HHS Public Access

Author manuscript

Am J Med Sci. Author manuscript; available in PMC 2018 June 01.

Published in final edited form as:

Am J Med Sci. 2017 June ; 353(6): 516–522. doi:10.1016/j.amjms.2016.09.013.

Assessment Tools for Use During Anesthesia-Centric Pediatric Advanced Life Support Training and Evaluation

Scott C. Watkins, MD, **Paul J. Nietert, PhD**, **Elisabeth Hughes, MD**, **Eric T. Stickles, MD**, **Tracy E. Wester, MD**, and **Matthew D. McEvoy, MD**

Department of Anesthesiology (SW, EH, ES, MM), Vanderbilt University School of Medicine, Nashville, TN; Department of Public Health Sciences (PN) and Department of Anesthesia and Perioperative Medicine (TW), College of Medicine, Medical University of South Carolina, Charleston, SC

Abstract

Background—Pediatric perioperative cardiac arrests are rare events that require rapid, skilled and coordinated efforts to optimize outcomes. We developed an assessment tool for assessing clinician performance during perioperative critical events termed Anesthesia-centric Pediatric Advanced Life Support (A-PALS). Here, we describe the development and evaluation of the A-PALS scoring instrument.

Methods—A group of raters scored videos of a perioperative team managing simulated events representing a range of scenarios and competency. We assessed agreement with the reference standard grading, as well as interrater and intrarater reliability.

Results—Overall, raters agreed with the reference standard 86.2% of the time. Rater scores concerning scenarios that depicted highly competent performance correlated better with the reference standard than scores from scenarios that depicted low clinical competence ($P < 0.0001$). Agreement with the reference standard was significantly ($P < 0.0001$) associated with scenario type, item category, level of competency displayed in the scenario, correct versus incorrect actions and whether the action was performed versus not performed. Kappa values were significantly ($P < 0.0001$) higher for highly competent performances as compared to lesser competent performances (good: mean = 0.83 [standard deviation = 0.07] versus poor: mean = 0.61 [standard deviation = 0.14]). The intraclass correlation coefficient (interrater reliability) was 0.97 for the raters' composite scores on correct actions and 0.98 for their composite scores on incorrect actions.

Conclusions—This study provides evidence for the validity of the A-PALS scoring instrument and demonstrates that the scoring instrument can provide reliable scores, although clinician performance affects reliability.

Keywords

Assessment; Anesthesia; Interdisciplinary Education; Simulation; Teamwork

Correspondence: Scott C. Watkins, MD, 2200 Children's Way, Suite 3115, Nashville, TN 37232 (E-mail: Scott.watkins@vanderbilt.edu).

INTRODUCTION

Perioperative pediatric critical events are rare and unpredictable occurrences that require rapid, skilled and coordinated efforts to optimize outcomes.[1,2] The rarity of these events makes it difficult, if not impossible, for clinicians to develop the skills required for optimal management and for training programs to assess provider performance during actual resuscitation events,[3,4] leading some to propose resuscitation guidelines and curricula tailored to the perioperative environment and suggesting that clinicians receive more frequent resuscitation training with skills assessment.[2–6] The literature contains reports of several scoring instruments for assessing the resuscitation performance of providers in managing both adult and pediatric patients in a variety of settings[4,7–9]; however, there are no valid published tools for the assessment of pediatric resuscitation skills in the perioperative setting with demonstrated reliability. Accordingly, we describe here our methods for developing a set of assessment tools for measuring participant competency during pediatric perioperative resuscitation scenarios and the process of assessing the reliability of scores produced by these tools.

METHODS

The human institutional review board at our institution granted this prospective, observational study exempt status.

Creation of Anesthesia-Centric Pediatric Advanced Life Support Scoring Instrument

We created an assessment checklist (hereafter “checklist”) for 4 pediatric perioperative critical events (hyperkalemia, local anesthetic toxicity, supraventricular tachycardia and anaphylaxis) and 4 cardiac arrest states mirroring the pediatric advanced life support algorithm (asystole, pulseless electrical activity, ventricular tachycardia and ventricular fibrillation) (Figure and Supplemental Digital Content 1–4). We based the methods for developing the assessment checklists on the methodology of similar studies and summarize our methods here.[4,9,10] We established the content of the checklists through a detailed review of relevant literature to identify evidence-based practices and further refined them through a modified Delphi process.[2,9–14] We used the checklists to score several pilot scenarios (described below) and made iterative changes based on observations from these pilot simulations. The checklist items were modified to ensure that each item represented an observable, objective measure of performance. An objective scoring criteria was created for each checklist item as a reference for raters. For example, the checklist item “Acknowledges quality of CPR” includes the criteria “verbalizes presence of end tidal CO2, pulse with compressions, proper rater and/or proper depth” to standardize scoring among raters. Observations from the pilot sessions were used to create additional items for each item on the checklist. For example, the development of the asystole checklist involved a review of current evidence-based resuscitation guidelines to generate a list of best practices and actions a clinician should perform when managing asystole, for example, “administer epinephrine.” Additional checklist items were created based on common errors observed during the pilot sessions. For “administer epinephrine,” additional items were added based on common errors associated with the administration of epinephrine observed during the

pilot sessions, for example, administering epinephrine too soon, too often, at the wrong dose (too much or too little) or when not indicated. This was performed for each item on the checklist. These additional checklist items add granularity to the scoring checklist and permit discrimination between experts and nonexperts. For example, an “expert” would administer epinephrine when indicated at the correct dose route and frequency. A “nonexpert” may know to administer epinephrine in an emergency but not possess detailed knowledge of the appropriate dose, route or frequency.

Creation of Simulated Perioperative Events

To create the 4 simulated scenarios, we paired 1 of the perioperative critical events with 1 of the pediatric advanced life support pathways based on the most likely clinical progression of each event. The 4 scenarios were (1) hyperkalemia progressing to ventricular fibrillation (HyperK > Vfib), (2) local anesthetic toxicity progressing to asystole (LAST > Asyst), (3) anaphylaxis progressing to pulseless electrical activity (Anaphy > PEA) and (4) supraventricular tachycardia progressing to unstable ventricular tachycardia (SVT > VT). A group of expert pediatric anesthesiologists created and vetted introductory stems containing the clinical background story and scenario progression. We revised the stems using an iterative process. Subsequently, anesthesia trainees tested the scenarios during regularly scheduled simulation-training sessions as part of their pediatric anesthesia rotation. We refined the clinical scenarios based on feedback from the trainees, as well as our review of the recordings of these sessions (henceforth referred to as pilot sessions). We chose some of these recorded sessions at random for later use in training the raters (henceforth referred to as training videos).

We recorded 2 scripted versions of each of the 4 scenarios for analysis by the raters in this study. We scripted the study videos to include a version demonstrating highly competent (“good”) management in which participants adhered closely to published guidelines, and another version demonstrating less competent (“poor”) management of the event in which participants deviated from published guidelines. For consistency, we used SimBaby (Laerdal Medical Corp, Stavanger, Norway) software and mannequin to program and perform the scenarios. We conducted all scenarios in our simulation laboratory in a room designed to replicate an operating room at our institution. We recorded all videos in 1 session using the same study personnel. To perform video recording of the scenario management, we used the B-Line system (SimCapture; B-Line Medical, LLC; Washington, DC).

Creation of Reference Standard

Two of the study authors (S.C.W. and M.D.M.) created a reference standard score for each of the recorded scenarios (study videos). The 2 authors are board certified anesthesiologists with greater than 5 years of experience in simulation education, training and research. The study authors scored each study video independently and then compared scores. When the study authors disagreed on individual items, they reviewed and discussed the recordings until an agreement could be reached. The reference score for each study video represents the consensus of the 2 expert raters.

Raters and Rater Training

We recruited 4 raters (2 faculty pediatric anesthesiologists and 2 senior pediatric anesthesia fellows) from the study institution to score the study scenarios. They received training in the manner previously described.[9] We used scores from the training videos to gauge the effectiveness of the rater training. The raters spent approximately 4 hours in a training period in which they viewed and scored 4 training videos and resolved all discrepancies between raters.

Scoring of Simulated Events

We presented videos to the raters in a counterbalanced method so that the order of video review was different for each rater. In addition, we did not inform raters as to which scenarios represented good performance and which scenarios represented poorer performance. The reviewing and scoring of each video occurred on 4 separate occasions—twice with pauses and twice without pauses. We had the raters score videos in this manner to assess whether the ability to pause and rewind the video permitted greater reliability in evaluating items that may have been missed during continuous evaluation and to allow for the assessment of intrarater reliability, that is, the reproducibility of the scoring instrument over time. In addition, the order in which raters scored videos was different for each rater to counter any potential learning effect.

Statistical Analysis

We used a variety of methods to assess agreement between raters and the reference standard, as well as interrater and intrarater reliability on checklist items. First, we used generalized linear mixed models (GLMMs) to assess factors associated with raters' agreement with the reference standard. GLMMs are ideal for modeling correlated binary outcomes, as was the case with our data given that each rater provided assessments on more than 1,900 individual checklist items during the experiment.[15] Within the GLMM, agreement between the raters and the reference standard served as the dependent variable. Fixed effects included scenario type, item category, pauses (allowed versus not allowed), whether the item was deemed to be a correct (versus incorrect) action, whether the expert (reference standard) scored the item as having been performed (versus not performed) and overall scenario performance (good versus poor). The GLMMs contained random rater effects and assumed a compound symmetry covariance structure; we used random effects to account for the fact that raters' item assessments within and between scenarios were correlated with one another. As the dependent variable was binary, we used a logit link function.

In addition, we calculated kappa statistics across checklist items for each rater's assessment of each scenario. As we could calculate kappa for each scored scenario, we constructed a separate general linear model to investigate whether agreement (as measured by kappa) was associated with scenario type, scenario performance (good versus poor), pauses (allowed versus not allowed) and rater.

Besides agreement on individual checklist items, it was also important to assess whether the checklist produced reliable composite scores (1 composite score for percentage of correct actions performed and another for the number of incorrect actions performed). First, we

calculated Lin's concordance correlation coefficients (CCCs) to compare raters' composite scores with those of the reference standard; CCCs are measures of agreement similar to Pearson's correlations, but CCCs also provide a "penalty" if a rater's score is systematically higher or lower than the reference standard.[16] We also calculated intraclass correlations that helped provide a sense of the degree to which scores are reproducible from rater to rater.[17] We conducted all analyses using SAS 9.4 (Cary, NC) and considered $P < 0.05$ statistically significant.

RESULTS

The 4 raters scored each of the 8 scenarios 4 times for a total of 128 recordings with 7,631 discrete items observed. Table 1 details the reference standard scores and the scores assigned by the raters. Overall, raters agreed with the reference standard 86.2% of the time. Raters' scores for "good" performances had a higher correlation with the reference standard than scores from "poor" performances ($P < 0.0001$). Table 2 highlights the results of the GLMM model examining the relationships between the experimental conditions of interest and agreement with the reference standard. Agreement with the reference standard was significantly ($P < 0.0001$) associated with scenario type, item category, good versus poor performance, correct versus incorrect actions and whether the action was performed versus not performed (Table 2). Although there was significant ($P < 0.0001$) variation in agreement between raters across all scenarios, agreement between raters and the reference standard was consistently high, ranging from 83.5-88.7% across raters.

Kappa Correlation of Individual Item Responses by Raters With Reference Standard

The overall average kappa between the raters and the reference standard was 0.72 across all scenarios, which qualifies as "good" agreement according to Shrout and Fleiss's criteria[17] and "substantial" according to the Landis and Koch's criteria.[18] Kappa values were relatively consistent, ranging from 0.67-0.77 across raters and ranging from 0.69-0.75 across scenario types. Kappas averaged 0.73 when raters could pause the video and 0.72 when they viewed the video without pauses. In a general linear model, we noted no significant association between kappa values and scenario type ($P = 0.06$), whether or not pauses were allowed ($P = 0.53$) or rater ($P = 0.24$); however, kappa values were significantly ($P < 0.0001$) higher for scenarios that were deemed to be "good" rather than "poor" performances (good: mean = 0.83 [standard deviation = 0.07] versus poor: mean = 0.61 [standard deviation = 0.14]).

Reliability of Checklist Composite Scores

For the composite scores for correct actions, test-retest reliability (intrarater reliability or within rater reliability), as assessed by Lin's CCC values, was moderately strong, averaging 0.68 across raters and ranging from 0.58-0.82. CCC values comparing the raters' scores on incorrect actions to the reference standard were moderately strong, averaging 0.76 and ranging from 0.62-0.88. Across the 4 raters and the reference standard, the interrater reliability (or between-raters reliability), as measured by the intraclass correlation coefficient, was 0.97 for the raters' composite scores on correct actions and 0.98 for their composite scores on incorrect actions.

DISCUSSION

In this study, we sought to develop an instrument for measuring pediatric perioperative resuscitation skills for use in training and evaluation, and offer evidence that the tool provides both valid and reliable scores. This article describes this process and makes a number of notable contributions to the growing field of simulation-based assessment of clinicians. The instrument appears to provide a valid measure of performance, and scores are equally reliable when used in real time, which is important if such tools are ever to be used for clinical assessment. In addition, the tools were developed using methodology that can be replicated to generate assessment instruments for other clinical scenarios or situations.

"Validity refers to the evidence presented to support or refute the meaning or interpretation assigned to assessment results …. without evidence of validity, assessments in medical education have little or no intrinsic meaning."[19] In the unified model of validity first proposed by Messick, evidence is compiled from 5 sources, such as (1) content, (2) response process, (3) internal structure, (4) relations with other variables and (5) consequences, into 1 argument supporting the validation of scores generated from an instrument.[20] We developed the A-PALS instrument through an iterative process by subject matter experts using available evidence-based resources, thus providing evidence for its *content.* The scores generated from the A-PALS tool permitted the raters to discriminate between high- and low-competency performances and between correct and incorrect actions, providing evidence for relations with other variables. The agreement between raters and the reference standard was largely a function of team performance and scenario tested rather than the rater or mode of scoring (with or without pauses), thus implying that the tools produce reliable scores under diverse conditions, which demonstrates response process evidence for the validity of the scoring instrument, as well as a low level of rater error contributing to the score variation. Variation in A-PALS scores was largely a function of scenario type, individual categories of the A-PALS tool and whether the performance being measured demonstrated low competency, providing additional evidence for response process. Although there was significant variation in scores from the ratings of low-competency performances, the scores generated by the raters were closely associated with the reference standard. Additionally, the scores were reproducible and consistent both between and within raters, providing evidence for internal structure. We created and programmed the scenarios and videos used for scoring in a uniform manner, thus ensuring that another investigator could easily reproduce them, providing additional evidence for the internal structure of the scoring instrument. Although, we did not determine a minimum passing score for each scenario, a score of 75% is often used as a cutoff value for "passing."[21] The scores provided by raters for "high" and for "low" competency performances were well above or below this value, thus providing consequence evidence. This would suggest that scoring variation imposed by scenario performance, while statistically significant, would not have affected the determination of a passing or failing score, that is, the scores were not educationally significant.

The A-PALS assessment tool provided highly reproducible scores as evidenced by the reliability within and between raters and the substantial and consistent agreement between raters and the reference standard. We can attribute most of the variance in scores from A-PALS to the performances being measured and the clinical scenario tested rather than to the

rater or scoring method (with or without pauses). Interestingly, agreement between raters and the reference standard was significantly better for scenarios representing high-competency performance when compared to low-competency performance, yet raters were more consistent when scoring incorrect actions than correct actions. However, the raters consistently assigned higher scores to scenarios representing high-competency performance and lower scores to scenarios representing low-competency performance. This would suggest that scores from the A-PALS tool can discriminate between different levels of performance, but that scoring of low-competency performances may require additional rater training or be inherently more difficult.

The current study is not without limitations. Firstly, we investigated error associated with time of rating only and did not assess task-sampling error (i.e., the number of scenarios or the number of raters needed to generate a reliable estimate of ability). Based on previous investigations, these may be more important sources of measurement error than interrater or intrarater reliability.[22–24] Secondly, the scoring instrument is specific to the unique scenarios studied, although our methodology provides a framework for others to replicate with different clinical scenarios. Thirdly, we did not perform item by item analyses to determine whether certain items could or should be removed from the checklists; such an approach could be the subject of future research. Finally, reliability and agreement between raters was lower for the scenarios representing lesser competency.

CONCLUSIONS

Clinically, perioperative resuscitations are rare events, highlighting the importance of developing simulation-based education and assessment strategies for ensuring competency. This study provides evidence for the validity of the A-PALS scoring instrument and demonstrates that the scoring instrument can provide reproducible scores with minimal rater training. Based on the validity evidence presented and the reliability of the scores generated by raters in this study, the A-PALS tool could be used for formative assessment of clinicians. Future studies should (1) identify and quantify potential sources of measurement error, (2) examine whether individual instrument items should be added or removed to further enhance the instrument and (3) gather additional evidence to support the validity of the ratings using a larger sampling of scenarios and clinicians.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Support for this study was provided by a research grant from the Anesthesia Patient Safety Foundation, United States, (Grant number 4-04-300-6125) (01/2014 to 08/2015). The content is solely the responsibility of the authors and does not necessarily reflect the views of the Anesthesia Patient Safety Foundation. Dr. Nietert was also funded, in part, by grants from the National Institutes of Health (National Center for Advancing Translational Sciences, United States, Grant number UL1TR001450 and National Institute of General Medical Sciences, United States, Grant number U54-GM104941).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at http://doi:10.1016/j.amjms.2016.09.013.

References

1. Ramamoorthy C, Haberkern CM, Bhananker SM, et al. Anesthesia-related cardiac arrest in children with heart disease: data from the Pediatric Perioperative Cardiac Arrest (POCA) registry. Anesth Analg. 2010; 110(5):1376–82. [PubMed: 20103543]
2. Shaffner DH, Heitmiller ES, Deshpande JK. Pediatric perioperative life support. Anesth Analg. 2013; 117(4):960–79. [PubMed: 24023023]
3. Hunt EA, Walker AR, Shaffner DH, et al. Simulation of in-hospital pediatric medical emergencies and cardiopulmonary arrests: highlighting the importance of the first 5 minutes. Pediatrics. 2008; 121(1):e34–43. [PubMed: 18166542]
4. Donoghue A, Nishisaki A, Sutton R, et al. Reliability and validity of a scoring instrument for clinical performance during pediatric advanced life support simulation scenarios. Resuscitation. 2010; 81(3):331–6. [PubMed: 20047787]
5. Watkins SC. The pharmacology of resuscitation training—time for a new treatment plan. Paediatr Anaesth. 2014; 24(12):1307–8. [PubMed: 25378042]
6. Moitra VK, Gabrielli A, Maccioli GA, et al. Anesthesia advanced circulatory life support. Canad J Anaesth. 2012; 59(6):586–603. [PubMed: 22528163]
7. Levy A, Donoghue A, Bailey B, et al. External validation of scoring instruments for evaluating pediatric resuscitation. Simul Healthc. 2014; 9(6):360–9. [PubMed: 25503530]
8. Grant EC, Grant VJ, Bhanji F, et al. The development and assessment of an evaluation tool for pediatric resident competence in leading simulated pediatric resuscitations. Resuscitation. 2012; 83(7):887–93. [PubMed: 22286047]
9. McEvoy MD, Smalley JC, Nietert PJ, et al. Validation of a detailed scoring checklist for use during advanced cardiac life support certification. Simul Healthc. 2012; 7(4):222–35. [PubMed: 22863996]
10. Morgan PJ, Lam-McCulloch J, Herold-McIlroy J, et al. Simulation performance checklist generation using the Delphi technique. Canad J Anaesth. 2007; 54(12):992–7. [PubMed: 18056208]
11. Kleinman ME, Chameides L, Schexnayder SM, et al. Part 14: pediatric advanced life support: 2010 American Heart Association Guidelines for Cardiopulmonary Resuscitation and Emergency Cardiovascular Care. Circulation. 2010; 122(18 suppl 3):S876–908. [PubMed: 20956230]
12. Kleinman ME, de Caen AR, Chameides L, et al. Part 10: pediatric basic and advanced life support: 2010 International Consensus on Cardiopulmonary Resuscitation and Emergency Cardiovascular Care Science With Treatment Recommendations. Circulation. 2010; 122(16 suppl 2):S466–515. [PubMed: 20956258]
13. Heitmiller, ES. Society for Pediatric Anesthesia Pediatric Critical Events Checklist. 2013. http://www.pedsanesthesia.org/newnews/Critical_Event_Checklists.pdf?201310291500
14. Neal JM, Bernards CM, Butterworth JF, et al. ASRA practice advisory on local anesthetic systemic toxicity. Reg Anesth Pain Med. 2010; 35(2):152–61. [PubMed: 20216033]
15. McCulloch, CSS. Generalized, Linear, and Mixed Models. New York: Wiley & Sons, Inc; 2001.
16. Lin LI. A concordance correlation coefficient to evaluate reproducibility. Biometrics. 1989; 45(1): 255–68. [PubMed: 2720055]
17. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979; 86(2):420–8. [PubMed: 18839484]
18. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; 33(1):159–74. [PubMed: 843571]
19. Downing SM. Validity: on meaningful interpretation of assessment data. Med Educ. 2003; 37(9): 830–7. [PubMed: 14506816]

20. Cook DA, Zendejas B, Hamstra SJ, et al. What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. Adv Health Sci Educ Theory Pract. 2014; 19(2):233–50. [PubMed: 23636643]
21. Wayne DB, Fudala MJ, Butter J, et al. Comparison of two standardsetting methods for advanced cardiac life support training. Acad Med. 2005; 80(10 suppl):S63–6. [PubMed: 16199461]
22. Murray DJ, Boulet JR, Kras JF, et al. A simulation-based acute skills performance assessment for anesthesia training. Anesth Analg. 2005; 101(4):1127–34. [PubMed: 16192533]
23. Weller JM, Robinson BJ, Jolly B, et al. Psychometric characteristics of simulation-based assessment in anaesthesia and accuracy of self-assessed scores. Anaesthesia. 2005; 60(3):245–50. [PubMed: 15710009]
24. Boulet JR, Smee SM, Dillon GF, et al. The use of standardized patient assessments for certification and licensure decisions. Simul Healthc. 2009; 4(1):35–42. [PubMed: 19212249]

A-PALS Checklists for Hyperkalemia	Performed (No, Yes)
General Assessment	
Review case including anesthetic record, PMH, Labs	
Check Vitals	
Check Pulse	
Assessed and verbalized anesthesia machine settings	
Performed Physical Exam	
Performed Timeout	
Crisis Specific	
Announce Problem	
Called for help	
Call for code cart/defibrillator	
Decrease Volatile Anesthetic	
Increase FiO2	
Review recent medications or fluids administered	
Send Labs (ABG, VBG, CBC, iSTAT, etc)	
Hyperkalemia Specific Management	
Assessed for Hyperkalemia	
Started IV fluid Bolus	
Gave Calcium Chorlide 10-20 mg/kg or Calcium Gluconate 30 MG/KG*	
Gave excessive dose of Calcium	
Gave inadequate dose of Calcium	
Gave Bicarb 1-2 meq/kg	
Gave excessive dose of Bicarb (>2 meq/kg)	
Gave inadequate dose of Bicarb (<1 meq/kg)	
Gave Insulin 0.1 unit/kg	
Gave excessive dose of Insulin (>0.5 units/kg)	
Administered Dextrose 0.5 gm/kg	
Administered excessive Dextrose	
Administered inadequate Dextrose	
Administered Albuterol (5-10 puffs)	
Considered Furosemide 0.5-1 mg/kg	
Gave excessive dose of Furosemide	
Gave inadequate dose of Furosemide	
Mentions or Stops Potassium Containing Fluids *	
Errors	
Gave insulin without glucose	
Other	

**Critical Actions*

FIGURE.

A-PALS checklists for hyperkalemia and ventricular fibrillation.

TABLE 1

Comparison of reference standard and trained raters' scores.

Scoring item	Reference standard	Trained raters
Composite score[a]	% (SD)	% (SD)
Good Performances	82.6% (±5.0%)	86.8% (±6.2%)*
Poor Performances	67.2% (±3.8%)	59.0% (±11.7%)
Incorrect Item Count[b]	Average (SD)	Average (SD)
Good Performances	0.50 (±0.87)	0.25 (±0.56)
Poor Performances	3.25 (±1.31)	4.55 (±2.32)

SD, Standard deviation.

*$P < 0.0001$

[a]Percentage of items done that should have been done.

[b]Count of the number of incorrect actions performed.

TABLE 2

Factors associated with agreement with the reference standard assessment of checklist items.

Factor	Unadjusted percent agreement with reference standard	Adjusted *P* Value from GLMM model
Scenario type		< 0.0001
Anaphylaxis	87.9	
Hyperkalemia	84.9	
LAST	84.6	
SVT	87.5	
Categories of items		< 0.0001
General assessment	82.2	
Crisis management	86.3	
ROSC management	81.7	
Management errors	91.1	
Perioperative events		
Anaphylactic-specific management	92.3	
Hyperkalemia-specific management	87.1	
LAST-specific management	85.4	
SVT-specific management	95.8	
PALS		
PEA management	80.5	
Asystole management	75.2	
VFib management	88.2	
VTach management	88.6	
Scoring method		
Pauses allowed	86.4	0.53
Pauses not allowed	85.9	
Performance of team		
Good performance	91.6	< 0.0001
Poor performance	80.7	
Actions		
Correct action	84.0	< 0.0001
Incorrect action	90.2	
Action performed (as assessed by reference standard)	86.2	< 0.0001
Action not performed (as assessed by reference standard)	86.1	
Raters		< 0.0001
Rater 1	83.5	
Rater 2	88.7	
Rater 3	86.9	
Rater 4	85.6	

GLMM, generalized linear mixed models; LAST, local anesthetic systemic toxicity; SVT, supraventricular tachycardia; ROSC, return of spontaneous circulation; PALS, pediatric advanced life support; PEA, pulseless electrical activity; VFib, ventricular fibrillation; VTach, ventricular tachycardia.