



OPEN ACCESS

Development and testing of an objective structured clinical exam (OSCE) to assess socio-cultural dimensions of patient safety competency

Liane R Ginsburg,¹ Deborah Tregunno,² Peter G Norton,³ Sydney Smees,⁴ Ingrid de Vries,⁴ Stefanie S Sebok,⁵ Elizabeth G VanDenKerkhof,⁶ Marian Luctkar-Flude,² Jennifer Medves²

► Additional material is published online only. To view please visit the journal online (<http://dx.doi.org/10.1136/bmjqs-2014-003277>).

For numbered affiliations see end of article.

Correspondence to

Dr Liane R Ginsburg, School of Health Policy & Management, HNES Bldg. Rm. 413, 4700 Keele Street, Toronto, Ontario, Canada M3J 1P3; lgins@yorku.ca

Received 5 June 2014

Revised 15 October 2014

Accepted 28 October 2014

ABSTRACT

Background Patient safety (PS) receives limited attention in health professional curricula. We developed and pilot tested four Objective Structured Clinical Examination (OSCE) stations intended to reflect socio-cultural dimensions in the Canadian Patient Safety Institute's Safety Competency Framework.

Setting and participants 18 third year undergraduate medical and nursing students at a Canadian University.

Methods OSCE cases were developed by faculty with clinical and PS expertise with assistance from expert facilitators from the Medical Council of Canada. Stations reflect domains in the Safety Competency Framework (ie, managing safety risks, culture of safety, communication). Stations were assessed by two clinical faculty members. Inter-rater reliability was examined using weighted κ values. Additional aspects of reliability and OSCE performance are reported.

Results Assessors exhibited excellent agreement (weighted κ scores ranged from 0.74 to 0.82 for the four OSCE stations). Learners' scores varied across the four stations. Nursing students scored significantly lower ($p<0.05$) than medical students on three stations (nursing student mean scores=1.9, 1.9 and 2.7; medical student mean scores=2.8, 2.9 and 3.5 for stations 1, 2 and 3, respectively where 1=borderline unsatisfactory, 2=borderline satisfactory and 3=competence demonstrated). 7/18 students (39%) scored below 'borderline satisfactory' on one or more stations.

Conclusions Results show (1) four OSCE stations evaluating socio-cultural dimensions of PS achieved variation in scores and (2) performance on this OSCE can be evaluated with high reliability, suggesting a single assessor per station would be

sufficient. Differences between nursing and medical student performance are interesting; however, it is unclear what factors explain these differences.

BACKGROUND

Patient safety (PS) competency among health professionals (HPs) nearing entry to practice is an area of growing interest. There are recommendations from numerous international bodies regarding the need to restructure HP education to ensure it equips students with the knowledge, skills and attitudes necessary to function safely.¹⁻⁵ Yet, PS receives limited attention in HP curricula.⁶ Moreover, when asked about their confidence in PS learning, new graduates in medicine, nursing and pharmacy identify a number of gaps in PS learning.⁷

The Canadian Patient Safety Institute (CPSI) Safety Competency Framework⁵ suggests six core competency domains that reflect the knowledge, skills and attitudes that enhance PS across the continuum of care (see [box 1](#)). Previously, we examined new graduates' self-reported confidence in each of these six competency areas.⁷ Self-reported PS competence provides one important and cost effective source of information about the effectiveness of the HP education process,¹ including aspects of an individual's PS knowledge that need attention.⁸⁻⁹ However, a recent systematic review concluded that physicians have a limited ability to accurately self-assess¹⁰ and this may be particularly true among

To cite: Ginsburg LR, Tregunno D, Norton PG, *et al.* *BMJ Qual Saf* Published Online First: [please include Day Month Year] doi:10.1136/bmjqs-2014-003277

Box 1 Canadian Patient Safety Institute Safety Competencies⁵

- ▶ *Domain 1: Contribute to a culture of patient safety*—Commitment to applying core patient safety knowledge, skills and attitudes to everyday work
- ▶ *Domain 2: Work in teams for patient safety*—Collaboration and interdependent decision-making with inter-professional teams and with patients to optimise patient safety and quality of care
- ▶ *Domain 3: Communicate effectively for patient safety*—Promote patient safety through effective healthcare communication
- ▶ *Domain 4: Manage safety risks*—Anticipate, recognise and manage situations that place patients at risk
- ▶ *Domain 5: Optimise human and environmental factors*—Manage the relationship between individual and environmental characteristics to optimise patient safety
- ▶ *Domain 6: Recognise, respond to and disclose adverse events*—Recognise the occurrence of an adverse event or close call and respond effectively to mitigate harm to the patient, ensure disclosure and prevent recurrence

the least skilled.^{10 11} Self-assessment alone is insufficient as it is not a stable skill (in some contexts it can be quite accurate but not in others). Instead, self-improvement is most likely to occur when a learner engages in internal reflection and also receives feedback from external sources.¹²

Objective Structured Clinical Examinations (OSCEs) provide external assessments of knowledge, skills and behaviours in the context of a low risk simulated clinical encounter and a growing literature encourages their use to assess PS competencies.^{13–15} Most PS OSCE studies have focused on clinical aspects of PS, such as hand hygiene, medication labelling¹⁶ and safe performance of procedures (eg, chest tube insertion),¹⁷ or on specific quality improvement techniques, such as root cause analysis.^{18 19} With the exception of those that focus on error disclosure, few studies use OSCEs to assess socio-cultural aspects of PS^{20 21} and most are discipline-specific (ie, only for physicians or only for nurses).

We developed and pilot tested a 4-station OSCE focused on the socio-cultural dimensions of the CPSI Safety Competency Framework⁵ that assessed both nursing and medical students at level 2 of educational outcomes, as outlined by Kirkpatrick²² and used by Barr.²³ Level 2 outcomes reflect the degree to which learners acquire knowledge and skills. Using Miller's framework for clinical assessment,²⁴ OSCEs and the use of standardised patients allow learners to demonstrate competence (knows how to do something) and performance (shows how).

METHODS

OSCE development

Expert facilitators from the Medical Council of Canada led a 1.5 day case writing workshop in June 2013. Seven faculty members from nursing and medicine with content expertise in PS participated and each one submitted a case by workshop close. These cases were revised by a trained case writer in conjunction with the study authors. The revised cases were then appraised by clinicians with expertise in OSCE or PS who had not previously seen the cases. This second group assessed each case based on (1) realism and (2) suitability for both nursing *and* medical students. Using their feedback, four cases were selected for pilot testing. In addition to their feedback, the number of PS competencies (box 1) reflected in each case, the balance of competencies across cases and the simulation requirements for each case were taken into account. Specifically, cases requiring standardised clinicians needed to be limited as these were deemed more challenging to portray and pilot.

Setting and subjects

All third and fourth year students enrolled in an undergraduate nursing programme or undergraduate medical programme at a Canadian university were invited to participate in a 'voluntary research study focused on PS knowledge and skills'. They were assured that their data would be anonymised and would in no way be connected to their educational record. Students were offered a \$20 coffee card and dinner as a modest inducement to participate. Eighteen third year students volunteered—eight nursing students and 10 medical students. The study received approval from the relevant research ethics board at the university which took part in the pilot and from the lead author's institution.

Stations

Each station was based on one case from the workshop and reflected at least four of the competency areas in the Safety Competency Framework.⁵ Station 1 required learners to uncover a deep vein thrombosis (DVT) near miss and then explain the system factors that led to the near miss to the patient's spouse. Station 2 involved team dynamics and communication with a patient around a complex discharge, while station 3 required learners to persist in an interaction with a dismissive, time-pressured staff physician. Station 4 required learners to discuss an insulin overdose with the patient including how it occurred and how similar events might be prevented in the future. Learners moved through each station one at a time. Competency area 2 (work in teams for PS) was evaluated based on the learner's interaction with the standardised patient/person (SP) as well as the extent to which the learner was respectful and professional in his or her description of the role other team members

Table 1 Station descriptions

Station	Case name	Safety context	PS competency areas assessed*†‡§¶	Standardised person
1	Jim the Roofer	Near miss (DVT)— <i>Learner is asked to deliver a normal test result; patient's wife has received discrepant information about the test</i>	Communication*, teamwork†, managing risk‡, R, R and disclose§, culture¶	Patient's wife
2	Bertha's not ready to go	Complex discharge— <i>Staff tells learner to discharge recovering stroke patient due to hospital capacity pressure; patient expresses several concerns about leaving so soon</i>	Communication*, teamwork†, managing risk‡, culture¶	Patient
3	Jackson is in pain	Challenging authority— <i>Learner has concerns about a patient's pain and requests staff physician's assessment; superior is busy and dismissive</i>	Communication*, teamwork†, managing risk‡, culture¶	Staff physician
4	Giny's insulin	Medication error disclosure— <i>Learner is asked to communicate an insulin overdose; patient expresses concerns and has questions</i>	Communication*, teamwork†, R, R and disclose§, culture¶	Patient

*Communicate effectively for PS.

†Work in teams for PS.

‡Manage safety risks.

§Recognise, respond to and disclose adverse events.

¶Contribute to a culture of safety.

DVT, deep vein thrombosis; PS, patient safety.

(not present in the station) played in the case (eg, they did not blame others). The safety context and competency areas of each station are provided in table 1. Standardised persons were recruited and trained as per the university simulation centre's protocol. Stations were 8 min long. In stations 1–3, a 6-min buzzer was incorporated and the last 2 min were reserved for specified oral questions administered by an assessor (as described below). There were 2 min between stations for learners to read the scenario and task(s) for the following station. Learners completed all four stations in 40 min.

Assessors and scoring

Each station was scored on four or five relevant PS competency dimensions (see table 1). A single 5-point global rating scale was adopted²⁵ and used for all: 0=COMPETENCE NOT DEMONSTRATED for a learner at this level—Skills: Clearly deficient; 1=BORDERLINE UNSATISFACTORY for a learner at this level—Skills: Somewhat deficient; 2=BORDERLINE SATISFACTORY for a learner at this level—Skills: Just adequate; 3=COMPETENCE DEMONSTRATED for a learner at this level—Skills: Good; 4=ABOVE the level expected of a learner at this level—Skills: Excellent. The decision to use rating scales alone aligns with the idea that mastery of the parts (ie, discrete skills on a checklist) does not indicate competency of the whole. While 'global ratings provide a more faithful reflection of expertise than detailed checklists'²⁶ and expert assessors are able to evaluate these holistic skills,²⁷ assessors were also provided with two to three concrete behaviours per competency to look for when rating learners. The intent was to improve rater agreement by providing assessors with a common framework for each case.

Scoring in each station was done by two assessors who were both in the exam room. Each assessor was a current or retired clinical faculty member from nursing or medicine. The primary assessor also had expertise in PS science and administered the oral questions. The oral questions allowed learners to demonstrate knowledge or behaviours that may not have been elicited in their interaction with the SP. For example in station 1, regarding the DVT near miss, the oral question explicitly asked about the system issues at play (which is core PS knowledge for the 'Contribute to a culture of patient safety' domain) since 'system factors' are not something the SP spouse could realistically question. Assessors were instructed to reflect on both the learner's interaction with the SP and on the learner's response to the oral questions when rating learner's competencies.

All assessors participated in a 1-h training session to familiarise them with the OSCE goals and the rating guidelines. To promote a common understanding of the rating scale anchors and to guide assessor discrimination between rating scale options, assessors were given a 'Guidelines to Rating Sheet' that provided non-case-specific competency definitions, boundaries, and descriptors of COMPETENT or ABOVE performance. They also received case-specific examples for each competency dimension to guide their observations. A calibration video (of a station 3 dry run) was scored by all assessors, and then discussed so they could calibrate their ratings. Assessor dyads were asked not to discuss their ratings with one another during the OSCE. For a more detailed description of one station, including the specific behaviours tied to each of the four competency areas assessed in the case, see the online supplementary case appendix.

During the OSCE, it became clear that one assessor in station 2 was in a conflict of interest with one

group of students. As a result, station 2 was rescored. The station 1 assessor dyad received refresher training and then rescored station 2 based on video recordings. Data presented here use the second set of scores for station 2 (see online supplementary technical appendix, Station 2 Reassessment for one exception).

Analysis

Reliability refers to the reproducibility of scores obtained from an assessment. We examined within-station internal consistency reliability (Cronbach's α).²⁸ Inter-rater reliability was examined using the equivalent²⁹ of a weighted κ .³⁰ A weighted κ is typically used for categorical data with an ordinal structure and takes into account the magnitude of a disagreement between assessors (eg, the difference between the first and second category is less important than a difference between the first and third category; see online supplementary technical appendix, Inter-rater reliability). Two-way mixed, average-measures, agreement intraclass correlation coefficient (ICC) is also reported as it provides a measure of the score reliability. For interpretation of ICCs, including weighted κ , Cicchetti's classification (inter-rater reliability less than 0.40 is poor; 0.40–0.59 is fair; 0.60–0.74 is good; 0.75–1.00 is excellent)³¹ was used.

Given that there was strong assessor agreement, the mean of the two assessors' scores became the station score for each candidate. Scores between nursing and medical students were compared with an independent samples t test for each station. Finally, the proportion of learners with above and below borderline competence scores for 1, 2, 3 and all 4 stations is reported.

RESULTS

Reliability

Table 2 shows inter-rater reliability (weighted κ) and station score reliability given two assessors (average

measures ICC). The weighted κ columns show that assessor agreement was good to excellent for half of the competency rating scales in all four stations. When calculated on the station scores (the mean of the competency area ratings within a station), assessor agreement was excellent for all four stations. Table 2 (average-measures ICC columns) indicates good score reliability for all four stations (Hallgren—personal communication) as all values exceed 0.8. Finally, within-station reliability was strong for all four stations (α s=0.881, 0.859, 0.893 and 0.767 for stations 1 through 4, respectively).

Station scores

The boxplot in figure 1 shows station score variation by HP group. The whiskers show the score range for the top and bottom 25% of learners, the box shows the IQR (the middle 50% of scores for the group), and the line in the box is the median. Figure 1 shows that the proportion of learners who achieved a station score above borderline satisfactory (rating=2) varied across stations and by HP learner group.

The independent samples t test comparing nursing and medical students was statistically significant for stations 1, 2 and 3 ($t(16)=-3.55$, $p<0.01$, $d=1.65$; $t(16)=-3.83$, $p<0.01$, $d=1.81$; $t(16)=-2.27$, $p<0.05$, $d=1.03$ for stations 1, 2 and 3, respectively). Table 3 shows that students in the nursing group scored lower than medical students on the DVT near miss (station 1), the complex discharge (station 2) and challenging authority (station 3). All of these effects exceed Cohen's (1988) convention for a large effect size ($d=0.80$). There was no difference in nursing and medical students' scores on station 4—medication error ($t(16)=-0.23$, $p=0.82$, $d=0.1$).

Last, overall performance across all four stations was considered. One student (6%) did not achieve a station score of 3 or higher on any of the stations

Table 2 Agreement of pairs of assessors and score reliability (across competency dimensions and stations)

	Station 1		Station 2		Station 3		Station 4	
	Near miss (DVT)		Complex discharge		Challenging authority		Medication error disclosure	
PS competency area*†‡§¶	Weighted κ	Avg measures ICC	Weighted κ	Avg measures ICC	Weighted κ	Avg measures ICC	Weighted κ	Avg measures ICC
Communication*	0.517	0.608	0.660	0.789	0.717	0.815	0.432	0.547
Teamwork†	0.665	0.809	0.135	0.247	0.496	0.659	0.438	0.621
Managing risk‡	0.794	0.888	0.637	0.692	0.869	0.933	NA	NA
R, R and disclose§	0.568	0.716	NA	NA	NA	NA	0.548	0.714
Culture¶	0.560	0.712	0.693	0.795	0.390	0.502	0.754	0.860
Station score	0.803	0.894	0.737	0.811	0.794	0.870	0.816	0.895

*Communicate effectively for PS.

†Work in teams for PS.

‡Manage safety risks.

§Recognise, respond to and disclose adverse events.

¶Contribute to a culture of safety.

DVT, deep vein thrombosis; ICC, intraclass correlation coefficient; PS, patient safety.

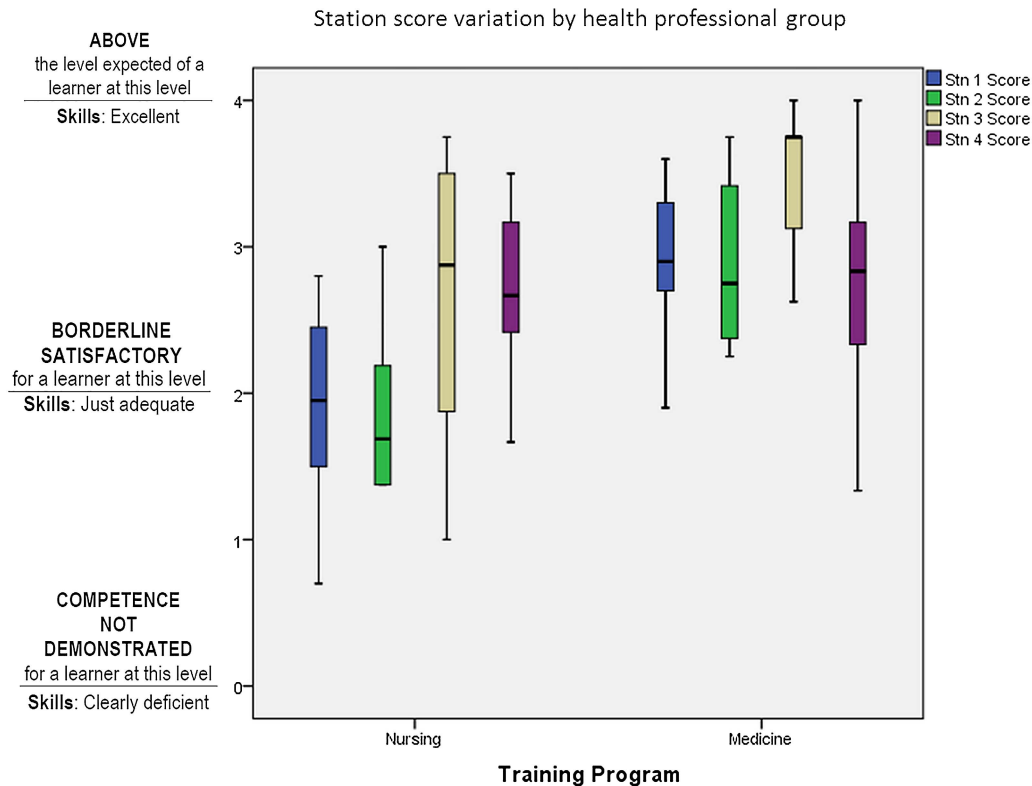


Figure 1 Station score variation by health professional group.

(a score of 3=COMPETENCE DEMONSTRATED for a learner at this level—Skills: Good) and eight students (44%) received a station score of 3 or higher on only one station. Five students (28%) received a station score of 3 or higher on two stations, and four students (22%) received a station score of 3 or higher on three or four stations. In all, 39% of students (7/18) received an overall station score below two on at least one station (a score of 2=BORDERLINE SATISFACTORY for a learner at this level—Skills: Just adequate).

DISCUSSION

OSCEs are a familiar, widely used method of assessing performance for HPs. Given the increasing interest in teaching PS in undergraduate HP curriculum, this

study explored the use of OSCE for assessing aspects of this domain. Our results demonstrate that sufficiently reliable station scores can be achieved in OSCE stations designed to assess performance on socio-cultural aspects of PS. High levels of inter-rater reliability for the overall station scores can also be achieved, indicating a minimal amount of measurement error was introduced by independent assessors. Given this finding, future studies may only require a single assessor in each station. Although the assessors were all clinical faculty, only one assessor per dyad had PS expertise. Sometimes, the PS expert was nursing faculty, sometimes medical faculty. Our experience suggests that clinical faculty in either nursing or medicine, with as little as 1 h of training, can provide reliable assessments of HP students' PS competence in an OSCE such as this one.

Our experience also suggests that the global ratings, although to some degree subjective, are not inherently unreliable.²⁷ From a validity standpoint, competency assessment in the context of PS requires global ratings of holistic expertise rather than evaluation of discrete skills.²⁷ Using an OSCE scored with global rating scales is consistent with two important trends identified by van der Vleuten and Schuwirth:²⁷ (1) a growing emphasis on general professional competencies (such as teamwork and awareness of the roles of culture and complex systems) in HP education and (2) recognition of the value of learning skills in an integrated fashion. The OSCE described here has a veneer of required

Table 3 OSCE Performance by station by learner group

Station	Medical students station score M (SD)	Nursing students station score M (SD)
1. Near miss (DVT)	2.90 (0.48)*	1.91 (0.70)
2. Complex discharge	2.92 (0.57)*	1.86 (0.60)
3. Challenging authority	3.49 (0.48)**	2.66 (1.03)
4. Medication error disclosure	2.78 (0.77)	2.71 (0.59)

*p<0.01, ** p<0.05.

DVT, deep vein thrombosis; OSCE, Objective Structured Clinical Examination.

clinical skills but asks learners to demonstrate competence in more multi-faceted socio-cultural aspects of PS.

The variability in scores across stations and between HP learner groups was seen as reasonable. Participants were third year medical and nursing students in the course of their regular training programme, with no additional PS curricular material. Their overall performance levels on this OSCE suggest that socio-cultural aspects of PS in undergraduate curricula may need to be enriched.

Differences between medical and nursing student performance were relatively large on three of the four stations. However, whether these differences stem from differences in their education (eg, most third year nursing students have less OSCE experience than third year medical students) or from the type of clinical encounters they are exposed to during training (eg, nursing students are more likely to be familiar with medication errors, the one context where nursing students' and medical students' scores did not differ) is not clear. Relatedly, differences in nursing and medical student scores may reflect differences in how they are/have been trained to fill their professional role (eg, nurses may not be trained to question authority or to consider and communicate about system issues).

The differences between medical and nursing student performance are interesting when compared with our previous work with new graduates.⁷ While medical student performance was higher than nursing student performance in the current study, the pattern is reversed for self-assessed confidence on the same socio-cultural dimensions (ie, self-assessed confidence was higher for new nurses than new physicians).⁷ This reversal fits with Eva and Regehr¹² who argue that self-assessment skills can be accurate in some contexts but not in others and therefore should be used alongside external assessments such as OSCEs.

Finally, the OSCE described here is suitable for students in nursing and medicine and may be of interest to training programmes seeking greater opportunities for inter-professional education. Our approach to station development could be relevant for other HP trainees while implementing feedback and debriefing opportunities around the OSCE would maximise the opportunity for inter-professional interactions.

Limitations/practical context for this work

With only four stations, we cannot report on the reliability of the total scores. Reliability at the OSCE level is largely a function of adequately sampling across conditions of measurement (eg, measuring the same competency in 10 or 15 different contexts (stations), with different assessors and standardised patients for each context²⁷). A second limitation of this study pertains to the rescoring of station 2 as there is a difference in the face-to-face scoring experience and video-based scoring.

There are challenges to implementing PS OSCEs as there is little space for 'additional material' in existing HP curricula. However, a high stakes PS OSCE could effectively drive what is taught and what is learned. Alternatively, the PS OSCE stations described could be embedded within existing OSCEs that measure aspects of socio-cultural competence, such as communication and teamwork.

Last, these OSCE cases surfaced the kinds of intimidating and obstructive negative behaviours that trainees routinely face in the clinical training environment and which shape a hidden curriculum that can threaten PS.³² Inclusion of a debriefing session in this socio-cultural PS OSCE would give learners a rare and rich opportunity to explicitly discuss negative aspects of the hidden curriculum. Without these kinds of opportunities, structural gaps in the curriculum that limit new graduates' capacity to improve care and prevent errors are likely to persist.³³

CONCLUSIONS

The OSCE is an underexplored methodology for assessing PS competencies. In this study, the OSCE assessed socio-cultural dimensions of PS based on realistic scenarios developed and reviewed by content experts. The scenarios were mapped to PS competencies and were applicable to both nursing and medicine. The station scores were highly reliable. Further, the score variation across stations and across nursing and medical students was appropriate. Of interest and still to be explored were the differences between nursing and medical student scores as it is unclear what factors would explain these differences. This study is the foundation for further work regarding the use of OSCEs for assessing competency in important socio-cultural dimensions of PS.

Author affiliations

¹School of Health Policy & Management, York University, Toronto, Ontario, Canada

²School of Nursing, Queen's University, Kingston, Ontario, Canada

³Department of Family Medicine (Emeritus), University of Calgary, Calgary, Canada

⁴Medical Council of Canada, Ottawa, Ontario, Canada

⁵Faculty of Education, Queen's University, Kingston, Ontario, Canada

⁶Department of Anesthesiology and Perioperative Medicine, School of Nursing, Queen's University, Kingston, Ontario, Canada

Acknowledgements We would like to acknowledge and thank our case writers, case reviewers and assessors: Ms Heather Campbell, Maryanne D'Arpino, Laura Goodfellow, Alicia Papanicolaou; Drs Sean Clarke, David Ginsburg, David Goldstein, Philip Hebert, Lindsey Patterson, and Lynfa Stroud; as well as the medical and nursing students who participated in the OSCE. We also thank Dr Nishan Sharma for observing and providing feedback and we are grateful to Dr Arthur Rothman for various suggestions pertaining to assessment and analysis.

Contributors LRG designed the study (including data collection tools and methods), monitored data collection, cleaned and analysed the data, and drafted and revised the paper. She is the guarantor. DT and PGN contributed to overall study design,

data collection as well as revising and approving the manuscript. SS and IdV contributed to OSCE design and analysis, revising and approving the manuscript. SSS contributed to assessment and analysis, revising and approving the manuscript. EGV, ML-F and JM contributed to overall study design, revising and approving the manuscript.

Funding This study was funded by a research grant from the Canadian Patient Safety Institute (CPSI), RFA09-1181-ON. CPSI did not play a role in the design, analysis, interpretation or write-up of the study.

Competing interests None.

Ethics approval The study received approval from the Human Participants Review Committee in the Office of Research Ethics at York University in Toronto and at Queen's University in Kingston (Health Sciences Research Ethics Board).

Provenance and peer review Not commissioned; externally peer reviewed.

Open Access This is an Open Access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>

REFERENCES

- Lucian Leape Institute. *Unmet needs: teaching physicians to provide safe patient care*. Boston: National Patient Safety Foundation, 2010.
- Department of Health. *Modernising medical careers. The new curriculum for the foundation years in postgraduate education and training*. London, UK: Department of Health, 2007.
- American Association of Colleges of Nursing. Hallmarks of quality and patient safety: recommended baccalaureate competencies and curricular guidelines to ensure high-quality and safe patient care. *J Prof Nurs* 2006;22:329–30.
- Cronenwett L, Sherwood G, Barnsteiner J, et al. Quality and safety education for nurses. *Nurs Outlook* 2007;55:122–31.
- Frank JR, Brien S; on behalf of The Safety Competencies Steering Committee. *The safety competencies: enhancing patient safety across the health professions*. Ottawa, ON: Canadian Patient Safety Institute, 2008.
- Wong BM, Etmells EE, Kuper A, et al. Teaching quality improvement and patient safety to trainees: a systematic review. *Acad Med* 2010;85:1425–39.
- Ginsburg LR, Tregunno D, Norton PG. Self-reported patient safety competence among new graduates in medicine, nursing and pharmacy. *BMJ Qual Saf* 2013;22:147–54.
- Ginsburg L, Castel E, Tregunno D, et al. The H-PEPSS: an instrument to measure health professionals' perceptions of patient safety competence at entry into practice. *BMJ Qual Saf* 2012;21:676–84.
- Regehr G, Hodges B, Tiberius R, et al. Measuring self-assessment skills: an innovative relative ranking model. *Acad Med* 1996;71(10 Suppl):S52–4.
- Davis DA, Mazmanian PE, Fordis M, et al. Accuracy of physician self-assessment compared with observed measures of competence: a systematic review. *JAMA* 2006;296:1094–102.
- Hodges B, Regehr G, Martin D. Difficulties in recognizing one's own incompetence: novice physicians who are unskilled and unaware of it. *Acad Med* 2001;76(10 Suppl):S87–9.
- Eva KW, Regehr G. Self-assessment in the health professions: a reformulation and research agenda. *Acad Med* 2005;80(10 Suppl):S46–54.
- Battles JB, Wilkinson SL, Lee SJ. Using standardised patients in an objective structured clinical examination as a patient safety tool. *Qual Saf Health Care* 2004;13(Suppl 1):i46–50.
- Salas E, Wilson KA, Burke CS, et al. Using simulation-based training to improve patient safety: what does it take?. *Jt Comm J Qual Patient Saf* 2005;31:363–71.
- Varkey P, Natt N. The objective structured clinical examination as an educational tool in patient safety. *Jt Comm J Qual Patient Saf* 2007;33:48–53.
- Pernar LI, Shaw TJ, Pozner CN, et al. Using an objective structured clinical examination to test adherence to Joint Commission National Patient Safety Goal-associated behaviors. *Jt Comm J Qual Patient Saf* 2012;38:414–18.
- Wagner DP, Hoppe RB, Lee CP. The patient safety OSCE for PGY-1 residents: a centralized response to the challenge of culture change. *Teach Learn Med* 2009;21:8–14.
- Gupta P, Varkey P. Developing a tool for assessing competency in root cause analysis. *Jt Comm J Qual Patient Saf* 2009;35:36–42.
- Varkey P, Natt N, Lesnick T, et al. Validity evidence for an OSCE to assess competency in systems-based practice and practice-based learning and improvement: a preliminary investigation. *Acad Med* 2008;83:775–80.
- Daud-Gallotti RM, Morinaga CV, Arlindo-Rodrigues M, et al. A new method for the assessment of patient safety competencies during a medical school clerkship using an objective structured clinical examination. *Clinics (Sao Paulo)* 2011;66:1209–15.
- Singh R, Singh A, Fish R, et al. A patient safety objective structured clinical examination. *J Patient Saf* 2009;5:55–60.
- Kirkpatrick DL. Evaluation of training. In: Craig R, Bittel L, eds. *Training and development handbook*. New York: McGraw Hill, 1967:87–112.
- Barr H. Competent to collaborate: towards a competency-based model for interprofessional education. *J Interprof Care* 1998;12:181–7.
- Miller GE. The assessment of clinical skills/competence/performance. *Acad Med* 1990;65(9 Suppl):S63–7.
- National Assessment Collaboration (NAC). <http://mcc.ca/examinations/nac-overview/> (accessed 13 May 2014).
- Regehr G, MacRae H, Reznick RK, et al. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Acad Med* 1998;73:993–7.
- van der Vleuten CJ, Schuwirth LW. Assessing professional competence: from methods to programmes. *Med Educ* 2005;39:309–17.
- Brannick MT, Erol-Korkmaz HT, Prewett M. A systematic review of the reliability of objective structured clinical examination scores. *Med Educ* 2011;45:1181–9.
- Hallgren KA. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutor Quant Methods Psychol* 2012;8:23–34.
- Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull* 1968;70:213–20.
- Cicchetti D. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol Assess* 1994;6:284–90.
- Hafferty FW. Beyond curriculum reform: confronting medicine's hidden curriculum. *Acad Med* 1998;73:403–7.
- Aron DC, Headrick LA. Educating physicians prepared to improve care and safety is no accident: it requires a systematic approach. *Qual Saf Health Care* 2002;11:168–73.