A CanMEDS Competency-Based Assessment Tool for High-Fidelity Simulation in Internal Medicine: The Montreal Internal Medicine Evaluation Scale (MIMES)

Patrice Chrétien Raymer MD, BSc, Jean-Paul Makhzoum MD, FRCPC, Robert Gagnon MPsy, Arielle Lévy MD, MMed, Jean-Pascal Costa, MD, MMEd, FRCPC

About the Authors:
Dr. Patrice Chrétien Raymer is an Internal Medicine Resident at the Université de Montréal, Montreal, Quebec
Dr. Jean-Paul Makhzoum, is an Assistant Professor, Hôpital Sacré-Coeur, Université de Montréal, Montreal, Quebec Robert Gagnon, is a Psychometrician, Assessment Office, Faculty of Medicine, Université de Montréal, Montreal, Quebec Dre. Arielle Lévy, is with the Department of Paediatrics, University of Montreal, Montreal, Canada
Dr. Jean-Pascal Costa is Assistant Professor, Centre Hospitalier de l'Université de Montréal, Montreal, Quebec Corresponding Author: Patrice.chretien.raymer@umontreal.ca
Submitted: February 22, 2018. Accepted: September 27, 2018. Published: November 9, 2018. DOI: 10.22374/cjgim.v13i4.280

ABSTRACT

High-fidelity simulation is an efficient and holistic teaching method. However, assessing simulation performances remains a challenge. We aimed to develop a CanMEDS competency-based global rating scale for internal medicine trainees during simulated acute care scenarios.

Methods

Our scale was developed using a formal Delphi process. Validity was tested using 6 videotaped scenarios of 2 residents managing unstable atrial fibrillation, rated by 6 experts. Psychometric properties were determined using a G-study and a satisfaction questionnaire.

Results

Most evaluators favourably rated the usability of our scale, and attested that the tool fully covered CanMEDS competencies. The scale showed low to intermediate generalization validity.

Conclusions

This study demonstrated some validity arguments for our scale. The best assessed aspect of performance was communication; further studies are planned to gather further validity arguments for our scale and to compare assessment of teamwork and communication during scenarios with multiple versus single residents.

RESUME

La simulation haute fidélité est une méthode d'enseignement efficace et holistique. Cependant, l'évaluation des performances en simulation demeure un défi. Nous avons cherché à développer une échelle d'évaluation globale fondée sur les compétences CanMEDS pour les résidents en médecine interne dans le cadre de simulations en soins aigus.

Méthodes

Notre échelle a été développée en utilisant un processus Delphi. Sa validité a été testée a l'aide de 6 scenarios filmés sur vidéo de 2 résidents prenant en charge un cas de fibrillation auriculaire instable, puis évalués par 6 experts. Les propriétés psychométriques de l'échelle ont été déterminées à l'aide d'une étude G et d'un questionnaire de satisfaction.

Résultats

La plupart des évaluateurs ont jugé favorablement l'utilisation de notre échelle et ont confirmé que l'outil couvrait pleinement les compétences CanMEDS. L'échelle a démontré une validité de généralisation faible a intermédiaire.

Conclusions

Cette étude a démontré certains arguments de validité pour notre échelle. Le meilleur aspect évalué était la communication. D'autres études sont prévues pour fournir d'autres arguments de validité pour notre échelle, ainsi que pour comparer la capacité de notre échelle à évaluer le travail d'équipe et la communication lors de scenarios avec un ou plusieurs résidents.

High-Fidelity simulation is a teaching and assessment method which has been rapidly introduced into many medical residency programs.[1] High-fidelity simulation is considered by many as an innovation for direct clinical observation[2] and punctual end-of-rotation evaluations.[3] Simulation has proven to be an effective and enjoyable teaching method, allowing for immediate and targeted feedback.[4] Simulation training has been associated with better subsequent performances in technical skills[3,5–7] and non-technical skills, the latter being often deficient in problematic students and challenging to evaluate and teach.[8,9]

To address these issues, global rating scales (GRS)[10] were developed based on recognized crisis resource management behavioural markers[11] to overcome the limitations of checklists-based evaluation. Studies have also reported GRSs to be especially useful in preparing direct observation and providing appropriate and organized feedback.[12] To be completely useful, such rating scales must be adapted to the specific medical specialty and complementary to the competency framework of the trainee within that program.[13–15]

The CanMEDS framework describes competencies required by Canadian physicians to properly tend to their patients.[16] CanMEDS competencies have clearly been identified as overlapping substantially with NTS's (non-technical skills); however, up to now, only one GRS (called the GIOSAT) has been based on CanMEDS, and is not adapted to internal medicine.[17]

In this study, we aimed to develop a GRS based on CanMEDS to assess internal medicine residents participating in acute care high-fidelity simulation scenarios. We also aim to study this scale's validity, as well as its usability, based on Kane's validity framework.[18–20]

Methods

Development of the rating scale: We initially performed a literature search to look for other NTS GRSs. We selected all articles for which the principal subject was evaluation of trainees in a simulation setting. Our search revealed several GRSs of interest.[10,15,17,21,22]

The Anesthetists' Nontechnical Skills (ANTS) was the first GRS developed for anesthesiologists,[13], and is one of the most studied, including in Canada.[6,23] We built an anchored 5-level Likert scale based on the ANTS, other GRSs, and the CanMEDS competency framework.[16] We further refined the scale using a 2-step Delphi method for which 4 recognized simulation-based

education experts were recruited. They were provided with a video of a pulmonary edema scenario managed by a junior resident and used our scale to assess performance. Testers filled in a formal questionnaire at each Delphi step, and provided informal comments. Our scale was modified at each step incorporating the comments provided. The final version of our scale was in French, but its English version can be found in Figure 1.

Preliminary Evaluation Study Design: To evaluate the MIMES, we developed a retrospective observational study involving internal medicine residents in a simulated acute care setting. Our protocol was evaluated and improved by our institution's Committee for Simulation-Based Medical Education and approved by our institution's Multifaculty Committee for Research Ethics.

We used 6 anonymized videos of different teams of 2 PGY-1 residents from our internal medicine program, managing a case of unstable atrial fibrillation. The simulated patient, portrayed by a high-fidelity mannequin, spoke neither French nor English, and had to be communicated with by speaking to his wife. The wife was scripted to be very anxious of the patient's state and inquisitive of the team's management plan.

Ratings: Six experienced acute care physicians and simulation-trained instructors were recruited to evaluate student performances. No evaluator training or calibration was provided for this tool.[15,24] Each evaluator evaluated the 2 students in each video independently using our GRS, after which they completed an online questionnaire to evaluate the scale`s validity and usability.

Statistical Analysis: We evaluated our scale using a generalizability theory model[25] G-study. Facets for our G-study were students (S), evaluators (E) and NTS categories (C). Design was fully crossed (S/RC); student and evaluator facets were random, while categories was a fixed facet (n=6). We also

Name : _____ Date : _____

Resident level : _____ Evaluator: _____

Task Management

	1	2	3	4	5	N/A
Prioritizing interventions	Does not prioritize CAB Does not stabilize patient		Prioritizes CAB and stabilization request adequate vascular access		Excellent prioritization and revaluates CAB adequately	
Organization	Organization lacking Non-systematic approach		Adequate structure; history, physical exam, investigation, treatment		Efficient, fluid structure Targeted, complete evaluation	
Adaptability	Disorganized or does not consider new information		Finds solutions, modifies approach with new information		Adapts rapidly, prepares for multiple outcomes in advance	
Calls for help	Does not call for help		Calls for help, recognizes limits		Calls for help with adequate transfer of information to consultant	

Medical Expertise

	1	2	3	4	5	N/A
Differential Diagnosis (Ddx)	Poor differential, Did not include correct diagnosis in Ddx		Adequate differential and most probable diagnosis included; Recognizes uncertainty		Targeted and ordered Ddx, active elaboration of Ddx during evaluation	
Investigations and Interpretation	Broad, unnecessary workup Wrong interpretation		Requests appropriate investigations Adequate interpretation		Organizes investigations based on level of urgency and differential diagnosis	
Therapeutics	Treatment lacking, inadequate, unsafe, not based on level of emergency		Adequate stabilization, treatment organized as per level of urgency		Rapid, accurate and evidence based treatment. Emergencies rapidly addressed	

Figure 1. The Montreal Internal Medicine Evaluation Scale (MIMES)

Collaboration and leadership

Team management

1	2	3	4	5	N/A
Works alone, lacks respect for other members		Integrates team in management, respects others during interventions		Uses team member strengths, does not overload	

Information Sharing

1	2	3	4	5	N/A
Does not share information Does not consider team interventions		Informs team of diagnosis and management plan		Verbalizes diagnostic and therapeutic process. Considers team input in management	

Communication

Listening Capacity

1	2	3	4	5	N/A
Does not listen to patient or family		Receptive to patient and family preoccupations		Establishes therapeutic contact, empathic, manages conflicts	

Vulgarisation

1	2	3	4	5	N/A
Uses unnecessary complex language		Uses common terms		Adapts to patient language, validates comprehension	

Global Evaluation

1	2	3	4	5
Inferior, unable to manage similar situations		Adequate, could manage similar situations		Superior, could likely manage more complex situations

Figure 1. (*continued*)

performed a (D) study for the number of evaluators [25]. The online questionnaire completed by our evaluators was analyzed using descriptive statistics.

Results

G and D studies: The result of our G-study is presented in Table 1; reliability of the overall scale was 0.64 with evaluations pooled from 6 evaluators. It is generally considered that a G coefficient of 0.6 is the limit for acceptable reliability, and that more than 0.8 represents near-perfect reliability. In addition, assessment of communication yielded the highest reliability (0.71).

Table 2 present the variance component of our G-study; the highest contributor to variance was evaluator-student interaction (49%). Results of the D studies performed are shown in Table 3.

Table 1: Estimated G-coefficients for Overall MIMES NTS Categories

Table 1. Relative G-Coefficient	
Overall Assessment Scale	0.64
Task Management	0.56
Medical Expertise	0.57
Collaboration/Leadership	0.61
Communication	0.71
Global Student Evaluation	0.59

Online questionnaire: Results from the survey can be found in Table 4; 100% of evaluators concluded that there were no superfluous elements in the MIMES. All evaluators concluded that MIMES is appropriate in length and is simple to use. 83% felt that the scale was potentially useful to provide immediate feedback to students, and 66% felt that it could be used to plan structured teaching. 66% of them agreed that the MIMES was not adapted to be used as a summative assessment tool. 100% of surveyed internal medicine specialists agreed that the MIMES contained all necessary elements when assessing performance of internal medicine residents during an acute care situation.

Discussion

In this study, we set out to create and evaluate an educational tool for internal medicine residents performing acute care simulation scenarios. Results from our study showed generalization validity arguments for the MIMES and most evaluators reported that the tool assessed all appropriate components of the CANMEDS framework, a recognized and comprehensive set of skills necessary for medical practice, without assessing superfluous categories. The formal Delphi method used to create the MIMES and its elaboration process which was based on recognized GRSs and well-recognized competency frameworks also adds to its scoring validity argument.

Table 2. Estimated Variance Contribution and Variance Components

Table 2: G-Study Variance		
Component	Contribution	Estimate
Students (S)	14%	59,50
Evaluator (E)	4%	19,65
Category (C)	2%	9,55
Student-Evaluator interaction (S-E)	49%	108,56
Student-Category interaction (S-C)	5%	17,62
Evaluator-Category interaction (E-C)	4%	10,89
S-E-C interaction (error)	22%	39,62

Table 3 D-study Estimated G-coefficient Results with Projected Changes in Number of Evaluators

Table 3: D-Studies		
D-Study - Evaluator projection	Number of Evaluators	Estimated G-Coefficient
	1	0,22
	2	0,37
	3	0,47
	4	0,54
	5	0,59
Actual Study	***6***	***0,64***
	7	0,64
	8	0,64

Table 4. Descriptive Statistical Analysis of the Usability and Validity Survey Completed by Raters

Table 4: Rater Questionnaire (n=6)			
Q1	Missing Elements? Yes 33% **No 67%**	Q6	Useful to provide immediate feedback? **Absolutely or yes** **83%** No really 17 %
Q2	Were there elements that seemed superfluous? Yes 0% **No 100%**	Q7	Useful for summative assessment? Yes 17% **Unsure or not really** **66%**
Q3	Were there elements difficult to evaluate? **Yes 67%** No 33%		
Q4	Simple to use? **Absolutely or Yes** **100%** Unsure, not really, not at all 0%	Q8	Appropriate lenght? **Yes** **100%** Unsure, not really, not at all 0%
Q5	Useful to plan structured teaching? **Yes** **66%** Unsure or not really 17%	Q9	All elements present for objectives being evaluated? **Yes** **67%** No 0% Hard to say, not an internist 33%

Usability was well evaluated, with all evaluators finding the tool easy to use. Evaluators felt that such a tool could be used for teaching and formative purposes and to assist educators with providing immediate feedback following simulation, an important component of the pedagogic value of such exercises.[26,27]

Using our actual experimental setup, reliability of our tool was moderate to low, even when combining 6 observers. Overall and aspect-specific G coefficients were not high enough for the MIMES to be considered reliable enough for proper summative assessment in similar scenarios. This is in accordance with the perception of evaluators reported in our survey.

The evaluator (E) component of variance in our G-study was very low, which demonstrates a relative insulation of our GRS to individual observer systematic biases such as hawk and dove effect or contrast effects.[28] Category (C) variance was low, indicating that good performances in one category correlated with good performances in others; this could be interpreted as redundancy in our GRS. However, usability was not affected by this, and division into such categories permits more oriented feedback to be given using the MIMES.

Student-Evaluator (S-E) interaction had the highest source of variance, meaning that the performance of a specific task performed by the same student was evaluated differently by the various evaluators. Modifications to our protocol will be necessary to gather additional validity evidence for the MIMES GRS in further studies. One important step not included for the sake of usability was evaluator training. Evaluator training can vary from as little as 2 hours to as long as 2 days,[13] and can greatly contribute to the reliability and scoring validity argument of an assessment tool.[20,24,28,29] Such processes will be implemented in upcoming validity studies of the MIMES, with the objective of attaining reliability scores close to ones published in similar GRS studies.[17,21,30]

The NTS that showed the highest reliability in our study was communication, while the second highest was collaboration/leadership. This may have been due to the scenario chosen where 2 students participated with the same role, instead of a single participant or multiple participants with distinct predetermined roles. The addition of communication challenges (anxious family member and language barrier with the simulated patient) also helped reach this goal. Using scenarios where teams of participants must manage communication and psychosocial challenges in addition to medical emergencies might help in adequately evaluating individual participant's communication and collaboration skills. This is particularly interesting considering that, most often, it is difficult to evaluate these specific NTSs even in OSCE settings.[31]

There is an increasing need to develop quality competency-based assessment tools in an era of evolving competency-based curriculum, such as those being implemented in Canadian medical education [32, 33]. We believe the MIMES to be one of those potentially useful tools for assessing clinical skills and NTS's required when managing acute care internal medicine cases.

Disclosure

None of the authors report any academic of financial conflicts of interest, relating to the results, interpretation, conclusions or expected impact of this article.

References

1. Boulet JR. Summative assessment in medicine: the promise of simulation for high-stakes evaluation. Acad Emerg Med 2008;15(11):1017–24.
2. Regehr G, et al. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. Acad Med 1998;73(9):993–7.
3. Doyle JD, Webber EM, Sidhu RS. A universal global rating scale for the evaluation of technical skills in the operating room. Am J Surg 2007;193(5):551–5.
4. Kneebone R, et al. An innovative model for teaching and learning clinical procedures. Med Educ 2002;36(7):628–34.
5. Bruppacher HR, et al. Simulation-based training improves physicians' performance in patient care in high-stakes clinical setting of cardiac surgery. Anesthesiology 2010;112(4):985–92.
6. Yee B, et al. Nontechnical skills in anesthesia crisis management with repeated exposure to simulation-based education. Anesthesiology 2005;103(2):241–8.
7. McGaghie WC, et al. Does Simulation-based medical education with deliberate practice yield better results than traditional clinical education? A meta-analytic comparative review of the evidence. Acad Med 2011;86(6):706–11.
8. Ker JS, et al. Can a ward simulation exercise achieve the realism that reflects the complexity of everyday practice junior doctors encounter? Med Teach 2006;28(4):330–4.
9. Khan K, Pattison T, Sherwood M. Simulation in medical education. Med Teach 2011;33(1):1–3.
10. Weller JM, et al. Evaluation of high fidelity patient simulator in assessment of performance of anaesthetists. Br J Anaesth 2003;90(1):43–7.
11. Klampfer B, et al. Enhancing Performance in High Risk Environments: Recommendations for the use of Behavioural Markers. In: Gottlieb Deimler and Kerl Benz foundation Kolleg group interaction in High Risk environnement (GIHRE). Behavioural Markers Workshop. Zurich: Swissair Training Center; 2001.
12. Patey R, et al. Developing a Taxonomy of Anesthetists' Nontechnical Skills (ANTS). In: Henriksen K, et al, eds. Advances in Patient Safety: From Research to Implementation, Volume 4: Programs, Tools, and Products. Rockville: Agency for Healthcare Research and Quality; 2005.
13. Flin R, et al. Anaesthetists' non-technical skills. Br J Anaesth 2010;105(1):38–44.
14. Yule S, et al. Development of a rating system for surgeons' non-technical skills. Med Educ 2006;40(11):1098–104.
15. Fletcher G, et al. Anaesthetists' Non-Technical Skills (ANTS): evaluation of a behavioural marker system. Br J Anaesth 2003;90(5):580–8.
16. Frank JR, Snell L, Sherbino J, editors. Can Meds 2015 Physician Competency Framework. Ottawa: Royal College of Physicians and Surgeons of Canada; 2015.
17. Neira VM, et al. "GIOSAT": a tool to assess CanMEDS competencies during simulated crises. Can J Anaesth 2013;60(3):280–9.
18. Hatala R, et al. Constructing a validity argument for the Objective Structured Assessment of Technical Skills (OSATS): a systematic review of validity evidence. Adv Health Sci Educ Theory Pract 2015;20(5):1149–75.

19. Cook DA, et al. A contemporary approach to validity arguments: a practical guide to Kane's framework. Med Educ 2015;49(6):560–75.
20. Tavares W, et al. Applying Kane's validity framework to a simulation based assessment of clinical competence. Adv Health Sci Educ Theory Pract 2017; [Epub ahead of print]
21. Hall AK, et al. Queen's simulation assessment tool: development and validation of an assessment tool for resuscitation objective structured clinical examination stations in emergency medicine. Simul Healthc 2015;10(2):98–105.
22. Hodges B, McIlroy JH. Analytic global OSCE ratings are sensitive to level of training. Med Educ 2003;37(11):1012–6.
23. Morgan PJ, et al. Nontechnical skills assessment after simulation-based continuing medical education. Simul Healthc 2011;6(5):255–9.
24. Preusche I, Schmidts M, Wagner-Menghin M. Twelve tips for designing and implementing a structured rater training in OSCEs. Med Teach 2012;34(5):368–72.
25. Cardinet J, Johnson S, and Pini G. Applying generalizability theory using EduG. New York:Taylor & Francis; 2010.
26. Welke TM, et al. Personalized oral debriefing versus standardized multimedia instruction after patient crisis simulation. Anesth Analg 2009;109(1):183–9.
27. Savoldelli GL, et al. Value of debriefing during simulated crisis management: oral versus video-assisted oral feedback. Anesthesiology 2006;105(2):279–85.
28. Feldman M, et al. Rater training to support high-stakes simulation-based assessments. J Contin Educ Health Prof 2012;32(4):279–86.
29. Boulet JR, Murray DJ. Simulation-based assessment in anesthesiology: requirements for practical implementation. Anesthesiology 2010;112(4):1041–52.
30. Kim J, et al. A pilot study using high-fidelity simulation to formally evaluate performance in the resuscitation of critically ill patients: The University of Ottawa Critical Care Medicine, High-Fidelity Simulation, and Crisis Resource Management I Study. Crit Care Med 2006;34(8):2167–74.
31. Setyonugroho W, Kennedy KM, Kropmans TJ. Reliability and validity of OSCE checklists used to assess the communication skills of undergraduate medical students: A systematic review. Patient Educ Couns 2015;98(12):1482–91.
32. Hamstra SJ. Keynote address: the focus on competencies and individual learner assessment as emerging themes in medical education research. Acad Emerg Med 2012;19(12):1336–43.
33. Harris K, Frank J, eds. Competence by design: Reshaping Canadian medical education. Ottawa:Royal College of Physicians and Surgeons of Canada; 2014.