Published in final edited form as:

Simul Healthc. 2014 October ; 9(5): 295–303. doi:10.1097/SIH.0000000000000048.

Validity and Reliability Assessment of Detailed Scoring Checklists for Use During Perioperative Emergency Simulation Training

Matthew D. McEvoy, MD,

Vice-Chair for Educational Affairs and Program Director Associate Professor Department of Anesthesiology Vanderbilt University Medical Center 2301 VUH Nashville, TN 37232
matthew.d.mcevoy@vanderbilt.edu

William R. Hand, MD,

Assistant Professor Department of Anesthesia and Perioperative Medicine Medical University of South Carolina

Cory M. Furse, MD, MPH,

Assistant Professor Department of Anesthesia and Perioperative Medicine Medical University of South Carolina furse@musc.edu

Larry C. Field, MD,

Associate Professor Department of Anesthesia and Perioperative Medicine Medical University of South Carolina field@musc.edu

Carlee A. Clark, MD,

Assistant Professor Department of Anesthesia and Perioperative Medicine Medical University of South Carolina clarca@musc.edu

Vivek K. Moitra, MD,

Clinical Associate Professor Department of Anesthesiology Columbia University Medical Center vm2161@columbia.edu

Paul J. Nietert, PhD,

Professor Department of Public Health Sciences Medical University of South Carolina 135 Cannon Street, Room 302C Charleston, SC 29425 nieterpj@musc.edu

Michael F. O'Connor, MD, and

Professor of Anesthesia & Critical Care Chief, Section of Critical Care Medicine Department of Anesthesia & Critical Care University of Chicago moc5@dacc.uchicago.edu

Mark E. Nunnally, MD

Associate Professor of Anesthesia & Critical Care Department of Anesthesia & Critical Care University of Chicago mnunnally@dacc.uchicago.edu

Abstract

(corresponding author)handw@musc.edu.

Introduction—Few valid and reliable grading checklists have been published for evaluation of performance during simulated high stakes perioperative event management. As such, the purpose of this study was to construct valid scoring checklists for a variety of perioperative emergencies and determine the reliability of scores produced by these checklists during continuous video review.

Methods—A group of anesthesiologists, intensivists, and educators created a set of simulation grading checklists for assessment of the following scenarios: severe anaphylaxis (ANAPH), cerebrovascular accident (CVA), hyperkalemic arrest (HYPERK), malignant hyperthermia (MH), and acute coronary syndrome (ACS). Checklist items were coded as critical or non-critical. Non-expert raters evaluated 10 simulation videos in a random order, with each video being graded four times. A group of faculty experts also graded the videos in order to create a reference standard to which non-expert ratings were compared. $P < 0.05$ was considered significant.

Results—Team leaders in the simulation videos were scored by the expert panel as having performed 56.5% of all items on the checklist (range = 43.8% to 84.0%), and 67.2% of critical items (range = 30.0% to 100%). Non-expert raters agreed with the expert assessment 89.6% of the time (95% CI: 87.2% to 91.6%). No learning curve development was found with repetitive video assessment or checklist use. The kappa values comparing non-expert rater assessments to the reference standard averaged 0.76 (95% CI: 0.71 to 0.81).

Conclusions—The findings indicate the grading checklists described are valid, reliable, and could be used in perioperative crisis management assessment.

INTRODUCTION

As one aspect of evaluating competency in medical training, educators strive to assess learner performance during the management of high stakes events with valid, reliable, defensible, and generalizable standards.¹ Similar to evaluating the ability of trainees to adhere to well-defined and generalizable treatment algorithms, such as advanced cardiac life support (ACLS), there is literature supporting validated and generalizable management algorithms encountered in the perioperative period.²⁻⁸ Although rare, examples of specific crises that every anesthesiologist should be able to treat include malignant hyperthermia, local anesthetic systemic toxicity, anaphylaxis, severe hyperkalemia, the management of a difficult airway, and many more. Management of perioperative emergencies requires practitioners to recall and apply diagnostic and therapeutic algorithms for causes beyond cardiac instability, including respiratory, neurologic, metabolic/endocrine, and toxin-mediated etiologies.⁵ When evaluating team and participant performance during simulations of such high stakes events, grading checklists that produce valid and reliable scores are necessary.

This is especially true if these grading checklists are ever to be used in high-stakes tests or certifications.⁵ Experts in simulation research and psychometric analysis have described testing of checklist formation and evaluation in an effort to validate a checklist independently.^{9,10} Several groups, including the authors of this research, have shown that checklists can be developed that produce reliable scores during simulation-based assessment of trainee performance.¹¹⁻¹⁴ However, many published studies that assess participant

performance in perioperative emergencies use raters who were involved in the checklist creation process.¹⁵⁻¹⁹ In order to show generalizability of any assessment tool, studies need to prove that reliable scores can be generated when evaluators who were not involved with checklist design use them.

The purpose of this study was to construct valid scoring checklists for a variety of perioperative emergencies and determine the reliability of scores produced by these checklists by non-expert raters. The authors hypothesize that if the checklists are reliable and generalizable, then it is postulated that trained, non-expert raters would have a high degree of inter- and intra-rater reliability, compared to expert raters. Additionally, the research evaluate the ability to use these checklists under different circumstances (i.e. while reviewing performances with and without the ability to pause). The authors used simulation videos with both excellent and poor team leader performances in hopes of demonstrating that checklist reliability was unchanged by participant performance.

METHODS

After this study was granted approval by the IRB (Medical University of South Carolina IRB-II), a group of experienced anesthesiologists, intensivists, and educators was constructed in order to oversee the creation of the grading checklists. The group contained individuals who have been involved with simulation education, training, and research for a minimum of three years and who teach crisis management simulation courses (e.g. ACLS, Anesthesia Crisis Resource Management, Maintenance of Certification in Anesthesiology), at least five times per year (M.D.M., W.R.H., L.C.F., C.M.F.). Additionally, the group contained experts in critical care medicine, resuscitation, Anesthesia-Advanced Cardiac Life Support (A-ACLS) content creation, and education (M.E.N., M.F.O., C.A.C., V.K.M.). The authors previously published work validating a scoring checklist for ACLS, and this new work relies heavily on the previously published methodology to provide additional validated checklists.¹⁴

Checklist and Video Creation

The methodology for the creation of the grading checklists and videos has been previously described by the authors.¹⁴ The major steps included: developing valid content of the checklists by literature review, determining the exact form of each item on the checklist using a modified Delphi technique, and organizing the items into chronological order and portions (e.g. initial evaluation). Each item was categorized as correct/incorrect, objective/subjective, and critical/non-critical. For example, administration of Dantrolene would be a correct, objective, and critical step for management of MH. The differences in this study, compared to the methods published previously, was the use of additional content experts during the validation process and that the videos used for checklist analysis were from a previously recorded bank of videos with resident participants as team leaders. The residents in these videos were unaware of the checklist analysis study and thus gave native, unscripted performances that ranged from expert to poor performance. Five checklists were formed for use during grading of the scenarios: anaphylaxis (ANAPH), cerebrovascular accident

(CVA), hyperkalemic arrest (HYPERK), malignant hyperthermia (MH), and acute coronary syndrome (ACS) (See Appendices A-E).

Four experienced raters (M.D.M., W.R.H., L.C.F, and C.M.F.) graded the reference videos of outstanding and poor performance using the checklists. This was done individually and then together in order to have a reference standard performance rating to which the evaluations of the non-expert raters could be compared, and this methodology is consistent with prior studies.^{14,16,17} Individual scores of the scenarios by the experienced faculty members were very similar. For any areas of discrepancy, which occurred on <3% of items, agreement was reached by watching the videos together with discussion of items of discrepancy and final agreement upon a single answer.

Training of Non-expert Raters

The non-expert raters in this study were four medical students with no prior experience in checklist creation or simulation-based assessment (M.D.P., G.L.K., S.K.G., L.A.A.). The students had previous clinical exposure in anesthesiology and participated in perioperative emergency simulations as trained confederates. During a single four-hour training session each of the raters was trained in a stepwise manner. They were all oriented to the software interface in which the grading was performed (SimMan 3G; Laerdal Medical Corp, Stavanger, Norway). Next, checklists were reviewed item-by-item ensuring the four non-expert raters understood what each item referenced. Next raters were given instructions on how to grade items classified as subjective. Finally a sample video was shown and the two viewing modes were demonstrated—continuous and with pauses. Any remaining questions were answered as a group discussion.

Checklist Evaluation

The four non-expert raters were trained on use of the checklists and evaluated each video (10 total) a total of four times in a random order. Each video was viewed twice allowing the rater to stop the video (“with pauses”) and twice without the ability to pause during grading (“continuous”), consistent with the manner our group has previously described.¹⁴ Appendix F provides an example of the sequence of videos graded and the method of viewing for each grader.

Simulation Scenarios

Detailed descriptions of the scenarios used can be found online (see document, Supplemental Digital Content 1, which contains descriptions of the scenarios and simulation set up). A brief description of each scenario follows:

1. Acute Coronary Syndrome: Patient is 58 y/o M awaiting urgent repair of a femur fracture. He has history of hypertension and hyperlipidemia and broke his femur falling from a ladder. He complained of shortness of breath upon arrival to preoperative holding bay. ECG will show anterior STEMI, subtle improvement with beta blockade and nitroglycerine. Confederate will serve as interventional cardiologist if consulted/called.

2. Anaphylaxis: 48 y/o male undergoing abdominal surgery becomes unstable immediately following administration of antibiotic. Presentation includes: hypotension, tachycardia, increased airway pressures, wheezes, and rash (photo to display). Patient will require epinephrine bolus, antigen removal, steroids, and epinephrine infusion to stabilize.
3. Cerebrovascular Accident: 68 y/o female underwent left total hip arthroplasty under epidural. Encounter occurs in postoperative setting with onset of altered mental status, Cheyne Stokes respiratory pattern and physical exam findings consistent with left MCA stroke. Management requires mobilization of neurology, intubation, and consideration of thrombolytics (discussed with neurology).
4. Hyperkalemic Arrest: Patient is 65 y/o male with diabetes, hypertension, coronary artery disease and chronic renal insufficiency undergoing left total knee replacement. Upon release of the tourniquet from the operative leg, the patient will demonstrate instability and ECG changes consistent with hyperkalemia. Based on management and timing, patient may become pulseless and require ACLS in addition to treatment of hyperkalemia. Critical steps include: patient assessment, ECG monitoring, administration of calcium, adherence to pulseless treatment algorithm (if pulseless state is reached).
5. Malignant Hyperthermia: Patient is 30 y/o M undergoing laparoscopic appendectomy. Participant has been called into the OR to give a break. The end tidal carbon dioxide and temperature are rising. Participant is to recognize signs of malignant hyperthermia. Critical Steps include: order ABG, administer adequate dose of Dantrolene (2.5mg/kg), hyperventilation, fluid bolus, and initiation of cooling measures.

Statistical Analysis

Using a previously described model, agreement of assessments between the raters and the experts was assessed using a generalized linear mixed model (GLMM) approach.^{14,20} The unit of analysis was each rater's individual checklist item assessment. The model allowed us to calculate agreement rates and their accompanying 95% confidence intervals, stratified for the levels of the various independent variables, while adjusting for all the other factors included in the model. All analyses were conducted using SAS v9.3 (Cary, NC) PROC GLIMMIX, and p-values < 0.05 were considered statistically significant.

RESULTS

In the ten videos used for this study, the team leaders in the simulation videos were scored by the expert panel as having performed 56.5% of all items on the checklist (range = 43.8% to 84.0%), and 67.2% of critical items (range = 30.0% to 100%). These team leader performance scores display variability that may be attributed to team leader skill or scenario difficulty. Despite absolute score variability, results of the GLMM indicate that non-expert raters agreed with the expert assessment 89.6% of the time (95% CI: 87.2% to 91.6%). Table 1 shows the results of the primary GLMM model evaluating independent associations between experimental conditions (factors) with agreement/disagreement to the reference

standard on individual checklist items. Agreement with the reference standard varied from 84.2% to 94.7%, with the lowest average level of agreement (84.2%) being within Hyperkalemic Arrest 1 and the highest agreement being within MH 1 (94.7%). Agreement was, on average, higher among items deemed objective as compared to subjective (91.1% vs. 86.3%, $p=0.001$), and higher among items considered by the expert panel not to have been performed compared to items considered to have been performed (92.4% vs. 87.1%, $p<0.0001$). No significant differences were noted when videos were viewed “continuous” versus “with pauses” (88.8% vs. 90.3%, $p=0.15$), and no association was noted between agreement and viewing order ($p=0.99$). Additionally, agreement on items deemed to be critical was not significantly different compared with items considered to be non-critical (88.4% vs. 91.1%, $p=0.22$).

Examining the presence of any potential rater learning curve development, we found that agreement between the raters and reference standard did not significantly change over time. Figure 1 demonstrates the lack of association between rater and round of grading or method of grading (i.e. continuous v. with pauses). Detailed breakdowns of the frequency of agreement between raters and the reference standard are shown in Table 1. The kappa values comparing non-expert rater assessments to the reference standard averaged 0.76 (95% CI: 0.71 to 0.81), representing excellent inter-rater agreement according to Fleiss's criteria.²⁰ Among critical items, kappa values averaged 0.70 (95% CI: 0.62 to 0.78), which was slightly lower than among non-critical items, where the kappa values averaged 0.77 (95% CI: 0.73 to 0.82). Overall, these findings illustrate the high level of inter-observer reliability was not affected by the scoring mode (continuous or with pauses) or item “type”. They corroborate that repeated evaluation of the same video did not significantly improve the level of agreement with the reference standard.

DISCUSSION

This study demonstrates that scores generated from a set of grading checklists used by non-expert raters have a high degree of reliability that is not affected by the grading technique. The findings from this study concerning performance assessment involving perioperative emergencies are important for several reasons.

The content contained in these assessment tools is relevant for the anesthesiologist in training. The Accreditation Council for Graduate Medical Education (ACGME) has announced the Next Accreditation System (NAS) as a significant update for competency-based assessment in GME.²¹ The Milestones will be the specific metrics by which the six core competencies of Patient Care (PC), Medical Knowledge (MK), Professionalism (Pro), Interpersonal and Communication Skills (ICS), Practice-Based Learning and Improvement (PBLI), and Systems-Based Practice (SBP) will be evaluated. The number of Milestones varies within each specialty and each competency, but the current proposal for anesthesiology GME involves a total of 25 separate Milestones that need to be evaluated on each resident each year. A large number of these can be evaluated through simulation-based assessment.²² In specific, there are six Milestones (PC5, PC6, PC8, ICS3, PBLI2, and PBLI3) that could be assessed through the high-fidelity simulation of clinical crises. Our group has recently published a reliable set of checklists for assessing performance during the

management of simulated in-hospital cardiac arrest in the non-operative setting (Milestone PC6).¹⁴ However, our specialty needs further valid and reliable assessment tools that can be used for generalizable standard setting if we are to adequately assess performance during crisis management specific to the perioperative arena. The current study reports on one set of grading checklists that produces reliable scores from performances at various levels in the spectrum of learning levels delineated by the ACGME (i.e. entry, junior, mid, senior, and advanced levels).

Additionally, the findings of this study are important with respect to training and assessment requirements of the ACGME Milestones Project. The ACGME recommends that simulation be employed in all anesthesiology residency-training programs.²² Since the assessment and debriefing associated with such exercises is of benefit to the learner, the ability of faculty to use reliable checklists without the need to pause and rewind is important. But, while the use of simulation training is a requirement, very few programs have faculty who are education experts trained in checklist design and evaluation. Furthermore, with clinical demands constantly pressing upon non-clinical duties in the academic medical center, the time that it takes to orient faculty to a new teaching and evaluation method is of paramount importance, as is the amount of time required to complete learner assessments. Assessment tools, such as those described in this study that can be learned in a short time period and can be reliably employed during continuous review of performances, are of value. The fact that these can be employed by personnel without educational expertise after a short orientation to the assessment methodology would suggest that these findings are generalizable to programs beyond the one in which they were tested, but future research needs to investigate this. The short training curriculum (~4 hours) given to non-expert raters in this study was sufficient to orient them to the use of the checklists, as evidenced by the absence of improvement in reliability scores over time (i.e. lack of a learning curve; see Figure 1). Overall, having clearly defined assessment criteria is of benefit to the learner and is of benefit for tracking trainee performance over time as they progress through the levels described in the ACGME Milestones.*

Beyond this use for the anesthesiologist in training, such grading checklists could benefit practicing anesthesiologists in formative assessment that occurs during the Maintenance of Certification in Anesthesiology (MOCA) Part 4 (Simulation) program. Several studies report that the performance of anesthesiologists during perioperative crisis management is suboptimal in both the simulation and clinical settings.^{23,24} In either setting, giving practicing anesthesiologists objective feedback about their performance that is based on highly reliable assessment tools is of great value for ongoing practice performance assessment and lifelong learning. Similar developments of assessment tools for perioperative emergency scenarios will make it possible to both teach and evaluate based on the disseminated learning materials.

While the results of this study advance upon prior knowledge in this domain, there are several limitations to note. First, this study was conducted at one institution and therefore

*Milestones for Anesthesiology will be released in March 2014. A draft of the Milestones was sent to all program directors and is expected to be similar to the final. An updated reference can be provided after this publication occurs.

whether these assessment tools would provide similar results in widespread use is not known. Second, this study did not define a minimum passing score for these scenarios or scores related to various levels of training. This is a separate issue related to performance, but needs to be undertaken in future research. Third, these checklists did not reliably document incorrect actions. Checklists by their nature cannot account for the innumerable potential treatment missteps. Similarly, because the authors could not conceive of a reliable way to determine how much temporal or sequential delay would make a checklist item change from done to “not done”, completion of a checklist item at any time during the simulation would result in credit for that item. There are some clinical scenarios that lend themselves to this type of evaluation (e.g. ventricular fibrillation requiring immediate defibrillation or perhaps even local anesthetic toxicity and administration of lipid emulsion), however that was not the case for the scenarios and treatment steps tested. Fourth, the reliability of the checklist has been studied only in terms of team leadership, rather than global team performance. Recent studies have investigated team performance and these checklists may need to be reassessed for reliability if used in a different manner, such as assessing an entire team instead of one person.^{16,17} Fifth, the perioperative emergency checklists used in this study were compiled from various sources and certainly will require updating as crisis management evolves with medical innovation. Thus, reassessment of reliability of these checklists as grading tools may be needed if/when the content of the checklist changes. Finally, using non-expert raters who had experience within the simulation experience and were motivated to perform well (mentees of authors) may have elevated performance. This strong performance may have caused something of a ceiling effect where good baseline performances were difficult to improve upon over time, thus erroneously allowing the authors to conclude there was not a significant learning curve over time.

Finally, these checklists were used in conjunction with specific scenarios. As the scenarios were constructed to progress in a certain fashion with respect to physiology and patient presentation, it is uncertain whether these assessment tools would yield reliable scores if used in conjunction with another clinical scenario of the same underlying disease, despite their content and internal structural validity. As such, the construct validity of the scenario that we presented as the test itself may have an important interaction with the grading checklist. It may be important to use a valid scenario presentation (scenario, mannequin, environment) that is paired with a particular checklist in order to obtain high reliability in the scores that are produced, but this is not currently known.

CONCLUSION

The findings of this study indicate that the set of grading checklists can reliably be used to assess simulated management of these clinical scenarios. This study produced reliability and content validity evidence for this new set of checklists, widening the scope of scenario-specific checklists for use in high-fidelity simulation training concerning perioperative crises. Future research needs to expand the set of grading tools and validated scenarios to be in line with the scope of perioperative emergencies that anesthesiologists must be competent to manage. The scenarios used for checklist validation in this research will provide a tool for future research on a wider scope of clinical scenarios, including several crises described in A-ACLS.⁵

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors wish to thank Matthew Parks, Samkon Gado, Louise Alexander, and Garrick Klaybor for their work in the performance of this study.

Funding:

Foundation for Anesthesia Education and Research (FAER), Research in Education Grant (PI: McEvoy) provided funded research time. FAER was not involved in the study design or data analysis.

South Carolina Clinical & Translational Research Institute, Medical University of South Carolina's CTSA, NIH/NCATS Grant Number UL1 TR000062 provided biostatistical resources.

FINANCIAL DISCLOSURES SUMMARY

The authors have no relevant conflict of interests related to this research, medical simulation, or the checklists themselves. The research was supported by Foundation for Anesthesia Education and Research (FAER), Research in Education Grant (PI: McEvoy) and South Carolina Clinical & Translational Research Institute, Medical University of South Carolina's CTSA, NIH/NCRR Grant Number UL1 TR000062. FAER and SCTRI were not involved in the study design or data analysis.

REFERENCES

- Wayne DB, Butter J, Cohen ER, McGaghie WC. Setting defensible standards for cardiac auscultation skills in medical students. *Academic medicine : journal of the Association of American Medical Colleges*. Oct; 2009 84(10 Suppl):S94–96. [PubMed: 19907398]
- Wayne DB, Siddall VJ, Butter J, et al. A longitudinal study of internal medicine residents' retention of advanced cardiac life support skills. *Academic medicine : journal of the Association of American Medical Colleges*. Oct; 2006 81(10 Suppl):S9–S12. [PubMed: 17001145]
- Wayne DB, Butter J, Siddall VJ, et al. Mastery learning of advanced cardiac life support skills by internal medicine residents using simulation technology and deliberate practice. *Journal of general internal medicine*. Mar; 2006 21(3):251–256. [PubMed: 16637824]
- Wayne DB, Butter J, Siddall VJ, et al. Graduating internal medicine residents' self-assessment and performance of advanced cardiac life support skills. *Medical teacher*. Jun; 2006 28(4):365–369. [PubMed: 16807178]
- Moitra VK, Gabrielli A, Maccioli GA, O'Connor MF. Anesthesia advanced circulatory life support. *Canadian journal of anaesthesia = Journal canadien d'anesthesie*. Jun; 2012 59(6):586–603.
- Neal JM, Mulroy MF, Weinberg GL. American Society of Regional A, Pain M. American Society of Regional Anesthesia and Pain Medicine checklist for managing local anesthetic systemic toxicity: 2012 version. *Regional anesthesia and pain medicine*. Jan-Feb;2012 37(1):16–18. [PubMed: 22189574]
- Larach MG, Gronert GA, Allen GC, Brandom BW, Lehman EB. Clinical presentation, treatment, and complications of malignant hyperthermia in North America from 1987 to 2006. *Anesthesia and analgesia*. Feb 1; 2010 110(2):498–507. [PubMed: 20081135]
- Bossaert L, O'Connor RE, Arntz HR, et al. Part 9: Acute coronary syndromes: 2010 International Consensus on Cardiopulmonary Resuscitation and Emergency Cardiovascular Care Science with Treatment Recommendations. *Resuscitation*. Oct; 2010 81(Suppl 1):e175–212. [PubMed: 20959169]
- Muller MJ, Dragicevic A. Standardized rater training for the Hamilton Depression Rating Scale (HAMD-17) in psychiatric novices. *Journal of affective disorders*. Oct; 2003 77(1):65–69. [PubMed: 14550936]

10. Sevdalis N, Lyons M, Healey AN, Undre S, Darzi A, Vincent CA. Observational teamwork assessment for surgery: construct validation with expert versus novice raters. *Annals of surgery*. Jun; 2009 249(6):1047–1051. [PubMed: 19474694]
11. Weidner AC, Gimpel JR, Boulet JR, Solomon M. Using standardized patients to assess the communication skills of graduating physicians for the comprehensive osteopathic medical licensing examination (COMLEX) level 2-performance evaluation (level 2-PE). *Teaching and learning in medicine*. Jan; 2010 22(1):8–15. [PubMed: 20391277]
12. Schmitz CC, Chipman JG, Luxenberg MG, Beilman GJ. Professionalism and communication in the intensive care unit: reliability and validity of a simulated family conference. *Simulation in healthcare : journal of the Society for Simulation in Healthcare*. 2008; 3(4):224–238. Winter. [PubMed: 19088667]
13. Zanetti M, Keller L, Mazor K, et al. Using standardized patients to assess professionalism: a generalizability study. *Teaching and learning in medicine*. Oct; 2010 22(4):274–279. [PubMed: 20936574]
14. McEvoy MD, Smalley JC, Nietert PJ, et al. Validation of a detailed scoring checklist for use during advanced cardiac life support certification. *Simulation in healthcare : journal of the Society for Simulation in Healthcare*. Aug; 2012 7(4):222–235. [PubMed: 22863996]
15. Neal JM, Hsiung RL, Mulroy MF, Halpern BB, Dragnich AD, Slee AE. ASRA checklist improves trainee performance during a simulated episode of local anesthetic systemic toxicity. *Regional anesthesia and pain medicine*. Jan-Feb;2012 37(1):8–15. [PubMed: 22157743]
16. Arriaga AF, Bader AM, Wong JM, et al. Simulation-based trial of surgical-crisis checklists. *The New England journal of medicine*. Jan 17; 2013 368(3):246–253. [PubMed: 23323901]
17. Ziewacz JE, Arriaga AF, Bader AM, et al. Crisis checklists for the operating room: development and pilot testing. *Journal of the American College of Surgeons*. Aug; 2011 213(2):212–217. e210. [PubMed: 21658974]
18. Harrison TK, Manser T, Howard SK, Gaba DM. Use of cognitive aids in a simulated anesthetic crisis. *Anesthesia and analgesia*. Sep; 2006 103(3):551–556. [PubMed: 16931660]
19. Nelson KL, Shilkofski NA, Haggerty JA, Saliski M, Hunt EA. The use of cognitive AIDS during simulated pediatric cardiopulmonary arrests. *Simulation in healthcare : journal of the Society for Simulation in Healthcare*. 2008; 3(3):138–145. Fall. [PubMed: 19088657]
20. Fleiss, JL.; Levin, BA.; Paik, MC. *Statistical methods for rates and proportions*. 3rd ed.. J. Wiley; Hoboken, N.J.: 2003.
21. Nasca TJ, Philibert I, Brigham T, Flynn TC. The next GME accreditation system--rationale and benefits. *The New England journal of medicine*. Mar 15; 2012 366(11):1051–1056. [PubMed: 22356262]
22. Wood, M. [April 18, 2013] ACGME RRC Update. <http://www.acgme.org/acgmeweb/Portals/0/PDFs/2013AEC-Presentations/ses040.pdf>.
23. Murray DJ, Boulet JR, Avidan M, et al. Performance of residents and anesthesiologists in a simulation-based skill assessment. *Anesthesiology*. Nov; 2007 107(5):705–713. [PubMed: 18073544]
24. Mhyre JM, Ramachandran SK, Kheterpal S, Morris M, Chan PS. American Heart Association National Registry for Cardiopulmonary Resuscitation I. Delayed time to defibrillation after intraoperative and periprocedural cardiac arrest. *Anesthesiology*. Oct; 2010 113(4):782–793. [PubMed: 20808215]

Figure 1.

This figure illustrates the percentage agreement of each of the four non-expert raters with the reference standard grade of the experts for all scenarios by the round (or order) in which it was graded. The figure illustrates that there was no learning curve with repeated viewing and grading of the videos or with the method in which the video was graded, that is, with or without pausing allowed (continuous viewing).

Table 1

Factors associated with agreement/disagreement with reference standard: unadjusted agreement levels and p-values from the generalized linear mixed models

Factor	Unadjusted Percent Agreement with Reference Standard	P-Value from GLMM
Scenario		
Anaphylaxis 1	87.8%	<0.0001
Anaphylaxis 2	85.0%	
CVA1	89.7%	
CVA2	89.1%	
Hyperkalemic Arrest 1	84.2%	
Hyperkalemic Arrest 2	85.6%	
MH 1	94.7%	
MH 2	92.4%	
STEMI 1	91.5%	
STEMI 2	90.5%	
Item considered not to have been performed	92.4	<0.0001
Item considered to have been performed	87.1	
Stopping allowed	90.2%	0.15
Stopping not allowed	88.7%	
Order of grading rounds		0.99
1 (continuous)	88.6%	
2 (with stops)	88.9%	
3 (continuous)	90.3%	
4 (with stops)	90.1%	0.001
Subjective item	86.3%	
Objective item	91.1%	
Critical actions	90.1%	0.22
Non-Critical actions	88.4%	
Rater 1	91.7%	0.002
Rater 2	90.6%	
Rater 3	88.7%	
Rater 4	86.9%	
Average of all 4 Raters	89.6%	

Percent agreement by scenario, item performed/not-performed designation, method of viewing, round of grading, subjective/objective designation, critical/non-critical designation, and by rater are shown. When controlling other variables, variation in percent agreement existed between raters and scenarios, subjective/objective items, and also when comparing items performed/not-performed. There was no difference in agreement between methods of viewing, order of grading, or critical/non-critical item designation. GLMM: Generalized linear mixed model, CVA: cerebral vascular accident, MH: malignant hyperthermia, STEMI: ST-elevation myocardial infarction.