CANADIAN MEDICAL EDUCATION JOURNAL | REVUE CANADIENNE DE L'ÉDUCATION MÉDICALE

2026, 17(2): https://doi.org/10.36834/cmej.81229

SCIENTIFIC REPORTS

The development of two Point of Care Ultrasound stations for Objective Structured Clinical Examinations in undergraduate medical education

Ryan Good,[1] Juliana Wilson,[2] Joshua C Kaine,[3] Vijay J Daniels,[4] Janeve Desy,[5] Joshua Lloyd,[6] Gillian Sheppard,[7] Linden Kolbenson,[8] Irene WY Ma,[5] Arthur Au,[9] Paul Olszynski[10]

*Author information is provided in the back matter of this manuscript

Abstract

Introduction: Point-of-care ultrasound (POCUS) is a valuable clinical skill that improves clinical care but requires substantial training. Validated assessment tools provide empirical evidence regarding trainee performance while also informing program-level evaluation. We developed two POCUS-specific stations for objective structured clinical examinations (OSCEs) to assess skill acquisition and inform best practices in undergraduate medical education.

Methods: A multidisciplinary group of POCUS educators identified two POCUS applications (pleural effusion and abdominal free fluid) well suited for the undergraduate level. A modified Delphi approach was used to develop POCUS-application-specific skill checklists and global rating scale. Two medical programs piloted the stations to inform reliability.

Results: Across two sites, 46 and 41 students participated in the pleural effusion and abdominal free fluid stations respectively. Checklists showed high internal reliability, with Cronbach's alpha of 0.85 (95% CI 0.71-0.93) for the pleural effusion station and 0.87 (95% CI 0.74-0.95) for the abdominal free fluid station. Krippendorff's alpha, a measure of inter-rater reliability, was also equally strong at 0.85 (95% CI 0.43-0.94) and 0.83 (95% CI 0.50-0.94) respectively.

Conclusion: Both POCUS OSCE stations demonstrated good internal and inter-rater reliability. Deployment of these OSCE stations at programs with integrated POCUS curricula may help refine programming and training expectations.

Développement de deux stations d'échographie au point d'intervention pour les examens cliniques objectifs structurés dans le cadre de la formation médicale de premier cycle

Résumé

Introduction : L'échographie au point d'intervention (POCUS) est une compétence clinique précieuse qui améliore les soins cliniques, mais qui nécessite une formation approfondie. Des outils d'évaluation validés fournissent des preuves empiriques concernant les performances des médecins résidents tout en éclairant l'évaluation au niveau du programme. Nous avons développé deux stations propres à POCUS pour les examens cliniques objectifs structurés (ECOS) afin d'évaluer l'acquisition des compétences et d'éclairer les meilleures pratiques dans l'enseignement médical de premier cycle.

Méthodes : Un groupe multidisciplinaire d'éducateurs POCUS a identifié deux applications POCUS (épanchement pleural et liquide libre abdominal) bien adaptées au niveau de premier cycle. Une approche Delphi modifiée a été utilisée pour développer des listes de contrôle des compétences spécifiques à l'application POCUS et une échelle d'évaluation globale. Deux programmes médicaux ont testé les stations afin d'en vérifier la fiabilité.

Résultats : Sur les deux sites, 46 et 41 étudiants ont respectivement participé aux stations consacrées à l'épanchement pleural et au liquide libre abdominal. Les listes de contrôle ont montré une fiabilité interne élevée, avec un coefficient alpha de Cronbach de 0,85 (IC à 95 % : 0,71-0,93) pour la station consacrée à l'épanchement pleural et de 0,87 (IC à 95 % : 0,74-0,95) pour la station consacrée au liquide libre abdominal. Le coefficient alpha de Krippendorff, qui mesure la fiabilité inter-évaluateurs, était également élevé, avec respectivement 0,85 (IC à 95 % : 0,43-0,94) et 0,83 (IC à 95 % : 0,50-0,94).

Conclusion : Les deux stations POCUS ECOS ont démontré une bonne fiabilité interne et inter-évaluateurs. Le déploiement de ces stations ECOS dans des programmes intégrant des cursus POCUS pourrait contribuer à affiner les attentes en matière de programmation et de formation.

Introduction

Point-of-care ultrasound (POCUS) improves diagnostic accuracy and facilitates procedures.[1] Most medical specialties have adopted POCUS as a valued clinical skill,[2] and several medical schools have integrated it into undergraduate training.[3,4] Over half of the medical schools in the United States and Canada include POCUS in their undergraduate medical curriculum.[5,6] However, programs vary widely in terms of POCUS instruction as no broadly recognized training requirements exist. Although consensus-based expert recommendations provide extensive lists of POCUS objectives for undergraduate medical education (UME),[7,8,9] they lack an empirical basis to inform breadth of instruction and training expectations.[10] Medical schools often develop (and deploy) assessments in isolation, lacking generalizability to inform expectations and best practices across institutions.[4,6,10] Insufficient evidence to describe students' POCUS performance leaves uncertainty about what POCUS skills and knowledge can realistically be learned (and assessed) within crowded medical curricula.

The objective structured clinical examination (OSCE) standardizes the assessment of learners' clinical skills, attitudes, problem-solving, and knowledge application.[11,12] OSCEs have proven feasible and reliable for both formative and summative assessment in UME[13] and often inform program evaluation. Assessment of POCUS skills should include a combination of modalities such as OSCEs to assess both image acquisition and interpretation.[10,14,15] However, only 18 of 122 medical schools in the United States use OSCEs to assess students' POCUS skills.[6]

Many postgraduate programs deem POCUS an entrustable professional activity (EPA).[16,17,18] The Association of American Medical Colleges and the Association of Faculties of Medicine of Canada also expect medical graduates to perform a range of clinical duties,[20,21] several of which warrant use of POCUS. Thus, integrating POCUS into UME is supported but the extent of instruction and associated training expectations remain undetermined. Validated assessment tools can provide empirical evidence to inform program-level expectations regarding realistic training outcomes. We brought together a multidisciplinary group of POCUS educators and assessment experts to develop two POCUS OSCE stations for use in OSCEs in UME. We aim to integrate these stations into OSCEs at our various institutions with the results informing both training outcomes and instructional resources.

Methods

Both the Colorado Multiple Institutional Review Board (COMIRB #17-0656) and the Indiana University Institutional Review Board (IRB #21754) deem this study exempt from full board review. We provided participants with information about the study but did not require them to sign consent.

OSCE station development

Following best practices,[11] we convened a multidisciplinary panel of POCUS and assessment experts from both Canada and the USA. The group communicated electronically via email and hosted virtual meetings (Zoom, USA). The group determined the most appropriate POCUS applications to assess, created a suitable marking scheme, adapted that marking scheme to reflect the spectrum of entrustability, and identified the knowledge and skills a minimally capable student should demonstrate for each level of entrustment.

Station checklists

We determined that two POCUS applications, the detection of pleural effusion and abdominal free fluid, were ubiquitous in clinical care and well-suited for assessment.[22] These applications include foundational POCUS concepts. The group first identified the knowledge items and tasks required to be entrustable in each application. After creating draft checklists, we conducted a modified Delphi process with a series of votes to determine which checklist items a minimally competent trainee should demonstrate at each level of entrustment. The levels ranged from needing frequent guidance while performing the task, to only needing occasional prompts, to independent performance (see Table 1). We set consensus at 80% and did not predetermine a maximum number of rounds.

Setting and subjects

We piloted the two OSCE stations with respective checklists at two US medical schools. An OSCE Working Group (Canadian site) also peer-reviewed the stations to help develop the global rating scale. Students had seven minutes to complete the station. All checklist items were weighted equally and scored in binary fashion (performed or not performed). We used feedback from these experiences to prepare the final versions of the OSCE stations, skills checklists, and global rating scale. We conducted the pilot studies following POCUS educational sessions as part of mandatory longitudinal POCUS programming and assessment at each institution.

Table 1. Expected entrustment score alignment with curriculum offering; Trainee POCUS experience at each entrustment score

Entrustment Score	Descriptor	Curriculum offering within UME	Trainee profile with regard to POCUS
1	Trainee cannot perform any part of the scan	No POCUS training	Trainee has had no experience with POCUS
2	Trainee requires constant guidance to perform the scan	Introductory session +/- brief hands-on experience	Trainee has minimal experience with this POCUS application
3	Trainee is able to perform the scan with help of occasional prompts	Introductory session and both supervised and unsupervised practice	Trainee has been introduced to and has practiced both during course time (supervised) and on their own
4	Trainee can perform the scan but lacks efficiency and/or confidence	Introductory session, supervised and unsupervised practice as well as practice in clinical setting	Trainee has practiced this POCUS application several times and has had opportunities for supervised clinical integration.
5	Trainee performs the scan well and efficiently. I did not need to be there	Introductory session, supervised and unsupervised practice, repeated use in clinical care with feedback	Trainee has had a lot of practice (including during clinical care) and is considered entrustable (outgoing resident)

Analysis

We analyzed the results using descriptive statistics, including the median and interquartile range (IQR). We tested for intra-station reliability using Cronbach's alpha to examine the internal structure of the OSCE station checklists during pilot testing. We also examined the inter-rater reliability (IRR) of the checklist using Krippendorff's alpha. After making minor modifications to the checklist following day one of the pilot, we conducted a second Cronbach's Alpha analysis to determine if internal validity remained.

Results

Determination of checklists (Modified Delphi)

We started by broadly describing and then matching entrustment scores to the amount of POCUS training needed to achieve the respective level of entrustment (see Table 1). This first step allowed us to then proceed with the Modified Delphi and create four distinct checklists for entrustment scores 2-5. After three rounds of voting, review, and discussion, we finalized the checklists (see Appendix A).

Pilot testing

A total of 46 students participated in the pleural effusion station, and 41 students participated in the abdominal free fluid station at the two pilot sites (see Table 2). At site A, Version 1 of each checklist demonstrated high internal reliability, with a Cronbach's alpha of 0.85 (95% CI 0.71 to 0.93) for pleural effusion and 0.87 (95% CI 0.74 to 0.95) for abdominal free fluid. At site B, where paired assessors attended each station, Krippendorff's alpha (inter-rater reliability) was also high, at 0.85 (95% CI 0.43 to 0.94) for pleural effusion and 0.83 (95% CI 0.50 to 0.94) for abdominal free fluid. The scores for each reviewer were similar for both the pleural effusion (7.0 v 8.0, $p = 0.17$) and abdominal free fluid stations (8.0 v 10.0, $p = 0.54$).

Table 2. Pilot testing of both POCUS OSCE stations showing trainee performance according to checklists.

Site	Station	Version*	Number of Examiners	Number of Students	Number of Correct Checklist Items, Median (IQR)
A	Pleural effusion	1	1	16	9.5 (5-11)
A	Abdominal free fluid	1	1	15	11 (7-13)
B	Pleural effusion	1	2	15	7.5 (6.5-8.5)
B	Abdominal free fluid	1	2	11	8.5 (6.8-11)
B	Pleural effusion	2	1	15	4.5 (2.5-6)
B	Abdominal free fluid	2	1	15	5 (3.5-7)

**Version 1 was the original OSCE station checklist developed through the modified Delphi process. Version 2 was the final version after the initial pilot test. Modifications included removal of 1 checklist item (interpretation of scan on volunteer) and adjustment of the ultrasound machine settings between each student to compel the students to select the correct probe and adjust the gain/depth*

Students in the first pilot performed several checklist items, including using the correct transducer, appropriate use of gain, and appropriate use of depth. However, it was unclear if this was deliberate or due to setting adjustments made during the previous trainee encounter. Similarly, nearly all students correctly interpreted the normal findings in the student volunteer and on the sample images provided. We modified and then repeated the pilot after removing the item on interpreting normal findings in the student volunteer and resetting the ultrasound settings (probe, depth, gain) with each student encounter. Version 2 of each checklist showed similar internal reliability to Version 1 with a Cronbach's alpha of 0.85 (95% CI 0.71 to 0.94) for pleural effusion and 0.85 (95% CI 0.68 to 0.94) for the abdominal free fluid. Item level statistics including difficulty and discrimination index are provided in Appendix A. The inter-rater reliability for checklist items was not repeated as we only removed one item and did not add any novel ones.

Global rating scale

The global rating scale was developed after the creation of the checklists to guide standard setting (borderline regression method) by providing a more detailed and wholistic description of borderline satisfactory performance beyond a simple entrustment score (see Appendix B). Using the global ratings scale (GRS) for each station, only 4 of 15 students were deemed "satisfactory" for pleural effusion and 6 of 15 students were deemed "satisfactory" for abdominal free fluid. Those with a "satisfactory" score on the GRS had significantly higher scores on the checklist compared to those with a "not satisfactory" for both the pleural effusion (9.0 v. 3.0, $p < 0.05$) and abdominal free fluid station (8.0 v 4.0, $p < 0.05$).

Peer-review and testing of the stations by OSCE working group members led to similar suggestions. They suggested changing the video interpretation of the pleural effusion station to include both small and large pleural effusions. Additionally, they recommended modifying the video interpretation yearly. We brought these changes from the pilots and peer review to the expert panel via email, and the group reached consensus on incorporating the above changes.

Discussion

The expert panel developed two POCUS OSCE stations both of which demonstrated high internal reliability (Cronbach's Alpha) as well as high inter-rater agreement (Krippendorff's alpha). Furthermore, we developed a global rating scale to assist with standard setting. Lessons from performing the pilots included adjusting image optimization settings after each encounter and randomizing image interpretation findings.

While multiple-choice questions, visual exams and bedside assessments of skills are taking place,[10,23] we found no reports of integrated POCUS OSCE

stations during middle and high-stakes assessments. This occurs despite its relevance to several proposed EPAs including 1, 2, 3 & 5.[21] Relatedly, it remains unclear whether POCUS should be considered a pre-entrustable or entrustable clinical skill at the UME level. If the latter, then there must be reasonable grounds for performance expectations,[24] including sufficient opportunities for guided practice in an already crowded curriculum. Alternatively, if pre-entrustable, POCUS performance can and has been assessed within only one or two domains (knowledge of indications, image generation, image interpretation, and clinical integration).[25] This type of assessments represents less authentic assessment[12] compared to our integrated approach, but may be sufficient until training expectations are further refined and standardized.

We aim to deploy these stations in OSCEs at medical schools where POCUS has been integrated into the undergraduate curriculum. While we already have good validity evidence from the high intra-station reliability (Cronbach alpha) and high inter-rater reliability (Krippendorff's alpha), our next step will be to focus on standard setting. Standard-setting will be further informed through a modified borderline regression process using our global rating scale. This process will also add additional empirical data regarding the checklist pass marks.[26] Other validity evidence we may seek is comparing self-report experience of POCUS with station scores. Results may inform whether currently dedicated time and resources for POCUS are adequate for the desired training outcomes.

Limitations

We followed best practices, but we acknowledge that a modified Delphi technique is subject to inherent bias from expert opinion. Secondly, we developed and conducted the checklists and pilots at sites with integrated undergraduate POCUS curricula, which may bias towards higher performance expectations for students. We acknowledged this during the derivation of the skills list and global rating scale, and the experts were particularly thoughtful as to what defines minimally competent performance. Third, while our study presents validity evidence based on content, and response process and internal structure, relationship with other variables and consequences of testing were not explored and should be evaluated in future studies.

Conclusion

The panel developed two POCUS OSCE stations with good internal and inter-rater reliability and corresponding global rating scale. Deployment of these OSCE stations at programs with integrated POCUS curricula may help refine programming and training expectations.

Author information:

1- Department of Pediatrics, University of Colorado, Colorado, USA

2- Department of Emergency Medicine, University of Colorado, Colorado, USA

3- Department of Emergency Medicine, Indiana University School of Medicine, Indiana, USA

4- Department of Medicine, University of Alberta, Alberta, Canada

5- Department of Medicine, Cumming School of Medicine, University of Calgary, Alberta, Canada

6- College of Medicine, University of Saskatchewan, Saskatchewan, Canada

7- Department of Emergency Medicine, Memorial University, Newfoundland, Canada

8- Department of Internal Medicine, University of Saskatchewan, Saskatchewan, Canada

9- Department of Emergency Medicine, Thomas Jefferson University Hospital, Pennsylvania, USA

10- Department of Emergency Medicine, University of Saskatchewan, Saskatchewan, Canada

Correspondence to:

Paul Olszynski

email: p.olszynski@usask.ca

Published ahead of issue:

Jun 3, 2025

Conflict of Interest:

The authors have no competing interests to declare.

Funding:

There was no funding associated with this study.

Acknowledgements:

The authors would like to thank Drs. David Lewis, Frances Russell, Kirsten Weerdenburg, Luke Devine, and Pamela Soriano for their contributions to the expert panel.

References:

1. Diaz-Gomez JL, Mayo PH, Koenig SJ. Point-of-care ultrasonography. *N Engl J Med.* 2021;385:1593-1602. https://doi.org/10.1056/nejmra1916062
2. Sena A, Alerhand S, Lamba S. Milestone approach to designing a point-of-care ultrasound curriculum for transition-to-residency programs in the United States. *Teach Learn Med.* 2020;33(3):270-81. https://doi.org/10.1080/10401334.2020.1814296
3. Tarique U, Tang B, Singh M, Kulasegaram KM, Ailon J. Ultrasound curricula in undergraduate medical education: a scoping review. *J Ultrasound Med.* 2018;37(1):69-82. https://doi.org/10.1002/jum.14333
4. Davis JJ, Wessner CE, Potts J, Au AK, Pohl CA, Fields JM. Ultrasonography in undergraduate medical education: a systematic review. *J Ultrasound Med.* 2018;37(11):2667-79. https://doi.org/10.1002/jum.14628
5. Steinmetz P, Dobrescu O, Oleskevich S, Lewis J. Bedside ultrasound education in Canadian medical schools: a national survey. *Can Med Educ J.* 2016;7(1):e78-e86. https://doi.org/10.36834/cmej.36646
6. Russell FM, Zakeri B, Herbert A, et al. The state of point-of-care ultrasound training in undergraduate medical education: findings from a national survey. *Acad Med.*

2022;97(5):723-7. https://doi.org/10.1097/ACM.0000000000004512
7. Hoppmann RA, Mladenovic J, Melniker L, et al. International consensus conference recommendations on ultrasound education for undergraduate medical students. Ultrasound J. 2022;14:31. https://doi.org/10.1186/s13089-022-00279-1
8. Ma IWY, Steinmetz P, Weerdenburg K, et al. The Canadian medical student ultrasound curriculum: a statement from the Canadian Ultrasound Consensus for Undergraduate Medical Education Group. *J Ultrasound Med.* 2020;39(7):1279-87. https://doi.org/10.1002/jum.15218
9. Dinh VA, Lakoff D, Hess J, et al. Medical student core clinical ultrasound milestones: a consensus among directors in the United States. *J Ultrasound Med.* 2016;35(2):421-34. https://doi.org/10.7863/ultra.15.07080
10. DeBiasio C, Pageau P, Shefrin A, Woo MY, Cheung WJ. Point-of-care-ultrasound in undergraduate medical education: a scoping review of assessment methods. *Ultrasound J.* 2023;15(1):30. https://doi.org/10.1186/s13089-023-00325-6
11. Khan KZ, Ramachandran S, Gaunt K, Pushkar P. The objective structured clinical examination (OSCE): AMEE guide no. 81. Part I: an historical and theoretical perspective. *Med Teach.* 2013;35(9):e1437-e46. https://doi.org/10.3109/0142159x.2013.818634
12. Miller GE. The assessment of clinical skills/competence/performance. Acad Med. 1990;65(9):S63-S7. https://doi.org/10.1097/00001888-199009000-00045
13. Patrício MF, Julião M, Fareleira F, Carneiro AV. Is the OSCE a feasible tool to assess competencies in undergraduate medical education? *Med Teach.* 2013;35(6):503-14. https://doi.org/10.3109/0142159x.2013.774330
14. Sheppard G, Williams KL, Metcalfe B, et al. Using Kane's framework to build an assessment tool for undergraduate medical student's clinical competency with point of care ultrasound. *BMC Med Educ.* 2023;23:43. https://doi.org/10.1186/s12909-023-04030-9
15. Damewood SC, Leo M, Bailitz J, et al. Tools for measuring clinical ultrasound competency: recommendations from the Ultrasound Competency Work Group. *AEM Educ Train.* 2019;4(Suppl 1):S106-S12. https://doi.org/10.1002/aet2.10368
16. Royal College of Physicians and Surgeons of Canada. *Objectives of training in the specialty of emergency medicine.* Available at: https://www.royalcollege.ca/rcsite/documents/ibd/emergency_otr_e.pdf. [Accessed Dec 10, 2024].
17. Royal College of Physicians and Surgeons of Canada. *Entrustable Professional Activities for anesthesiology Version 3.0.* Available from: https://www.royalcollege.ca/rcsite/documents/cbd/epa-guide-anesthesiology-v3-e.pdf. [Accessed Dec 10, 2024].
18. Emergency ultrasound guidelines. *Ann Emerg Med.* 2009;53:550-70. https://doi.org/10.1016/j.annemergmed.2008.12.013
19. Blehar DJ, Barton B, Gaspari RJ. Learning curves in emergency ultrasound education. *Acad Emerg Med.* 2015;22(5):574-82. https://doi.org/10.1111/acem.12653
20. Association of American Medical Colleges. *Core EPAs guiding principles.* Core Entrustable Professional Activities for Entering Residency Curriculum Developers' Guide. Available at: https://store.aamc.org/downloadable/download/sample/sample_id/63/%20. [Accessed Dec 10, 2024].
21. Association of Faculties of Medicine of Canada. AFMC Entrustable Professional Activities for the Transition from Medical School to Residency. Available at: https://www.afmc.ca/wp-content/uploads/2022/10/AFMC_Entrustable-Professional-Activities_EN_Final.pdf. [Accessed Nov 30, 2024].
22. Steinmetz P, Oleskevich S, Dyachenko A, McCusker J, Lewis J. Accuracy of medical students in detecting pleural effusion using lung ultrasound as an adjunct to the physical examination. *J Ultrasound Med.* 2018;37:2545-52. https://doi.org/10.1002/jum.14612
23. Olszynski P, Russell M, Neufeld A, Malin G. The clinical ultrasonography elective in clerkship (CUSEC): a pilot elective for senior clerkship students at the University of Saskatchewan. *Can Med Educ J.*

2020;11(1):e144-6. https://doi.org/10.36834/cmej.61810
24. Kern DE, Thomas PA, Hughes MT. *Curriculum development for medical education: a six-step approach*. Baltimore: Johns Hopkins University Press; 2009.
25. Bahner DP, Hughes D, Royall NA. I-AIM: a novel model for teaching and performing focused sonography. *J Ultrasound Med.* 2012;31(2):295-300. https://doi.org/10.7863/jum.2012.31.2.295
26. Hejri SM, Jalili M, Muijtjens AM, Van Der Vleuten CP. Assessing the reliability of the borderline regression method as a standard setting procedure for objective structured clinical examination. *J Res Med Sci.* 2013;18(10):887-91.