# Relationship between ratings of performance in the simulated and workplace environments among emergency medicine residents

Nicholas Prudhomme, MD[*]; Michael O'Brien, MD, MSc.PT[*]; Meghan M. McConnell, PhD[†‡]; Nancy Dudek, MD, MEd[§]; Warren J. Cheung, MD, MMEd[*¶]

## CLINICIAN'S CAPSULE

**What is known about the topic?**
Resuscitation entrustable professional activities (EPAs) are assessed in workplace and simulated environments, but limited validity evidence exists for these assessments in either setting.

**What did this study ask?**
Do EPA F1 ratings improve over time, and is there an association between ratings in the workplace versus simulation environment?

**What did this study find?**
EPA ratings improved over time in both environments, but no correlation was observed. Ratings were higher in the workplace setting.

**Why does this study matter to clinicians?**
There is some validity evidence for EPA assessments in the simulated environment, but further studies are needed.

## ABSTRACT

**Objectives**: The Emergency Medicine (EM) Specialty Committee of the Royal College of Physicians and Surgeons of Canada (RCPSC) specifies that resuscitation entrustable professional activities (EPAs) can be assessed in the workplace and simulated environments. However, limited validity evidence for these assessments in either setting exists. We sought to determine if EPA ratings improve over time and whether an association exists between ratings in the workplace v. simulation environment.

**Methods**: All Foundations EPA1 (F1) assessments were collected for first-year residents (n = 9) in our program during the 2018–2019 academic year. This EPA focuses on initiating and assisting in the resuscitation of critically ill patients. EPA ratings obtained in the workplace and simulation environments were compared using Lin's concordance correlation coefficient (CCC). To determine whether ratings in the two environments differed as residents progressed through training, a within-subjects analysis of variance was conducted with training environment and month as independent variables.

**Results**: We collected 104 workplace and 36 simulation assessments. No correlation was observed between mean EPA ratings in the two environments (CCC(8) = -0.01; p = 0.93). Ratings in both settings improved significantly over time ($F(2,16) = 18.8$; $p < 0.001$; $\eta^2 = 0.70$), from 2.9 ± 1.2 in months 1–4 to 3.5 ± 0.2 in months 9–12. Workplace ratings (3.4 ± 0.1) were consistently higher than simulation ratings (2.9 ± 0.2) ($F(2,16) = 7.2$; $p = 0.028$; $\eta^2 = 0.47$).

**Conclusions**: No correlation was observed between EPA F1 ratings in the workplace v. simulation environments. Further studies are needed to clarify the conflicting results of our study with others and build an evidence base for the validity of EPA assessments in simulated and workplace environments.

## RÉSUMÉ

**Objectifs**: D'après le comité de spécialité en médecine d'urgence (MU) du Collège royal des médecins et chirurgiens du Canada, il est possible d'évaluer les activités professionnelles confiables (APC) de réanimation, tant en milieu de travail qu'en contexte de simulation. Toutefois, il existe peu de données sur la validité de ce type d'évaluation, dans l'un ou l'autre des deux environnements mentionnés. L'étude visait donc à déterminer si les évaluations des APC s'amélioraient au fil du temps et s'il existait une relation entre les évaluations réalisées en milieu de travail et celles effectuées en contexte de simulation.

**Méthode**: Toutes les évaluations des résidents de première année (n = 9) en MU, relatives à l'APC1 des fondements (F1) de la formation ont été recueillies au cours de l'année

From the [*]Department of Emergency Medicine, University of Ottawa, Ottawa, ON; [†]Department of Innovation in Medical Education, University of Ottawa, Ottawa, ON; [‡]Department of Anesthesiology and Pain Medicine, University of Ottawa, Ottawa, ON; [§]Division of Physical Medicine & Rehabilitation, University of Ottawa, Ottawa, ON; and the [¶]Ottawa Hospital Research Institute, The Ottawa Hospital, Ottawa, ON.

**Correspondence to:** Dr. Nicholas Prudhomme, Department of Emergency Medicine, The Ottawa Hospital, Civic Campus, 1053 Carling Avenue, E-Main, Room EM-206, Ottawa, Ontario K1Y 4E9; Email: nprudhomme@toh.ca

 *CJEM* 2020;22(6):811–818 DOI 10.1017/cem.2020.388

universitaire 2018-2019. Cette APC porte principalement sur l'amorce des manœuvres de réanimation chez les patients gravement malades et sur l'assistance des résidents. Les évaluations de l'APC, effectuées en milieu de travail et en contexte de simulation ont été comparées à l'aide du coefficient de corrélation de concordance (CCC) de Lin. Afin de déterminer si les évaluations, dans les deux types de milieu, changeaient à mesure que les résidents progressaient dans leur formation, l'équipe a procédé à une analyse de la variance entre sujets, en considérant le milieu de la formation et les mois écoulés comme des variables indépendantes.

**Résultats**: Ont été recueillies 104 évaluations en milieu de travail et 36, en contexte de simulation. Aucune corrélation n'a été établie entre la moyenne des évaluations de l'APC effectuées dans les deux types de milieu (CCC [8] = –0,01; $p = 0,93$). Une amélioration significative des évaluations a été observée au fil du temps, dans les deux milieux ($F [2,16] = 18,8$; $p < 0,001$; $\eta^2 = 0,70$); elles sont passées de 2,9 ± 1,2, du 1er au 4e mois inclusivement, à 3,5 ± 0,2, du 9e au 12e mois inclusivement. Enfin, les évaluations réalisées en milieu de travail (3,4 ± 0,1) étaient toujours plus élevées que celles effectuées en contexte de simulation (2,9 ± 0,2) ($F [2,16] = 7,2$; $p = 0,028$; $\eta^2 = 0,47$).

**Conclusion**: Il n'existe pas de corrélation entre les évaluations de l'APC F1, réalisées en milieu de travail et celles effectuées en contexte de simulation. Aussi faudrait-il rechercher les raisons pour lesquelles les résultats obtenus dans cette étude divergent de ceux obtenus dans d'autres études, et constituer une base de données probantes sur la validité des évaluations des APC réalisées en milieu de travail et celles effectuées en contexte de simulation.

**Keywords:** Education, emergency medicine, simulation

## INTRODUCTION

Postgraduate medical education in Canada is restructuring to a competency-based model of education called Competency by Design, which emphasizes demonstration of competencies required for patient care.[1] A cornerstone of Competency by Design is the concept of entrustable professional activities (EPAs), which are tasks specific to a discipline and stage of training.[2] Assessment of EPAs requires supervisors to document performance on EPA assessment forms, which use a rating scale that incorporates entrustment anchors.[3] Each EPA is assessed using a different EPA assessment form designed for national use by each discipline's specialty committee.[4] Despite the widespread implementation of these forms across specialties and training programs, validity evidence for their use is lacking.

The Royal College of Physicians and Surgeons of Canada (RCPSC) Emergency Medicine (EM) Specialty Committee indicates that certain EPAs may be assessed in either simulated or workplace environments.[4,5] Simulation provides learners with structured educational experiences to promote deliberate practice and feedback in a safe learning environment without risk to patient safety.[6] It has been shown to be an effective instructional method in health care education,[7] and a growing body of literature supports the translational outcomes of simulation-based training. Two recent reviews demonstrated that simulation-based mastery learning can lead to improved patient care practices and outcomes.[8,9] More recently, simulation has increasingly been used for low- and high-stakes assessment of clinical competence across medical specialties,[10] including EM.[11] While there are well-established benefits to learning in the simulation setting, a systematic review by Cook et al. reported that the validity evidence for simulation-based assessment is sparse and concentrated within specific specialties and assessment tools.[12] Additionally, there are few studies directly correlating simulation-based assessments with performance in authentic, workplace environments.[11] A recent multicenter study demonstrated a weak to moderate correlation between simulation-based assessments and in-training rotation evaluations.[13] Another study found a moderately positive correlation between simulation and workplace assessments of resuscitation skills using a locally derived assessment tool with limited validity evidence.[14]

There is limited validity evidence for the use of EM EPA assessment forms in the simulated and workplace environments, and it remains unclear whether EPA ratings in simulation reflect real-world performance. As Competency by Design curricula increasingly incorporate elements of simulation-based assessment, it is important to begin collecting evidence for the validity of EPA ratings in both settings. Applying modern validity theory using Kane's framework,[15] this study sought evidence to support an extrapolation inference (ratings in the "test world" reflect real-world performance) by examining whether (a) EPA ratings in the simulated and workplace settings correlate and (b) EPA ratings improve with progression through training.



https://doi.org/10.1017/cem.2020.388 Published online by Cambridge University Press

## METHODS

### Study design and setting

We conducted a prospective observational study to compare ratings of resident resuscitation performance in both the workplace and simulated environments. This study was conducted at The Ottawa Hospital Department of Emergency Medicine. This study was deemed exempt from ethics review by the Ottawa Health Science Network Research Ethics Board.

### Population

All first-year residents (n = 9) enrolled in the RCPSC-EM program at the University of Ottawa during the 2018–2019 academic year were included.

### Clinical workplace assessments

The EM Foundations of Training EPA F1 focuses on the early stages of resuscitation, including the initial management of patients experiencing shock, dysrhythmias, respiratory distress, altered mental status, and cardiopulmonary arrest.[5] When residents complete an assessment of a critically ill patient under direct observation by their supervisor, they are eligible and encouraged to have an assessment of EPA F1 completed. This assessment can be initiated by either the resident or the supervising physician in the resident's electronic portfolio, and details of the case, including patient demographics, case complexity, and clinical presentation, are documented (see the online Supplemental Appendix A). The supervisor assigns a global rating of the observed performance using the 5-point rating scale adopted by the Royal College to rate EPA performance.[16,17] The supervisor also provides and documents targeted feedback to the resident guided by the EPA milestones (the component skills required to perform the EPA). Milestone ratings are not required.

### Simulated environment assessments

In the first year of training, the study cohort was scheduled for six high-fidelity simulation sessions. At each session, simulation cases were run in parallel rooms. Three residents per room each led a unique simulated scenario designed to optimize the conceptual, physical, and experiential realism of the case.[18] Cases included resuscitation of simulated patients presenting with shock, dysrhythmia, respiratory distress, traumatic injury, altered level of consciousness, and cardiopulmonary arrest (Supplemental Appendix B). The team leader for each case was observed and their performance rated at the end of the scenario by two independent assessors (one staff simulation educator, one simulation fellow) in the same manner as in the workplace setting using the EPA F1 form. All ratings were documented before the case debriefing.

The RCPSC-EM Specialty Committee anticipates that residents will progress through the Foundations stage of training during their first year of residency. Therefore, EPA F1 ratings assigned in both the workplace and simulated environments during the 2018–2019 academic year were anonymized and exported into a spreadsheet for analysis.

### Data analysis

Data analysis was conducted using SPSS Statistics version 26. Descriptive statistics were calculated including means, standard deviations, and number of assessments per resident.

Reliability of EPA ratings was examined in several ways. First, an intraclass coefficient (ICC) was calculated to examine interrater reliability between EPA simulation ratings across the two raters. Second, generalizability theory (G-theory)[19] was used to estimate the overall reliability of EPA ratings obtained in both workplace and simulation environments. G-theory and the interpretation of G-coefficients is described in Supplemental Appendix C. Given the low-stakes, formative nature of EPAs, a dependability analysis was conducted to determine the number of assessments per resident needed to obtain a reliability of 0.6.

To examine the relationship between EPA ratings obtained in the workplace and those obtained in simulation environments, we used Lin's concordance correlation coefficient (CCC).[20] A detailed description of this analysis is provided in Appendix C.

To determine whether mean EPA ratings from the simulated and workplace environments differed as residents progressed through their training, data were collapsed into three 4-month blocks: months 1–4, 5–8, and 9–12. A within-subjects analysis of variance (ANOVA) was conducted using the mean ratings as the dependent variable and environment (simulation, workplace) and training month (months 1–4, months 5–8,



https://doi.org/10.1017/cem.2020.388 Published online by Cambridge University Press

**Table 1. EPA ratings in each learning environment**

| | Workplace environment | | | | Simulation environment | | | |
|---|---|---|---|---|---|---|---|---|
| Resident ID | Mean (SD) rating | Range | Median (IQR) rating | No. of assessments | Mean (SD) rating | Range | Median (IQR) rating | No. of assessments |
| 1 | 3.9 (1.2) | 2-5 | 4.0 (2.0) | 7 | 3.4 (0.5) | 3-4 | 3.5 (1.0) | 4 |
| 2 | 3.3 (0.8) | 2-4 | 3.0 (1.0) | 7 | 2.2 (0.4) | 2-3 | 3 (2.0) | 5 |
| 3 | 3.5 (1.1) | 1-5 | 4.0 (1.0) | 19 | 3.9 (0.9) | 3-5 | 4 (1.0) | 4 |
| 4 | 4.1 (0.4) | 4-5 | 4.0 (0.0) | 8 | 2.5 (0.7) | 2-3 | 4 (1.0) | 2 |
| 5 | 3.0 (0.5) | 2-4 | 3.0 (0.0) | 9 | 2.9 (0.3) | 2.5-3 | 3 (0.0) | 4 |
| 6 | 3.4 (0.9) | 2-5 | 3.0 (1.0) | 21 | 3.3 (0.5) | 3-4 | 3 (1.0) | 4 |
| 7 | 3.6 (0.9) | 2-5 | 4.0 (1.0) | 16 | 2.6 (0.5) | 2-3 | 3 (1.0) | 4 |
| 8 | 3.5 (1.3) | 2-5 | 3.0 (2.0) | 4 | 2.9 (0.9) | 2-4 | 3 (2.0) | 4 |
| 9 | 3.6 (0.5) | 3-4 | 4.0 (1.0) | 13 | 2.7 (0.7) | 1.5-3 | 3.5 (1.0) | 5 |
| Total | 3.5 (0.9) | 1-5 | 4.0 (1.0) | 104 | 2.9 | 1.5-4 | 3 (0.75) | 36 |

months 9–12) as independent variables. An explanation of this factorial design is provided in Supplemental Appendix C. Bonferroni corrections were applied to all multiple pairwise comparisons. Effect sizes were calculated using partial eta-squared ($\eta^2$) for ANOVAs and Cohen d for *t* tests. The magnitude of these effect sizes was interpreted using classifications proposed by Cohen.[21,22]

## RESULTS

Table 1 reports the mean EPA ratings and number of assessments for each resident in both training environments. A mean of 12 workplace and 4 simulation assessments were collected per resident. Levene's Test for Equality of Variance demonstrated no significant difference in variability of mean EPA ratings across workplace and simulation settings (F = 1.802; $p = 0.20$). The interrater reliability of simulation assessments was high (ICC = 0.863). Generalizability (G) coefficients for workplace EPA ratings and simulation EPA ratings were 0.35 and 0.75, respectively. Thirty-three workplace EPA assessments and three in the simulated environment would be required to achieve a reliability of 0.6.

There was no evidence of a relationship between EPA ratings in the simulation and workplace learning environments (CCC(8) = -0.01; 95% CI, -0.31–0.29; $p = 0.93$). The mean EPA ratings as a function of time and learning environment are shown in Figure 1. There was a main effect of month of training ($F(1,8) = 18.79$; $p < 0.001$; $\eta_p^2 = 0.70$). Subsequent comparisons revealed that mean ratings for months 1–4 (mean(SD) = 2.9 (1.2)) were significantly lower than for months 5–8 (3.1 (0.1), t(8) = 2.9; $p = 0.06$; d = 0.6) and months 9–12 (3.5 (0.2); t(8) = 5.3; $p = 0.002$; d = 1.3); similarly, mean ratings for months 5–8 (3.1(0.1)) were significantly lower than for months 9–12 (3.5(0.2); t(8) = 3.0; $p = 0.018$; d = 0.8). A main effect of environment was also identified ($F(1,8) = 7.16$; $p = 0.028$; $\eta_p^2 = 0.47$), indicating that mean workplace EPA ratings were consistently higher than mean simulation EPA ratings (3.4(0.1) v. 2.9(0.2), respectively). There was no interaction between time and environment ($p = 0.80$), indicating that the observed difference between workplace and simulation ratings remained constant over time.

## DISCUSSION

We compared resident performance on the initial stages of resuscitation as assessed with EPA F1 in the workplace and simulation environments. Ratings in each environment improved over time and were consistently lower in the simulation setting. There was high interobserver reliability among simulation educators. No correlation was observed between ratings of performance in both environments.

### *Improvement in ratings over time*

Performance ratings in both settings improved over time. This is expected as residents gain knowledge and expertise throughout their training. Applying modern validity theory, this observation supports an extrapolation inference for the validity of the assessment in either



https://doi.org/10.1017/cem.2020.388 Published online by Cambridge University Press

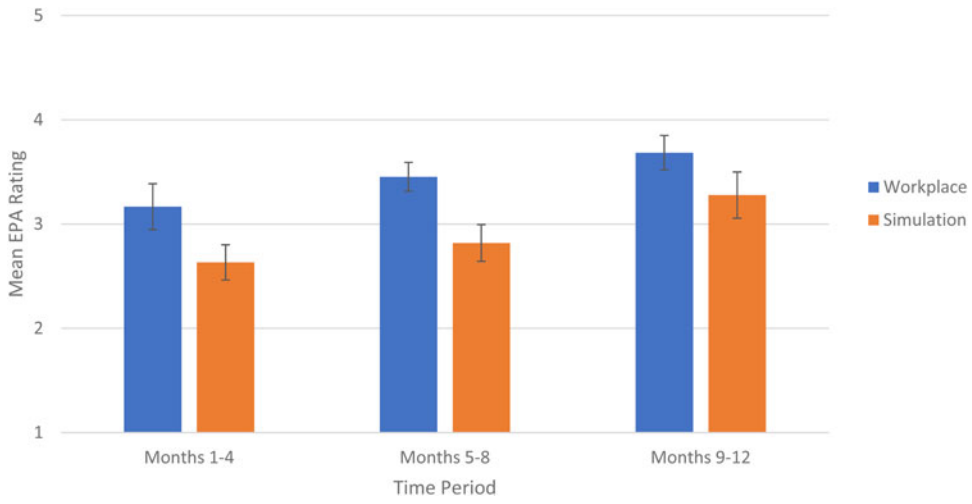


**Figure 1.** Improvement in mean EPA ratings in the workplace v. the simulation environments.

setting.[15] Weersink et al. observed a similar improvement in ratings based on resident training year, with more experienced residents scoring higher on assessments in the simulation setting.[14] Cheung et al. also observed a significant main effect of training level when residents were assessed in the workplace environment.[23] Our findings suggest that several months of training can yield sufficient data to observe improvements in resuscitation skills and potentially map the trajectory of performance for a given resident cohort. This may facilitate early identification of residents who are falling off the curve and subsequent implementation of modified learning plans.

### *Rating reliability*

The moderate to high reliability of EPA ratings in the simulation setting and high ICC observed among our simulation assessors further supports the validity of assessing resuscitation EPAs in the simulation lab. In contrast, an American study demonstrated poor interrater reliability of milestone ratings among faculty assessors who observed EM resident resuscitation performance in the simulated setting.[24] This difference may be partly related to the rating scales used. The Accreditation Council of Graduate Medical Education EM milestones are rated using scales that incorporate descriptive performance anchors unique to each milestone. However, there is a paucity of validity evidence for these scales.[25] EPAs in our study were rated using the O-SCORE scale, which incorporates entrustment anchors that reflect increasing levels of independence.[16] Several studies have demonstrated multiple sources of validity evidence for the use of this scale in different workplace and simulated contexts including the ED.[16,23,26–29]

### *Ratings in the workplace v. simulation environment*

Based on a recent study comparing ratings in the simulated and workplace environment,[14] we expected that there would be a positive correlation in ratings. However, our study observed no correlation between EPA F1 ratings in the two settings. One explanation may be the innate variability in workplace exposure experienced by each resident. Cases that residents typically experience in the simulated environment are critically ill patients in extremis or scenarios that are rare or infrequently experienced in the workplace setting.[30,31] On the other hand, we observed the use of EPA F1 in the workplace setting across a highly variable breadth of acuity, including minor trauma patients, hypotensive patients responsive to fluid administration, as well critically ill patients with multisystem injuries or cardiac arrest. Therefore, the acuity and complexity of cases reflected in each resident's simulated versus workplace EPA F1 assessments may have been variable, making it challenging to determine a correlation in performance between the two settings. If the assessments in each setting truly reflected different case types, the observed differences in mean EPA ratings and the lack of correlation between the two environments in our study may actually represent a form of discriminant validity evidence for EPA assessments.



https://doi.org/10.1017/cem.2020.388 Published online by Cambridge University Press

Similarly, the observed lack of correlation in ratings may have been related to a resident's tendency to select particular cases to be assessed. Ratings in the simulation lab were significantly lower than those in the workplace. Our competency committee observed that most workplace EPA assessments were triggered by residents as opposed to their supervisors. It is possible that residents may preferentially request their supervisors to document assessments in which they performed well, thus systematically biasing workplace case selection and workplace EPA ratings. This form of "gaming the system" has been previously described in the literature, and faculty development resources have been designed to help programs mitigate this assessment bias.[32,33]

### Implications for progression through training

By the end of the Foundations stage of training (months 9–12), residents were receiving mean ratings of 3.7 and 3.3 in the workplace and simulated environments, respectively. These ratings suggest that residents were not yet able to perform EPA F1 independently without supervision, a prerequisite for promotion to the next stage of training. Nevertheless, all residents were promoted. Our competence committee observed a discrepancy between EPA ratings and their associated narrative comments. The latter consistently reflected residents' ability to perform the EPA without supervision, but were associated with ratings that suggested they could not perform the task independently. Based on the narrative comments documented, we suspect the discrepant ratings were due to supervisors rating residents based on their performance of the *entire* resuscitation, requiring more complex skills and abilities, rather than assessing the specific EPA task of *initiating* and *assisting* in the resuscitation. A correlation between ratings in the simulated and workplace settings may not have been observed because faculty were misinterpreting the EPA task. Misinterpretation of an EPA task by supervisors is a potential threat to the validity of these types of assessments and highlights an important ongoing faculty development need within Competency by Design.

### Limitations

We observed variation in numbers of workplace assessments between residents. This likely reflects the resident-driven nature of EPA assessments. We attempted to account for this variability by conducting a within-subject analysis and applying G-theory to determine reliability. In a controlled study, all residents would have ideally had similar numbers of workplace and simulated assessments uniformly distributed over the study period. However, our pragmatic, observational design took advantage of real-world implementation of EPAs in a residency program. The observed variation in number of assessments is not unique to our institution and represents a major challenge associated with the implementation of Competency by Design.[34] Numbers of workplace EPAs also varied over each study block, while those in simulation remained constant. However, our within-subject analysis of variance was able to account for this difference. Furthermore, this was a single center study conducted over a short time frame, and results may have been influenced by local cultural norms and assessment patterns, thus limiting the transferability of our findings. Last, all workplace and simulation assessors were unblinded to each participant. Prior experience with each learner carries the risk biasing current and future assessments,[35] and our methodology did not allow for blinded external assessment of performance.

## CONCLUSION

There was no correlation between ratings of resident skills in the initial resuscitation of critically ill patients in the workplace and simulated environments as assessed by EPA F1. Ratings improved over time and higher ratings were observed in the workplace settings. Factors such as variable case complexity, case selection, and misinterpretation of the assessment task make it challenging to compare ratings of performance in the two environments. Given the conflicting results of this study with others, it remains unclear whether resuscitation performance in a simulated setting reflects performance in the clinical workplace. As greater emphasis is being placed on simulation as a modality for assessing clinical competence, future studies are needed to clarify these differences and establish an evidence base for the validity of EPA assessments in both environments.

**Supplemental material:** The supplemental material for this article can be found at https://doi.org/10.1017/cem.2020.388.

**Competing interests:** There are no conflicts of interest to report.

**Author Contributions:** N.P., M.O., and W.J.C. conceived the idea. W.J.C. supervised the conduct of the trial and data collection. N.P. collected and managed the data. M.M. conducted statistical analysis of the data. N.P. drafted the manuscript. N.P.,

M.O., M.M., N.D., and W.J.C. reviewed the manuscript and contributed substantially to its revision. N.P. takes responsibility for the study as a whole.

**Financial support:** This study was funded by a TOH Department of Emergency Medicine Academic Grant.

## REFERENCES


1. Royal College of Physicians and Surgeons of Canada. *CBD Speaking Points*. Ottawa: Royal College of Physicians and Surgeons; 2015.
2. Royal College of Physicians and Surgeons of Canada. EPAs and milestones. 2019. [cited May 20, 2019]. Available at: http://www.royalcollege.ca/rcsite/cbd/implementation/cbd-milestones-epas-e (accessed May 15, 2020).
3. Rekman J, Gofton W, Dudek N, Gofton T, Hamstra SJ. Entrustability scales: outlining their usefulness for competency-based clinical assessment. *Acad Med* 2016;91(2):186-90.
4. Sherbino J, Bandiera G, Doyle K, et al. The competency-based medical education evolution of Canadian emergency medicine specialist training. *CJEM* 2020;22(1):95–102.
5. Emergency Medicine Specialty Committee. *EPA Guide: Emergency Medicine*. Ottawa: Emergency Medicine Specialty Committee; 2017.
6. Ziv A, Wolpe PR, Small SD, Glick S. Simulation-based medical education: an ethical imperative. *Simul Healthc* 2006;1(4):252-6.
7. Cook DA, Brydges R, Hamstra SJ, et al. Comparative effectiveness of technology-enhanced simulation versus other instructional methods: a systematic review and meta-analysis. *Simul Healthc* 2012;7(5):308-20.
8. McGaghie WC, Issenberg SB, Barsuk JH, Wayne DB. A critical review of simulation-based mastery learning with translational outcomes. *Med Educ* 2014;48(4):375-85.
9. Brydges R, Hatala R, Zendejas B, Erwin PJ, Cook DA. Linking simulation-based educational assessments and patient-related outcomes: a systematic review and meta-analysis. *Acad Med* 2015;90:246-56.
10. Chiu M, Tarshis J, Antoniou A, et al. Simulation-based assessment of anesthesiology residents’ competence: development and implementation of the Canadian National Anesthesiology Simulation Curriculum (CanNASC). *Can J Anaesth* 2016;63(12):1357-63.
11. Hall AK, Chaplin T, McColl T, et al. Harnessing the power of simulation for assessment: Consensus recommendations for the use of simulation-based assessment in emergency medicine. *CJEM* 2020;22(2):194–203.
12. Cook DA, Brydges R, Zendejas B, Hamstra SJ, Hatala R. Technology-enhanced simulation to assess health professionals: a systematic review of validity evidence, research methods, and reporting quality. *Acad Med* 2013;88(6):872-83.
13. Hall AK, Damon Dagnone J, Moore S, et al. Comparison of Simulation-based Resuscitation Performance Assessments With In-training Evaluation Reports in Emergency Medicine Residents: A Canadian Multicenter Study. *AEM Educ Train* 2017;1(4):293–300.
14. Weersink K, Hall AK, Rich J, Szulewski A, Dagnone JD. Simulation versus real-world performance: a direct comparison of emergency medicine resident resuscitation entrustment scoring. *Adv Simul* (*Lond*) 2019;4:9.
15. Kane M. Validation. *Educational Measurement* (ed. Brennan R). Westport, CT: Praeger; 2006, 17–64.
16. Gofton WT, Dudek NL, Wood TJ, Balaa F, Hamstra SJ. The Ottawa Surgical Competency Operating Room Evaluation (O-SCORE): a tool to assess surgical competence. *Acad Med* 2012;87(10):1401-7.
17. Gofton W, Dudek N, Barton G, Bhanji F. Workplace-based assessment implementation guide: formative tips for medical teaching practice. 1st ed. Ottawa; 2017. Available at: http://www.royalcollege.ca/rcsite/documents/cbd/wba-implementation-guide-tips-medical-teaching-practice-e.pdf
18. Rudolph JW, Simon R, Raemer DB. Which reality matters? Questions on the path to high engagement in healthcare simulation. *Simul Healthc* 2007;2(3):161-3.
19. Brennan R. Generalizability theory. *Educ Meas Issues Pract* 1992;11(4):27–34.
20. Lin L. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 1989;45:255-68.
21. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. Revised Edition. Hillsdale, NJ: Laurence Earlbaum Associates Inc; 1988.
22. Cohen J. Eta-squared and partial Eta-Squared in fixed factor Anova designs. *Educ Psychol Meas* 1973;33(1):107-12.
23. Cheung WJ, Wood TJ, Gofton W, Dewhirst S, Dudek N. The Ottawa Emergency Department Shift Observation Tool (O-EDShOT): A New Tool for Assessing Resident Competence in the Emergency Department. Burkhardt JC, editor. AEM Educ Train 2019 December 19, 2019. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1002/aet2.10419 (accessed May 15, 2020).
24. Wittels K, Abboud ME, Chang Y, Sheng A, Takayesu JK. Inter-rater relaiability of select emergency medicine milestones in simulation. *J Emerg Intern Med* 2017. Available at: http://www.imedpub.com/emergency-and-internal-medicine/ (accessed May 15, 2020).
25. Schott M, Kedia R, Promes SB, et al. Direct observation assessment of milestones: problems with reliability. *West J Emerg Med* 2015;16(6):871-6.
26. MacEwan MJ, Dudek NL, Wood TJ, Gofton WT. Continued validation of the O-SCORE (Ottawa Surgical Competency Operating Room Evaluation): use in the simulated environment. *Teach Learn Med* 2016;28(1):72-9.
27. Voduc N, Dudek N, Parker CM, Sharma KB, Wood TJ. Development and validation of a bronchoscopy competence assessment tool in a clinical setting. *Ann Am Thorac Soc* 2016;13(4):495–501.
28. Rekman J, Hamstra SJ, Dudek N, Wood T, Seabrook C, Gofton W. A new instrument for assessing resident competence in surgical clinic: The Ottawa Clinic Assessment Tool. *J Surg Educ* 2016;73(4):575-82.
29. Halman S, Rekman J, Wood T, Baird A, Gofton W, Dudek N. Avoid reinventing the wheel: implementation of the Ottawa Clinic Assessment Tool (OCAT) in internal medicine. *BMC Med Educ* 2018;18(1):218.




https://doi.org/10.1017/cem.2020.388 Published online by Cambridge University Press

30. Motola I, Devine LA, Chung HS, Sullivan JE, Issenberg SB. Simulation in healthcare education: a best evidence practical guide. AMEE Guide No. 82. *Med Teach* 2013;35(10):e1511-30.
31. Petrosoniak A, Ryzynski A, Lebovic G, Woolfrey K. Cricothyroidotomy in situ simulation curriculum (CRIC Study). *Simul Healthc* 2017;12(2):76–82.
32. Pinsk M, Karpinski J, Carlisle E. Introduction of competence by design to Canadian nephrology postgraduate training. *Can J Kidney Health Dis* 2018;5 :2054358118786972.
33. Oswald A, Cheung WJ, Bhanji F, Ladhani MB, Hamilton J. Mock competence committee cases for practice deliberation. Royal College of Physicians and Surgeons of Canada. Available at: http://www.royalcollege.ca/mssites/casescenarios_en/story_html5.html (accessed May 15, 2020).
34. Chan T, Paterson Q, Hall A, et al. Outcomes in the age of competency-based medical education: recommendations for EM training in Canada from the 2019 symposium of academic emergency physicians. *Can J Emerg Med* 2020;22 (2):204-11.
35. Gauthier G, St-Onge C, Tavares W. Rater cognition: review and integration of research findings. *Med Educ* Blackwell Publishing Ltd; 2016;50:511-22.



https://doi.org/10.1017/cem.2020.388 Published online by Cambridge University Press