Journal of Healthcare Simulation

Check for updates

Journal of Healthcare Simulation

HOW TO

How to... measure teamwork in healthcare simulation

Jan B. Schmutz[1], Mirko Antino[2]

[1]Department of Psychology, University of Zurich, Zurich, Switzerland
[2]Departamento de Psicobiologia y Metodología en Ciencias del Comportamiento, Universidad Complutense de Madrid, Madrid, España

Corresponding author: Jan B. Schmutz, jan.schmutz@psychologie.uzh.ch

https://johs.org.uk/article/doi/10.54531/QMJP1895

ABSTRACT

Teamwork is vital for patient safety and clinical performance, yet measuring it reliably in healthcare simulation remains challenging. This article offers a practical overview of three common approaches to assess teamwork: survey measures, systematic behavioural observation (SBO) and Behaviourally Anchored Rating Scales (BARS). Surveys capture how team members perceive their teamwork and are best suited for shared psychological states such as trust or cohesion. SBO records what teams actually do in real time, allowing detailed analysis of interaction patterns. BARS combine observation with evaluative judgement, providing structured and feedback-oriented assessments of teamwork quality. The article also highlights key issues in measurement quality, including reliability, validity, rater training and level of analysis. No single method captures teamwork fully; the choice depends on purpose and resources. Combining methods often yields the most complete understanding and supports targeted debriefing, improved training and ultimately, better patient care.

Key points

- Teamwork is essential for clinical performance and patient safety but challenging to measure reliably in simulation.
- Three complementary measurement methods can be used: surveys, systematic behavioural observation (SBO) and Behaviourally Anchored Rating Scales (BARS).
- Surveys assess perceptions and shared psychological states; behavioural observation captures real actions; BARS evaluate the quality of observed behaviours.
- High-quality teamwork measurement requires clear definitions and checks for validity and reliability, including rater training.
- Combining multiple methods provides the most comprehensive picture of teamwork and supports targeted debriefing and improvement.

Introduction

Teamwork is crucial for medical teams. Research has shown that teamwork is strongly related to clinical performance [1], while teamwork failures can have devastating consequences for patients. At the same time, training teams leads to better clinical outcomes [2,3]. Teamwork is also often the central focus of healthcare simulation training, most prominently in the form of crew resource management programmes [4].

Despite its importance, educators and researchers often struggle to effectively measure teamwork in simulation settings. Two main reasons contribute to

Submission Date: 28 October 2025
Accepted Date: 13 May 2026
Published Date: 11 June 2026

this challenge. First, an underlying framework and clear definitions of what should be measured are often missing. Second, many educators lack methodological knowledge about how teamwork can be assessed systematically.

The goal of this article is to provide a practical overview of methods for measuring teamwork in healthcare simulation. While we cannot provide a comprehensive account of all possible methods and analytical considerations, we focus on the three most common approaches: survey measures, systematic behavioural observation (SBO) and Behaviourally Anchored Rating Scales (BARS). We also refer to further literature that can guide the choice of a specific method.

Relevance to healthcare simulation

The economist Peter Drucker famously stated: 'You can't manage what you don't measure'. This insight is highly relevant for simulation-based education. Teamwork and interprofessional collaboration are often the primary focus of simulation training; yet surprisingly little emphasis is placed on systematically and reliably measuring these aspects during training.

We identify four reasons why measuring teamwork can offer significant benefits. First, teamwork assessments *help reveal training needs* by making teamwork visible. They highlight specific strengths and weaknesses, which in turn enable educators to tailor debriefings and future training sessions more effectively. Second, systematic measurement can *enhance the quality of debriefings*. Data from surveys, observations or rating scales provide objective reference points to enrich discussions. In particular, BARS offer concrete behavioural examples of effective teamwork that can be directly incorporated into training. Third, measurement allows educators to *monitor progress over time*. Repeated assessments can show whether a team improves across training cycles or whether a specific intervention has been successful. Finally, appropriate teamwork measures *support research and benchmarking*. They enable comparisons across teams, hospitals or training centres and facilitate scientific studies on the effectiveness of simulation-based education.

Defining what to measure

The first step in measuring teamwork is to decide which aspects are most relevant. The term *teamwork* is often used loosely in the literature, and clear definitions are sometimes lacking. To provide structure, we draw on the input–process–output (IPO) model of team performance, which describes how inputs (e.g. knowledge, team composition [5]) are transformed into outcomes through interdependent actions within the team [6]. In healthcare simulation, the focus is typically on these interdependent actions – so-called team processes. It is useful, however, to distinguish team processes from emergent states. Team processes refer to verbal (e.g. communication) and behavioural (e.g. coordination) interactions among team members that help organize and carry out clinical tasks, transforming individual contributions into effective collective performance aimed at a shared goal, such as high-quality patient care. In contrast, emergent states describe the shared cognitive, motivational and emotional characteristics of a team that develop and change over time, depending on the team's context, interactions and experiences. Unlike team processes, which focus on how members interact, emergent states reflect the qualities of the team. Examples include psychological safety, the shared belief that it is safe to speak up [7] or team cohesion, a shared bond and commitment among team members that reflects their attraction to the team and their motivation to achieve its goals together.

This distinction is important because it directly affects which measurement methods are suitable. Emergent states, by definition, are internal to team members and cannot be directly observed. To measure them, surveys and self-reports are necessary. Team processes, however, involve observable actions and communication, making them accessible to external observers who can code and quantify them.

In the following, we introduce three common approaches for measuring teamwork: survey measures, SBO and BARS. We also provide guidance on how to apply these methods effectively in healthcare simulation. Table 1 provides a summary of the three methods, including advantages, limitations and examples.

Across all three approaches, measurement quality is a central concern. Regardless of whether teamwork is assessed through self-report surveys, SBO or BARS, the resulting data must be both reliable and valid to support meaningful interpretation. Reliability refers to the consistency of measurement: a method should produce similar results when applied under comparable conditions, such as across repeated simulations or different raters observing the same team [8]. Validity concerns whether a measure actually captures the aspect of teamwork it is intended to assess [8,9]. This requires clear conceptual definitions, appropriate operationalization of constructs or behaviours and careful attention to potential sources of bias, such as rater subjectivity, recall bias or social desirability. Without adequate reliability and validity, observed differences in teamwork may reflect measurement artefacts rather than true differences in team functioning. For this reason, considerations of measurement quality are essential when selecting and applying any teamwork assessment method in healthcare simulation.

Survey measures

Surveys are a common tool to assess teamwork [10]. They consist of several questions, called items, that together measure different specific teamwork-related concepts such as psychological safety or team cohesion. Each of these concepts is known as a construct, and each item represents one small aspect of it [9]. When combined, the items form a scale that provides an overall score for how strongly a team or individual demonstrates that construct. Although surveys can, in principle, be completed by external raters, we use the term 'survey' here exclusively to refer to self-reported questionnaire measures that capture team members' subjective perceptions, attitudes or experiences.

Table 1: Overview of three teamwork measurement approaches

Aspect	Survey measures	Systematic behavioural observation	Behaviourally anchored rating scales
Purpose	Capture team members' perceptions, attitudes or experiences	Capture observable behaviours and interactions in real time	Assess the *quality* of observable teamwork behaviours
Data source	Team members (individuals' perception aggregated to team level)	Trained observers that rate individual and/or team behaviours	Trained raters/observers (usually directly referred to team level constructs)
Data type	Subjective self-assessed ratings	Behavioural frequencies, sequences or categories	Judgement-based ratings anchored in behavioural exemplars
Timing of assessment	After a simulation (post-scenario questionnaires) or training (post-training questionnaire)	During the simulation (real-time coding or video analysis)	During or immediately after the simulation (performance assessment)
Advantages	Easy to administer, can cover a broad spectrum of psychological constructs (attitudes, behaviour, emergent states etc.), efficient for large samples and longitudinal designs	Rich behavioural detail, real-time data, high ecological validity because the behaviour of interest is directly observed	Combines objectivity with evaluative judgement, provides structured feedback, shared performance language and benchmarking
Limitations	Response bias, limited behavioural validity	Labour-intensive, requires coding schemes, can only capture behaviour that is visible from the outside	Time-consuming to develop and validate, requires training, can only evaluate behaviour that is visible from the outside
Example	Psychological safety [33], Teamwork [11]	Coding speaking up behaviour during a simulated case [34]	ANTS (Anaesthetists' Non-Technical Skills), NOTECHS [25]

Most teamwork constructs are emergent states that develop through interaction and are shared among team members. Because these states or climate variables cannot be observed directly, surveys rely on the perceptions of the team members themselves. For example, to assess cohesion, each member might rate statements such as 'Our team works together as one unit'. These individual responses are valuable because teamwork is built from the shared experiences and perceptions of the people who form the team.

Once all members have completed the survey, their responses can be combined to represent the team as a whole. The simplest way to do this is to calculate the team's average score, which reflects the group's shared view. Educators can also look at differences within the team (e.g. standard deviation). Large differences in how team members respond may suggest that not everyone experiences teamwork in the same way. Both the average score and the variation between members can offer useful insights for training and debriefing.

Examples of established survey tools to measure teamwork in healthcare are the self-assessment teamwork tool (SATT) [11] or TeamSTEPPS [12]. Many survey items are phrased to capture general perceptions of teamwork (e.g. 'Team members communicate clearly and effectively with one another'). When surveys are used in simulation settings, however, it is necessary to adapt the *frame of reference* of each item. This is a crucial step to ensure construct validity, particularly when the aim is to assess teamwork during a specific scenario rather than general team functioning [13]. In such cases, items should explicitly reference the simulation episode. For example, the item above can be adapted to: '*During the last scenario*, team members *communicated* clearly and effectively with one another'. This contextualization should be applied consistently across all items to ensure a shared and unambiguous reference point.

The same principle applies to measures of emergent states. For instance, an item from Amy Edmondson's psychological safety scale [14] – 'Members of this team are able to bring up problems and tough issues' – can be adapted to a simulation context as: '*During the scenario*, members of this team *were* able to bring up problems and tough issues'. When selecting which aspects of teamwork to measure, it is also important to consider whether the construct is likely to change within the observed time frame. Teamwork behaviours, such as communication, can reasonably be expected to improve across repeated simulation scenarios and debriefings (e.g. from scenario one to scenario three). In contrast, emergent states such as psychological safety are typically more stable within the same team and context and are therefore less likely to fluctuate substantially across closely spaced scenarios, as they usually require more time and repeated experiences to change.

Measurement quality considerations of survey measures

For survey-based teamwork measures, measurement quality primarily concerns the internal consistency of the scale, the adequacy of construct coverage and the appropriateness of aggregating individual responses to the team level.

Reliability in surveys is most: assessed as internal consistency, indicating the degree to which items intended to measure the same construct are interrelated. High internal consistency suggests that items capture a coherent

underlying concept, such as psychological safety or team cohesion. This is typically quantified using indices such as Cronbach's α or McDonald's ω [9]. In addition, test–retest reliability may be relevant when the same team completes a survey repeatedly under comparable conditions, for example across identical simulation scenarios, to ensure that observed changes reflect meaningful differences rather than measurement noise.

Survey validity concerns whether the survey measures what it claims to measure. A valid teamwork scale includes items that cover all important aspects of the construct. For example, a scale designed to assess teamwork effectiveness should include questions about communication, coordination and decision-making. If one of these elements is missing, the measure probably does not fully capture the concept (depending on the underlying theory or framework). Validity also means that the scores are not strongly influenced by unrelated factors such as social desirability (answering in a way that appears socially acceptable rather than truthful) or recall bias (difficulty remembering accurately what happened, especially in high-pressure scenarios).

When interpreting team survey data, it is important to remember that responses come from individuals who work within teams. Team members' answers are often similar because they share the same environment and experiences [15]. Statistical checks, such as intra-class correlations (ICC(1), ICC(2); see Bliese [16]), can help determine whether it is appropriate to combine individual responses into a single team score. In practice, this means that before interpreting a team's average level of cohesion or psychological safety, educators should confirm that team members generally agree in their perceptions [16].

Systematic behavioural observation

SBO is a powerful method for studying real-time team processes, particularly in dynamic environments such as acute care. It involves systematically counting and recording the occurrence of specific, predefined behaviours during a simulation or debriefing, for example, instances of speaking up, coordination or reflection. To do this, researchers use a coding scheme that defines each behaviour in clear, observable terms. Such a scheme can be adapted from existing frameworks or developed specifically for the context under study [17]. Coding schemes can also be hierarchical, meaning that coders first identify who performs the behaviour (e.g. nurse, resident) and then specify what behaviour was shown. Using this structured approach, observers capture and code behaviours, typically from video or audio recordings, to quantify how interaction patterns unfold over time [18]. Compared with survey measures, systematic observation allows researchers to assess actual, context-embedded behaviour rather than perceptions or retrospective judgements. For example, the Co-ACT framework quantifies explicit and implicit coordination patterns in acute care teams [19] and the TuRBO system captures in-action reflection cycles during ongoing care episodes [20]. Together, these approaches demonstrate how SBO enables researchers to examine complex, emergent team processes such as coordination or reflection directly from behaviour in context.

In practice, SBO requires several design decisions concerning what to code (for example, reflection, coordination or leadership behaviours), how to code (for example, continuous versus interval coding) and how to train coders [18]. In interval coding, the observation period is divided into fixed time windows (for example, 10 seconds), and raters record whether a behaviour occurred within each interval (for example, 'Behaviour A' in interval 1: yes/no). In continuous event coding, by contrast, each behaviour is coded exactly when it occurs and marked with a time stamp (for example, 'Behaviour A' at 11 minutes 22 seconds). The choice of method depends on the level of detail needed and the study's purpose. Video recordings allow for richer and more complex coding schemes because coders can rewatch sequences, clarify ambiguities and capture overlapping behaviours. Live coding, in contrast, is more practical in training settings but typically requires a smaller number of clearly defined categories to observe to remain manageable.

Measurement quality and analysis considerations of systematic behavioural observations

Measurement quality depends strongly on coder training, precise behavioural definitions and interrater reliability, which is commonly assessed with Cohen's κ, with values of 0.70 or higher considered acceptable [18,21]. Reliability is further enhanced through piloting and iterative calibration sessions to align coders' interpretations. Tools such as *Noldus Observer XT* (Noldus Information Technology, Wageningen, The Netherlands) and *Mangold INTERACT* (Mangold International GmbH, Arnstorf, Germany) support the timestamping and systematic extraction of behavioural data. However, a simple Excel sheet (for video coding) or a paper–pencil version might suffice for simple coding schemes. Further, observers need to rehearse with example videos, resolve coding disagreements in calibration meetings and update definitions when ambiguities emerge.

Behavioural observation schemes are typically characterized by high face validity, as they rely on clearly labelled and explicitly defined, observable behaviours that appear directly linked to the intended teamwork construct [22]. Because of this strong face validity, additional forms of validity evidence are sometimes neglected in observational research. However, like survey measures, face validity alone is insufficient and further validity evidence (e.g. links to performance or convergence with other teamwork measures) is desirable [20].

Once the data is coded, it can be analysed in several ways. Researchers can aggregate the frequencies or durations of specific behaviours, compare high- and low-performing teams or examine interaction patterns over time using advanced approaches such as lag-sequential or pattern recognition analyses [23,24]. These analyses provide detailed insights into how team processes unfold and how they relate to performance in simulated or clinical environments.

Behaviourally anchored rating scales

Building on the distinction above between survey methods that capture self-rated perceptions and SBO that records what teams actually do or say, BARS translate those observations into an assessment of how well teamwork behaviours are performed. Like systematic observation, an external rater observes and evaluates behaviour, but instead of simply counting its occurrence, BARS require a qualitative judgement of performance (for example, 1 = poor to 5 = excellent). Rather than asking for a general rating such as 'good' or 'poor', BARS provide concrete behavioural examples – called anchors – at each scale point that illustrate what low, moderate and high levels of performance look like. These anchors improve interrater consistency, make evaluations more transparent and support targeted feedback.

Widely used BARS frameworks such as Anaesthetists' Non-Technical Skills (ANTS) [25,26] and Non-Technical Skills for Surgeons (NOTECHS) [27,28] illustrate the approach: they define core dimensions (e.g. teamwork, decision-making, situation awareness, task management) and offer anchored descriptors (e.g. 'Team members participate actively in checking procedures') that help faculty judge the quality of behaviours during scenarios and debriefings. Because anchors make expectations explicit, BARS are especially useful for guiding debriefs ('what high-quality teamwork looked like here') and for monitoring improvement across sessions. BARS complement surveys and SBO by offering a practical, reliable way to convert observed actions into standardised judgements of teamwork competence.

Measurement quality considerations of BARS

The quality of BARS data depends on how the instruments are developed and how they are applied. One option is to develop a new BARS tailored to a specific teamwork process, which is a time-intensive process involving expert interviews, critical incident analysis and extensive pilot testing to ensure content validity [29,30]. Alternatively, educators can use or adapt existing instruments such as ANTS or NOTECHS to their specific simulation context, which allows for a more practical and resource-efficient approach.

Because BARS rely on concrete behavioural anchors that illustrate different levels of performance, they typically exhibit high face validity, as the anchors appear directly linked to the teamwork behaviours being assessed. Like SBO, however, face validity alone might be insufficient, and additional validity evidence (e.g. construct or criterion validity) is needed to support meaningful interpretation of the ratings.

Regardless of the path chosen, rater training is essential so that observers interpret behavioural anchors consistently and apply them in the same way across teams. Regular assessments of interrater reliability using indices such as intraclass correlations help verify consistency across raters [31]. Ongoing calibration sessions and clear scoring guidelines further support reliability, while testing for construct and criterion validity.

Choosing the right level of assessment: Team or individual?

An important yet often overlooked question when measuring teamwork is at what level the assessment should take place: the team or the individual. The choice depends on the purpose of the evaluation and the underlying theoretical assumptions about what the intervention or measurement aims to capture.

When evaluating a team-level intervention (e.g. a simulation-based team training), the appropriate level of analysis is the team, because it is expected that the intervention has an impact on the whole team. Here, data are aggregated across team members, and comparisons (e.g. pre–post training) are made using team-level scores (i.e. one teamwork score per team). Consequently, the sample size corresponds to the number of teams rather than individual participants. This approach reflects the assumption that teamwork phenomena, such as shared mental models and coordination, are emergent properties that exist at the team level rather than as individual traits or behaviours.

In contrast, when training or measurement focuses on individual skills, attitudes or behaviours, the unit of analysis should be the individual. For example, individual speaking-up behaviour, such as how often or how effectively a single team member voices concerns or challenges decisions during a simulation, can be assessed for each participant separately. In this case, pre–post comparisons are conducted at the individual level, with the number of participants representing the sample size.

However, because individuals are nested within teams, their responses are not statistically independent; members of the same team often influence each other. For instance, an individual's performance might differ depending on whether they work in a highly experienced or an unfamiliar team, although the individual's skills remain the same. Similarly, individual speaking-up behaviour may vary across team contexts: the same person may voice concerns more readily in a psychologically safe team than in a hierarchical or unfamiliar team. As a result, two individuals working in the same team are likely to show more similar levels of speaking-up behaviour or performance than two individuals from different teams, even when their individual skills are comparable. To account for this dependency, multilevel analytical techniques such as hierarchical linear modelling (HLM) or mixed-effects models should be used, as these explicitly partition variance between the individual and team levels [32].

The issue of levels is particularly relevant for survey measures, since each team member typically completes the questionnaire individually. However, similar considerations apply to SBO and BARS. In systematic observation, for example, coders can use a hierarchical scheme that records both *who* performs a behaviour and *what* the behaviour is, thereby allowing analyses at both the individual and team levels. Alternatively, observers can focus on collective team behaviour without distinguishing between members, resulting in team-level data. Likewise, most existing BARS instruments in healthcare are designed to evaluate teamwork as a collective performance (e.g. 'the team

reevaluated its procedures'), but it is also possible to adapt them to assess individual contributions using the same anchored scales (e.g. 'Person A reevaluated the procedures').

Selecting the right method for simulation research and training

Choosing the most suitable method to measure teamwork depends on what aspect of teamwork you want to understand and how the results will be used. Survey measures are ideal for capturing how team members feel about their teamwork, such as their sense of trust in the team, cohesion or psychological safety. SBO is best when the focus is on what teams do in practice, allowing researchers or educators to record and quantify real behaviours as they unfold during a simulation (e.g. speaking up behaviour). BARS bridge these two perspectives by combining observation with evaluative judgement, offering a structured way to assess the quality of observed teamwork behaviours.

No single approach is superior in every situation. Instead, each method provides a different lens for understanding teamwork. Educators may use surveys to explore how team members experience collaboration, while researchers might combine BARS with behavioural observation to gain a richer picture of interaction patterns and performance. Using several methods together can provide the most complete insight into team functioning and learning.

Ultimately, the method should match both the purpose of the measurement and the resources available. A short survey may be most practical for routine evaluation, whereas a detailed observational study can offer a deeper understanding in research projects. The overview in Table 1 should help educators and researchers choose the right approach. Whichever method is chosen, the key is to apply it systematically, ensure measurement quality and use the results to strengthen teamwork training and, ultimately, patient care.

Suggestions for further reading

- Brauner E, Boos M, Kolbe M, editors. The Cambridge handbook of group interaction analysis. Cambridge: Cambridge University Press. 2018.
- Fletcher G, Flin R, McGeorge P, Glavin R, Maran N, Patey R. Anaesthetists' Non-Technical Skills (ANTS) system: handbook v1.0. Aberdeen: University of Aberdeen. 2012. Available from: https://research.abdn.ac.uk/wp-content/uploads/sites/14/2019/03/ANTS-Handbook-2012-1.pdf [Accessed 29 May 2026].
- Flin R, Patey R. Non-technical skills for anaesthetists: developing and applying ANTS. Best Practice & Research Clinical Anaesthesiology. 2011;25(2):215–227.
- González-Romá V, Hernández A. Conducting and evaluating multilevel studies: recommendations, resources, and a checklist. Organizational Research Methods. 2022;25(5):795–825.
- Tschan F, Zimmermann J, Semmer NK. Rules for coding scheme development. In: Brauner E, Boos M, Kolbe M, editors. The Cambridge handbook of group interaction analysis. Cambridge: Cambridge University Press. 2018. pp. 191–207.

Acknowledgements

The Association for Simulated Practice in Healthcare (ASPiH) has supported the publication of this work through their fee waiver member benefit.

Declarations

Authors' contributions

JBS conceived the manuscript, developed the initial ideas, and wrote the first draft. Mirko Antino contributed to the conceptual development of the manuscript and critically revised the manuscript for important intellectual content. Both authors reviewed and approved the final version of the manuscript.

Funding

This work has been funded by the 'Swiss National Science Foundation' [grant number PCEFP1_203374].

Availability of data and materials

Not applicable.

Ethics approval

Not applicable.

Competing interests

The authors have no conflict of interest to disclose.

References

1. Schmutz JB, Meier LL, Manser T. How effective is teamwork really? The relationship between teamwork and performance in healthcare teams: a systematic review and meta-analysis. BMJ Open. 2019;9:e028280. doi: 10.1136/bmjopen-2018-028280
2. Hughes AM, Gregory ME, Joseph DL, Sonesh SC, Marlow SL, Lacerenza CN, et al. Saving lives: a meta-analysis of team training in healthcare. The Journal of Applied Psychology. 2016;101:1266–1304. doi: 10.1037/apl0000120
3. Wittig J, Krogh K, Blanchard EE, Xing K, Kushner J, Bichmann A, et al. A systematic review on conditions before and after training of teamwork competencies and the effect on transfer of skills to the clinical workplace. Simulation in Healthcare: The Journal of the Society for Simulation in Healthcare. 2025;20:111–117. doi: 10.1097/SIH.0000000000000809
4. Gaba DM, Howard SK, Fish KJ, Smith BE, Sowb YA. Simulation-based training in anesthesia crisis resource management (ACRM): a decade of experience. Simulation & Gaming. 2001;32:175–193. doi: 10.1177/104687810103200206
5. Bichmann A, Blanchard EE, Wittig J, McEwan D, Cooper D, Tannenbaum S, et al. Impact of team composition on learning outcomes following simulation-based training of teamwork competencies: a systematic review. Simulation in Healthcare: The Journal of the Society for Simulation in Healthcare. 2025;20:259–266. doi: 10.1097/SIH.0000000000000865
6. Marks MA, Mathieu JE, Zaccaro SJ. A temporally based framework and taxonomy of team processes. The Academy of Management Review. 2001;26:356–376. doi: 10.5465/amr.2001.4845785
7. Edmondson AC, Lei Z. Psychological safety: the history, renaissance, and future of an interpersonal construct. Annual Review of Organizational Psychology. 2014;1:23–43. doi: 10.1146/annurev-orgpsych-031413-091305

8. Streiner DL, Norman GR, Cairney J. Health measurement scales: a practical guide to their development and use. Oxford: Oxford University Press. 2024.
9. Hambleton RK. Guidelines for adapting educational and psychological tests. Washington, DC: U.S. Department of Education. 1996.
10. Mathieu JE, Luciano MM, D'Innocenzo L, Klock EA, LePine JA. The development and construct validity of a team processes survey measure. Organizational Research Methods. 2020;23:399–431. doi: 10.1177/1094428119840801
11. Roper L, Shulruf B, Jorm C, Currie J, Gordon CJ. Validation of the self-assessment teamwork tool (SATT) in a cohort of nursing and medical students. Medical Teacher. 2018;40:1072–1075. doi: 10.1080/0142159X.2017.1418849
12. Chen AS, Yau B, Revere L, Swails J. Implementation, evaluation, and outcome of TeamSTEPPS in interprofessional education: a scoping review. Journal of Interprofessional Care. 2019;33:795–804. doi: 10.1080/13561820.2019.1594729
13. Chan D. So why ask me? Are self-report data really that bad? In: Lance CE, Vandenberg RJ, editors. Statistical and methodological myths and urban legends. 1st edition. New York: Routledge. 2010. pp. 329–356.
14. Edmondson AC. A safe harbor: social psychological conditions enabling boundary spanning in work teams. In: Mannix B, Neale M, Wageman R, editors. Research on Groups and Teams. Greenwich, CT: JAI Press. 1999. pp. 179–200.
15. Kim ES, Dedrick RF, Cao C, Ferron JM. Multilevel factor analysis: reporting guidelines and a review of reporting practices. Multivariate Behavioral Research. 2016;51:881–898. doi: 10.1080/00273171.2016.1228042
16. Bliese PD. Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In: Klein KJ, Kozlowski SWJ, editors. Multilevel theory, research, and methods in organizations: foundations, extensions, and new directions. Jossey-Bass, San Francisco. 1st edition. 2000. pp. 349–381.
17. Tschan F, Zimmermann J, Semmer NK. Rules for coding scheme development. In: Brauner E, Boos M, Kolbe M, editors. The Cambridge handbook of group interaction analysis. Cambridge: Cambridge University Press. 2018. 1st edition. pp. 191–207.
18. Waller MJ, Kaplan SA. Systematic behavioral observation for emergent team phenomena key considerations for quantitative video-based approaches. Organizational Research Methods. 2016;21:1094428116647785. doi: 10.1177/1094428116647785
19. Kolbe M, Burtscher MJ, Manser T. Co-ACT—a framework for observing coordination behaviour in acute care teams. BMJ Quality & Safety. 2013;22:596–605. doi: 10.1136/bmjqs-2012-001319
20. Schmutz JB, Lei Z, Eppich WJ. Reflection on the fly: development of the Team Reflection Behavioral Observation System (TuRBO) for Acute Care Teams. Academic Medicine. 2021;9:1337–1345. doi: 10.1097/ACM.0000000000004105
21. McHugh ML. Interrater reliability: the kappa statistic. Biochemia Medica. 2012;22:276–282. doi: 10.11613/BM.2012.031
22. Heyman RE, Lorber MF, Eddy JM, West TV. Behavioral observation and coding. New York: Cambridge University Press. 2000.
23. Lei Z, Waller MJ, Hagen J, Kaplan S. Team adaptiveness in dynamic contexts: Contextualizing the roles of interaction patterns and in-process planning. Group and Organizational Management. 2016;41:491–525. doi: 10.1177/1059601115615246
24. Sharpe T, Koperwas J. Behavior and sequential analyses: principles and practice. Thousand Oaks, CA: SAGE Publications; 2003. doi: 10.4135/9781412983518
25. Flin R, Patey R, Glavin R, Maran N. Anaesthetists' non-technical skills. British Journal of Anaesthesia. 2010;105:38–44. doi: 10.1093/bja/aeq134
26. Flin R, Patey R. Non-technical skills for anaesthetists: developing and applying ANTS. Best Practice & Research: Clinical Anaesthesiology. 2011;25:215–227. doi: 10.1016/j.bpa.2011.02.005
27. Mishra A, Catchpole K, McCulloch P. The Oxford NOTECHS system: Reliability and validity of a tool for measuring teamwork behaviour in the operating theatre. Quality and Safety in Health Care. 2009;18:104–108. doi: 10.1136/qshc.2007.024760
28. Sevdalis N, Davis R, Koutantji M, Undre S, Darzi A, Vincent CA. Reliability of a revised NOTECHS scale for use in surgical teams. The American Journal of Surgery. 2008;196:184–190. doi: 10.1016/j.amjsurg.2007.08.070
29. Georganta E, Brodbeck FC. Capturing the four-phase team adaptation process with Behaviorally Anchored Rating Scales (BARS). European Journal of Psychological Assessment. 2018;36:1–12. doi: 10.1027/1015-5759/a000503
30. Bernardin HJ, Smith PC. A clarification of some issues regarding the development and use of behaviorally anchored ratings scales (BARS). Journal of Applied Psychology. 1981;66:458–463. doi: 10.1037/0021-9010.66.4.458
31. LeBreton JM, Senter JL. Answers to 20 questions about interrater reliability and interrater agreement. Organizational Research Methods. 2008;11:815–852. doi: 10.1177/1094428106296642
32. González-Romá V, Hernández A. Conducting and evaluating multilevel studies: recommendations, resources, and a checklist. Organizational Research Methods. 2022;26:109442812110607. doi: 10.1177/10944281211060712
33. Roussin CJ, Larraz E, Jamieson K, Maestre JM. Psychological safety, self-efficacy, and speaking up in interprofessional health care simulation. Clinical Simulation in Nursing. 2018;17:38–46. doi: 10.1016/j.ecns.2017.12.002
34. Weiss M, Kolbe M, Grote G, Spahn DR, Grande B. We can do it! Inclusive leader language promotes voice behavior in multi-professional teams. The Leadership Quarterly. 2018;3:389–402. doi: 10.1016/j.leaqua.2017.09.002