JOURNAL OF Surgical Simulation

Journal of Surgical Simulation (2025) **12**, 10–22
DOI: https://doi.org/10.1102/2051-7726.2025.0002

REVIEW ARTICLE

Comparing the validity and reliability of different teamwork assessment tools within a surgical context

Wenjing He,[a,*] Princess Maglanque,[a] Krista Hardy,[a] Eliz Malaso,[a] Bin Zheng[b] and Ashley Vergis[a]

[a]*Department of Surgery, Max Rady College of Medicine, University of Manitoba, Winnipeg, MB, Canada;* [b]*Department of Surgery, Faculty of Medicine and Dentistry, University of Alberta, Edmonton, AB, Canada*

**Corresponding author at: St Boniface General Hospital, Z3008-409 Taché Avenue, Winnipeg, Manitoba, Canada R2H 2A6. Email: wenjing.he112@gmail.com*

Date accepted for publication: 17 March 2025

Abstract

Background: Effective teamwork is essential for preventing adverse events during surgery. Over the past two decades, various teamwork assessment tools have been developed to evaluate the non-technical performance of surgical teams. The purpose of this literature review is to present a summary of tools including the validity and reliability of each tool. **Methods:** To identify teamwork assessment tools, a literature search was performed through the Medline and Embase databases, along with forward citation tracking across Scopus and Web of Science. **Results:** Forty-nine articles were selected for review. Sixteen original team assessment tools identified have been employed in various surgical procedures. Fourteen out of sixteen tools were designed by point scale and two were event coding of surgical videos. The Non-Technical Skills for Surgeons (NOTSS) assessment was found to possess the highest level of validity. The Oxford Non-technical Skills System (NOTECHS) was the most validated tool on content, concurrent, predictive, and convergent validation evidence. **Conclusions:** Several teamwork assessment tools such as ANTS, NOTECHS, NOTSS, OTAS and SPLINTS demonstrated good validity and reliability. However, tools developed more recently tend to show less validity and reliability evidence, possibly due to a lack of research to date. Deficits in the current integration of teamwork training and assessment into surgical training are observed. The incorporation of non-technical skill training and formal training of assessors are recommended in surgical programs.

Keywords: *teamwork skills; surgery; non-technical performance; validity; reliability*

Introduction

Effective team performance is a cornerstone for surgical success,[1] wherein a surgical team typically comprises a surgeon, a surgeon's assistant, an anesthetist, and a circulating nurse. A robust understanding of team dynamics necessitates the utilization of valid and reliable tools for surgical team assessment.[1]

Over the past two decades, a variety of teamwork assessment tools have been developed to evaluate non-technical skills in surgery. These tools commonly measure communication, coordination, leadership, and situational awareness, all of which collectively affect team effectiveness. Tools such as the Observational Teamwork Assessment for Surgery (OTAS) and Non-Technical Skills for Surgeons (NOTSS) have been widely studied and used for their focus on evaluating teamwork.[1–11] By capturing the behavioral metrics of teamwork, these tools facilitate the identification of areas of improvement and support interventions aimed at enhancing team performance.

The validity of an assessment is the extent to which it measures what it was designed to measure.[12] There are three main types of internal test validity: construct, content, and criterion, each of which can be subdivided. *Construct validity* concerns the extent to which a measure accurately assesses what it is supposed to.[12] It is demonstrated by linkages between the expected and acquired measurements; for example, more experienced participants achieve higher test scores. The measures are correlation-based such as Pearson's *r* or experimentally based hypothesis testing specifying between group differences (analyses of variance, etc). Construct validity subtypes include convergent validity and discriminant validity. *Content validity*, subjectively determined by experts, concerns the degree to which a test evaluates all aspects of the construct that it is designed to measure.[12] *Face validity,* a subtype of content validity, concerns the test content being suitable for its aims. *Criterion validity* concerns the results accurately measuring the

concrete outcome that the tool is designed to measure. *Concurrent validity,* a subtype of criterion validity, is illustrated by the correlation of test scores with simultaneous results from a previously validated tool measuring the same construct. *Predictive validity*, a subtype of criterion validity, is the ability of a test to predict future outcomes.[12] The outcome in this review could be a behavior/performance, which is accurately predicted following an initial assessment with a time period between tests.

Reliability is the extent to which a measurement gives consistent results.[12] Reliability is measured by test–retest (measures the consistency of the same test over time), inter-rater (measures the consistency of the same test conducted by different people), intra-rater (measures how consistent an individual is at measuring), parallel forms (measures different versions of an equivalently designed test) and internal consistency (measures the consistency of the individual items of a test).

Given the intricate interplay of teamwork dynamics, the validation and reliability of the assessment tools are of paramount importance. Notably, the most recent comprehensive literature review on this subject was conducted in 2015.[13] Considering the dynamic evolution of surgical practices and the evolving understanding of team dynamics, it is imperative to revisit and update the available knowledge. Multiple systematic reviews[13–15] have been published to summarize these tools, but none have provided a review for both the validity and reliability of each tool. This literature review aims to bridge the temporal gap since the 2015 review and present a contemporary summary of the current teamwork assessment tools within surgical contexts. This literature review will serve to inform readers of the design of the assessment tools, which surgical environment the tools were tested in, and the available validity and reliability evidence for each tool.

Methods

To begin the search for relevant literature, a list of keywords was generated, followed by the implementation of a search strategy. Forward citation tracking was employed to broaden the search. Databases such as Medline, Embase, Scopus and Web of Science were explored for this literature review. The results generated were screened and retrieved based on the article title and abstract using the inclusion criteria specified below.

Search strategy

A literature search was conducted through the Medline (1946 to December 2024) and Embase (1974 to December 2024) databases using a combination of the following keywords: "team OR teams OR teamwork," "surgery* OR surgical* OR operat*," "surgical procedures/operative," "data collection," "focus groups," "interviews," "narration," "surveys and questionnaires," "assess* OR evaluat*," "non?technical skill," "co-operation OR cooperation," "team collaboration," "team coordination," "team communication," "operating room OR operating theatre OR operating theater OR surg*," "rating," "scale," "measure," "communication," "observ*." These keywords were formed through a collection of keywords obtained from the author's (W.H.) thesis chapter reference list,[16] and with the guidance of a Health Sciences Librarian from the University of Manitoba. Initial screening for eligible articles was performed by reviewing each article title generated from the search results and obtaining the abstract and full article to provide further clarification when needed.

Citation tracking

A reference list of articles regarding teamwork in surgery was obtained from the first author's (W.H) thesis chapter[16] to initiate forward citation tracking in the Scopus and Web of Science databases. In addition, PubMed was used when no results were generated, or if articles were inaccessible in previously mentioned databases. In Scopus, search categories were limited to 'Medicine,' 'Multidisciplinary,' 'Health Professions,' and applied "surg*," "team*," and "assess*" to filter results. Web of Science results were limited to the category "surgery" when available and keywords "team?work*," "non?technical skill*," "surg*," "tool*," and "assess*" were used to further refine the results. Articles were screened by reviewing the title and abstract, retrieving only those that meet the criteria.

Articles obtained from the search strategy and citation tracking were combined and deduplicated in EndNote X9. The following inclusion criteria were developed to identify relevant papers: 1) evaluation of non-technical skill teamwork or cooperation, 2) teamwork assessment in a surgical setting such as operating room, robotic surgery, and surgical or video simulations, 3) original research analyzing the reliability and validity of the assessment tool, and 4) translation and cultural validation of modified or revised version of the teamwork assessment tool. Review articles, book chapters, articles in languages other than English, and team assessment conducted outside of surgical setting were excluded from the review. Measures of teamwork performance based on patients' perspectives towards the surgical team, and study participants' attitudes and team perceptions were also omitted.

The identified tools were observational assessment of teamwork and other non-technical skills in which trained expert

raters assessed the team dynamic in operation and simulation settings. No self-report measures were found that assess teamwork performance.

Validity and reliability data presentation

The validity and reliability evidence are summarized in the results, including the types of validity (construct, content, face, concurrent, and predictive) and reliability evidence (inter/intra-rater, internal consistency, and test–retest) with the number of tools demonstrating such evidence. Descriptive statistics with reported values such as intraclass correlation coefficient (ICC) and Cronbach's alpha (intra-rater reliability and internal consistency) were used in the literature. To better represent the validity and reliability data, cutoff values for the various validity and reliability measures were applied in three categories: low, moderate, and high levels. For validity, the Spearman's rank correlation coefficient (ρ) values are categorized as follows: low validity is defined by $\rho < 0.25$, moderate validity by ρ between 0.25 and 0.50, and high validity by $\rho > 0.75$.[17]

For reliability or agreement, the cutoff values are as follows: low reliability is reflected by ICC < 0.50, kappa < 0.41, Cronbach's alpha < 0.5, Krippendorff's alpha < 0.67, and Spearman's $\rho < 0.25$ (small correlation).[18,19] Moderate reliability is represented by ICC between 0.25 and 0.50, kappa values between 0.41 and 0.60, Cronbach's alpha between 0.5 and 0.8, Krippendorff's alpha between 0.67 and 0.79, and Spearman's ρ between 0.25 and 0.50.[18,19] High reliability or agreement is represented by ICC and within-group inter-rater coefficient (R_{wg}) values of 0.70 or higher, kappa $>$ 0.61, Cronbach's alpha > 0.8, Krippendorff's alpha ≥ 0.80, and Spearman's $\rho > 0.75$.[18,19]

Results

Of 349 articles retrieved from citation tracking, 183 were duplicates and 166 articles were selected. The literature search generated 1007 articles; 939 articles were deemed irrelevant and 68 articles were retrieved. Articles retrieved from citation tracking and electronic search were combined and deduplicated in EndNote X9, removing 51 articles from the list. With the addition of the inclusion criteria, 117 articles failed to meet the criteria, and 66 articles were deemed to be relevant. Following a full text assessment, a final total of 49 articles were selected for review (Fig. 1).

Sixteen original team assessment tools identified have been employed in various surgical procedures such as in general surgery, vascular, orthopedic, maxillofacial, pediatric neurosurgery, ophthalmic ambulatory, urological, cardiac, neurosurgery, gynecology, venous, and robotic surgery. Ten team assessment tools were modified, revised, translated, or culturally adapted versions of the NOTSS, NOTECHS, ANTS, and OTAS tools. It is important to note that the Human Factors Rating Scales—Modified (HFRS-M) tool was included as an original article in this review as it was modified from aviation training and assessment to application in a surgical setting.[20] These identified instruments use a three- to seven-point numerical or Likert scale design to measure team interactions and performance. However, the assessment tools Simultaneous Observation of Distractors and Communication in the Operating Room (SO-DIC-OR) and Behavioral Marker System for Assessing Neurosurgical Non-Technical Skills (BMS-NNTS) were the only instruments to utilize event coding in team assessment (Table 1).

Participants included in the studies were surgical teams consisting of surgeons, anesthetists, and nurses. In some cases, medical residents/students/trainees, surgical assistants/technicians, neurophysiologists, physiotherapists, and operating department practitioners were involved in the study assessment (Table 1). The validity and reliability of each tool are shown in Table 2.

Non-Technical Skills for Surgeons (NOTSS)

NOTSS serves as an indispensable rating instrument employed for team assessments occurring in both live operations and in simulations such as video scenarios. The NOTSS system provides a framework and common terminology for rating and giving feedback on non-technical skills. Distinguished by its multi-dimensional structure, the tool comprises four domains: situational awareness, communication and teamwork, decision making, and leadership. The evaluation transpires on a comprehensive four-point rating scale that ranges from 1 (poor) to 4 (good), with the inclusion of an 'N/A' option to accommodate scenarios where assessment may not be applicable.[22–24]

Originating within the dynamic framework of the UK surgical environment, NOTSS has transcended geographical boundaries and cultural contexts, leading to adaptations that tailor its applicability to diverse surgical landscapes. Notably, the tool has been adapted to the US surgical context, emerging as NOTSS-US.[44,45] In addition, the Danish setting has witnessed the development of (NOTSSdk), further attesting to the tool's adaptive versatility.[46] Moreover, NOTSS has found its place in a Japanese cancer center, encapsulating its global influence and utility.[10]

The robustness of NOTSS is fortified by a spectrum of validation evidence. Its construct validity is fortified by a spectrum of validation evidence. Its construct validity is underscored by impressive comparative fit indices of 0.94 and 0.92,[23] affirming the tool's capacity to accurately represent the intended dimensions. Furthermore, content validity, a cornerstone of

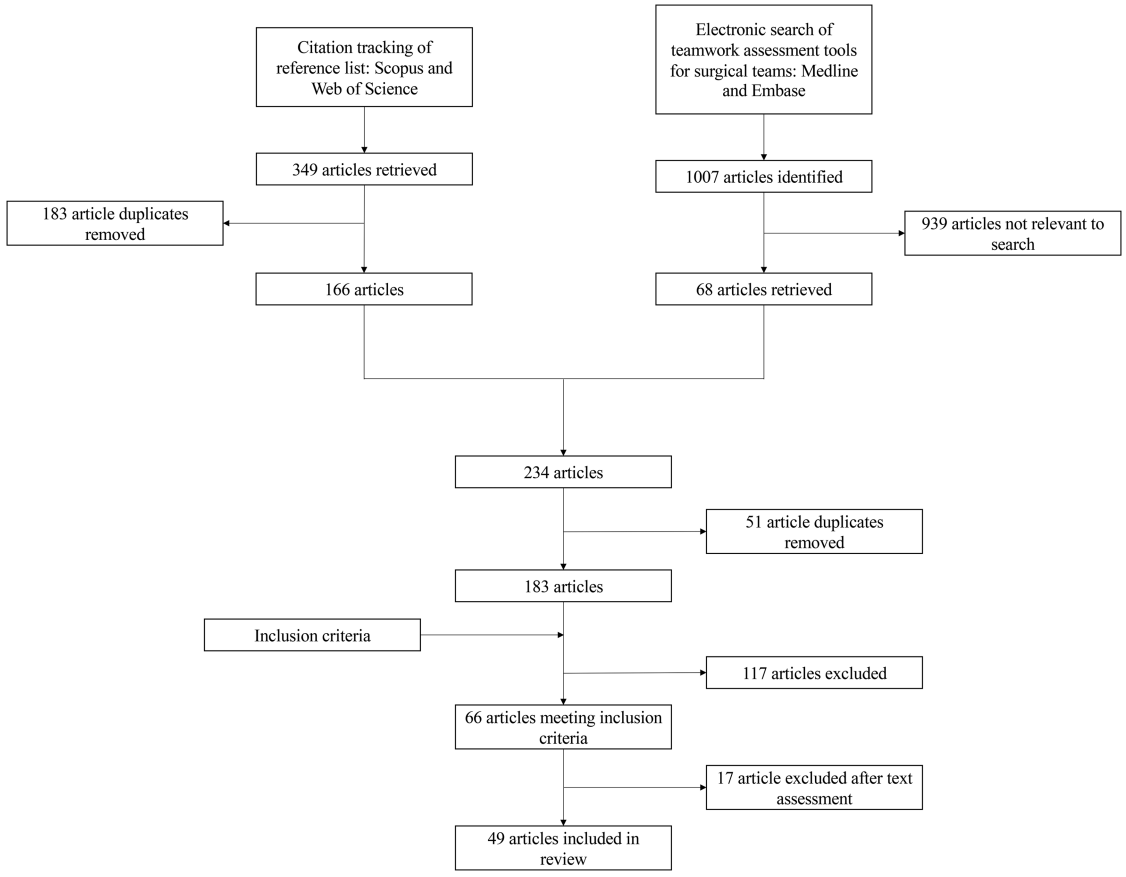

Figure 1. Flow chart of literature review.

assessment precision, is derived through expert consensus, bolstering the tool's authenticity and relevance.[24] In the realm of concurrent validity, NOTSS exhibits a commendable correlation coefficient of 0.86 ($r = 0.86$),[3,23] affirming its alignment with other established metrics. Face validity, an essential aspect of initial perception, also underscores NOTSS, consolidating its intuitive and representative nature.[47]

The progression and diffusion of NOTSS across varied healthcare contexts attest to its adaptability, relevance, and robustness. As a linchpin in assessing non-technical skills within surgical teams, NOTSS stands as a beacon of comprehensive evaluation and continual improvement, transcending boundaries to enhance patient safety and surgical excellence.

Oxford Non-Technical Skills (NOTECHS)

NOTECHS stands as a valuable tool for the assessment of the non-technical skills of a surgical team, operating in both the dynamic environment of the OR and simulated scenarios. The assessment framework employs a four-point rating system (1 = poor, 2 = marginal, 3 = acceptable, 4 = good) to evaluate a range of critical domains that shape effective team functioning. In its evaluation, NOTECHS comprehensively encompasses key dimensions such as leadership and management, which gauge the team's ability to lead and coordinate tasks efficiently. The skillful application of problem solving and decision making is another crucial facet evaluated by NOTECHS, shedding light on the team's aptitude for making sound judgments under pressure. Moreover, NOTECHS delves into teamwork and cooperation, discerning the team's capability to collaborate harmoniously, communicate effectively, and share responsibilities seamlessly. Situation awareness, a pivotal aspect of team performance, is also examined by NOTECHS to ascertain the team's perceptiveness, attentiveness to evolving circumstances, and overall cognitive understanding of the surgical context.[48–52]

Drawing a parallel with other established assessment tools such as NOTSS, OTAS, and ANTS, NOTECHS is presumed to possess a content validity akin to these counterparts.[28]

This underlying assumption underscores the expectation that NOTECHS, like its contemporaries, effectively measures

Table 1. Teamwork assessment tools that have been used in surgical environment.

Tool Name	Domain	Design	Participants	Procedure	Setting
NOTSS	Situational awareness, communication and teamwork, decision making, leadership	4-point scale	Surgical team*, surgical assistants and trainees,[2–4,21], surgical residents,[22] and surgeons[3,23,24]	General surgery,[3,4,21–24] vascular surgery,[4] hemorrhage and airway emergency,[2] orthopedic surgery[24]	OR,[2–4] simulation[21–24]
NOTECHS	Leadership and management, teamwork and cooperation, problem solving and decision making, situation awareness	4-point scale	Surgical team[25–27]	General surgery,[25,27] maxillofacial surgery, pediatric neurosurgery, vascular surgery[26]	OR,[25,26,28] simulation[27]
ANTS	Task management, team working, situation awareness	4-point scale	Surgical team,[2,4,21] anesthesiologists[29]	General surgery,[4,21,29] vascular surgery,[4] hemorrhage and airway emergency[2]	OR,[4] simulation[2,21,29]
SPLINTS	Situation awareness, communication and teamwork, task management	4-point scale	Surgical team and surgical technicians,[4,21] scrub practitioners[30–32]	General surgery,[4,21,31] ophthalmic ambulatory surgery[30]	OR,[4,30] simulation[21,31]
OTAS	Communication, coordination, cooperation and backup behavior, leadership, team monitoring and situation awareness	7-point scale	Surgeons,[10] surgical team[5–9,11,21,33]	General surgery,[5,7,8,10,21] urological surgery,[5,7,9,11] cardiac surgery[6]	OR,[1,5–11] simulation[21]
BMS-NNTS[34]	Cooperation and teamwork, situational awareness, explicit coordination, decision making, leadership, other	Coding and evaluation of explicit oral communication, including silences	Surgical team, neurophysiologist, nurse anesthetist and physiotherapist	Neurosurgery	OR
Cannon-Bowers Scale[22]	Overall planning and strategy, monitoring, affect and attitude management, motivation building, adaptability, shared mental models	5-point scale	Surgical residents	General surgery	Simulation
CATME[35]	Work quality, communication, team effectiveness	5-point scale	Surgical residents	General surgery	Simulation
CATS[36]	Coordination, cooperation, situational awareness, communication	3-point scale	Surgical team	General surgery, cesarean section	OR
HUFOES[14]	Teamwork and communication, leadership, decision making, situational awareness, professionalism	5-point scale	N/R	N/R	N/R
HFRS-M[20]	Communication and interaction, vigilance/situational awareness, team skills, leadership and management skills, decision making—crisis	6-point scale	Surgical team, operating departmental practitioner	Venous surgery, general surgery	Simulation
ICARS[37]	Checklist and equipment, interpersonal skills, cognitive skills, resource skills	5-point scale	Surgeons	Robotic surgery, urology	Simulation
ORTAS[38]	Two scales: individual performance and overall teamwork Overall teamwork domains: shared mental mode, adaptive communication and response	6-point scale	Medical students, nursing students, nurse anesthesia students	General surgery	Simulation
OSANTS[39]	Situation awareness, decision making, teamwork, communication, leading and directing, professionalism, managing and coordinating	5-point scale	Surgical residents	General surgery and crisis scenarios	OR, simulation
SO-DIC-OR[40]	Distractors, communication/teamwork, contextual codes	Event coding	Surgical team	General surgery	OR
The Surgical Teamwork Tool[41]	Clinical leadership, communication, coordination, respect, assertiveness	5-point scale	Surgical team	N/R	OR

NOTSS: Non-Technical Skills for Surgeons, NOTECHS: Oxford Non-Technical Skills, ANTS: Anesthetists' Non-Technical Skills, SPLINTS: Scrub Practitioners' List of Intraoperative Non-Technical Skills, OTAS: Observational Teamwork Assessment for Surgery, BMS-NNTS: Behavioral Marker System for Assessing Neurosurgical Non-Technical Skills, CATME: Comprehensive Assessment of Team Member Effectiveness, CATS: Communication and Teamwork Skills, HUFOES: Human Factors in Intraoperative Ophthalmic Emergencies Scoring System, HFRS-M: Human Factors Rating Scales—Modified, ICARS: Interpersonal and Cognitive Assessment for Robotic Surgery, ORTAS: Operating Room Teamwork Assessment Scales, OSANTS: Objective Structured Assessment of Non-technical Skills, SO-DIC-OR: Simultaneous Observation of Distractions and Communications in the Operating Room, OR: operating room, N/R: not reported, Surgical team*: includes surgeon, anesthetist, nurses.

Adapted and expanded from Whittaker et al.[13]

Table 2. Comparison of the validity evidence for the surgical teamwork assessment tools

Tool Name	Validity
NOTSS	**Construct**: High[23] **Content**: Derived from its systematic development process with subject matter experts[24] **Concurrent**: High[3,23]
NOTECHS	**Content**: Assumed content validity as NOTECHS is similar in content to OTAS, NOTSS, and ANTS[28] **Concurrent**: Moderate[28] **Predictive**: High[28] **Convergent**: High[28]
ANTS	**Content**: Developed from attitude survey, incident analysis, observations in theatre, critical incident interviews with consultants, and a prototype system[42]
SPLINTS	**Content**: Derived from systematic development by SMEs[31] (CVI = 0.93; CEI = 0.91)[30] **Concurrent**: High[30]
OTAS	**Construct**: High[9] **Content**: Exemplar behaviors assessed by expert OR personnel were judged as relevant to the tool (CVM = 8.31)[7]
BMS-NNTS	**Content**: Prototype based on literature and observation tools developed in other specialties, observations during neurosurgical operations, and preliminary evaluation of video recordings of actual operations[34]
Cannon-Bowers Scale	**Concurrent**: High[22]
CATME	**Content**: Created based on the previously validated CATME, which used teamwork literature to create potential items and was tested using two surveys of college students[35,43]
CATS	**Content**: Behavior markers were selected from CRM behavior-based markers adapted to healthcare; items common to ANTS and OTAS deemed applicable to all health professions were also incorporated[36]
HUFOES	**Content**: Focus group, literature review, and questionnaires distributed to ophthalmologists were used to gain feasibility and validity[14]
HFRS-M	**Content**: Developed using briefings in simulated crisis scenarios[20] **Predictive**: Not observed[20]
ICARS	**Content**: 86% expert panel agreement for application in robotic surgical environment[37] **Concurrent**: Low[37]
ORTAS	**Predictive**: Observed[38]
OSANTS	**Content**: Existing evidence-based rating systems of NTS in the OR, training requirements of several programs and colleges, and pilot testing were implemented to develop the tool[39] **Concurrent**: High[39]
SO-DIC-OR	**Content**: Developed based on expert interviews, observations of surgical procedures, and literature review[40]
The Surgical Teamwork Tool	**Content**: Developed through literature review, expert consultation, and end-user testing[41]

NOTSS: Non-Technical Skills for Surgeons, NOTECHS: Oxford Non-Technical Skills, ANTS: Anesthetists' Non-Technical Skills, SPLINTS: Scrub Practitioners' List of Intraoperative Non-Technical Skills, OTAS: Observational Teamwork Assessment for Surgery, BMS-NNTS: Behavioral Marker System for Assessing Neurosurgical Non-Technical Skills, CATME: Comprehensive Assessment of Team Member Effectiveness, CATS: Communication and Teamwork Skills, HUFOES: Human Factors in Intraoperative Ophthalmic Emergencies Scoring System, HFRS-M: Human Factors Rating Scales—Modified, ICARS: Interpersonal and Cognitive Assessment for Robotic Surgery, ORTAS: Operating Room Teamwork Assessment Scales, OSANTS: Objective Structured Assessment of Non-technical Skills, SO-DIC-OR: Simultaneous Observation of Distractions and Communications in the Operating Room, NTS: non-technical skills.

the intended constructs and skills, ensuring that its results are indicative of the non-technical skills under evaluation.

The utilization of NOTECHS as part of the comprehensive repertoire of assessment tools enriches our understanding of surgical team dynamics and non-technical proficiencies. Through its incorporation of critical domains within a four-point rating system, NOTECHS serves as a robust instrument for assessing and advancing the non-technical expertise that underpins efficient and cohesive surgical team performance.

Anesthetists' Non-Technical Skills (ANTS)

ANTS is a teamwork assessment tool designed specifically for anesthetists.[53] The authors kept the ANTS in the review as anesthetists are an indispensable part of a surgical team. ANTS systematically categorizes observed behaviors into key domains encompassing task management, team working, situation awareness, and decision making. It employs a well-structured four-point rating system (1 = poor, 2 = marginal, 3 = acceptable, 4 = good), offering a granular assessment of anesthetists' contributions and interactions.[21,29]

Scrub Practitioners' List of Intraoperative Non-Technical Skills (SPLINTS)

The SPLINTS framework represents a behavioral assessment system conceived through collaborative efforts involving a diverse assembly of professionals, including OR nurses, surgeons, anesthetists, and psychologists in Scotland. SPLINTS

is designed to assess the non-technical skills of scrub practitioners using a four-point scale (1 = poor, 2 = marginal, 3 = acceptable, 4 = good). The ICC value of 0.63 across raters suggests good reliability.[21]

Observational Teamwork Assessment for Surgery (OTAS)

OTAS assesses domains on communication, coordination and backup behavior, leadership, team monitoring and situation awareness on a seven-point scale.[21,54–56] It is a popular tool that has been used in general surgery, urological surgery, and cardiac surgery. Construct and content validity have been shown with r = 0.74 (0.72–0.76) between expert raters and r = 0.26 (0.08–0.60) between expert and novice raters.[9] Exemplar behaviors assessed by expert OR personnel were judged as relevant to the tool (CVM = 8.31), ICC = 0.70 (0.64–0.77);[7] Cohen's κ = 0.46 (0.38–0.60) across domains.[7]

Behavioral Marker System for Assessing Neurological Non-Technical Skills (BMS-NNTS)

BMS-NNTS quantifies the non-technical skills (NTS) categories of cooperation and teamwork, situation awareness, explicit coordination, decision making, and leadership from verbal communications that occur during surgical procedures in the OR. Evidence of content validity stems from its development through literature reviews, observation tools in other specialties, observations during neurosurgical operations, and video recordings during the operations. The average ICC value of 0.65 across domains indicates good to moderate inter-rater reliability.[34]

Cannon-Bowers Scale

The Cannon-Bowers scale is designed to assess surgical residents during a stimulation using a five-point rating system (1 = not performed, 2 = poor performance, 3 = average performance, 4 = good performance, and 5 = excellent performance). These are the following categories that are assessed: motivation building, overall planning and strategy, monitoring, adaptability, affect and attitude management, and shared mental models. When looking at the internal consistency, this tool has a Cronbach's α value of 0.8, which indicates a good level of reliability.[22]

Comprehensive Assessment of Team Member Effectiveness (CATME)

CATME was developed to collects data on team-member effectiveness in five areas that research has shown to be important: contributing to team's work; having relevant knowledge, skills, and abilities (KSAs); expecting quality; keeping the team on track; and interacting with teammates. CATME evaluates the work quality, communication and team effectiveness of surgical residents under stimulation.[9,35,43] This tool is rated on a five-point scale: 1 = not performed, 2 = poor performance, 3 = average performance, 4 = good performance, and 5 = excellent performance. The internal consistency had a Cronbach's α value of 0.81, suggesting good reliability.[35]

Communication and Teamwork Skills (CATS)

CATS is a behavior-based tool that assesses a surgical team in an OR.[36] Behavior markers are grouped into the following categories: coordination, cooperation, situational awareness, and communication. Scores of a team are based on the quality and occurrence of the behaviors and are weighted as follows: 1 = observed and good, 0.5 = variation in quality and 0 = expected but not observed.

Human Factors in Intraoperative Ophthalmic Emergencies Scoring System (HUFOES)

The HUFOES is a NTS assessment system that uses a five-point scale (1 = strongly disagree, 2 = somewhat disagree, 3 = neutral, 4 = somewhat agree and 5 = strongly agree). The application of the following NTS are assessed: teamwork and communication, leadership, professionalism, decision making, situation awareness.[14] The inter-rater agreement of this tool was overall excellent, with an average Cronbach's α value of 0.83 found across the domains.[14]

Interpersonal and Cognitive Assessment for Robotic Surgery (ICARS)

ICARS is the first NTS rating system developed for robotic surgery. The assessment includes checklist and equipment, interpersonal skills, cognitive skills and resource skills on a five-point scale. Content and concurrent validities were confirmed: 86% of experts agreed for the application in robotic surgical environments.[37] The Bland–Altman analysis 95% CI for the correlation of ICARS to NOTSS had a Z score of −0.66 to 0.65.[37]

Human Factors Rating Scales—Modified (HFRS-M)

HFRS-M assesses communication and interaction, vigilance/situational awareness, team skills, leadership and management skills and decision making—crisis on a six-point scale.[20] It was developed using briefings in simulated crisis scenarios. No overall effect of training on NTS was observed. There was also no reliability evidence reported.

Operating Room Teamwork Assessment Scales (ORTAS)

ORTAS uses two scales: individual performance and overall teamwork; the overall teamwork domains are shared mental

model and adaptive communication and response, based on a six-point scale under simulation. Predictive validity is demonstrated but no other validity was reported.[38] No reliability evidence was shown.

Objective Structured Assessment of Non-technical Skills (OSANTS)

OSANTS assesses situational awareness, decision making, teamwork, communication, leading and directing professionalism, managing and coordinating based on a five-point scale in the general surgery team. Content and concurrent validities are demonstrated for simulation videos and live observations in the OR.[39] The inter-rater reliability was excellent with ICC = 0.95 both in the OR and simulation setting. The inter-rater consistency had a Cronbach's α value of 0.80, suggesting good reliability.

Simultaneous Observation of Distractions and Communication in the Operating Room (SO-DIC-OR)

SO-DIC-OR uses event-coding to observe distractors, communication/teamwork, and contextual codes in the OR.[40] Content validity of SO-DIC-OR relies on its development based on expert interviews, observations of surgical procedures, and literature reviews. An average Cohen's κ value of 0.85 indicates very good interobserver agreement.[40] No other validity or reliability measures have been demonstrated from other studies.

The Surgical Teamwork Tool

The Surgical Teamwork Tool involves clinical leadership, communication, coordination, respect and assertiveness on a five-point scale. It was developed through literature review, expert consultation and end-user testing.[41] No other validity was observed. Inter-rater reliability had an average of 0.73 (0.63–0.92) across domains.[41]

Discussion

Both technical and non-technical skills are requisites to proficiently performing both novel and familiar surgical procedures.[15,57–59] Effective surgical training and credentialing are critical to ensure high standards of surgical care. Pertinently, the most recent review of teamwork assessment tools was conducted by George Whittaker et al. in 2015.[13] During their investigation, the authors identified a total of eight tools for assessing surgical teamwork. Notably, the Non-Technical Skills for Surgeons (NOTSS) assessment was found to possess the greatest amount of evidence for validity.

It is worth noting that the present study, in contrast to the antecedent examination, has identified a more extensive array of assessment tools, totaling 16 distinct instruments. Teamwork assessment tools have been associated with various surgical procedures, as depicted in Table 1. Some tools, such as NOTSS and ANTS, have been utilized in general surgery, vascular surgery, and emergencies, such as hemorrhage and airway emergencies. Other tools, such as NOTECHS, OTAS, and CATS, have found application in a range of surgical specialties, including maxillofacial surgery, pediatric neurosurgery, urological surgery, and cardiac surgery. The specific procedures may vary depending on the study or evaluation context, and, in some cases, the specific procedure with which certain tools were employed has not been reported.

Furthermore, it is evident that the teamwork assessment tools have predominantly been employed in the operating room (OR) and simulation environments. Notably, several tools, including NOTSS, NOTECHS, ANTS, SPLINTS, OTAS, and the Surgical Teamwork Tool, have demonstrated their applicability in both the OR and simulation settings. On the other hand, certain tools, such as BMS-NNTS, Cannon-Bowers Scale, CATME, CATS, HFRS-M, ICARS, ORTAS, OSANTS, and SO-DIC-OR, have primarily been utilized in simulation environments. It is important to acknowledge that the specific settings may vary depending on the study or evaluation context, and some tools may not have reported the specific setting in which they were employed. The amount and variety of validation and reliability evidence varied greatly between each tool. Among the original 16 tools, five were established surgical teamwork assessment tools that were developed at least ten years ago, including the Non-Technical Skills for Surgeons (NOTSS) tool. These tools have been evaluated multiple times, affording them greater validity and reliability evidence, as illustrated in Table 2 with the example of NOTSS. It is also intriguing to note that, of the 26 total tools included in the review, only two tools were designed to assess team behaviors through event coding, rather than using a numerical or Likert-type scale. These two tools are the Simultaneous Observation of Distractors and Communication in the Operating Room (SO-DIC-OR) and Behavioral Marker System for Assessing Neurosurgical Non-Technical Skills (BMS-NNTS).

Newer assessment tools, such as HFRS-M and SO-DIC-OR, lack validity or reliability evidence aside from their original development article. Regarding SO-DIC-OR, no explicit validity evidence was mentioned; thus, the sole validity evidence for content validity was inferred from its development using other tools, such as NOTSS, as reference (Table 2).

Table 3. Comparison of the reliability evidence for the surgical teamwork assessment tools.

Tool Name	Reliability			
	Inter-rater	Intra-rater	Internal consistency	Test–retest
NOTSS	Moderate[3,21,24]–High[2,24]	-	Low[24]; High[22,23]	-
NOTECHS	Low[27]; High[25,26,28]	-	-	-
ANTS	Low[21,29]; High[2]	-	-	-
SPLINTS	Moderate[21]–High[30,31]	-	Low[31]–Moderate[30]	High[30]
OTAS	Moderate[7,21]–High[7,11]	-	-	-
BMS-NNTS	Moderate[34]	-	-	-
Cannon-Bowers Scale	-	-	High[22]	-
CATME	-	-	High[35]	-
CATS	-	-	-	-
HUFOES	High[14]	-	-	-
HFRS-M	-	-	-	-
ICARS	Low–Moderate[37]	-	High[37]	-
ORTAS	-	-	-	-
OSANTS	High[39]	-	High[39]	-
SO-DIC-OR	High[40]	-	-	-
The Surgical Teamwork Tool	Moderate–High[41]	-	-	-

NOTSS: Non-Technical Skills for Surgeons, NOTECHS: Oxford Non-Technical Skills, ANTS: Anesthetists' Non-Technical Skills, SPLINTS: Scrub Practitioners' List of Intraoperative Non-Technical Skills, OTAS: Observational Teamwork Assessment for Surgery, BMS-NNTS: Behavioral Marker System for Assessing Neurosurgical Non-Technical Skills, CATME: Comprehensive Assessment of Team Member Effectiveness, CATS: Communication and Teamwork Skills, HUFOES: Human Factors in Intraoperative Ophthalmic Emergencies Scoring System, HFRS-M: Human Factors Rating Scales—Modified, ICARS: Interpersonal and Cognitive Assessment for Robotic Surgery, ORTAS: Operating Room Teamwork Assessment Scales, OSANTS: Objective Structured Assessment of Non-technical Skills, SO-DIC-OR: Simultaneous Observation of Distractions and Communications in the Operating Room, -: not reported.

For HFRS-M, no evidence of reliability was found in the development article, nor were other articles located that evaluated the tool's reliability (Table 3). Another concern arose regarding whether the reported value r = 0.86 from Jung and colleagues' article would classify as an instance of construct or concurrent validity, as this was evidence of construct concurrent validity for NOTSS as per the authors.[3] Ultimately, it was placed under the concurrent validity category as the authors reported concurrent validity was assessed by calculating the Pearson correlation coefficients between the NOTSS ratings of surgical team and those of individual attending surgeons.

The average statistical value and range for reliability evidence were calculated whenever enough information was provided to perform the calculation. It was frequently observed that the calculated average statistical value did not correspond to the article's stated average statistical value for inter-rater reliability. In the case of Phitayakorn and colleagues, the stated range for inter-rater reliability evidence for Anesthetists' Non-Technical Skills (ANTS) did not align with the information provided in a table of absolute agreement between pairs of observers for each operating room team.[21] As various methods were applied to evaluate reliability, direct comparison of reliability evidence across tools proved challenging.

Given that most tools, except ANTS, Non-Technical Skills (NOTECHS), NOTSS, Observational Teamwork Assessment for Surgery (OTAS), and Scrub Practitioners' List of Intraoperative Non-Technical Skills (SPLINTS), were developed less than 10 years ago, insufficient time has elapsed for a comprehensive review of their validity and reliability. This includes certain tools that were translated, modified, revised, or adapted from the aforementioned instruments. As most recent tool development articles primarily serve as the source for validity and reliability evidence, further studies must be conducted to comprehensively evaluate these tools. Additionally, other forms of validity and reliability, apart from content validity and inter-rater reliability, respectively, should be assessed to thoroughly evaluate each tool. In some instances, other studies have been conducted to assess the validity, reliability, and usability of the recently developed tools, albeit often performed by the same group of authors.

As it stands, ANTS, NOTECHS, NOTSS, OTAS, and SPLINTS can be considered more suitable for deployment within their respective environments compared to more recently developed assessment tools. Novel tools intended to assess teams in specific surgical specialties or to introduce new approaches for assessing non-technical skills and teamwork in a surgical setting lack sufficient validity and reliability evidence for recommendation.

Limitations

Citation tracking could only feasibly be performed using Scopus and Web of Science, which may limit the number of articles, the years covered and the scope of the review.

Recommendations and future direction

Among the assessed tools, ANTS, NOTECHS, NOTSS, OTAS, and SPLINTS, which were developed more than 10 years ago, appear to be better suited for their respective environments compared to the newer tools. Recent tool innovations, particularly those tailored to specific surgical specialties or offering innovative methodologies for assessing non-technical skills and teamwork, necessitate further validation before meriting widespread adoption. Subsequent research endeavors should focus on addressing gaps in validity and reliability evidence while exploring additional dimensions of assessment tool performance, such as usability, to ensure the effective team evaluation within surgical contexts.

Furthermore, the validation of recently developed teamwork assessment tools and the enhancement of the objectivity of the assessment warrant continued investigation. Leveraging room cameras for video analysis can enhance the reliability of teamwork assessment by facilitating assessment by multiple evaluators.[60] Video recordings can aid in the identification and quantification of specific human behaviors that contribute to effective team collaboration, such as gaze overlap[61] and anticipatory movements reflecting changes in communication patterns and collaboration without following verbal requests.[62] Additionally, the use of OR Black Box technology, which captures audiovisual data, allows for the retrospective analysis of teamwork, offering potential contributions to more objective teamwork assessments in the future.[63]

To enable a direct comparison of teamwork assessment across surgical team and settings, the implementation of techniques such as eye tracking and electroencephalography (EEG) among surgical team members holds promise. This approach would obviate reliance on human raters and mitigate the variability in assessment scores. Given the current variability and low inter-rater reliability (e.g. <0.02) observed in certain domains of some tools, there is a compelling need for more objective measures in the evaluation of teamwork.

Conclusion

In conclusion, while established assessment tools such as ANTS, NOTECHS, NOTSS, OTAS, and SPLINTS have demonstrated good validity and reliability in general surgery, newer tools developed in recent years require further evaluation to establish their validity and reliability. There is a need to integrate these tools into surgical training and conduct additional research to gather evidence for their effectiveness. Objective assessments play a vital role in enhancing reliability and objectivity in surgical teamwork evaluation, ultimately leading to more accurate and standardized assessments. Continued exploration and innovation in assessment methodologies are essential to drive advancements in the field and improve the quality of surgical teamwork assessment.

Conflict of interest

The authors have no conflicts of interest to declare.

Data availability

Data sharing is not applicable to this article as no new data were created or analyzed.

References

1. Healey AN, Undre S, Vincent CA. Developing observational measures of performance in surgical teams. Qual Saf Health Care 2004; 13 Suppl 1: i33-40. https://doi.org/10.1136/qhc.13.suppl_1.i33
2. Doumouras AG, Hamidi M, Lung K, Tarola CL, Tsao MW, Scott JW, et al. Non-technical skills of surgeons and anaesthetists in simulated operating theatre crises. Br J Surg 2017; 104 (8): 1028-1036. https://doi.org/10.1002/bjs.10526
3. Jung JJ, Yule S, Boet S, Szasz P, Schulthess P, Grantcharov T. Nontechnical skill assessment of the collective surgical team using the Non-Technical Skills for Surgeons (NOTSS) system. Ann Surg 2020; 272(6): 1158-1163. https://doi.org/10.1097/SLA.0000000000003250
4. Siu J, Maran N, Paterson-Brown S. Observation of behavioural markers of non-technical skills in the operating room and their relationship to intra-operative incidents. Surgeon 2016; 14(3): 119-128. https://doi.org/10.1016/j.surge.2014.06.005
5. Aouicha W, Tlili MA, Limam M, Sněne M, Ben Dhiab M, Chelbi S, et al. Evaluation of the impact of intraoperative distractions on teamwork, stress, and workload. J Surg Res 2021; 259: 465-472. https://doi.org/10.1016/j.jss.2020.09.006
6. Dellefield ME, Verkaaik CA. Using the observational teamwork assessment in surgery instrument to measure RN teamwork during cardiac surgery: lessons learned. J Nurs Care Qual 2021; 36(2): 162-168. https://doi.org/10.1097/NCQ.0000000000000497
7. Hull L, Arora S, Kassab E, Kneebone R, Sevdalis N. Observational teamwork assessment for surgery: content validation and tool refinement. J Am Coll Surg 2011; 212(2): 234-243. e5. https://doi.org/10.1016/j.jamcollsurg.2010.11.001

8. Russ S, Hull L, Rout S, Vincent C, Darzi A, Sevdalis N. Observational teamwork assessment for surgery: feasibility of clinical and nonclinical assessor calibration with short-term training. Ann Surg 2012; 255(4): 804-809. https://doi.org/10.1097/SLA.0b013e31824a9a02
9. 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. https://doi.org/10.1097/SLA.0b013e3181a50220
10. Tsuburaya A, Soma T, Yoshikawa T, Cho H, Miki T, Uramatsu M, et al. Introduction of the non-technical skills for surgeons (NOTSS) system in a Japanese cancer center. Surg Today 2016; 46(12): 1451-1455. https://doi.org/10.1007/s00595-016-1322-8
11. 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. https://doi.org/10.1007/s00268-007-9053-z
12. Heale R, Twycross A. Validity and reliability in quantitative studies. Evid Based Nurs 2015; 18(3): 66-67. https://doi.org/10.1136/eb-2015-102129
13. Whittaker G, Abboudi H, Khan MS, Dasgupta P, Ahmed K. Teamwork assessment tools in modern surgical practice: a systematic review. Surg Res Pract 2015; 494827. https://doi.org/10.1155/2015/494827
14. Wood TC, Maqsood S, Zoutewelle S, Nanavaty MA, Rajak S. Development of the HUman Factors in intraoperative Ophthalmic Emergencies Scoring System (HUFOES) for non-technical skills in cataract surgery. Eye (Lond) 2021; 35(2): 616-624. https://doi.org/10.1038/s41433-020-0921-1
15. Szasz P, Louridas M, Harris KA, Aggarwal R, Grantcharov TP. Assessing technical competence in surgical trainees: a systematic review. Ann Surg 2015; 261(6): 1046-1055. https://doi.org/10.1097/SLA.0000000000000866
16. He W. Surgical team and team assessment: psychomotor evidence [Doctoral dissertation]. Edmonton (AB): University of Alberta; 2019.
17. Portney L, Watkins M. Foundations of clinical research: applications to practice. 3rd ed. New Jersey: Prentice Hall; 2007.
18. Balk E, Gazula A, Markozannes G, Kimmel HJ, Saldanha IJ, Resnik LJ, et al. Lower limb prostheses: measurement instruments, comparison of component effects by subgroups, and long-term outcomes [Internet]. Rockville (MD): Agency for Healthcare Research and Quality (US); 2018; Comparative Effectiveness Review, No. 213. Table 2, Metrics for evaluation of reliability, validity, and related psychometric properties. Available from: https://www.ncbi.nlm.nih.gov/books/NBK531518/table/ch3.tab1/. https://doi.org/10.23970/AHRQEPCCER213
19. Weber M, Van Ancum J, Bergquist R, Taraldsen K, Gordt K, Mikolaizak AS, et al. Concurrent validity and reliability of the Community Balance and Mobility scale in young-older adults. BMC Geriatr 2018; 18(1): 156. https://doi.org/10.1186/s12877-018-0845-9
20. Koutantji M, McCulloch P, Undre S, Gautama S, Cunniffe S, Sevdalis N, et al. Is team training in briefings for surgical teams feasible in simulation? Cogn Technol Work 2008; 10(4): 275-285. https://doi.org/10.1007/s10111-007-0089-5
21. Phitayakorn R, Minehart R, Pian-Smith MCM, Hemingway MW, Milosh-Zinkus T, Oriol-Morway D, et al. Practicality of intraoperative teamwork assessments. J Surg Res 2014; 190(1): 22-28. https://doi.org/10.1016/j.jss.2014.04.024
22. Pugh CM, Cohen ER, Kwan C, Cannon-Bowers JA. A comparative assessment and gap analysis of commonly used team rating scales. J Surg Res 2014; 190(2): 445-450. https://doi.org/10.1016/j.jss.2014.04.034
23. Yule S, Gupta A, Gazarian D, Geraghty A, Smink DS, Beard J, et al. Construct and criterion validity testing of the Non-Technical Skills for Surgeons (NOTSS) behaviour assessment tool using videos of simulated operations. Br J Surg 2018; 105 (6): 719-727. https://doi.org/10.1002/bjs.10779
24. 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. https://doi.org/10.1007/s00268-007-9320-z
25. McCulloch P, Mishra A, Handa A, Dale T, Hirst G, Catchpole K. The effects of aviation-style non-technical skills training on technical performance and outcome in the operating theatre. Qual Saf Health Care 2009; 18(2): 109-115. https://doi.org/10.1136/qshc.2008.032045
26. Catchpole KR, Dale TJ, Hirst DG, Smith JP, Giddings TAEB. A multicenter trial of aviation-style training for surgical teams. J Patient Saf 2010; 6(3): 180-186. https://doi.org/10.1097/PTS.0b013e3181f100ea
27. Raîche I, Moloo H, Schoenherr J, Boet S. Interprofessional simulation: the challenges of teamwork training. Perioper Care Oper Room Manag 2021; 24: 100180. https://doi.org/10.1016/j.pcorm.2021.100180
28. Mishra A, Catchpole K, McCulloch P. The Oxford NOTECHS system: reliability and validity of a tool for measuring teamwork behaviour in the operating theatre. Qual Saf Health Care 2009; 18(2): 104-108. https://doi.org/10.1136/qshc.2007.024760
29. 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. https://doi.org/10.1097/SIH.0b013e31821dfd05
30. Loh HP, De Korne DF, Yin SQ, Ang E, Lau Y. Assessment of Scrub Practitioners' List of Intraoperative Non-Technical Skills (SPLINTS) in an Asian ambulatory surgical setting. AORN J 2019; 109(4): 465-476. https://doi.org/10.1002/aorn.12640
31. Mitchell L, Flin R, Yule S, Mitchell J, Coutts K, Youngson G. Evaluation of the Scrub Practitioners' List of Intraoperative

Non-Technical Skills system. Int J Nurs Stud 2012; 49(2): 201-211. https://doi.org/10.1016/j.ijnurstu.2011.08.012

32. Mitchell L, Flin R, Yule S, Mitchell J, Coutts K, Youngson G. Development of a behavioural marker system for scrub practitioners' non-technical skills (SPLINTS system). J Eval Clin Pract 2013; 19(2): 317-323. https://doi.org/10.1111/j.1365-2753.2012.01825.x
33. Undre S, Healey AN, Darzi A, Vincent CA. Observational assessment of surgical teamwork: a feasibility study. World J Surg 2006; 30(10): 1774-1783. https://doi.org/10.1007/s00268-005-0488-9
34. Michinov E, Jamet E, Dodeler V, Haegelen C, Jannin P. Assessing neurosurgical non-technical skills: an exploratory study of a new behavioural marker system. J Eval Clin Pract 2014; 20(5): 582-588. https://doi.org/10.1111/jep.12152
35. Andrew B, Plachta S, Salud L, Pugh CM. Development and evaluation of a decision-based simulation for assessment of team skills. Surgery 2012; 152(2): 152-157. https://doi.org/10.1016/j.surg.2012.02.018
36. Frankel A, Gardner R, Maynard L, Kelly A. Using the Communication and Teamwork Skills (CATS) assessment to measure health care team performance. Jt Comm J Qual Patient Sa 2007; 33(9): 549-558. https://doi.org/10.1016/S1553-7250(07)33059-6
37. Raison N, Wood T, Brunckhorst O, Abe T, Ross T, Challacombe B, et al. Development and validation of a tool for non-technical skills evaluation in robotic surgery—the ICARS system. Surg Endosc 2017; 31(12): 5403-5410. https://doi.org/10.1007/s00464-017-5622-x
38. Paige JT, Garbee DD, Kozmenko V, Yu Q, Kozmenko L, Yang T, et al. Getting a head start: high-fidelity, simulation-based operating room team training of interprofessional students. J Am Coll Surg 2014; 218(1): 140-149. https://doi.org/10.1016/j.jamcollsurg.2013.09.006
39. Dedy NJ, Szasz P, Louridas M, Bonrath EM, Husslein H, Grantcharov TP. Objective structured assessment of nontechnical skills: reliability of a global rating scale for the in-training assessment in the operating room. Surgery 2015; 157(6): 1002-1013. https://doi.org/10.1016/j.surg.2014.12.023
40. Seelandt JC, Tschan F, Keller S, Beldi G, Jenni N, Kurmann A, et al. Assessing distractors and teamwork during surgery: developing an event-based method for direct observation. BMJ Qual Saf 2014; 23(11): 918-929. https://doi.org/10.1136/bmjqs-2014-002860
41. Huang LC, Conley D, Lipsitz S, Wright CC, Diller TW, Edmondson L, et al. The surgical safety checklist and teamwork coaching tools: a study of inter-rater reliability. BMJ Qual Saf 2014; 23(8): 639-650. https://doi.org/10.1136/bmjqs-2013-002446
42. Flin R, Maran N. Identifying and training non-technical skills for teams in acute medicine. Qual Saf Health Care 2004; 13 (Suppl 1): i80-84. https://doi.org/10.1136/qhc.13.suppl_1.i80
43. Loughry ML, Ohland MW, DeWayne Moore D. Development of a theory-based assessment of team member effectiveness. Educ Psychol Meas 2007; 67(3): 505-524. https://doi.org/10.1177/0013164406292085
44. Yule S, Gupta A, Blair PG, Sachdeva AK, Smink DS, American College of Surgeons Committee on Non-Technical S. Gathering validity evidence to adapt the Non-technical Skills for Surgeons (NOTSS) assessment tool to the United States context. J Surg Educ 2021; 78(3): 955-966. https://doi.org/10.1016/j.jsurg.2020.09.010
45. Pradarelli JC, Gupta A, Lipsitz S, Blair PG, Sachdeva AK, Smink DS, et al. Assessment of the Non-Technical Skills for Surgeons (NOTSS) framework in the USA. Br J Surg 2020; 107 (9): 1137-1144. https://doi.org/10.1002/bjs.11607
46. Spanager L, Lyk-Jensen HT, Dieckmann P, Wettergren A, Rosenberg J, Østergaard D. Customization of a tool to assess Danish surgeons' non-technical skills in the operating room. Dan Med J 2012; 59(11). PMID: 23171747.
47. 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. https://doi.org/10.1111/j.1365-2929.2006.02610.x
48. Gillespie BM, Harbeck E, Kang E, Steel C, Fairweather N, Chaboyer W. Correlates of non-technical skills in surgery: a prospective study. BMJ Open 2017; 7(1): e014480 https://doi.org/10.1136/bmjopen-2016-014480
49. Sevdalis N, Davis R, Koutantji M, Undre S, Darzi A, Vincent CA. Reliability of a revised NOTECHS scale for use in surgical teams. Am J Surg 2008; 196(2): 184-190. https://doi.org/10.1016/j.amjsurg.2007.08.070
50. Robertson ER, Hadi M, Morgan LJ, Pickering SP, Collins G, New S, et al. Oxford NOTECHS II: a modified theatre team non-technical skills scoring system. PLoS One 2014; 9(3): e90320. https://doi.org/10.1371/journal.pone.0090320
51. Kalantari R, Zanjirani Farahani A, Garosi E, Badeli H, Jamali J. Translation and psychometric properties of the Persian version of Oxford Non-technical Skills 2 System: assessment of surgical teams' non-technical skills in orthopedic surgery wards. Arch Bone Jt Surg 2019; 7(2): 173-181. PMID: 31211196.
52. Schreyer J, Koch A, Herlemann A, Becker A, Schlenker B, Catchpole K, et al. RAS-NOTECHS: validity and reliability of a tool for measuring non-technical skills in robotic-assisted surgery settings. Surg Endosc 2021; 36(3): 1916-1926. https://doi.org/10.1007/s00464-021-08474-2
53. Lyk-Jensen HT, Jepsen RM, Spanager L, Dieckmann P, Østergaard D. Assessing nurse anaesthetists' non-technical skills in the operating room. Acta Anaesthesiol Scand; 2014; 58 (7): 794-801. https://doi.org/10.1111/aas.12315
54. Hull L, Bicknell C, Patel K, Vyas R, Van Herzeele I, Sevdalis N, et al. Content validation and evaluation of an endovascular teamwork assessment tool. Eur J Vasc Endovasc Surg 2016; 52 (1): 11-20. https://doi.org/10.1016/j.ejvs.2015.12.044
55. Passauer-Baierl S, Hull L, Miskovic D, Russ S, Sevdalis N, Weigl M. Re-validating the observational teamwork assessment for surgery tool (OTAS-D): cultural adaptation, refinement,

and psychometric evaluation. World J Surg 2014; 38(2): 305-313. https://doi.org/10.1007/s00268-013-2299-8

56. Amaya Arias AC, Barajas R, Eslava-Schmalbach JH, Wheelock A, Gaitán Duarte H, Hull L, et al. Translation, cultural adaptation and content re-validation of the observational teamwork assessment for surgery tool. Int J Surg 2014; 12(12): 1390-1402. https://doi.org/10.1016/j.ijsu.2014.10.001
57. Bhatti NI, Cummings CW. Competency in surgical residency training: defining and raising the bar. Acad Med 2007; 82(6): 569-573. https://doi.org/10.1097/ACM.0b013e3180555bfb
58. Satava RM, Gallagher AG, Pellegrini CA. Surgical competence and surgical proficiency: definitions, taxonomy, and metrics. J Am Coll Surg 2003; 196(6): 933-937. https://doi.org/10.1016/S1072-7515(03)00237-0
59. Sharma B, Mishra A, Aggarwal R, Grantcharov TP. Non-technical skills assessment in surgery. Surg Oncol 2011; 20(3): 169-177. https://doi.org/10.1016/j.suronc.2010.10.001
60. He W, Zheng B. Collaborative performance in laparoscopic teams: behavioral evidences from simulation. Surg Endosc 2016; 30: 4569-4574. https://doi.org/10.1007/s00464-016-4794-0
61. He W, Jiang X, Zheng B. Synchronization of pupil dilations correlates with team performance in a simulated laparoscopic team coordination task. Simul Healthc 2021; 16(6): e206-e213. https://doi.org/10.1097/SIH.0000000000000548
62. Zheng B, Swanström LL, MacKenzie CL. A laboratory study on anticipatory movement in laparoscopic surgery: a behavioral indicator for team collaboration. Surg Endosc 2007; 21(6): 935-940. https://doi.org/10.1007/s00464-006-9090-y
63. Incze T, Pinkney SJ, Li C, Hameed U, Hallbeck MS, Grantcharov TP, et al. Using the operating room black box to assess surgical team member adaptation under uncertainty: an observational study. Ann Surg 2024; 280(1): 75-81. https://doi.org/10.1097/SLA.0000000000006191