The Development and Validation of a Concise Instrument for Formative Assessment of Team Leader Performance During Simulated Pediatric Resuscitations

Lindsay D. Nadkarni, MD;
Cindy G. Roskind, MD;
Marc A. Auerbach, MD, MSc;
Aaron W. Calhoun, MD;
Mark D. Adler, MD;
David O. Kessler, MD, MSc

Aim: The aim of this study was to assess the validity of a formative feedback instrument for leaders of simulated resuscitations.

Methods: This is a prospective validation study with a fully crossed (person × scenario × rater) study design. The Concise Assessment of Leader Management (CALM) instrument was designed by pediatric emergency medicine and graduate medical education experts to be used off the shelf to evaluate and provide formative feedback to resuscitation leaders. Four experts reviewed 16 videos of in situ simulated pediatric resuscitations and scored resuscitation leader performance using the CALM instrument. The videos consisted of 4 pediatric emergency department resuscitation teams each performing in 4 pediatric resuscitation scenarios (cardiac arrest, respiratory arrest, seizure, and sepsis). We report on content and internal structure (reliability) validity of the CALM instrument.

Results: Content validity was supported by the instrument development process that involved professional experience, expert consensus, focused literature review, and pilot testing. Internal structure validity (reliability) was supported by the generalizability analysis. The main component that contributed to score variability was the person (33%), meaning that individual leaders performed differently. The rater component had almost zero (0%) contribution to variance, which implies that raters were in agreement and argues for high interrater reliability.

Conclusions: These results provide initial evidence to support the validity of the CALM instrument as a reliable assessment instrument that can facilitate formative feedback to leaders of pediatric simulated resuscitations.

(*Sim Healthcare* 13:77–82, 2018)

Key Words: Simulation, Resuscitation, Team leader.

Pediatric resuscitations are infrequent but high-stakes events, providing scarce opportunities for trainees to achieve proficiency in leading these scenarios.[1–6] Teamwork is critical to success in resuscitations, and effective leadership is integral to both improved team performance and high-quality patient care.[7–13] The current resuscitation guidelines support leadership training as a part of advanced life support training.[7]

Simulation is increasingly used as a tool to increase trainee resuscitation experience, skills, and teamwork.[14–17] Prompt feedback is a vital component of simulation-based medical education, often guided by standardized assessment instruments.[16–19] However, standardized assessments of resuscitation leader performance are lacking. Concise, "off the shelf" instruments that are easy to use in real time can allow supervisors to assess and give immediate formative feedback to learners after resuscitation-leading experiences. Many existing instruments do not focus on individual team leader performance but rather the performance of the entire team.[20–25] Other instruments that have been created evaluate individual performance of pediatric resuscitation team leaders in the research setting but may be cumbersome or require extensive training to use them, thus limiting their practical use in the clinical or educational environment.[23–29]

We developed the Concise Assessment of Leader Management (CALM) instrument as a pragmatic instrument to help educators provide formative feedback to resuscitation leaders after simulated pediatric resuscitations. The CALM instrument was designed to require minimal user training and be used to efficiently collect real-time assessment data to inform immediate formative feedback to learners. For this validation study, we aim to demonstrate initial evidence to support content and internal structure (reliability) validity for the CALM instrument.

METHODS

Study Design

In this prospective validation study, experts were recruited from the International Network for Simulation-based

From the Sidney Kimmel Medical College at Thomas Jefferson University (L.D.N.), Philadelphia, PA; Department of Pediatrics (C.G.R., D.O.K.), Division of Pediatric Emergency Medicine, Morgan Stanley Children's Hospital of NY Presbyterian, Columbia University Medical Center, New York, NY; Department of Pediatrics (M.A.A.), Division of Pediatric Emergency Medicine, Yale University School of Medicine, New Haven, CT; Department of Pediatrics (A.W.C.), Division of Pediatric Critical Care, University of Louisville School of Medicine, Louisville, KY; and Department of Pediatrics (M.D.A.), Division of Pediatric Emergency Medicine, Northwestern University Feinberg School of Medicine, Ann & Robert H. Lurie Children's Hospital of Chicago, Chicago, IL.

Reprints: Lindsay D. Nadkarni, MD, Sidney Kimmel Medical College at Thomas Jefferson University, 1025 Walnut St #100, Philadelphia, PA 19107 (e-mail: lindsay.nadkarni@gmail.com).

The authors declare no conflict of interest.

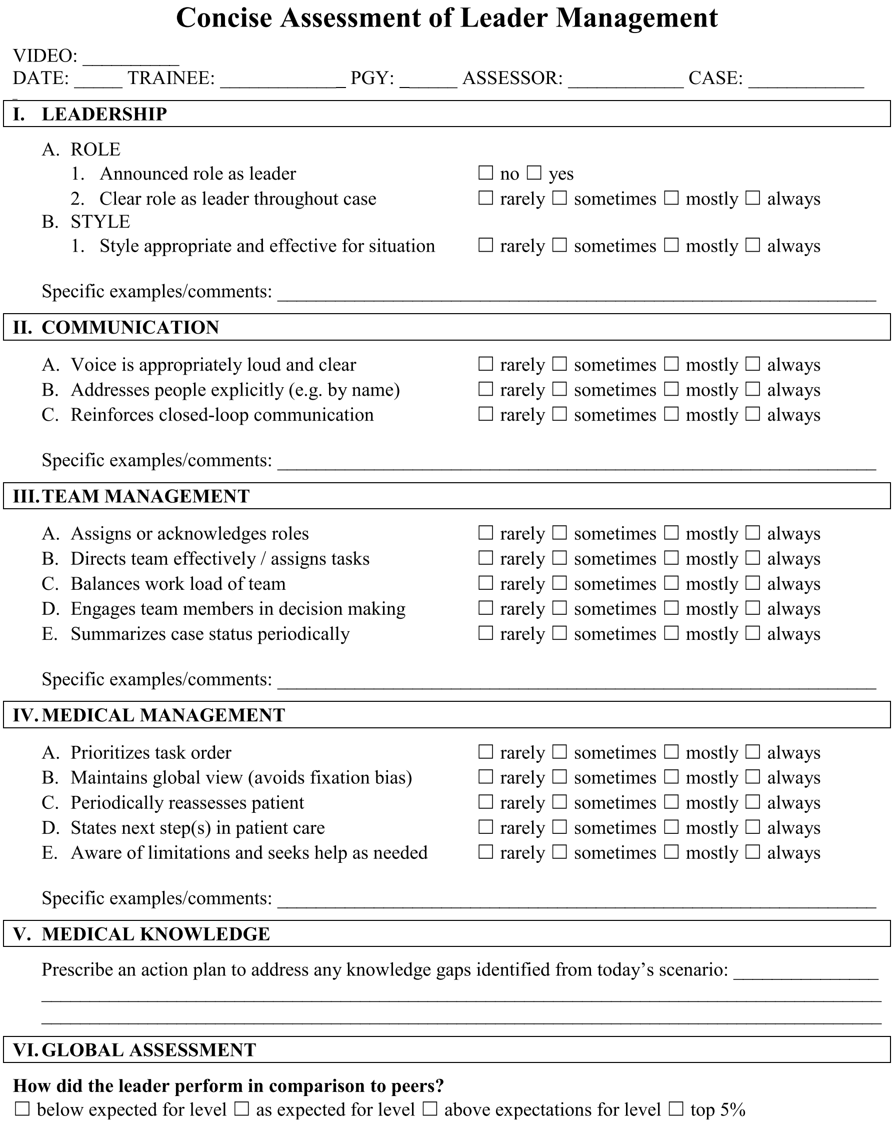
Copyright © 2017 Society for Simulation in Healthcare
DOI: 10.1097/SIH.0000000000000267

Pediatric Research, Innovation, and Education (INSPIRE)[30] to review videos of simulated resuscitations and score the performance of the resuscitation leader using the CALM instrument. Videos were selected from the archive of the Improving Pediatric Acute Care Through Simulation (ImPACTS) group with institutional review board approval obtained through Yale University.[31]

Instrument Development Process

The CALM instrument was developed using existing assessment instruments in the literature, professional experience, and expert consensus. The goal was to create a concise and pragmatic instrument that could be implemented by educators with minimal training. Three experts (including authors D.O.K. and C.G.R.) in graduate medical education and pediatric emergency medicine (PEM) met bimonthly over 3 months to review existing instruments and articles that reported validity data supporting their use in the assessment of leader and team performance.[20,28,29,32–35] Questions/elements/themes were abstracted in their original wording. Duplicates were then consolidated, phrasing was iteratively refined for simplicity, and questions were prioritized via a modified Delphi process resulting in an initial 18-item assessment instrument. The initial CALM instrument was then pilot tested by PEM attendings at 1 institution over a 3-month period to assess resuscitation leaders during mock resuscitations in the emergency department. During the pilot, raters received no specific training on the use of the CALM instrument, because the goal was to generate a user-friendly instrument that required no training; they were simply instructed to use the tool to assess the resuscitation leader's performance. Feedback from pilot raters was incorporated, and the final CALM instrument was developed.

The final CALM instrument consists of 15 four-point Likert scale items and 1 dichotomous behavioral item divided into 4 overall domains based on the 4 major elements of leadership in an acute resuscitation scenario: (1) leadership (role/style), (2) communication, (3) team management, and (4) medical management. Additional questions were added to aid in formative feedback (but were not included in the CALM score), including a free text item that asks about specific gaps in medical knowledge, and a global rating scale item assessing comparison with peers (Fig. 1).



Concise Assessment of Leader Management

VIDEO: __________
DATE: _____ TRAINEE: ____________ PGY: _____ ASSESSOR: ___________ CASE: ___________

I. LEADERSHIP

A. ROLE	
1. Announced role as leader	☐ no ☐ yes
2. Clear role as leader throughout case	☐ rarely ☐ sometimes ☐ mostly ☐ always
B. STYLE	
1. Style appropriate and effective for situation	☐ rarely ☐ sometimes ☐ mostly ☐ always

Specific examples/comments: ____

II. COMMUNICATION

A. Voice is appropriately loud and clear	☐ rarely ☐ sometimes ☐ mostly ☐ always
B. Addresses people explicitly (e.g. by name)	☐ rarely ☐ sometimes ☐ mostly ☐ always
C. Reinforces closed-loop communication	☐ rarely ☐ sometimes ☐ mostly ☐ always

Specific examples/comments: ____

III. TEAM MANAGEMENT

A. Assigns or acknowledges roles	☐ rarely ☐ sometimes ☐ mostly ☐ always
B. Directs team effectively / assigns tasks	☐ rarely ☐ sometimes ☐ mostly ☐ always
C. Balances work load of team	☐ rarely ☐ sometimes ☐ mostly ☐ always
D. Engages team members in decision making	☐ rarely ☐ sometimes ☐ mostly ☐ always
E. Summarizes case status periodically	☐ rarely ☐ sometimes ☐ mostly ☐ always

Specific examples/comments: ____

IV. MEDICAL MANAGEMENT

A. Prioritizes task order	☐ rarely ☐ sometimes ☐ mostly ☐ always
B. Maintains global view (avoids fixation bias)	☐ rarely ☐ sometimes ☐ mostly ☐ always
C. Periodically reassesses patient	☐ rarely ☐ sometimes ☐ mostly ☐ always
D. States next step(s) in patient care	☐ rarely ☐ sometimes ☐ mostly ☐ always
E. Aware of limitations and seeks help as needed	☐ rarely ☐ sometimes ☐ mostly ☐ always

Specific examples/comments: ____

V. MEDICAL KNOWLEDGE

Prescribe an action plan to address any knowledge gaps identified from today's scenario: ______________
____
____

VI. GLOBAL ASSESSMENT

How did the leader perform in comparison to peers?
☐ below expected for level ☐ as expected for level ☐ above expectations for level ☐ top 5%

FIGURE 1. The final CALM instrument that was distributed to raters.

TABLE 1. The 16 Videos Sent to Raters, Consisting of 4 Leaders Each Performing in 4 Resuscitation Scenarios

	Scenario A (Child Cardiac Arrest—Drowning)	Scenario B (Infant Respiratory Arrest—Foreign Body)	Scenario C (Infant Seizure—Hypoglycemia)	Scenario D (Infant Sepsis—Bacteremia)
Leader A	Video 1	Video 2	Video 3	Video 4
Leader B	Video 5	Video 6	Video 7	Video 8
Leader C	Video 9	Video 10	Video 11	Video 12
Leader D	Video 13	Video 14	Video 15	Video 16

Video Assessment and Data Collection

A total of 16 unique videos were abstracted from the ImPACTS database to include videos of 4 different resuscitation team leaders each performing in 4 separate scenarios (Table 1). These 16 videos were distributed to 4 independent raters.

The videos selected from the ImPACTS database captured the performance of actual interprofessional teams of health care providers caring for 4 simulated pediatric resuscitation scenarios: (1) child cardiac arrest (drowning), (2) infant respiratory arrest (foreign body), (3) infant seizure (hypoglycemia), and (4) infant sepsis (bacteremia). Scenarios were diverse, requiring different amounts of teamwork and sophistication. Teams were composed of clinicians (no trainees), involving 1 or 2 physicians (board certified in PEM or emergency medicine), 3 to 5 nurses, and 2 to 3 certified nursing assistants or emergency medicine technicians. The videos of each team performing in the 4 scenarios were obtained during a single 2.5-hour simulation session and filmed from a standard angle using the B-line Live Capture Ultraportable system (B-Line Medical, Washington, DC).[31]

We selected 4 independent raters from within the INSPIRE network who were PEM fellowship directors representing different academic institutions across the country. The order of the 16 videos was randomized for each rater with access provided via a password-protected Web-based file-sharing application.[36,37] Raters were instructed to use the CALM instrument to rate the resuscitation leader in each of the 16 videos to the best of their ability without any further specific instructions on how to use the instrument. Each video was reviewed only once, without pausing or rewinding during the playback, viewed in order of randomization. Pauses were permitted between videos.

Validity Framework

We followed Messick's framework for validity and report on content and internal structure validity.[38–40] Content validity refers to whether the content of the instrument measures its intended constructs. This was assessed based on the steps taken to develop the CALM tool. Internal structure validity assesses whether the instrument has acceptable reliability. This was assessed by generalizability analysis, which identifies the amount variation attributable to the person (leader), rater, and scenario and combinations of those factors and yields a generalizability coefficient (g-coefficient). A decision study (D-study) looks at the stability of the g-coefficient when different study design parameters are hypothetically changed (i.e. what the g-coefficient would be if greater or fewer raters or scenarios were used).[41,42]

Statistics

We conducted a fully crossed, person × scenario × rater (p × s × r) design using generalizability analysis to evaluate individual factor and factor interactions relating to score variance in CALM scores.[41,42] For each instrument, 4 raters scored each of the 4 scenarios for each leader. Variance components were obtained using IBM SPSS 22 (Armonk, NY) VARCOMP command (restricted [residual] maximum likelihood [REML] method). Generalizability (G) and decision (D) coefficients were calculated based on these components.

RESULTS

All 4 raters completed ratings for each video on each of the leadership elements on the CALM instrument.

Content Validity

The CALM instrument was developed by experts in pediatric graduate medical education and PEM and was based off of existing resuscitation leader assessment instruments. These were then subjected to a modified Delphi process with iterative revisions and then pilot tested by PEM attendings, supporting content validity. Themes were identified and categorized into 4 major domains of resuscitation leadership: (1) leadership (role/style), (2) communication, (3) team management, and (4) medical management.

Internal Structure Validity

Table 2 shows the mean CALM score and SD for each of the 16 videos. Results of the generalizability analysis are given in Table 3. The main contributors of score variability were the person (33%) and interaction of scenario × rater (16%) and person × rater (14%). Importantly, person was the largest contributor to variance. This indicates that the score variation largely reflects inter-subject variation in performance, which may be attributable to the inherent differences in knowledge and skill levels between subjects. The substantial person × scenario component indicates that there was variation for a given subject across the scenarios, also indicating that

TABLE 2. Mean CALM Score (of a Total Possible Score of 74) and SD for Each of the 16 Videos

Video	Mean Score	SD
1	45.0	9.6
2	54.5	3.1
3	50.8	5.4
4	56.5	4.4
5	38.3	2.8
6	42.8	3.0
7	40.0	4.2
8	45.3	4.7
9	45.0	9.9
10	42.3	6.4
11	44.5	5.1
12	46.0	9.9
13	41.0	6.4
14	36.8	9.6
15	40.5	3.3
16	44.0	2.4

TABLE 3. G-Study Results With the Estimate of Variance Attributable to Each Component (Person, Rater, and Scenario) and the Interaction of These Components

Variance Component	Estimate	% of Total Variance
Person	38.3	32.9
Rater	0	0.0
Scenario	1.9	1.6
Person × scenario	5.4	4.6
Person × rater	16.3	14.0
Scenario × rater	19.1	16.4
Error	35.5	30.5

leaders may have been more familiar with one scenario than another.

The rater facet had virtually no contribution to variance (0%), which implies that the raters were in agreement about the assessment of the various leaders and argues for high inter-rater reliability. The g-study for 4 raters and 4 subjects resulted in an absolute generalizability coefficient of 0.80. The D-study, which shows the theoretical effect of changing the number of raters or scenarios on the generalizability coefficient, is shown in Figure 2.

Of note, the error variance contributed 31% to the overall variance in scores. This represents possible triple order interactions (i.e. the interaction of person, scenario, and rater together) as well as other unidentified factors, possibly due to incomplete capture of scenarios by video or differences in camera angles, and bears further investigation.

DISCUSSION

In this prospective validation study, we present initial evidence on content and internal structure validity to support the use of the CALM instrument as a reliable tool to provide formative feedback to leaders of simulated pediatric resuscitations. The instrument was rigorously developed based off of existing tools, professional experience, and expert consensus and subjected to modified Delphi process and pilot testing. The generalizability study yielded a generalizability coefficient of 0.80, which is above the acceptable range of 0.70 to 0.79 for formative assessments and is consistent with the performance assessment literature.[43–45]

The CALM instrument is a concise, easy-to-use instrument that requires minimal rater training to assess team leaders of simulated pediatric resuscitations for the provision of formative feedback. Several other tools to address resuscitation leaders exist, although none of them are as brief and focused on the leader as ours is. The Simulation Team Assessment Tool, while excellent for research, may be cumbersome in practice, with 94 discrete tasks evaluating multiple domains and not exclusive to the team leader.[25] It was validated using raters who received 4 hours of training and practice along with very detailed definitions and was not intended for real-time evaluation. The Resuscitation Team Leader Evaluation is another tool that was designed to comprehensively assess resuscitation team leaders but may similarly be considered unwieldy for real-time use.[27] Another instrument was developed to assess clinical performance during Pediatric Advanced Life Support simulated scenarios.[21] This instrument is designed to be used for specific scenarios and therefore may not be as generalizable as our instrument, which was applied across a variety of scenarios.

Validation of assessment instruments is increasingly important because simulation and assessments guiding feedback are being used frequently in medical education. It is important to understand that validation is a continual process, whereby validity evidence is collected for an intended use. For results and conclusions to be valid, the validity data must be continually reassessed with regard to context and application. In a recent article outlining important principles in interpreting and assessing validity arguments, Cook and Hatala[44] conclude that validation studies gather validation evidence, but one study will not support all aspects of validity. Rather, it is important to identify gaps and the context in which the instrument should be used.

We validated our instrument in the context that it is intended to be used in, which is real time, "off the shelf" with minimal rater training. In its current iteration, the instrument is intended primarily as a means of providing formative feedback. Thus, although the long-term effects of the instrument's use on learner behavior were not assessed, the psychometrics presented previously are adequate to support this usage, implying an appropriate consequence validity when applied in formative situations. Applying the instrument

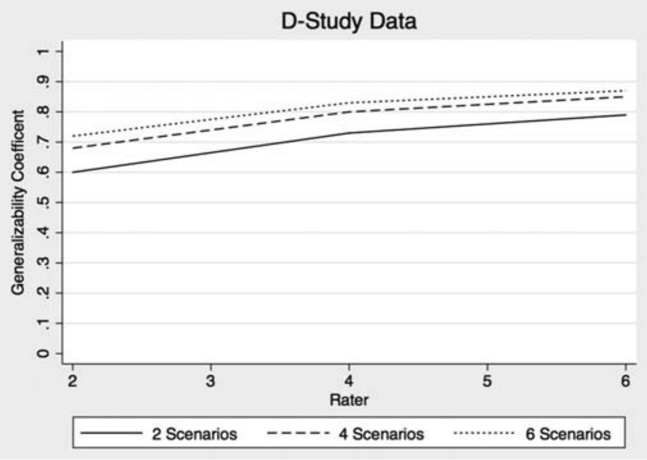


FIGURE 2. The D-study data showing the theoretical effect of changing the number of scenarios or raters in the study on the generalizability coefficient.

in more high-stakes scenarios, however, would require additional study focusing on the relationship between the instrument scores and long-term clinical performance of the residents assessed.

LIMITATIONS

The major limitation in this study was the use of videos. Although the videos were required for feasibility of the study, and were the closest to "real-time" possible, some actions may have been hard to hear or see simply because of the way they were recorded. For example, the leader may have "announced role as leader" before the videotape began. This likely was a contributing factor to the large percentage of variance attributed to error in the generalizability study. In addition, the phrasing of the tool, although concise, may have allowed for multiple interpretations of the same response options, also contributing to the error variance. For example, if a leader asked for input from other team members once during the simulation, a rater may have had difficulty determining whether they should receive credit for "always" "engaging team members in decision making," or if that would better be classified as "mostly" or "sometimes." It may be beneficial to add a few brief sentences to future iterations of the tool to define the anchored rating scale so that there is a more cohesive understanding of the meaning of each response. Another limitation is that raters were all PEM fellowship directors and experts in leadership. This may affect the generalizability of our study, such that nonexperts in leadership may not rate leaders using the CALM instrument similarly. The small sample size, with only 16 videos, is also a limitation. Although the generalizability study was fully crossed (4 leaders, 4 scenarios, and 4 raters), a larger sample size may alter the generalizability and φ coefficients. This underscores the preliminary nature of this validation study. In addition, we did not gather learner feedback regarding the usefulness of the formative data provided by the instrument. This will be a key area of further research, because such data are needed to support the instrument's stated purpose.

CONCLUSIONS

These results provide initial evidence to support the validity of the CALM instrument as a reliable assessment instrument that can guide the provision of formative feedback to leaders of pediatric simulated resuscitations. Although further validation data is needed, we recommend the initial usage of the instrument in this manner and offer it to the simulation community in the hope that it assists facilitators to shape their learners' future crisis resource management practice.

REFERENCES

1. Nadel FM, Lavelle JM, Fein JA, Giardino AP, Decker JM, Durbin DR. Assessing pediatric senior residents' training in resuscitation: fund of knowledge, technical skills, and perception of confidence. *Pediatr Emerg Care* 2000;16(2):73–76.

2. Chen EH, Cho CS, Shofer FS, Mills AM, Baren JM. Resident exposure to critical patients in a pediatric emergency department. *Pediatr Emerg Care* 2007;23(11):774–778.

3. Guilfoyle FJ, Milner R, Kissoon N. Resuscitation interventions in a tertiary level pediatric emergency department: implications for maintenance of skills. *CJEM* 2011;13(2):90–95.

4. Chen EH, Shofer FS, Baren JM. Emergency medicine resident rotation in pediatric emergency medicine: what kind of experience are we providing? *Acad Emerg Med* 2004;11(7):771–773.

5. Knudson JD, Neish SR, Cabrera AG, et al. Prevalence and outcomes of pediatric in-hospital cardiopulmonary resuscitation in the United States: an analysis of the Kids' Inpatient Database*. *Crit Care Med* 2012; 40(11):2940–2944.

6. Topjian AA, Nadkarni VM, Berg RA. Cardiopulmonary resuscitation in children. *Curr Opin Crit Care* 2009;15(3):203–208.

7. Bhanji F, Donoghue AJ, Wolff MS, et al. Part 14: Education: 2015 American Heart Association Guidelines Update for Cardiopulmonary Resuscitation and Emergency Cardiovascular Care. *Circulation* 2015; 132(18 Suppl 2):S561–S573.

8. Cooper S, Wakelam A. Leadership of resuscitation teams: "Lighthouse Leadership". *Resuscitation* 1999;42(1):27–45.

9. Gilfoyle E, Gottesman R, Razack S. Development of a leadership skills workshop in paediatric advanced resuscitation. *Med Teach* 2007;29(9): e276–e283.

10. Hunziker S, Buhlmann C, Tschan F, et al. Brief leadership instructions improve cardiopulmonary resuscitation in a high-fidelity simulation: a randomized controlled trial. *Crit Care Med* 2010;38(4): 1086–1091.

11. Fernandez Castelao E, Boos M, Ringer C, Eich C, Russo SG. Effect of CRM team leader training on team performance and leadership behavior in simulated cardiac arrest scenarios: a prospective, randomized, controlled study. *BMC Med Educ* 2015;15:116.

12. Marasch SC, Müller C, Marquardt K, Conrad G, Tschan F, Hunziker PR. Human factors affect the quality of cardiopulmonary resuscitation in simulated cardiac arrests. *Resuscitation* 2004;60(1):51–56.

13. Yeung JH, Ong GJ, Davies RP, Gao F, Perkins GD. Factors affecting team leadership skills and their relationship with quality of cardiopulmonary resuscitation. *Crit Care Med* 2012;40(9):2617–2621.

14. Weller J, Boyd M, Cumin D. Teams, tribes and patient safety: overcoming barriers to effective teamwork in healthcare. *Postgrad Med J* 2014;90: 149–154.

15. Nishisaki A, Nguyen J, Colborn S, et al. Evaluation of multidisciplinary simulation training on clinical performance and team behavior during tracheal intubation procedures in a pediatric intensive care unit. *Pediatr Crit Care Med* 2011;12(4):406–414.

16. Issenberg SB, McGaghie WC, Petrusa ER, Lee Gordon D, Scalese RJ. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. *Med Teach* 2005;27(1): 10–28.

17. Cheng A, Goldman RD, Aish MA, Kissoon N. A simulation-based acute care curriculum for pediatric emergency medicine fellowship training programs. *Pediatr Emerg Care* 2010;26(7):475–480.

18. Doughty CB, Kessler DO, Zuckerbraun NS, et al. Simulation in pediatric emergency medicine fellowships. *Pediatrics* 2015;136(1):e152–e158.

19. McGaghie WC, Issenberg SB, Petrusa ER, Scalese RJ. A critical review of simulation-based medical education research: 2003–2009. *Med Educ* 2010;44:50–63.

20. Cooper S, Cant R, Porter J, et al. Rating medical emergency teamwork performance: development of the Team Emergency Assessment Measure (TEAM). *Resuscitation* 2010;81(3):446–452.

21. Donoghue A, Nishisaki A, Sutton R, Hales R, Boulet J. Reliability and validity of a scoring instrument for clinical performance during Pediatric Advanced Life Support simulation scenarios. *Resuscitation* 2010; 81(3):331–336.

22. Kim J, Neilipovitz D, Cardinal P, Chiu M. A comparison of global rating scale and checklist scores in the validation of an evaluation tool to assess performance in the resuscitation of critically ill patients during simulated emergencies (abbreviated as "CRM simulator study IB"). *Simul Healthc* 2009;4(1):6–16.

23. Brett-Fleegler MB, Vinci RJ, Weiner DL, Harris SK, Shih MC, Kleinman ME. A simulator-based tool that assesses pediatric resident resuscitation competency. *Pediatrics* 2008;121(3):e597–e603.

24. Donoghue A, Ventre K, Boulet J, et al. Design, implementation, and psychometric analysis of a scoring instrument for simulated pediatric resuscitation: a report from the EXPRESS pediatric investigators. *Simul Healthc* 2011;6(2):71–77.

25. Reid J, Stone K, Brown J, et al. The Simulation Team Assessment Tool (STAT): development, reliability and validation. *Resuscitation* 2012; 83(7):879–886.

26. Lockyer J, Singhal N, Fidler H, Weiner G, Aziz K, Curran V. The development and testing of a performance checklist to assess neonatal resuscitation megacode skill. *Pediatrics* 2006;118(6):e1739–e1744.

27. Grant EC, Grant VJ, Bhanji F, Duff JP, Cheng A, Lockyer JM. The development and assessment of an evaluation tool for pediatric resident competence in leading simulated pediatric resuscitations. *Resuscitation* 2012;83(7):887–893.

28. LeFlore JL, Anderson M, Michael JL, Engle WD, Anderson J. Comparison of self-directed learning versus instructor-modeled learning during a simulated clinical experience. *Simul Healthc* 2007; 2(3):170–177.

29. LeFlore JL, Anderson M. Alternative educational models for interdisciplinary student teams. *Simul Healthc* 2009;4(3):135–142.

30. International Network for Simulation-based Pediatric Innovation, Research, & Education Website. Available at: http://inspiresim.com. Accessed February 29, 2016.

31. Yale School of Medicine Web site. Improving Pediatric Acute Care Through Simulation. Available at: http://medicine.yale.edu/lab/impacts/. Accessed March 29, 2016.

32. Calhoun AW, Boone M, Miller KH, Taulbee RL, Montgomery VL, Boland K. A multirater instrument for the assessment of simulated pediatric crises. *J Grad Med Educ* 2011;3(1):88–94.

33. Zajano EA, Brown LL, Steele DW, Baird J, Overly FL, Duffy SJ. Development of a survey of teamwork and task load among medical providers: a measure of provider perceptions of teamwork when caring for critical pediatric patients. *Pediatr Emerg Care* 2014;30(3):157–160.

34. Jelovsek JE, Kow N, Diwadkar GB. Tools for the direct observation and assessment of psychomotor skills in medical trainees: a systematic review. *Med Educ* 2013;47(7):650–673.

35. Hunt EA, Walker AR, Shaffner DH, Miller MR, Pronovost PJ. Simulation of in-hospital pediatric medical emergencies and cardiopulmonary arrests: highlighting the importance of the first 5 minutes. *Pediatrics* 2008;121(1):e34–e43.

36. Box Web site. Available at: http://box.com. Accessed November 2014.

37. Random.org Web site. Available at: http://www.random.org/lists/. Accessed December 26, 2014.

38. Messick S. Validity. In: Linn RL, ed. *Educational Measurement*. 3rd ed. New York, NY: American Council on Education and Macmillan; 1989.

39. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: a practical guide to Kane's framework. *Med Educ* 2015;49(6):560–575.

40. Brennan RL. Performance assessments from the perspective of generalizability theory. *Appl Psychol Meas* 2001;24(4):339–353.

41. Brennan RL. Generalizability Theory. *Educ Meas Issues Prac* 1992; 11(4):27–34.

42. Cronbach LJ. *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. New York, NY: Wiley; 1972.

43. Boulet JR. Summative assessment in medicine: the promise of simulation for high-stakes evaluation. *Acad Emerg Med* 2008;15(11):1017–1024.

44. Cook DA, Hatala R. Validation of educational assessments: a primer for simulation and beyond. *Adv Simul* 2016;1:31.

45. Downing SM. Reliability: on the reproducibility of assessment data. *Med Educ* 2004;38:1006–1012.