OPEN

Improving the Quality of Evaluation Data in Simulation-Based Healthcare Improvement Projects

A Practitioner's Guide to Choosing and Using Published Measurement Tools

Chiara M. Santomauro, PhD;

Andrew Hill, PhD;

Tara McCurdie, PhD;

Hannah L. McGlashan, MPsych

Summary Statement: Simulation is increasingly being used in healthcare improvement projects. The aims of such projects can be extremely diverse. Accordingly, the outcomes or participant attributes that need to be measured can vary dramatically from project-to-project and may include a wide range of nontechnical skills, technical skills, and psychological constructs. Consequently, there is a growing need for simulation practitioners to be able to identify suitable measurement tools and incorporate them into their work. This article provides a practical introduction and guide to the key considerations for practitioners when selecting and using such tools. It also offers a substantial selection of example tools, both to illustrate the key considerations in relation to choosing a measure (including reliability and validity) and to serve as a convenient resource for those planning a study. By making well-informed choices, practitioners can improve the quality of the data they collect, and the likelihood that their projects will succeed.
(*Sim Healthcare* 15:341–355, 2020)

Key Words: Healthcare improvement, quality improvement, simulation, evaluation, measurement tools, self-report measures, behavioral marker systems, psychometrics, reliability, validity.

The use of simulation in healthcare, which has ancient roots, has seen a gradual resurgence since it was "rediscovered" in the latter half of the 20th century.[1] Traditionally, simulation has been used for a range of purposes including education and training of clinical students and professionals, assessment of clinical competencies, and research.[2] More recently, simulation has also been used in healthcare improvement projects to evaluate proposed interventions, work processes, and systems before real-world implementation[3] and to detect latent safety threats (ie, unrecognized system flaws that potentially threaten patient safety) in new and existing health contexts.[4–6] Consequently, there is a growing need for simulation practitioners to be skilled in collecting evaluation data in relation to healthcare improvement work.

Broadly speaking, healthcare improvement projects focus on bettering the delivery of healthcare through the design, evaluation, and implementation of new and innovative methods. Simulation can serve as a powerful evaluation method for healthcare improvement projects, particularly when it may be impractical, inappropriate, or risky to test proposed changes or innovations in a live clinical setting. For example, Geis et al[4] conducted laboratory and in situ simulations before the opening of a new satellite emergency department. The simulations, which involved staff selected for transfer to the new facility, were used to detect potential latent safety threats (eg, having to use a single medication station for multipatient care) that otherwise may not have been detected until after the facility had opened and become fully operational. Consequently, the clinical staff and leadership were able to obtain resources and to calibrate and refine role responsibilities and environmental layouts, before the opening of the facility.

Because the outcomes of simulation-based healthcare improvement activities may lead to the implementation of changes in clinical workplaces (with potential implications for patient safety, clinical throughput, or clinician well-being), practitioners must have confidence in the data upon which those outcomes are based. More broadly, the healthcare simulation community is increasingly acknowledging the importance of tangible high-quality evaluation data.[7,8] For example, *Simulation in Healthcare* no longer accepts manuscripts for technical reports, case reports, or simulation scenarios without formal evaluations. The *Society for Simulation in Healthcare* also discourages sole reliance on low-level data (such as subjective participant reactions[9]) when forming conclusions about simulation activities.

One way to collect tangible, meaningful, and potentially publishable data is to incorporate the use of high-quality measurement tools into your evaluations. Measurement tools (also referred to as *assessment tools*, *evaluation tools*, *instruments, or measures*) can be used to gather objective or subjective data on participants' performance, behaviors, and experiences

From the Clinical Skills Development Service (C.M.S., A.H., T.M., H.L.M.), Metro North Hospital and Health Service; and School of Psychology (C.M.S., A.H., T.M.), The University of Queensland, Brisbane, Queensland, Australia.

Correspondence to: Andrew Hill, PhD, Clinical Skills Development Service, Metro North Hospital and Health Service, Herston QLD 4029, Australia (e-mail: andrew.hill@health.qld.gov.au).

The authors declare no conflict of interest.

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 (www.simulationinhealthcare.com).

DOI: 10.1097/SIH.0000000000000442

during simulation activities. Such information can potentially be used to evaluate the impact of proposed workplace interventions or changes on teams or individual roles, by assessing changes over time (eg, before vs. after implementation) or between groups or sites (eg, intervention vs. control). In the present context, these tools typically take the form of either: (*a*) self-report questionnaires completed by participants; or (*b*) behavioral marker systems (ie, measures of "concrete and observable examples of some aspect of effective or ineffective performance"[10(p1031)]) and/or global rating scales (which quantify high-level impressions or judgements[11]) completed by trained observers rating individual participants, teams, or systems of care. Even when a measurement tool is used to gather data about individuals, the results are typically aggregated across groups of participants if the purpose is to evaluate a proposed intervention, work process, or system. When used correctly, high-quality measurement tools can provide useful feedback to practitioners (eg, was intervention X successful at improving outcome Y?) without resorting to potentially unreliable or inaccurate anecdotal evidence.

A substantial number of measurement tools for assessing a broad range of outcomes relevant to healthcare improvement activities are described in the literature, with varying degrees of quality. This may be overwhelming for some simulation practitioners, who may not yet have the background knowledge or experience required to: (*a*) choose suitable, high-quality tools for their healthcare improvement projects, and reject questionable ones or (*b*) use these measurement tools appropriately. Consequently, simulation-based healthcare improvement projects may fail to yield meaningful or interpretable outcome data, potentially leading to a range of negative consequences. Most importantly, an opportunity for healthcare improvement could be lost or a substandard intervention implemented, negatively impacting the clinical workplace. Practitioners may also lose the support of coalface clinicians or jeopardize opportunities to conduct similar work in the future. A related issue is that simulation projects are typically complex and time- and resource-intensive; therefore, practitioners must be equipped to make meaningful interpretations and conclusions based on supporting data to ensure an appropriate return on that investment. Finally, projects that lack meaningful evaluation data are less likely to be published, potentially reducing the wider uptake of valuable innovations.

Because many simulation practitioners lack extensive training in experimental design and research methods,[7,12,13] the *Society for Simulation in Healthcare* (SSH) recently published a report providing guidance on research and data collection, which includes checklists and cognitive aids to support practitioners to conduct high-quality simulation studies.[7] This is an excellent resource for practitioners but does not provide recommendations of specific measurement tools for use in simulations or discuss considerations for choosing appropriate tools. To our knowledge, no prior peer-reviewed publication provides practitioners designing simulation-based healthcare improvement activities with details of a wide selection of relevant, high-quality measurement tools and guidance on how to select and administer such tools.

Importantly, all healthcare improvement initiatives, including simulation-based activities, need to be considered in the context of healthcare as a complex adaptive system or a "system of systems" that comprises "an integrated and adaptive set of people, processes and products."[14(p346)] One consequence of this complexity is that the behavior of such systems can be nonlinear and therefore unpredictable.[15] This means that in some cases, it may not be possible to simulate all aspects of the surrounding system (or systems) that, in real clinical situations, may impact on the execution or effects of an innovation. Therefore, the outcomes of simulation-based healthcare improvement exercises should be interpreted cautiously with this in mind even when every effort has been made to collect high-quality evaluation data.

Aim and Scope of the Article

This article is targeted toward simulation practitioners who may not have extensive research experience, but who nevertheless wish to conduct more rigorous evaluations of their simulation-based healthcare improvement projects, and who may even want to publish their work in peer-reviewed journals. Overall, we aim to guide these practitioners in choosing and using high-quality measurement tools for a wide variety of *constructs* relevant to healthcare improvement activities. The term "construct" can be used to refer to any intangible attribute of a person, group of people, or system. In the context of healthcare improvement, such attributes may include nontechnical skills (eg, teamwork), technical skills (eg, surgical skills), and pertinent psychological constructs (eg, cognitive workload). To achieve our aim, we provide explanations of key concepts in relation to the selection of measurement tools, including *reliability* and *validity*. To illustrate these concepts and to provide a useful resource for practitioners, we also include tabulated examples of relevant measurement tools from the literature with a summary of published reliability and validity evidence, as well as examples of potential uses for each tool in the healthcare improvement context. In addition, we provide guidance on key aspects of the data collection process that are directly related to the use of measurement tools (such as the order of tool administration).

It is also important to highlight what is not covered in this article. First, when choosing and using measurement tools for educational or vocational assessment purposes (rather than healthcare improvement), the concepts discussed herein remain relevant but additional considerations apply that are beyond the scope of this article. Second, we do not provide guidance on the development of new measurement tools. Creating a new tool from scratch can be prohibitively time-consuming and resource intensive and requires a level of expertise to which most practitioners are unlikely to have access.[16] Third, this article is not a comprehensive guide on how to conduct a simulation project—such information, including reporting and publishing guidelines, can be found elsewhere.[7,12,13,17] Fourth, this article is not a systematic review and does not provide an exhaustive list of all measurement tools that could plausibly be used in simulation-based healthcare improvement projects. Rather, we outline a selection of reasonably high-quality tools and provide guidance on selecting appropriate tools so that practitioners are better prepared when reviewing the literature for additional measures. Lastly, because our focus is on the tools, rather than on the studies that used

them, we do not assess the content or quality of studies that we cite as examples.

KEY CONSIDERATIONS FOR CHOOSING A MEASUREMENT TOOL

Once the decision has been made to conduct a simulation-based healthcare improvement project, careful consideration of the project's aims and hypotheses (or research questions) is necessary to determine which outcome(s) must be measured to evaluate whether or not the intervention was a success. The appropriate unit of analysis (eg, individual participants, teams, or systems of care) must also be determined. Although some outcomes may be relatively straightforward to measure (eg, error rates or completion times), others will require the selection of a suitable measurement tool for the *construct of interest* (ie, in most cases, a specific self-report measure, behavioral marker system, or global rating scale).

To make an informed choice, practitioners must be aware of the pros and cons associated with different types of measures. When considering a specific measurement tool, it is also essential that practitioners are able to appraise its psychometric soundness, which essentially means assessing *reliability* and *validity* for the specific context in which the tool will be used. Broadly speaking, *reliability* is the degree to which a measure can produce stable and consistent results, and *validity* is the degree to which our interpretations of scores derived from the measure are justified and defensible for their intended use.[18–22] In this section, we will address each of these considerations in turn.

Self-report Measures Versus Behavioral Marker Systems and Global Rating Scales

Self-report measures (or *self-report questionnaires*) require the participant to provide responses to a series of questions (sometimes referred to as *questionnaire items*, *test items*, or simply *items*). There are several different response formats that can be used for questionnaire items, including: dichotomous (ie, selecting between 2 response options, such as indicating whether a statement is true or false); multiple choice (ie, selecting a response from a range of options); rating scales (eg, indicating one's level of agreement with a statement); or open ended (ie, providing a written response to a question). The items within a self-report measure usually use a single consistent response format.

A unique benefit of self-report measures is that they can be used to investigate constructs that are difficult to assess objectively with physiological or behavioral measures. Hence, they are frequently used to measure psychological constructs (eg, safety attitudes), and sometimes nontechnical skills (eg, teamwork), but are not typically used to quantify technical skills. When administered appropriately, self-report measures can also offer participants anonymity, which encourages honest responses.[16] Self-report measures are also one of the easiest ways to collect data in a timely fashion, particularly if a large participant sample is required, and especially if the questionnaire is administered online, eliminating the need for time-consuming data entry.[23] Web-based survey platforms (eg, Qualtrics, SurveyMonkey) can be used to administer questionnaires quickly and easily and provide simple data exportation methods.

Despite their benefits, self-report measures also have several drawbacks that should be considered when selecting a measurement tool. First, they require participants to have a high level of self-insight; however, what people think of themselves may not reflect how they truly are.[24] Second, self-report measures are more prone to inaccurate responses because of *social desirability bias* compared with objective measures. Socially desirable responding occurs when participants deliberately choose a response that is socially acceptable but not necessarily truthful.[25] However, offering participants credible assurances of anonymity, and never looking at participants' responses in front of them, may reduce socially desirable responding.[16] Third, responses on self-report measures may be prone to priming effects, which occur when factors in people's social environment influence their thoughts and behaviors.[26] A hallmark study on susceptibility to stereotypes provides convincing evidence that priming can alter behavior: Asian-American females who completed a mathematics test performed more poorly if they had been asked questions about their gender (making their gender salient) before taking the test, but performed better if they had instead been asked questions about their ethnic identity (making their ethnic identity salient).[27] Fourth, missing data (ie, unanswered questions) can complicate analyses and lead to biased results. For example, if a subset of participants completing the *Safety Attitudes Questionnaire* (SAQ)[28] are disgruntled employees, and a substantial proportion choose not to answer items about perceptions of management because their anonymity has not been sufficiently assured, then the average of responses to this subscale would overestimate employees' satisfaction with their leaders. Finally, a problem with rating scales in particular is that they are prone to *response bias* (also known as "acquiescent responding"), that is, some people tend to gravitate toward one end of a scale (eg, positive) or one type of response (eg, agree).[29] However, many well-designed self-report tools address this issue by including "reverse-worded" (or "negative-worded") items. These items are written in such a way that a "positive" response (eg, responding *very much so* to the statement "I feel calm" in the *State-Trait Anxiety Inventory*[30]) has a similar meaning to a "negative" response to the other items (eg, responding *not at all* to the statement "I am tense"). Such items, which need to be reverse scored, can reduce participants' propensity to respond a certain way regardless of the item content and make it more apparent when they do.[31]

Behavioral marker systems contrast with self-report measures in 2 important respects. First, the focus is on concrete, observable behaviors as performance indicators.[10] Hence, they are most frequently used in healthcare to assess technical skills and nontechnical skills but are not suited to measuring psychological constructs. Second, the ratings are not completed by participants but by trained observers who rate individuals or teams on their behaviors and performance in authentic or simulated clinical scenarios, either live or on video. Rating scales are the most common response format used for behavioral marker systems, but they can also use checklists or frequency counts.[10]

Behavioral marker systems are often used for assessment or feedback purposes. In the context of healthcare improvement

work, they can also be used to generate group-level data to evaluate the impact of interventions. By assessing actual behaviors, rather than relying on participants' self-insight (which is prone to biases, as discussed previously), behavioral marker systems offer a greater degree of objectivity compared with self-report questionnaires. Measurement error can also be reduced by increasing the number of trained observers (and averaging their ratings) or by increasing the number of simulated cases or events that they observe and rate.

One of the biggest disadvantages of behavioral marker systems is that they can be significantly more resource intensive to use than self-report questionnaires. For example, the training time for observers can range from a few hours to a few days,[10] in addition to the time required to observe and code the scenarios. Another disadvantage is that observers' ratings may be influenced by cognitive biases (eg, the *halo effect*, in which positive or negative judgments made about a person with respect to one attribute can contaminate subsequent judgments about other attributes[32,33]), although there is some evidence that training can potentially reduce their impact.[10,34] In addition, simulation-based activities can trigger inauthentic behavior because of participants' awareness of being observed and/or recorded (known as the *Hawthorne effect*[35]), and this may be exacerbated by the use of behavioral marker systems because the process for obtaining participant consent may draw additional attention to the fact that their behavior is being assessed.

Some measurement tools take the form of *Global Rating Scales* (GRSs), which share similar properties (including advantages and disadvantages) with behavioral marker systems, except that they measure raters' *high-level* impressions or judgments about participants' behavior, performance, or skill-level (either *overall*, in the case of a single-item GRS, or for each of a set of subskills).[11] For example, the *Ottawa Crisis Resource Management Global Rating Scale* measures crisis resource management skill across several subskills (eg, leadership) using 7-point rating scales with different descriptive anchors for each subskill. Raters do not evaluate participants on specific tasks or components of tasks, but rather their global ability to demonstrate each subskill throughout the scenario.

One benefit of GRSs is that they can potentially be used to assess performance on a wide variety of different tasks or scenarios, because they are not tailored to specific tasks. However, being task-agnostic means that GRSs can be more susceptible to the *halo effect* (than checklists, for example), resulting in an increased likelihood of similar responses across items, which may artificially inflate the tool's internal consistency.[36,37] It is also worth noting that a GRS is sometimes embedded within a behavioral marker system. For example, the *Objective Structured Assessment of Technical Skills*[38] contains both a behavioral marker checklist and a multi-item GRS; and the *Team Emergency Assessment Measure*[39] contains a series of behavioral marker rating scales followed by a single-item GRS to capture impressions of overall nontechnical performance.

Reliability: Can the Measure Produce Stable and Consistent Results?

When we think about measuring something abstract, the analogy that springs most readily to mind is the measurement of physical dimensions or distances. For example, without necessarily thinking about it, we use a spatial metaphor in everyday speech when we talk about a period of time being "long" or "short."[40] Thus, let us consider the reliability (ie, consistency or replicability[18]) of measurements made using a ruler. If the ruler is made of wood and we measure the length of an object—for example, a pencil—repeatedly, we would expect to get the same result every time, provided that we perform the task diligently. The measurement tool itself is not contributing any unreliability to the measurements in this case. However, what if the ruler were made of rubber instead of wood? It might still give us a reasonable estimate of the pencil's length, but we might find that the estimate varies slightly from one measurement occasion to the next, although the actual length of the pencil is the same. There would be more measurement error and the reliability of our measurements would be reduced: The more elastic the rubber, the less stable and consistent the measurements would be. One consequence of this reduced reliability is that our measurements are less able to detect small differences. For example, if we carefully measure 2 pencils—one that has never been used and one that has been sharpened a few times—the measurements from the wooden ruler should correctly indicate which pencil is shorter, whereas the less reliable measurements from the rubber ruler might lead us to conclude that the 2 pencils are the same length or even that the sharpened pencil is longer than the unsharpened one.

Unfortunately, when we need to measure something less tangible, such as a nontechnical skill or a psychological attribute, even the best tools available may be more like a rubber ruler than a wooden one. To further complicate matters, just because a measurement tool *can* yield reliable measurements in some circumstances, it does not follow that it always *will* produce reliable data. This is because, from the perspective of *generalizability theory*, reliability is not simply a static attribute of a tool (or the measurements derived from it); rather, reliability depends on the context in which the tool is intended to be used.[41–43] Therefore, for any given measure you may wish to use, multiple potential sources of error variation (ie, unreliability) should be considered in assessing whether the tool is likely to yield reliable scores in the proposed *measurement situation*.[18] These potential sources of error variation are sometimes referred to as *facets*.[42] In addition to attributes of the measurement tool itself (eg, the test items, or the version of the test used if alternate forms exist), the facets of a measurement situation can also include the following: aspects of the study design (eg, the testing occasions, if participants are tested more than once); characteristics of the conditions under which participants are tested (eg, the specific tasks or patient cases that participants complete before or during the study); and the raters, if applicable (eg, for behavioral marker systems).[42–44]

A range of techniques have been developed for assessing reliability, and test developers (and other researchers) routinely publish reliability evidence. As "end-users" of measurement tools, practitioners who know how to interpret reliability evidence and select the most reliable tools for use in their particular measurement situation are more likely to be able to detect interpretable differences in their results (eg, between groups or time points).[16]

Appraising Reliability Evidence

Traditional techniques for assessing reliability center primarily on statistics referred to generically as "reliability coefficients." Different types of reliability coefficient are used to quantify different potential sources of unreliability, but all take the format of a single number with a maximum value of 1. Values closer to 1 (for the same statistic) indicate greater reliability (ie, less error variation) and values closer to 0 indicate lower reliability (ie, more error variation). Some reliability coefficients (as well as other statistics you may find reported) are accompanied by a test of statistical significance, the result of which is expressed as a *P* value (ie, probability value). In social science (including psychometrics), the conventional threshold for a result to be regarded as "statistically significant" is $P < 0.05$.[45] However, statistical significance alone is no guarantee of practical importance,[45] and the magnitude of the reliability coefficient is much more important than the associated *P* value.[18] Indeed, authors frequently choose not to report *P* values alongside relevant reliability coefficients, in which case it is (usually) safe to assume that they are statistically significant.

For our present purposes, we will consider 3 traditional forms of reliability evidence that you are likely to encounter in the literature. These address potential sources of unreliability by quantifying the degree to which: (*a*) participants give consistent responses to items intended to measure the same construct (ie, "internal consistency"); (*b*) participants' scores on a measure are similar when they complete it multiple times (ie, "test-retest reliability"); and (*c*) multiple observers using a measure provide similar ratings of participants (ie, "interrater reliability"). Table 1 provides a basic guide to interpreting these forms of evidence, including lists of terms that are sometimes used to label them in the literature, specific tips for interpretation,[46,47,50,51] and relevant examples.[48,49,52]

When applying the tips outlined in Table 1 to appraise reliability evidence, it is important to understand that the "rules of thumb" provided do not represent hard-and-fast cutoffs. For example, the most commonly reported index of internal consistency is Cronbach α (pronounced "alpha"; also known as "coefficient alpha"), which estimates the internal consistency of a scale (or subscale) from the strength of the

TABLE 1. Potential Sources of Unreliability Commonly Considered in Published Studies of Relevant Measures, Typical Traditional Forms of Reliability Evidence, Terminology, Tips for Interpretation, and Examples From Relevant Literature

Potential Source of Unreliability	Typical Traditional Form of Reliability Evidence	Relevant Terms Used in the Literature	Tips for Interpreting Reliability Evidence for Research/Evaluation Purposes	Example From Relevant Literature
Items	The degree to which participants completing a measure give consistent responses to items intended to tap into the same construct.	• Internal consistency • Internal reliability • Inter-item consistency • Alpha reliability	If the tool incorporates several subscales, each measuring a different construct (or subconstruct), the internal consistency of each subscale should be evaluated. The measure (or subscale) should have adequate internal consistency (typically, Cronbach $\alpha \geq 0.70$[46,47]).*	Participants completed the 12-item HuFSHI. Scores were consistent across items ($\alpha = 0.92$), suggesting that the 12 items were assessing the same underlying construct.[48]
Occasions	The degree to which participants' scores on a measure are similar when they complete it multiple (≥2) times.	• Test-retest reliability (or consistency) • Stability coefficient • Coefficient of stability	The time points should be sufficiently spaced out to prevent participants from relying on memory. If an event has occurred after the first time point (eg, training, intervention), subsequent attempts at the measure may be affected. The correlation between scores across multiple time points should be adequate [typically, correlation coefficient (Pearson *r* or Spearman ϱ) ≥ 0.70[47]]. The correlation should also be statistically significant ($P < 0.05$, if reported).	Participants completed the NTS on 2 occasions, 2 weeks apart. Scores were consistent across the 2 time points ($r = 0.92$ for the overall measure, and 0.77–0.87 for the 5 subscales, all P's < 0.05), suggesting that the instrument and its subscales produce stable scores over time.[49]
Raters	The degree to which multiple observers (≥2) using a measure provide similar ratings of participants.	• Interrater (or rater) reliability (or agreement) • Interscorer (or scorer) reliability (or agreement) • Interobserver (or observer) reliability (or agreement) • Interjudge (or judge) reliability (or agreement)	Observers should be appropriate for the context and may require training to use the measure. Several statistics are commonly used to assess interrater reliability. Suggested minimum values for adequate agreement are as follows (but values closer to 1 are highly desirable): • Cohen $\kappa \geq 0.60$[50†] • ICC ≥ 0.50[51‡] • Correlation coefficient (Pearson *r* or Spearman ϱ) ≥ 0.70[47] The statistic should also be statistically significant ($P < 0.05$, if reported).	Three observers rated the same participants on teamwork using the CTS. Their ratings were similar (ICC = 0.98), suggesting that the measure allows for consistent ratings.[52]

* DeVellis[46] suggests the following ranges for interpreting α: <0.60, unacceptable; 0.60–0.65, undesirable; 0.65–0.70, minimally acceptable; 0.70–0.80, respectable, 0.80–0.90, very good; and >0.90, excellent but indicating possible redundancy.
† McHugh[50] suggests the following ranges for interpreting Cohen κ: 0–0.20, no agreement; 0.21–0.39, minimal agreement; 0.40–0.59, weak agreement; 0.60–0.79, moderate agreement; 0.80–0.90, strong agreement; and ≥0.90, almost perfect agreement. In addition, McHugh[50] argues that κ values of less than 0.60 indicate inadequate agreement.
‡ Koo and Li[51] suggest the following ranges for interpreting an ICC: <0.50, poor agreement; 0.50–0.75, moderate agreement; 0.75–0.90, good agreement; and >0.90, excellent agreement.
CTS, Clinical Teamwork Scale; HuFSHI, Human Factors Skills for Healthcare Instrument; ICC, intraclass correlation; NTS, Nursing Teamwork Survey.

intercorrelations between participants' responses to the individual items.[53] Many sources suggest that α values of 0.70 or greater represent adequate internal consistency for research or evaluation purposes.[46,54,55] Nevertheless, 0.70 does not represent a fixed boundary between "reliable" and "unreliable" (for this potential source of unreliability) because reliability is relative, not absolute.[45] All else being equal, it is preferable to select a measure with an α value of, say, 0.90 over one with an α value of 0.70, although they both satisfy the rule of thumb. Similar arguments could also be made about the rules of thumb listed for other types of reliability coefficient.

It should also be noted that the rules of thumb presented in Table 1 are inappropriate for purposes other than research and evaluation, where the emphasis is usually on group-level data. For example, much higher levels of reliability should be required of a test if important decisions about an individual's future are to be made on the basis of their score, such as their suitability to practice medicine.[18,43,47] In addition, different types of reliability coefficient cannot be directly compared with one another; for example, it is meaningless to say that a measure's α reliability is higher than its test-retest reliability.

It is also important to understand that any given reliability coefficient essentially addresses only one potential source of unreliability (although it may be contaminated by other sources) and is typically derived from one particular measurement situation, which may differ in important respects from your own. However, for some measurement tools, researchers have also published studies that take an explicit *generalizability theory* perspective, assessing multiple sources of unreliability simultaneously.[18] In such research, several facets of the measurement situation are varied systematically to assess how each facet contributes to variation in the data (this is known as a *generalizability study* or *G study*), and these results can be used to determine how to improve measurement (in a *decision study* or *D study*, often published in the same article).[18,42,56] For example, for a behavioral marker system that has been subjected to a *G study*, a *D study* might yield recommendations about the number of raters required,[57,59] and/or the number of medical procedures that need to be observed,[57,60] to obtain sufficiently reliable measurement. In this context, reliability is indicated by the strength of the *generalizability coefficient*, where $G > 0.8$ is a commonly accepted rule of thumb for sufficient reliability for high-stake judgments.[61] Data from *D studies* can be extremely valuable because factors such as these can potentially even have a larger impact on reliability than the measurement tool chosen, or the number of items it contains. At the time of writing, such studies are rare compared with those that report only the more traditional forms of reliability evidence[18]; however, they are becoming more common.

When choosing between measurement tools for your project, it is advisable to carefully appraise all the available reliability evidence for each option. In so doing, you should consider all potential sources of unreliability relevant to your study and the applicability of any published reliability evidence, taking into account both (*a*) the similarities and differences between the measurement situation in the published study and your proposed study, and (*b*) relevant attributes of the study participants (eg, occupation or other pertinent demographic factors). *D* studies may also inform your decisions about how to conduct the study (eg, how many raters to use or how many procedures participants should complete), if available. Having considered all of this information, you will then be in the best position to make an informed judgment as to which measure, on balance, is most likely to yield sufficiently reliable data in your particular measurement situation.

Validity: How Defensible Are Our Interpretations of Scores Derived From the Measure?

Reliability can be regarded as a precondition for validity in the sense that unreliable measurements cannot be valid, but good reliability in itself is no guarantee of validity.[18,43,47] One key reason for this is that we cannot automatically assume that a published measure "does what it says on the tin." If we choose a measurement tool on the basis of reliability evidence alone, there is a risk that we may end up reliably measuring the wrong construct (or no construct at all). In this context, "the wrong construct" could be a related construct, an unrelated construct, or a mishmash of several constructs. For example, a questionnaire that purports to measure respondents' subjective level of cognitive workload may actually measure their subjective stress level instead. This issue could arise because workload and stress are likely to be correlated (ie, participants with higher workload also tend to have higher stress), and items that are designed to measure workload may be worded similarly to items designed to measure stress (or vice versa). Alternatively, the wrong construct could be a subset or superset of the right construct.[62] An example of this is a tool that purports to measure emotional intelligence[63] but actually only measures one dimension of the construct, such as emotional self-regulation. For the data produced using a measurement tool to be meaningful and interpretable, it is crucial that it appropriately reflects the relevant attribute of interest.

However, validity is about more than just measuring the right construct. According to the *Standards for Educational and Psychological Testing*, validity is "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests."[64(p11)] This definition, which draws heavily on the work of Samuel Messick, encapsulates the dominant contemporary conceptualization of validity.[65–70] Unpacking it highlights 3 important aspects of the current view.

First, validity is not a property of the measure itself or even of the scores derived from the measure. Rather, it is ultimately a question of whether our *interpretation* of the scores derived from the measure is justified and defensible.[19–21,65,66,68,69]

Second, how easily we can justify our interpretation of the scores will depend on the *proposed use* (also known as *intended use*).[19–21,65,66,68,69] This means that like reliability, validity should also be regarded as context dependent. For example, if the same surgical skills measure were used to (*a*) evaluate the impact of a training program on clinical performance based on group-level data and (*b*) assess a particular clinician's suitability to practice based on individual-level data, different (and stronger) validity evidence would be required to justify the use of the measure for the latter purpose. Equally, the interpretation of scores from a particular measurement tool may no longer be regarded as valid if the clinical context is substantially different from the one it was originally designed for and evaluated in.[71] For example, a tool developed to assess

teamwork in emergency department settings may not yield meaningfully interpretable scores for radiology teams.

Third, even when the intended use has been considered, validity is not an "all-or-nothing" concept. As with reliability evidence, validity evidence lies on a spectrum; that is, the metrics that measurement tools yield are not simply "valid" or "invalid," but rather they can be said to have higher or lower levels of evidence to support or refute the validity of their interpretation for particular uses.[71,72] The better, the more comprehensive, and the more logically and theoretically consistent the validity evidence is, the more confident we can be that the tool actually measures the underlying construct that it was designed to measure in the kinds of contexts in which it has been evaluated. This is crucial if we are to interpret the data appropriately and reach sound conclusions about causal relationships involving the measured construct (eg, its effect on clinicians' performance in a simulation or the effect of a simulated task on the construct). Given such complexities, it is essential that practitioners are well equipped to appraise published validity evidence to select the best available measurement tool for their project and to interpret the scores appropriately.

Appraising Validity Evidence

The current edition of the *Standards*[64] adopts a version of Michael Kane's *argument-based approach* to validation.[19–21] For measurement tool developers, this is essentially a 2-step process. The first step is to specify how the scores are proposed to be interpreted in the context of their intended use. This involves establishing a conceptual framework that describes the construct the tool is designed to measure, specifies which aspects of the construct are to be represented in the tool, and indicates how the construct is theorized to relate to other variables of interest.[64] From this framework, it is possible to specify the chain of assumptions and inferences required to go from a score on the measure to the intended use.[19] The second step involves constructing a *validity argument* by using empirical evidence (new or existing) and logical reasoning to assess the plausibility of each of these assumptions and inferences.[19,64] A claim of validity requires that every link in the chain is demonstrated to be plausible (ie, the chain is only as strong as its weakest link).

Because the chain of assumptions and inferences depends on the intended use, the evidence required to support the validity argument is likely to vary for different proposed uses.[64] This is why it can be risky to use an established measure for a purpose that is materially different from the way it was used when the existing validity evidence was gathered (eg, in a different clinical setting or for predicting future performance when all prior evidence was based on performance at around the same time that the measure was administered). As a test user, you need to satisfy yourself that the validity argument for your particular intended use is adequately supported and be able to justify this conclusion, which may also mean gathering additional validity evidence if your proposed context or score interpretation is substantially different.[64] The quality and quantity of evidence required to conclude that the validity argument is supported also depend on the stakes associated with the intended use.[64] Stronger, more rigorous evidence is required for higher-stake decisions (eg, those that involve substantial investments of money or which may impact an individual's career path or certification).

The *Standards* recognize 5 sources of evidence that can be used as part of the validity argument to support the proposed interpretation of scores for their intended uses: (*a*) the content of the measure; (*b*) response processes; (*c*) internal structure; (*d*) relations to other variables; and (*e*) the consequences of testing.[64] Notably, these categories do not represent distinct "types" of validity (nor do any subcategories). For example, relations to other variables were traditionally described using a range of terms, such as "predictive validity" (indicating a theoretically consistent statistical relationship with future performance), but such terms are now regarded as outdated and potentially misleading.[70] This is because the contemporary conceptualization regards validity as a unitary concept.[65–68] To continue the "predictive validity" example, it is now more appropriate to refer to "predictive evidence of validity" as part of an overall validity argument. Nevertheless, you will still encounter these outdated terms in older publications about established measures, and many researchers and practitioners continue to use them.[62,70] It should also be noted that "face validity" (ie, whether the measure seems valid to participants) is not really a form of validity evidence at all, although poor face validity may impact on participants' motivation.[18,73]

In the remainder of this section, we will discuss the 5 sources of validity evidence in turn. Unfortunately, you are likely to find that some authors do not link all of the validity evidence they present to an explicit validity argument, particularly in older articles.

Content of the Measure

In relation to a measurement tool, "content" refers to the themes or concepts reflected in the items or tasks, and the format and wording used.[64] The central concern for content-related evidence of validity is the extent to which the content of the tool reflects the construct that it is intended to measure.[64] Measurement tool developers may use a wide range of sources, including observational work, expert opinion, theory, and prior research, to define the content domain (ie, the complete set of skills, knowledge, attitudes, or other characteristics that make up the attribute of interest).[64] When publishing validity evidence, developers should make an explicit case for the adequacy of the correspondence between the content domain and the content of the measure.[64]

In particular, it is important that developers show that they have avoided *construct underrepresentation* (also known as *construct deficiency*), where the scores derived from the tool measure less than the intended construct (as in the emotional intelligence example hereinabove) and *construct-irrelevant variance* (also known as *construct contamination*), where they measure more.[62,64,68] An example of the latter is a test of laparoscopic surgical skills with excessively complicated English-language instructions. If inadequate comprehension detrimentally affects the scores of nonnative English speakers, then the test instructions are contributing construct-irrelevant variance. This also illustrates the general principle that the format and wording of a measure must be appropriate for the full range of potential participants.[64] However, issues with item quality can threaten validity even in the absence of equity concerns. For example, if a self-report measure contains ambiguous

or poorly worded items, this may adversely affect the defensibility of inferences made about all respondents.[72]

Response Processes

In the context of healthcare improvement work, this form of evidence is most likely to be relevant to the validity argument if your study uses a behavioral marker system or global rating scale. When observers record and assess the performance or behavior of participants using a measurement tool, we must be confident that the processes they follow are consistent with the proposed interpretation of the resulting scores; for example, we would want reassurance that the observers are applying the defined criteria associated with the measure and are not influenced by extraneous factors, such as participants' likeability.[62,64] At a minimum, there should be a defensible rationale for the response processes and clearly defined scoring criteria. As an end-user, it is your responsibility to ensure that observers are adequately trained to adhere to the relevant processes and apply the criteria appropriately.

Internal Structure

The conceptual framework for a measurement tool should include information about its hypothesized internal structure. The developers may intend the measure to be either (*a*) *unidimensional*, with all items measuring a single construct, or (*b*) *multidimensional* (also known as *multifactorial*), with different sets of items measuring different constructs or different dimensions (or *subconstructs*) of a construct.

For measurement tools intended to be *unidimensional*, the most common form of internal structure related validity evidence provided by test developers is an index of internal consistency (typically Cronbach α), as discussed in the reliability section hereinabove. Such measures can be interpreted as an indication of the extent to which the test items all tap into the same underlying construct. They can also be used to assess inter-item consistency for individual dimensions of *multidimensional* measures. However, more sophisticated methods are required to assess the degree to which the overall internal structure of a multidimensional measure is consistent with the hypothesized model. The most common of these in the context of validation studies is confirmatory factor analysis (CFA). Confirmatory factor analysis can be used to test how well a sample of participant data fits the model by assessing whether items that are closely related in theory are highly intercorrelated in practice, forming distinct "factors" as expected.[76] The results should support the hypothesized number of factors, and the specific items that load onto each factor should make theoretical sense. For example, data from the SAQ suggested that its items could be categorized into 6 factors, which mapped onto the 6 components of patient safety the measure was designed to capture.[28] The results of a CFA may also prompt post hoc model modifications (eg, deletion of items) to improve fit.[74]

Relations to Other Variables

In the context of validation studies, researchers often investigate the statistical relationships between scores generated using a measurement tool and other relevant variables. In general terms, the purpose is to gather empirical evidence to assess whether the observed relationships are consistent with the proposed score interpretation in terms of assumptions made about the identity of the underlying construct and its hypothesized relationships with external criteria, such as outcome measures.

In these investigations, the key statistic is usually a correlation coefficient. When interpreting these "validity coefficients" (as opposed to reliability coefficients), it is important to understand that a correlation can either be "positive" or "negative." A correlation coefficient represents the strength of the relationship between 2 variables, represented as a value between −1 and +1. A value of −1 indicates a perfect negative relationship (ie, higher scores on one variable are associated with lower scores on the other), 0 indicates no relationship, and +1 indicates a perfect positive relationship (ie, higher scores on one variable are associated with higher scores on the other). For example, evidence for the validity of scores on a cognitive workload measure might include a strong positive correlation with task complexity (ie, the higher the task complexity, the higher the cognitive workload) and a strong negative correlation with working memory capacity (ie, the larger a person's working memory capacity, the lower their perceived cognitive workload). Hence, it is necessary to consider whether the sign of the correlation (+ or −) is consistent with the proposed score interpretation.

In this article, we consider 4 main forms of validity evidence based on relationships with other variables that you are likely to encounter in the literature. Specifically, these are the degree to which scores on a measure (*a*) are associated with scores on an established measure intended to assess the same construct, (*b*) are associated with scores on an established measure intended to assess a similar or a theoretically related construct, (*c*) are associated with performance on a relevant outcome measure (or external criterion), and (*d*) can distinguish between different participant groups (eg, experienced vs. novice), in a manner consistent with the proposed score interpretation. Table 2 provides a basic guide to appraising these forms of evidence, including lists of associated terms used in the literature, tips for interpretation,[64–68,70–72,75] and relevant examples.[78–81] As per Table 1, the rules of thumb provided do not represent hard-and-fast cutoffs and may be inappropriate for purposes other than research and evaluation.

Consequences of Testing

Analysis of whether a measurement tool is valid for its intended use can also encompass consideration of the consequences of testing, both intended and unintended, in which any potential negative consequences are weighed against potential positive consequences.[64] In particular, unintended consequences can threaten the proposed interpretation of scores if they stem from a source of invalidity.[64] This is illustrated by the earlier example in which nonnative English speakers were systematically disadvantaged on a test of laparoscopic surgical skills due to construct-irrelevant variance.

Examples of Relevant Measurement Tools for Simulation-Based Healthcare Improvement Projects

To provide a starting point for practitioners seeking suitable measurement tools, and to illustrate of some of the reliability and validity considerations outlined in this article, we present summaries of a range of measures that are potentially relevant for use in simulation-based healthcare improvement projects, and for which (in the case of healthcare-specific

TABLE 2. Common Forms of Validity Evidence Based on Relations to Other Variables, Terminology, Tips for Interpretation, and Examples From Relevant Literature

Form of Validity Evidence	Relevant Terms Used in the Literature*	Tips for Interpreting Validity Evidence for Research/Evaluation Purposes	Example From Relevant Literature
The degree to which scores on the measure are associated with scores on an *established measure intended to assess the same construct* (in a way that is consistent with the proposed score interpretation).	• Convergent evidence (of validity) • Convergent validity (evidence)	The previously established measure must itself produce reliable scores that can be argued to be valid for the proposed score interpretation. Ideally, the correlation between the 2 measures should be large (correlation coefficient [Pearson r or Spearman ϱ] $\geq$ |0.50|[75†]). The correlation should also be statistically significant ($P < 0.05$).	Scores on the NOTSS positively correlated with scores on the ANTS—a previously established measure of nontechnical skills ($r = 0.92$, $P < 0.001$).[76]
The degree to which scores on a measure are associated with scores on an *established measure intended to assess a similar or theoretically related construct* (in a way that is consistent with the proposed score interpretation).	• Convergent evidence (of validity) • Convergent validity (evidence)	The previously established measure must itself produce reliable scores that can be argued to be valid for the proposed score interpretation. Ideally, the correlation between the 2 measures should be large [correlation coefficient (Pearson r or Spearman ϱ) $\geq$ |0.50|[75†]]. The correlation should also be statistically significant ($P < 0.05$).	Scores on the SAGAT positively correlated with scores on a traditional checklist assessment of task performance ($r = 0.81$, $P < 0.01$).[77]
The degree to which scores on a measure are associated with performance on a *relevant outcome measure or external criterion* (in a way that is consistent with the proposed score interpretation).	• Criterion-related evidence (of validity) • Criterion(-related) validity (evidence) • Concurrent evidence (of validity) • Concurrent validity • Predictive evidence (of validity)[‡] • Predictive validity[‡]	The criterion measure must itself produce reliable scores that can be argued to be valid for the proposed score interpretation. The reported statistics will vary depending on how the outcome is measured, but the relationship between the 2 measures should be statistically significant ($P < 0.05$). If applicable, the correlation should ideally be large [correlation coefficient (Pearson r or Spearman ϱ) $\geq$ |0.50|[75†]].	Scores on the NASA-TLX—a measure of subjective workload—positively correlated with the time taken to complete a task ($r = 0.75$, $P < 0.01$).[78]
The degree to which scores on a measure can distinguish between *different participant groups, eg, experienced vs. novice* (in a way that is consistent with the proposed score interpretation).	• Contrasted groups • Criterion-group approach (or strategy) • Criterion-related evidence of validity • Criterion(-related) validity (evidence) • Construct validity (in the traditional sense)[§]	The difference between the groups must be in the expected direction (eg, participants who are more experienced in the content domain would be predicted to receive better scores than novices). The difference should also be statistically significant ($P < 0.05$). Statistics reported will vary, but may include a t test, ANOVA, or Mann-Whitney U test.	An ANOVA on *Clinical Performance Tool* scores revealed that participants with more clinical experience scored higher than participants with less clinical experience ($P < 0.05$).[79]

* Many of the terms listed in this column are now regarded as outdated and potentially misleading because they imply that there are multiple "types" of validity (eg, predictive validity),[70] which is inconsistent with the contemporary unitary conceptualisation.[65–68] They have been included only to assist with the interpretation of published work that uses these terms and not as an endorsement of their continued use.

† Cohen[75] suggests the following benchmarks for classifying the strength of a correlation in the context of validity testing: |0.10| = small, |0.30| = medium, and |0.50| = large.

‡ If the outcome (or external criterion) is measured at a meaningfully later point in time, the association may be regarded as predictive evidence of validity, rather than concurrent evidence.

§ This term has been used inconsistently over time, shifting from its traditional constrained meaning (as a "type" of validity) to a general term encompassing all validity evidence,[71,72] and then to the present situation where it no longer appears in the *Standards* and its use is discouraged entirely.[64,70] Unfortunately, inconsistent usage of this term continues to afflict the literature.[70]

ANOVA, analysis of variance; ANTS, Anesthetists' Non-Technical Skills; NOTSS, Non-Technical Skills for Surgeons; SAGAT, Situation Awareness Global Assessment Technique.

tools) validity evidence has been collected using samples of clinicians. These include measures of: nontechnical skills[4,5,39,48,49,52,58–60,76,77,80–117] (see Table 3, Supplemental Digital Content 1, which outlines 17 measures, http://links.lww.com/SIH/A516); technical skills and clinical performance[40,59,81,102,112,120–127] (see Table 4, Supplemental Digital Content 2, which outlines 6 measures, http://links.lww.com/SIH/A517); and psychological constructs[4,6,28,30,78,111,126–140] (see Table 5, Supplemental Digital Content 3, which outlines 7 measures, http://links.lww.com/SIH/A518).

For each measurement tool, the tables specify what the tool is designed to measure (eg, subjective workload) and what type of tool it is (eg, self-report questionnaire). They also summarize the tool's response format, reliability evidence, and quantitative evidence of validity. However, if you are considering one of these tools for use in your own work, you will also need to consult the original articles cited in the tables to compare and contrast the context(s) in which the tool was designed and evaluated with your own measurement situation and intended use to make a reasoned judgment about the extent to which the published reliability and validity evidence is likely to generalize. The tables also contain relevant usage examples that have been drawn directly from published simulation-based healthcare improvement studies, where possible. However, we have also included several tools that have not yet been used in published healthcare improvement projects, but which have evidence of sound psychometric properties and may be appropriate for such purposes.

Much of the reliability and validity evidence presented in Tables 3 to 5 is derived from studies published by the authors of the tools. In cases where we cite evidence published by other researchers, reliability and validity coefficients may be influenced by the extent to which those researchers complied with the original instructions for administering and/or scoring the measures. Evidence from third parties should therefore be interpreted cautiously and weighted appropriately, informed by qualitative information provided by the authors about compliance with the established response processes.

KEY CONSIDERATIONS FOR USING A MEASUREMENT TOOL

Having selected a suitable measurement tool for your project, it is important to know how to use it appropriately in order for the data it produces to be meaningful and interpretable. It is beyond the scope of this article to provide exhaustive guidance on the process of designing and executing a simulation-based healthcare improvement study; rather, we outline several key considerations for successfully integrating measurement tools into your project and avoiding common pitfalls.

Adhering to the Measurement Tool Instructions

Authors of measurement tools typically provide instructions on how to administer the measure (eg, what to say to participants) and score the data (eg, how to score reverse-worded items and create subscale scores). Some measurement tools are also accompanied by advice on the training required for raters. Any deviation from these instructions may result in compromised data: At best, the data may be contaminated with random "noise" that makes real differences in the construct more difficult to detect and, at worst, the data may be completely meaningless. When relying on published reliability and validity evidence to support your choice of tool, the validity argument for your intended use can only be supported if you administer the measure as per the authors' instructions, and you can demonstrate (or reason) that the measure is appropriate for your participants.

Modifying or Customizing a Measurement Tool

There may be a temptation to modify an established measurement tool in a way that is perceived to better suit the needs of your study. For example, you might consider removing specific items that are irrelevant to the intended population or context, or rewording items to improve relevance. However, any such changes may have deleterious effects on reliability and/or validity. Consequently, when measurement tools are modified, it is common practice to collect new reliability and validity data to examine how the adaptions have affected the tool's psychometric properties. For example, the *Oxford Non-Technical Skills Scale* was originally designed for the aviation context but was modified for use in surgical contexts.[80] Subsequently, it was further modified for trauma contexts and a new scale was created—the *Trauma Non-Technical Skills scale*.[90] Reliability and validity were thoroughly assessed with each modification of the measure.

Whether you make changes to your chosen measurement tool or not, it is essential to assess whether the scores that it yields are likely to be reliable in your measurement situation and whether a defensible validity argument can be made for your intended use (taking into account any changes). In addition, it is advisable to collect as much empirical reliability and validity evidence from your own sample of participants as possible to confirm that the tool is performing as expected in your study. However, in practice, it is often only practicable to assess internal consistency,[18] and even this cannot be done meaningfully with the small participant samples often used in healthcare improvement projects.

One common modification seen in research and practice is to use one or more subscales from a measurement tool, rather than administering the entire measure. For example, Kalisch et al[51] correlated scores produced by their novel teamwork measure with scores on the teamwork subscale of an established measure, the SAQ.[28] Some common reasons for excluding particular subscales when administering a measure are time constraints and irrelevance to the project. When considering this approach, it is important to evaluate any available reliability and validity evidence that is specific to the subscale(s) you intend to use. Because removing a subscale from the context of the test it belongs to may affect participants' responses, it is again advisable to collect empirical reliability and validity evidence from your own sample if possible.

If multiple self-report measures are administered sequentially, one minor modification that is unlikely to negatively affect reliability or validity (although this can still be tested) is to ensure that the response options are presented in a consistent direction for all measures. For example, if one measure uses a rating scale with anchors ranging from "strongly disagree" to "strongly agree" and the next uses a similar rating scale with anchors presented in the opposite direction (ie, from "strongly agree" to "strongly disagree"), then flipping the direction of the scale for the second measure may avoid potential confusion and/or unintended responses from participants who may cease to pay attention to the column headers after completing several pages of questions.

Using Multiple Measurement Tools

Depending on the study, multiple measurement tools may be used to assess preexisting attributes of the participants and/or outcome measures. Where multiple outcomes of interest are measured, the data can potentially be used to analyze relationships between them. For example, Gilfoyle et al[100] tested the effectiveness of a team training intervention on pediatric resuscitation teams' clinical performance and teamwork with 2 behavioral marker systems (the *Clinical Performance Tool*[79] and the *Clinical Teamwork Scale*[52]) and assessed the relationship between these 2 measures.

However, the use of multiple measurement tools is not without its costs. An important consideration when planning any study involving time-poor clinicians as participants is how long it will take them to complete. In addition, factors such as varying levels of language and comprehension skills may affect completion times and should be taken into account. The use of multiple behavioral marker systems can also create complications. If the scoring is conducted off-line (eg, by coding video recordings), the use of multiple measures will increase the time it takes observers to complete the scoring, and therefore the financial cost of the study. Worse still, any attempt by

individual observers to complete multiple behavioral marker systems concurrently (whether in real time during a simulation event, or during video-based coding) is likely to negatively impact the quality of the data, given the competing demands on their attention.

Finally, regardless of what type of measure is used, it is important to resist any temptation to include multiple measures of the same construct or large numbers of measures for which you have no specific hypotheses or research questions, in the hope that one of them will produce a significant result.[141] Such "fishing expeditions" are likely to yield seemingly significant results that are actually the product of chance and are therefore not reproducible. Nevertheless, it may be legitimate to use multiple measures of the same construct if their inclusion is supported by a convincing rationale (eg, to combine the data from multiple measures to increase overall data quality).

Sequencing Measurement Tools and Other Tasks

When using multiple measurement tools, the order of administration should be carefully chosen to reduce the likelihood of carryover effects (eg, due to priming, as discussed earlier). For example, Lee and Grant[142] manipulated the location of an item assessing self-rated health in a questionnaire and found that responses differed depending on whether it was presented before or after items about participants' chronic health conditions. If the act of responding to "Questionnaire A" may influence responses to "Questionnaire B," but not vice versa, then "Questionnaire B" should be administered first. However, if 2 (or more) measurement tools are likely to have similar carryover effects on one another, then the order of administration should be counterbalanced (ie, arranged so that the same number of participants in each group or condition receive each possible order) or randomized if counterbalancing is not possible (eg, because the number of possible orders exceeds the sample size).

Practitioners should bear in mind that other tasks or procedures that participants complete during the study (ie, not just measurement tools) can also potentially create or be subject to carryover effects. For example, a demographic questionnaire that asks participants questions about their professional background might make their membership of particular occupational group (eg, nurses or surgeons) more salient, potentially impacting on their subsequent performance in a simulation-based multidisciplinary teamwork task. Hence, if there is any plausible potential for a carry-over effect, it would be safer to collect such data at the end of the study if carryover cannot occur in the opposite direction (eg, a participant's occupation will not change because they participate in a teamwork exercise).

In addition to minimizing carryover effects, tasks and measurement tools must also be sequenced and timed appropriately to address the hypotheses or research questions of the study. For example, if we used the NASA Task Load Index (NASA-TLX) to assess surgical team members' subjective cognitive workload during a simulated procedure, it would be important to administer the measure as soon as practicable after completion of the procedure, while the experience is still fresh in participants' minds. If we then wanted to compare their workload during another procedure (eg, to test the impact of an alternative workflow), we would need to administer the *NASA-TLX* in exactly the same way at exactly the same time relative to the simulation, in order for the data to be comparable.

More generally, we recommend having participants complete *all* self-report measures during the testing session, rather than afterward, even if they are administered online. As well as providing a better reflection of participants' current state, this also ensures that data collection is completed in a timely manner and eliminates the need to expend additional resources following-up with participants. This approach is also likely to maximize the response rate (and therefore the final sample size) and potentially avoid biased data because of systematic attrition (eg, if participants who performed poorly in a simulation were less likely than others to complete the self-report measures at home).

Linking Data From Measurement Tools

To ensure that all potential options for data analysis are available, each participant should be assigned their own unique code to link all of their data together, which may include demographics, responses to self-report measures, data from behavioral marker systems, and performance data (eg, task completion times), among other things. For example, in a surgical simulation study where cognitive workload is an outcome of interest, it might be necessary to link each participant's role in the simulation (eg, instrument nurse, surgeon) to their scores on the outcome measures (eg, the *NASA-TLX*) if the intervention under investigation is likely to have different impacts on different roles. More generally, standard statistical methods for evaluating relationships between 2 or more variables in a study assume that each participant's data on the relevant variables can be linked.

To maintain participant confidentiality, it is necessary to use participant codes that do not incorporate the participants' names or other personal information that could reveal their identity. The simplest solution is to assign sequential numbers to participants. However, if data are to be collected at multiple time points, it can be useful to use a system for creating codes that can be replicated reliably by the participants themselves (eg, the first 3 letters of their mother's maiden name, followed by 2 digits representing their birth month). Nevertheless, complete data linkage will only be possible if the codes are recorded in all project documents (paper or electronic) containing study data, including self-report measures, demographic questionnaires, video filenames, and coding sheets. Because participant confidentiality is also an ethical issue, it is important to comply with local human research ethics board guidelines surrounding data linkage and storage (eg, they may require researchers to store video securely because participants could be identified[7]).

CONCLUSIONS

As simulation is increasingly being used in healthcare improvement projects, there is a growing need for simulation practitioners to be well equipped to incorporate appropriate measurement tools into their work. This article provides a practical introduction and guide to the key considerations of relevance to practitioners when selecting and using such tools. It also offers a substantial selection of example tools, both to illustrate the key considerations in relation to choosing a

measure (including reliability and validity) and to serve as a convenient practical resource for those planning a study. Although no single article could ever provide a comprehensive discussion of all matters related to the reliability, validity, and appropriate use of measurement tools (which could fill an entire book), this article should suffice to equip practitioners with enough knowledge to make better decisions and therefore to increase the quality of the data they collect and the likelihood that their simulation-based healthcare improvement projects will be successful. It may also stimulate further interest, reading, and professional development.

REFERENCES

1. Owen H. *Early Examples of Simulation in Training and Healthcare. Simulation in Healthcare Education*. Cham, Switzerland: Springer; 2016:9–19.
2. Aggarwal R, Mytton OT, Derbrew M, et al. Training and simulation for patient safety. *BMJ Qual Saf* 2010;19(suppl 2):i34–i43.
3. Barlow M, Dickie R, Morse C, Bonney D, Simon R. Documentation framework for healthcare simulation quality improvement activities. *Adv Simul (Lond)* 2017;2:19.
4. Geis GL, Pio B, Pendergrass TL, Moyer MR, Patterson MD. Simulation to assess the safety of new healthcare teams and new facilities. *Simul Healthc* 2011;6(3):125–133.
5. Patterson MD, Geis GL, Falcone RA, LeMaster T, Wears RL. In situ simulation: detection of safety threats and teamwork training in a high risk emergency department. *BMJ Qual Saf* 2013;22(6):468–477.
6. Medwid K, Smith S, Gang M. Use of in-situ simulation to investigate latent safety threats prior to opening a new emergency department. *Saf Sci* 2015;77:19–24.
7. Calhoun A, Cendan J, Dong C, et al. Empowering the inexperienced researcher: a summary report and expert recommendations. *Soc Simul Healthc* 2017.
8. Scerbo MW. Some changes in store for the journal. *Simul Healthc* 2017;12(2):67–68.
9. Kirkpatrick D, Kirkpatrick J. *Evaluating Training Programs: The Four Levels*. 2nd ed. San Francisco, CA: Berrett-Koehler Publishers; 2006.
10. Dietz AS, Pronovost PJ, Benson KN, et al. A systematic review of behavioural marker systems in healthcare: what do we know about their attributes, validity and application? *BMJ Qual Saf* 2014;23(12):1031–1039.
11. Ilgen JS, Ma IW, Hatala R, Cook DA. A systematic review of validity evidence for checklists versus global rating scales in simulation-based assessment. *Med Educ* 2015;49(2):161–173.
12. Kardong-Edgren S, Dieckmann P, Phero JC. Simulation Research Considerations. In: Palaganas JC, Maxworthy JC, Epps CA, Mancini ME, eds. *Defining Excellence in Simulation Programs*. South Holland, the Netherlands: Wolters Kluwer; 2015:615–623.
13. White ML, Peterson DT. Research in healthcare simulation. In: Palaganas JC, Maxworthy JC, Epps CA, Mancini ME, eds. *Defining Excellence in Simulation Programs*. South Holland, the Netherlands: Wolters Kluwer; 2015:604–614.
14. Brathwaite B, Gamage E, Hall S, Rajagopalan K, Tybinski M, Kopec D. Hospitals and health-care systems: methods for reducing errors in medical systems. In: Gorod A, White BE, Vernon I, Gandhi SJ, Sauser B, eds. *Case Studies in System of Systems, Enterprise Systems, and Complex Systems Engineering*. Boca Raton, FL: Taylor & Francis; 2015:345–375.
15. Gao J, Dekker S. Concepts and models of safety, resilience, and reliability. In: Sanchez JA, Barach P, Johnson JK, Jacobs JP, eds. *Surgical Patient Care*. New York, NY: Springer; 2017:25–37.
16. Marshall G. The purpose, design and administration of a questionnaire for data collection. *Radiography* 2005;11(2):131–136.
17. Cheng A, Kessler D, Mackinnon R, et al. Reporting guidelines for health care simulation research: extensions to the CONSORT and STROBE statements. *Adv Simul (Lond)* 2016;1:25.
18. Hogan TP. *Psychological Testing: A Practical Introduction*. 3rd ed. Hoboken, NJ: John Wiley & Sons; 2015.
19. Kane M, Burns M. The argument-based approach to validation. *School Psychol Rev* 2013;42(4):448–457.
20. Kane MT. An argument-based approach to validity. *Psychol Bull* 1992;112(3):527–535.
21. Kane MT. Current concerns in validity theory. *J Educ Mes* 2001;38(4):319–342.
22. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: a practical guide to Kane's framework. *Med Educ* 2015;49(6):560–575.
23. Jones S, Murphy F, Edwards M, James J. Doing things differently: advantages and disadvantages of web questionnaires. *Nurse Res* 2008;15(4):15–26.
24. Paulhus DL, Vazire S. The self-report method. In: Robins RW, Fraley C, Krueger RF, eds. *Handbook of Research Methods in Personality Psychology*. New York, NY: Guilford Press; 2007:224–239.
25. Gittelman S, Lange V, Cook WA, et al. Accounting for social-desirability bias in survey sampling: a model for predicting and calibrating the direction and magnitude of social-desirability bias. *J Advert Res* 2015;55(3):242–254.
26. Molden DC. Understanding priming effects in social psychology: what is "social priming" and how does it occur? *Soc Cogn* 2014;32:1–11.
27. Shih M, Pittinsky TL, Ambady N. Stereotype susceptibility: identity salience and shifts in quantitative performance. *Psychol Sci* 1999;10(1):80–83.
28. Sexton JB, Helmreich RL, Neilands TB, et al. The safety attitudes questionnaire: psychometric properties, benchmarking data, and emerging research. *BMC Health Serv Res* 2006;6:44.
29. Hinz A, Michalski D, Schwarz R, Herzberg PY. The acquiescence effect in responding to a questionnaire. *Psychosoc Med* 2007;4:Doc07.
30. Spielberger CD, Gorsuch RL, Lushene RE, Vagg PR, Jacobs GA. *Manual for the State-Trait Anxiety Inventory STAI (form Y)*. Palo Alto, CA: Consulting Psychologists Press; 1983.
31. Woods C. Careless responding to reverse-worded items: implications for confirmatory factor analysis. *J Psychopathol Behav Assess* 2006;28(3):186–191.
32. Thorndike EL. A constant error in psychological ratings. *J Appl Psychol* 1920;4(1):25–29.
33. Nisbett RE, Wilson TD. The halo effect: evidence for unconscious alteration of judgments. *J Pers Soc Psychol* 1977;35(4):250–256.
34. Flin R, Martin L. Behavioral markers for crew resource management: a review of current practice. *Int J Aviat Psychol* 2001;11(1):95–118.
35. McCambridge J, Witton J, Elbourne DR. Systematic review of the Hawthorne effect: new concepts are needed to study research participation effects. *J Clin Epidemiol* 2014;67(3):267–277.
36. Bould MD, Crabtree NA, Naik VN. Assessment of procedural skills in anaesthesia. *Br J Anaesth* 2009;103(4):472–483.
37. Chong L, Taylor S, Haywood M, Adelstein BA, Shulruf B, Huh S. The sights and insights of examiners in objective structured clinical examinations. *J Educ Eval Health* 2017;14(34):1–8.
38. Reznick R, Regehr G, MacRae H, Martin J, McCulloch W. Testing technical skill via an innovative "bench station" examination. *Am J Surg* 1997;173(3):226–230.
39. Cooper S, Cant R, Porter J, et al. Rating medical emergency teamwork performance: development of the Team Emergency Assessment Measure (TEAM). *Resuscitation* 2010;81(4):446–452.
40. Starr A, Srinivasan M. Spatial metaphor and the development of cross-domain mappings in early childhood. *Dev Psychol* 2018;54(10):1822–1832.
41. Cronbach LJ, Gleser GC, Nanda H, Rajaratnam N. *The Dependability of Behavioural Measurements: Theory of Generalizability for Scores and Profiles*. New York, NY: John Wiley & Sons; 1972.

42. Shavelson RJ, Webb NM. Generalizability Theory. In: Green JL, Camilli G, Elmore PB, eds. *Handbook of Complementary Methods in Education Research*. Mahwah, NJ: Lawrence Erlbaum Associates; 2012:302–322.
43. Murphy KR, Davidshofer CO. *Psychological Testing: Principles and Applications*. Upper Saddle River, NJ: Prentice Hall; 1988.
44. Boulet JR. Generalizability Theory: Basics. In: Everitt BS, Howell DC, eds. *Encyclopedia of Statistics in Behavioral Science* vol. Volume 2. Chichester, United Kingdom: John Wiley & Sons; 2005:704–711.
45. Field A. *Discovering Statistics Using SPSS*. 3rd ed. Thousand Oaks, CA: Sage Publications; 2009.
46. DeVellis RF. *Scale Development: Theory and Applications*. Thousand Oaks, CA: Sage Publications; 2012.
47. Kaplan RM, Saccuzzo DP. *Psychological Assessment and Theory: Creating and Using Psychological Tests*. Boston, MA: Wadsworth Cengage Learning; 2013.
48. Reedy GB, Lavelle M, Simpson T, Anderson JE. Development of the human factors skills for healthcare instrument: a valid and reliable tool for assessing interprofessional learning across healthcare practice settings. *BMJ Simul Technol Enhanc Learn* 2017;3(4):135–141.
49. Kalisch BJ, Lee H, Salas E. The development and testing of the nursing teamwork survey. *Nurs Res* 2010;59(1):42–50.
50. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)* 2012;22(3):276–282.
51. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med* 2016;15(2):155–163.
52. Guise JM, Deering SH, Kanki BG, et al. Validation of a tool to measure and promote clinical teamwork. *Simul Healthc* 2008;3(4):217–223.
53. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika* 1951;16(3):297–334.
54. Scholtes VA, Terwee CB, Poolman RW. What makes a measurement instrument valid and reliable? *Injury* 2011;42(3):236–240.
55. Kirk RE. *Statistics: An Introduction*. Austin, TX: Holt, Rinehart, and Winston. Inc.; 1990.
56. Webb NM, Shavelson RJ. Generalizability Theory: Overview. In: Everitt BS, Howell DC, eds. *Encyclopedia of Statistics in Behavioral Science* vol. Volume 2. Hoboken, NJ: John Wiley & Sons; 2005:717–719.
57. Beard JD, Marriott J, Purdie H, Crossley J. Assessing the surgical skills of trainees in the operating theatre: a prospective observational study of the methodology. *Clin Govern Int J* 2011;16(3):1–162.
58. Crossley J, Marriott J, Purdie H, Beard J. Prospective observational study to evaluate NOTSS (Non-Technical Skills for Surgeons) for assessing trainees' non-technical performance in the operating theatre. *Br J Surg* 2011;98(7):1010–1020.
59. Spanager L, Beier-Holgersen R, Dieckmann P, Konge L, Rosenberg J, Oestergaard D. Reliable assessment of general surgeons' non-technical skills based on video-recordings of patient simulated scenarios. *Am J Surg* 2013;206(5):810–817.
60. Spanager L, Konge L, Dieckmann P, Beier-Holgersen R, Rosenberg J, Oestergaard D. Assessing trainee surgeons' nontechnical skills: five cases are sufficient for reliable assessments. *J Surg Educ* 2015;72(1):16–22.
61. Crossley J, Davies H, Humphris G, Jolly B. Generalisability: a key to unlock professional assessment. *Med Educ* 2002;36(10):972–978.
62. Miller LA, Lovler RL. *Foundations of Psychological Testing: A Practical Approach*. 5th ed. Thousand Oaks, CA: Sage Publications; 2016.
63. O'Connor PJ, Hill A, Kaya M, Martin B. The measurement of emotional intelligence: a critical review of the literature and recommendations for researchers and practitioners. *Front Psychol* 2019;10:1116.
64. Association AER, Association AP, Education NCoMi, (U.S.) JCoSfEaPT. *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association; 2014.
65. Messick S. Test validity and the ethics of assessment. *Am Psychol* 1980;35(11):1012–1027.
66. Messick S. Evidence and ethics in the evaluation of tests. *Educ Res* 1981;10:9–20.
67. Messick S. Validity. In: Linn RL, ed. *Educational Measurement*. 3rd ed. New York, NY: American Council on Education/Macmillan; 1989:13–103.
68. Messick S. Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am Psychol* 1995;50(9):741–749.
69. Messick S. Standards of validity and the validity of standards in performance asessment. *Educ Meas* 1995;14(4):5–8.
70. Newton PE, Shaw SD. Standards for talking and thinking about validity. *Psychol Methods* 2013;18(3):301–319.
71. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med* 2006;119(2):166.e7–166.e16.
72. Downing SM. Validity: on the meaningful interpretation of assessment data. *Med Educ* 2003;37(9):830–837.
73. Cohen RJ, Swerdlik ME. *Psychological Testing and Assessment: An Introduction to Tests and Measurement*. 9th ed. New York, NY: McGraw-Hill Education; 2013.
74. Brown TA. *Confirmatory Factor Analysis for Applied Research*. 2nd ed. New York, NY: Guilford Publications; 2014.
75. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Earlbaum Associates; 1988.
76. Saleh GM, Wawrzynski JR, Saha K, et al. Feasibility of human factors immersive simulation training in ophthalmology: the London pilot. *JAMA Ophthalmol* 2016;134(8):905–911.
77. Hogan MP, Pace DE, Hapgood J, Boone DC. Use of human patient simulation and the situation awareness global assessment technique in practical trauma skills assessment. *J Trauma Acute Care Surg* 2006;61(5):1047–1052.
78. Rubio S, Díaz E, Martín J, Puente JM. Evaluation of subjective mental workload: a comparison of SWAT, NASA-TLX, and workload profile methods. *Appl Psychol* 2004;53(1):61–86.
79. Donoghue A, Nishisaki A, Sutton R, Hales R, Boulet J. Reliability and validity of a scoring instrument for clinical performance during Pediatric Advanced Life Support simulation scenarios. *Resuscitation* 2010;81(3):331–336.
80. Mishra A, Catchpole K, McCulloch P. The Oxford NOTECHS system: reliability and validity of a tool for measuring teamwork behaviour in the operating theatre. *BMJ Qual Saf* 2009;18(2):104–108.
81. Moorthy K, Munz Y, Forrest D, et al. Surgical crisis management skills training and assessment: a simulation-based approach to enhancing operating room performance. *Ann Surg* 2006;244(1):139–147.
82. Nicksa GA, Anderson C, Fidler R, Stewart L. Innovative approach using interprofessional simulation to educate surgical residents in technical and nontechnical skills in high-risk clinical scenarios. *JAMA Surg* 2015;150(3):201–207.
83. Briggs A, Raja AS, Joyce MF, et al. The role of nontechnical skills in simulated trauma resuscitation. *J Surg Educ* 2015;72(4):732–739.
84. Yule S, Flin R, Maran N, Rowley D, Youngson G, Paterson-Brown S. Surgeons' non-technical skills in the operating room: reliability testing of the NOTSS behavior rating system. *World J Surg* 2008;32(4):548–556.
85. Yule S, Flin R, Paterson-Brown S, Maran N, Rowley D. Development of a rating system for surgeons' non-technical skills. *Med Educ* 2006;40(11):1098–1104.
86. Fletcher G, Flin R, McGeorge P, Glavin R, Maran N, Patey R. Rating non-technical skills: developing a behavioural marker system for use in anaesthesia. *Cogn Technol Work* 2004;6(3):165–171.
87. 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(5):580–588.
88. Jankouskas T, Bush MC, Murray B, et al. Crisis resource management: evaluating outcomes of a multidisciplinary team. *Simul Healthc* 2007;2(2):96–101.
89. Morgan PJ, Kurrek MM, Bertram S, LeBlanc V, Przybyszewski T. Nontechnical skills assessment after simulation-based continuing medical education. *Simul Healthc* 2011;6(5):255–259.

90. Steinemann S, Berg B, DiTullio A, et al. Assessing teamwork in the trauma bay: introduction of a modified "NOTECHS" scale for trauma. *Am J Surg* 2012;203(1):69–75.

91. Steinemann S, Berg B, Skinner A, et al. In situ, multidisciplinary, simulation-based teamwork training improves early trauma care. *J Surg Educ* 2011;68(6):472–477.

92. Endsley MR. Measurement of situation awareness in dynamic systems. *Hum Fac* 1995;37(1):65–84.

93. Lavelle M, Abthorpe J, Simpson T, Reedy G, Little F, Banerjee A. MBRRACE in simulation: an evaluation of a multi-disciplinary simulation training for medical emergencies in obstetrics (MEmO). *J Obstet Gynaecol* 2018;38(6):781–788.

94. Morgan P, Tregunno D, Brydges R, et al. Using a situational awareness global assessment technique for interprofessional obstetrical team training with high fidelity simulation. *J Interprof Care* 2015; 29(1):13–19.

95. Ballangrud R, Persenius M, Hedelin B, Hall-Lord ML. Exploring intensive care nurses' team performance in a simulation-based emergency situation, - expert raters' assessments versus self-assessments: an explorative study. *BMC Nurs* 2014;13(1):47.

96. Kim J, Neilipovitz D, Cardinal P, Chiu M. A comparison of global rating scale and checklist scores in the validation of an evaluation tool to assess performance in the resuscitation of critically ill patients during simulated emergencies (abbreviated as "CRM simulator study IB"). *Simul Healthc* 2009;4(1):6–16.

97. Kim J, Neilipovitz D, Cardinal P, Chiu M, Clinch J. A pilot study using high-fidelity simulation to formally evaluate performance in the resuscitation of critically ill patients: the University of Ottawa Critical Care Medicine, High-Fidelity Simulation, and Crisis Resource Management I Study. *Crit Care Med* 2006;34(8):2167–2174.

98. Malec JF, Torsher 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(1):4–10.

99. Fransen AF, van de Ven J, Merién AE, et al. Effect of obstetric team training on team performance and medical technical skills: a randomised controlled trial. *BJOG* 2012;119(11):1387–1393.

100. Gilfoyle E, Koot DA, Annear JC, et al. Improved clinical performance and teamwork of pediatric interprofessional resuscitation teams with a simulation-based educational intervention. *Pediatr Crit Care Med* 2017;18(2):e62–e69.

101. Letchworth PM, Duffy SP, Phillips D. Improving non-technical skills (teamwork) in post-partum haemorrhage: a grouped randomised trial. *Eur J Obstet Gynecol Reprod Biol* 2017;217:154–160.

102. Miller D, Crandall C, Washington CIII, McLaughlin S. Improving teamwork and communication in trauma care through in situ simulations. *Acad Emerg Med* 2012;19(5):608–612.

103. Frengley RW, Weller JM, Torrie J, et al. The effect of a simulation-based training intervention on the performance of established critical care unit teams. *Crit Care Med* 2011;39(12):2605–2611.

104. Millward L, Ramsay K. *Measuring Team Performance: From a Cognitive and Motivational Perspective-A Pilot Study of an Evaluation Tool*. Guildford, UK: Centre for Employee Research University of Surrey; 1998.

105. Millward LJ, Jeffries N. The team survey: a tool for health care team development. *J Adv Nurs* 2001;35(2):276–287.

106. Weller J, Frengley R, Torrie J, et al. Evaluation of an instrument to measure teamwork in multidisciplinary critical care teams. *BMJ Qual Saf* 2011;20:216–222.

107. Weller J, Shulruf B, Torrie J, et al. Validation of a measurement tool for self-assessment of teamwork in intensive care. *Br J Anaesth* 2013;111(3):460–467.

108. Couto TB, Kerrey BT, Taylor RG, FitzGerald M, Geis GL. Teamwork skills in actual, in situ, and in-center pediatric emergencies: performance levels across settings and perceptions of comparative educational impact. *Simul Healthc* 2015;10(2):76–84.

109. Rovamo L, Nurmi E, Mattila MM, Suominen P, Silvennoinen M. Effect of a simulation-based workshop on multidisplinary teamwork of newborn emergencies: an intervention study. *BMC Res Notes* 2015;8:671.

110. Rubio-Gurung S, Putet G, Touzet S, et al. In situ simulation training for neonatal resuscitation: an RCT. *Pediatrics* 2014;134(3):e790–e797.

111. Sørensen JL, van der Vleuten C, Rosthøj S, et al. Simulation-based multiprofessional obstetric anaesthesia training conducted in situ versus off-site leads to similar individual and team outcomes: a randomised educational trial. *BMJ Open* 2015;5(10):e008344.

112. Healey AN, Undre S, Vincent CA. Developing observational measures of performance in surgical teams. *Qual Saf Health Care* 2004;13(Suppl 1): i33–i40.

113. Sevdalis N, Lyons M, Healey AN, Undre S, Darzi A, Vincent CA. Observational teamwork assessment for surgery: construct validation with expert versus novice raters. *Ann Surg* 2009; 249(6):1047–1051.

114. Undre S, Sevdalis N, Healey AN, Darzi A, Vincent CA. Observational teamwork assessment for surgery (OTAS): refinement and application in urological surgery. *World J Surg* 2007;31(7):1373–1381.

115. Morgan PJ, Tregunno D, Pittini R, et al. Determination of the psychometric properties of a behavioural marking system for obstetrical team training using high-fidelity simulation. *BMJ Qual Saf* 2012;21(1):78–82.

116. Tregunno D, Pittini R, Haley M, Morgan P. Development and usability of a behavioural marking system for performance assessment of obstetrical teams. *Qual Saf Health Care* 2009;18(5):393–396.

117. Kalisch BJ, Aebersold M, McLaughlin M, Tschannen D, Lane S. An intervention to improve nursing teamwork using virtual simulation. *West J Nurs Res* 2015;37(2):164–179.

118. 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(5):631–639.

119. Brogaard L, Hvidman L, Hinshaw K, et al. Development of the team OBS-PPH–targeting clinical performance in postpartum hemorrhage. *Acta Obstet Gynecol Scand* 2018;97(6):677–687.

120. Rovamo L, Mattila MM, Andersson S, Rosenberg P. Assessment of newborn resuscitation skills of physicians with a simulator manikin. *Arch Dis Child Fetal Neonatal Ed* 2011;96(5):F383–F389.

121. van der Heide PA, van Toledo-Eppinga L, van der Heide M, van der Lee JH. Assessment of neonatal resuscitation skills: a reliable and valid scoring system. *Resuscitation* 2006;71(2):212–221.

122. Reid J, Stone K, Brown J, et al. The simulation team assessment tool (STAT): development, reliability and validation. *Resuscitation* 2012;83(7):879–886.

123. McReynolds MC, Mullan PB, Fitzgerald JT, Kronick S, Oh M, Andreatta P. On-site simulation training improves nurses' first-responder cardiopulmonary resuscitation performance: traveling simulation program. *Ann Behav Sci Med Educ* 2013;19(1):8–13.

124. Kara CO, Mengi E, Tümkaya F, Ardıç FN, Şenol H. Adaptation of "objective structured assessment of technical skills" for adenotonsillectomy into Turkish: a validity and reliability study. *Turk Arch Otorhinolaryngol* 2019;57(1):7–13.

125. Levy A, Donoghue A, Bailey B, et al. External validation of scoring instruments for evaluating pediatric resuscitation. *Simul Healthc* 2014;9(6):360–369.

126. Harvey A, Nathens AB, Bandiera G, LeBlanc VR. Threat and challenge: cognitive appraisal and stress responses in simulated trauma resuscitations. *Med Educ* 2010;44(6):587–594.

127. LeBlanc VR, MacDonald RD, McArthur B, King K, Lepine T. Paramedic performance in calculating drug dosages following stressful scenarios in a human patient simulator. *Prehosp Emerg Care* 2005;9(4):439–444.

128. Battiste V, Bortolussi M. Transport pilot workload: a comparison of two subjective techniques. *Proc Hum Factors Ergon Soc Annu Meet* 1988;32(2):150–154.

129. Hart SG, Staveland LE. Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. *Adv Psychol* 1988;52:139–183.

130. Tsang PS, Velazquez VL. Diagnosticity and multidimensional subjective workload ratings. *Ergonomics* 1996;39(3):358–381.

131. Xiao Y, Wang Z, Wang M, Lan Y. The appraisal of reliability and validity of subjective workload assessment technique and NASA-Task Load Index. *Chin J Ind Hygi Occup Dis* 2005;23(3):178–181.

132. Weigl M, Stefan P, Abhari K, et al. Intra-operative disruptions, surgeon's mental workload, and technical performance in a full-scale simulated procedure. *Surg Endosc* 2016;30(2):559–566.

133. Wilson MR, Poolton JM, Malhotra N, Ngo K, Bright E, Masters RS. Development and validation of a surgical workload measure: the surgery task load index (SURG-TLX). *World J Surg* 2011;35(9):1961–1969.

134. Deilkås ET, Hofoss D. Psychometric properties of the Norwegian version of the Safety Attitudes Questionnaire (SAQ), generic version (short form 2006). *BMC Health Serv Res* 2008;8:191.

135. Paltved C, Bjerregaard AT, Krogh K, Pedersen JJ, Musaeus P. Designing in situ simulation in the emergency department: evaluating safety attitudes amongst physicians and nurses. *Adv Simul (Lond)* 2017;2:4.

136. Patterson MD, Geis GL, LeMaster T, Wears RL. Impact of multidisciplinary simulation-based training on patient safety in a paediatric emergency department. *BMJ Qual Saf* 2013;22(5):383–393.

137. Riley W, Davis S, Miller K, Hansen H, Sainfort F, Sweet R. Didactic and simulation nontechnical skills team training to improve perinatal patient outcomes in a community hospital. *Jt Comm J Qual Patient Saf* 2011;37(8):357–364.

138. Baker DP, Amodeo AM, Krokos KJ, Slonim A, Herrera H. Assessing teamwork attitudes in healthcare: development of the TeamSTEPPS teamwork attitudes questionnaire. *Qual Saf Health Care* 2010;19: e49–e52.

139. Costa AC, Anderson N. Measuring trust in teams: development and validation of a multifaceted measure of formative and reflective indicators of team trust. *Eur J Work Organ Psychol* 2011;20(1):119–154.

140. Sawyer T, Laubach VA, Hudak J, Yamamura K, Pocrnich A. Improvements in teamwork during neonatal resuscitation after interprofessional TeamSTEPPS training. *Neonatal Netw* 2013;32(1):26–33.

141. Wicherts JM, Veldkamp CL, Augusteijn HE, Bakker M, Van Aert RC, Van Assen MA. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: a checklist to avoid p-hacking. *Front Psychol* 2016;7:1832.

142. Lee S, Grant D. The effect of question order on self-rated general health status in a multilingual survey context. *Am J Epidemiol* 2009;169(12):1525–1530.