Prospective Comparison of Live Evaluation and Video Review in the Evaluation of Operator Performance in a Pediatric Emergency Airway Simulation

JOSEPH B. HOUSE, MD
SUZANNE DOOLEY-HASH, MD
TERRY KOWALENKO, MD
ATHINA SIKAVITSAS, DO
DESIREE M. SEEYAVE, MB, BS
JOHN G. YOUNGER, MD
STANLEY J. HAMSTRA, PhD
MICHELE M. NYPAVER, MD

Abstract

Introduction Real-time assessment of operator performance during procedural simulation is a common practice that requires undivided attention by 1 or more reviewers, potentially over many repetitions of the same case.

Objective To determine whether reviewers display better interrater agreement of procedural competency when observing recorded, rather than live, performance; and to develop an assessment tool for pediatric rapid sequence intubation (pRSI).

Methods A framework of a previously established Objective Structured Assessment of Technical Skills (OSATS) tool was modified for pRSI. Emergency medicine residents (postgraduate year 1–4) were prospectively enrolled in a pRSI simulation scenario and evaluated by 2 live raters using the modified tool. Sessions were videotaped and reviewed by the same raters at least 4 months later. Raters were blinded to their initial rating.

Interrater agreement was determined by using the Krippendorff generalized concordance method.

Results Overall interrater agreement for live review was 0.75 (95% confidence interval [CI], 0.72–0.78) and for video was 0.79 (95% CI, 0.73–0.82). Live review was significantly superior to video review in only 1 of the OSATS domains (Preparation) and was equivalent in the other domains. Intrarater agreement between the live and video evaluation was very good, greater than 0.75 for all raters, with a mean of 0.81 (95% CI, 0.76–0.85).

Conclusion The modified OSATS assessment tool demonstrated some evidence of validity in discriminating among levels of resident experience and high interreviewer reliability. With this tool, intrareviewer reliability was high between live and 4-months' delayed video review of the simulated procedure, which supports feasibility of delayed video review in resident assessment.

Joseph B. House, MD, is Clinical Instructor at University of Michigan Health System Department of Emergency Medicine; **Suzanne Dooley-Hash, MD**, is Clinical Assistant Professor at University of Michigan Health System Department of Emergency Medicine; **Terry Kowalenko, MD**, is Clinical Professor at University of Michigan Health System Department of Emergency Medicine; **Athina Sikavitsas, DO**, is Assistant Professor at University of Michigan Health System Department of Emergency Medicine; **Desiree M. Seeyave, MB, BS**, is Clinical Assistant Professor at University of Michigan Health System Department of Emergency Medicine; **John G. Younger, MD**, is Professor at University of Michigan Health System Department of Emergency Medicine and Michigan Critical Injury and Illness Research Center; **Stanley J. Hamstra, PhD**, is Acting Assistant Dean, Academy for Innovation in Medical Education, and Research Director, University of Ottawa Skills and Simulation Centre, at University of Ottawa Faculty of Medicine, Ottawa, Ontario, Canada; and **Michele M. Nypaver, MD**, is Clinical Assistant Professor at University of Michigan Health System Department of Emergency Medicine.

Funding: The authors report no external funding source for this study.

Corresponding author: Joseph B. House, MD, University of Michigan Health System, Emergency Medicine, TC B1 380, 1500 East Medical Center Drive, SPC 5305, Ann Arbor, MI 48109, 734.763.4964, joshhouse@umich.edu

Received June 1, 2011; revisions received November 13, 2011 and November 29, 2011; accepted January 9, 2012.

DOI: <http://dx.doi.org/10.4300/JGME-D-11-00123.1>

Editor's note: The online version of this article contains the Objective Structured Assessment of Technical Skills for Pediatric Emergency Endotracheal Intubation tool used for assessment of simulated procedure in this study.

Background

Simulation has become an important part of competency assessment in postgraduate medical education.^{1–4} Although there is no consensus on specific tools or methodologies to reliably assess emergency skill performance, a standard approach involves expert observers evaluating provider performance in real time.^{4–7} Few instruments for pediatric emergency medicine skills have been rigorously tested for reliability or validity.^{7–10} As a consequence, evaluators use scenario-specific data collection tools, which are often institution-specific and may lack validity.

Training and coordinating a limited number of expert evaluators with many learners during traditional hours is

challenging. Both faculty availability and reviewer fatigue are significant. Video-based assessment tools may add value to detailed review of learner performance.^{11,12} No study to date has evaluated the reliability of video compared to live assessment during pediatric rapid sequence intubation (pRSI).¹³ We evaluated emergency medicine residents' performance during simulation of critical procedures to determine whether the reviewer actually needed to be present. Our specific aims were to demonstrate that (1) modification of a previously validated assessment tool provides reliable performance data in both live and video review environments and that (2) this tool is able to discriminate learner level of training.

Methods

We conducted a prospective comparison of faculty evaluations of emergency medicine resident performance during pediatric rapid sequence induction of anesthesia and intubation in live compared to video-playback formats. This educational research project was reviewed and found exempt by the University of Michigan Medical School Institutional Review Board.

Case Development

An expert panel composed of emergency medicine and pediatric emergency medicine faculty and 1 pediatric emergency medicine fellow defined the necessary features of a successful pediatric airway case involving emergent induction of general anesthesia. The case was designed to capture available published procedural guidelines.^{14,15} A scripted scenario describing a 7-month-old infant (Laerdal SimBaby, Wappingers Falls, NY) in respiratory failure was read by a facilitator and trainees were allowed 20 minutes to complete the task. The scenario also required recognition and correction of a procedural complication, inappropriate right main stem bronchus intubation. The simulated environment was prepared with equipment, materials, and resources typically found in the pediatric emergency department of our institution.

Assessment Tool

An Objective Structured Assessment of Technical Skills (OSATS) instrument¹⁶ with established validity and reliability in surgical procedural education was modified to collect performance data for pRSI (the tool is provided as online supplemental material). Seven critical content domains were identified (Time and Situational Awareness, Preparation, Medication, Preintubation Process, Intubation Technique, Time to Endotracheal Tube Placement, and Confirmation of Tube Placement) and quantified by 5-point Likert scales. An additional domain was added (Correction of Right Main Stem Intubation) to assess

What was known

The accuracy of ratings of real or simulated performance is important to judgments about, and feedback on, residents' developing clinical skills.

What is new

A modified Objective Structured Assessment of Technical Skills (OSATS) tool for pediatric emergency management showed acceptable interrater reliability and stability of ratings over time in a comparison of live observation with review of a video tape.

Limitations

Single-site, single-specialty study (emergency medicine) may limit generalizability.

Bottom line

The modified OSATS is useful as an assessment tool, demonstrating interrater reliability and some validity in discriminating performance of residents at different levels of training.

knowledge. The expert panel anchored critical actions to the Likert scale for each domain by using available evidence and where evidence was unavailable, by consensus.^{15,17} The assessment was designed by the expert panel with established guidelines to ensure content/construct validity evidence. It was determined a priori by the expert panel that competency to perform pRSI required a minimum score of 3 in all domains.

Five raters received 2 hours of instruction, including review of tool structure, content, and practice scoring of a live and videotaped pRSI session. Feedback from instructor training directed revisions to the final tool consisting of anchored performances (1 = does not know, 3 = independent with minor errors, and 5 = independent with no errors/expert) and an accompanying detailed checklist of critical performance actions assigned to numerical anchors. Raters made notes during the scenario, but completed the written tool individually at the conclusion of the scenario.

Data Collection

We prospectively enrolled 49 emergency medicine residents (12 residents each from postgraduate year [PGY]–1 through PGY-3 and 13 from PGY-4). Residents were recruited through e-mail and during didactic meetings, and also given an incentive to participate (including gift cards and educational credit). Prior simulation, pediatric intubation, and neonatal intensive care unit (NICU) experience were documented. A facilitator read and answered questions from a scripted scenario and informed trainees that 2 live raters would observe and videotape them. No more than 6 residents were evaluated during any session. Two raters assessed each participant during live and video sessions. Residents received feedback from faculty raters during a short, verbal debrief at the end of the scenario.

TABLE 1

COGNITIVE DOMAIN PERFORMANCE COMPARED TO EMERGENCY MEDICINE RESIDENT POSTGRADUATE YEAR (PGY) AND PRIOR NEONATAL INTENSIVE CARE UNIT (NICU) EXPERIENCE

Feature	Association With PGY Level		Association With Prior NICU Training		
	Spearman ρ	P Value	Before NICU Experience, Median (IQR)	After NICU Experience, Median (IQR)	P Value
Time/Situational Awareness	0.26	.08	3 (3–3.5)	3.5 (3–3.5)	.09
Preparation	0.16	.30	2.75 (2.13–3)	3 (2.5–3)	.29
Medication	−0.02	.90	3.5 (3.25–4)	3.5 (3.25–4)	.77
Preintubation Process	0.36	.02	3 (2.5–3)	3 (3–4)	.02
Intubation Technique	0.36	.01	3.5 (3.5–4)	4 (3.5–4.5)	.01
Time to ETT Placement	−0.04	.79	4.5 (3.5–5)	4.5 (3.5–5)	.99
Confirmation of ETT Placement	0.12	.42	4.5 (4–5)	4.5 (4.5–5)	.32
Time to Correction of Complication	0.07	.63	4.5 (4–5)	4.5 (4.25–5)	.65

Abbreviations: ETT, endotracheal tube; IQR, interquartile range.

Between 4 and 10 months later, the same 2 evaluators rerated participant performance by using the digital video recording of the sessions without reference to their initial scoring documents.

Data Analysis

Summary statistics were calculated in the standard fashion. We used the Krippendorff α method¹⁸ to compare the interrater agreement between faculty evaluating participants in real-time and during video review. The α statistic is analogous to the more familiar Cohen κ statistic but accommodates multiple reviewers as well as instances, such as in our design, where not every reviewer evaluates every case. With this method, statistics were generated for global interrater agreement for the real-time and video review strategy and also for each of the individual domains (eg, Time/Situational Awareness, Medication). The kripp.alpha function in the Concord statistical package for R 2.7.2 was used¹⁹ with all default settings for ordinal scoring systems. As written, the kripp.alpha routine does not generate confidence bounds around agreement estimates, thus preventing statistical comparison between the real-time and video review methods. To overcome this issue, we used bootstrap resampling of the original reviewer data to estimate confidence intervals. For each comparison to be made, 1000 bootstrap simulations were performed for each method.²⁰

For time to completion of key tasks, the simulated case was segmented into 3 intervals: initiation of case to placement of the laryngoscope blade, placement of the laryngoscope blade to first successful breath, and first

successful breath to correction of complication (right main stem bronchial intubation). Each of these intervals, as well as the time needed to complete the case, were analyzed via Cox proportional hazard modeling with predictors such as participant's postgraduate year and completion of a 1-month rotation in the NICU. The Spearman method also was used to correlate rated performance with postgraduate year and with the time to critical events.

Results

Of the 49 residents who completed the simulation, 46 participated in the study. Three residents were excluded from video review because of poor quality of audio recordings.

Two domains of the modified OSATS instrument (Preintubation Process and Intubation Technique) were able to discriminate trainees by level of training and neonatal experience (TABLE 1). Time to performance of predetermined tasks did not show a significant difference by postgraduate year or as a function of prior NICU experience (TABLE 2).

Interrater agreement of learner performance as assessed by postsimulation video review (0.79; 95% CI, 0.73–0.82) produced concordant scores as those achieved by live evaluation (0.75; 95% CI, 0.72–0.78). Intrarater agreement between the live evaluation and later video evaluation was greater than 0.75 for all raters with a mean of 0.81 (95% CI, 0.76–0.85). Raters agreed significantly more often in live format for Preparation (TABLE 3). There was no significant difference for the remaining domains in either review format.

TABLE 2 TIME TO COMPLETION OF TASKS IN MINUTES (25TH PERCENTILE, 75TH PERCENTILE)

	Median Time t ₁ ^a	Median Time t ₂ ^b	Median Time t ₃ ^c
PGY-1	6 (4.9, 7.2)	0.5 (0.4, 0.7)	1 (0.5, 1.4)
PGY-2	6 (5.1, 6.7)	0.5 (0.4, 0.8)	0.8 (0.6, 0.9)
PGY-3	6.4 (5.0, 6.9)	0.6 (0.5, 0.8)	0.7 (0.5, 1.2)
PGY-4	5 (4.6, 6.52)	0.5 (0.4, 0.63)	0.5 (0.4, 1.1)
Pre NICU	6.2 (5.2, 7.0)	0.5 (0.4, 0.6)	0.8 (0.5, 1.2)
Post NICU	5.6 (4.6, 6.7)	0.6 (0.4, 0.7)	0.7 (0.4, 1.1)

Abbreviations: NICU, neonatal intensive care unit; PGY, postgraduate year.

^at₁ = Time from start of scenario to blade in mouth.

^bt₂ = Blade in mouth to first successful breath.

^ct₃ = First successful breath to correction of complication.

Discussion

Our study demonstrated 2 main findings. First, video evaluation of procedural simulation provided reviewer performance at least as reliable as live review and has substantial potential in simplifying OSATS use when evaluation of many operators is anticipated. Second, the modified OSATS (for pRSI) instrument demonstrated both intrarater and interrater reliability in the assessment of emergency medicine residents' procedural skill in pRSI, supporting its use as an effective evaluation tool.

Retrospective video review of emergency medicine resident performance of pRSI produced assessment metrics (eg, agreement between reviewers of resident performance) of equal quality to traditional real-time assessment. These findings support an evaluative model of brief, limited face-to-face debriefing immediately after the procedure and options for more detailed asynchronous evaluation of resident performance, using video capture. Debriefing by

an expert at the time of event is necessary to deliver immediate feedback to residents. However, delayed detailed review by faculty allows opportunities for more specific feedback to learners with time for discussion and self-assessment by the learner. Video review of scenarios also frees raters to evaluate residents on demand, and reduces the number of evaluators needed.

The second finding of the study concerns the development of the OSATS tool. We found that the Intubation Process and Intubation Technique domains were able to differentiate the residents by both postgraduate year and prior NICU experience. The content of these 2 domains required demonstration of technical ability and highlighted key differences between an adult and pediatric airway. These skills were thought to reflect the development of expertise in complex skills and are similar to previous reports using the OSATS to assess surgical procedural technique. The modified pRSI OSATS tool was unable to

TABLE 3 COMPARISON OF INTERRATER AGREEMENT PERFORMANCE OF LIVE AND VIDEO-BASED OBSERVATIONS^a

Feature	Live Observation	Video Observation
Time/Situational Awareness	0.19 (−0.8, 0.46)	0.07 (−0.24, 0.36)
Preparation	0.46 (0.24, 0.69)	0.13 (−0.20, 0.43) ^b
Medication	0.26 (−0.01, 0.52)	0.30 (−0.01, 0.58)
Preintubation Process	0.64 (0.46, 0.78)	0.48 (0.21, 0.68)
Intubation Technique	0.49 (0.22, 0.68)	0.59 (0.36, 0.76)
Time to Endotracheal Tube Placement	0.65 (0.41, 0.82)	0.70 (0.44, 0.87)
Confirmation of Endotracheal Tube Placement	0.40 (0.14, 0.64)	0.51 (0.22, 0.72)
Time to Complication Correction	0.62 (0.40, 0.79)	0.67 (0.38, 0.88)

^a By Krippendorff concordance method, which is bounded by −1 (perfect disagreement) to +1 (perfect agreement).

^b Live is statistically better, as 95% CI for difference does not include 0.

differentiate between training levels for the other domains. All participants performed well or poorly in some domains (TABLE 1) and there was not significant variability. For example, participants were only scored highly (5) if they knew doses and equipment without references, such as the Broselow Tape, a reference typically available to our trainees in their emergency department practice.²¹ For the most part, emergency medicine resident participants were not confident in their ability to determine pediatric drug doses and endotracheal tube size without resources, regardless of their level of training. This may be due to gaps in the training curriculum or need for better discriminators in the scoring definitions of the pRSI OSATS tool.

Time to completion of predetermined tasks was unable to demonstrate differences between junior and senior residents (TABLE 3). While success in pediatric emergency airway management is considered time sensitive, the relationship between time and competency has not been established. Decreased time to completion of a procedure may be due to missed procedural steps that may or may not indicate expert skill.²² Our findings may have resulted from a scenario not sufficiently difficult to capture subtle time differences in task completion and/or mechanical limitations of the mannequin model.

This study is a single institutional study, which limits its validity in other settings. At this institution, intubation of patients in the emergency department is not video recorded, thus limiting our ability to assess learners' performance in their natural setting. Technology may have also limited raters' ability to detect subtle movements of the laryngoscope blade, despite overhead microphones and 2 video cameras. Suggested improvements by raters included a standard camera on the intubation blade and/or a camera attached to the head of the trainee. Sound was determined to be very important to the success of understanding the communication surrounding the trainees as they completed the pRSI task; 3 sessions were excluded from video evaluation owing to problems with sound capture. Personal microphones may reduce background noise and improve raters' ability to evaluate residents.

Conclusion

We demonstrated that a modified OSATS instrument for pediatric emergency airway management proved to be reliable as measured by raters during live and video methodologies. Further research is needed to improve discriminators for level of trainee experience. A procedural assessment model using 1 facilitator and delayed video review demonstrated reliable capture of procedural performance data as measured by consistent interrater and intrarater reliability compared to live assessment.

References

- Johnson L, Patterson MD. Simulation education in emergency medical services for children. *Clin Pediatr Emerg Med*. 2006;7(2):121–127.
- Adler MD, Trainor JL, Siddall VJ, McGaghie WC. Development and evaluation of high-fidelity simulation case scenarios for pediatric resident education. *Ambul Pediatr*. 2007;7(2):182–186.
- Hunt EA, Heine M, Hohenhaus SM, Luo X, Frush KS. Simulated pediatric trauma team management: assessment of an educational intervention. *Pediatr Emerg Care*. 2007;23(11):796–804.
- 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(1):11–15.
- Shayne P, Gallahue F, Rinnert S, Anderson CL, Hern G, Katz E. Reliability of a core competency checklist assessment in the emergency department: the Standardized Direct Observation Assessment Tool. *Acad Emerg Med*. 2006;13(7):727–732.
- LaMantia J, Kane B, Yarris L, Tados A, Ward MF, Lesser M, et al.; SDOT Study Group II. Real-time inter-rater reliability of the Council of Emergency Medicine residency directors standardized direct observation assessment tool. *Acad Emerg Med*. 2009;16(suppl 2):S51–S57.
- Kogan JR, Holmboe ES, Hauer KE. Tools for direct observation and assessment of clinical skills of medical trainees: a systematic review. *JAMA*. 2009;302(12):1316–1326.
- Brett-Fleegler MB, Vinci RJ, Weiner DL, Harris SK, Shih M-C, Kleinman ME. A simulator-based tool that assesses pediatric resident resuscitation competency. *Pediatrics*. 2008;121(3):e597–e603.
- Chipman JG, Schmitz CC. Using objective structured assessment of technical skills to evaluate a basic skills simulation curriculum for first-year surgical residents. *J Am Coll Surg*. 2009;209(3):364–370.e362.
- Adler MD, Vozenilek JA, Trainor JL, Eppich WJ, Wang EE, Beaumont JL, et al. Comparison of checklist and anchored global rating instruments for performance rating of simulated pediatric emergencies. *Simul Healthc*. 2011;6(1):18–24.
- Dath D, Regehr G, Birch D, Schlachta C, Poulin E, Mamazza J, et al. Toward reliable operative assessment: the reliability and feasibility of videotaped assessment of laparoscopic technical skills. *Surg Endosc*. 2004;18(12):1800–1804.
- Laeq K, Infusino S, Lin SY, Reh DD, Ishii M, Kim J, et al. Video-based assessment of operative competency in endoscopic sinus surgery. *Am J Rhinol Allergy*. 2010;24(3):234–237.
- Williams JB, McDonough MA, Hilliard MW, Williams AL, Cuniowski PC, Gonzalez MG. Intermethod reliability of real-time versus delayed videotaped evaluation of a high-fidelity medical simulation septic shock scenario. *Acad Emerg Med*. 2009;16(9):887–893.
- McGill J, Clinton J. Tracheal intubation. In: Roberts JR, Hedges JR, Chanmugam AS, eds. *Clinical Procedures in Emergency Medicine*, 4th ed. Philadelphia, PA: W.B. Saunders; 2004:69–83.
- Tschudy M, Arcara K. Emergency management. In: Gunn, VL, Barone, MA, John Hopkins Hospital. Children's Medical and Surgical Center. *The Harriet Lane Handbook: A Manual for Pediatric House Officers*, 16th ed. Philadelphia, PA: Mosby; 2002:3–7.
- Reznick R, Regehr G, MacRae H, Martin J, McCulloch W. Testing technical skill via an innovative "bench station" examination. *Am J Surg*. 1997;173(3):226–230.
- Ludwig S, Lavelle J. Resuscitation—pediatric basic and advanced life support. In: Fleisher GR, Ludwig S, eds. *Textbook of Pediatric Emergency Medicine*. 6th ed. Philadelphia, PA: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2010:1–31.
- Krippendorff K. *Content Analysis: An Introduction to Its Methodology*. 2nd ed. Thousand Oaks, CA: SAGE; 2003.
- Lemon J, Fellows I. concord: concordance and reliability, 2007. In: R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2008.
- Efron B, Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat Sci*. 1986;1(1):54–75.
- Luten RC, Wears RL, Broselow J, Zaritsky A, Barnett TM, Lee T, et al. Length-based endotracheal tube and emergency equipment in pediatrics. *Ann Emerg Med*. 1992;21(8):900–904.
- Grober ED, Hamstra SJ, Wanzel KR, Reznick RK, Matsumoto ED, Sidhu RS, et al. The educational impact of bench model fidelity on the acquisition of technical skill: the use of clinically relevant outcome measures. *Ann Surg*. 2004;240(2):374–381.