Going DEEP:

Guidelines for Building Simulation-based Team Assessments

James A. Grand, Ph.D. [1]

Marina Pearce, M. A.[2]

Tara A. Rench, M.A.[2]

Georgia T. Chao, Ph.D.[3]

Rosemarie Fernandez, M.D.[4]

Steve W.J. Kozlowski, Ph.D. [2]

[1]College of Health Professions & Psychology Department, The University of Akron, Akron, OH, USA
[2]Department of Psychology; Michigan State University; East Lansing, MI, USA
[3]Department of Management; Michigan State University; East Lansing, MI, USA
[4]Division of Emergency Medicine; Department of Medicine; University of Washington; Seattle, WA, USA

Citation:
Grand, J.A., Pearce, M., Rench, T., Fernandez, R., Chao, G.T., & Kozlowski, S.W.J. (2013). Going DEEP: Guidelines for building simulation-based team assessments. *BMJ Quality & Safety*, *22*, 436-448.

ABSTRACT

Background: Whether for team training, research, or evaluation, making effective use of simulation-based technologies requires robust, reliable, and accurate assessment tools. Extant literature on simulation-based assessment practices has primarily focused on scenario and instructional design; however, relatively little direct guidance has been provided regarding the challenging decisions and fundamental principles related to assessment development and implementation.

Objective: The objective of this manuscript is to introduce a generalizable assessment framework supplemented by specific guidance on how to construct and ensure valid and reliable simulation-based team assessment tools. The recommendations reflect best practices in assessment and are designed to empower healthcare educators, professionals, and researchers with the knowledge to design and employ valid and reliable simulation-based team assessments.

Overview: Information and actionable recommendations associated with creating assessments of team processes (non-technical "teamwork" activities) and performance (demonstration of technical proficiency) are presented which provide direct guidance on how to **D**istinguish the underlying competencies one aims to assess, **E**laborate the measures used to capture team member behaviors during simulation activities, **E**stablish the content validity of these measures, and **P**roceduralize the measurement tools in a way that is systematically aligned with the goals of the simulation activity while maintaining methodological rigor (**DEEP**).

Summary: The DEEP framework targets fundamental principles and critical activities that are important for effective assessment, and should benefit healthcare educators, professionals, and researchers seeking to design or enhance any simulation-based assessment effort.

INTRODUCTION

Over a decade has passed since the Institute of Medicine's seminal report *To Err is Human* cited communication breakdown and teamwork failures as primary threats to patient safety.[1,2] Throughout this time span, multiple efforts have been spearheaded within the healthcare field to improve the effectiveness of interdisciplinary medical teams. These efforts have targeted a variety of areas and topics, including team training (e.g., TeamSTEPPS), team design (e.g., rapid response teams), and protocol development (e.g., sepsis care policies).[3-5] While investigations and implementations of such programs represent significant advancements and achievements [6-8], team-based assessments necessitate highly accurate and reliable methodologies in order to pinpoint team member behaviors and determine whether an intervention influences these behaviors.[9,10] Consequently, continued development of training initiatives and research streams directed toward improving healthcare team effectiveness must be accompanied by equally critical advancements in the rigor and sophistication of measurement techniques used to assess these improvements.[11,12] Failure to maintain such precision in the evaluation of team effectiveness not only lessens one's ability to assess the potency of viable interventions, but also poses a significant threat to the validity of conclusions drawn by educators and researchers.

Ensuring rigorous assessment takes on even greater significance given the increasing prevalence of high fidelity human patient simulation technologies in medical education and research.[13-20] Simulation-based technologies represent a powerful platform for administering assessments of medical teams as they enable one to deliver realistic, scientifically valid (i.e., systematic, reliable, and replicable), and practically useful experiential opportunities to participants. Despite these advantages, the greater degree of flexibility and equifinality inherent in simulation-based environments poses some significant challenges for accurately capturing and interpreting the quality of observed activities.[21] Coupled with the already complex demands of assessing team outcomes, capitalizing on the unique strengths of conducting team training and research in simulation-based environments further reinforces the need for precise and rigorous behavioral assessment systems.[12,14]

With respect to providing medical educators and researchers with guidance on these issues, a survey of the current literature reveals several reviews that outline best practices related to administering team-based training in simulation environments.[7,13,14,21-24] These resources generally provide guidance regarding broader system design issues, such as the need to construct learning environments in a manner that aligns educational objectives with desired teamwork competencies.[25-26] While some of these works also make note of and provide some insight into assessment during the development of simulation-based team training and research, the requirements of sound assessment design and implementation are often far more granular in nature and typically demand significantly greater attention than one might expect. For example, questions such as “Which particular behaviors should I assess?,” “How should I craft an assessment tool to capture those behaviors?,” and “Does my assessment accurately reflect what is intended?” can be surprisingly challenging to answer in team-based assessments.[27]

To this end, the present manuscript provides guidance regarding best practices, approaches, and techniques relevant to creating and employing simulation-based team assessment tools. The recommendations advanced in the present work are organized into four core activities that are especially critical for accurate and reliable assessments of team functioning within simulation-based environments—**D**istinguish, **E**laborate, **E**stablish, and **P**roceduralize (DEEP). The purpose of the guiding principles summarized within this framework is twofold: (1) to provide specific, meaningful, and actionable recommendations regarding fundamental facets of assessment design and evaluation, and (2) to advance an approach that is applicable to all possible uses of simulation-based team assessment (e.g., training, research, and evaluation).

Key Definitions and a Framework for Assessment

Before delving into the specifics of DEEP, it is pertinent to briefly define a few key terms as well as provide some context for understanding the present work as it relates to the broader domain of simulation-based assessment. Most team training interventions and related research investigations seek to improve *team effectiveness*, or the quality with which teams of individuals combine their capabilities and resources to meet environmentally-driven task demands.[27-29] As with many other disciplines, the primary questions of interest within healthcare domains typically involve assessing two critical components of team effectiveness: team processes and team performance.[30] *Team processes* refer to non-technical skills possessed by team members that collectively organize, orient, structure, or facilitate a team's efforts (e.g., coordination, back-up behavior, etc.).[31-35] *Team performance*, by comparison, characterizes actions and outcomes directly associated with the demonstration and application of technical knowledge, skills, and abilities (KSAs) towards the resolution of task goals (e.g., error reduction, decreased hospital stay, adherence to standard treatment guidelines).[12,21] More colloquially, team processes reflect the notion of "teamwork," whereas team performance reflects aspects of "taskwork."[31] Team training and research interventions frequently attempt to target team processes because they are often more diagnostic of a team's strengths and weaknesses and provide insight into what, why, and how certain team performance outcomes were (or were not) achieved.[13,14] Consequently, appropriately distinguishing team processes from team performance is critical for virtually all team-based assessment efforts.[27]

As noted previously, a number of methodological frameworks and approaches have been advanced which provide guidance regarding issues related to assessment within simulation-based training environments.[7,11-13] For example, Rosen et al.'s Simulation Module for Assessment of Resident Targeted Event Responses (SMARTER)[11] outlines an 8-step procedure for the development of simulation-based training intended to promote (1) consistency among core training competencies, simulation experiences, and performance measures; (2) collection of data which facilitates corrective feedback; and (3) ensuring that provided opportunities to perform adequately span certain desired competencies. Thus, although assessment is recognized as integral component within frameworks like SMARTER, these guidelines have primarily emphasized procedures for constructing training environments in which assessments takes place and/or the manner by which assessment activities should be woven into scenario design (e.g., constructing experiential learning opportunities aligned with desired competencies, employing event-based methodologies, identifying behavioral markers). Such practices are both essential and critical to effectively using simulation-based technologies; however, when considered only within the context of instructional and simulation design, many of the critical decisions, foundational practices, and challenging nuances central to developing and employing rigorous assessment instruments can be overshadowed.

Understanding how to construct highly controlled simulations is but one of the critical skillsets necessary for carrying out precise and meaningful team-based simulation activities. Equally important is comprehension of the key principles and practices which can be leveraged to improve the validity and reliability of assessments used for education, research, formal evaluation/certification, or validation purposes within simulation environments.[36]

Consequently, the recommendations highlighted in the DEEP framework seek to provide focused insight into the fundamental machinery underlying sound assessment. Figure 1 offers a high-level procedural framework summarizing the primary stages and critical decisions to be addressed during the design, refinement, and implementation of simulation-based assessment systems. This general blueprint follows globally recognized best practices and, as such, shares features that appear similar to other approaches (e.g., SMARTER). However, the emphasis of this procedural framework and the specific activities encapsulated by DEEP differs from extant frameworks in a few notable ways: (A) they highlight the foundational practices of simulation-based assessment development and implementation applicable to a wide variety of purposes (e.g., training, evaluation, research, etc.); (B) they make explicit how one's decisions related to assessment permeate throughout all phases of development; and (C) they emphasize the iterative nature of assessment development and the critical importance of revisiting one's decisions throughout to ensure that the desired goals of assessment are being achieved. A comprehensive description of all the activities encompassed within each of the stages shown in Figure 1 far exceeds the scope of this article. In light of this, the principles of DEEP provide guidance regarding some of the most critical activities within these phases that also tend to be particularly challenging or overlooked, but which hold significant potential for improving the validity and reliability of simulation-based assessment efforts.[36]

METHODOLOGY

Assessing Team Effectiveness: Going DEEP

In the sections that follow, specific guidelines for **D**istinguishing, **E**laborating, **E**stablishing, and **P**roceduralizing simulation-based team assessments are discussed. Figure 2 provides a complete summary of these guidelines. As an overall point of reference, the underlying goal of all assessments is to provide meaning to an occurrence, perception, or observation from the environment.[37] The confidence one is able to place in the interpretation of an assessment is tied to the precision and reliability possessed by the tools used to record those events.[36] Consequently, the goal of "going deep" with team assessments by adhering to the principles elaborated below is to *ensure that consistency and specificity of assessment activities are maintained throughout all stages of development and implementation*.

Distinguish

As shown in Figure 1, the initial stages of assessment begin with explicating the phenomena/outcomes of interest (*Identify Focal Constructs*) and describing how those focal concepts manifest in the observable environment (*Specify Focal Constructs*). Thus, in the context of simulation-based team assessments, the first set of core activities associated with the DEEP framework—**D**istinguish—involves explicitly defining the team processes and performance outcomes that one wishes to capture. Within this stage, two integral components are described: *delineating team performance from team processes* and *operationalizing constructs*.

Delineating team performance from team processes

Although team performance and team process are intimately intertwined aspects of team effectiveness, their meaning and underlying motivations are unique.[27,30,31] Therefore, separate assessment tools that capture team performance behaviors that are distinct from team process behaviors are needed. In constructing these tools, one will need to make decisions regarding how to distinguish team activities that represent either team performance or team processes. To this end, we propose the following guidelines:

- If an action produces change in the status, trajectory, or characteristics of the team's task objectives (e.g., stabilizing a patient, administering treatment, etc.), then that action should be captured as *team performance*.
- If an action produces change in the collaborative interactions, awareness, or immediate activities of team members, then that action should be captured as *team process*.

Given the singularly critical importance of delineating between process and performance in measurements of team effectiveness, additional guidance and exemplars relevant to this decision point are provided in Table 1. To reiterate, team performance should reflect activities which are directly linked to clinical/task outcomes and which have objective and/or agreed upon standards of proficiency (i.e., can be evaluated as correct/incorrect); they should not assess the quality of team member interactions (though they may capture whether the content of team member communication is objectively accurate), the appropriateness/purpose of team discussions, or interpretations of team members' intentions. Alternatively, team processes should reflect communications and activities geared towards monitoring situational demands, making decisions, ensuring important roles and jobs are being fulfilled, identifying and/or clarifying problems and strategies, and assisting other team members in need of help; they should not assess the accuracy of teams' actions/decisions nor whether a technical/clinical KSA was adequately demonstrated.

Operationalizing constructs

Once team performance and team process have been conceptually distinguished, the next step involves operational distinction. In psychometrics, *operationalization* describes the process by which "fuzzy" theoretically defined concepts (constructs) are converted into something capable of being measured, quantified, and interpreted through empirical observation (variables). More specifically, operationalization refers to the creation and application of specific logic, rules, and/or definitions which describe precisely how an observable behavior should be interpreted in relation to an unobservable construct.[38] For example, consider a physician who conducts a routine physical examination to determine a person’s overall level of health. Although the patient’s “overall health” is not a readily identifiable entity, it represents something the physician wants to be able to capture and discuss (a construct). Thus, she takes a variety of concrete, observable measurements (variables; e.g., blood pressure, weight, cardiac/pulmonary auscultation, etc.) that are representative of overall health. The physician then takes the data/information gathered from the measurement of these variables and compares it against a rule of thumb, medical definition, or some other standard to assess whether that observed measurement (e.g., blood pressure = 120/80, weight = 160 lbs., etc.) is indicative of good or bad overall health. In this fashion, the assessment of the patient’s overall health is described (i.e., operationalized) according to the rules of thumb, definitions, or standards used to interpret the meaning of a particular observation.

This same basic idea is used in the measurement of team performance and team process constructs; variables should be identified which are representative of either team performance or team process, and then operational definitions should be established which describe how these variables uniquely relate to each construct. The selection of specific team performance and team process constructs will vary depending upon the simulation scenario, and there are a number of resources in the clinical and team training literatures that can be used to identify relevant variables of interest.[7,12,21,31,32] However, there is far less guidance regarding how to operationalize constructs in a manner that enables one to answer “The team did X – is this characteristic of team performance A or B (or team process C or D)?" Consequently, we offer the following guidelines:

- An operational definition should possess, at minimum, a clear description of the *content, action, expression, and/or criterion* which must be observed in order to state that an occurrence of the team performance or team process variable has been witnessed.
- To the extent possible, the operational definition should reflect *which team members can perform the action* (e.g., any member, team leader, etc.), *to whom or to what can the action be performed* (e.g., another team member, something in the environment, etc.), and *under what circumstances should the action typically be performed* (e.g., within a certain time frame, in response to a particular event, etc.).
- As a general rule, *a good operational definition trades generality and breadth for objectivity and specificity*. During assessment, if it becomes clear that classifying team activities as representative of only a single team performance or process variable is difficult/ambiguous, the likely culprit is an operational definition that is too broad.

For example, *Mission Analysis* is a team process that describes members' activities related to the interpretation and evaluation of the team's main tasks, environmental conditions, and available resources.[31] An operational definition in this case must provide a description of the specific observable actions that count towards a team's display of Mission Analysis processes. Thus, in the context of a simulated healthcare scenario, one might operationalize this concept as "any action performed by a member of the team that is related to the gathering and communication of information about patient diagnosis (i.e., the team's main task) or conditions relevant to the operational environment (i.e., time pressures, availability of equipment, drugs, consults)." The bi-directional arrows between the *Identify Focal Constructs* and *Specify Focal Constructs* stages in Figure 1 highlight the importance of revisiting operational definitions to ensure their precision and representativeness of the focal concepts of interest. This is an iterative process that will likely continue as experience and knowledge of team behaviors improve.

Elaborate

Once an adequate representation of the desired measurement targets has been achieved, crafting the actual measurement tools (and items) can proceed (*Measurement Design* in Figure 1). The second phase of the DEEP framework, **E**laborate, focuses on the construction of the actual measurement tools/items used to assess the team performance and team process variables identified in the previous stage. A key goal of this effort is to ensure that the data collected during simulation-based activities provides rich and informative information useful for conducting a thorough assessment of a team. In this section we describe two critical areas of focus: *creating clear behavioral indicators of team performance and team process* and *targeting the lowest levels of measurement*.

Creating clear behavioral indicators of team performance and team process

There are a variety of measurement methodologies (e.g., rating scales, self-reports, behavior checklists, etc.) available to capture team performance and processes during simulation-based assessments, each with their own advantages and disadvantages.[7,14] In general, we echo the sentiments of others [39-41] who recommend the use of measures which focus on recording the occurrence of observable behaviors representative of team performance and processes whenever possible. Such data provide a rich, verifiable source of information that is oftentimes easier to translate into tangible feedback that participants can actively apply in subsequent experiences.[10] Although it may still be useful to supplement behavioral measures with self-report and other similar rating scales, these methodologies suffer from a number of undesirable biases and thus should not be relied upon as the sole source of data within a simulation-based environment.[42,43] In either case, it is strongly recommended that assessors compare each measurement item used to capture team performance or team process against the previously established operational definitions to ensure that the assessment tool is representative of and distinctly associated with its intended focal construct (see Figure 1).

When constructing behavioral items for team performance and processes, it is critical to concentrate on creating specific, easily identifiable events as opposed to generic descriptions of team activities.[11-14] The more difficult a behavior is to precisely identify, the greater the chances that data accuracy and reliability can be adversely affected by contamination from observer errors and/or rater disagreement.[36] Consequently, it is imperative to avoid ambiguous, confusing, or subjective language in behavioral item descriptions. Drawing from the substantial literatures on survey design and the development of behavioral marker systems[39-41,44,45], we present three guidelines helpful for constructing unambiguous behavioral items:

- In addition to stating the target behavior, *provide clear examples of acceptable or likely demonstrations of that behavior in each item*. For example, if assessing whether team members order medications during a scenario simulation, also provide the names of medications likely to be ordered (e.g., the sedatives and paralytics most often administered before intubation) to provide observers with concrete markers that the action has occurred.
- *Avoid double-barreled items* that specify that more than one behavior occurs in a single item. For example, the item “Team member verbalizes need to intubate to others or

begins intubation without first discussing with team mates” should be separated into two distinct items so that the specific behavior that occurred can be precisely identified.

- *Steer clear of subjective terminology*. Behavioral items including words that could be interpreted in multiple ways can lower the reliability and validity of data. For example, “interpreting an x-ray” might manifest via a lone team member reviewing results silently, one team member making facial or nonverbal expressions conveying the extent of an injury to another team member, or verbal discussion among all team members. Rather than leaving it to the observer to decide what “interpretation” entails, use precise wording that specifies exactly what the item is intended to capture (e.g., “Team member verbally communicates meaning of x-ray results with teammates”).

Targeting the lowest levels of measurement

In addition to precisely clarifying the content of one’s behavioral indicators, the level of specificity for the target of measurement is also an important consideration for team-based assessments.[10] In this case, there are two important decisions to be made concerning the desired level of measurement for one’s items: the specificity of the targeted behavioral action and the referent of the assessment item. With regards to the former, items can be tailored to capture behaviors at either more macro (e.g., “Team members were coordinated in their efforts to manage the patient's airway”) or more micro (e.g., "One team member prepared materials while another prepared the patient for intubation", "Team member assisted another with endotracheal tube placement") levels of activity. In general, our recommendation on this matter is as follows:

- *Capture multiple micro-level behaviors as opposed to fewer global behaviors* in one’s measurement of team performance and team process variables.

Not only will narrowly construing items minimize the likelihood for observer biases, but the inclusion of more rather than fewer items for a given variable can greatly improve the flexibility, accuracy, and reliability of one's measurement tool.[38] Data collected with broad measures cannot later be broken down to identify specific indicators of behavioral events, whereas data from specific behavioral events can always be aggregated to form broader indices.

With respect to the referent of an assessment item, one has the option of measuring behaviors at either the team-level (i.e., items capture whether *anyone* on the team completed a particular behavior) or the individual-level (i.e., items capture *which particular team member* completed a particular behavior). In line with the previous recommendation, we propose the following guideline:

- *Capture behaviors at the individual-level as much as possible* in one's measurement of team performance and team process variables.

Although we advise adhering to this guideline in most cases, the choice regarding individual- versus team-level measurement may be shaped by the specific purpose of assessment (i.e., only team-level effects may be of interest in some cases).[27] Nevertheless, capturing behavioral data at the individual-level offers the greatest flexibility for examining pertinent issues related to team composition and member contribution and is also better suited for interventions designed to target feedback/training towards those members who need it most.[10] Again, individual-level data can always be clustered to form team-level aggregates later, while the reverse is impossible.

Establish

Once the instruments have taken shape, it is desirable to confirm the adequacy of the measures with other knowledgeable experts to demonstrate that they sufficiently capture the intended phenomena/outcomes (*Measurement Validation,* Figure 1). Thus, the third component of DEEP, **E**stablish, involves solidifying and validating the content of the measurement tool *prior* to implementation. In doing so, adjustments to either the specification or design (or both) of the assessment instrument can be made (see Figure 1) in an effort to address any weaknesses or oversights before observational data is collected. To this end, we focus on two tasks relevant to this goal: *determining inclusion of behavioral indicators* and *demonstrating evidence of content validity using subject matter experts*.

Determining inclusion of behavioral indicators

When constructing measures of team performance and process a variety of decisions must be made that balance competing interests of methodological rigor against practical limitations of implementation.[40] In the previous section, the merits of including multiple narrow behavioral indicators to measure a given variable versus fewer broad indicators were discussed. But how many items are "too many?" Member interactions evaluated in team research can be extremely complex; consequently, one should seemingly attempt to measure as many of the nuanced ways members may behave in the team setting as possible.[46] One reason for this preference is purely statistical; more items means greater ability to demonstrate construct validity and measurement reliability.[47] Another is the ability to ask questions up-front and remove them later if they prove to be outliers or inconsequential (rather than neglecting to ask questions up-front and lamenting the lack of information later).

However, there may be cases where capturing more items is worse (or at least no better and therefore unnecessary) than asking fewer items. More specifically:

- *Team performance and process measures should generally exclude items that have low variance across teams*; in other words, when *all* or *no* teams engage in certain behaviors, those behaviors are not diagnostic of effective team performance or process and therefore should be excluded from a measure.

This is especially important for assessment within simulation-based environments that are highly structured.[13] Behaviors driven by demand characteristics in the simulation or those that are unlikely to occur given the simulation's design are primary candidates for exclusion. Measuring such behavioral items simply asks observers to invest resources in recording information that is ultimately not useful; additionally, including such items in one's analyses can diminish the power to detect effects (e.g., via biased estimates of scale reliability and relationships with covariates). One notable exception, however, is if a particular activity is considered essential and therefore there are practical reasons to assess it; nevertheless, if teams always complete this behavior, it adds no diagnostic value to the assessment.

Demonstrating evidence of content validity using subject matter experts

An important first step in validation is determining whether the content reflected in a measure of team performance or team process is representative of the task outcomes or teamwork competencies (respectively) most relevant in the scenario.[36] For this purpose, it is common to use subject matter experts (SMEs) to review the scenario design, measurement approach, and assessment tools developed for team training. Although much can be said about the use of SMEs in assessment-related activities, the present focus specifically centers on how to structure an SME's involvement and who to select.[48]

SMEs can fulfill a variety of roles in the assessment development process, including writing/contributing items or resolving ambiguities within one's operational definitions. At a minimum, however, we recommend that SMEs be used to content validate (i.e., evaluate the representativeness, consistency, and/or importance of) variables of interest and their relevance to the demonstration of effective team performance or process. To maximize the utility of SME involvement during this process, we propose the following guidelines:

- *Provide SMEs with detailed information about the assessment environment, simulation features, and scenario script* so they understand the context in which teams will be operating and will be able to interpret the representativeness of the team performance and process variables.
- *Provide SMEs with a rating scale and/or standardized questions for evaluating the representativeness of the scenario design, team process or performance variables, and specific behavioral indicators*. SMEs have the expertise to evaluate the content of an assessment, but they do not necessarily possess the skill to convey that knowledge. Thus, providing SMEs with even simple items such as "How important do you believe [some variable] is to a team's ability to effectively complete the requirements of this scenario?" helps provide a common language to validate the content of the assessments.

With respect to SME selection, any individual who possesses knowledge of, experience with, and/or insight into a particular domain or procedure could potentially serve as an SME. In assessments designed to uniquely capture both team performance and team processes, we also recommend the following when deciding who and how to select SMEs for purposes of content validation:

- *Use separate SMEs to validate team performance outcomes and team processes within the simulation environment.* In the case of team performance, SMEs should be chosen who possess a background in the technical KSAs most relevant to the training context (e.g., physicians, nurses, surgeons, etc.); for team processes, SMEs should be chosen who possess familiarity with the meaning and demonstration of non-technical teamwork competencies (team training instructors, team process researchers, etc.) relevant to the training context.

Identifying colleagues in the healthcare profession who could serve as evaluators of team performance may pose less difficulty than locating qualified SMEs to validate assessments of team processes. As a starting place, we recommend contacting authors of research articles in both the medical and psychological literatures whose research centers on team effectiveness or team functioning. Such individuals are likely to have access to greater networks of team process SMEs and can assist in the identification of willing and able experts in this domain.

Proceduralize

With the measurement content established, the next step is to ensure that the assessment tool is properly calibrated for the purposes of the assessment context (*Implementation*, Figure 1), recognizing that it may be necessary to make adjustments to either the instrument or simulation (*Simulation Design*) to accommodate this task (or both). At this stage, the team performance and team process assessments should be differentiated, well explicated, and largely validated; the final phase of the DEEP framework, **P**roceduralize, involves fine-tuning one's assessment tool to ensure it is implemented in a manner consistent with the goals of the simulation exercise. To this end, we focus specifically on two key areas: *finalizing the desired level of precision for assessment activities* and *adjusting the precision of the assessment instrument.*

Finalizing the desired level of precision for assessment activities

A significant advantage of simulation environments is that a single scenario or application can be used for a variety of purposes—such as training, research, or formal evaluation—so long as the scenario context permits participants to express focal constructs behaviorally.[49,50] However, each of these applications places different demands on the level of precision required by the assessment system. For example, in many training contexts, a measurement tool is needed that can be used to collect and interpret data quickly and in real-time to provide corrective feedback on team processes and performance.[40] In research or formal evaluations though, immediate data interpretation and feedback may not be critical; instead the most important need is for highly comprehensive and accurate coverage of specific dimensions of team performance outcomes and processes.

Consequently, an important issue is explicitly recognizing the tradeoffs inherent in determining “how deep” one goes with an assessment instrument during implementation. A measurement tool with very specific behavioral items permits a greater level of precision, which generally improves the content and construct validity of the instrument and allows one to generate more concrete and diagnostic feedback or interpretations.[11,36] However, such instruments can be resource-intensive and place significant cognitive demands on observers, often making them difficult to employ in real-time. In contrast, techniques which rely on more global assessment instruments (such as those described in the development of behavioral marker systems[39-41]) are generally easier to implement and less demanding on observers; such tools may be preferable for simulations that are simple, shorter, or require educators/professionals to provide immediate feedback following an exercise. The disadvantage of such assessments, however, is that they are often too imprecise to draw definitive conclusions about the strengths, weaknesses, and capabilities of a team; furthermore, a coarser assessment tool makes it challenging (if not impossible) to provide feedback regarding what specific activities teams engaged in that were desirable or undesirable and, therefore, provide concrete behavioral anchors for teams to improve in subsequent efforts. Lastly, assessments that make use of more global evaluations can be more prone to certain troublesome observer biases (e.g., halo effects, confirmation biases)—a critical concern if the assessment is to be used for high-stakes evaluations such as certification.

Relevant to these points, we offer the following pieces of guidance:

- *Base decisions regarding the level of assessment precision on the desired conclusions to be drawn and not solely on the ease with which the assessment can be obtained*. Practical limitations of observers and/or the simulation environment are considerations that must be recognized when designing/implementing an assessment instrument. Nevertheless, the interpretations one is able to draw, justify to external audiences, and/or provide as meaningful feedback from an assessment are only as good as the quality of the data acquired. Thus, every effort must be made to obtain data from a simulation exercise that is as comprehensive and accurate as possible. If it is not feasible to obtain the level of precision needed to support valid, reliable, and informative conclusions with traditional real-time observational methods, we recommend supplementing these efforts with alternative techniques (e.g., off-line behavioral coding of video/audio recordings, speech processing software, etc.) that permit evidence-based interpretations of team functioning.

- *Always favor greater precision in an assessment instrument when the simulation is to be used for research, formal evaluation, or training validation*. In these applications, precise and accurate data are of the highest priority.[51,52] Consequently, we also highly recommend recording (with video and audio equipment) team behavior in the simulation so that observers have the capability to conduct a thorough assessment at their own pace. Using video and audio recording also allows the same or new raters to re-check initial coding work as needed.

Adjusting the precision of the assessment instrument

A key characteristic of a "deep" assessment tool is that the precision/accuracy of observations is a function of the instrument itself; that is, a deep assessment tool is one in which its indicators, items, checklists, etc. are specific enough that virtually any observer could recognize and record the targeted outcome with little to no need for subjective interpretation (e.g., "Team member requests another member to take over chest compressions"[13]). Alternatively, the precision/accuracy of observations in a "shallow" assessment are far more dependent on the expertise and skill of the observer; that is, the items on a shallow assessment do not define the specific conditions, actions, etc. which define whether a targeted outcome occurred and thus rely on observer's interpretations of the situation ("Team members considered the requirements of others before acting"[40]). In general, we highly recommend the development and use of "deeper" assessment instruments whenever possible, recognizing that it is much easier to scale back the level of detail in an assessment than to ramp up its specificity. Should it be desirable or necessary to "go shallower" with one's assessment tools though, we offer the following guidelines:

- *Leverage rater training practices to offset tradeoffs in assessment precision*. A full discussion of the considerations involved in developing and instituting training that prepares observational coders/raters to collect data in a simulation exercise is far beyond the scope of this paper (we encourage readers to consult the plethora of resources on the topic[53-59]). Nevertheless, it is important to recognize that rater training activities are a powerful tool that can be used to maximize the validity of an assessment activity when one is adjusting the depth of an assessment instrument. As noted above, deep assessment instruments are highly structured and precise; consequently, inexperienced or less knowledgeable raters can be employed to carry out observations of the simulation, but these individuals should receive training that focuses on what and how to identify the targeted behavioral indicators and how to properly use the instrument as it is intended. Conversely, because shallower assessments allow for more subjectivity, raters must be more experienced and/or knowledgeable in the domain to ensure that the appropriate behaviors are being attended to. In these situations, it is important for rater training to focus on establishing a shared, explicit understanding of what the targeted behavioral indicators are and how to avoid common rater biases.[57,58]
- *Utilize SME judgments to justify removal or revision of items kept in an assessment.* Just as SME evaluations can be used to validate whether an assessment instrument adequately represents a focal construct, similar judgments can be used to aid decisions regarding which items (or which cluster of items) are most important, representative, and/or critical to keep in a measurement instrument and which items could be potentially removed without significant loss of information.

- *Adjust precision by reducing the number of items in an assessment instrument, but do not reduce the level of detail in the items*. Although limiting which behaviors observers should be watching for sacrifices the scope of an assessment and thus a tool's representativeness of the focal construct as a whole, lessening the specificity of the targeted observations runs the greater risk of contaminating the assessment with ambiguities and therefore impacting confidence in its construct validity.[36] This is especially true when observers are less experienced, well-trained, or knowledgeable.

CONCLUSION

The goal of this manuscript is to equip healthcare educators, researchers, and professionals with a targeted and impactful set of recommendations capable of significantly improving the validity and interpretability of assessment activities in which simulation-based systems are employed to examine aspects of team effectiveness (e.g., developmental training, formal evaluation, or experimental research). In our view, there are a number of resources available that describe related practices for developing high quality simulation-based healthcare training.[11-14] However, viewing assessment practices solely through the lens of training can obscure certain fundamental principles and techniques central to measurement design and implementation. The guidance offered by the DEEP framework to team assessment is intended to provide greater clarity to certain of these practices which, in turn, can be used to enhance the quality of assessment tools within any simulation-based application.

Although the recommendations summarized in the DEEP framework address critical aspects of assessment, there are certain caveats worth noting. First, the guidelines advanced herein do not cover all aspects of a simulation-based assessment system exhaustively nor do they cover each aspect comprehensively. For example, considerations of simulation design were not addressed in the present discussion; fortunately, there is a wealth of excellent information already available on this topic which can be easily integrated alongside the present guidelines.[7,11-13] It should also be noted that the final stages of assessment shown in Figure 1 concerning the application of appropriate quantitative/qualitative examinations of the recorded observations (*Analysis*) and subsequently determining the appropriate conclusions one can draw on the basis of those efforts (*Interpretation*) were not covered in this paper. These phases present a number of unique challenges—especially when dealing with team-based applications[60,61]—for which we believe further guidance and similar efforts to summarize best practices would be beneficial.

Second, the present framework is not, nor was it intended to be, a procedural "cookbook" that describes in detail the exact needs and steps for assessment development and implementation at every stage. For instance, the brief mention of rater training in the Proceduralize section only scratches the surface of the myriad methods, techniques, and principles one could apply to improve assessment implementation. The focused and directive recommendations summarized by the DEEP framework were purposefully chosen to provide insight into a select number of highly critical yet particularly challenging/overlooked assessment activities that, if adhered to during the course of one's assessment-related activities, can greatly enhance the precision and rigor of those efforts. We believe this is a worthwhile and important contribution of the DEEP framework.

Finally, although the process summarized in Figure 1 and emphasized in the guidelines shown in Figure 2 appears relatively sequential, the various stages of assessment are highly interrelated and will often need to be revisited or considered in parallel to effectively meet the needs of the assessment context. Planning and carrying out an effective simulation-based team assessment requires multiple iterations though each phase of assessment activity as new decisions are made. However, by understanding the important considerations, functions, and consequences of one’s assessment-related choices, the likelihood of designing a reliable and valid assessment instrument is greatly enhanced.

The development and implementation of high quality simulation-based assessment tools—particularly those which seek to capture complex events involving teams—is likely to require multiple iterations through each of the core activities emphasized within DEEP to ensure that consistency and rigor is maintained between the objectives of the evaluative context, the structure of the simulation and scenario content, and the desired information to be extracted from those experiences. The concepts and recommendations highlighted by DEEP are consistent with state-of-the-art standards in assessment practices. By elucidating and providing concise recommendations regarding their application to simulation-based team assessments, we hope they will contribute to the development of robust, reliable, and valid assessment instruments in healthcare disciplines.

Table 1

Distinguishing features and additional behavioral examples of team performance outcomes and team processes

Team performance outcomes INCLUDE...	Examples
• Demonstration of correct technical or clinical KSAs in a task situation	• Chest X-ray is correctly diagnosed as pneumonia
• Actions which produce an observable change in patient health or status	• Pelvis binding is placed (defined as time first knot tied)
• Accuracy of requested orders, treatments administered, or protocols followed	• Medications ordered for hospital-acquired pneumonia (imipenem or mirapenem)
• Quality or efficiency with which a particular team task or goal was completed	• Time patient spent in VFib (measured from time of onset to reestablishment of sinus)
Team performance outcomes DO NOT INCLUDE...	Examples
• Suggestions or discussions of possible courses of action	• Team member(s) discuss delivering oxygen to patient via non-rebreather before O2 saturation falls below 90%
• Communications or activities related to other team member's progress, status, or actions	• Team member(s) corrected someone else who wanted to give the patient paralytics first (sedatives should be given first)
• Act of sharing information between teammates	• Parapalegic status verbalized to team members (e.g., "He has a spinal fracture")
• Act of assisting teammates perform a procedure or complete a task	• Team member delivering chest compressions was replaced by a different team member after 2 minutes of performing CPR
Team processes INCLUDE...	Examples
• Communications which change the effort, goals, or awareness of other team members	• Change in heart rhythm is communicated to all members of team
• Actions which directly facilitate and/or influence the behaviors of other team members	• If one member is applying oxygen, another member holds mask in place on patient's face (pre-intubation)
• Asking for or providing verification of accuracy/appropriateness of course of action	• Verification that IV fluids are being administered to the patient (e.g., "Is the IV running?", "Did we start the IV?", "How much fluid is in?")
• Eliciting discussion, suggestions, or opinions from other team members	• Discussed which intubation medications (sedatives, paralytics) to administer (e.g., Etomidate, Versed, Ativan, Propofol, Succinylcholine, Rocuronium, Vecuronium)
Team processes DO NOT INCLUDE...	Examples
• Activities performed independently of and whose purpose/results are not shared with others	• Parapalegic status identified but not verbalized to other team members
• Speed with which team members complete task requirements	• Time until IV confirmed (asking nurse "does pt have IV" or acknowledging from sheet that patient has IV)
• Accuracy or quality of decisions made by team	• Rhythm is assessed to be tachycardia
• Dialogue/activities unrelated to team task or situation	• Team members converse about difficulty of procedure

Figure 1. Summary of procedural framework for developing and implementing simulation-based assessment tools

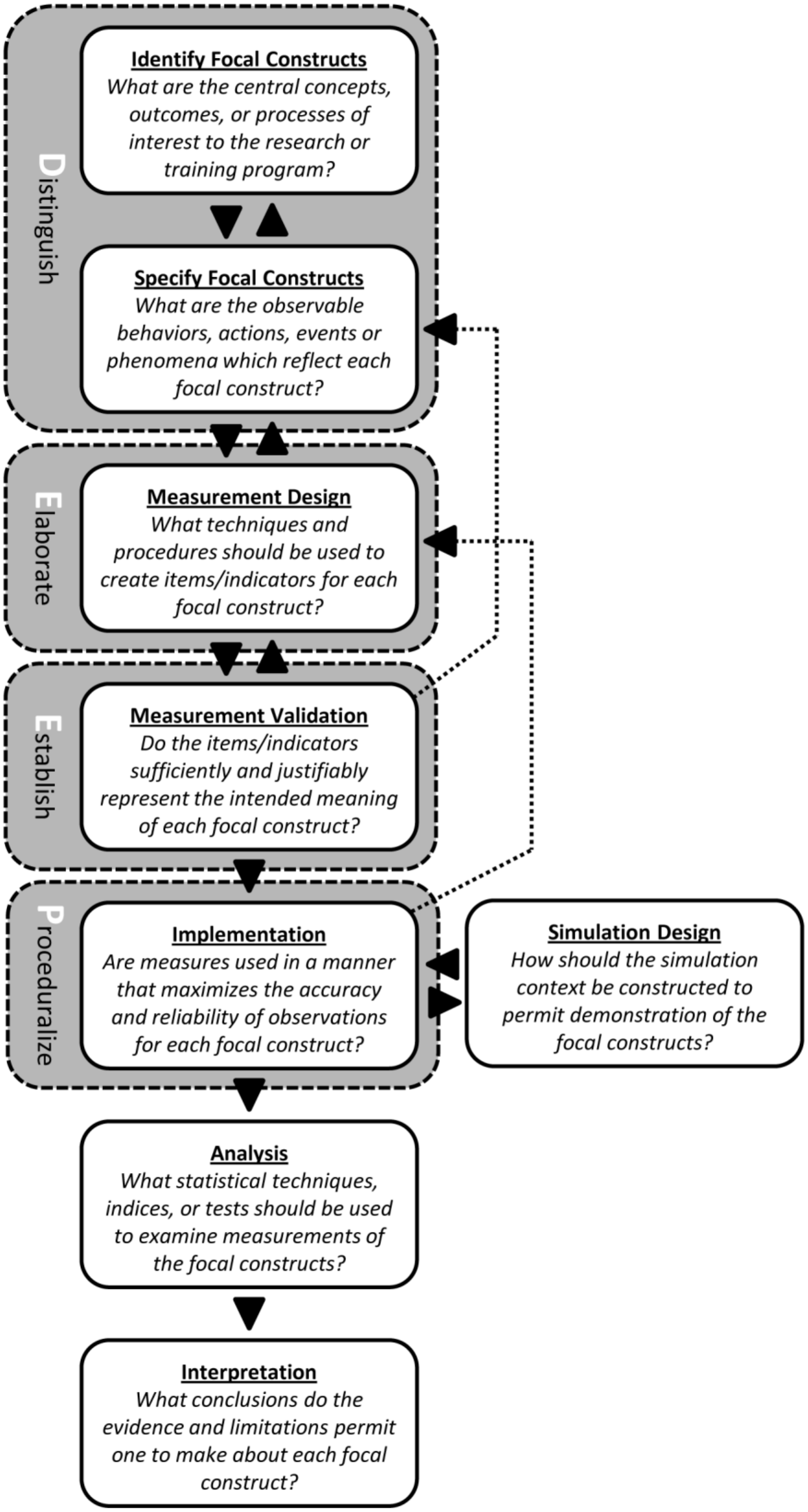
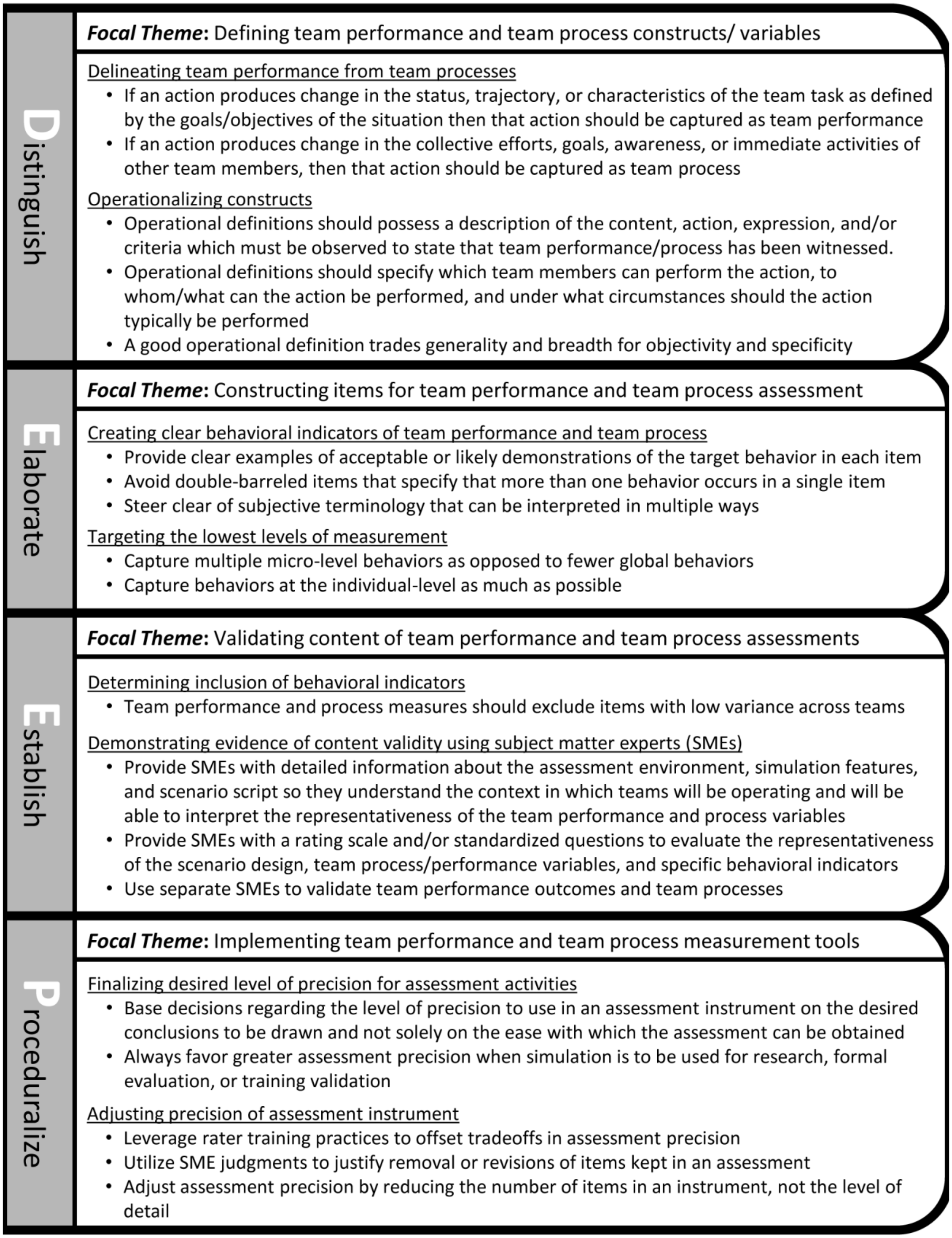

Distinguish
Identify Focal Constructs
What are the central concepts, outcomes, or processes of interest to the research or training program?
Specify Focal Constructs
What are the observable behaviors, actions, events or phenomena which reflect each focal construct?
Elaborate
Measurement Design
What techniques and procedures should be used to create items/indicators for each focal construct?
Establish
Measurement Validation
Do the items/indicators sufficiently and justifiably represent the intended meaning of each focal construct?
Proceduralize
Implementation
Are measures used in a manner that maximizes the accuracy and reliability of observations for each focal construct?
Simulation Design
How should the simulation context be constructed to permit demonstration of the focal constructs?
Analysis
What statistical techniques, indices, or tests should be used to examine measurements of the focal constructs?
Interpretation
What conclusions do the evidence and limitations permit one to make about each focal construct?

Figure 2. Summary of guidelines and recommendations from DEEP assessment framework



Distinguish

Focal Theme: Defining team performance and team process constructs/ variables

Delineating team performance from team processes

- If an action produces change in the status, trajectory, or characteristics of the team task as defined by the goals/objectives of the situation then that action should be captured as team performance
- If an action produces change in the collective efforts, goals, awareness, or immediate activities of other team members, then that action should be captured as team process

Operationalizing constructs

- Operational definitions should possess a description of the content, action, expression, and/or criteria which must be observed to state that team performance/process has been witnessed.
- Operational definitions should specify which team members can perform the action, to whom/what can the action be performed, and under what circumstances should the action typically be performed
- A good operational definition trades generality and breadth for objectivity and specificity

Elaborate

Focal Theme: Constructing items for team performance and team process assessment

Creating clear behavioral indicators of team performance and team process

- Provide clear examples of acceptable or likely demonstrations of the target behavior in each item
- Avoid double-barreled items that specify that more than one behavior occurs in a single item
- Steer clear of subjective terminology that can be interpreted in multiple ways

Targeting the lowest levels of measurement

- Capture multiple micro-level behaviors as opposed to fewer global behaviors
- Capture behaviors at the individual-level as much as possible

Establish

Focal Theme: Validating content of team performance and team process assessments

Determining inclusion of behavioral indicators

- Team performance and process measures should exclude items with low variance across teams

Demonstrating evidence of content validity using subject matter experts (SMEs)

- Provide SMEs with detailed information about the assessment environment, simulation features, and scenario script so they understand the context in which teams will be operating and will be able to interpret the representativeness of the team performance and process variables
- Provide SMEs with a rating scale and/or standardized questions to evaluate the representativeness of the scenario design, team process/performance variables, and specific behavioral indicators
- Use separate SMEs to validate team performance outcomes and team processes

Proceduralize

Focal Theme: Implementing team performance and team process measurement tools

Finalizing desired level of precision for assessment activities

- Base decisions regarding the level of precision to use in an assessment instrument on the desired conclusions to be drawn and not solely on the ease with which the assessment can be obtained
- Always favor greater assessment precision when simulation is to be used for research, formal evaluation, or training validation

Adjusting precision of assessment instrument

- Leverage rater training practices to offset tradeoffs in assessment precision
- Utilize SME judgments to justify removal or revisions of items kept in an assessment
- Adjust assessment precision by reducing the number of items in an instrument, not the level of detail

FUNDING

This work was supported, in part, by grant 1R18HS020295-01 (RF,SWJK) from the Agency for Healthcare Research and Quality.

REFERENCES

1. Kohn LT, Corrigan JM, Donaldson MS. *To err is human*. Washington, DC: National Academies Press 1999.

2. Risser DT, Rice MM, Salisbury ML, et al. The potential for improved teamwork to reduce medical errors in the emergency department. *Ann Emer Med* 1999;**34**:373-83.

3. Clancy CM, Tornberg DN. TeamSTEPPS: Assuring optimal teamwork in clinical settings. *Am J Med Qual* 2007;**22**:214-17.

4. Cheung DS, Kelly JJ, Beach C, et al. Improving handoffs in the emergency department. *Ann Emer Med* 2010;**55**:171-80.

5. Walker S, Brett S. Oiling the wheels of intensive care to reduce "machine friction:" The best way to improve outcomes? *Crit Care Med* 2010;**38**:S642-48.

6. Alonso A, Baker D, Holtzman A, et al. Reducing medical error in the military health system: How can team training help? *Human Resource Management Review* 2006;**16**:396-415.

7. Baker DP, Salas E, King H, et al. 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

8. Salas E, DiazGranados D, Weaver SJ, et al. Does team training work? Principles for health care. *Acad Emerg Med* 2008;**15**:1002-9.

9. Kendall D, Salas E. Measuring team performance: Review of current methods and consideratioon of future needs. In: Ness JW, Tepe V, Ritzer D, eds.*The Science and Simulation of Human Performance: Advances in Human Performance and Cognitive Engineering Research*. Philadelphia, PA: Elsevier 2004:307–26.

10. Kozlowski SWJ, Klein KJ. A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes. In: Klein KJ, Kozlowski SWJ, eds *Multilevel Theory, Research, and Methods in Organizations: Foundations, Extensions, and New Directions*. San Francisco, CA: Jossey-Bass 2000:3-90.

11. Rosen MA, Salas E, Silvestri S, et al. A measurement tool for simulation-based training in emergency medicine: The Simulation Module for Assessment of Resident Targeted Event Responses (SMARTER) approach. *Simul Healthc* 2008;**3**:170-79.

12. Shapiro MJ, Gardner R, Godwin SA, et al. Defining team performance for simulation-based training: Methodology, metrics, and opportunities for emergency medicine. *Acad Emerg Med* 2008;**15**:1088-97.

13. Rosen MA, Salas E, Wu TS, et al. Promoting teamwork: An event-based approach to simulation-based teamwork training for emergency medicine residents. *Acad Emerg Med* 2008;**15**:1190–98.

14. Rosen MA, Salas E, Wilson KA, et al. Measuring team performance in simulation-based training: Adopting best practices for healthcare. *Simul Healthc* 2008;**3**:33-41.

15. Bond WF, Lammers RL, Spillane LL, et al. The use of simulation in emergency medicine: A research agenda. *Acad Emerg Med* 2007;**14**:353-63.

16. Ziv A, Wolpe PR, Small SD, et al. Simulation-based medical education: An ethical imperative. *Acad Med* 2003;**78**:783-88.

17. Gaba DM. The future vision of simulation in health care. *Qual Saf Health Care* 2004;**13**(Suppl 1):i2-10.

18. Shapiro MJ, Morey JC, Small SD, et al. Simulation based teamwork training for emergency department staff: Does it improve clinical team performance when added to an existing didactic teamwork curriculum? *Qual Saf Health Care* 2004;13:417-21.

19. Salas E, Wilson KA, Burke CS, et al. Using simulation-based training to improve patient safety: What does it take? *Jt Comm J Qual Patient Saf* 2005;**31**:363-71.

20. Gaba DM, Howard SK, Flanagan B, et al. Assessment of clinical performance during simulated crises using both technical and behavioral ratings. *Anesthesiology* 1998;**89**:8-18.

21. Baker DP, Gustafson S, Beaubien J, et al. *Medical Teamwork and Patient Safety: The Evidence-based Relation*. Publication No. 05-053, Agency for Healthcare Research and Quality, 2005. Available at: http://www.ahrq.gov/qual/medteam/. Accessed Feb, 2012.

22. Henriksen K, Battles JB, Keyes MA, Grady ML, eds. *Advances in patient safety: New directions and alternative approaches* (*Vol. 3. Performance and Tools*). Publication No. 08-0034-3, Agency for Healthcare Research and Quality, 2008. Available at: http://www.ahrq.gov/qual/advances2/#v3. Accessed: Feb, 2012.

23. Salas E, Rosen MA, King H. Managing teams managing crises: Principles of teamwork to improve patient safety in the emergency room and beyond. *Theoretical Issues in Ergonomics Science* 2007;8(5):381-94.

24. Salas E, Wilson-Donnelly K, Sims D, et al. Teamwork training for patient safety: Best practices and guiding principles. In: Carayon P, ed. *Handbook of Human Factors and Ergonomics in Health Care and Patient Safety*. Mahwah, NJ: Erlbaum 2007:803-22.

25. Cannon-Bowers JA, Tannenbaum SI, Salas E, Volpe CE. Defining competencies and establishing team training requirements. In: Guzzo R, Salas E, editors. *Team Effectiveness and Decision Making in Organizations*. San Francisco, CA: Jossey-Bass 1995:333-80.

26. Cannon-Bowers JA, Salas E. A framework for developing team performance measures in training. In: Brannick MT, Salas E, Prince C, eds. *Team Performance and Measurement: Theory, Methods, and Application*. Mahwah, NJ: Erlbaum; 1997:45-62.

27. Brannick MT, Prince C. An overview of team performance measurement. In: Brannick MT, Salas E, Prince C, eds. *Team Performance and Measurement: Theory, Methods, and Application*. Mahwah, NJ: Erlbaum; 1997:3-16.

28. Gladstein, DL. Groups in context: A model of task group effectiveness. *Adm Sci Q* 1984;**29**:499-517.

29. Kozlowski SWJ, Ilgen DR. Enhancing the effectiveness of work groups and teams. *Psychol Sci Public Interest* 2006;**7**:77-124.

30. Fernandez R, Kozlowski SWJ, Shapiro MJ, et al. Toward a definition of teamwork in emergency medicine. *Acad Emerg Med* 2008;**15**:1104-12.

31. Marks MA, Mathieu JE, Zaccaro SJ. A temporally based framework and taxonomy of team processes. *Acad Manage Rev* 2001;**26**:356-76.

32. Salas E, Sims DE, Burke CS. Is there a “big five” in teamwork? *Small Group Research* 2005;**36**:555-99.

33. Yule S, Flin R, Paterson-Brown S, et al. Non-technical skills for surgeons in the operating room. A review of the literature. *Surgery* 2006;**139**:140-49.

34. Hull L, Arora S, Kassab E, et al. Observational teamwork assessment for surgery: Content validation and tool refinement. *J Am Coll Surg* 2011;**212**:234-43.

35. Flin R, Maran N. Identifying and training non-technical skills for teams in acute medicine. *Qual Saf Health Care* 2004;**13(**Suppl 1):i80-84.

36. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: Theory and application. *Am J Med* 2006;**119**:166.e7-.e16.

37. Murphy KR, Davidshofer CO. *Psychological Testing: Practices and Applications*. Upper Saddle River, NJ: Pearson 2005.

38. Hoyle R, Harris M, Judd C. *Research Methods in Social Relations*. Belmont, CA: Wadsworth 2002.

39. Rosen MA, Bedwell WL, Wildman, JL, et al. Managing adaptive performance in teams: Guiding principles and behavioral markers for measurement. *Human Resource Management Review* 2011;**21**:107-22.

40. Fletcher G, Flin R, McGeorge R, et al. Rating non-techincal skills: Developing a behavioural marker system for use in anaesthesia. *Cogn Technol Work* 2004;**6**:165-71.

41. Flin R, Martin L. Behavioural markers for crew resource management: A survey of current practice. *Int J Aviat Psychol* 2001;**11**:95-118

42. Krueger J. Enhancement bias in descriptions of self and others. *Pers Soc Psychol Bull* 1998;**24**:505-16.

43. Paulhus DL. Two-component models of socially desirable responding. *J Pers Soc Psychol* 1984;**46**:598-609.

44. Hinkin TR. A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods* 1998;**1**:104-21.

45. Spector PE. *Summated Rating Scales*. Thousand Oaks, CA: Sage 1992.

46. Smith-Jentsch KA, Johnston JH, Payne SC. Measuring team-related expertise in complex environments. In: Cannon-Bowers JA, Salas E, eds. *Making Decisions Under Stress: Implications for Individual and Team Training*. Washington, DC: American Psychological Association 1998:61-87.

47. Prince A, Brannick MT, Prince C, et al. The measurement of team process behaviors in the cockpit: Lessons learned. In: Brannick MT, Salas E, Prince C, eds. *Team performance Assessment and Measurement: Theory, Methods, and Applications*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers 1997:289-310.

48. *Subject matter experts and VV&A*. United States Department of Defense, 2000. Available at: http://vva.msco.mil/Special_Topics/SME/sme-pr.pdf. Accessed Feb, 2012.

49. Alinier G. Developing high-fidelity health care simulation scenarios: A guide for educators and professionals. *Simul Gaming* 2011;**42**:9-26.

50. Kozlowski SWJ, DeShon RP. A psychological fidelity approach to simulation-based training: Theory, research, and principles. In: Schiflett SG, Elliot LR, Salas E, Coovert MD, eds. *Scaled worlds: Development, Validation and Applications*. Burlington, VT: Ashgate Publishing 2004:75-99.

51. Neily J, Mills PD, Young-Xu YN, et al. Association between implementation of a medical team training program and surgical mortality. *JAMA* 2010;**304**:1693-1700.

52. Salas E, Weaver SJ, DiazGranados D, et al. Sounding the call for team training in health care: Some insights and warnings. *Acad Med* 2009;**84**:S128-S131.

53. Hoyt WT. Rater bias in psychological research: When is it a problem and what can we do about it? *Psychol Methods* 2000;**5**:64-86.

54. Lievens F, Sanchez JI. Can training improve the quality of inferences made by raters in competency modeling? A quasi-experiment. *J Appl Psychol* 2007;**92**:812-9.

55. Uggerslev KL, Sulsky LM. Using frame-of-reference training to understand the implications of rater idiosyncrasy for rating accuracy. *J Appl Psychol* 2008;**93**:711-19.

56. Day DV, Sulsky LM. Effects of frame-of-reference training and information configuration on memory organization and rating accuracy. *J Appl Psychol* 1995;**80**:158-67.

57. Bernardin HJ, Buckley MR. Strategies in rater training. *Acad Manage Rev* 1981;**6**:205-12.

58. Thornton GC, Zorich, S. Training to improve observer accuracy. J Appl Psychol 1980;65:351-354.

59. Castorr AH, Thompson KO, Ryan JW, et al. The process of rater training for observational instruments: Implications for intterater reliability. *Res Nurs Health* 1990;**13**:311-318.

60. Bliese P. Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In: Klein KJ, Kozlowski SWJ, eds. *Multilevel Theory, Research, and Methods in Organizations: Foundations, Extensions, and New Directions* San Francisco, CA: Jossey-Bass 2000:349-381.

61. Klein KJ, Bliese P, Kozlowski SWJ, et al. Multilevel analytical techniques: Commonalities, differences, and continuing questions. In: Klein KJ, Kozlowski SWJ, eds. *Multilevel Theory, Research, and Methods in Organizations: Foundations, Extensions, and New Directions* San Francisco, CA: Jossey-Bass 2000:512-556.