Published in final edited form as:

Simul Healthc. 2013 October ; 8(5): 292–303. doi:10.1097/SIH.0b013e318290a022.

Standardized Assessment for Evaluation of Team Skills:

Validity and Feasibility

Melanie C. Wright, PhD,

From the Research Institute, Saint Alphonsus Health System, Boise, ID; Trinity Institute for Health and Community Benefit, North Carolina Hospital Association, Cary, NC.

Noa Segall, PhD,

Trinity Health, Livonia, MI; Department of Anesthesiology, and Human Simulation and Patient Safety Center, North Carolina Hospital Association, Cary, NC.

Gene Hobbs, CHT,

Trinity Health, Livonia, MI; Department of Anesthesiology, and Human Simulation and Patient Safety Center, North Carolina Hospital Association, Cary, NC.

Barbara Phillips-Bute, PhD,

Trinity Health, Livonia, MI; Department of Anesthesiology, North Carolina Hospital Association, Cary, NC.

Laura Maynard, MDiv, and

School of Medicine, Duke University, Durham, NC; NC Center for Hospital Quality and Patient Safety, North Carolina Hospital Association, Cary, NC.

Jeffrey M. Taekman, MD

Trinity Health, Livonia, MI; Department of Anesthesiology, and Human Simulation and Patient Safety Center, North Carolina Hospital Association, Cary, NC.

Abstract

Introduction: The authors developed a Standardized Assessment for Evaluation of Team Skills (SAFE-TeamS) in which actors portray health care team members in simulated challenging teamwork scenarios. Participants are scored on scenario-specific ideal behaviors associated with assistance, conflict resolution, communication, assertion, and situation assessment. This research sought to provide evidence of the validity and feasibility of SAFE-TeamS as a tool to support the advancement of science related to team skills training.

Methods: Thirty-eight medical and nursing students were assessed using SAFE-TeamS before and after team skills training. The SAFE-TeamS pretraining and posttraining scores were compared, and participants were surveyed. Generalizability analysis was used to estimate the variance in scores associated with the following: examinee, scenario, rater, pretraining/

Reprints: Melanie C. Wright, PhD, Saint Alphonsus Regional Medical Center, 1055 N Curtis Rd, Boise ID 83706 (wrightmcc@sarmc.org).

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (<http://www.simulationinhealthcare.com>).

The authors declare no conflict of interest.

posttraining, examinee type, rater type (actor-live vs. external rater-videotape), actor team, and scenario order.

Results: The SAFE-TeamS scores reflected improvement after training and were sensitive to individual differences. Score variance due to rater was low. Variance due to scenario was moderate. Estimates of relative reliability for 2 raters and 8 scenarios ranged from 0.6 to 0.7. With fixed scenarios and raters, 2 raters and 2 scenarios, reliability is greater than 0.8. Raters believed SAFE-TeamS assessed relevant team skills. Examinees' responses were mixed.

Conclusions: The SAFE-TeamS was sensitive to individual differences and team skill training, providing evidence for validity. It is not clear whether different scenarios measure different skills and whether the scenarios cover the necessary breadth of skills. Use of multiple scenarios will support assessment across a broader range of skills. Future research is required to determine whether assessments using SAFE-TeamS will translate to performance in clinical practice.

Keywords

Team skills; TeamSTEPPS; Assessment; Teamwork; Communication; Conflict resolution; Assertion; Situation awareness; Situation assessment; Assistance

There is clear evidence to support the importance of team coordination skills in health care.¹⁻³ However, we know little about how to assess an individual's team skill performance or the relative quality of training programs.⁴ Researchers have sought to define specific teamwork skills of interest, and although there is no clear consensus regarding a specific set of skills or behaviors, there is general agreement on certain constructs.^{5,6} TeamSTEPPS, a large collaborative effort of the Department of Defense and the Agency for Healthcare Research and Quality, for example, identifies important behaviors associated with 5 key teamwork principles as follows: team structure, leadership, situation monitoring, mutual support, and communication.⁷

In addition, there is mounting evidence that training can improve team skills and patient outcomes.⁸⁻¹⁰ However, the training programs that have shown success in changing outcomes generally have involved a combination of teamwork skills training and specialty-specific process-related changes (eg, briefings, debriefings, and checklist use in the intervention of Neily et al⁸). Thus, it is not possible to determine whether improvements in outcome are related to individuals' improvement in skills such as communication or assertiveness or are primarily related to the tools and processes developed to improve teamwork in context. Teamwork-related failings linked to sentinel events involve problems across a variety of clinical contexts and are sometimes related to individuals' inability to manage challenging teamwork situations (eg, high workload, distraction, authority gradients, and conflict). Medical schools and nursing schools have begun to implement a variety of programs intended to teach individuals skills that are broadly applicable to their future health care careers. However, we have limited knowledge related to whether it is possible to assess and teach general behaviors (across specialties and roles) related to managing challenging teamwork scenarios. Expanding our knowledge in this area has the potential to significantly reduce the frequency of such sentinel events.

Methods for assessing teamwork skills generally involve the use of observer-based rating scales, which often are specific to the health care task at hand.^{11–14} Some tools focus on the performance of an entire team rather than an individual, which is useful for assessing impacts of process changes or education of a team in general but not useful for individual assessment purposes. Capturing instances of challenging teamwork scenarios in the context of observation during clinical care is difficult because this requires extended observation periods to capture situations of interest. Schraagen et al¹⁵ attempted to observe and code nonroutine events in cardiac surgery. Although interrater reliability was high with respect to the way events were coded when the 2 raters scored the same nonroutine event, the 2 raters identified the same situations as a nonroutine event for only 32% of events documented (many situations were coded as a nonroutine event by only 1 rater). The presence of an observer may also reduce the likelihood that some critical situations, such as one in which an individual of authority asks a subordinate to take an action that is unethical or unsafe, will occur. Simulation provides an opportunity to expose care providers to challenging teamwork scenarios and observe their response. Simulation also provides the opportunity to develop metrics based on observation of scenario-specific expected behaviors that are easily identified by the rater.

Researchers have attained contradictory results with respect to the validity and reliability of observational techniques such as behaviorally anchored rating scales.^{11,12,15–19} For example, Fletcher et al,¹¹ in an evaluation of the anesthetists nontechnical skills rating scale, report moderate interrater agreement with correlation coefficients of between 0.55 and 0.67 at the level of individual elements of ratings and correlation coefficients of 0.56 to 0.65 at the level of teamwork categories. Morgan et al¹⁸ presented less promising results associated with a general rating scale and advocate for the development of a domain-specific scale for obstetrics. From ratings of simulated obstetric care scenarios, the single-rater intraclass correlation coefficient for external observers was 0.34 using a 45-item scale with 5 team skill constructs and 0.45 for a global rating scale with a single 5-point rating of teamwork from unacceptable to acceptable. Reliability of self-ratings was low and did not correlate highly with the observer ratings. Scores were not consistent across different scenarios. The authors evaluated a behaviorally anchored rating scale for interrater agreement to assess reliability and correlation with a clinical performance measure to provide evidence of content validity.²⁰ In simulated care, moderate interrater agreement was upheld for a combined overall teamwork score, but interrater agreement at the level of individual skill categories was low. Simulated scenarios required participants to work together to treat the patient; however, interaction between team members was without the types of complications that frequently impact communication difficulties associated with adverse events. Although the evidence for the use of general rating scales is promising for purposes other than high-stakes testing, this has generally been at the level of a single overall teamwork rating (frequently averaged from other ratings or scales) in the context of uncomplicated field settings or simulation scenarios. It is not clear whether such scales will exhibit the sensitivity and specificity necessary to differentiate the effectiveness of different training programs with respect to teaching learners to manage challenging teamwork situations.

OBJECTIVE

A Standardized Assessment for Evaluation of Team Skills (SAFE-TeamS) was developed primarily to serve as a tool to compare the effectiveness of different training programs targeting individual team skills applicable across multiple health care roles and environments (eg, team skills training as a part of medical and nursing education). Studies making such comparisons (eg, see Hobgood et al²¹) have provided evidence related to training impact on knowledge and attitudes but have fallen short in assessing the impact of training on skills. It is important to understand whether training has provided learners with the ability to translate knowledge into skills, particularly in the context of challenging team situations, such as standing up to a superior, where knowledge of the correct behavior is inherently simpler than performing it.

The primary objective of this project was to provide evidence of the validity and feasibility of SAFE-TeamS. Downing²² defines 5 sources of evidence of validity in medical education assessment including the following: content, relationship to other variables, internal structure, response process, and consequences. In this study we evaluated evidence for validity of content by determining whether participants, actors, and raters believe the measure assesses relevant team skills. We evaluated relationship to other variables by determining whether the measure distinguishes performance of participants before and after team skills training. We evaluated reliability, an important aspect of internal structure, through generalizability analysis to identify the degree to which the individual, raters, scenarios, and other variables such as rater type and participant type account for the variance in scores achieved. We expect that if the measure is reliable, the variance due to examinee and pretraining/posttraining will be high compared with the variance due to other variables. With regard to response process, our generalizability study provides evidence related to responses of raters by determining whether variability in scores is associated with response process variables such as whether the rater was blind to training condition or rated the sessions live or from videotape. We discuss the implementation SAFE-TeamS and relevant controls for reducing errors associated with test administration. Although we did not specifically evaluate consequences in this research, we provide a brief discussion of this in relation to SAFE-TeamS.

Our primary criteria for evaluating feasibility were the number and qualifications of individuals and the time required to perform the assessment. We conducted a decision study using generalizability analysis to determine the number of scenarios, number of raters, and type of raters required to achieve a reliable rating. The generalizability analysis provides evidence regarding whether actors playing roles in the scenarios are reliable raters. We also describe the time it took to train actors and raters and to prepare and implement the scenarios.

SAFE-TEAMS

We have developed an assessment tool that combines the use of standardized team members (actors playing the roles of other health care team members) in the context of challenging teamwork scenarios. Scoring of performance is observer-based rating of scenario-specific

ideal team skill behaviors. The SAFE-TeamS has the advantages of (1) assessing a trainee in a standardized scenario that is the same for each trainee (not dependent on the performance of other trainees), (2) using scenarios that stress critical teamwork skills in challenging scenarios, and (3) basing scoring on observable behaviors that are easily identified within the context of each scenario.

The SAFE-TeamS involves placing a single examinee in a short (up to 10-minute), structured scenario with 1 or 2 actors playing standardized team members. Actors and a facilitator assume the roles of care providers, patients, or family members and exhibit challenging behaviors, for example, an overbearing attending unwilling to recognize his or her mistakes or a timid coworker overwhelmed by multiple, simultaneous tasks. Students are expected to demonstrate good team skills associated with assistance, communication, conflict resolution, assertion, and situation assessment. Although grounded in clinical contexts, the scenarios integrate relevant clinical background information as necessary to accommodate a wide audience with varied experience or from different clinical domains. Twelve scenarios were developed for initial validation (see Document, Supplemental Digital Content 1, <http://links.lww.com/SIH/A81>, which contains detailed description of the simulation scenarios and scoring). Several of the scenarios were adapted from scenarios published by the TeamSTEPPS collaborative.⁷

The SAFE-TeamS was developed using the following process. Based on a review of the literature and experience with multiple teamwork skills observation tools,^{7,9,11,14,16,20,23,24} we identified specific teamwork skills and constructs that were relevant to managing challenging teamwork scenarios. We then selected several scenarios we believed were representative of challenging team situations from among Team-STEPPS content⁷ and drafted additional scenarios based on (a) personal experiences of clinicians (J.M.T. and Duke colleagues) and (b) literature describing case reports or root cause analyses of preventable adverse events.^{1,2,9,25–28} We selected scenarios that, if not handled with skill, could result in errors or poor-quality care. Specifically, we selected and developed scenarios that would include conditions in which there is high workload, unclear roles and responsibilities, multiple distractions, conflict among team members, or authority gradients/personalities that make it difficult to communicate dissent or stand one's ground. The scenarios include introductory context, relevant props, definition of medical terms (for actors), and a detailed actor script with scripted reactions based on expected examinee behaviors.

We elected to use multiple short scenarios so that we could, in a reasonably short span of time, attain multiple measures of team skill performance. There is a growing trend in simulation-based training to develop short scenarios that address key content areas so that it is possible to teach and/or assess multiple situations across a relatively short period (eg, see Waldrop et al²⁹ and Henrichs et al³⁰). Owing to the nature of teamwork skills (there is rarely a single “correct” approach to managing a situation) and because we were interested in evaluating multiple aspects of team skill performance, we expected that multiple measures may be required to obtain reliable results.

We then reviewed these scenarios to identify scenario-specific behaviors that were both observable and key to skilled performance (ie, behaviors we would expect experts to

exhibit). We used our previous review of team skills literature and Team-STEPPS content, including specific strategies such as checkback, 2-challenge rule, and SBAR (Situation, Background, Assessment, Recognition) to identify these behaviors. We matched the scenario-specific observable behaviors we identified to overarching team skill constructs. Thus, our review of team skills literature drove the selection and design of scenarios, which in turn drove our selection of scenario-specific behaviors and linked team skill constructs that were most important to managing the challenging team skill situations we were targeting. A set of 6 overarching constructs that were associated with observable behaviors were repeated throughout the scenarios (Table 1).

We created questions and specific observable responses that would allow raters to classify performance of each scenario-specific behavior as 0 for a behavior not performed, 1 for a behavior performed in some form but not ideal, and 2 for ideal performance of the behavior. For example, observers are not required to determine whether the participant “was assertive,” but rather whether he or she did not push for patient to be sent to radiology immediately (scored as “0”), voiced this concern once or twice (scored as “1”), or voiced concern at least twice and asked to speak with another authority (scored as “2”). We found that the scenarios naturally lent themselves to the assessment of 3 or 4 specific key behaviors. We then edited the scenarios to attain relatively even coverage across the 6 constructs identified and to allow for 4 scenario-specific behavior ratings in each scenario (see Document, Supplemental Digital Content 1, <http://links.lww.com/SIH/A81>, which also provides a table that matches specific team skills constructs to specific scenarios). We chose to combine the 4 ratings into a single score by summing them. For each scenario, participants can achieve a total score that ranges from 0 to 8, with 8 representing perfect performance across all teamwork behaviors assessed. This decision was based on previous experiences with teamwork rating scales that revealed better interrater reliability associated with a score combined across multiple measures of team constructs compared with scores on individual team constructs.^{20,31}

METHODS

Overview

This study was approved by the Duke University Medical Center Institutional Review Board for research with human subjects.

We collected data for the purposes of assessing relationship to other variables by comparing pretraining and posttraining scores to determine whether SAFE-TeamS was sensitive to improvement in team skills associated with assistance, conflict resolution, communication, assertion, and situation assessment after training. Using the same data set, we conducted generalizability analysis^{32,33} to determine the variability in scores associated with several important variables outlined in Table 2. The variance components calculated through generalizability analysis provide evidence of reliability (eg, within subjects across scenarios or across raters or rater types) and, with decision analyses, can be used to calculate the number of repetitions required to attain targeted reliability levels to provide data that can be used to estimate time and cost as an evaluation of feasibility. We also surveyed examinees,

actor raters, and external raters at the conclusion of the study as an evaluation of content validity.

Participants

Examinees were medical and nursing students in their final year of study. We set a target sample size of 30 participants for the generalizability analysis because 95% confidence intervals for a broad range of correlation coefficients have a width of ± 0.10 with $n = 30$.³³ Because we encountered sound recording problems for some participants in the first year of data collection, we collected data in the spring of 2 consecutive years, resulting in a final sample size of 38 examinees, with 38 completing the pretraining session and 33 completing the posttraining session. We elected to sample both medical and nursing students to determine whether SAFE-TeamS was both valid and reliable across a broad sample of health care learners.

Team Skill Training

To assess the validity of SAFE-TeamS with respect to sensitivity to training targeting relevant teamwork skills, we assessed examinees before and after teamwork training delivered as a part of their normal curriculum. The training was a capstone activity that occurred in the final year of the curriculum. The training included lectures, video presentations exemplifying good and poor teamwork, and interdisciplinary small group activities based on TeamSTEPPS content. The training was similar in content and presentation across the 2 years of the study. Individuals responsible for designing and delivering the training content were not part of the research team, and their goals for the training were focused on teamwork education of medical students and nursing students. They facilitated the research through announcements for recruiting but were not invested in the outcomes of the research project. The training targeted content relevant to SAFE-TeamS in that SAFE-TeamS was designed to assess skills that are taught in TeamSTEPPS content, particularly as the skills apply to managing challenging teamwork situations.

SAFE-TeamS Scenarios and Experimental Procedures

Participants completed 6 SAFE-TeamS scenarios before training and 6 SAFE-TeamS scenarios after training. All participants who completed both assessment sessions encountered all 12 scenarios. Scenario presentation order was randomized. The number of scenarios used was selected to provide a large enough sample for a realistic representation of an assessment tool involving multiple vignettes but still few enough to easily train actors and to be able to assess examinees in a 1-hour session.

Participants were asked for consent and were given an explanation of the study and instructions for the assessment procedure. Once a scenario was staged and the actor team and experimenters were prepared, participants were asked to enter the room and were given a brief introduction to the scenario including their role, the clinical context, and the names and roles of the actors in the scenarios (see Document, Supplemental Digital Content 1, <http://links.lww.com/SIH/A81> scenario descriptions and Video, Supplemental Digital Content 2, <http://links.lww.com/SIH/A82>, of a medical student interacting with actors in the “Transfusion Reaction” SAFETeamS scenario). The experimenter would then signal the

beginning of the scenario, and the actors would begin the scenario as scripted. Each scenario took between 5 and 10 minutes to complete and was videotaped. The experimenter would signal the end of a scenario when it had progressed to the point required for scoring all questions.

Actor Teams, Actor Raters, and External Raters

One variable that may impact scores is how actors portray their roles in the scenarios. Therefore, we designed the study to expose participants to different actor teams and controlled the makeup of the teams, so that we could evaluate actor teams as a variable in the generalizability analysis. We trained 5 actors who were combined into 4 unique 2-person actor teams. Actors underwent approximately 4 half-day training and rehearsal sessions covering a subset of TeamSTEPPS content, practice in the 12 scenarios, practice and study of relevant medical terminology, and practice using the rating forms. The actor teams were exposed to 8, 14, 17, and 18 different participants each (either before training, after training, or both). Although our goal was to expose the teams to similar numbers of participants, one actor withdrew from the study prematurely such that 1 team was used for a relatively small sample compared with the other teams.

Actors completed their scoring of examinees immediately after each scenario. In accordance with SAFE-TeamS scoring, they scored each of 4 questions on the scenario as 0, 1, or 2, by circling the appropriate response. Actor raters were asked to complete their scoring independently and were monitored to ensure compliance to the degree possible. Because actors saw some participants both before and after training, they were not blind to either participant type or pretraining/posttraining. Each actor scored every participant they encountered; thus, all participants were scored by 2 actors. The 5 different actors rated 8, 17, 22, 32, and 35 participants each.

Three external raters scored participants. External raters independently scored participants from recorded video and were blind to both pretraining/posttraining and participant type. One external rater was a social scientist familiar with the TeamSTEPPS content. The other 2 raters were a social scientist and a health care provider who were familiar with team training content in general and received a briefing on TeamSTEPPS content and the SAFE-TeamS tool. External raters rated 18, 30, and 32 participants each. External raters were exposed to a subset of examinees because of recording problems (missing audio, failure to record, or poor recording quality). One external rater rated fewer participants because of an unanticipated change in the time available to participate.

Analysis for Sensitivity to Pretraining/Posttraining

To evaluate whether SAFE-TeamS was sensitive to improvement due to training, we generated a single pretraining score and a single posttraining score for each participant that was the average of their 6 scenario scores (an average of 6 ordinal scores on a scale from 0 to 8, resulting in a single score treated as a continuous variable). Because actor raters rated participants live and were not blinded to pretraining/ posttraining and examinee type while external raters rated participants from videotape and were blinded to pretraining/ posttraining and examinee type, we compiled 2 separate data sets for analysis, one that

averaged the scores from the 2 actor raters and one that averaged the scores from the 2 external raters who rated 30 or more participants. The sample for the actor raters included 33 participants scored by 2 actors pretraining and posttraining. The sample for the external raters included 27 participants scored by the same 2 external raters pretraining and posttraining. For both data sets, scenario was randomly assigned within pretraining/posttraining. For the data set using actor ratings, raters varied depending on the actor team used. For the data set using external raters, the same 2 raters rated all participants. We then conducted paired t tests comparing pretest and posttest scores for both data sets.

Postexperiment Survey

Survey questions assessed impressions of examinees, actors, and external raters with respect to how well they believed the scenarios assessed team skills, how realistic the scenarios and actors were, and whether they perceived any learning effect due to participation. We conducted descriptive analyses of these survey data.

Generalizability Analysis

For the generalizability analysis, our dependent variable is the teamwork skill score for each individual scenario scored, an ordinal measure, ranging from 0 to 8. We assessed multiple independent variables as listed in Table 2. We included in the analysis all of the data collected across 5 actor raters, 3 external raters, 12 scenarios, and 38 participants. The total number of scores in the data set (across scenarios, raters, and examinees) was 1823. Although the research design is complex, generalizability theory is capable of assessing reliability in incomplete block designs such as this.

Because of multiple variables, it was not possible to conduct a fully crossed generalizability analysis incorporating all 8 variables and interactions. Therefore, we first analyzed each variable in the data set in a single variable (plus error) model to determine which of the 8 variables were important to consider in assessing the reliability of SAFE-TeamS. From this analysis, we eliminated from further analysis any variables that accounted for less than 2% of the variance of the scores (with one exception for the decision study, explained later). We then conducted a generalizability analysis, without interactions, including all of the variables accounting for more than 2% of the variance in the model.

Decision Study

To assess both reliability and feasibility, we conducted a fully crossed 3-way decision study³⁴ using the variables examinee, scenario, and rater to predict the number of raters and number of scenarios required in the use of SAFE-TeamS. We chose this set of variables because these are the variables most likely to be controlled in the use of SAFE-TeamS. To control for other variables with significant variance (pretraining/posttraining) or under particular scrutiny (rater type; although rater type accounted for <2% of the variance in the initial analysis, only external raters were blinded to pretraining/posttraining and had no opportunity for other bias, such as overhearing comments of other actors or experimenters), we conducted the analysis separately for 4 data sets that encompassed the pretraining and posttraining conditions and actor and external raters. We performed both random-effects and fixed-effects analyses. We also used the variance component estimates to estimate the

reliability associated with the individual scores used for the *t* test comparisons of pretraining and posttraining scores.

RESULTS

In the analysis comparing pretraining and posttraining scores, one participant had a difference between pretraining and posttraining scores that was greater than 3 SDs from the mean (in both the actor rater and the external rater data sets); therefore, this examinee's data were removed for the analysis. With the removal of these data, both data sets were normally distributed. The mean (SD) as scored by 2 actors ($n = 32$) was 4.2 (1.2) before training and 5.1 (1.1) after training. The mean (SD) as scored by the 2 external raters ($n = 26$) was 4.0 (1.0) before training and 4.8 (1.2) after training. Based on *t* tests on both samples, the score after training was significantly greater than the score before training at $P < 0.001$. The pretest and posttest scores of the participants rated by 2 external raters are displayed in Figure 1.

Thirty-seven of the 38 examinees, 3 of 3 external raters, and 4 of 5 actors completed the survey after their participation in the assessment exercise. The results are shown in Table 3.

In the single-variable model generalizability analyses for each of the initial set of 8 variables, the variables examinee type, rater type, actor team, and order (Table 1) each accounted for less than 2% of the variance in the scores. With the exception of separating the data sets by rater type for the decision analyses, the remainder of the analyses included only the variables that accounted for more than 2% of the variance in the scores: examinee, rater, scenario, and pretraining/posttraining. Variance components estimates and percentage of total variance for a model including the variables examinee, scenario, rater, and pretraining/posttraining without interactions are displayed in Table 4.

For the decision study, the results of the 3-way fully crossed model with examinee, rater, and scenario for the 4 data sets including pretraining and posttraining rated by actors and external raters are presented in Table 5. Figure 2 presents plots of generalizability coefficient estimates for both relative reliability and fixed reliability for 1 to 4 raters across number of scenarios for the external rater, pretest data set (improvement in reliability with increasing raters was similar for the other 3 data sets). To provide an understanding of the potential variability of generalizability coefficient estimates across rater types and different characteristics of trainees, Figure 3 presents plots of generalizability coefficient estimates for both relative reliability and fixed reliability assuming 2 raters for the 4 data sets based on rater type and pretraining/posttraining. Estimates of relative reliability for 2 raters and 8 scenarios range from 0.63 to 0.71. With fixed scenarios and raters, reliability is greater than 0.8 with 2 raters and 2 scenarios.

For estimating the reliability of scores used in our comparison of pretraining and posttraining scores, we used the pretraining variance estimates (the more conservative estimates). For the actor data set, a single examinee's score based on 6 random scenarios and 2 random raters has an estimated reliability of 0.58. For the external rater data set, a single

examinee's score based on 6 random scenarios and 2 fixed raters has an estimated reliability of 0.68.

DISCUSSION

The results provide evidence of validity associated with content, relationship to other variables, internal structure, and the rating component of response process of SAFE-TeamS. The SAFE-TeamS demonstrated validity with respect to differentiating pretraining and posttraining performance and individual differences in team skills. Raters reliably scored participants independent of actor team, rater type, scenario, examinee, order, examinee type, and training condition. Although raters perceptions were positive in support of SAFE-TeamS, participant support was mixed. Discussion of evidence and threats to the 5 sources of validity follows. We also discuss feasibility and implications for use.

Content

Our primary evidence for content-related validity was our survey of raters and participants. All 7 of the raters who completed the survey indicated that they believed the scenarios assessed important teamwork skills. Participants, however, had mixed reactions to both the realism and the value of SAFE-TeamS as a valid measure of teamwork skills. Although 75% of the examinees described the experience as positive, half indicated that they did not believe the measure assessed important teamwork skills. Based on the comments of some of the examinees, part of their reaction may have been related to the fact that they were sometimes asked to take on different roles (eg, medical students playing the roles of nurses and vice versa). One way to manage this would be to use only within-role scenarios; however, cross-training in other roles may be a valuable component of teamwork training^{35,36} and, thus, useful in an assessment tool. A second comment noted by participants was that the actors may have "overacted," portraying behaviors that are beyond what would normally be expected in a care environment. Another potential problem area with respect to realism was that the depth of the actors' clinical knowledge was limited and this was sometimes apparent when they were pressed for information that was not scripted. In addition, because the scenarios stressed challenging teamwork scenarios, the experience may have been psychologically distressing for some participants. Some of the scenarios represent extreme situations that include standardized team members asking examinees to do things that may be unethical or unprofessional (eg, shouting in the presence of a patient, signing that they witnessed something they did not see). It is not clear whether this may have impacted their ratings in a negative way.

Evidence for content validity can also be assessed by comparing the content to other sources such as descriptions of relevant content in the literature. Within our scenarios, we map specific behaviors to the 6 constructs defined in Table 1. These 6 constructs are frequently included in descriptions of key team skill behaviors and other measures of team skill.^{4,6,9,11,20,23,24,28} Constructs that are present in other sources that may not be well covered by SAFE-TeamS are those associated with leadership, team structure, and planning. These constructs did not arise as critical behaviors in the context of scenarios portraying

challenging team situations that can be driven by behaviors of actors playing the roles of team members.

The process by which we designed SAFE-TeamS involved working from scenarios selected as challenging teamwork scenarios toward the development of objective questions targeting key team skills. There is evidence of convergent validity in this process in that, with the exceptions mentioned previously, the resulting scoring mechanisms aligned well with established team skills constructs. However, the development process also resulted in scenarios that did not individually cover all 6 team skill constructs. We expected that multiple scenarios would be required to assess performance across all skill categories to be assessed. This is a threat to SAFE-TeamS content-related validity, and future efforts may be targeted toward developing scenarios that cover each of the targeted skills. Another threat to content validity for SAFE-TeamS is that, although clinician experts and team skills experts developed the tool, we did not use a rigorous method of expert review by clinicians and team skill experts before evaluating SAFE-TeamS with examinees.

Relationship to Other Variables

We specifically assessed relationship to other variables by comparing participants before and after team skills training. This revealed that SAFE-TeamS, using a sample of 6 scenarios and 2 raters, was sensitive to changes in performance associated with managing challenging teamwork skills after training. Although training condition only accounted for 4% of the variance in scores, for the 27 participants scored by the same 2 external raters who were blinded to training condition, 21 had higher scores posttraining compared with pretraining, and the mean score increased after training by 10%. Thus, although training did not account for a significant portion of variance in the score, it certainly impacted scores in the manner predicted. It is clear, however, that individual differences account for more variance in SAFE-TeamS scores than at least the form of team skills training we evaluated (group training of 1-day duration). We believe it is likely that nontechnical teamwork skills are either largely personality dependent or learned at earlier stages of development (eg, as children and young adults, both in the home and at school). Thus, training can have some impact, but to see changes in scores that are more on the order of the differences in scores identified between individuals, more significant training and practice may be required.

The degree to which exposure to the SAFE-TeamS assessment pretraining may have impacted posttraining performance is not clear. Although no feedback was given to participants based on their scores and the lack of an order effect within the pretraining and posttraining set of 6 scenarios suggest that exposure to SAFE-TeamS scenarios alone had little measurable impact on performance, it is possible that participating in the exercise made participants more aware of challenging scenarios that could occur during their work in health care and perhaps more receptive of the content delivered in their team skills education session.

We did not evaluate SAFE-TeamS in relation to other measures of teamwork skill. However, training similar to that used in our study has previously shown to increase scores on teamwork knowledge tests,¹⁸ so we may infer at least some level of agreement between SAFE-TeamS scores and knowledge tests. However, our goal was specifically to develop a

tool that provided information beyond what can be attained through knowledge tests. Ultimately, triangulation with other metrics will be required to fully assess the validity of SAFE-TeamS. In particular, it is not yet possible based on these findings to assert generalizability of performance scores to team skill performance in true patient care situations.

Response Process

Participants were given instructions regarding what was expected in the scenarios (eg, to simply act as they would playing a particular role given the scenario context). This type of assessment was novel for participants. Given that participation was by paid volunteers with no consequences for their performance, the degree to which each participant took the assessment seriously is not clear. Qualitatively, we observed some hesitation in the first few moments of the first scenario, followed by engagement by most participants. We also qualitatively observed some variability in participants' ability or choice to suspend disbelief and "play along." Only 1 or perhaps 2 participants clearly expressed discontent and acted in a manner that would suggest they were not taking their role seriously.

With regard to the response process of raters, raters perceived the tool as easy to use (Table 3). External raters completed their scoring individually and were blind to training condition. Actor raters were asked to complete their scoring independently and were monitored for this during data collection. That the generalizability analysis revealed less than 2% of variance in scores accounted for by rater type suggests that there was no pretest/posttest bias on the part of actor raters and that participating in the scenarios and scoring live did not negatively impact external raters' ability to score participants. In addition, the relatively low impact of rater on variability in scores and consistent variance due to raters between external and actor raters also reflects positively on the response process for both types of raters.

After the experiment, we sought feedback from external and internal raters regarding the scoring tool. In general, the feedback was positive with a few comments on specific queries, which are documented in Document, Supplementary Digital Content 1, <http://links.lww.com/SIH/A81>.

Threats to validity with respect to response process included (a) external raters rated a subset of the participants rated by actor raters, (b) differences in the number of participants rated by raters, (c) variability in the distribution of actor teams across participants, and (d) attaining data from 2 years of classes, rather than the one originally designed because of technical recording difficulties.

Internal Structure

Our analysis of internal structure of SAFE-TeamS focused primarily on generalizability analysis to quantify the variance in scores associated with variables we believed may serve as threats to internal structure. These included the makeup of the actor team, order, rater type (includes rating process), examinee type (because SAFE-TeamS was designed to work across health care roles), and error. We expected to attain measurable variance due to examinee, training condition, rater, and scenario although, for purposes of feasibility with

respect to required numbers of raters and scenarios, lower variance components within these variables is ideal.

The generalizability analysis revealed that the actor team, order of scenario presentation, rater type and process, and examinee type had little impact on the SAFE-TeamS score. Although external raters and experimenters perceived that actors played their roles in scenarios somewhat differently from one another, this finding suggests that SAFETeamS can be implemented with different actors. The absence of variability across examinee type (medical vs. nursing student) and rater type suggests that different health care professionals can be rated live or from video recordings, with little impact to the overall reliability of the rating. Actor raters were not blind to training condition, which represents a threat to internal structure. However, the lack of variability associated with rater type as described in the discussion of response process suggests that different rater types and process had little impact on variance in scores.

The variance in the SAFE-TeamS score due to the examinee was 16%, accounting for more variance than any other single variable in the analysis. This is a positive finding in that we expect the individual being examined to influence the score. Variance due to training condition was discussed previously in the section on relationship to other variables.

Rater accounted for about 3% of the variance in the model. This is a relatively small impact of rater compared with other rating tools and is positive in that reliability can be high with just a few raters. This finding suggests that we achieved our goals of designing an objective tool through the technique of rating easily observable scenario-specific behaviors. This finding is more in line with assessment tools targeted toward scoring clinical skills³⁷ compared with those targeted toward assessing teamwork skills, which tend to achieve lower rates of interrater agreement.^{11–20} Based on the decision study analyses (Figs. 1 and 2), simply increasing the number of raters from 1 to 2 (eg, by using both actors) greatly increases the reliability of the data yielded from the measure and the use of a greater number of raters beyond 2 has limited return on investment.

With regard to the variance in the score accounted for by scenario, the main effect of 9% is significant but not surprising. Because our measure was designed with the intent of using multiple scenarios, it is not surprising that the influence of scenario on score variance was somewhat high. The scenarios are designed to be relatively short for this reason. These findings are consistent with those of Weller et al in an analysis of the reliability of an assessment of performance in anesthesia crisis management in which they found that 12 to 15 cases were required to reliably rank trainees.³⁸ Using multiple scenarios in an assessment increases both reliability and confidence regarding validity.

The most problematic finding in our generalizability analysis was the high contribution of error or unknown sources to variance in scores. The decision study (Table 5) revealed that a significant portion of the variance was contributed to an examinee by scenario interaction. It is not clear whether we can interpret this as scenarios assessing different skills (ie, some individuals did well on certain scenarios but poorly on others) or, since each participant performed each scenario only once, whether this is simply measurement error. The high

proportion of variance accounted for by examinee suggests that there is consistency in scores within subjects across scenarios. However, further analyses are required to fully understand how scenarios and the skills assessed within scenarios may differently impact different individuals.

Consequences

Based on our initial goals for SAFE-TeamS, there are minimal consequences for individuals with respect to SAFE-TeamS. Instead, SAFE-TeamS is intended as a tool to assess group level effects and compare training methods. As such, validity failings in the use of SAFE-TeamS may have the unintended impact of wrongly promoting or devaluing specific training content or methods. However, in comparison with the state of the science with respect to evaluating such training (eg, making conclusions based on short-term changes in knowledge or attitudes), risks associated with the use of SAFE-TeamS for this purpose are minimal.

Feasibility

According to the decision study, with 6 scenarios and 2 raters, relative reliability (both scenarios and raters may vary) estimates range from 0.58 to 0.71. With 12 scenarios and 2 raters, relative reliability estimates range from 0.69 to 0.83. Using a fixed set of scenarios and/or raters increases reliability substantially (a reliability of nearly 0.8 can be achieved with only 1 scenario and 2 raters). However, with fewer than 6 fixed scenarios, although the measure may be reliable, it is not clear whether the rating may reflect performance on only a subset of the intended teamwork constructs of assistance, conflict resolution, communication, assertion, and situation assessment, depending on which scenarios are selected for use.

Each scenario takes 6 to 10 minutes to run with about 5 minutes required for transitions to set up new props and for actors to score the previous scenario. We ran our assessments with 2 experimenters. Experimenters were responsible for (1) communicating with examinees, (2) narrating the scenarios and playing some roles such as individuals on a telephone call or the patient, (3) helping actors keep track of ratings, (4) managing props, and (5) video recording. With practice, it is possible to run SAFE-TeamS sessions with 1 coordinator and 2 actors. It is also possible for a coordinator to serve as a third rater with limited impact on resources or time. Thus, one could assess participants across 6 scenarios in about an hour with two actors and 1 or 2 facilitators.

Because scoring was designed to be simple, based on easily observable scenario-specific behaviors, external rater training time was minimal, requiring about 1 hour of interaction with one of the authors (M.C.W.) and an additional hour or two reviewing scenarios and questions. External rater scoring time was roughly equivalent to scenario performance time. External raters had the ability to, and did, play back parts of scenarios but were also able to fast-forward through parts of the scenarios that were not relevant to scoring (such as the introduction and setup time).

Actor training time was significant, requiring about 16 hours of instruction and rehearsal. Although our generalizability analysis indicates that SAFE-TeamS was feasibly implemented with actors who have traditionally played standardized patient roles, we found

that the actors sometimes struggled with medical terminology and with clinical details when pressed for information by examinees. About half of the medical and nursing trainees found their performance to be unrealistic. In future use of SAFE-TeamS, we will evaluate the use of health care providers or trainees in these roles, with actors hired to facilitate training of these individuals in acting skills.

With respect to feasibility related to resources, we designed SAFE-TeamS to allow for rapid setup and transition between scenarios, low cost and portability with respect to equipment, and flexibility in terms of location. We conducted our experiment in 2 different locations (in a typical simulation laboratory and across the hall in a conference room with makeshift walls). Thus, to some extent, we sacrificed environmental fidelity and focused primarily on the fidelity of the team interaction. We did not evaluate whether our limitations with respect to environmental fidelity impacted our findings.

Implications for Use

We designed SAFE-TeamS primarily as a tool to measure the impact of different training methods on improving the management of challenging teamwork simulations. The findings of our study suggest high levels of reliability. Estimates of reliability for designs incorporating different numbers or types of raters and different numbers of scenarios can be attained using the variance components estimated in this study. With respect to support for content related validity or generalization to other applications or settings, further evidence may be required.

Other potential applications of SAFE-TeamS may include (1) providing an objective benchmark with which to validate less resource-intensive metrics of assessing individual teamwork skills such as supervisor ratings or observational metrics and (2) providing formative assessment of trainees. With regard to the latter, we found that a somewhat common reaction of participants was a desire for feedback on their performance. With opportunity for learners to view their own scores and/or debrief their video-recorded performance with a facilitator, SAFE-TeamS may prove to be an effective tool for improving their teamwork skills. For use both as a formative assessment and as a benchmark to evaluate other metrics, the consequences associated with using SAFE-TeamS are minimal. Thus, the findings of this research may be applicable for these purposes.

Before using SAFE-TeamS, one should consider important threats to content validity evidence. It is not clear the extent to which SAFE-TeamS covers the breadth of important teamwork constructs. The SAFE-TeamS is limited to the assessment of assistance, closed loop communication, structured communication, situation assessment, assertion, and conflict resolution. The mapping of scenario content to these constructs has not been independently evaluated. The use of multiple scenarios increases confidence that the score attained reliably predicts performance on other SAFE-TeamS scenarios. However, there is limited evidence to suggest that this translates to performance in clinical practice. The perceptions of medical and nursing students regarding the realism and relevance of SAFE-TeamS scenarios add to this concern, as does the high degree of variance that could not be accounted for in our generalizability analysis. It is recommended that users of SAFE-TeamS select scenarios that

are most appropriate to the goals of training or assessment. Practice or pilot testing (perhaps with different actors) before use is strongly recommended.

In addition to the potential use of SAFE-TeamS for low consequence purposes, this study provides general knowledge to inform future development of tools or techniques intended for a variety of purposes including high stakes assessment. Basing scoring on observable scenario-specific behaviors had the intended outcome of supporting high levels of reliability between raters. However, this method also requires significant effort in the development of appropriate and effective scenarios. Using actors in scenarios to provide individual assessment had the intended outcome of isolating individual performance in controlled situations. However, actor training is time intensive and, in comparison with team-based assessments, it is not possible to evaluate multiple participants simultaneously. Our use of multiple short scenarios also achieved our intended objective of allowing us to attain multiple measures on relevant team constructs across different clinical scenarios in a relatively short period. Although preparation is more intensive than other techniques such as observation in clinical settings, SAFE-TeamS provides high levels of control for the purposes of assessing individual performance in challenging teamwork situations. Future research is required to determine whether assessments using SAFE-TeamS will translate to performance in clinical practice.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

The authors wish to acknowledge the hard work of the actors and external raters who participated in this project: Jim Babel, Barbie Iszlar, Andrea Bloch, Amy Magurno, Bradley Norden, Kevin Silva, and Angela Ray. We would like to thank the Duke University School of Medicine (Dr. Colleen Grochowski, Ms. Susan Eudy, Dr. Edward Buckley, and Mr. David Gordon), Duke University School of Nursing (Dr. Dori Sullivan and Ms. Nancy Short), the Duke University Health System Patient Safety Office (Dr. Karen Frush), and TeamSTEPPS® participants and leaders (Dr. Heidi King, Department of Defense and Dr. James Battles, Agency for Healthcare Research and Quality), Madojutola Dawodu, and the National Board of Medical Examiners for their ongoing support of this project. We thank the Simulation in Healthcare reviewers and editors for their thoughtful contributions to the final manuscript. During the time that Ms. Laura Maynard worked on this project, she was employed by the Duke University Health System Patient Safety Office. During data collection and analysis for this project, Dr. Wright was Assistant Professor, Department of Anesthesiology, Duke University School of Medicine.

Supported by grants from the Edward J. Stemmler, MD, Medical Education Research Fund at the National Board of Medical Examiners (0708–086) and the Agency for Healthcare Research and Quality (U18 HS01653–01). Also partially supported by the Agency for Healthcare Research and Quality (K02 HS015704–01) (to M.C.W).

The study sponsors had no role in any aspect of study design, data collection and analysis, or manuscript preparation.

REFERENCES

1. Commision Joint. Update: sentinel event statistics. *Jt Comm Perspect* 2006;26:14–15.
2. Lingard L, Espin S, Whyte S, et al. Communication failures in the operating room: an observational classification of recurrent types and effects. *Qual Saf Health Care* 2004;13:330–334. [PubMed: 15465935]

3. Helmreich RL, Davies J. Human factors in the operating room: interpersonal determinants of safety, efficiency, and morale In: Aitkenhead AA, ed. Balliere's Clinical Anaesthesiology: Safety and Risk Management in Anaesthesia. London, England: Balliere Tindall; 1996:277–296.
4. Gaba DM. Crisis resource management and teamwork training in anaesthesia. *Br J Anaesth* 2010;105:3–6. [PubMed: 20551023]
5. Baker DP, Salas E, King H, Battles J, Barach P. The role of teamwork in the professional education of physicians: current status and assessment recommendations. *Jt Comm J Qual Patient Saf* 2005;31:185–202.
6. Salas E Is there a “big five” in teamwork? *Small Group Res* 2005;36:555–559.
7. Agency for Healthcare Research and Quality. TeamSTEPPS Team Strategies & Tools to Enhance Performance & Patient Safety. Rockville, MD: Agency for Healthcare Research and Quality Available at: <http://teamstepps.ahrq.gov/> Accessed February 3, 2011.
8. Neily J, Mills PD, Young-Xu Y, Carney BT, et al. Association between implementation of a medical team training program and surgical mortality. *JAMA* 2010;304:1693–1700. [PubMed: 20959579]
9. Baker DP, Gustafson S, Beaubien J, Salas E, Barach P. Medical Teamwork and Patient Safety: The Evidence-Based Relation. Washington, DC: American Institutes for Research; 2003.
10. Salas E, Wilson KA, Burke CS, Wightman DC. Does crew resource management training work? An update, an extension, and some critical needs. *Hum Factors* 2006;48:392–412. [PubMed: 16884057]
11. 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]
12. Healey AN, Undre S, Vincent CA. Developing observational measures of performance in surgical teams. *Qual Saf Health Care* 2004; 13(Suppl 1):i33–i40. [PubMed: 15465953]
13. Thomas EJ, Sexton JB, Helmreich RL. Translating teamwork behaviours from aviation to healthcare: development of behavioural markers for neonatal resuscitation. *Qual Saf Health Care* 2004; 13(Suppl 1):i57–i64.
14. Frankel A, Gardner R, Maynard L, Kelly A. Using the Communication and Teamwork Skills (CATS) Assessment to measure health care team performance. *Jt Comm J Qual Patient Saf* 2007;33:549–558. [PubMed: 17915529]
15. Schraag JM, Schouten T, Smit M, et al. Assessing and improving teamwork in cardiac surgery. *Qual Saf Health Care* 2010;19:1–6.
16. Malec JF, Tosher LC, Dunn WF, et al. The Mayo high performance teamwork scale: reliability and validity for evaluating key crew resource management skills. *Simul Healthc* 2007;2:4–10. [PubMed: 19088602]
17. Moorthy K, Munz Y, Adams S, Pandey V, Darzi A. A human factors analysis of technical and team skills among surgical trainees during procedural simulations in a simulated operating theatre. *Ann Surg* 2005;242:631–639. [PubMed: 16244534]
18. Morgan PJ, Pittini R, Regehr G, Marrs C, Haley MF. Evaluating teamwork in a simulated obstetric environment. *Anesthesiology* 2007;106:907–915. [PubMed: 17457121]
19. 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;89:8–36. [PubMed: 9667288]
20. Wright MC, Petrusa ER, Griffin KL, Phillips-Bute BG, Hobbs GW, Taekman JM. Assessing teamwork in medical education and practice: relating behavioural teamwork ratings and clinical performance. *Med Teach* 2009;31:30–38. [PubMed: 18825572]
21. Hobgood C, Sherwood G, Frush K, et al. Teamwork training with nursing and medical students: does the method matter? Results of an interinstitutional, interdisciplinary collaboration. *Qual Saf Health Care* 2010;19:e25.
22. Downing SM. Validity: on meaningful interpretation of assessment data. *Med Educ* 2003;37:830–837. [PubMed: 14506816]
23. Wright MC, Luo X, Richardson WJ, et al. Piloting Team Coordination Training at Duke University Health System. Proceedings of the HFES 50th Annual Meeting San Diego, CA: Human Factors and Ergonomics Society; 2006:949–953.

24. Baker DP, Salas E. Principles for measuring teamwork skills. *Hum Factors* 1992;34:469–475.
25. Behaviors that undermine a culture of safety. Sentinel Event Alert [Internet]. 2008; (40). Available at: http://www.jointcommission.org/sentinel_event_alert_issue_40_behaviors_that_undermine_a_culture_of_safety/.
26. Leape LL, Bates DW, Cullen DJ, et al. Systems analysis of adverse drug events. ADE Prevention Study Group. *JAMA* 1995;274:35–43. [PubMed: 7791256]
27. Leape LL, Brennan TA, Laird N, et al. The nature of adverse events in hospitalized patients. Results of the Harvard Medical Practice Study NEngl J Med 1991;324:377–384.
28. Leonard M, Graham S, Bonacum D. The human factor: the critical importance of effective teamwork and communication in providing safe care. *Qual Saf Health Care* 2004;13(Suppl 1):i85–i90. [PubMed: 15465961]
29. Waldrop WB, Murray DJ, Boulet JR, Kras JF. Management of anesthesia equipment failure: a simulation-based resident skill assessment. *Anesth Analg* 2009;109:426–433. [PubMed: 19608813]
30. Henrichs BM, Avidan MS, Murray DJ, et al. Performance of certified registered nurse anesthetists and anesthesiologists in a simulation-based skills assessment. *Anesth Analg* 2009;108:255–262. [PubMed: 19095860]
31. Flowerdew L, Gaunt A, Spedding J, et al. A multicentre observational study to evaluate a new tool to assess emergency physicians' non-technical skills. *Emerg Med J* 2012;[Epub ahead of print].
32. Brennan R Generalizability Theory. New York, NY: Springer-Verlag; 2001.
33. Shavelson R, Webb N. Generalizability Theory: A Primer. Newbury Park, CA: Sage Publications, Inc; 1991.
34. Brennan RL. Generalizability Theory. ITEMS; The Instructional Topics in Educational Measurement Series [Internet]. 2008–2010 November 1, 2010; Module 14 (Winter 1992):225–232. Available at: <http://www.ncme.org/pubs/items.cfm>.
35. Travillian KK, Volpe CE, Canon-Bowers JA, Salas E. Cross-Training Highly Interdependent Teams: Effects on Team Processes and Team Performance. Proceedings of the Human Factors and Ergonomics Society 37th Annual Meeting Santa Monica, CA: Human Factors and Ergonomics Society; 1993:1243–1247.
36. Salas E, Nichols DR, Driskell JE. Testing three team training strategies in intact teams: a meta-analysis. *Small Group Res* 2007;38:471–488.
37. Fehr JJ, Boulet JR, Waldrop WB, Snider R, Brockel M, Murray DJ. Simulation-based assessment of pediatric anesthesia skills. *Anesthesiology* 2011;115:1308–1315. [PubMed: 22037637]
38. Weller JM, Robinson BJ, Jolly B, et al. Psychometric characteristics of simulation-based assessment in anaesthesia and accuracy of self-assessed scores. *Anaesthesia* 2005;60:245–250. [PubMed: 15710009]

FIGURE 1.

Scores of 27 participants scored by the same 2 external raters pretraining and posttraining. The black line represents the mean score pretraining, and the gray line represents the mean score posttraining.

FIGURE 2.

Estimates of generalizability coefficient for 1 to 4 raters across number of scenarios for external raters using the pretraining data set. Relative reliability assumes scenarios, and raters are selected at random. Fixed reliability assumes that the raters and scenarios are fixed (eg, all participants receive the same scenarios and are rated by the same raters).

FIGURE 3.

Estimates of generalizability coefficient assuming 2 raters for 4 different samples of data (rated by actors before and after training and rated by external raters before and after training). Relative reliability assumes scenarios, and raters are selected at random. Fixed reliability assumes that the raters and scenarios are fixed (eg, all participants receive the same scenarios and are rated by the same raters).

TABLE 1.**Team Skills Constructs Assessed in SAFE-TeamS**

Construct	Description
Assistance	Offering task assistance, following through with task assistance, or requesting assistance on tasks.
Closed loop communication	The use of closed loop communication techniques such as repeating back verbal information provided or communicating relevant next steps that make clear that a specific message was received.
Structured communication	Communication using clear and concise language. This includes the use of the SBAR (Situation, Background, Assessment, and Recommendation) technique but does not require SBAR as long as participants clearly and concisely convey the information important in the context of the simulation.
Situation assessment and advocacy	Assessing relevant clinical details in the scenario (eg, that a patient is becoming unstable), verbalizing this assessment, and/or advocating for redirection of attention or specific actions that are in the best interest of the patient with respect to safe high quality care.
Assertion	Maintaining and asserting a position in the face of opposition. This includes the use of the “2-challenge” rule, which states that individuals should state their opposition twice and if it is not heeded, raise their concern to a higher level.
Debrief/conflict resolution	Managing conflict, for example, through the use of techniques such as verbally acknowledging the concerns of other individuals and respectfully refocusing other team members to relevant patient care concerns. This category also describes performance in the context of debriefing, which requires similar skills with respect to identifying problems that occurred and offering constructive feedback.

TABLE 2.

Variables Analyzed in the Generalizability Analysis

Variable	Description
Examinee	Medical students and nursing students in their final year of training.
Examinee type	Medical vs. nursing student is a between-subjects variable.
Pretraining/posttraining	Pretraining vs. posttraining is a repeated-measures variable, with students assessed both before and after teamwork skill training.
Scenario	12 unique teamwork scenarios made up a repeated-measures variable, blocked by pretraining/posttraining. Students experienced all 12 scenarios, 6 before training and 6 after training. Order of scenario presentation was randomized.
Actor teams	4 unique teams of 2 actors was an incomplete block design. Multiple teams are required to assess reliability associated with actor teams and to provide coverage for scheduling. Actors were exposed to a subset of examinees and multiple replications of all scenarios.
Rater	3 trained external observers scored recorded scenarios from videotape. External raters were blind to pretraining/posttraining and student type. The 5 actors who made up the 4 actor teams also served as raters. Each actor scored each participant that they encountered. Each rater scored multiple replications of all 12 scenarios.
Rater type	Rater type (actor live vs. external from video) is a repeated measure, with most examinees being evaluated by both rater types.
Order	Order indicates whether the scenario was encountered first, second, through sixth within the pretraining or posttraining set of scenarios administered.
Error	The remaining variance in scores can be attributed to error, which will include any uncontrolled variables.

TABLE 3.

Survey Responses

No. Responses That Were*	Positive	Neutral	Negative
Examinee questions (n = 36)			
The behavior of the standardized team players was realistic.	18	4	14
The scenarios were realistic.	19	3	14
The exercise was long enough to accurately assess my team skills.	13	10	13
The scenarios assessed important team skills.	15	3	18
I acquired important team skills in participating in this exercise.	14	10	12
Overall team skills assessment experience.	27	N/A	9
Rater and actor questions (n = 7)			
The scenarios assessed important team skills.	7		
The assessment tool accurately assessed participants' performance.	5	2	
The assessment tool consistently assessed participants' performance.	6	1	
Overall experience rating participant teamwork skills.	6	N/A	1
The assessment tool was easy to use.	6	1	
Actor questions (n = 4)			
I portrayed my roles accurately.	4		
I portrayed my roles consistently with different participants.	3	1	
My acting partners portrayed their roles accurately.	3	1	
My acting partners portrayed their roles consistently with different participants.	3	1	
Rater questions (n = 3)			
The behavior of the standardized team players was realistic.	2	1	
The acting of the standardized team players was consistent across different acting teams.	1		2
The acting of the standardized team players was consistent across different trainee sessions.	2		1

* Positive is a response of agree or strongly agree (for the overall experience question, good or excellent). Negative is a response of disagree or strongly disagree (for the overall experience question, fair or poor).

N/A indicates not applicable.

TABLE 4.

Variance Components for a Model Including the 4 Variables That Accounted for Greater Than 2% of the Variance

Variable	Variance Component	Percentage of Variance Accounted for by Variable
Examinee	0.80	16.4%
Scenario	0.46	9.3%
Pretraining/posttraining	0.20	4.1%
Rater	0.14	2.9%
Error	3.30	67.3%

TABLE 5.

Variance Components for 3-Way Fully Crossed Models With Examinee, Scenario, and Rater for 4 Separate Data Sets Including Actor Pretest, Actor Posttest, External Pretest, and External Posttest

Variable	Estimated Variance (Percentage of Variance Accounted for by Variable)					
	Actor Raters		External Raters			
	Pretest	Posttest	Pretest	Posttest	Pretest	Posttest
Examinee	0.77	(16.2%)	0.90	(20.1%)	0.67	(16.1%)
Scenario	0.34	(7.1%)	0.39	(8.7%)	0.39	(9.4%)
Rater	0.19	(4.0%)	0.09	(2.0%)	0.19	(4.6%)
Examinee $\times$ scenario	2.22	(46.6%)	1.50	(33.6%)	1.58	(37.9%)
Examinee $\times$ rater	0.23	(4.8%)	0.00	(0%)	0.16	(3.8%)
Scenario $\times$ rater	0.22	(4.6%)	0.19	(4.3%)	0.16	(3.8%)
Examinee $\times$ scenario $\times$ rater and error	0.79	(16.6%)	1.40	(31.3%)	1.02	(24.5%)
Relative error variance	3.24		2.90		2.76	
G-coefficient (1 rater and 1 scenario)	0.19		0.24		0.20	
						3.15
						0.23