to replace pharmacist-assessors.
- The competence standard setting methodology applied to holistic scales yielded consistent and reliable passing scores.
- Stations contribute most significantly to candidates' score variance (especially using analytical scoring); stations can and do vary in difficulty, and care should be taken in balancing OSCEs using domain specification and task sampling techniques.
- Assessment of knowledge through written tests and performance-based assessment complement one another but do not replace one another.

While this study has established the reliability, validity, and generalizability of an entry-to-practice OSCE for pharmacy, implementation at a national level of such an examination poses significant logistical and procedural challenges. Issues related to co-ordination of such an exam across multiple examination centres, cities, and time zones were not considered as part of this study. Further, results were based, in large part, on the participation of a relatively homogenous group of entry-to-practice candidates, all from the same part of the country (British Columbia) and from the same educational institution (the University of British Columbia). Regional variation across Canada, coupled with inclusion of those educated outside Canada, may affect results.

Overall, the OSCE appears to be an important and complementary tool for assessment of entry-to-practice competency, adding breadth and depth to the existing written examination. Based on these findings and subsequent confirmatory research, the Pharmacy Examining Board of Canada has implemented a national entry-to-practice OSCE as part of the licensure process for Canadian pharmacists. In so doing, Canada has become the first country in the world to adopt the use of a multi-site entry-to-practice OSCE for pharmacy.

Acknowledgements

The authors wish to acknowledge the significant contributions made by the following groups and individuals to the successful development and piloting of this first national performance-based assessment for certification and licensure of pharmacists in Canada,

Members of the Joint Project Steering Committee: Dr Jim Blackburn (PEBC), Dr David Blackmore (The Medical Council of Canada Jim Dunsdon, NAPRA), Dr David Fielding (Faculty of Pharmaceutical Sciences, University of British Columbia), Bruce Millin (CPBC), Mits Miyata (CPBC, PEBC), Midge Monaghan (Ontario College of Pharmacists), Scott McLeod (PEBC), Barbara Thompson (CPBC) and Pharmacists and staff members of the CPBC, particularly Linda Lytle, Registrar Members of the Board of Examiners of the CPBC, particularly Kathy McInnes and Chair, Panel Assessment Committee.

References

Abedi, J. (2004). The no child left behind act and english language learners: Assessment and accountability issues. *Education Research*, 33(1), 4–14.

Austin, Z., Croteau, D., Marinia, A., & Violato, C. (2003). Continuous professional development: The Ontario experience in professional self-regulation through quality assurance and peer review. *American Journal of Pharmaceutical Education*, 62(2), 56–63.

Benson, J. A. (1991). Certification and recertification: One approach to professional accountability. *Annals of Internal Medicine*, 114, 228–232.

Brennan, R. L. (1995). Generalizability of performance assessments. *Educational Measurement: Issues and Practice*, 9–12, Winter.

Campbell, C., Parboosingh, J., & Slotnick, H. B. (1999). Outcomes related to physicians practice-based learning. *Journal of Continuing Education of the Health Professions*, 19(4), 234–241.

Carpenter, J. L. (1995). Cost analysis of OSCEs (Review). *Academic Medicine*, 70, 828–833.

Colliver, J. A., & Swartz, M. H. (1997). Assessing clinical performance with standardized patients. *JAMA*, 278, 790–791.

Dupras, D. M., & Li, J. T. (1995). Useof an OSCE to determine clinical competence. *Academic Medicine*, 70, 1029–1034.

Fielding, D. W., Page, G. G., Fevang, L. C., & Thomas, N. S. (1981). Competency assessment: A progress report on British Columbia's program. *American Journal of Pharmaceutical Education*, 45, 178–183.

Fielding, D. W., Page, G. G., Schulzer, M., Rogers, W. T., & O'Byrne, C. C. (1992). Assuring continuing competency: Identification and validation of parctice-based assessment blueprint. *American Journal of Pharmaceutical Education*, 56, 21–29.

Fielding, D., Page, G., Rogers, W., O'Byrne, C., Schulzer, M., Moody, K. G., & Dyer, S. (1997). Application of objective structured clinical examinations in an assessment of pharmacists' continuing competency. *American Journal of Pharmaceutical Education*, 61, 117–126.

Fielding, D. W., Rogers, W. T., Tench, E., O'Byrne, C. C., Page, G. G., & Schulzer, M. (2001). Predictors of continuing competence. *American Journal of Pharmaceutical Education*, 65, 107–118.

Grand'Maison, P., Brailovsky, C. A., Lescop, J. J., & Rainsberry, P. C. (1997). Using standardized patients in licensing/certification examinations: A comparison of two tests in Canada. *Family Medicine*, 29, 27–32.

Regeher, G., MacRae, H., Reznick, R. K., & Szalay, D. (1998). Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Academic Medicine*, 73(9), 993–997.

Robb, K. V., & Rothman, A. (1985). Assessment of clinical skills in general medical residents: Comparison of the objective structured clinical examination to a conventional oral examination. *Annals of the Royal College of Physicians and Surgeons of Canada.*, 18, 235–238.

Singer, P. A., Robb, A., Cohen, R., Norman, G., & Turnbull, J. (1996). Performance-based assessment of clinical ethics using an objective structured clinical examination. *Academic Medicine*, 21, 495–498.

Sloan, D. A., Donnelly, M. B., Schwartz, R. W., & Strodel, W. E. (1995). The objective structured clinical examination: The new gold standard for evaluating postgraduate clinical performance. *Annals of Surgery*, 222(6), 735–742.

Solomon, D. J., Szauter, K., Rosebraugh, C. J., & Callaway, M. R. (1999). Global ratings of student performance in a standardized patient examination: Is the whole more than the sum of the parts? *Advances in Health Sciences Education*, 4, 1–10.

Woodburn, J., & Sutcliffe, N. (1996). The reliability, validity and evaluation of the objective structured clinical examination in podiatry (chiropody). *Assessment and Evaluation in Higher Education*, 21, 131–146.

Author Queries

JOB NUMBER: 102517
JOURNAL: GPHE

Q1 Barrows (1993) is cited inside the text, but is missing in Ref. list, please check.