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

Liane R Ginsburg, Deborah Tregunno, Peter G Norton, Sydney Smee, Ingrid de Vries, Stefanie S Sebok, Elizabeth G VanDenKerkhof, Marian Luctkar-Flude, Jennifer Medves

ABSTRACT
Background Patient safety (PS) receives limited attention in health professional curricula. We developed and pilot tested four Objective Structured Clinical Examination (OSCE) stations 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. Yet, PS receives limited attention in HP curricula. Moreover, when asked about their confidence in PS learning, new graduates in medicine, nursing and pharmacy identify a number of gaps in PS learning.

The Canadian Patient Safety Institute (CPSI) Safety Competency Framework 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. Self-reported PS competence provides one important and cost effective source of information about the effectiveness of the HP education process, including aspects of an individual’s PS knowledge that need attention. However, a recent systematic review concluded that physicians have a limited ability to accurately self-assess and this may be particularly true among...
the least skilled. 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.

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. Most PS OSCE studies have focused on clinical aspects of PS, such as hand hygiene, medication labelling and safe performance of procedures (eg, chest tube insertion), or on specific quality improvement techniques, such as root cause analysis. With the exception of those that focus on error disclosure, few OSCE studies have focused on clinical aspects of PS, and many have engaged in internal reflection and also received feedback from external sources.

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. Most PS OSCE studies have focused on clinical aspects of PS, such as hand hygiene, medication labelling and safe performance of procedures (eg, chest tube insertion), or on specific quality improvement techniques, such as root cause analysis. With the exception of those that focus on error disclosure, few OSCE studies have focused on clinical aspects of PS, and many have engaged in internal reflection and also received feedback from external sources.

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.
was to improve rater agreement by providing assessors with two to three concrete behaviours per competency area. A single 5-point global rating scale was adopted and used for all: 0 = COMPETENCE NOT DEMONSTRATED for a learner at this level; 1 = BORDERLINE UNSATISFACTORY for a learner at this level; 2 = BORDERLINE SATISFACTORY for a learner at this level; 3 = COMPETENCE DEMONSTRATED for a learner at this level—Skills: Just adequate; 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’ and expert assessors are able to evaluate these holistic skills, 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 $\alpha$).28 Inter-rater reliability was examined using the equivalent29 of a weighted $\kappa$.30 A weighted $\kappa$ 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 reliability. T wo-way mixed, average-measures, agreement ICC is also reliability intraclass correlation coefficient (ICC) is also reliability intraclass correlation coefficient (ICC). T wo-way mixed, average-measures, agreement ICC is also reliability intraclass correlation coefficient (ICC). T wo-way mixed, average-measures, agreement ICC is also reliability intraclass correlation coefficient (ICC). T wo-way mixed, average-measures, agreement ICC is also reliability intraclass correlation coefficient (ICC). T wo-way mixed, average-measures, agreement ICC is also reliability intraclass correlation coefficient (ICC). T wo-way mixed, average-measures, agreement ICC is also reliability intraclass correlation coefficient (ICC). T wo-way mixed, average-measures, agreement ICC is also reliability intraclass correlation coefficient (ICC). T wo-way mixed, average-measures, agreement ICC is also reliability intraclass correlation coefficient (ICC).

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 $\kappa$) and station score reliability given two assessors (average measures ICC). The weighted $\kappa$ 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 ($\alpha=0.881, 0.859, 0.893$ 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 exceeded 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)

<table>
<thead>
<tr>
<th>PS competency area*†‡¶</th>
<th>Station 1</th>
<th>Station 2</th>
<th>Station 3</th>
<th>Station 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Near miss (DVT)</td>
<td>Complex discharge</td>
<td>Challenging authority</td>
<td>Medication error disclosure</td>
</tr>
<tr>
<td></td>
<td>Weighted $\kappa$</td>
<td>Avg measures ICC</td>
<td>Weighted $\kappa$</td>
<td>Avg measures ICC</td>
</tr>
<tr>
<td>Communication*</td>
<td>0.517</td>
<td>0.608</td>
<td>0.660</td>
<td>0.789</td>
</tr>
<tr>
<td>Teamwork†</td>
<td>0.665</td>
<td>0.809</td>
<td>0.135</td>
<td>0.247</td>
</tr>
<tr>
<td>Managing risk‡</td>
<td>0.794</td>
<td>0.888</td>
<td>0.637</td>
<td>0.692</td>
</tr>
<tr>
<td>R, R and disclose§</td>
<td>0.568</td>
<td>0.716</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Culture¶</td>
<td>0.560</td>
<td>0.712</td>
<td>0.693</td>
<td>0.795</td>
</tr>
<tr>
<td>Station score</td>
<td>0.803</td>
<td>0.894</td>
<td>0.737</td>
<td>0.811</td>
</tr>
</tbody>
</table>

*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. From a validity standpoint, competency assessment in the context of PS requires global ratings of holistic expertise rather than evaluation of discrete skills. Using an OSCE scored with global rating scales is consistent with two important trends identified by van der Vleuten and Schuwirth: (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...
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. 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). This reversal fits with Eva and Regehr who argue that self-assessed 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 context). A second limitation of this study pertains to the rescaling 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. 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.

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**


Station Example - Supplemental On-line Only Appendix

This supplemental appendix provides a high-level breakdown of one of the four stations to show how the case was developed to assess four competency domains. We provide a summary of the key aspects of the case and links to the specific domains assessed in this station.

Station Purpose. To assess communication, risk assessment, safety culture and teamwork skills required to optimize safe patient care in the context of a complex discharge and interdisciplinary conflict.

Case Overview. The student instructions explain the student has just come from interdisciplinary rounds on a stroke rehab floor in an acute care setting in which the student presented an update on a 78-year medically stable female (Bertha Waller) who suffered a stroke three weeks earlier (The student’s supervisor had been called away). Mrs. Waller lives alone, has right-sided weakness and some short-term memory problems as a result of the stroke but is progressing well with therapy. At the close of rounds the plan is for discharge in 5-7 days. However, after rounds the student is told that, due to “pressure on inpatient beds”, Mrs. Waller needs to be discharged tomorrow instead of next week. This student is looking after Mrs. Waller today and has to share this information with her. The student is instructed to enter Mrs. Waller’s room and notify her that they want to discharge her tomorrow and explore any concerns she may have. There are two buzzer questions at 6 minutes that ask about (1) next steps for the patient and (2) the factors contributing to this patient safety challenge.

The SP. The SP is given several things she ‘must say/ask’ including that she is not ready to go, she was told by the physiotherapist she wouldn’t be ready to go home before next week, she doesn’t understand who makes these decisions, she questions why she was told she’d stay another week, she is afraid she can’t manage at home, etc.

Competency Domains. Four competency domains are assessed in this station. Each domain is listed in the table below along with specific things the assessor can ‘look for’ to help in arriving at a rating for each of these four domains.

<table>
<thead>
<tr>
<th>Competency Domains Assessed</th>
<th>“Look For” linked to each competency domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Communicate effectively for patient safety</td>
<td>• Reassures patient they won’t go home unless safety is ensured</td>
</tr>
</tbody>
</table>
| • Work in teams for patient safety    | • Candidate listens attentively to patient’s concerns  
• Is respectful of team members’ divergent perspectives in conversation with patient (e.g. does not blame others) |
| • Manage safety risks                 | • Advocates for individual patient in conversation with assessor  
• States actions required to ensure safe discharge                                                           |
| • Contribute to a culture of patient safety | • Explains system issues that put pressure on discharges                                                         |
Inter-rater Reliability
Inter-rater reliability was examined using the equivalent of a weighted Kappa—a two-way mixed, single-measures, consistency ICC. A weighted Kappa takes into account the magnitude of a disagreement between raters and is typically used for categorical data with an ordinal structure. Quadratic weights are used if the difference between the first and second category is less important than a difference between the first and third category, etc., (which was the case with our data).

Station 2 Reassessment
All data presented in the paper for station 2 are data taken from the reassessment of station 2 that was carried out (due to the fact that one assessor was in a conflict of interest position and realized it was difficult to provide objective assessments during the OSCE). However, due to the fact that we did not use student identifiers that could be linked back to the students in this study, we were unable to link the station 2 reassessment scores with students’ scores on the other three stations. This did not pose a problem for our analyses because this pilot had only four stations and we were therefore not seeking to comment on the reliability of the OSCE as a whole. Instead, our analyses tend to be within station analyses. One exception are the data presented in Table 3 on the overall performance of the trainees on the group of stations. We wanted to be able to report on the proportion of trainees that scored very high (and very low) on several stations. In order to do this, we have, in table 3, used data from the initial rater in station 2 who was not in a conflict position.

Table 3 Calculation
Table 3 provides data on overall performance on the group of stations and shows that ~=25% of trainees scored greater than 3 on three or four stations. If a trainee had a mean station score >= 2.95 the score was rounded to 3 and they were included as scoring in the 3+ range for the station.

References