Design, Implementation, and Psychometric Analysis of a Scoring Instrument for Simulated Pediatric Resuscitation: A Report from the EXPRESS Pediatric Investigators

Aaron Donoghue, MD, MSCE;

Kathleen Ventre, MD;

John Boulet, PhD;

Marisa Brett-Fleegler, MD;

Akira Nishisaki, MD;

Frank Overly, MD;

Adam Cheng, MD;

for the EXPRESS Pediatric Simulation Research Investigators

Introduction: Robustly tested instruments for quantifying clinical performance during pediatric resuscitation are lacking. Examining Pediatric Resuscitation Education through Simulation and Scripting Collaborative was established to conduct multicenter trials of simulation education in pediatric resuscitation, evaluating performance with multiple instruments, one of which is the Clinical Performance Tool (CPT). We hypothesize that the CPT will measure clinical performance during simulated pediatric resuscitation in a reliable and valid manner.

Methods: Using a pediatric resuscitation scenario as a basis, a scoring system was designed based on Pediatric Advanced Life Support algorithms comprising 21 tasks. Each task was scored as follows: task not performed (0 points); task performed partially, incorrectly, or late (1 point); and task performed completely, correctly, and within the recommended time frame (2 points). Study teams at 14 children's hospitals went through the scenario twice (PRE and POST) with an interposed 20-minute debriefing. Both scenarios for each of eight study teams were scored by multiple raters. A generalizability study, based on the PRE scores, was conducted to investigate the sources of measurement error in the CPT total scores. Inter-rater reliability was estimated based on the variance components. Validity was assessed by repeated measures analysis of variance comparing PRE and POST scores.

Results: Sixteen resuscitation scenarios were reviewed and scored by seven raters. Inter-rater reliability for the overall CPT score was 0.63. POST scores were found to be significantly improved compared with PRE scores when controlled for within-subject covariance ($F_{1,15} = 4.64$, $P < 0.05$). The variance component ascribable to rater was 2.4%.

Conclusions: Reliable and valid measures of performance in simulated pediatric resuscitation can be obtained from the CPT. Future studies should examine the applicability of trichotomous scoring instruments to other clinical scenarios, as well as performance during actual resuscitations.

(*Sim Healthcare* 6:71–77, 2011)

Key Words: Pediatric, Resuscitation, Pediatric advanced life support.

Resuscitating critically ill and injured children remains among the most challenging clinical areas in which residents and fellows need to be trained. The rarity of pediatric resuscitation events, combined with progressive limitations in allowable work hours, makes it difficult for physicians to acquire adequate resuscitation management experience during their training.[1–3] In addition, the challenges of capturing and tracking data from resuscitations make it difficult to design appropriately powered trials that are capable of evaluating the efficacy of training interventions on resuscitation team performance and overall patient outcomes. Addressing existing deficiencies in the performance of fully credentialed resuscitation teams will ultimately depend on developing robust tools to evaluate individual and collective team performance during resuscitation events.[4]

Simulation training has gained widespread acceptance as a technique for training pediatric and adult health care teams to conduct resuscitations in accordance with established guidelines. A growing body of evidence has evaluated the effectiveness of simulation training compared with traditional training for resuscitation, with mixed results.[5–7] In 2007, the Examining Pediatric Resuscitation Education through Simulation and Scripting (EXPRESS) Research Collaborative was established among several pediatric centers in the United States and Canada to facilitate a prospective, multicenter, randomized trial examining the impact of manne-

From the Divisions of Emergency Medicine (A.D.) and Critical Care Medicine (A.D., A.N.), The Children's Hospital of Philadelphia, Philadelphia, PA; Division of Critical Care Medicine (K.V.), University of Utah, Salt Lake City, UT; Foundation for Advancement of International Medical Education and Research (J.B.), Philadelphia, PA; Division of Emergency Medicine (M.B.-F.), Children's Hospital, Boston, MA; Division of Emergency Medicine (F.O.), Hasbro Children's Hospital, Providence, RI; and Division of Emergency Medicine (A.C.), British Columbia Children's Hospital, Vancouver, BC, Canada.

The authors declare no conflicts of interest.

The EXPRESS Pediatric Simulation Research Investigators are listed at the end of this article.

Reprints: Aaron Donoghue, MD, MSCE, Divisions of Emergency Medicine and Critical Care Medicine, The Children's Hospital of Philadelphia, 34th Street and Civic Center Boulevard, Philadelphia, PA 19104 (e-mail: donoghue@email.chop.edu).

Copyright © 2011 Society for Simulation in Healthcare
DOI: 10.1097/SIH.0b013e31820c44da

quin fidelity and debriefing technique on performance outcomes during a simulated pediatric cardiac arrest scenario. A major goal of the EXPRESS investigators was to develop and validate a set of instruments to measure provider performance in the clinical/psychomotor, behavioral, and cognitive domains during simulated pediatric resuscitation.

The Clinical Performance Tool (CPT) was one of the instruments developed by the EXPRESS Collaborative. We designed the CPT to quantitatively measure team clinical and psychomotor performance during a simulated resuscitation scenario. We sought to evaluate the reliability and validity of the CPT scores for use in the EXPRESS trial.

METHODS

The EXPRESS Research Collaborative and Trial

The EXPRESS Collaborative was first established in February 2007, at a meeting where 25 pediatric simulation and Pediatric Advanced Life Support (PALS) experts gathered to discuss strategies to effectively incorporate simulation-based education into future PALS courses. This meeting was held at the Children's Hospital of Philadelphia and served as a springboard for discussion of simulation-based research ideas. It also served to confirm the pressing need for collaborative, multicenter research in the field of simulation. The meeting culminated with a commitment to undertake the EXPRESS trial, a large prospective, multicenter, randomized trial to evaluate the impact of simulation fidelity and debriefing technique of novice instructors on performance of health care teams during simulated pediatric resuscitations.

For the EXPRESS trial, teams of health care professionals were recruited from 14 participating tertiary care pediatric hospitals by written or e-mail invitation by study personnel from their respective institutions. Institutional review board approval was obtained locally at each participating institution. Study teams consisted of one code team leader (a resident or nurse practitioner in pediatrics, emergency medicine, or anesthesia/critical care) and three to four other team members, made up of residents, nurses, or respiratory therapists. All participants gave written informed consent before completing the study. Teams were randomized into one of four study arms: (1) low-fidelity simulation and nonscripted debriefing; (2) low-fidelity simulation and scripted debriefing; (3) high-fidelity simulation and nonscripted debriefing; and (4) high-fidelity simulation and scripted debriefing. Scripted debriefings were conducted using an expert-designed script to guide the debriefer through a step-by-step review of critical medical and team performance behaviors. Trained raters measured team performance during each simulated clinical scenario by video review, using multiple instruments that were developed to assess specific competency domains (clinical, behavioral, and cognitive knowledge). CPT described in this article was one such instrument. The present analysis of the CPT involved assessing the instrument independent of study arm randomization, and the multirater assessment of the present analysis does not distinguish the underlying experimental conditions.

The scenario for the EXPRESS trial involved the resuscitation of a simulated infant with hypotension and shock, deteriorating into ventricular fibrillation. The scenario was divided into three distinct stages (Fig. 1): hypotension and shock (2 minutes in duration), ventricular fibrillation arrest (8 minutes maximum duration), and return of spontaneous circulation (2 minutes in duration). The scenario was allowed to progress through all three stages irrespective of how the teams performed. However, the duration of stage 2 could potentially be shortened if the team attempted to defibrillate the patient three times within 8 minutes.

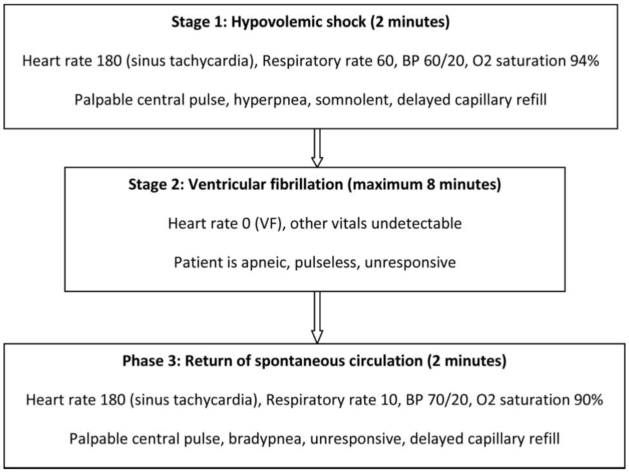

Stage 1: Hypovolemic shock (2 minutes)
Heart rate 180 (sinus tachycardia), Respiratory rate 60, BP 60/20, O2 saturation 94%
Palpable central pulse, hyperpnea, somnolent, delayed capillary refill
Stage 2: Ventricular fibrillation (maximum 8 minutes)
Heart rate 0 (VF), other vitals undetectable
Patient is apneic, pulseless, unresponsive
Phase 3: Return of spontaneous circulation (2 minutes)
Heart rate 180 (sinus tachycardia), Respiratory rate 10, BP 70/20, O2 saturation 90%
Palpable central pulse, bradypnea, unresponsive, delayed capillary refill

Figure 1. Resuscitation scenario.

Each study team managed the simulated clinical scenario twice (PRE-debriefing and POST-debriefing). After a 20-minute debriefing, the scenario was presented with a unique case stem, so as to suggest an entirely new scenario for participants, although the simulated patient in the second scenario demonstrated the exact physiologic states as the patient in the first scenario. The study diagram is shown in Figure 2.

Development of the Trichotomous CPT Scoring Instrument

The scoring methodology for the CPT was adapted directly from scoring instruments used in a previously published trial of resuscitation education using patient simulation.[1] In this trial, four brief task-based instruments were synthesized according to PALS algorithms, where tasks were assigned one of three possible scores (ie, 0, 1, or 2 points). Scores on individual tasks reflected whether each task was performed in a correct technical fashion, in proper sequence, and in a timely manner. Analysis of the scoring instruments from this trial demonstrated an overall inter-rater reliability (IRR) of 0.81. Residents in postgraduate year 2 had significantly better scores than those in postgraduate year 1, providing evidence to support construct validity. A fully crossed generalizability study demonstrated a minimal impact of the rater as a source of variance in scores.[8] Based on these performance characteristics, the EXPRESS investigators chose to synthesize the CPT based on similar methodology adapted specifically for the resuscitation scenario used in the EXPRESS trial.

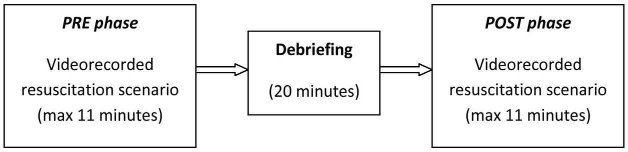

PRE phase
Videorecorded resuscitation scenario (max 11 minutes)
Debriefing
(20 minutes)
POST phase
Videorecorded resuscitation scenario (max 11 minutes)

Figure 2. Schematic of study team participation.

Simulated Scenario and Creation of Items and Definitions

Three investigators who are board-certified in pediatric emergency medicine or pediatric critical care medicine (A.D., K.V., M.B.-F.) used a consensus process to develop operational definitions for each element of the CPT. For each stage of the scenario, the investigators identified a series of key tasks that they expected the resuscitation team to perform within the corresponding time period. Tasks selected for inclusion in the scoring instrument were those emphasized in the management guidelines most recently published by the American Heart Association, which could also be reliably captured and rated through video review of a fully contextualized simulated resuscitation. Performance rating options for each task were collapsed into three categories: task not performed (0 points); task performed partially, incorrectly, or late (1 point); and task performed completely, correctly, and within the recommended time frame (2 points). For example, during stage 1 of the scenario, a team could earn two points for performing each of the eight key tasks if they opened the airway and assessed respiratory effort, work of breathing, and pulses within 30 seconds; if they properly connected the simulated patient to electrocardiogram monitoring and pulse oximetry within 60 seconds; and if they obtained a blood pressure, offered supplemental oxygen by nonrebreather mask, placed an intravenous or intraosseous line, and ordered a 20 mL/kg isotonic fluid bolus. The consensus scoring instrument along with anchoring behavioral descriptions for each rating option is shown in Table 1. The maximum possible score a resuscitation team could earn during the simulated scenario was 42 points. An option of "cannot tell" was available for each item; it was decided a priori that "cannot tell" responses would be eliminated from the total score possible (eg, one such response would reduce the maximum possible score from 42 points to 40 points).

CPT Training, Video Review, and Data Collection

Eight expert video reviewers composed of pediatric emergency medicine and pediatric critical care medicine physicians were trained to use the CPT before using the instrument to rate performance in the videotaped simulated resuscitation scenarios. Two separate training sessions were conducted. The first was a face-to-face training session conducted 12 months before initiation of the video review process. During this training session, all prospective video reviewers were given a 45-minute lecture that contained some background information, outlined the rationale for the tool, and provided an overview of the features of the trichotomous scoring system. The prospective video reviewers then watched two videos of healthcare teams resuscitating a simulated patient using the scenario employed in the trial. One of the videos depicted poor clinical performance and the other depicted excellent clinical performance. Reviewers were asked to rate the performance using the CPT after each video, and a subsequent group discussion allowed each reviewer to calibrate his or her ratings against those of other raters, with additional individualized feedback to clarify uncertainty or inconsistency in application of scoring definitions where appropriate. The second CPT training session was conducted 11 months later or 1 month before the raters were expected to begin scoring performance on the EXPRESS trial videos. This session was conducted over a conference call and included a brief review of the CPT instrument, followed by review and rating of health care team performance as depicted in three sample videos. As in the first training session, the raters then received individualized feedback on their scoring choices relative to the "gold standard" scoring for each sample video.

After completing both CPT training sessions, the video reviewers were given access to the web-based, password-protected EXPRESS research portal where the study videos would be posted. Each of the eight trained video reviewers used the CPT to rate eight pairs of videos, representing the PRE-debriefing and POST-debriefing performance of eight unique health care teams. These video recordings were randomly selected from 8 of the 14 study sites (one per site); reviewers were blinded to study site and study phase. The videos used to validate the CPT were subsequently removed from analysis for the main EXPRESS study. In total, 16 common videos were distributed for scoring by the eight video reviewers. CPT ratings were entered online and collected by the research portal for subsequent data analysis.

Statistical Analysis

Mean overall scores on the CPT were calculated for each PRE- and POST-debriefing video. A person (team) by rater generalizability study was conducted to partition the sources of variability in the CPT total scores. As team ability was expected to become more homogeneous after debriefing, this analysis was based only on the PRE scores. The estimated variance components were used to calculate an overall measure of IRR. This represents the average correlation in scores between any two randomly selected raters.

A repeated measures analysis of variance with study phase (ie, PRE- or POST-debriefing) as an independent variable was performed to gather evidence to support the validity of the scores.

RESULTS

Seven raters completed review and scoring of 16 videos, one PRE/POST-debriefing pair of videos for each of the eight study groups. By unadjusted univariate analysis, mean overall performance scores across all study teams were significantly better in the POST phase compared with the PRE phase (29 ± 3 points vs. 25 ± 4 points, Wilcoxon rank-sum $P < 0.0001$).

Based on the estimated variance components (PRE scores), and the seven raters who completed all rating tasks, IRR was calculated to be 0.63. The variance attributable to the rater was small (2.4% of the total variance), indicating that, over the eight teams, the raters provided similar mean scores. By repeated measures analysis of variance, POST-debriefing scores were found to be significantly improved when controlled for within-subject covariance ($F_{1,15} = 4.64, P < 0.05$).

DISCUSSION

Our study of the CPT showed that the instrument had moderate IRR of 0.63. In addition, the statistically significant increase in overall performance scores for study groups undergoing their second assessment suggests evidence of construct validity, although the strength of this inference is tem-

Table 1. The Clinical Performance Tool (CPT)

Task	0 Points	1 Point	2 Points
STAGE 1: Shock (tachycardia, hypotension, tachypnea) – 2 minutes			
Open Airway/assess airway patency	● Not done	● >30 secs	● Opens airway ● Recognizes child is vocalizing ● <30 secs
Assess breathing	● Not done	● >30 sec ● Done by auscultation only (no recognition of diminished chest wall movement)	● <30 sec, auscultation, tachypnea and WOB assessed
Oxygen	● Not done	● Nasal cannula or blow-by O2 (not 100%)	● 100% O2 applied
Monitors	● Not done	● >60 secs ● Pulse ox OR cardiac monitors applied (not both)	● <60 sec and proper placement ● Pulse ox and cardiac monitors applied
Pulse Check	● Not done	● >30 secs	● <30 sec
Blood pressure	● Not done	● >120 seconds (ie. Done in stage 2)	● <120 seconds
IV/IO	● Not done	● One PIV / IO done in >120 seconds (ie. Done in stage 2)	● One PIV / IO done in <120 seconds
Fluid bolus	● Not done	● Wrong fluid ordered ● Wrong amount ordered	● 20 cc/kg isotonic fluid ordered
STAGE 2: VF arrest or Pulseless Vtach arrest – 8 minutes			
Pulse check	● Not done	● >30 seconds after VF occurs (scenario time clock >2:30) ● Peripheral pulse checked	● <30 sec after VF occurs AND central pulse checked (scenario time <2:30)
Rhythm Identification	● Not done	● Does not verbalize rhythm but demonsrates awareness of rhythm ● Verbalizes incorrect rhythm	● Verbalizes correct rhythm
Effective ventilation	● Not done	● >30 seconds after apnea occurs (scenario time >2:30) ● Improper ventilation rate	● <30 secs after apnea occurs (scenario time <2:30) ● Proper ventilation rate and ratio (if not intubated)
CPR	● Not done	● >30 sec after pulselessness recognized (scenario time >2:30) ● Poor CPR technique (wrong hand position, improper rate, disruptions in CPR, does not check pulse with CPR)	● <30 sec after pulselessness recognized (scenario time <2:30) AND good CPR technique AND checks pulse with CPR
Defibrillation (first)	● Not done ● Attempted but electricity not delivered to patient (e.g. pads not on, etc.)	● Wrong dose ● Wrong mode ● >90 sec after rhythm change (scenario time >3:30)	● <90 sec after rhythm change (scenario time <3:30) AND correct dose AND correct mode
CPR continued	● Not done	● Delayed for >30 seconds ● Poor CPR technique (wrong hand position, improper rate, disruptions in CPR, does not check pulse with CPR)	● Initiated immediately following first shock with no delay and no pulse check, AND good CPR technique AND checks pulse with CPR
Defibrillation (second)	● Not done ● Attempted but electricity not delivered to patient (e.g. pads not on, etc.)	● Wrong dose ● Wrong mode ● >120 sec (or >5 cycles of CPR) after last shock (scenario time >5:30)	● Done 120 sec or 5 cycles of CPR after last shock (ie. scenario time <5:30) AND correct dose AND correct mode
Rhythm Identification	● Not done	● Does not verbalize rhythm but demonsrates awareness of rhythm ● Verbalizes incorrect rhythm	● Verbalizes correct rhythm
Pulse Check	● Not done	● Done >30 secs after second shock ● Done in incorrect sequence ● Peripheral Pulse checked	● Done immediately after second shock ● Central pulse checked
CPR continued	● Not done	● Delayed for >30 secs ● Poor CPR technique (wrong hand position, improper rate, disruptions in CPR, does not check pulse with CPR)	● Initiated immediately after pulse check and rhythm identification (<30 secs) AND good CPR technique AND checks pulse with CPR
Epinephrine	● Not given	● Incorrect dose ● Suboptimal route (ETT) ● Given prior to second defibrillation	● IV / IO epi dose given ● Correct dose given ● Given following second defibrillation
Defibrillation (third)	● Not done ● Attempted but electricity not delivered to patient (e.g. pads not on, etc.)	● Wrong dose ● Wrong mode ● >120 sec or >5 cycles of CPR after last shock (ie. scenario time >7:30)	● Done 120 sec or 5 cycles of CPR after last shock (ie. scenario time <7:30) AND correct dose AND correct mode
STAGE 3: ROSC – 2 minutes			
Pulse check	● Not done	● Done >60 seconds after state change (ie. Scenario time >11:00)	● Done within 60 sec of state change (ie. Scenario time <11:00)

pered by the lack of a control group or preexisting "gold standard" to which it can be compared.

Donoghue et al[8] recently published the results of an analysis of the reliability and validity of a set of scoring instruments used in PALS scenarios, where a notably higher overall IRR of 0.81 was noted. Several differences between that trial and the EXPRESS trial bear mentioning as important potential factors accounting for such a difference. The original trial involved a series of four brief scenarios, scored with four to seven tasks each, performed by individual subjects at one center; four raters scored the video recordings as part of data collection. By contrast, the EXPRESS trial design involved a single scenario of 21 tasks undertaken by teams of subjects at multiple institutions scored by eight raters.

Instruments intending to provide a quantitative assessment of human performance have been developed for a variety of acute care skills, including pediatric airway management,[9] neonatal resuscitation,[10] pediatric resuscitation,[11] and pediatric trauma care.[12] Although these reports are among the most important early contributions to the pediatric simulation literature, they also illustrate some important obstacles to generating experimental results that would be applicable in a variety of settings and circumstances. For example, each of these groups of investigators developed a unique set of outcome measures, all in the form of checklists of dichotomous items that were specifically suited to their own unique study conditions. Psychometric data were also largely absent from these studies, making it difficult to draw strong inferences about the reliability and validity of the reported outcomes. Hunt et al[4] summarized this issue in a recent report on pediatric house staff performance during simulated cardiac arrest scenarios, aptly stating that "a rigorously validated tool to measure errors during resuscitation is not available."

Given that arrest states arise uncommonly in pediatric patients,[13–15] and that survival rates from out of hospital and in-hospital arrests remain poor,[16,17] measuring clinical performance during pediatric resuscitations by survival outcomes alone becomes extremely difficult. Therefore, it is important to develop ways of assessing resuscitation team performance that are not dependent on patient survival statistics. Wayne et al[18] published the results of a study demonstrating that internal medicine residents who underwent simulation training exhibited better care delivery during in-hospital adult cardiac arrests, as measured by degree of adherence to Advanced Cardiac Life Support protocols and time to defibrillation. Importantly, this study did not demonstrate a difference in patient survival but did show a measurable difference in care delivery, as indicated by increased adherence to published treatment algorithms among simulation-trained residents, as compared with traditionally trained residents. We believe that our CPT would provide a feasible methodologic framework for measuring the impact of simulation training on pediatric team performance in the clinical setting.

The EXPRESS Research Collaborative consists of an international group of investigators who are recognized experts in pediatric resuscitation and simulation. Among the benefits of this multicenter collaboration is the ability to design and test robust performance rating instruments. The CPT was developed to facilitate improvements in pediatric resuscitation training. We believe that we have developed a reliable and valid instrument to evaluate health care provider team performance. Deriving the scoring metrics based on observable tasks with clear operational definitions was important for optimizing IRR. Vigorous expert rater training was another critical component for successful use of the CPT. Development of the task list from the widely accepted and rigorously developed resuscitation guidelines was essential to achieve robust face validity and content validity. Finally, the construct validity of our CPT instrument was supported by the salutary effect of debriefing on team performance.

Limitations

We acknowledge several important limitations of our study. The IRR achieved in our study (0.63) is moderate and notably different than that of the instrument(s) used in the original trial from which the CPT was derived. As discussed above, several potential reasons for this discrepancy exist; among them are differing study conditions, a greater number of raters, differences in the set of tasks in the instruments, and sampling error. Importantly, the generalizability study suggests that the variance component attributable to raters was minimal (2.4%), suggesting that the addition of raters would not be beneficial in terms of increasing reliability further.

The CPT was scored by multiple raters through video capture of simulated bedside resuscitations. Although all scenarios were recorded by a minimum of two video capture devices, it remains possible that scoring of tasks was inaccurate or impossible as a result of the limitations of static video and audio recording. Although the investigators attempted to standardize the simulated clinical environment at each participating study center, small differences between centers in video camera angles and the overall quality of audio/video recording were noted. Each item on the CPT had an option of "cannot tell" in an effort to track the frequency with which the video was not able to capture whether a particular skill was performed. Ultimately, the "cannot tell" response was chosen in only 3/2520 (0.1%) item responses, suggesting that, despite these potential pitfalls, the tasks selected for inclusion in the CPT were relatively easy to capture on video. Some, but not all, of the individual tasks on the CPT could be documented from the events logs of the simulator software and the sensing capability of the simulators themselves (eg, pulse check, chest compressions, and positive pressure ventilation). Nonetheless, the investigators decided to use only video review for scoring, despite these inherent simulator capabilities. It is not clear whether our approach (using video alone for scoring purposes), or a combination of both approaches (video review supplemented by simulator event log review), would be optimal for accurate assessment of task completion during simulated high-stakes, rapidly evolving clinical situations. In addition, it should be noted that this instrument is not well suited to application in real time, making its use in extemporaneous debriefing on clinical performance difficult.

All the subjects in this trial were pediatric house staff and nurses who were novices at resuscitation. It is not clear whether the CPT would perform equally well if used to rate the performance of more expert teams, who may perform critical tasks simultaneously or surreptitiously, making each task more difficult to reliably capture on video. Although the performance scores recorded during the EXPRESS trial were not high enough to suggest a "ceiling effect" for the CPT, additional study will be needed to determine whether the CPT would be a valid way to assess performance of the most expert providers. Further study will also be necessary to determine whether this type of instrument may be suitable for defining minimum competency standards for trainees.

The tasks scored by the CPT are each given equivalent weight toward the overall performance score for the scenario. Although the tasks were chosen based on established treatment guidelines, no attempt was made to weight specific tasks based on scientific evidence for the importance of their completion during the scenario. In addition, the CPT scoring rubric made no attempt to identify and deduct points for actions that could be potentially harmful. Ventre et al[19] recently published the results of a trial of a set of PALS scoring instruments for a computer-based PALS simulator, in which points were deducted for incorrect, delayed, or potentially harmful actions. Advantages exist for both approaches to scoring performance; further study will be needed to elucidate which types of instruments are easier to construct, administer, and score, as well as which are capable of measuring actual clinical performance in the most valid manner. Multiple complementary tools may be needed to provide the most thorough estimation of competency for high-stakes performance evaluations conducted across the spectrum of professional capability. With larger participant samples, detailed item analysis could be used to empirically identify which actions were most important in the determination of overall team ability.

Finally, while evidence to suggest that valid and reliable inferences can be made based on the CPT scores, the psychometric investigation of the instrument is far from complete. The reliability of the scores was estimated based on a single simulation scenario. As such, only measurement error associated with the rater could be investigated. Additional studies focusing on task sampling variability are warranted.[20] As the CPT was designed around a specific resuscitation scenario, the generalizability of the scores to other pediatric resuscitation events is not known. Further study of the validity of this methodology might be more ideally conducted through treatment versus control design and accounting for within-group covariance.[21]

CONCLUSIONS

A significant need exists for instruments that provide a quantitative assessment of clinical performance during resuscitation. Our CPT demonstrated acceptable reliability and construct validity. Effectiveness of simulation-based pediatric resuscitation education will be quantitatively assessed with this tool, among others, in future collaborative multicenter trials conducted by the EXPRESS investigators. Future studies should evaluate the applicability of these scoring metrics to performance assessment during actual pediatric resuscitations.

EXPRESS Pediatric Simulation Research Investigators

EXPRESS Pediatric Simulation Research Investigators: Elizabeth A. Hunt, MD; Kristen Nelson, MD; Judy Leflore, PhD; JoDee Anderson, MD; Walter Eppich, MD; Robert Simon, EdD; Jenny Rudolph, PhD; Vinay Nadkarni, MD; Mike Moyer, BS, MS; Monica Kleinman, MD; Matthew Braga, MD; Susanne Kost, MD; Glenn Stryjewski, MD; Steve Min, MD; John Podraza, MD; Joseph Lopreiato, MD; Melinda Fiedor Hamilton, MD; Jonathan Duff, MD; Jeffrey Hopkins, RN; Kimberly Stone, MD, Jennifer Reid, MD, Douglas Leonard, MD; Laura Corbin, MD; Kristine Boyle, MS; Marino Festa, MBBS; Stephanie Sudikoff, MD, Takanari Ikeyama, MD, Louis Halamek, MD; Stephen Schexnayder, MD; John Gosbee, MD; Laura Gosbee, MASc; and Matthew Richard, BSc.

REFERENCES

1. Donoghue AJ, Durbin DR, Nadel FM, Stryjewski GR, Kost SI, Nadkarni VM. Effect of high-fidelity simulation on Pediatric Advanced Life Support training in pediatric house staff: a randomized trial. *Pediatr Emerg Care* 2009;25:139–144.
2. Hunt EA, Patel S, Vera K, Shaffner DH, Pronovost PJ. Survey of pediatric resident experiences with resuscitation training and attendance at actual cardiopulmonary arrests. *Pediatr Crit Care Med* 2009;10:96–105.
3. 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;16:73–76.
4. 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:e34–e43.
5. Ali J, Al Ahmadi K, Williams JI, Cherry RA. The standardized live patient and mechanical patient models—their roles in trauma teaching. *J Trauma* 2009;66:98–102.
6. Hoadley TA. Learning advanced cardiac life support: a comparison study of the effects of low- and high-fidelity simulation. *Nurs Educ Perspect* 2009;30:91–95.
7. Owen H, Mugford B, Follows V, Plummer JL. Comparison of three simulation-based training methods for management of medical emergencies. *Resuscitation* 2006;71:204–211.
8. 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;81:331–336.
9. Overly FL, Sudikoff SN, Shapiro MJ. High-fidelity medical simulation as an assessment tool for pediatric residents' airway management skills. *Pediatr Emerg Care* 2007;23:11–15.
10. Halamek LP, Kaegi DM, Gaba DM, et al. Time for a new paradigm in pediatric medical education: teaching neonatal resuscitation in a simulated delivery room environment. *Pediatrics* 2000;106:E45.
11. 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:e597–e603.

12. Holcomb JB, Dumire RD, Crommett JW, et al. Evaluation of trauma team performance using an advanced human patient simulator for resuscitation training. *J Trauma* 2002;52:1078–1085; discussion 1085–1086.

13. Richard J, Osmond MH, Nesbitt L, Stiell IG. Management and outcomes of pediatric patients transported by emergency medical services in a Canadian prehospital system. *CJEM* 2006;8:6–12.

14. Wang HE, Abo BN, Lave JR, Yealy DM. How would minimum experience standards affect the distribution of out-of-hospital endotracheal intubations? *Ann Emerg Med* 2007;50:246–252.

15. Slonim AD, Patel KM, Ruttimann UE, Pollack MM. Cardiopulmonary resuscitation in pediatric intensive care units. *Crit Care Med* 1997;25:1951–1955.

16. Donoghue AJ, Nadkarni V, Berg RA, et al. Out-of-hospital pediatric cardiac arrest: an epidemiologic review and assessment of current knowledge. *Ann Emerg Med* 2005;46:512–522.

17. Nadkarni VM, Larkin GL, Peberdy MA, et al. First documented rhythm and clinical outcome from in-hospital cardiac arrest among children and adults. *JAMA* 2006;295:50–57.

18. Wayne DB, Didwania A, Feinglass J, Fudala MJ, Barsuk JH, McGaghie WC. Simulation-based education improves quality of care during cardiac arrest team responses at an academic teaching hospital: a case-control study. *Chest* 2008;133:56–61.

19. Ventre KM, Collingridge DS, DeCarlo D, Schwid HA. Performance of a consensus scoring algorithm for assessing pediatric advanced life support competency using a computer screen-based simulator. *Pediatr Crit Care Med* 2009;10:623–635.

20. Boulet JR, Murray DJ. Simulation-based assessment in anesthesiology: requirements for practical implementation. *Anesthesiology* 2010;112:1041–1052.

21. Barnett AG, van der Pols JC, Dobson AJ. Regression to the mean: what it is and how to deal with it. *Int J Epidemiol* 2005;34:215–220.

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.

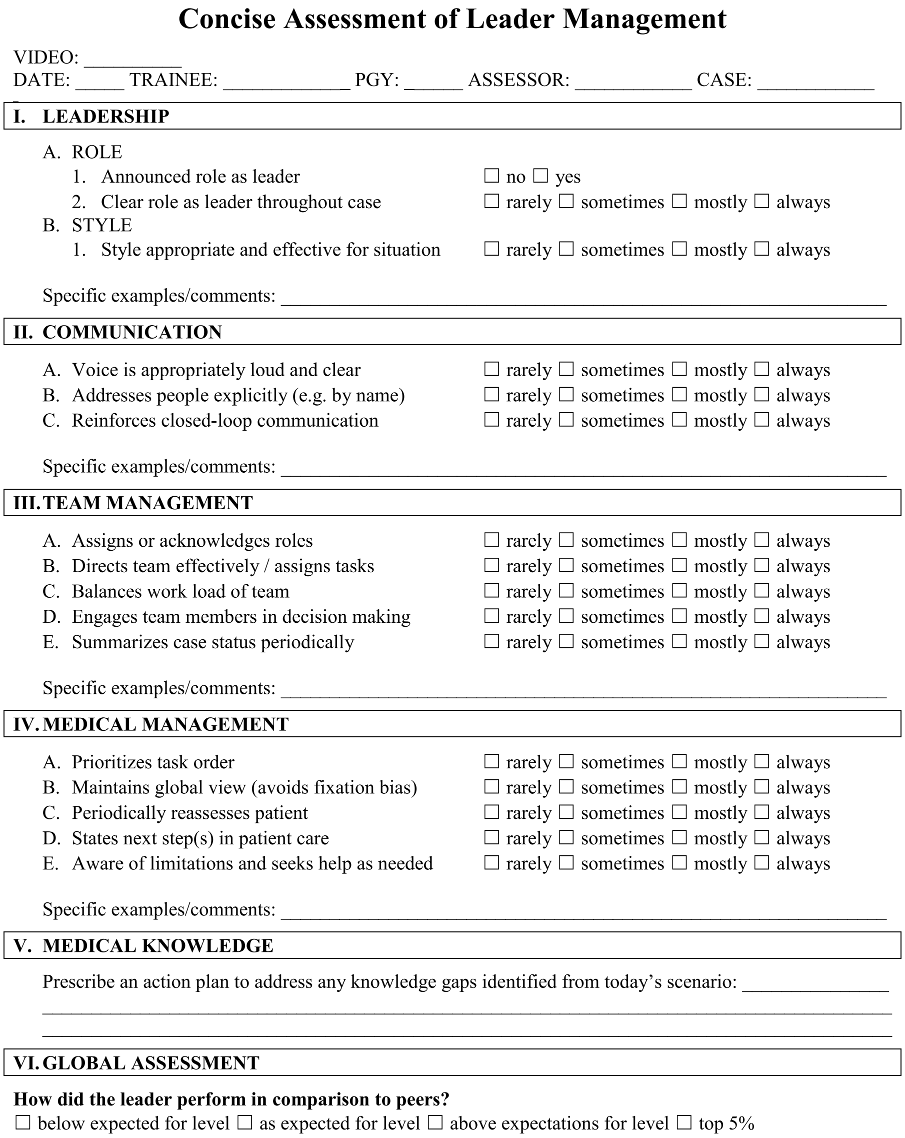
Copyright © 2017 Society for Simulation in Healthcare
DOI: 10.1097/SIH.0000000000000267

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

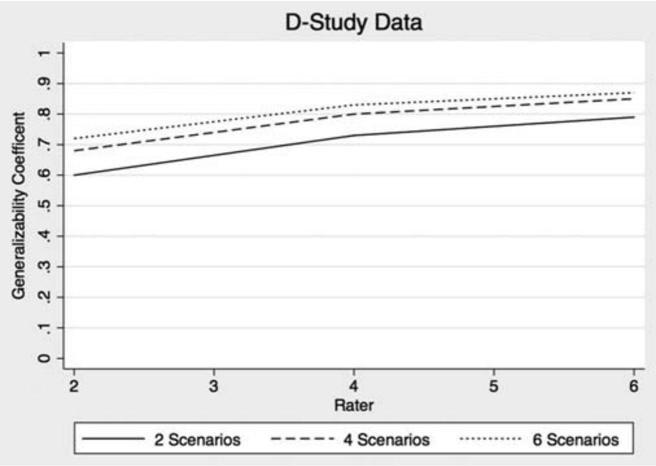


FIGURE 2. The D-study data showing the theoretical effect of changing the number of scenarios or raters in the study on the generalizability coefficient.

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.