Published in final edited form as:

Simul Healthc. 2012 August ; 7(4): 222–235. doi:10.1097/SIH.0b013e3182590b07.

Validation of a Detailed Scoring Checklist for Use During Advanced Cardiac Life Support Certification

Matthew D. McEvoy, MD [Associate Professor and Vice-Chair for Education],

Department of Anesthesia and Perioperative Medicine, Assistant Dean for Patient Safety and Simulation, Medical University of South Carolina, 167 Ashley Avenue, Suite 301, Charleston, SC 29425, 843.792.2322 (phone), 843.792.2726 (fax), mcevoymd@musc.edu

Jeremy C. Smalley, MD [Resident],

Department of Anesthesia and Perioperative Medicine, Medical University of South Carolina, smalley@musc.edu

Paul J. Nietert, PhD [Associate Professor],

Department of Medicine, Division of Biostatistics and Epidemiology, Medical University of South Carolina, nieterpj@musc.edu

Larry C. Field, MD [Assistant Professor],

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

Cory M. Furse, MD [Assistant Professor],

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

John W. Blenko, MD [Associate Professor],

Department of Anesthesiology, University of Maryland School of Medicine, jblenko@umm.edu

Benjamin G. Cobb, MD [Resident],

Department of Anesthesiology & Pain Medicine, University of Washington, bgcobb@uw.edu

Jenna L. Walters, MD [Resident],

Department of Anesthesiology, Vanderbilt University Medical Center, jenna.l.walters@vanderbilt.edu

Allen Pendarvis, MD [Resident],

Department of Anesthesia and Perioperative Medicine, Medical University of South Carolina, pendarva@musc.edu

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Conflict of Interest Statement:

Dr. Schaefer contributed mentorship, review of videos, guidance concerning checklist design, and manuscript review for this study, but pursuant to the Medical University of South Carolina Conflict of Interest (COI) policy, he did not participate in data collection, reduction or analysis related to this study due to his potential COI, which includes simulator patent and copyright royalties. Dr. Schaefer receives patent royalties from Laerdal Medical Corporation (SimMan/Baby/3G) and he is a non-majority owner of SimTunes, which is a commercial outlet for Medical University of South Carolina licensed, copyrightable simulation training products. These amount to <0.5% of Dr. Schaefer's annual income.

The full contents of this information are included in the attached COI forms. In brief, Dr. McEvoy and Mr. Smalley have non-majority equity interest in Patient Safety Strategies, LLC. This is a company that markets medical applications for iOS-compatible devices. Neither of these authors have received any remunerations from the company, nor were any company products tested during this investigation. Dr. Schaefer has Patent royalties from Laerdal Medical Corp. (SimMan/Baby/3G) and a non-majority ownership of SimTunes (an outlet copyrightable simulation material).

Nishita S. Dalal, MD [Resident], and

Department of Anesthesia and Perioperative Medicine, Medical University of South Carolina,
dalal@musc.edu

John J. Schaefer III, MD [Lewis Blackman Chair and Professor]

Department of Anesthesia and Perioperative Medicine, Director of Clinical Effectiveness and
Patient Safety Center, Medical University of South Carolina, schaefer@musc.edu

Abstract

Introduction—Defining valid, reliable, defensible, and generalizable standards for the evaluation of learner performance is a key issue in assessing both baseline competence and mastery in medical education. However, prior to setting these standards of performance, the reliability of the scores yielding from a grading tool must be assessed. Accordingly, the purpose of this study was to assess the reliability of scores generated from a set of grading checklists used by non-expert raters during simulations of American Heart Association (AHA) MegaCodes.

Methods—The reliability of scores generated from a detailed set of checklists, when used by four non-expert raters, was tested by grading team leader performance in eight MegaCode scenarios. Videos of the scenarios were reviewed and rated by trained faculty facilitators and by a group of non-expert raters. The videos were reviewed “continuously” and “with pauses.” Two content experts served as the reference standard for grading, and four non-expert raters were used to test the reliability of the checklists.

Results—Our results demonstrate that non-expert raters are able to produce reliable grades when using the checklists under consideration, demonstrating excellent intra-rater reliability and agreement with a reference standard. The results also demonstrate that non-expert raters can be trained in the proper use of the checklist in a short amount of time, with no discernible learning curve thereafter. Finally, our results show that a single trained rater can achieve reliable scores of team leader performance during AHA MegaCodes when using our checklist in continuous mode, as measures of agreement in total scoring were very strong (Lin’s Concordance Correlation Coefficient = 0.96; Intraclass Correlation Coefficient = 0.97).

Discussion—We have shown that our checklists can yield reliable scores, are appropriate for use by non-expert raters, and are able to be employed during continuous assessment of team leader performance during the review of a simulated MegaCode. This checklist may be more appropriate for use by Advanced Cardiac Life Support (ACLS) instructors during MegaCode assessments than current tools provided by the AHA.

Keywords

simulation; education; checklist; reliability; ACLS

Introduction

Defining valid, reliable, defensible, and generalizable standards for the evaluation of learner performance is a key issue in assessing both baseline competence and mastery in medical education.¹ Regarding Advanced Cardiac Life Support (ACLS), the standards for performance are the published guidelines for patient management. Several studies have previously described the process by which an ACLS checklist can be validated and the process by which minimum passing scores for team leader performance can be determined for ACLS testing.^{2,3,4,5} Furthermore, the processes of checklist development, validation, and reliability testing have been described for other areas of medical training and assessment, including cardiac auscultation, objective structured clinical examinations, and anesthesia

simulation scenarios.^{6,7,8} However, the literature to date leaves several important questions unanswered regarding the reliability of scores yielded from grading tools for ACLS testing.

First, are these checklists objective and simple enough to be used by non-expert raters as opposed to faculty facilitators trained in grading simulation performances? Prior publications for some checklists describe how professionals with training in simulation research and psychometric analysis tested the checklist being studied.^{9,10} But, the final end user of these published checklists will likely not have such training. Other publications have demonstrated that non-expert raters, such as standardized patients or lay persons with no medical training, can use checklists to produce reliable scores that enable valid judgments of trainees.^{11,12,13} However, this has not been demonstrated with ACLS checklists. As such, ACLS checklists intended for use in simulation testing should undergo evaluation in order to ensure that a high level of reliability in scores is obtained when they are used by instructors not involved in checklist design and who have not received targeted instruction in simulation training. Second, if the checklists can be used by a non-expert rater, is training for proper checklist use of such duration that it could be given routinely to ACLS instructors? Third, is the checklist able to be employed during real-time (i.e., continuous) evaluation of the team leader managing an American Heart Association (AHA) MegaCode, or does its use require additional time for video review after the competency exam? Accordingly, the purpose of this study was to assess the reliability of a checklist used as a continuous grading tool by non-expert raters during the review of simulations of AHA MegaCodes.

Methods

This study was submitted to our Institutional Review Board and granted 'exempt' status.

Participants

For the purposes of this paper, the term 'experienced' refers to individuals who have been involved with simulation education, training, and research for at least three years, and who teach crisis management simulation courses, including ACLS, at least five times a year (MM, LF, CF, JS). The non-expert raters in this study were four 4th year medical students (BC, JW, AP, ND). These students had previously received ACLS training and certification via high-fidelity simulation and were familiar with the SimMan® (SimMan3G, Laerdal Inc., Stavanger, Norway) software interface. However, none of the four non-expert raters had received training in SimMan® programming, checklist design, or simulation facilitation, grading, or debriefing. None of the four students had received training as an ACLS Instructor.

Checklist Creation

The ACLS checklists used in this study were constructed following the steps outlined in previous checklist development studies, as well as those from authors experienced in checklist design.^{1,2,3,4,14,15} First, the checklist items were constructed through a detailed review of the AHA 2005 Guidelines for Cardiopulmonary Resuscitation (CPR) and Emergency Cardiovascular Care (ECC), thus assuring content validity.¹⁶ Second, this content was evaluated by the group of faculty experts using a modified Delphi technique in order to determine the exact form of the items on the checklist.¹⁷ Third, the checklists were then divided into several portions, including the initial evaluation of the patient, items specific to the three patient conditions managed during the MegaCode simulation (e.g., pulseless electrical activity (PEA), stable tachycardia, ventricular fibrillation (VFIB)), and a section on common/possible errors. Fourth, four final checklists were constructed that would represent specific checklists to be used during the grading of MegaCode scenarios

programmed within the SimMan® software (see Appendices 1–4). These MegaCode scenarios contained the following patient state sequences:

1. Unstable bradycardia – VFIB – asystole
2. Stable tachycardia – Pulseless Ventricular Tachycardia (PVT) – PEA
3. Unstable bradycardia – PVT – asystole
4. Stable tachycardia – VFIB – PEA

After the checklists were completed in terms of content and order of the items, each item was then further categorized by the faculty experts in a number different ways: correct/incorrect actions, objective/subjective checklist item, critical/non-critical action. Items were first categorized as a correct or incorrect action. These varied for each checklist depending on the patient state. For example, defibrillation is a correct step in the VFIB protocol, but an incorrect step in the PEA or asystole pathways. Each of the checklist consisted of 63 to 68 correct items and 32 to 34 incorrect items. Items were also classified as subjective or objective, depending upon whether any interpretation was needed by the rater. Examples of objective items involved discrete actions, such as “defibrillated at 200J biphasic” or “gave Epinephrine 1mg IV.” Examples of subjective items often involved assessment and communication steps, such as “assessed reversible causes of arrest,” “assessed patient stability,” and “assigned team member roles.” These steps are more complex because “assessed reversible causes of arrest” could be done by reviewing them out loud with the team, thinking about them without verbalizing the condition in the differential, or a combination of the two methods. For example, the team leader may order a treatment for hyperkalemia without actually stating hyperkalemia in the differential. This can also be true for “assessed patient stability” because asking the patient certain questions is important in this step (e.g., “Are you short of breath? Do you have chest pain?”), as is assessing the ECG tracing and the vital signs; and this complex assessment can be done in a variety of combinations. Thus, while rules were given during the training period to account for this variety in the subjective items, grading these items often required some subjectivity on the part of the rater. The variations in the number of items, the correct or incorrect nature of each item, the objective or subjective nature of each item, and whether the item was deemed “critical” can be seen in the full checklists shown in Appendices 1–4.

Creation of Videos Used for Rater Training and Checklist Analysis

Four training videos of simulated ACLS MegaCodes were created for use in the checklist training curriculum for the non-expert raters. One video was made for each of the four checklists (see Appendices 1–4). After this, eight ACLS MegaCode simulation videos were created so that scores yielded from the checklists could be tested for reliability. Two videos were made for each checklist. The scenarios were all programmed within the SimMan® software and executed using the SimMan3G manikins. All scenarios were executed in our simulation laboratory in a patient care room that duplicates an in-patient hospital room at our institution, and with a code cart replicating those in use in our hospital. The team leader performance during the videos was scripted to include both correct and incorrect actions so that the raters would not be evaluating only expert performance. The team leaders in the training videos (N=4) were ACLS-certified faculty at our institution, and the team leaders in the testing videos (N=8) were ACLS-certified resident volunteers in our anesthesiology training program. The code team members (i.e., confederates) in both video sets consisted of trained simulation staff members who played the roles of two nurses performing CPR, a nurse managing the defibrillator, a pharmacist, and an airway manager. All of the videos were roughly 10–12 minutes. Video recording of the scenario management was performed using the B-Line system® (SimCapture, B-line Medical, LLC, Washington, DC).

After the creation of both sets of videos, two of the experienced faculty (MM and JS) graded the performance in the videos 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. Of note, the experienced raters provided very similar ratings of the video performances. The most problematic items were those considered subjective, such as “assessed reversible causes of the arrest.” For these few areas, the two faculty members watched the videos together and reached an agreement on the grading of the particular discrepant/discordant items. The reference standard was created so that the non-expert raters would be compared not only to themselves and one another (intra-rater and inter-rater reliability), but also to experienced simulation facilitators, as one of the aims of this study was to evaluate whether non-expert raters could use the checklist as accurately as faculty proficient in assessment and grading of simulation training and ACLS.

Training of Non-expert Ratets

The training of non-expert raters consisted of several parts. First, they received instruction on the content of the checklists and the meaning of each item. Second, they were given instructions on how to grade items that were classified as subjective. Third, they were given a chance to ask questions about the checklist items and grading criteria. Fourth, the group was shown one of the four training videos. This video was played continuously and then with pauses to illustrate the two methods by which the MegaCode performances were to be graded. That is, the non-expert raters were to practice grading the team leader first in “continuous” mode (i.e., letting the video play continuously without the ability to stop or rewind, and grading the team leader performance while viewing the video). Then they were to practice grading the team leader performance “with pauses” (i.e., having the freedom to pause, stop, and/or rewind the video during the grading process). This second method was allowed in order to make sure that the raters felt confident of their evaluation of every item on the checklist in relation to the performance of the team leader in the video.

Following this group session, each non-expert rater individually graded the four training videos, including the one that they had viewed as a group. Each video was graded once continuously and once with pauses. The raters were debriefed on their checklist evaluations individually and as a group. That is, one experienced facilitator (MM) would do a 1:1 debriefing with each non-expert rater after the rater had completed a grading session. This facilitator clarified questions with respect to grading and discussed areas where videos were graded differently by the expert raters. This helped to clarify grading criteria for the set of eight videos to be evaluated in the study. For instance, criteria were set for how to grade items such as “stated diagnosis of unstable bradycardia,” as there were questions regarding how specific the wording must be from the team leader (i.e., did the leader have to state “the diagnosis is unstable bradycardia,” or could various combinations of words be used as long as the leader indicated the correct diagnosis to the team?). Finally, the group met to discuss any final questions about the checklist and grading criteria.

The checklists were presented to the non-expert raters in Microsoft Excel® (Excel 2007, Microsoft, Inc., Seattle, WA) format for review during the training and study periods. They recorded their assessment of the team leader in this format during the training period and during the performance of the study. Their performance evaluations were entered directly into the Excel program on the computer. The Excel spreadsheet was designed to allow most of the checklist items to be seen at any one time (when displayed on a 19” desktop monitor). More importantly, the items were arranged top to bottom in a fashion that mirrored the design of the scenarios, and thus, the order in which they would likely be assessed. For instance, one scenario started in stable tachycardia and then advanced to PVT and then to PEA. In the Excel spreadsheet, the items related to stable tachycardia would be the first set of items on the checklist, then those for PVT and finally those for PEA (see Appendices 1–

4). All of the relevant checklist items for any patient state within the scenario (e.g., VFIB) could be seen at one time by the rater. Thus, all items being rated at a given time were visually accessible to the rater on the computer screen without the need for scrolling. The total training time on the use of the checklist was approximately four hours, including the scoring of the four videos used for training on use of the checklists. All of the training was done in a single half-day session.

Checklist Evaluation

The non-expert graders were then informed about the order in which to evaluate the eight MegaCode simulation videos. This video evaluation sequence was created by a random order generator and was different for each rater. The scenarios were presented to them in random order because each video was to be rated four times by each rater – twice in “continuous” mode and twice “with pauses” (see Table 1). The purpose of rating the videos in this manner was two-fold. First, it was to assess whether there was greater reliability in scoring if the rater was able to pause and rewind the video to evaluate items that may have been missed during continuous evaluation (“with pauses” versus “continuous”). Second, the videos were presented in random order to reduce repetition bias. Video evaluations by the non-expert raters were compared to the expert video evaluation to test whether there was an improvement in agreement with repeated analysis of the videos. This was to test whether the training curriculum was sufficient, or if there was an ongoing learning curve with continued use of the checklist. The video scoring was accomplished over three days. Day 1 included two half-days in which the videos were reviewed and scored once in continuous mode in the morning, and then once in continuous mode in the afternoon. Days 2 and 3 included reviewing and grading the videos “with pauses” once each day. The “with pauses” method of grading took about twice as long compared to the “continuous” mode.

Statistical analysis

The inter-rater and intra-rater reliability of the checklist scores were primarily assessed using a generalized linear mixed model (GLMM) approach (see Appendix 5 for a complete description). The first GLMM used individual checklist items as the unit of analysis. This model used binary dependent variables reflecting whether or not a given rater’s assessment agreed with the expert assessment. The GLMM included the following independent variables: scenario; “continuous” vs. “with pauses”; round (1 or 2); whether the checklist item was a correct/incorrect action; whether the item was subjective/objective; sequential order of the scenario (1 through 7); whether the expert claimed the item had been performed; and whether the experts viewed each item as “critical.” A random effect was included to account for the fact that within-rater assessments are likely correlated with one another. The GLMM used a logit link function and was conducted using SAS v9.2 (SAS, Cary, NC). This modeling process is somewhat analogous to the “G-Theory” approach for evaluating performance assessments on continuous assessment scores.^{18,19} We also used a separate GLMM to examine inter-rater and intra-rater reliability across all items within a given ACLS MegaCode scenario. In other words, we assessed the level of agreement between the raters and the expert across all checklist items (~n=100 per scenario, including possible incorrect actions), and modeled that value (using an identity link function) as a function of scoring method (“continuous” vs. “with pauses”) and grading round (1 or 2), again using random rater effects to control for the correlated nature of the data. We also summarized agreement using the kappa statistic (κ). Shrout and Fleiss suggest that κ greater than 0.75 is indicative of excellent agreement, κ between 0.40 and 0.75 is fair to good, and κ below 0.40 is considered poor.²⁰

In addition to the measures of agreement for individual items within the checklist, we also measured Lin’s Concordance Correlation Coefficient (CCC) and the intraclass correlation

coefficient (ICC) for the code leader composite scores, as determined by four raters and the reference standard.^{20,21,22} Separate composite measures were created for “correct” and “incorrect actions” performed by the code leader (as assessed by the rater). The correct actions composite scores were determined by summing the total of all possible correct actions performed divided by the total number of possible correct actions (i.e., % correct), while the incorrect actions composite scores were determined by summing the total number of incorrect actions performed. The CCC ranges from -1.0 to +1.0 and reflects overall agreement between any two sets of scores. It is similar to the traditional Pearson’s correlation coefficient, but the CCC also includes a penalty for any systematic bias. We were able to use CCC values to assess test-retest reliability (comparing raters’ scores of the same scenarios on two different occasions) and to assess accuracy (i.e., agreement between each rater and the reference standard). The ICC ranges from 0% to 100% and reflects the degree to which score variability is attributable to the code leader as opposed to inter-rater variability.

Under the assumption of low variability across raters in agreement with the reference standard (which was confirmed in our data), our sample of four raters assessed against a reference standard in grading eight scenarios four times each (two rounds in “continuous” mode and two rounds “with pauses”) provided sufficient power (>80%) to determine if there was a greater than 2% inter-round difference between experienced and non-expert raters, as well as between non-expert raters, in the scoring of team leader actions during ACLS Megacode simulations.

Results

An online supplement (Table, Supplemental Digital Content 1, <http://links.lww.com/SIH/A30>) shows a complete listing of raw scores by rater, by grading round, and by scenario. On average, in the eight videos used for this study, the team leaders performed correct actions 66.2% of the time (SD: 10.4%; range 53.7% to 82.1%) and performed 2.6 (SD: 2.0; range 0 to 6) incorrect actions, as assessed by the reference standard. These performances reflect a range of poor to highly competent performance. Overall, the non-expert raters agreed with the reference standard on individual checklist items 88.9% of the time. Table 2 shows the results of the primary GLMM model examining independent associations between experimental conditions (factors) with agreement/disagreement with the reference standard on individual checklist items. Agreement with the reference standard varied considerably across the eight scenarios, with levels of agreement ranging from 83.3% to 91.43%. Agreement was significantly higher on items that were deemed to be incorrect actions vs. correct actions, on items that were deemed to be objective vs. subjective, and on “critical” items vs. “not critical.” We observed significant ($p=0.03$) variation across raters; however, the range of agreement only varied by 1.86 percentage points. No significant variation was observed when comparing the scoring methods, grading rounds, or comparing actions deemed to be performed by the reference standard vs. actions deemed not performed. Similar results (i.e., significant variability across scenarios, significant but small variability among raters, and no significant differences comparing scoring methods or grading rounds) were observed when the total percent agreement score was used as the dependent variable in the GLMM.

In our GLMM model examining the existence of any potential learning curve over the eight scenarios, we found that agreement between the raters and the reference standard did not significantly change from the early assessments to the later. After adjusting for scenario type, scoring method, and grading round, the estimated change in percent agreement from the 1st to the 8th scenarios was 0.2%. Figure 1 illustrates this lack of association, for example, among the non-expert raters’ “continuous mode” round 1 assessments.

Details of the level of agreement between raters and the reference standard, as well as kappa values (κ) summarizing agreement with the reference standard for rater responses on individual items, are provided as an online supplement in Table, Supplemental Digital Content 2, <http://links.lww.com/SIH/A31>. In summary, the overall average kappa between the raters and the reference standard was 0.78 across rounds, qualifying as excellent agreement according to Fleiss's criteria.²² Kappas were relatively consistent across experimental conditions, ranging only from 0.74 to 0.80 for continuous grading for first and second rounds (mean difference between rounds=0.01, $p=0.29$), and ranging from 0.74 to 0.81 for both rounds of the video grading with pauses (mean difference between rounds = 0.02, $p=0.17$). These findings again suggest that the high level of inter-observer reliability was not affected by scoring mode ("continuous," "with pauses"). They also illustrate that the level of agreement did not substantially improve with repeated evaluation of the same tapes. There was no significant difference between these values for evaluations conducted in continuous mode for round one, continuous mode in round 2, with pauses in round 1, and with pauses in round 2, wherein the average κ values only ranged from 0.77 to 0.78 (intra-rater reliability). When kappas were averaged across round and mode of evaluation (continuous vs. with pauses), the variation in agreement between the raters and the reference standard was minimal. In this analysis, the average κ values ranged from 0.76 to 0.80, indicating high agreement with the reference standard overall.

For the correct actions composite scores, test-retest reliability CCC values were high, averaging 0.96 across the four raters and ranging from 0.95 to 0.98. CCC values comparing raters' scores to the reference standard were also quite strong, averaging 0.87 and ranging from 0.83 to 0.94. Across the four raters and the reference standard, the ICC was 0.97 for the correct actions composite scores, indicating extremely low inter-rater variability. For the incorrect actions composite scores, test-retest reliability CCC values were moderately high, averaging 0.83 across the four raters and ranging from 0.79 to 0.86. CCC values comparing raters' incorrect actions composite scores to the reference standard were moderately strong, averaging 0.59 and ranging from 0.54 to 0.69. Across the four raters and the reference standard, the ICC was 0.92, again indicating low inter-rater variability.

As noted above, grading "with pauses" took about twice as long as "continuous" grading. We did not record the frequency with which raters changed their initial answers when using the "with pauses" mode. However, they reported that this occurred at a very low rate.

Discussion

This study was undertaken to assess the reliability of scores generated from ACLS checklists used as grading tools by non-expert raters during the review of simulations of AHA MegaCodes. Previous studies have shown that ACLS checklists can be used to make valid judgments for the purpose of developing minimum passing score criteria in competency testing.²⁻⁵ However, it was noted above that these studies left several questions unanswered if such checklists are to be employed for general use in the setting of ACLS training. The results of this present study illustrate that we have developed a set of checklists that yield reliable scores and that can be employed by non-expert raters with little variability in scores across raters and with a high level of agreement between non-expert and expert raters. The fact that non-expert, non-ACLS certified raters can produce reliable scores using these detailed checklists suggests that these checklists could produce reliable scores for AHA-certified ACLS instructors who would not be considered experts in education or research. Further studies will be needed to elucidate whether ACLS instructors who are already receiving instruction on the use of the AHA checklists could be taught to use this checklist in a similar amount of time. Taken together, these findings suggest that the checklists described in this study could be generalizable for widespread use in ACLS training without

substantially increasing the training needed for course instructors with respect to the evaluation of team leader performance during MegaCode management. Overall, this study has produced reliability evidence, content validity evidence, and some response-process evidence (raters are able to do this both with and without stops) for the checklists under consideration. In the future, instruments that yield reliable data and enable valid judgments of performance need to be developed for other skills addressed during ACLS courses, such as intubation.

The results of this study also show that a short training curriculum given to the non-expert raters was sufficient to properly instruct them in the use of the checklist, as no improvement in reliability of scores occurred with repeated use of the checklists during the testing phase (Figure 1). It should be noted that the level of agreement between the raters and the reference standard was higher for objective compared to subjective questions. This was a result of the range of communication possibilities of some items that make scoring these items as a binary "yes/no" more difficult. These checklist elements may never reach the same level of reliability in scoring as the measurement of discrete, objective actions. It is well recognized in the anesthesia literature that assessing non-technical skills, such as situational awareness and team communication, presents a unique challenge.²³ Until discrete communication language is considered a requirement for correct MegaCode management, such as clearly stating the precise words, "the patient is in $\times$ condition" or "let's evaluate all of the reversible causes of the arrest," the assessment of team leader communication and patient assessment may retain some element of subjectivity. Therefore, the nature of the checklist item could influence the grading performance for a rater, as objective items are less prone to variation in evaluation. Thus, it is possible this is an area in which training in the use of the checklists could be improved. Additional research will be needed to complete the validity assessment of these checklists so that they can be employed in summative assessments for high-stakes testing.

Additional work needs to be performed in order to determine the validity of these checklists, and to set standards for passing. Methods of setting minimum passing scores in ACLS have been previously described.⁵ Furthermore, prior studies have shown a link between simulation training and improved adherence to ACLS guidelines in the clinical setting.²⁴ However, this prior work did not show that there was improvement in patient outcome. This could be due to the fact that even in the group that received high-fidelity simulation (versus standard teaching), only 68% of actions taken adhered to published guidelines. Yet, studies continue to show that adherence to guidelines does affect patient outcomes.^{25,26} Therefore, future work will need to address the best methodologies for ensuring a high rate of adherence to guidelines. Additionally, this granularity in testing could be performed every few months, rather than at random, in order to ensure that providers who attend in-hospital cardiac arrests have demonstrated an ability to lead such an event properly.²⁷

Our results demonstrate a high level of agreement between raters, and show that overall inter-rater reliability was not affected by scoring mode. Additionally, we demonstrated high intra-rater reliability and excellent agreement between non-expert and expert rater evaluations. However, incorrect actions were scored with less reliability. The level of reliability in scoring incorrect items was still in the range of what is considered 'good.' However, this may remain a weakness of our checklist, as the correct actions were all arranged in a logical, temporal sequence, and thus it was probably easier for the rater to follow and score these items. On the other hand, the incorrect actions do not follow a logical sequence, and it is impossible to place them in a pre-determined order.

The major limitations of our findings have been discussed above. However, three further points deserve mention. First, it should be noted that these checklists were tested only for

team leader evaluation, not the team as a whole. Second, the non-expert raters recorded their assessment of team leader performance into an Excel® spreadsheet while watching a video recording on a computer screen. This may be a much simpler task (keeping the head still and looking between two computer screens) than watching a room of several participants in continuous action while also looking up and down at a computer interface in order to score performance. Operating the simulator user interface and simultaneously recording participant actions on an Excel spreadsheet is impractical. Therefore, if the content of these checklists are to be used in a real-time (i.e., continuous) manner, future research will need to be performed by testing them within a simulator software interface (see Figure 2). Third, the checklists used in this study were based on the 2005 AHA ACLS Update. However, the only change in the 2010 update that would affect the checklists would be the removal of atropine administration from the asystole and bradycardic PEA pathways. These changes would not substantively change the reliability assessment in this study, and the checklists could still be applicable for use under the new guidelines.

Conclusions

In conclusion, our findings advance the knowledge in this arena of educational research in at least four key aspects. First, we have produced reliable checklists that yield reliable scores for evaluating correct and incorrect team leader performance during the review of simulated ACLS MegaCodes. Second, faculty experienced in checklist design and ACLS instruction are *not* needed for the use of these checklists. Third, a short training curriculum on the proper use of the checklist is effective. Fourth, by extension of these first three points, these checklists are likely to be appropriate for use in continuous grading of team leader performance during ACLS MegaCode skills testing. Future studies need to address the feasibility of using these checklists within a simulator interface during live simulations and the generalizability of these checklists for widespread use in ACLS certifications. These two problems would best be addressed by a multi-site study where our checklists are used by ACLS instructors during MegaCode testing and compared to the grading checklists currently provided by the AHA. Overall, the results of this study indicate that one continuous evaluation (by a non-expert rater) of team leader performance during the review of an ACLS MegaCode simulation is adequate to achieve an accurate rating when employing this checklist.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to thank Dr. JG Reves and Dr. Scott T. Reeves for their assistance in manuscript review. The authors would also like to thank all of the staff at the MUSC Clinical Effectiveness and Patient Safety Center for their assistance during this project.

Funding Sources: FAER Research in Education Grant (7/2009–6/2011) and the National Center for Research Resources (UL1RR029882). The project described was supported in part by funding from the National Center for Research Resources. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Center for Research Resources or the National Institutes of Health.

Attribute work to: MUSC Department of Anesthesia and Perioperative Medicine and FAER Research in Education Grant

References

1. Wayne DB, Butter J, Cohen ER, McGaghie WC. Setting defensible standards for cardiac auscultation skills in medical students. *Acad Med.* 2009; 84:S94–S96. [PubMed: 19907398]
2. Wayne DB, Siddall VJ, Butter J, Fudala MJ, Wade LD, Feinglass J, McGaghie WC. A longitudinal study of internal medicine residents' retention of advanced cardiac life support skills. *Acad Med.* 2006; 81:S9–S12. [PubMed: 17001145]
3. Wayne DB, Butter J, Siddall VJ, Fudala MJ, Wade LD, Feinglass J, McGaghie WC. Graduating internal medicine residents' self-assessment and performance of advanced cardiac life support skills. *Med Teach.* 2006; 28:365–369. [PubMed: 16807178]
4. Wayne DB, Butter J, Siddall VJ, Fudala MJ, Wade LD, Feinglass J, McGaghie WC. Mastery learning of advanced cardiac life support skills by internal medicine residents using simulation technology and deliberate practice. *J Gen Intern Med.* 2006; 21:251–256. [PubMed: 16637824]
5. Wayne DB, Fudala MJ, Butter J, Siddall VJ, Feinglass J, Wade LD, McGaghie WC. Comparison of two standard-setting methods for advanced cardiac life support training. *Acad Med.* 2005; 80:S63–S66. [PubMed: 16199461]
6. Hatala R, Scalese RJ, Cole G, Bacchus M, Kassen B, Issenburg SB. Development and Validation of a Cardiac Findings Checklist for Use With Simulator-Based Assessments of Cardiac Physical Examination Competence. *Sim Healthcare.* 2009; 4:17–21.
7. Tudiver F, Rose D, Banks B, Pfortmiller D. Reliability and Validity Testing of an Evidence-based Medicine OSCE Station. *Fam Med.* 2009; 41:89–91. [PubMed: 19184682]
8. Morgan PJ, Cleave-Hogg D, Guest CB. A Comparison of Global Rating and Checklist Scores from an Undergraduate Assessment Using an Anesthesia Simulator. *Acad Med.* 2001; 76:1053–1055. [PubMed: 11597848]
9. Müller MJ, Dragicevic A. Standardized rater training for the Hamilton Depression Rating Scale (HAM-D-17) in psychiatric novices. *J Affect Disord.* 2003; 77:65–69. [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 non-expert raters. *Ann Surg.* 2009; 249:1047–1051. [PubMed: 19474694]
11. Weidner AC, Gimpel JR, Boulet JR, Solomon M. Using Standardized Patients to Assess the Communication Skills of Physicians for the Comprehensive Osteopathic Medical Licensing Examination (COMLEX) Level 2-Performance Evaluation. *Teach Learn Med.* 2010; 22:8–15. [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. *Simul Healthc.* 2008; 3:224–238. [PubMed: 19088667]
13. Zanetti M, Keller L, Mazor K, Carlin M, Alper E, Hatem D, Gammon W, Pugnaire M. Using standardized patients to assess professionalism: a generalizability study. *Teach Learn Med.* 2010; 22:274–279. [PubMed: 20936574]
14. <http://www.projectcheck.org/checklist-for-checklists.html>.
15. Stufflebeam DL. Guidelines for Checklist Development and Assessment. http://www.wmich.edu/evalctr/archive_checklists/guidelines.htm.
16. 2005 American Heart Association Guidelines for Cardiopulmonary Resuscitation and Emergency Cardiovascular Care. *Circulation.* 2005; 112 IV-1-149.
17. Morgan PJ, Lam-McCulloch J, Herold-McIlroy J, Tarshis J. Simulation performance checklist generation using the Delphi technique. *Can J Anaesth.* 2007; 54:992–997. [PubMed: 18056208]
18. Cronbach LJ, Nageswari R, Gleser GC. Theory of generalizability: A liberation of reliability theory. *The British Journal of Statistical Psychology.* 1963; 16:137–163.
19. Shavelson, Richard J.; Webb, Noreen M.; Rowley, Glenn L. Generalizability theory. *American Psychologist.* 1989; 44:922–932.
20. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin.* 1979; 86:420–428. [PubMed: 18839484]
21. Lin LI. A concordance correlation coefficient to evaluate reproducibility. *Biometrics.* 1989; 45:255–268. [PubMed: 2720055]

22. Fleiss, JL. Statistical methods for rates and proportions. 2nd ed.. New York: John Wiley; 1981.
23. Fletcher G, Flin R, McGeorge P, Glavin R, Maran N, Patey R. Anaesthetists' Non-Technical Skills (ANTS): evaluation of a behavioural marker system. *Br J Anaesth*. 2003; 90:580–588. [PubMed: 12697584]
24. 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. [PubMed: 17573509]
25. Chan PS, Krumholz HM, Nichol G, Nallamothu BK. Delayed time to defibrillation after in-hospital cardiac arrest. *N Engl J Med*. 2008; 358:9–17. [PubMed: 18172170]
26. Myhre JM, Ramachandran SK, Kheterpal S, Morris M, Chan PS. for the American Heart Association National Registry for Cardiopulmonary Resuscitation Investigators. Delayed time to defibrillation after intraoperative and periprocedural cardiac arrest. *Anesthesiology*. 2010; 113:782–793. [PubMed: 20808215]
27. Andreatta P, Saxton E, Thompson M, Annich G. Simulation-based mock codes significantly correlate with improved pediatric patient cardiopulmonary arrest survival rates. *Pediatr Crit Care Med*. 2011; 12:33–38. [PubMed: 20581734]

Appendix 1

Unstable Brady/VFIB/Asystole

	Critical?	Objective/ Subjective
Universal Management of Initial Presentation		
Primary Survey - BLS		
Assessed patient responsiveness	No	Obj
Checked airway patency	Yes	Obj
Assessed breathing	Yes	Obj
Assessed circulation/pulse	Yes	Obj
If apneic, gave two rescue breaths	No	Obj
Communication		
If pulseless, called for help (Code/May Day/Activate Emergency Response System) and AED	Yes	Obj
Introduced oneself as Team Leader (if others present)	No	Obj
Assigned Team Member Roles to Each Person (monitors, CPR, etc)	No	Subj
Initial Management/Monitoring		
If pulseless, commenced CPR (30:2)	Yes	Obj
Attached ECG leads	No	Obj
Attached defibrillator	Yes	Obj
Placed pulse oximeter	No	Obj
Measured BP (if patient has pulse)	Yes	Obj
Attached facemask O2 if breathing or BMV if apneic (>10L/min if SaO2<93%)	Yes	Obj
Attached facemask or NC O2 (2–6L/min) if SaO2 93–96%	No	Obj
Assessed Instability (SOB, hypotension, AMS, angina, etc.)	Yes	Subj
Asked for recent labs	No	Obj
Asked for past medical history	No	Obj
Intravascular Access		

	Critical?	Objective/ Subjective
Checked existing IV function or started IV if absent	Yes	Obj
Possible Incorrect Actions		
Any action prior to Primary BLS survey (if first responder)	No	Subj
Unstable Bradycardia		
Correct Actions		
Assessed rhythm, pulse, and stability.	Yes	Subj
Stated unstable brady dx	No	Obj
Gave Atropine 0.5mg IV	Yes	Obj
Considered transcutaneous pacing (TCP)	Yes	Obj
Considered IV chronotropic support (dopamine, epinephrine, etc) if atropine not effective	Yes	Obj
Consulted cardiology	No	Obj
Considered transvenous pacer for 2°/3° AV block	No	Obj
Considered H's and T's	No	Subj
Ordered ABG	No	Obj
Ordered other labs (BMP, CMP, CBC, etc)	No	Obj
Gave IV fluid bolus if hypotensive	No	Obj
Possible Incorrect Actions		
Gave amiodarone	No	Obj
Gave lidocaine	No	Obj
Gave CCB or Beta-blocker	Yes	Obj
Gave vasopressin	No	Obj
Gave phenylephrine	No	Obj
Gave nitroglycerin	Yes	Obj
Gave wrong dose of a drug in this pathway	No	Obj
Gave other wrong drug for this pathway	No	Obj
Incorrectly performed TCP	No	Subj
Delivered NON-SYNCEd DC-CV	Yes	Obj
Delivered SYNCEd DC-CV	Yes	Obj
Ventricular Fibrillation (VFIB)/Pulseless Ventricular Tachycardia (VTACH)		
Correct Actions		
Assessed rhythm, pulse, and stability (N/A if presenting rhythm)	Yes	Subj
Stated diagnosis of VFIB	No	Obj
Started CPR (30:2) until defibrillator attached	Yes	Obj
Declared all clear prior to defib	No	Obj
Immediately defibrillated @ 150–200J biphasic after rhythm recognition	Yes	Obj
Resumed CPR × 5 cycles (30:2)	Yes	Obj
Rhythm and pulse checked (<10 sec pause)	No	Obj
Declared 'all clear' prior to defibrillation	No	Obj
Defibrillated at 200J biphasic - SHOCK 2	Yes	Obj

	Critical?	Objective/ Subjective
Resumed CPR × 5 cycles (30:2)	Yes	Obj
Gave Epi 1mg or Vaso 40U IV	Yes	Obj
Rhythm and pulse checked (<10 sec pause)	No	Obj
Declared 'all clear' prior to defibrillation	No	Obj
Defibrillated at 200J biphasic - SHOCK 3	Yes	Obj
Resumed CPR × 5 cycles (30:2)	Yes	Obj
Gave Amiodarone 300mg or Lidocaine 1.5mg/kg IV	Yes	Obj
Considered H's and T's	No	Subj
Ordered ABG	No	Obj
Possible Incorrect Actions		
Gave any drug prior to Shock 2	Yes	Obj
Gave wrong dose of a drug in this pathway (Epi or Vaso, Amio or Lido)	No	Obj
Gave wrong sequence of Epi/Vaso	No	Obj
Gave wrong sequence of Amiodarone/Lidocaine	No	Obj
Gave Atropine (any dose)	No	Obj
Gave other wrong drug for this pathway	No	Obj
>1 min from rhythm recognition to 1 st shock	No	Obj
>2 min from rhythm recognition to 1 st shock	No	Obj
>3 min from rhythm recognition to 1 st shock	Yes	Obj
Gave shock <150 j on biphasic	Yes	Obj
Gave shock >200 j on biphasic	No	Obj
Asystole		
Correct Actions		
Assessed rhythm, pulse, and stability.	Yes	Subj
Confirmed asystole in 2 leads	No	Obj
Stated diagnosis of asystole	No	Obj
Resumed CPR × 5 cycles (30:2)	Yes	Obj
Gave Epi 1mg or Vaso 40U IV	Yes	Obj
Gave Atropine 1mg IV	No	Obj
Rhythm and pulse checked (<10 sec pause)	No	Obj
Resumed CPR × 5 cycles (30:2)	Yes	Obj
Rhythm and pulse checked (<10 sec pause)	No	Obj
Considered H's and T's and other Diff Dx	Yes	Subj
Ordered ABG	No	Obj
Possible Incorrect Actions		
Gave amiodarone in this pathway	No	Obj
Gave lidocaine in this pathway	No	Obj
Administered shock(s) in this pathway	Yes	Obj
Gave the wrong dose of a drug indicated in this pathway (i.e. Atropine 2mg IV)	No	Obj

	Critical?	Objective/ Subjective
Gave adenosine, beta-blockers, or calcium channel blockers in this pathway	No	Obj
Gave other wrong drug for this pathway	No	Obj
Assessment of CPR (Every 5 cycles)		
Correct Actions		
Place backboard under patient	Yes	Obj
30:2 ratio	Yes	Obj
Rate of 100–120 compressions/minute verified or corrected if improper	Yes	Subj
Compression depth 1.5–2" verified or corrected if improper.	Yes	Obj
Hand placement: midline, lower half of sternum verified or corrected if improper	Yes	Subj
Possible Incorrect Actions		
CPR not started within 60 sec of recognizing pulseless state	No	Obj
CPR not started within 180 sec of recognizing pulseless state	Yes	Obj
CPR delayed for >10 seconds at pulse & rhythm check	No	Obj
Airway Management		
Correct Actions		
Confirmed chest rise with BMV	Yes	Obj
Requested ETT placement if BMV insufficient	Yes	Obj
Requested ETT placement for secure airway prior to transport, if not already present	No	Obj
Possible Incorrect Actions		
CPR delayed for ETT placement:> 15 sec	No	Obj
CPR delayed for ETT placement:> 30 sec	No	Obj
CPR delayed for ETT placement:> 60 sec	Yes	Obj

Appendix 2

Stable Tachycardia/Pulseless VTACH/PEA

	Critical?	Objective/ Subjective
Universal Management of Initial Presentation		
Primary Survey - BLS		
Assessed patient responsiveness	No	Obj
Checked airway patency	Yes	Obj
Assessed breathing	Yes	Obj
Assessed circulation/pulse	Yes	Obj
If apneic, gave two rescue breaths	No	Obj
Communication		
If pulseless, called for help (Code/May Day/Activate Emergency Response System) and AED	Yes	Obj
Introduced oneself as Team Leader (if others present)	No	Obj

	Critical?	Objective/ Subjective
Assigned Team Member Roles to Each Person (monitors, CPR, etc)	No	Subj
Initial Management/Monitoring		
If pulseless, commenced CPR (30:2)	Yes	Obj
Attached ECG leads	No	Obj
Attached defibrillator	Yes	Obj
Placed pulse oximeter	No	Obj
Measured BP (if patient has pulse)	Yes	Obj
Attached facemask O2 if breathing or BMV if apneic (>10L/min if SaO2<93%)	Yes	Obj
Attached facemask or NC O2 (2–6L/min) if SaO2 93–96%	No	Obj
Assessed Instability (SOB, hypotension, AMS, angina, etc.)	Yes	Subj
Asked for recent labs	No	Obj
Asked for past medical history	No	Obj
Intravascular Access		
Checked existing IV function or started IV if absent	Yes	Obj
Possible Incorrect Actions		
Any action prior to Primary BLS survey (if first responder)	No	Subj
Stable Tachycardia		
Correct Actions		
Assessed rhythm, pulse, and stability.	Yes	Subj
Stated stable tachycardia as dx	No	Obj
Classified as wide or narrow complex	Yes	Obj
Classified as regular or irregular	Yes	Obj
Obtained cardiology consult	No	Obj
Obtained 12-lead ECG	No	Obj
Ordered ABG & BMP or Long Panel	No	Obj
Considered H's & T's	No	Subj
Gave IV fluid bolus if hypotensive	No	Obj
Narrow Complex, Regular Rhythm		
Vagal Maneuver (s)	No	Obj
If vagal maneuver(s) failed, gave Adenosine 6mg IV (1st dose)	Yes	Obj
Adenosine 12mg IV 2nd dose	No	Obj
Adenosine 12mg IV 3rd dose	No	Obj
If Adenosine failed, gave β -blocker or CCB	Yes	Obj
Wide Complex, Regular Rhythm		
Considered VTACH v. SVT with aberrancy	No	Obj
If stable VTACH, gave amiodarone 150mg IV over 10 minutes	Yes	Obj
If amiodarone failed to convert VTACH, prepared for SYNC DCCV	Yes	Obj
If wide complex SVT with aberrancy, gave Adenosine 6mg IV (repeated 6mg, 12mg IV as indicated)	No	Obj

	Critical?	Objective/ Subjective
Narrow Complex, Irregular Rhythm		
Considered Afib vs Aflutter vs MAT	No	Obj
If Afib/Aflutter/MAT, gave β -blocker or CCB for rate control	Yes	Obj
Wide Complex, Irregular Rhythm		
Considered expert consultation	No	Obj
If stable Torsades de Pointes, gave MgSO ₄ , 1–2 Gm IV over >5 min	No	Obj
If Afib w/ aberrancy, gave β -blocker or CCB for rate control	No	Obj
If Afib + WPW, gave amiodarone 150mg IV over 10 mins	No	Obj
Possible Incorrect Actions		
Failed to listen for carotid bruit if carotid massage performed	No	Obj
Bilateral simultaneous carotid massage	No	Obj
Administered Synced DCCV to stable tachycardia other than sustained		
VTACH	No	Obj
Administered non-Synced shock to stable tachycardia	Yes	Obj
Gave incorrect dose of Adenosine	No	Obj
Gave Atropine	No	Obj
Gave Amiodarone	No	Obj
Gave other incorrect drug	No	Obj
Ventricular Fibrillation (VFIB)/Pulseless Ventricular Tachycardia (VTACH)		
Correct Actions		
Assessed rhythm, pulse, and stability (N/A if presenting rhythm)	Yes	Subj
Stated diagnosis of VFIB/pulseless VTACH	No	Obj
Started CPR (30:2) until defibrillator attached	Yes	Obj
Declared all clear prior to defib	No	Obj
Immediately defibrillated 150–200J biphasic after rhythm recognition	Yes	Obj
Resumed CPR $\times$ 5 cycles (30:2)	Yes	Obj
Rhythm and pulse checked (<10 sec pause)	No	Obj
Declared 'all clear' prior to defibrillation	No	Obj
Defibrillated at 200J biphasic - SHOCK 2	Yes	Obj
Resumed CPR $\times$ 5 cycles (30:2)	Yes	Obj
Gave Epi 1mg or Vaso 40U IV	Yes	Obj
Rhythm and pulse checked (<10 sec pause)	No	Obj
Declared 'all clear' prior to defibrillation	No	Obj
Defibrillated at 200J biphasic - SHOCK 3	Yes	Obj
Resumed CPR $\times$ 5 cycles (30:2)	Yes	Obj
Gave Amiodarone 300mg or Lidocaine 1.5mg/kg IV	Yes	Obj
Considered H's and T's	No	Subj
Ordered ABG	No	Obj
Possible Incorrect Actions		

	Critical?	Objective/ Subjective
Gave any drug prior to Shock 2	Yes	Obj
Gave wrong dose of a drug in this pathway (Epi or Vaso, Amio or Lido)	No	Obj
Gave wrong sequence of Epi/Vaso	No	Obj
Gave wrong sequence of Amiodarone/Lidocaine	No	Obj
Gave Atropine (any dose)	No	Obj
Gave other wrong drug for this pathway	No	Obj
>1 min from rhythm recognition to 1 st shock	No	Obj
>2 min from rhythm recognition to 1 st shock	No	Obj
>3 min from rhythm recognition to 1 st shock	Yes	Obj
Gave shock <150 j on biphasic	Yes	Obj
Gave shock >200 j on biphasic	No	Obj
PEA		
Correct Actions		
Assessed rhythm, pulse, and stability.	Yes	Subj
Started/continued CPR (30:2)×5 cycles	Yes	Obj
Stated PEA as dx	No	Obj
Pulse & rhythm check <10s	No	Obj
Resumed CPR×5 cycles	Yes	Obj
Gave epinephrine 1mg IV	Yes	Obj
Gave Atropine 1mg IV if HR<60bpm	No	Obj
Considered H's & T's	Yes	Subj
Possible Incorrect Actions		
Gave atropine if HR >60bpm	No	Obj
Gave amiodarone in this pathway	No	Obj
Gave lidocaine in this pathway	No	Obj
Administered shock(s) in this pathway	Yes	Obj
Gave the wrong dose of a drug indicated in this pathway (i.e. Atropine 2mg IV)	No	Obj
Gave adenosine, beta-blockers, or calcium channel blockers in this pathway	No	Obj
Gave other wrong drug for this pathway	No	Obj
Assessment of CPR (Every 5 cycles)		
Correct Actions		
Place backboard under patient	Yes	Obj
30:2 ratio	Yes	Obj
Rate of 100–120 compressions/minute verified or corrected if improper	Yes	Subj
Compression depth 1.5–2" verified or corrected if improper.	Yes	Subj
Hand placement: midline, lower half of sternum verified or corrected if improper	Yes	Subj
Possible Incorrect Actions		
CPR not started within 60 sec of recognizing pulseless state	No	Obj
CPR not started within 180 sec of recognizing pulseless state	Yes	Obj

	Critical?	Objective/ Subjective
CPR delayed for >10 seconds at pulse & rhythm check	No	Obj
Airway Management		
Correct Actions		
Confirmed chest rise with BMV	Yes	Obj
Requested ETT placement if BMV insufficient	Yes	Obj
Requested ETT placement for secure airway prior to transport, if not already present	No	Obj
Possible Incorrect Actions		
CPR delayed for ETT placement:>15 sec	No	Obj
CPR delayed for ETT placement:>30 sec	No	Obj
CPR delayed for ETT placement:>60 sec	Yes	Obj

Appendix 3

Unstable Bradycardia/Pulseless VTACH/Asystole

Universal Management of Initial Presentation
See Appendix 1 for list of items in this section
Unstable Bradycardia
See Appendix 1 for list of items in this section
Pulseless Ventricular Tachycardia (VTACH)
See Appendix 2 for list of items in this section with correct diagnosis being 'Pulseless VTACH'
Asystole
See Appendix 1 for list of items in this section
Assessment of CPR (Every 5 cycles)
See Appendix 1 for list of items in this section
Airway Management
See Appendix 1 for list of items in this section

Appendix 4

Stable Tachycardia/VFIB/PEA

Universal Management of Initial Presentation
See Appendix 1 for list of items in this section
stable Tachycardia
See Appendix 2 for list of items in this section
Ventricular Fibrillation (VFIB)
See Appendix 1 for list of items in this section with correct diagnosis being 'Ventricular Fibrillation'
Pulseless Electrical Activity (PEA)
See Appendix 2 for list of items in this section

Assessment of CPR (Every 5 cycles)
See Appendix 1 for list of items in this section
Airway Management
See Appendix 1 for list of items in this section

Appendix 5

Use of generalized linear mixed models (GLMMs) to assess inter-rater and intra-rater reliability

There are several statistical methods for assessing the impact of various factors on reliability of a given measurement; generalizability theory (or “G theory”) is a particularly useful for examining the impact of multiple sources of variation and was introduced in 1963 (Cronbach 1963). G theory uses an analysis of variance (ANOVA) approach to quantify the degree to which individual factors (e.g. raters, time, experimental setting) affect measurement reliability. While this type of approach is ideal for continuous measurements, some adaptations are necessary for binary observations, such as the individual checklist item responses summarized in the current paper. To study the impact of item characteristics on inter-rater and intra-rater reliability of the scoring checklist, we employed a generalized linear mixed model (GLMM) approach with a binary response and a logit link function. The GLMM we used is similar to a logistic regression model; however the traditional logistic regression model assumes independent errors across observations, whereas our GLMM allowed for correlated errors within observations made by the same rater. Our model can be expressed as:

$$\begin{aligned}
 \ln\left(\frac{p_{ij}}{1 - p_{ij}}\right) &= \alpha + \sum_{k=1}^7 \beta_k \cdot \text{Scenario}_k \\
 &+ \beta_8 \cdot \text{Continuous} \\
 &+ \beta_9 \cdot \text{Round} \\
 &+ \beta_{10} \cdot \text{CorrectAction} \\
 &+ \beta_{11} \cdot \text{SubjectiveItem} \\
 &+ \beta_{12} \cdot \text{ActionPerformed} \\
 &+ \beta_{13} \cdot \text{ActionEssential} \\
 &+ \beta_{14} \cdot \text{Order} + \gamma_j + \varepsilon_{ij}
 \end{aligned}$$

In the model above, p_{ij} represents the probability of agreement with the reference standard on item i by rater j ; α is the estimated overall intercept; the β 's are estimated regression parameters for the specified fixed effect variables of interest (see descriptors in Table 1), γ_j is a random intercept for reader j , and ε_{ij} is an error term for the i^{th} item as assessed by the j^{th} rater. In this GLMM we assumed that variation in agreement with the reference standard had a compound symmetry error structure, meaning that assessments made by the same rater being assumed to be correlated with one another and assessments made by different raters being assumed to be independent from one another. We constructed our GLMM using SAS v9.2 (Cary, NC) PROC GLIMMIX.

The GLMM is an ideal model for simultaneously studying a number of sources of variation in raters' agreement with the reference standard. By being able to incorporate *random* rater effects rather than *fixed* rater effects (as would be the case using an ANOVA approach), inference about reliability can be made about a population of raters, not just the specific raters who provided assessments. In other words, the GLMM approach provides greater generalizability than the ANOVA approach. Additionally, a likelihood ratio test (using the COVTEST statement) can be used to test whether or not the variance component associated with the random rater effect is zero, providing a means for formally assessing inter-rater reliability. Since the "Round" variable distinguishes whether a rater's assessments were made during Round 1 vs. Round 2, its parameter estimate and associated standard error provide a means for assessing intra-rater reliability over time. Since other factors (e.g. questionnaire item characteristics, rater's viewing order) may also contribute to the rater's agreement with the reference standard, they are easily added to the model as fixed effects, and F-tests based on the associated parameters and standard errors are conducted by default. Working with GLMMs generally requires knowledge of general linear models (e.g. linear regression, ANOVA) but also requires additional knowledge of more advance statistical topics such as link functions, error covariance structures, and numerical quadrature; thus collaboration with a biostatistician is recommended. McCulloch & Searle (2001) provide an excellent discussion of these topics.

Cronbach, L.J., Nageswari, R., & Gleser, G.C. (1963). Theory of generalizability: A liberation of reliability theory. *The British Journal of Statistical Psychology*, 16, 137–163.

McCulloch CE & Searle SR (2001). *Generalized, Linear, and Mixed Models*. New York: John Wiley & Sons, Inc.

Percent Agreement by Grading Order with Stratification by Rater

Figure 1.

This figure shows the agreement with the reference standard by rater and by the order in which the video was graded during the first round of grading. The 8 videos were graded in different orders by each non-expert rater. As this figure represents the agreement with the reference standard during the first round of ‘continuous’ grading, it shows that no significant learning curve existed after the checklist and video training program was complete.

Figure 2.

An example screenshot from the SimMan software interface that will need to be tested in the future in order to determine if these checklists can be reliably used in this format while simulations are actually occurring.

Table 1

This table shows the order in which each video was graded by round for Rater 1. Each video was assigned a letter A through H. The orders were determined by random order. A similar grading scheme was randomly generated for each rater.

Example of Grading Template for Novice Raters	
	Order of Video Evaluation
Round 1 – Real-Time	F C E G H A D B
Round 2 – Real-Time	G A F B E H C D
Round 3 – With Stops	H C F G D A B E
Round 4 – With Stops	A H D F G C B E

* Videos were assigned letters A–G in grading template.

Table 2

Factors associated with agreement/disagreement with the reference standard on individual checklist items: unadjusted agreement levels and p-values from the GLMM model reflecting each factors independent association with agreement. P-values reflect significant variation across categories within factors. The model also included a factor reflecting viewing order, which did not exhibit significant variation (p=0.81)

Factor	Unadjusted percent agreement with reference standard	P-value from GLMM model
Scenario		<0.0001
A	83.31%	
B	91.43%	
C	91.18%	
D	89.47%	
E	89.27%	
F	88.50%	
G	88.61%	
H	89.90%	
Stopping allowed	89.00%	0.85
Stopping not allowed (i.e. on the fly)	88.89%	
Round 1	88.64%	0.27
Round 2	89.26%	
Correct action	85.57%	<0.0001
Incorrect action	94.68%	
Subjective item	74.31%	<0.0001
Objective item	90.81%	
Action performed (as assessed by reference standard)	87.04%	0.29
Action not performed (as assessed by reference standard)	90.49%	
Action is "Essential"	90.36%	<0.0001
Action is not "Essential"	87.86%	
Rater 1	88.09%	0.032
Rater 2	89.81%	
Rater 3	89.88%	
Rater 4	88.02%	