Five Steps to Develop Checklists for Evaluating Clinical Performance: An Integrative Approach

Jan Schmutz, MSc, Walter J. Eppich, MD, MEd, Florian Hoffmann, MD, Ellen Heimberg, MD, and Tanja Manser, PhD

Abstract

Purpose
The process of developing checklists to rate clinical performance is essential for ensuring their quality; thus, the authors applied an integrative approach for designing checklists that evaluate clinical performance.

Method
The approach consisted of five predefined steps (taken 2012–2013). *Step 1:* On the basis of the relevant literature and their clinical experience, the authors drafted a preliminary checklist. *Step 2:* The authors sent the draft checklist to five experts who reviewed it using an adapted Delphi technique. *Step 3:* The authors devised three scoring categories for items after pilot testing. *Step 4:* To ensure the changes made after pilot testing were valid, the checklist was submitted to an additional Delphi review round. *Step 5:* To weight items needed for accurate performance assessment, 10 pediatricians rated all checklist items in terms of their importance on a scale from 1 (not important) to 5 (essential).

Results
The authors have illustrated their approach using the example of a checklist for a simulation scenario of infant septic shock. The five-step approach resulted in a valid, reliable tool and proved to be an effective method to design evaluation checklists. It resulted in 33 items, most consisting of three scoring categories.

Conclusions
This approach integrates published evidence and the knowledge of domain experts. A robust development process is a necessary prerequisite of valid performance checklists. Establishing a widely recognized standard for developing evaluation checklists will likely support the design of appropriate measurement tools and move the field of performance assessment in health care forward.

Editor's Note: A commentary by M.A. Rosen and P.J. Pronovost appears on pages 963–965.

Assessing clinical performance in health care is important for many reasons.[1] Doing so helps to characterize the abilities of clinicians and identify potential performance gaps. Further, assessing performance augments debriefings and forms the basis of scientific studies investigating factors influencing clinical performance. Because performance is a complex concept[2] and no single "best" performance measure exists, the reliable and valid assessment of performance is challenging for educators and researchers alike.

Please see the end of this article for information about the authors.

Correspondence should be addressed to Mr. Schmutz, University of Fribourg, Department of Psychology, Industrial Psychology and Human Factors, Rue de Faucigny 2, 1700 Fribourg, Switzerland; telephone: (+41) 26-300-7484; e-mail: j@n-schmutz.ch.

Acad Med. 2014;89:996–1005.
First published online May 13, 2014
doi: 10.1097/ACM.0000000000000289

Systematic Performance Assessment

Within organizational psychology, performance is viewed as a multidimensional concept that comprises a process and an outcome component.[3,4] *Process performance* refers to what an individual or a team *does* in the work situation (e.g., performing a treatment task), whereas *outcome performance* refers to the *result* of this behavior (e.g., treatment-related patient outcomes).[2,4] Outcome performance measures, such as infection rate or mortality, can be assessed objectively but cannot always be directly attributed to clinical performance. For example, a patient might die despite a team's optimal performance in the resuscitation. Furthermore, in training settings, educators focus on correct clinical behaviors because outcomes are often not available and feedback on performance can modify trainees' behaviors. Thus, measures of process performance play a central role when assessing a trainee's clinical competence.

Process performance can be assessed by subjective and objective measures.[1] *Subjective measures*, which include global expert ratings of specific behavioral aspects or of overall performance, are mainly based on the clinical expertise of the rater. *Objective measures* are based on predefined scoring categories in the form of listed key items (i.e., evaluation checklists).[5]

This report focuses on checklists for evaluating clinical performance rather than on checklists supporting procedural task execution (e.g., central line placement). Checklists for evaluating clinical performance are structured tools outlining criteria to consider for a specific process.[6] They ensure that the assessment includes all important tasks during a process and, thus, force the rater to focus on predefined items. In the evaluation process, defining the specific criteria for the evaluation is crucial.[7] These criteria help to reduce observation biases (e.g., halo effect, confirmation bias),[8] and they can increase reliability among different evaluators.[9,10]

A classic evaluation checklist uses simple dichotomous items (done/not done). Because dichotomous items are frequently not sufficient for the assessment of complex tasks, this format has been extended to include more categories (e.g., done/done incorrectly/not done).[11–13] Other investigators have weighted checklist items to differentiate between essential

and less important actions.[14] A few have defined specific actions as mandatory, which renders a total performance score of zero when the mandatory actions are not executed even if other actions are performed correctly.[15] One frequent criticism of evaluation checklists is that they reward thoroughness without considering the timeliness of actions.[16,17] Some researchers have acknowledged this by integrating time frames.[11,18] Factors such as weighting and time frames help create a more refined assessment of performance and should thus be considered in the development of future evaluation checklists.

Developing Checklists for Evaluating Clinical Performance

The development process of an evaluation checklist affects its quality.[19] The literature provides methodological recommendations for developing effective evaluation checklists: They should be based on (1) professional experience,[6,19] (2) primary literature sources or peer-reviewed guidelines,[19] and (3) the consensus of experts in the field of interest.[8,19] Table 1 provides examples of checklists, all of which have incorporated methodological recommendations from the literature and most of which also relied on expert opinion.

Not all studies in Table 1 include a description of a structured procedure for checklist development (i.e., defining a series of steps to be completed), nor do they all follow an overall systematic approach (i.e., defining guidelines or criteria for each of those steps). In fact, because the main focus of these studies is the evaluation of the checklist itself, only a few of them explicitly describe a structured and systematic approach to the checklist's development.[14,20,21] The later steps in the development process, such as weighting checklist items and integrating feedback from pilot testing, seem especially underemphasized (see Table 1).

Given the state of research and current practice in checklist development, we believe that researchers need a clear outline of the methodological steps to develop checklists for evaluating clinical performance, an outline that integrates existing recommendations into a more comprehensive approach. This integrated approach will support researchers in evaluating the suitability of checklists for different contexts, in designing performance assessment tools for specific clinical scenarios that reflect precisely what the task demands of the clinicians, and in either adapting existing checklists or generating new ones.

The aim of this study is to examine such an integrative approach in the development of checklists for evaluating clinical performance. Using the example of a simulated sepsis scenario, we have applied a five-step approach to checklist development that includes an adapted Delphi process and yields more than a classical dichotomous checklist by integrating timeliness and weights indicating the importance of different actions. In doing so we have used existing guidelines and methods and have integrated them into a comprehensive development process.

Method

This study was exempt from ethics review, per Swiss law. Figure 1 outlines the five steps of our systematic approach for the development of performance checklists. We developed and tested the five-step approach between May 2012 and June 2013.

Step 1: Development of a draft checklist

We developed an evaluation checklist for the simulated scenario of septic shock in a six-month-old boy. Three experienced acute care pediatricians and simulation educators (E.H., F.H., and W.J.E.) drafted an initial checklist of critical treatment tasks for this scenario based on published European Resuscitation Council guidelines,[22] the literature, and their own clinical experience.

Step 2: Delphi review rounds

We sent the draft checklist to five experts, whom we had chosen on the basis of established selection criteria, for review. The experts used an adapted Delphi technique to review the draft checklist. The Delphi technique is a consensus-based method through which experts respond to questionnaires and receive anonymous group feedback.[23] The main advantage of this method is the application of "collective intelligence," which is the combined ability of group members to jointly produce better results than anyone in the group could produce on his or her own.[24] The procedure consists of multiple review rounds until consensus is achieved. When Delphi rounds are conducted by mail or e-mail, the process is anonymous, allowing each expert to make suggestions without fear of losing face. The anonymity also reduces the impact of common group biases like conformity or power influences.[23] The Delphi technique is well established in social science and increasingly used in health care research for various purposes.[25–27]

Selection of experts. Recommended sample sizes for experts for a Delphi study range from 5 to 30 depending on the research question.[28] In line with these recommendations, we felt a sample of 5 experts would be sufficient because the treatment of septic shock mostly follows established, standardized algorithms. The validity of the Delphi technique depends strongly on the selection of the experts; thus, we required all experts to be board-certified physicians with at least 10 years of clinical practice after medical school including at least 6 years in pediatric care.

Procedure. Figure 1 provides a visual representation of the Delphi review rounds. Experts received the draft checklist as well as a short history of the simulated patient and the current sepsis scenario (Box 1) by e-mail. We instructed the experts not only to delete irrelevant actions, add missing but relevant actions, or reformulate already-listed items but also to include a comment explaining all additions, deletions, and reformulations as information for all experts in the next review round. After the first round, all expert feedback was integrated into one modified list, and the source of all edits was deidentified. All suggestions were clearly highlighted.

In round 2, we asked participants to confirm whether or not an added item should remain in the list and if they agreed with the proposed deletions. If a majority (three of five experts) recommended an addition or deletion, we included the change. We repeated this procedure until the experts achieved a consensus and made no more suggestions.

Step 3: Design of the final checklist and pilot testing

In the third step, three clinicians, including two of us (E.H. and F.H.),

Table 1

Characteristics of Procedures Used for Designing Performance Checklists in the Literature

Study	Scenario evaluated by checklist	Literature- and/or guideline based	Expert opinion based	Structured development process[a]	Overall systematic approach[b]
Chopra et al, 1994[43]	-Anaphylactic shock -Malignant hyperthermia	✓			
Gaba et al, 1998[15]	-Malignant hyperthermia -Cardiac arrest	✓	✓		
Lockyer et al, 2006[20]	-Neonatal resuscitation	✓	✓	✓	✓
Scavone et al, 2006[21]	-General anesthesia for emergency cesarean delivery	✓	✓ (Delphi)	✓	✓
Thomas et al, 2006[44]	-Neonatal resuscitation	✓			
Tschan et al, 2006[45]	-Cardiac arrest	✓			
Adler et al, 2007[46]	-Apnea -Asthma -Supraventricular tachycardia -Sepsis	✓	✓	✓ (CDC)[19]	
Morgan et al, 2007[14]	Anesthesia induction for patient with: -Laparoscopic cholecystectomy -Laparotomy for large bowel obstruction	✓	✓ (Delphi)	✓	✓
Brett-Fleegler et al, 2008[47]	-Near-drowning child -Child with asthma -Child with tricyclic antidepressant overdose	✓	✓ (Delphi and development session)	✓	
Adler et al, 2009[48]	-Infant in shock -Tachycardia -Altered mental status -Trauma	✓	✓	✓ (CDC)[19]	
Carlson et al, 2009[49]	-Acute dyspnea	✓			
Donoghue et al, 2009[11]	-Asystole -Tachydysrhythmia -Respiratory arrest -Shock	✓	✓	✓	✓ (Approach by Lockyer)[20]
Manser et al, 2009[50]	-Anesthesia induction in malignant hyperthermia	✓			
Burtscher et al, 2010[51]	-Standard anesthesia induction	✓	✓ (Delphi)	✓	
Donoghue et al, 2010[18]	-Asystole -Dysrhythmia -Respiratory arrest -Shock	✓	✓	✓	✓ (Approach by Lockyer)[20]
Westli et al, 2010[52]	-Trauma	✓			
Adler et al, 2011[16]	-Shock -Unexplained altered mental status -Multisystem trauma	✓	✓	✓ (CDC)[19]	
Burtscher et al, 2011[40]	-General anesthesia induction	✓	✓ (Delphi)	✓	
Lambden et al, 2013[32]	-Respiratory failure -Sepsis -Meningitis with raised intracranial pressure	✓	✓		

Abbreviations: CDC indicates Checklists Development Checklist.

[a]Following a well-predefined process.

[b]Methodical approach predefined and replicable through a step-by-step procedure.

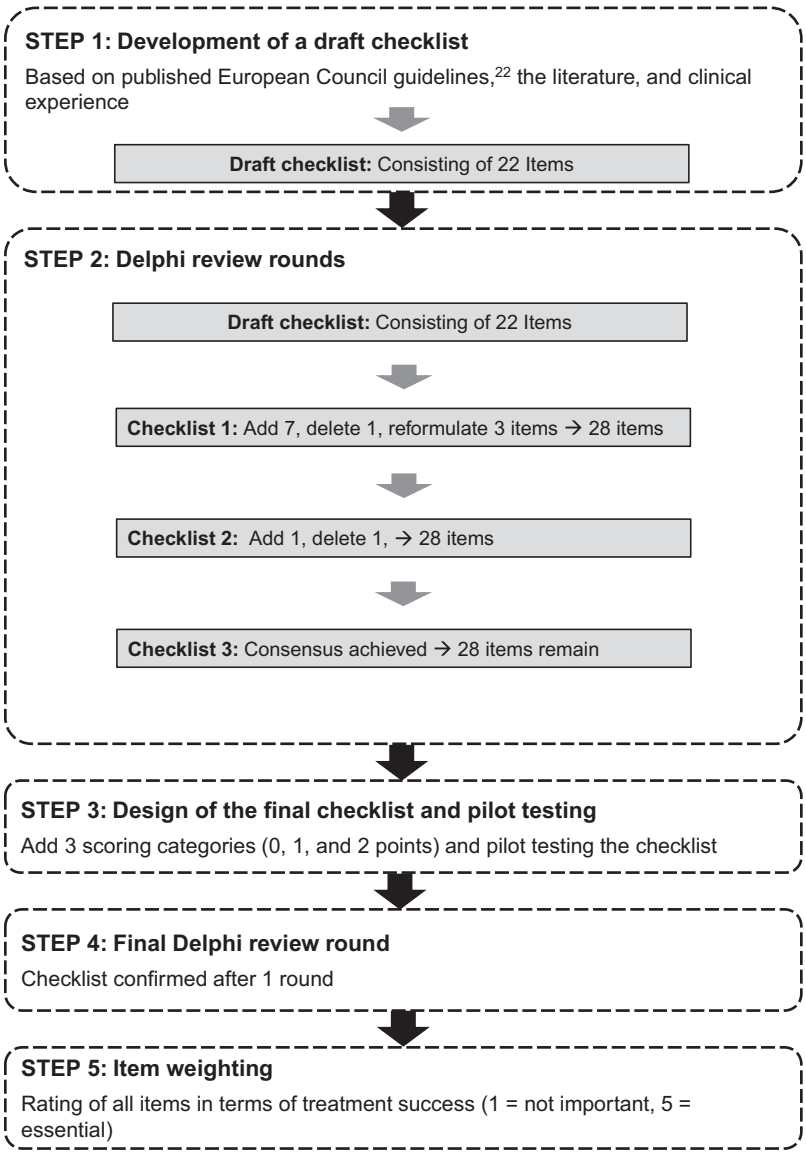

Figure 1 The five steps to develop checklists for evaluating clinical performance.

pilot tested the checklist by rating the videotaped management of six simulated pediatric septic shock scenarios. Through this process, we identified items that were formulated ambiguously, items that could not clearly be observed (e.g., items referring to cognitive processes), and problems in the order or grouping of items.

To increase the accuracy of the evaluation of a performance, we followed the example of the Clinical Performance Tool[29] and specified, for each checklist item, three scoring categories: task not performed (zero points); task performed partially, incorrectly, or with delay (one point); and task performed completely, correctly, and within the recommended time frame (two points). For example, a team would score two points for calling for help in the first five minutes but only one point for doing so after five minutes.

Step 4: Final Delphi review round

To ensure expert consensus concerning changes made in step 3, we sent the revised checklist for review, asking the original five pediatrician experts, as before, to delete irrelevant actions, add missing items, and/or edit listed items.

Step 5: Item weighting

Not every item in a checklist is equally important for the treatment to be successful. A checklist differentiating between essential and less important items is likely to provide more accurate performance assessments. Thus, in a final development step, we sent the checklist to 10 pediatricians and pediatric anesthetists from Switzerland, Germany, and Australia. We instructed them to rate all checklist items in terms of their importance for the success of the treatment from 1 (not important) to 5 (essential). The mean importance score of each item served as its weight.

Internal consistency and validity

We tested internal consistency of the final checklist by having three pediatricians, including two of us (E.H. and F.H.) and an independent rater, assess four videotaped samples of managing the simulated pediatric septic shock scenario. We assessed Cohen kappa[30] to measure agreement (for all four videos) between the independent rater and either E.H. or F.H.

We assessed *content* validity (the extent to which the checklist includes all relevant items) through a detailed discussion at an international workshop for simulation in medicine.

Evidence for *construct* validity would mean that the checklist score positively but not excessively correlated with the external constructs.[31] We applied two external constructs commonly used for validation[12,32]: (1) a global expert performance rating from 1 to 10 (given by E.H., F.H., and the independent rater before rating the video with the evaluation checklist) and (2) the experience level of the team leader (assessed by a questionnaire; in an emergency scenario, a team with a more experienced leader should get higher scores than a team with an inexperienced leader). We tested construct validity using a sample of 22 teams performing a septic shock scenario in a simulated setting.

Box 1

Patient History of the Simulated Septic Shock Scenario

A 6-month-old male infant presents with several hours of fever and vomiting. The fever responds poorly to antipyretics, and the infant becomes progressively lethargic and responds only to painful stimulus. Skin exam reveals scattered nonblanching petechiae. Two minutes after initial evaluation, the infant becomes unresponsive.

Results

Step 1: Development of a draft checklist

Three of us (E.H., F.H., and W.J.E.) developed a draft checklist consisting of 22 potential items.

Step 2: Delphi review rounds

Experts. The five experts included in the Delphi process had 14 to 28 years of general medical experience and had worked 6 to 27 years in pediatric care in different Swiss and German hospitals.

Delphi rounds and checklist changes. During the first review round, experts made the following suggestions: seven items for addition; one item for deletion; and three items for reformulation.

In the second round, all the experts agreed to add six of the seven items newly suggested for addition in round 1. Four of the five experts did not agree with the seventh addition, so this one item was excluded. The majority (n = 4) of experts disagreed with the proposal to exclude the one item suggested for deletion in round 1; thus, this item was included again. All the experts agreed with the proposed rewording of three items. Furthermore, one new additional item was proposed to add to the list.

After the third round, all the experts agreed to add the additional item suggested in round 2 and had no further suggestions. Also, the two changes which were not accepted by the majority in round 2 (deleting one item and adding one item) were, at this point, accepted by the corresponding expert who had proposed the changes based on the detailed comments of opposing experts. So the five experts achieved consensus about all the items in the list after three review rounds. The list, after step 2, contained 28 items.

Step 3: Design of the final checklist and pilot testing

At this step, we determined which of the 28 items made sense to rate with the three scoring categories. For 7 of the items, the scoring option "partially done" made no sense (e.g., check pulse, check temperature), leaving the option of only zero or two points.

Using the checklist, two of us (E.H. and F.H.) individually rated the video-recorded management of six simulated cases of septic shock and took notes about problems with specific scoring categories. Next, we discussed possible improvements to the checklist. We identified four types of adjustments to enhance the usability of the checklist. Table 2 provides the four types of adjustments made and the corresponding scoring categories. After the pilot phase, the checklist contained 33 items.

Step 4: Final Delphi review round

All five experts agreed with all adjustments made in step 3.

Step 5: Item weighting

The average experience after medical school of the 10 experts participating in step 5 was 16 years (standard deviation [SD] = 9.9), and in a pediatric field specifically, it was 11.4 years (SD = 8).

Internal consistency and validity

The mean score of the ratings ranged from 3 to 5. In general, the SD was small. Only 6 of the 33 items had an SD of more than 1.0. The two items least specifically related to the immediate treatment of septic shock generated the highest disagreement: "Put on gloves before procedure" (SD = 1.73) and "Early planning for other treatment" (SD = 1.35). The final list including the rounded weighting of each item can be seen in Appendix 1.

Interrater reliability analyses of the resulting checklist revealed "substantial" to "almost perfect" kappa coefficients[33] (κ range: 0.65–0.95).

Our thorough, integrative development process through which we derived the items provides content validity.[1] Further, participants of an international workshop for simulation in medicine agreed that the content of the checklist includes all necessary items. The correlation between the checklist score and team leader experience ($r = 0.48$, $P < .05$) and the global rating ($r = 0.68$, $P < .05$) were both significant. Thus, the checklist yields valid results.

Discussion

To design effective training interventions, valid and reliable performance measures must be developed systematically. In this report, we describe a systematic approach to designing checklists for evaluating clinical performance that integrates the published evidence and the knowledge of domain experts. Through its clearly predefined procedure, our method reduces opportunities for subjective interpretations and thus minimizes rater biases. Our approach consists of five easy-to-apply, predefined steps that integrate the following: current checklist development guidelines, an expert consensus method, pilot testing, an additional expert consensus round, and a

Table 2

The Four Types of Adjustments Made After Pilot Testing the Checklist for Taking Care of an Infant With Septic Shock, 2013

Type of adjustment	Problem	Old item	Solution or new item(s)
Specifying	No clear and objective definition of item	"Connect monitors"	Connect ECG, SpO_2, and blood pressure monitors
Splitting	Items are formulated too broadly	"Order and give fluid bolus 3 times"	-Order first fluid bolus -Give first fluid bolus -Order second fluid bolus -Give second fluid bolus -Order third fluid bolus -Give third fluid bolus
Eliminating redundancy	Different items include the same actions	"ABC evaluation"	Item deleted because the following items were already in the list: -Assess airway/breathing -Assess mental status
Changing the order of items	Inconvenient order of items due to thematic grouping instead of grouping according to the course of events		The order should correspond to the expected course of events as much as possible so as to minimize rater search time

Abbreviations: ECG indicates electrocardiogram; SpO_2, oxygen saturation; ABC, airway, breathing, circulation.

survey to get importance ratings for the checklist items.

Our approach has some advantages over other systematic approaches.[14,20,21] Lockyer and colleagues[20] proposed a method to develop a checklist in three stages. In stage 1, the authors created an evaluation checklist and then published it on the Web site of the Neonatal Resuscitation Program (NRP) for additional review. Then the NRP recruited volunteers to review the list by mail. In stage 3, a pilot test was conducted in which experienced instructors used the list to rate specific video clips. After each step the checklist was modified. Although Lockyer et al[20] obtained feedback from a large number of responders and conducted a pilot test, it is unclear how they modified the list after every step and how they dealt with conflicting comments. Further consensus methods were not applied.

Morgan et al[14] and Scavone et al[21] both used a well-defined Delphi technique. They required experts participating in their Delphi technique to agree not only with the items included in the checklist but also to a weight (of 1 to 5) for each item. These weights could be problematic because the two analyses did not consider the small sample size and because the final checklist used the mean score of the expert ratings for the items for which no consensus could be achieved. Therefore, we strongly suggest conducting a separate step (following our step 5) to obtain the weights of the checklist items, allowing for an adequate sample size, at least for those items for which no consensus is achieved.

All three aforementioned studies included a pilot phase through which raters tested the checklist as an assessment tool.[14,20,21] This step is indispensable; by applying the checklist to a set of different examples, the raters experience the applicability of the items and the usability of the rating scale. Each item has to be formulated in a clear and observable way. If it is not, then it must be excluded or modified so that it does not threaten interrater agreement. For example, the item "Equipment check" seems absolutely reasonable. However, to get reliable ratings, the checklist must define *what* equipment has to be checked (e.g., oxygen connector, ventilation bag). Another problem arises with the rating of behaviors that are hard to detect or are executed mentally (e.g., "Check breathing"—whether or not the trainee has perceived the lifting and lowering of the chest is unclear if he or she does not verbalize doing so).

After the pilot phase, we conducted a final Delphi review round (step 4). To our knowledge, no other study has included additional expert feedback *after* a pilot phase. This step is important because it ensures that the adjustments made after the first testing are recognized by experts and not biased in any way by raters' personal opinions or by experiences that are not generally valid.

Lessons learned

Not only the Delphi review rounds but also the inquiry about the item weights can take a long period of time. Content experts in the field are often very busy clinicians, and responding to the inquiries is not their first priority. If possible, checklist developers and investigators should consider creating individual incentives for the experts to enhance their commitment (e.g., free access to the final product).

Further, we noted some process issues: In one case, we detected a lack of expert diligence in providing feedback, and in another an expert overlooked some items and did not comment on them. Soliciting the missing comments lengthened the time of that particular Delphi review round. Thus, we recommend emphasizing the importance of the experts' contribution in the first communication and indicating a reasonable expectation for response time so that experts can reject the invitation immediately rather than dropping out later.

Areas of application

We demonstrated our approach using the example of a simulated sepsis scenario. Our approach, though, is not limited to one scenario; it is generalizable. We have since successfully applied this five-step process to other clinical scenarios (i.e., pulseless ventricular tachycardia, bronchiolitis, and near-drowning). In doing so, we have created checklists which correspond to the specific context in question and which, in some cases, differ considerably from the initial checklist drafted by the research team.

Evaluation checklists are generally most suitable for training purposes or for simulated scenarios in which no patient outcome measures are available. Our approach to checklist development may be particularly useful to design evaluation checklists for situations that have a certain degree of standardization and are frequently covered by guidelines. Because there are national differences in the treatment of specific clinical scenarios, our approach can also be employed to include expert feedback when adapting existing checklists for a new national setting. Our approach would also be useful in any setting for updating a list to account for changes in guideline regulations. For less standardized situations, in which the actions depend highly on the particular situation, assessing performance with a checklist may not be suitable; for example, a checklist would not capture the many skills necessary for managing a critically ill child with a complex past medical history and dealing with end-of-life issues related to "do not resuscitate" orders or withdrawal of intensive care. In such situations, another form of assessment, such as a behaviorally anchored rating scale, a global rating tool, or patient-focused outcomes, may augment performance assessment.

Physicians and physician educators can use our five-step procedure not only to design checklists for performance assessment but also for the development of cognitive aids that help ensure all necessary tasks are completed.[34] The use of checklists has been demonstrated to reduce error by standardizing specific processes in surgery,[35] anesthesia,[36] handover,[37] and inpatient care.[38]

Limitations

Despite the advantages of our approach, we note some limitations. The development of an evaluation checklist according to our approach requires significant time and effort. Patient outcomes, specific performance markers (e.g., time to key interventions,[39] decision latency[40]), and global rating scales are often easier to assess and do not require a long development process. Thus, some researchers propose global rating scales as the preferred assessment tool.[16,41] However, patient outcomes or global ratings often do not provide comprehensive evaluations of the treatment process and, thus, cannot provide process feedback to augment debriefings.

One notable limitation concerns the fact that two raters involved in pilot testing (E.H., F.H.) were both also involved in the development process. The testing of a new tool should ideally be done by independent potential users.[42] In our case, we were able to show good interrater agreement with a third, independent rater during the validation process. Therefore, we believe that this limitation had no negative effect on the final checklist. Nevertheless, we recommend independent raters during pilot testing.

Other limitations are related to the Delphi technique in general. Although this process facilitates reaching expert consensus, it does not necessarily mean that this consensus is "correct." Although the possibility of the consensus being influenced by a single expert's opinion is small, the possibility still cannot be ruled out completely. In our case, there was no serious disagreement about whether an individual item should be included in the checklist or not; thus, we assume that this issue did not influence our results.

Further, the country of origin and professional background of the experts could influence their responses. Medical guidelines may vary on a regional or national basis; even local factors at an individual hospital can result in different expert opinions. Although completely controlling for the background of every expert is almost impossible, we tried to counteract cultural differences by selecting the experts from regions where the final checklist should be applied (Germany, Switzerland, Australia). Differences in culture and regions should be kept in mind when choosing the experts for future studies.

Finally, we developed the evaluation checklist for a specific pediatric sepsis scenario as we use it in our simulation trainings. This local context might have influenced the development process in a way that may preclude adopting the final checklist for other sepsis scenarios without making small adjustments.

Future research

In future studies, other formats and venues for performing the Delphi review rounds and their impact on the quality of the final checklist should be explored. A consensus meeting instead of e-mail inquiry may result in a more dynamic and deliberate discussion of the checklist and would speed up the process. Clearly, a disadvantage of such a consensus meeting could be the higher risk of group biases because the experts would no longer be anonymous. To avoid this potential drawback, a consensus meeting could be held online in which experts could maintain anonymity in a virtual chat room.

Conclusions

Assessment of clinical performance is fundamental to further enhance patient safety. Only reliable and valid process performance measures that are less influenced by unknown variables (than are clinical outcomes) will allow medical educators to accurately evaluate the behavioral effects of training interventions and, in turn, leverage and modify the training.

A systematic development process is a necessary prerequisite of valid checklists for reliably assessing process performance. However, no universally agreed guideline for the systematic development of evaluation checklists exists. With this report, we describe a comprehensive integrative approach that may be used in future studies. We are convinced that a widely recognized standard for developing evaluation checklists, such as the one we applied, would advance the field of performance assessment in health care.

Acknowledgments: The authors are grateful to all experts for their participation in the development process of the checklist. They also thank Lauren Clack for her help with the design of the figures and Julia Keil for her participation as an independent rater for the reliability analysis.

Funding/Support: This work was funded by the Swiss National Science Foundation (grant number PP00P1_128616).

Other disclosures: None reported.

Ethical approval: This type of research is exempt from ethics review in Switzerland.

Previous presentations: Oral presentation at the 19th Annual Meeting of the Society in Europe for Simulation Applied to Medicine (SESAM), June 13 to 15, 2013, Paris, France.

Mr. Schmutz is research associate, Industrial Psychology and Human Factors, Department of Psychology, University of Fribourg, Fribourg, Switzerland.

Dr. Eppich is assistant professor of pediatrics and medical education, Northwestern University Feinberg School of Medicine, Chicago, Illinois.

Dr. Hoffmann is senior pediatrician, Dr. von Hauner University Children's Hospital, Munich, Germany.

Dr. Heimberg is pediatrician, University Children's Hospital, Tübingen, Germany.

Dr. Manser is associate professor for industrial psychology and human factors, Department of Psychology, University of Fribourg, Fribourg, Switzerland.

References

1 Boulet JR, Murray D. Review article: Assessment in anesthesiology education. Can J Anaesth. 2012;59:182–192.
2 Campbell JP. Modeling the performance prediction problem in industrial and organizational psychology. In: Dunnette MD, Hough LM, eds. Handbook of Industrial and Organizational Psychology. Palo Alto, Calif: Consulting Psychologists Press; 1990:687–732.
3 Sonnentag S, Frese M. Performance concepts and performance theory. In: Sonnentag S, ed. Psychological Management of Individual Performance. West Sussex, UK: John Wiley & Sons, Ltd.; 2002:1–25.
4 Anderson N, Ones DS, Sinangil HK, Viswesvaran C. Handbook of Industrial, Work and Organizational Psychology: Personnel Psychology. Vol 1. London, UK: Sage Publications Ltd.; 2001.
5 Boulet JR, Jeffries PR, Hatala RA, Korndorffer JR Jr, Feinstein DM, Roche JP. Research regarding methods of assessing learning outcomes. Simul Healthc. 2011;6(suppl):S48–S51.
6 Hales B, Terblanche M, Fowler R, Sibbald W. Development of medical checklists for improved quality of patient care. Int J Qual Health Care. 2008;20:22–30.
7 Davidson JE. Evaluation Methodology Basics: The Nuts and Bolts of Sound Evaluation. Thousand Oaks, Calif: Sage; 2004.
8 Scriven M. The logic and methodology of checklists. December 2007. http://www.wmich.edu/evalctr/archive_checklists/papers/logic&methodology_dec07.pdf. Accessed March 25, 2014.
9 Morgan PJ, Cleave-Hogg D, Guest CB. A comparison of global ratings and checklist scores from an undergraduate assessment using an anesthesia simulator. Acad Med. 2001;76:1053–1055.
10 Bakeman R, Gottman JM. Observing Interaction: An Introduction to Sequential Analysis. 2nd ed. Cambridge, UK: Cambridge University Press; 1997.
11 Donoghue JD, Durbin DR, Nadel FM, Stryjewski GR, Kost SI, Nadkarni VM. Effect of high-fidelity simulation on pediatric advanced life support training in pediatric house staff. Pediatr Emerg Care. 2009;25:139–144.
12 Devitt JH, Kurrek MM, Cohen MM, et al. Testing internal consistency and construct validity during evaluation of performance in a patient simulator. Anesth Analg. 1998;86:1160–1164.
13 Reid J, Stone K, Brown J, et al. The Simulation Team Assessment Tool (STAT): Development, reliability and validation. Resuscitation. 2012;83:879–886.
14 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.
15 Gaba DM, Howard SK, Flanagan B, Smith BE, Fish KJ, Botney R. Assessment of clinical performance during simulated crises using both technical and behavioral ratings. Anesthesiology. 1998;89:8–18.
16 Adler MD, Vozenilek JA, Trainor JL, et al. Comparison of checklist and anchored global rating instruments for performance rating of simulated pediatric emergencies. Simul Healthc. 2011;6:18–24.
17 Cohen R, Rothman AI, Poldre P, Ross J. Validity and generalizability of global ratings in an objective structured clinical examination. Acad Med. 1991;66:545–548.
18 Donoghue A, Nishisaki A, Sutton R, Hales R, Boulet J. Reliability and validity of a scoring instrument for clinical performance during pediatric advanced life support simulation scenarios. Resuscitation. 2010;81:331–336.
19 Stufflebeam DL. Guidelines for developing evaluation checklists: The checklists development checklist (CDC). July 2000. http://www.wmich.edu/evalctr/archive_checklists/guidelines_cdc.pdf. Accessed March 25, 2014.
20 Lockyer J, Singhal N, Fidler H, Weiner G, Aziz K, Curran V. The development and testing of a performance checklist to assess neonatal resuscitation megacode skill. Pediatrics. 2006;118:e1739–e1744.
21 Scavone BM, Sproviero MT, McCarthy RJ, et al. Development of an objective scoring system for measurement of resident performance on the human patient simulator. Anesthesiology. 2006;105:260–266.
22 Biarent D, Bingham R, Eich C, et al. European Resuscitation Council Guidelines for Resuscitation 2010 Section 6. Paediatric life support. Resuscitation. 2010;81:1364–1388.
23 Gordon TJ. The Delphi Method. American Council for the United Nations University (AC/UNU) Millenium Project. 1994. http://fpf.ueh.edu.vn/imgnews/04-Delphi.pdf. Accessed March 25, 2014.
24 Turoff M, Hiltz SR. Computer based Delphi processes. In: Adler M, Ziglio E, eds. Gazing Into the oracle: The Delphi Method and Its Application to Social Policy and Public Health. London, UK: Kingsley Publishers Ltd.; 1996:56–88.
25 Ferguson ND, Davis AM, Slutsky AS, Stewart TE. Development of a clinical definition for acute respiratory distress syndrome using the Delphi technique. J Crit Care. 2005;20:147–154.
26 Huang HC, Lin WC, Lin JD. Development of a fall-risk checklist using the Delphi technique. J Clin Nurs. 2008;17:2275–2283.
27 Naik VN, Perlas A, Chandra DB, Chung DY, Chan VW. An assessment tool for brachial plexus regional anesthesia performance: Establishing construct validity and reliability. Reg Anesth Pain Med. 2007;32:41–45.
28 Clayton MJ. Delphi: A technique to harness expert opinion for critical decision-making tasks in education. Educ Psychol. 1997;17:373–386.
29 Donoghue A, Ventre K, Boulet J, et al; EXPRESS Pediatric Simulation Research Investigators. Design, implementation, and psychometric analysis of a scoring instrument for simulated pediatric resuscitation: A report from the EXPRESS pediatric investigators. Simul Healthc. 2011;6:71–77.
30 Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20:37–46.
31 Cronbach LJ, Meehl PE. Construct validity in psychological tests. Psychol Bull. 1955;52:281–302.
32 Lambden S, DeMunter C, Dowson A, Cooper M, Gautama S, Sevdalis N. The Imperial Paediatric Emergency Training Toolkit (IPETT) for use in paediatric emergency training: Development and evaluation of feasibility and validity. Resuscitation. 2013;84:831–836.
33 Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159–174.
34 Hales BM, Pronovost PJ. The checklist—a tool for error management and performance improvement. J Crit Care. 2006;21:231–235.
35 Haynes AB, Weiser TG, Berry WR, et al; Safe Surgery Saves Lives Study Group. A surgical safety checklist to reduce morbidity and mortality in a global population. N Engl J Med. 2009;360:491–499.
36 Myburgh JA, Chapman MJ, Szekely SM, Osborne GA. Crisis management during anaesthesia: Sepsis. Qual Saf Health Care. 2005;14:e22.
37 Foster S, Manser T. The effects of patient handoff characteristics on subsequent care: A systematic review and areas for future research. Acad Med. 2012;87:1105–1124.
38 Wolff AM, Taylor SA, McCabe JF. Using checklists and reminders in clinical pathways to improve hospital inpatient care. Med J Aust. 2004;181:428–431.
39 Marsch SC, Müller C, Marquardt K, Conrad G, Tschan F, Hunziker PR. Human factors affect the quality of cardiopulmonary resuscitation in simulated cardiac arrests. Resuscitation. 2004;60:51–56.
40 Burtscher MJ, Manser T, Kolbe M, et al. Adaptation in anaesthesia team coordination in response to a simulated critical event and its relationship to clinical performance. Br J Anaesth. 2011;106:1–6.
41 Ma IW, Zalunardo N, Pachev G, et al. Comparing the use of global rating scale with checklists for the assessment of central venous catheterization skills using simulation. Adv Health Sci Educ Theory Pract. 2012;17:457–470.
42 Bevan N, Macleod M. Usability measurement in context. Behav Inf Technol. 1994;13:132–145.

References that appear in Table 1 only

43 Chopra V, Gesink BJ, de Jong J, Bovill JG, Spierdijk J, Brand R. Does training on an anaesthesia simulator lead to improvement in performance? Br J Anaesth. 1994;73:293–297.
44 Thomas EJ, Sexton JB, Lasky RE, Helmreich RL, Crandell DS, Tyson J. Teamwork and quality during neonatal care in the delivery room. J Perinatol. 2006;26:163–169.
45 Tschan F, Semmer NK, Gautschi D, Hunziker P, Spychiger M, Marsch SU. Leading to recovery: Group performance and coordinative activities in medical emergency driven groups. Hum Per. 2006;19:277–304.
46 Adler MD, Trainor JL, Siddall VJ, McGaghie WC. Development and evaluation of high-fidelity simulation case scenarios for pediatric resident education. Ambul Pediatr. 2007;7:182–186.
47 Brett-Fleegler MB, Vinci RJ, Weiner DL, Harris SK, Shih MC, Kleinman ME. A simulator-based tool that assesses pediatric resident resuscitation competency. Pediatrics. 2008;121:e597–e603.
48 Adler MD, Vozenilek JA, Trainor JL, et al. Development and evaluation of a simulation-based pediatric emergency medicine curriculum. Acad Med. 2009;84:935–941.
49 Carlson J, Min E, Bridges D. The impact of leadership and team behavior on standard of care delivered during human patient simulation: A pilot study for undergraduate medical students. Teach Learn Med. 2009;21:24–32.
50 Manser T, Howard SK, Gaba DM. Identifying characteristics of effective teamwork in complex medical work environments: Adaptive crew coordination in anaesthesia. In: Flin R, Michell L, eds. Safer Surgery: Analysing Behaviour in the Operating Theatre. Aldershot, UK: Ashgate; 2009:223–239.
51 Burtscher MJ, Wacker J, Grote G, Manser T. Managing nonroutine events in anesthesia: The role of adaptive coordination. Hum Factors. 2010;52:282–294.
52 Westli HK, Johnsen BH, Eid J, Rasten I, Brattebø G. Teamwork skills, shared mental models, and performance in simulated trauma teams: An independent group design. Scand J Trauma Resusc Emerg Med. 2010;18:47.

Appendix 1

Checklist Developed[a] to Evaluate Care of an Infant With Septic Shock, 2012

Stage of care (timing in minutes)	Item no.	Item description	Scoring (Check the box for 0, 1, or 2 points, as appropriate)			Weighting
			Not done (0 points)	Partially or incorrectly done or not done in a timely manner (1 point)	Done correctly, completely, and in a timely manner (2 points)	
General tasks (0–5)	1-1	Put on gloves before procedure	☐	Some, but not all persons involved in procedure put on gloves ☐	Every person who is involved in procedure puts on gloves ☐	3
	1-2	Equipment check	☐	Incomplete: Oxygen connected *or* ventilator bag checked ☐	Oxygen connected, ventilation bag checked ☐	3.5
	1-3	Connect ECG, SpO_2, BP	☐	Incomplete (only 1 or 2 items) ☐	All complete ☐	4.5
	1-4	Call for help (senior physician)	☐	Not in time (i.e., after actors' recommendation) ☐	Done in time ☐	5
	1-5	Inform team members about diagnosis	☐	Not done in time *or* incomplete information related to vital signs (i.e., only "tachycardia" or only "low blood pressure") ☐	Complete information about diagnosis "shock" ☐	4
Evaluation (0–5)	2-1	Assess airway and breathing	☐	Only bilateral auscultation *or* assess breathing frequency *or* work of breathing ☐	Bilateral auscultation *and* assess breathing frequency *and* work of breathing ☐	5
	2-2	Check SpO_2	☐	Done ☐	Done in time and verbalized ☐	3.5
	2-3	Check pulse	☐		Done in time ☐	4
	2-4	Check ECG	☐	Done ☐	Done in time and verbalized ☐	3.5
	2-5	Check CRT	☐		Done in time ☐	4.5
	2-6	Check BP	☐	Done ☐	Done in time and verbalized ☐	4
	2-7	Check temperature	☐		Done in time ☐	3
	2-8	Assess mental status	☐		Done (explicit question about mental status, e.g., AVPU) ☐	4.5
Treatment (0–5)	3-1	Apply oxygen	☐	Nasal cannula ☐	100% O_2 applied ☐	4.5
	3-2	Establish IV/IO access	☐	More than 2 attempts for IV access *or* not in time ☐	Successful in maximum of two IV attempts ☐	5
	3-3	Order first fluid bolus	☐	Wrong fluid *or* wrong amount ordered *or* not in time ☐	Right dose (20 mL/kg) and right fluid ☐	5
	3-4	Start giving first fluid bolus	☐	IV pump *or* rapid IV push not in time ☐	Rapid IV push ☐	5

(*Appendix Continues*)

Appendix 1
(Continued)

Stage of care (timing in minutes)	Item no.	Item description	Scoring (Check the box for 0, 1, or 2 points, as appropriate)			Weighting
			Not done (0 points)	Partially or incorrectly done or not done in a timely manner (1 point)	Done correctly, completely, and in a timely manner (2 points)	
Treatment and assessment (5–15)	4-1	Reassess circulation (CRT, BP, HR)	☐	Incomplete (checked only 1 or 2) ☐	All complete ☐	4.5
	4-2	Reassess breathing (SpO_2, breathing frequency)	☐	Incomplete 1 (checked only 1) ☐	All complete ☐	5
	4-3	Order second fluid bolus	☐	Wrong fluid *or* wrong amount ordered ☐	Right dose and right fluid ☐	4.5
	4-4	Give second fluid bolus	☐	IV pump ☐	Rapid IV push ☐	4.5
	4-5	Reassess circulation (CRT, BP, HR)	☐	Incomplete ☐	All complete ☐	5
	4-6	Reassess breathing (SpO_2, breathing frequency)	☐	Incomplete ☐	All complete ☐	5
	4-7	Order third fluid bolus	☐	Wrong fluid *or* wrong amount ordered ☐	Right dose and right fluid ☐	4.5
	4-8	Give third fluid bolus	☐	IV pump ☐	Rapid IV push ☐	4.5
	4-9	Reassess circulation (CRT, BP, HR)	☐	Incomplete ☐	All complete ☐	4.5
	4-10	Reassess breathing (SpO_2, breathing frequency)	☐	Incomplete	All complete ☐	4.5
Reassessment and planning for other treatment (5–15)	5-1	Reassess mental status	☐		All complete ☐	4
	5-2	Consider vasoactive agents	☐		Done ☐	4
	5-3	Consider antibiotics	☐	Administration discussed ☐	Ceftriaxon or Cefotaxim ordered ☐	5
	5-4	Draw blood culture and BGA glucose	☐	Incomplete ☐	Includes at least BGA, blood culture, electrolytes ☐	4
	5-5	Prepare for advanced airway management	☐	Incomplete ☐	Suction, medication, bag mask, laryngoscope ☐	4.5
	5-6	Early planning for other treatment	☐		Done ☐	3.5

Abbreviations: ECG indicates electrocardiogram; SpO_2, oxygen saturation; CRT, capillary refill time; BP, blood pressure; IV, intravenous; IO, intraosseous; HR, heart rate; AVPU, alert, responds to voice, responds to pain, unresponsive; BGA, blood gas analysis.

[a]The authors used an integrated, five-step approach to develop the checklist.