Observer-based tools for non-technical skills assessment in simulated and real clinical environments in healthcare: a systematic review

Helen Higham,[1] Paul R Greig,[1] John Rutherford,[2] Laura Vincent,[1] Duncan Young,[1] Charles Vincent[3]

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136bmjqs-2018-008565)

[1]Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, UK
[2]Department of Anaesthetics, Dumfries and Galloway Royal Infirmary, Dumfries, UK
[3]Department of Experimental Psychology, University of Oxford, Oxford, UK

Correspondence to
Dr Helen Higham, Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford OX3 9DU, UK; helen.higham@ndcn.ox.ac.uk

Received 11 July 2018
Revised 17 April 2019
Accepted 23 April 2019

To cite: Higham H, Greig PR, Rutherford J, *et al*. *BMJ Qual Saf* Epub ahead of print: [please include Day Month Year]. doi:10.1136/bmjqs-2018-008565

ABSTRACT

Background Over the past three decades multiple tools have been developed for the assessment of non-technical skills (NTS) in healthcare. This study was designed primarily to analyse how they have been designed and tested but also to consider guidance on how to select them.

Objectives To analyse the context of use, method of development, evidence of validity (including reliability) and usability of tools for the observer-based assessment of NTS in healthcare.

Design Systematic review.

Data sources Search of electronic resources, including PubMed, Embase, CINAHL, ERIC, PsycNet, Scopus, Google Scholar and Web of Science. Additional records identified through searching grey literature (OpenGrey, ProQuest, AHRQ, King's Fund, Health Foundation).

Study selection Studies of observer-based tools for NTS assessment in healthcare professionals (or undergraduates) were included if they: were available in English; published between January 1990 and March 2018; assessed two or more NTS; were designed for simulated or real clinical settings and had provided evidence of validity plus or minus usability. 11,101 articles were identified. After limits were applied, 576 were retrieved for evaluation and 118 articles included in this review.

Results One hundred and eighteen studies describing 76 tools for assessment of NTS in healthcare met the eligibility criteria. There was substantial variation in the method of design of the tools and the extent of validity, and usability testing. There was considerable overlap in the skills assessed, and the contexts of use of the tools.

Conclusion This study suggests a need for rationalisation and standardisation of the way we assess NTS in healthcare and greater consistency in how tools are developed and deployed.

INTRODUCTION

Evidence that errors in non-technical skills (NTS) are common in adverse incidents in healthcare has been accruing over the past two decades.[1–5] NTS have been defined as 'the cognitive, social, and personal resource skills that complement technical skills, and contribute to safe and efficient task performance.'[6] They include such attributes as communication, teamwork, situation awareness, decision-making, task allocation and stress and fatigue management. It is worth highlighting that concern exists around the use of the term NTS[7] to describe such important aspects of professional clinical practice; however, while there is currently no universally agreed substitute[8] the term NTS will be used for this study.

Interest in evaluating and enhancing NTS in multiprofessional teams of healthcare workers has been increasing in line with concerns highlighted in studies of error in healthcare and a number of tools are now available for measuring them with many of the early examples adapted from the civil aviation field.[9–12] Concerns about the measurement properties of these tools (including their validity and reliability) have been raised by educational and research communities.[13–17] Assessment of healthcare professionals, particularly in high stakes settings such as examinations or interviews, requires rigorous attention to the quality of the tool being used to make that assessment if it is to be objective and fair. Furthermore, the choice of an appropriate tool for NTS assessment may be hampered by the large number available for different settings in healthcare.

This systematic review of the NTS assessment tools in healthcare seeks to provide a clearer understanding of the range, purpose, evidence of validity and usability of published tools.

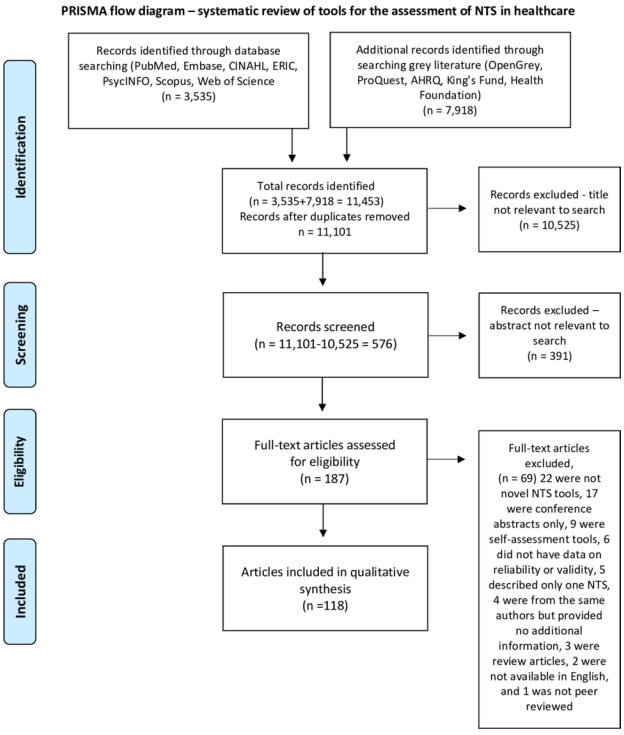

Figure 1 Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) diagram for non-technical skills (NTS) assessment tools.

Objectives

The objectives were:

- To provide an overview of observer-based assessment tools for performance of NTS in healthcare professionals or students in simulated or clinical environments.
- To describe the methods used in developing the tools.
- To explore the evidence provided for the validity and usability (including training required) of the tools.

METHODS

This systematic review was registered with PROSPERO (Ref No: CRD42017055445). Peer-reviewed studies were identified by search of the electronic bibliographic databases Medline, Embase, CINAHL, PsycINFO, Scopus and ERIC. A search of the grey literature was made via Google Scholar, ProQuest and OpenGrey. A manual search of the reference list of identified relevant articles was also conducted. No further searches were conducted after March 2018.

All reviewed articles were assessed using criteria defined by Hawker *et al* for mixed qualitative and quantitative research studies[18] (https://www.crd.york.ac.uk/prospero/#aboutpage). The inclusion and exclusion criteria are included below and the assessment questionnaire (as per Hawker) and a detailed search strategy are included as online supplementary appendix 1.

Inclusion criteria

Papers were eligible for inclusion where:

- They were published in the English language, or translation was available.
- The population studied comprised healthy adults working in healthcare settings.
- The publication date was between January 1990 and March 2018.
- They described a tool designed to assess NTS and included more than one of the following domains: communication, teamwork, situation awareness, decision-making and task allocation/management.
- They described a tool designed for use by direct observation or review of audiovisual files in a simulated or real clinical setting.
- Peer-reviewed papers were preferred but if a tool had been developed and only published as, for example, a thesis, this was highlighted.

Exclusion criteria

Papers were excluded where:

- Ethical approval of the study or informed consent from participants was not described.
- No data describing evidence of the tool's validity or reliability were available.
- The tool was designed for self-assessment only.
- The tool did not analyse performance under more than one of the key non-technical domains of: communication, situation awareness (sometimes described as vigilance), decision-making or task allocation/management.
- They described a tool used for the study of technical skills only.

Synthesis of results

Papers with potential for inclusion in the review on the initial search were first screened for relevance, by review of the title and by abstract review (see figure 1 for the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) review process). Papers with a relevant title and abstract were retained for full review. Papers without any assessment of validity or reliability for the NTS tool being used were discarded. Where papers were not retained for review, their reason for non-inclusion was recorded.

The first stage of the screening process was conducted for all papers in pairs (HH and PRG; HH and JR or PRG and JR)—where any disagreement was encountered a decision was made by the reviewer who was not a member of the original pair. Full-text articles were acquired for all abstracts put forward for further analysis. These were divided between the three reviewers for initial assessment and any ambiguities arising regarding inclusion were discussed and agreed together. The final in-depth analysis was then undertaken by HH and PRG with JR acting as final arbiter. All first authors were contacted by email, on two separate occasions, to seek additional unpublished information.

Most of the tools had already been given a name (eg, Team Emergency Assessment Measure—TEAM[19])

and, if not, we devised a name based on an approximation of the purpose of the tool (eg, anaesthetic trainee NTS[20]). A list of acronyms for all the tools in this review can be found in the online supplementary appendix 2.

The NTS assessed by the tools were usually described in categories, for example, communication, teamwork, leadership, and so on, which were underpinned by behavioural markers (eg, TEAM, Observational Teamwork Assessment for Surgery (OTAS), Oxford NOnTECHnical Skills (Oxford NOTECHS), Non-Technical Skills for Surgeons (NOTSS) and Ottawa CRM Global Rating Scale[19 21–24]) but some described an inventory of behaviours relevant to the context or professional group being analysed (eg, University of Texas Behavioural Markers for Neonatal Resuscitation, Mayo High Performance Teamwork Scale (MHPTS), Teamwork Behavioural Rater (TBR)[11 25 26]). We classified NTS into the five most commonly occurring categories: communication, leadership and/or teamwork, situation awareness, decision-making and task management. We also included an 'other' section to capture elements not ascribable to one of these categories. Examples where additional behaviours were assessed included: professionalism,[27 28] 'environment in the room'[29] and stress and distractors.[30] Where descriptors of behaviour were essentially a subcategory of one of the five domains they were included under the relevant heading, for example, cooperation was included under teamwork and vigilance under situation awareness.

Studies were analysed over three broad domains: method of development, the applicability and context of use of the tool and the evidence provided for validity of the tool (including any assessment of usability and training requirements). Where the *original* development and evidence of validity of a tool was described in more than one publication the data from all relevant papers were analysed, as long as at least one member of the original research team was involved.

Evidence of validity was classified (where possible) into domains described by the American Educational Research Association[31] which consider all forms of validity under the overarching term, 'construct validity':

- Content (ie, test items are representative of the construct of interest).
- Relations to other variables such as the ability to discriminate between learner characteristics (eg, between a good or a poor performance, or between levels of experience or professional groups) or relationships with separate measures (eg, that results from the assessment tool are related to those from a tool measuring another, similar construct, often called concurrent or convergent validity in the studies in this review).
- Internal structure (including: rater reliability and item correlations).
- Response process (ie, evidence of data integrity including methods for scoring and data entry).
- Consequences (intended or unintended consequences of an assessment—rarely reported).

Cook *et al*[17] have highlighted the difficulty of applying instruments used for clinical studies such as Standards for Reporting Diagnostic Accuracy[32] and Guidelines for Reporting Reliability and Agreement Studies[33] in the context of assessing tools for educational assessment. To provide some assistance to educators in selecting tools for NTS assessment we have categorised tools in terms of context of use, method of design, evidence of validity and assessment of usability (see table 1). The attributes we assessed were developed by the authors and informed by: the initial study assessment questionnaire (see above and online supplementary appendix 1); the iterative analysis of 118 studies; our experience as clinicians and educators; and guidance on design of educational assessment tools[34] (including validity and reliability[35–39] and team training assessments[40]).

Risk of bias

Data analysis and interpretation was undertaken with an awareness of the risk of bias. Repeated reflection on potential sources of bias in the context of personal beliefs and values (researcher reflexivity[41]) was integral to the iterative review of the studies. Study selection bias was minimised through use of a systematic search method.

Potential bias for the authors in reviewing the assessment tools included:

- Familiarity bias: four of the authors are active educators in simulation-based education (JR was the author of one of the tools (Anaesthetic NTS-Anaesthetic Practitioners[42]), CV has been involved in the development of other tools for NTS assessment[43–45]). The lead authors (HH, PRG and JR) have been trained to use the Anaesthetists' Non-Technical Skills (ANTS) assessment tool.
- Availability heuristics: the lead authors (HH, PRG and JR) are practising anaesthetists, as such our training and clinical experience is largely in theatre and intensive care unit settings.
- Anchoring bias: the order in which we reviewed the papers and the organisation of information presented in each study may influence decisions made in assessing the tools.

  Mitigations for these risks included development of a list attributes for analysis of tools (to provide a more objective framework for describing them, see table 1), review by more than one author and repeated re-examinations of the papers in random order.

RESULTS

The screening process is described in figure 1 as per PRISMA guidance. All articles included for review were observational studies of healthcare professionals or students in simulated or real clinical settings.

Table 1 Attributes assessed during analysis of 76 tools for the measurement of NTS in 118 papers

Applicability and context of use	
Applicability/environment	Participants rated (eg, intensive care MDT, surgeons, medical students) Individual or team assessment Clinical context of use (eg, ward, operating theatre etc and so) Simulated or real environments or both
Method of design and evidence of validity	
Subject matter experts involved in tool development	Relevant (multidisciplinary where appropriate) clinical expertise and any additional human factors or psychology expertise
Validity: content	Relevant literature review Evidence of structured, iterative (Delphi-type) process Formal task analysis used to determine relevant NTS
Validity: relationships with other variables	Learner characteristics (eg, level of experience, pretraining and post-training) Separate measures (eg, comparison with another NTS tool or a technical skills score)
Validity: internal structure	Internal consistency Inter-rater reliability Intrarater reliability Test–retest reliability
Usability	
Usability of tool	Description of training required to use the tool (eg, length of time taken to train; description of course content, any online materials) Quantitative assessment of usability (eg, time taken to complete assessment, completeness of score sheet) Qualitative assessment of usability (eg, informal feedback from raters, questionnaires completed by raters, interviews with raters)

Validity is described as per standards from the American Educational Research Association.[31]
Attributes were defined by the authors in an iterative process as described above.
MDT, multidisciplinary team; NTS, non-technical skills.

We identified 76 unique tools for the assessment of NTS in healthcare that were suitable for inclusion in the review. These were described in 118 papers. The first tool was developed by Gaba *et al*[9] in North America. Subsequently, most tools have been developed in North America (35 tools), followed by Europe (31 tools) and Australasia (8 tools). One tool was developed in Colombia[46] and one in Israel[47] (country of origin is shown in table 2 and the online supplementary appendix 3).

Most tools were developed de novo, but some were explicitly based on tools developed by other groups[48–51] and some relied on data gathered in the original tool. Self-assessment tools were excluded because, while they may be useful in formative settings, self-assessment of NTS is inaccurate and unsuitable for use in high stakes settings.[52]

Considerable variability was found in method of tool development, applicability, context of use and evidence of validity in this study, in line with previous systematic reviews of assessment.[17 53 54]

Methods of tool design and context of use

Methods of reporting observations varied. For example, number of observations made using the tool (eg, Behavioural Marker System - Neurosurgical Non-Technical Skills (BMS-NNTS)[55] and Explicit Professional Oral Communication (EPOC)[56] include an assessment of frequency of interactions), or number of participants or teams observed (some had large numbers of observations or participants[56–58] and others fewer[49 59 60]), and some were individual or team assessments or both, as shown in table 2. Consequently, it was difficult to make meaningful inferences between the studies.

Most assessment tools (37 (49%)) had been designed for use with multidisciplinary teams; 27 (36%) were for single specialty postgraduate healthcare professionals; 8 (10%) were for the assessment of healthcare students; and 4 (5%) were for multispecialty postgraduate doctors (see table 2 and online supplementary appendix 3).

The environments in which the tools were designed and tested varied but fell under two broad domains—simulated or real clinical settings, and context of use included seven clinical domains: adult inpatient (7 tools (9%)); adult intensive/emergency care (21 tools (28%)); obstetrics (4 tools (5%)); operating theatres (adult and paediatric—25 tools (33%)); paediatric intensive/emergency care (5 tools (7%)); prehospital care (3 tools (4%)); and generic healthcare settings (3 tools (4%)). Tools for the assessment of NTS in undergraduates (8 tools (10%)) were put in a separate category from postgraduate tools (because the authors did) but there were not enough to warrant further subdivision by clinical domain.

NTS categories assessed were also variable. Communication was assessed in every tool although not always as an isolated category (eg, Oxford NOTECHS and ANTS). Teamwork and leadership were the next most commonly included categories (74 (97%) of tools), situation awareness was assessed in 66 (87%), task management in 61 (80%) and decision-making in 36 (47%).

Data for 30 tools grouped by context of use as described above are shown in table 2 (tools are listed

Table 2 Description of environment, context of use and scoring for 30 tools for the assessment of NTS in healthcare

NTS tool name—grouped by context of use (acronym if used; country of origin), author, year of publication	Environment of use	Participants rated	Score: individual or team rating	Comments
Operating theatre				
Anaesthetists' Non-Technical Skills (ANTS; UK), Fletcher *et al*,[10 135 136] 2003	Simulated and real	Anaesthetists	**Individual** video review and real-time rating (4-point scale)	ANTS based on NOTECHS aviation NTS tool. Research group have developed many tools.*
Observational Teamwork Assessment for Surgery (OTAS; UK), Healey *et al*,[21 116 127 137] 2004	Simulated and real	Operating theatre MDT	**Subteams**: nurses, surgeons, anaesthetists and **global team score,** video review and real-time rating (7-point scale)	Technical and NTS scores across three time periods (preoperative, intraoperative, postoperative). Research group have developed many tools.†
Revised Non-Technical Skills scale (Revised NOTECHS; UK), Moorthy *et al*,[43 125 138 139] 2005	Simulated	Operating theatre MDT	Whole team, real-time rating (6-point scale)	Research group have developed many tools.†
Non-Technical Skills for Surgeons (NOTSS; UK), Yule *et al*,[23 108 115 130] 2006	Simulated and real	Surgeons	**Individual**, video review and real-time rating (4-point scale)	Research group have developed many tools.*
Oxford NOnTECHnical Skills (Oxford NOTECHS; UK), Mishra *et al*,[12 22 140] 2009	Simulated and real	Operating theatre MDT	**Subteams**: surgical, anaesthetic, nursing and **whole team**, video review and real-time rating (8-point scale)	Scoring system revised in 2014
Non-technical skills for anaesthetic trainees (UK), Gale *et al*,[20 141] 2010	Simulated and real	Anaesthetic trainees	**Individual**, real-time rating (4-point scale)	Designed to assist in the recruitment process for anaesthetists then tested in the workplace
Paediatric Cardiac Surgery Teamwork classification tool (PCST; Netherlands), Schraagen *et al*,[126 142] 2010	Real	Paediatric cardiac surgery MDT	**Subteams**: surgeons, anaesthetists, perfusionists, nurses; video review and real-time rating (7-point scale)	Tool used by non-clinical raters (human factor experts). Subteam scores aggregated to whole team score.
Scrub Practitioners List of NTS (SPLINTS; UK), Mitchell *et al*,[110 143] 2011	Simulated and real	Scrub practitioners	**Individual**, video review (4-point scale)	
Nurse Anaesthetists' Non-Technical Skills-Denmark (NANTSdk), Lyk-Jensen *et al*,[114 144] 2014	Simulated and real	Nurse anaesthetists in Denmark	**Individual**, video review (7-point scale)	Tool based on ANTS
Objective Structured Assessment of Non-Technical Skills (OSANTS; Canada), Dedy *et al*,[28] 2015	Simulated and real	Surgical trainees	**Individual**, video review and real-time rating (5-point scale)	Tool designed to assess surgeons in training
WHO Behaviourally Anchored Rating Scale (WHOBARS; New Zealand), Devcich *et al*,[145] 2015	Simulated and real	Operating theatre MDT for WHO checklist	**Whole team**, video review and real-time rating (7-point scale)	Tool only for WHO checklist
Anaesthetists' NTS-Denmark (ANTSdk), Jepsen *et al*,[113 146 147] 2015	Simulated	Anaesthetists in Denmark	**Individual**, video review (5-point scale for categories and elements, 7-point scale for global score)	Based on ANTS. Research group also developed NANTSdk and NOTSSdk.
Anaesthetic NTS-Anaesthetic Practitioners (ANTS-AP; UK), Rutherford *et al*[42], 2015	Simulated and real	Anaesthetic practitioners	**Individual**, video review (4-point scale)	Research group have developed many tools.*
Interpersonal and Cognitive Assessment for Robotic Surgery (ICARS; UK[c]), Raison *et al*,[30] 2017	Simulated	Surgeons: robotic surgery only	**Individual**, video review (5-point scale)	Highly specific tool for use during robotic surgery
Adult intensive/emergency care				
Team Dimensions Rating Form (USA), Morey *et al*,[58] 2002	Real	ED MDT	**Whole team**, real-time rating (7-point scale)	Designed for large in situ study of team training for ED teams (MedTeam). Early NTS tool (based on aviation NTS tool).
Ottawa CRM Global Rating Scale (Ottawa GRS; Canada), Kim *et al*,[148 149] 2006	Simulated	Trainee doctors (any specialty)	**Individual**, video review (7-point scale)	Based BAR for ACRM, designed for CRM course
Mayo High Performance Teamwork Scale (MHPTS; USA), Malec *et al*,[25] 2007	Simulated	Any MDT in CRM training	**Whole team**, real-time rating (3-point scale)	Originally used as self-assessment scale, but MHPTS subsequently used to develop observer-based tools
Team Emergency Assessment Measure (TEAM; Australia), Cooper *et al*,[19 150–152] 2010	Simulated and real	ED MDT	**Whole team**, video review and real-time ratings (4-point scale for items, 10-point scale for global score)	Used as basis for other tools
Observational Skill-based Clinical Assessment tool for Resuscitation (OSCAR; UK), Walker *et al*,[153] 2011	Simulated and real	Resuscitation MDT	**Subteams**: anaesthetic, physician and nurse, video review (6-point scale)	Research group have developed many tools.†
Trauma NOTECHS (T-NOTECHS; USA), Steinemann *et al*,[154] 2012	Simulated and real	Trauma MDT	**Whole team**, video review and real-time rating (5-point scale)	Tool based on OTAS and revised NOTECHS

Continued

Table 2 Continued

NTS tool name—grouped by context of use (acronym if used; country of origin), author, year of publication	Environment of use	Participants rated	Score: individual or team rating	Comments
Adult inpatient				
Team Functioning Assessment Tool (TFAT; Australia), Sutton *et al*,[155 156] 2011	Simulated and real	Ward MDT	**Whole team**, video review and real-time rating (7-point scale)	Tool modified from 61 to 40 behavioural items
Teamwork Mini-clinical Evaluation Exercise (T-MEX; Australia), Olupeliyawa *et al*,[111] 2014	Real	Medical students or trainee doctors	**Individual**, real-time rating (5-point scale)	Tool also used for self-assessment
Surgical Ward-round Assessment Tool (SWAT; UK), Ahmed *et al*,[157] 2015	Simulated and real	Surgeons on ward rounds	**Individual**, video review and real-time rating (5-point scale)	Tool combines NTS with task checklist. Research group have developed many tools.†
Paediatric intensive/emergency care				
University of Texas Behavioural Markers for Neonatal Resuscitation (UTBMNR; USA), Thomas *et al*,[11 158 159] 2004	Simulated and real	Neonatal MDT	**Whole team**, video review and real-time rating (5-point scale)	Team had extensive experience in the design of tools for aviation.
Obstetrics				
Assessment of Obstetrical Team Performance (AOTP) and Global AOTP (Canada), Tregunno *et al*,[29 107 160] 2009	Simulated	Obstetric MDT	**Whole team**, video review (5-point scale)	Two systems evaluated: AOTP 18 items, Global AOTP (GAOTP) 6 items
Perinatal Emergency Team Response Assessment (PETRA; Canada), Balki *et al*,[109 161] 2017	Simulated and real	Obstetric MDT	**Whole team**, video review (5-point scale)	New tool, not yet tested in real environment
Prehospital care				
Aero-NOnTechnical Skills (AeroNOTS; New Zealand), Myers *et al*,[162] 2016	Simulated	Doctors in aeromedical transport	**Individual**, video review (5-point scale for elements and categories, 7-point global rating)	Based on ANTS, score adapted
Generic healthcare environment				
Clinical Teamwork Scale (CTS; USA), Guise *et al*,[117] 2008	Simulated and real	Any healthcare MDT	**Whole team**, video review (10-point scale)	One of the few generic tools. Based on aviation CRM assessment tool.
Undergraduate				
Standardised Assessment For the Evaluation of Team Skills (SAFE-TeamS; USA), Wright *et al*,[163] 2013	Simulated	Medical and nursing undergraduates	**Individual**, real-time rating (2-point scale)	Designed to be applicable to medical and nursing education in standardised simulated settings
Individual Teamwork Observation and Feedback Tool (iTOFT; Australia[c]), Thistlethwaite *et al*,[112] 2016	Simulated and real	Undergraduate MDT	**Individual**, real-time assessment (3-point scale)	Basic and advanced iTOFT for junior and senior students, respectively. Tool can be used for self-assessment.

'[c]' tool designed with international collaboration.
*Research group developed: ANTS, NOTSS, SPLINTS, FoNTS, ANTS-AP.
†Research group developed: Revised NOTECHS, OTAS, OSCAR, Endo-OTAS, Imperial Paediatric Emergency Training Toolkit (IPETT), SWAT and Emergency Dr NTS.
ACRM, Anaesthesia Crisis (Crew) Resource Management; BAR, Behaviourally Anchored Rating; CRM, crisis (crew) resource management; ED, Emergency Department; FoNTS, Foundation Non-Technical Skills; MDT, multidisciplinary team; NTS, non-technical skills.

chronologically, and if there were more than one tool from the same year, in alphabetical order of the author's name). Space constraints prevent all 76 being shown here. Those that are shown presented more detail on method of development and the greatest amount of evidence for validity (including reliability), requirements for training and usability. Data for the remaining 46 tools are available as online supplementary appendix 3, and the references are shown below, categorised by context of use (all papers describing tools are included):

- Operating theatres (11 tools[9 46 49 55 61–69]).
- Adult intensive/emergency care (15 tools[26 45 47 51 56 59 70–81]).
- Adult inpatient (4 tools[82–85]).
- Paediatric intensive/emergency care (4 tools[60 86–90]).
- Obstetrics (2 tools[91 92]).
- Prehospital care (2 tools[93 94]).
- Generic health environment (2 tools[95–97]).
- Undergraduate education (6 tools[27 57 98–102]).

Evidence of validity and description of training requirements and usability

The argument-based approach to validity[35 103 104] was used to assess the tools but this was limited by the variability in the provision of evidence and because the majority of papers referred to validity using more traditional terms. Validity was classified (where possible) into domains described by the American Educational Research Association[31]: content, response process, internal structure and relations to other variables and consequences. All tools assessed content validity in some form and the next most common assessment was relation to learner characteristics such as experience or educational level of participants (47 tools (62%)). Tests of relationships with separate measures including tools measuring similar, related constructs (25 (33%)) were more common than those testing tools against others measuring the same construct (frequently these tests were termed convergent or concurrent validity) (11

(14%)) and only three groups considered predictive validity in the sense of ability to predict future performance.[105–107] Some tools contained a technical as well as an NTS assessment but not all of them assessed the relationship with the NTS items (see table 3).

Reliability was most commonly assessed with inter-rater testing (61 tools (80%)) or internal consistency (41 tools (54%)). Only 11 studies (14%) considered intra-rater or test–retest reliability.

Some authors went to great lengths to analyse usability and generated qualitative and quantitative data from questionnaires or interviews (which informed the development and deployment of their assessment tools).[10 29 30 42 108–114]

Recommendations for training were described in very different ways, from those who have designed bespoke courses for their tools (eg, NOTSS,[115] OTAS[116] and Multiprofessional Inventory for Non-Technical Skills in the Delivery Room (MINTS-DR)[92]) to those where a tool was designed with a specific remit of not requiring much training to use it (TEAM,[19] MHPTS,[25] Perinatal Emergency Team Response Assessment (PETRA)[109] and Clinical Teamwork Scale (CTS)[117]). Table 3 provides an overview of validity evidence, training requirements and usability assessments for the same 30 tools in table 2 (the same information for the remaining 46 tools is found in the online supplementary appendix 3).

DISCUSSION

We have analysed the growing array of NTS assessment tools in healthcare since the first was developed in 1998 by Gaba *et al*.[9] Box 1 highlights what this study adds to the field.

Method of development

The importance of measures which assess whole team performance has been highlighted by several authors;[40 118 119] while the training and assessment of NTS in individuals is important[120] some tools allowed more flexibility (ie, they could be used for more than one profession or environment).

Instruments varied in their intended purpose, some assessed routine teamwork while others focused on management of crisis scenarios. Simulated settings allow control of scenarios and reliable depiction of behaviours (often by actors). However, it has been suggested that it is not truly representative of a real clinical environment where there may be long periods of relative calm with short bursts of intense activity, whereas a video of a simulated crisis will only focus on the 15 min or so of high pressure.[121] It would, therefore, seem desirable to develop tools that might be used in both settings to provide meaningful assessments during training and real clinical practice and in routine as well as emergency situations.

The NTS domains assessed were broadly similar across all the tools, suggesting that they are relevant in a wide variety of clinical settings with the appropriate context-specific adaptations, which begs the question: why are there so many? Authors frequently stated that the reason for the development of a new tool was the lack of one relevant to their specific need. The answer may also be found, to a degree, in the necessity for compromise highlighted by van der Vleuten,[34] who described five key components in considering the utility of assessment methods: educational impact, validity, reliability, cost and acceptability (both to examiners and examinees). He stressed that 'choosing an assessment method inevitably entails compromise and the type of compromise varies for each specific assessment context' and 'perfect utility is a utopia.'

Usability

The issue of usability and cost of NTS assessment tools is not trivial, and has been brought into sharp relief by the current staff shortages in healthcare and difficulties in releasing staff to train.[122]

A formative training event may benefit from the use of a tool which requires little training to implement and brings additional richness to the debriefing. However, in high stakes settings evidence of validity and reliability for an assessment tool must be robust and those using it must be trained and experienced in so doing.

Most of the in-depth analysis of usability has occurred in tools developed in the past 5 years, suggesting a heightened awareness of the need to consider the practical use of such assessments.

Training requirements

The challenges of assessing NTS accurately and reliably have been enumerated by Flin *et al*[120] and Smith-Jentsch *et al*[123] (eg, difficulty seeing and hearing all the relevant information; difficulty interpreting cognitive skills and rare but important behaviours that may be missed because they are not categorised). Many of the research teams who have designed these tools pointed out the challenges of using them and suggestions for best practice have been put forward by an expert group from aviation and healthcare.[124] Furthermore, Gaba *et al*,[9] Moorthy *et al*[125] and Schraagen *et al*[126] highlight the value of simplifying the number of NTS domains analysed by a tool in order to improve the reliability of the observers.

While this approach may be more cost-effective, Sevdalis *et al* showed the value of psychologist or human factor expert raters in using OTAS[127] but also recognised the resource implications. A later paper using OTAS showed that it was possible to train clinical staff to assess behaviours reliably in a short space of time.[116] Guidelines for the training of faculty in NTS assessment have since been published[128] and they stress the importance of training to ensure reliability, particularly for high stakes settings. The authors suggest a minimum requirement of 2 days' training

Table 3 Evidence of validity, training requirements and assessment of usability for 30 tools for the assessment of NTS in healthcare

NTS tool name (year of publication)	Content	Internal structure/ reliability	Relations with other variables	Training required to use tool and assessment of tool usability
Operating theatre				
ANTS, Fletcher *et al*[10 135]	LR, Delphi, SME plus, TA	Internal consistency, inter-rater reliability	Learner characteristics (experience level)	Training course designed with handbook and online materials. Quantitative and qualitative assessment of usability.
OTAS, Healey *et al*[137]	LR, Delphi, SME plus, TA	Inter-rater reliability, intrarater reliability	Learner characteristics, separate measures (task checklist)	Training course designed with handbook and online materials. Quantitative and qualitative assessment of usability.
Revised NOTECHS, Moorthy *et al*[125 139]	LR, Delphi, SME plus	Internal consistency, inter-rater reliability	Learner characteristics (experience level, professional group), separate measures (technical skills, communication frequency, NTS self-assessment)	Training not described. No formal assessment of usability.
NOTSS, Yule *et al*[108 115 130]	LR, Delphi, SME plus, TA	Internal consistency, inter-rater reliability	Learner characteristics	Training course designed with handbook and online resources. Quantitative and qualitative assessment of usability.
Oxford NOTECHS, Mishra *et al*[22]	LR, Delphi, SME plus, TA	Internal consistency, inter-rater reliability, intrarater reliability	Learner characteristics (professional group), separate measures (error count, comparison with OTAS, WHO time out performance, safety attitudes questionnaire)	Training clearly described. Qualitative assessment of usability.
Non-technical skills for anaesthetic trainees, Gale *et al*[20 106]	Delphi, SME	Internal consistency, inter-rater reliability	Learner characteristics. Separate measures (prediction of performance at future point).	Two-day training for selection processes including use of assessment tool. No formal assessment of usability.
PCST, Schraagen *et al*[126 142]	LR, Delphi, SME plus, TA	Inter-rater reliability	Learner characteristics. Separate measures (surgical team assessment record; non-routine event count).	Training clearly described. Qualitative assessment of usability.
SPLINTS, Mitchell *et al*[110 143]	LR, Delphi, SME plus	Internal consistency, inter-rater reliability	Learner characteristics	Training course designed, handbook available. Quantitative and qualitative assessment of usability.
NANTSdk, Lyk-Jensen[114 144]	Delphi, SME plus	Internal consistency, inter-rater reliability, intrarater reliability	Learner characteristics. Separate measures (expert reference ratings).	Training clearly described. Quantitative and qualitative assessment of usability.
OSANTS, Dedy *et al*[28]	LR, Delphi, SME	Internal consistency, inter-rater reliability	Learner characteristics. Separate measures (compared with NOTSS).	Training clearly described. No formal assessment of usability.
WHOBARS, Devcich *et al*[145]	Delphi, SME plus	Internal consistency, inter-rater reliability, intrarater reliability	Learner characteristics	Training clearly described. Quantitative assessment of usability.
ANTSdk, Jepsen *et al*[113 146]	LR, Delphi, SME plus	Internal consistency, inter-rater reliability	Learner characteristics. Separate measures (technical skills).	Training course designed, handbook available. Quantitative and qualitative assessment of usability.
ANTS-AP, Rutherford *et al*[42]	LR, Delphi, SME plus, TA	Internal consistency, inter-rater reliability, test–retest reliability	Learner characteristics	Training course designed, handbook available. Quantitative and qualitative assessment of usability.
ICARS, Raison *et al*[30]	LR, Delphi, SME, TA	Internal consistency, inter-rater reliability	Learner characteristics (experience level). Separate measures (compared with NOTSS).	Designed for use with limited training. Qualitative and quantitative analysis of usability.
Adult intensive/ emergency care				
Team Dimensions Rating Form, Morey *et al*[58]	LR, Delphi, SME plus	Internal consistency, inter-rater reliability	Learner characteristics (trained and untrained), separate measures (subjective workload, error rate, staff attitudes, patient satisfaction)	Training clearly described. No formal assessment of usability.
Ottawa GRS, Kim *et al*[24 148 149]	LR, Delphi, SME plus	Internal consistency, inter-rater reliability, intrarater reliability	Learner characteristics (experience level), separate measures (checklist)	Training clearly described. Quantitative and qualitative assessment of usability.
MHPTS, Malec *et al*[25]	LR, Delphi, SME plus	Internal consistency, inter-rater reliability	Learner characteristics (pretraining and post-training)	Designed for use with limited training. No formal assessment of usability.
TEAM, Cooper *et al*[19 131 150 151]	LR Delphi, SME plus, TA	Internal consistency, inter-rater reliability, intrarater reliability	Learner characteristics (experience level). Separate measures (knowledge, team performance checklist).	Designed for use with limited training, online resources available. Quantitative and qualitative assessment of usability.
OSCAR, Walker *et al*[153]	LR, Delphi, SME plus	Internal consistency, inter-rater reliability	Learner characteristics (professional subgroup)	Designed for use with limited training. No formal assessment of usability.

Continued

Table 3 Continued

NTS tool name (year of publication)	Content	Internal structure/ reliability	Relations with other variables	Training required to use tool and assessment of tool usability
Trauma NOTECHS, Steinemann *et al*[154]	LR, Delphi, SME	Internal consistency, inter-rater reliability	Learner characteristics (pretraining and post-training). Separate measures (task completion, timed tasks, self-assessment score).	Training clearly described. No formal assessment of usability.
Adult inpatient				
TFAT, Sutton *et al*[155 156]	LR, Delphi, SME plus, TA	Inter-rater reliability	Learner characteristics. Separate measures (self-assessment scores for job characteristic variables).	Training course designed. Qualitative assessment of usability.
T-MEX, Olupeliyawa *et al*[111]	LR, Delphi, SME	Internal consistency, inter-rater reliability	Learner characteristics. Separate measures (self-assessment score). Consequences (educational impact).	Training clearly described, with accompanying materials. Quantitative and qualitative assessment of usability.
SWAT, Ahmed *et al*[157]	LR, Delphi, SME plus, TA	Inter-rater reliability, test–retest reliability	Learner characteristics (experience level)	Training not described. Quantitative assessment of usability.
Paediatric intensive/ emergency care				
UTBMNR, Thomas *et al*[11 158 159]	LR, Delphi, SME plus, TA	Internal consistency, inter-rater reliability	Learner characteristics (trained and untrained teams), separate measures (resuscitation guidelines)	Training clearly described. No formal assessment of usability.
Obstetrics				
AOTP and GAOTP, Tregunno *et al*[29]	LR, Delphi, SME plus, TA	Internal consistency, inter-rater reliability	Learner characteristics (pretraining and post-training). Separate measures (prediction of performance at future point).	Training described. Quantitative and qualitative assessment of usability.
PETRA, Balki *et al*[109 161]	LR, Delphi, SME	Internal consistency, inter-rater reliability	Learner characteristics. Separate measures (team performance score).	Designed for use with limited training. Qualitative and quantitative assessment of usability.
Pre-hospital care				
AeroNOTS, Myers *et al*[162]	LR, Delphi, SME, TA	Inter-rater reliability	Learner characteristics (level of experience). Separate measures (general performance rating, NTS self-assessment score).	Training briefly described. Qualitative assessment of usability.
Generic healthcare environment				
CTS, Guise *et al*[117]	SME plus	Inter-rater reliability	Learner characteristics	Designed for use with limited training. Quantitative assessment of usability.
Undergraduate				
SAFE-TeamS, Wright *et al*[99 163]	LR, Delphi, SME plus	Internal consistency, inter-rater reliability	Learner characteristics (pretraining and post-training, professional group). Separate measures (self-assessment score).	Training clearly described. Quantitative and qualitative assessment of usability.
iTOFT, Thistlethwaite *et al*[112]	LR, Delphi, SME	Internal consistency	Learner characteristics	Training: self-directed use of extensive resource pack. Qualitative and quantitative assessment of usability.

Learner characteristics denote the ability to discriminate between a good and a poor performance (where tools considered additional characteristics these are included in parenthesis). Tool acronyms are defined in table 1 and online supplementary appendix 2.
ACRM, Anaesthesia Crisis (Crew) Resource Management; ANTS, Anaesthetists' Non-Technical Skills; ANTS-AP, Anaesthetic NTS-Anaesthetic Practitioners; ANTSdk, Anaesthetists' NTS-Denmark; AOTP, Assessment of Obstetrical Team Performance; AeroNOTS, Aero-NOnTechnical Skills; BAR, Behaviourally Anchored Rating ; CTS, Clinical Teamwork Scale; Delphi, evidence of iterative process of tool development; ED, Emergency Department; FoNTS, Foundation Non-Technical Skills; GAOTP, Global AOTP; ICARS, Interpersonal and Cognitive Assessment for Robotic Surgery; LR, literature review; MHPTS, Mayo High Performance Teamwork Scale; NANTSdk, Nurse Anaesthetists' Non-Technical Skills-Denmark; NOTSS, Non-Technical Skills for Surgeons; NTS, non-technical skills; OSANTS, Objective Structured Assessment of Non-Technical Skills; OSCAR, Observational Skill-based Clinical Assessment tool for Resuscitation; OTAS, Observational Teamwork Assessment for Surgery; Ottawa GRS, Ottawa CRM Global Rating Scale; Oxford NOTECHS, Oxford NOnTECHnical Skills; PCST, Paediatric Cardiac Surgery Teamwork classification tool; PETRA, Perinatal Emergency Team Response Assessment; Revised NOTECHS, Revised Non-Technical Skills; SAFE-TeamS, Standardised Assessment For the Evaluation of Team Skills; SME, subject matter experts involved; SME plus, clinically relevant subject matter experts plus additional input from psychologists or human factor experts; SPLINTS, Scrub Practitioners List of NTS; SWAT, Surgical Ward-round Assessment Tool; TA, formal task analysis undertaken with interviews±observations; TEAM, Team Emergency Assessment Measure; TFAT, Team Functioning Assessment Tool; T-MEX, Teamwork Mini-clinical Evaluation Exercise; UTBMNR, University of Texas Behavioural Markers for Neonatal Resuscitation; WHOBARS, WHO Behaviourally Anchored Rating Scale; iTOFT, Individual Teamwork Observation and Feedback Tool.

and a robust process of revalidation which has clear cost implications in practice.

Choosing an NTS assessment tool

This review has revealed the multiplicity of NTS assessment tools available in healthcare, highlighting clear challenges for the educator in healthcare in trying to choose which is most appropriate for their training purposes. The process of categorising the tools in this review highlighted three initial decisions to be made:

- Is the training for a multidisciplinary team or for a single group, for example, medical students?
- Is the training in a real or simulated environment?

Box 1 What this study adds

- There are 76 published tools for the measurement of non-technical skills (NTS) in healthcare across seven clinical areas with widely differing methods of scoring for either individuals or teams.
- The methods of development and rigour of assessments of validity vary widely among these tools.
- Recommendations for training also vary greatly and pragmatic assessment of usability is scarce.
- A standardised approach to the development and testing of tools for the measurement of NTS would assist both educators and researchers.
- There is currently no pre-eminent tool for the measurement of NTS which we can recommend.

- What is the setting for the training, for example, ward based, critical care or obstetrics?

Table 2 has been configured to highlight these key features with the aim of providing a means of selecting a tool for a particular setting. It is hoped that the additional information provided in table 3, where practical issues such as training required to use the tool are described, will further support the selection process for educators in healthcare.

Study limitations

The authors recognise the difficulty of excluding bias and that using the techniques described above can mitigate but not remove it.

Some of the variability described in this review can be ascribed to the following issues:

- Tools which were published in the early days of NTS research in healthcare were often based on tools from aviation and provided less evidence of validity due to lack of available reference points.
- Tools only recently published may not have had time to undertake rigorous reliability testing.
- Tools based on those developed earlier (eg, for use in a different language/culture) did not describe method of design as they relied on data from the original work.

This study was designed to provide an objective analysis of the observer-based tools for assessment of NTS in healthcare, including evidence of validity and an assessment of ease of use. The analysis of attributes allowed for some discrimination between tools but the variability described throughout the review precluded meaningful analysis of, for example, quality of method of design or how long it took before a tool could be used reliably. This is an area deserving further analysis.

Although we contacted authors via email to ask for further information it is possible that we do not have a complete data set for each tool.

We restricted the study to considering only papers that were contiguous with the original development of the tool and did not include data from groups who had used the tools in different settings.

CONCLUSION

This review has shown that there is variability in the method of design and testing of tools for the assessment of NTS and that consideration of these features is not always complete. Recommendations for designing and training to use tools for the assessment of NTS made by Klampfer *et al*[124] and Hull *et al*[128] may be regarded as the gold standard but acceptability and cost implications remain a considerable barrier. Similarities between systems have also been highlighted[49 129]—strengthening support for a more unified approach to NTS teaching and a rationalisation of assessment tools.

Finally, previous reviews of NTS tools have provided an overview of available assessment techniques in different areas but have not provided a means of discriminating between them.[130–134]

We have devised a system for categorising tools for the assessment of NTS which could be useful to both novice and expert educators in simulation-based education.

The ideal tool for NTS assessment in healthcare does not yet exist. Further research is required to determine if a more generic tool for use in any healthcare context with the appropriate subject matter expertise to guide assessment of validity and reliability, task analysis and deployment is feasible and brings us closer to that goal.

Acknowledgements Grateful thanks go to Neal Thurley at Bodleian libraries for his generous assistance in structuring and refining the search criteria for this review. The authors do not seek to criticise the extensive research and effort which has gone into designing these tools, merely to highlight that we may now need to consider how best to rationalise and standardise future work in this field. We are also very grateful to the researchers who took the time to reply to our requests for additional information about the tools they had designed, particularly those who spoke to HH in person. The information they provided assisted in the analysis of their own tools and added richness to the discussion section of the paper. Finally, I dedicate this article to Dr Denis O'Leary whose insights and gentle wisdom added enormously to this study and so much more.

Contributors HH was responsible for the study concept and drafting the initial review manuscript. HH, PRG, JR, LV, DY and CV were responsible for design, analysis, interpretation and preparation of the completed manuscript.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data relevant to the study are included in the article or uploaded as supplementary information.

REFERENCES

1 Brennan TA, Leape LL, Laird NM, *et al*. Incidence of adverse events and negligence in hospitalized patients. Results

of the Harvard Medical Practice Study I. *N Eng J Med* 1991;324:370–6.
2 Vincent C, Neale G, Woloshynowych M. Adverse events in British hospitals: preliminary retrospective record review. *BMJ* 2001;322:517–9.
3 Wilson RM, Runciman WB, Gibberd RW, *et al*. The Quality in Australian health Care study. *Med J Aust* 1995;163:458–71.
4 Makary M, Daniel M. Medical error—the third leading cause of death in the US. *BMJ* 2016;353.
5 Michel P, Brami J, Chanelière M, *et al*. Patient safety incidents are common in primary care: a national prospective active incident reporting survey. *PLoS One* 2017;12:e0165455.
6 Flin R, Martin L, Goeters K-M, *et al*. Development of the NOTECHS (non-technical skills) system for assessing pilots' CRM skills. *Hum Factors Aerosp Saf* 2003;3:95–117.
7 Nestel D, Walker K, Simon R, *et al*. Nontechnical skills: an inaccurate and unhelpful descriptor? *Simul Healthc J Soc Simul Healthc* 2011;6:2–3.
8 Gaba DM. Training and nontechnical skills: the politics of terminology. *Simul Healthc J Soc Simul Healthc* 2011;6:8–10.
9 Gaba DM, Howard SK, Flanagan B, *et al*. Assessment of clinical performance during simulated crises using both technical and behavioral ratings. *Anesthesiology* 1998;89:8–18.
10 Fletcher G, Flin R, McGeorge P, *et al*. Anaesthetists' Non-Technical Skills (ANTS): evaluation of a behavioural marker system. *Br J Anaesth* 2003;90:580–8.
11 Thomas EJ, Sexton JB, Helmreich RL. Translating teamwork behaviours from aviation to healthcare: development of behavioural markers for neonatal resuscitation. *Qual Saf Heal Care* 2004;13:57–64.
12 Catchpole KR, Giddings AEB, Hirst G, *et al*. A method for measuring threats and errors in surgery. *Cogn Technol Work* 2008;10:295–304.
13 Flin R, Martin L. Behavioral markers for Crew Resource management: a review of current practice. *Int J Aviat Psychol* 2001;11:95–118.
14 Weaver SJ, Lyons R, DiazGranados D, *et al*. The anatomy of health Care team training and the state of practice: a critical review. *Acad Med* 2010;85:1746–60.
15 Salas E, Rosen AM. Building high reliability teams: progress and some reflections on teamwork training. *BMJ Qual Saf* 2013;22:369–73.
16 Grand AJ, Pearce M, Rench AT, *et al*. Going deep: guidelines for building simulation-based team assessments. *BMJ Qual Saf* 2013;22:436–48.
17 Cook DA, Brydges R, Zendejas B, *et al*. Technology-Enhanced simulation to assess health professionals: a systematic review of validity evidence, research methods, and reporting quality. *Acad Med* 2013;88:872–83.
18 Hawker S, Payne S, Kerr C, *et al*. Appraising the evidence: reviewing disparate data systematically. *Qual Health Res* 2002;12:1284–99.
19 Cooper S, Cant R, Porter J, *et al*. Rating medical emergency teamwork performance: development of the team emergency assessment measure (team). *Resuscitation* 2010;81:446–52.
20 Gale TCE, Roberts MJ, Sice PJ, *et al*. Predictive validity of a selection centre testing non-technical skills for recruitment to training in anaesthesia. *Br J Anaesth* 2010;105:603–9.
21 Undre S, Healey AN, Darzi A, *et al*. Observational assessment of surgical teamwork: a feasibility study. *World J Surg* 2006;30:1774–83.
22 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 Heal Care* 2009;18:104–8.
23 Yule S, Flin R, Paterson-Brown S, *et al*. Development of a rating system for surgeons' non-technical skills. *Med Educ* 2006;40:1098–104.
24 Kim J, Neilipovitz D, Cardinal P, *et al*. A pilot study using high-fidelity simulation to formally evaluate performance in the resuscitation of critically ill patients: the University of Ottawa critical care medicine, high-fidelity simulation, and crisis resource management I study. *Crit Care Med* 2006;34:2167–74.
25 Malec JF, Torsher LC, Dunn WF, *et al*. The Mayo high performance teamwork scale: reliability and validity for evaluating key Crew Resource management skills. *Simul Healthc* 2007;2:4–10.
26 Weller J, Frengley R, Torrie J, *et al*. Evaluation of an instrument to measure teamwork in multidisciplinary critical care teams. *BMJ Qual Saf* 2011;20:216–22.
27 Blackhall LJ, Erickson J, Brashers V, *et al*. Development and validation of a collaborative behaviors objective assessment tool for end-of-life communication. *J Palliat Med* 2014;17:68–74.
28 Dedy NJ, Szasz P, Louridas M, *et al*. Objective structured assessment of nontechnical skills: reliability of a global rating scale for the in-training assessment in the operating room. *Surg* 2015;157:1002–13.
29 Tregunno D, Pittini R, Haley M, *et al*. Development and usability of a behavioural marking system for performance assessment of obstetrical teams. *Qual Saf Heal Care* 2009;18:393–6.
30 Raison N, Wood T, Brunckhorst O, *et al*. Development and validation of a tool for non-technical skills evaluation in robotic surgery-the ICARS system. *Surg Endosc* 2017;31:5403–10.
31 American Educational Research Association, American Psychological Association, National Council on Measurement in Education. *Standards for educational and psychological testing*. American Educational Research Association, 1999.
32 Bossuyt PM, Reitsma JB, Bruns DE, *et al*. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *BMJ* 2003;326:41–4.
33 Kottner J, Audigé L, Brorson S, *et al*. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *J Clin Epidemiol* 2011;64:96–106.
34 van der Vleuten CPM, Schuwirth LWT. Assessing professional competence: from methods to programmes. *Med Educ* 2005;39:309–17.
35 Messick S. Validity. In: Linn R, ed. *Educational measurement*. Macmillan, 1988.
36 Mokkink LB, Terwee CB, Patrick DL, *et al*. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol* 2010;63:737–45.
37 Masys AJ, Olafsen R, Brathen K, *et al*. Situation awareness. *Hum performance Situat Aware Autom Curr Res trends* 2004;1:39–171.
38 Cook DA, Zendejas B, Hamstra SJ, *et al*. What counts as validity evidence? examples and prevalence in a systematic review of simulation-based assessment. *Adv Heal Sci Educ* 2014;19:233–50.

39 Cook DA, Hatala R. Validation of educational assessments: a primer for simulation and beyond. *Adv Simul* 2016;1.
40 Salas E, Almeida SA, Salisbury M, *et al*. What are the critical success factors for team training in health care? *Jt Comm J Qual Patient Saf* 2009;35:398–405.
41 Ritchie J, Lewis J, McNaughton Nicholls C, *et al*. *Qualitative research practice : a guide for social science students and researchers*. London: Sage Publications Inc, 2014.
42 Rutherford JS, Flin R, Irwin A, *et al*. Evaluation of the prototype anaesthetic non-technical skills for anaesthetic practitioners (ANTS-AP) system: a behavioural rating system to assess the non-technical skills used by staff assisting the anaesthetist. *Anaesthesia* 2015;70:907–14.
43 Sevdalis N, Davis R, Moorthy K, *et al*. Reliability of a revised NOTECHS scale for use in surgical teams. *Am J Surg* 2008;196:184–90.
44 Undre S, Healey AN, Darzi A, *et al*. Observational assessment of surgical teamwork: a feasibility study. *World J Surg* 2006;30:1774–83.
45 Flowerdew L, Brown R, Vincent C, *et al*. Development and validation of a tool to assess emergency physicians' nontechnical skills. *Ann Emerg Med* 2012;59:376–85.
46 Amaya Arias AC, Barajas R, Eslava-Schmalbach JH, *et al*. Translation, cultural adaptation and content re-validation of the observational teamwork assessment for surgery tool. *Int J Surg* 2014;12:1390–402.
47 Amiel I, Simon D, Merin O, *et al*. Mobile in situ simulation as a tool for evaluation and improvement of trauma treatment in the emergency department. *J Surg Educ* 2016;73:121–8.
48 Maignan M, Koch F-XF-X, Chaix J, *et al*. Team emergency assessment measure (team) for the assessment of non-technical skills during resuscitation: validation of the French version. *Resuscitation* 2016;101:115–20.
49 Passauer-Baierl S, Hull L, Miskovic D, *et al*. Re-validating the observational teamwork assessment for surgery tool (OTAS-D): cultural adaptation, refinement, and psychometric evaluation. *World J Surg* 2014;38:305–13.
50 Spanager L, Lyk-Jensen HT, Dieckmann P, *et al*. Customization of a tool to assess Danish surgeons' non-technical skills in the operating room. *Dan Med J* 2012;59:1–6.
51 Munroe B, Curtis K, Murphy M, *et al*. A structured framework improves clinical patient assessment and nontechnical skills of early career emergency nurses: a pre-post study using full immersion simulation. *J Clin Nurs* 2016;25:2262–74.
52 Davis DA, Mazmanian PE, Fordis M, *et al*. Accuracy of physician self-assessment compared with observed measures of competence. *JAMA* 2006;296:1094–102.
53 Byrne AJ, Greaves JD. Assessment instruments used during anaesthetic simulation: review of published studies. *BJA Br J Anaesth* 2001;86:445–50.
54 Kardong-Edgren S, Adamson KA, Fitzgerald C. A review of currently published evaluation instruments for human patient simulation. *Clin Simul Nurs* 2010;6:e25–35.
55 Michinov E, Jamet E, Dodeler V, *et al*. Assessing neurosurgical non-technical skills: an exploratory study of a new behavioural marker system. *J Eval Clin Pract* 2014;20:582–8.
56 Kemper PF, Noord van I, de Bruijne M, *et al*. Development and reliability of the explicit professional oral communication observation tool to quantify the use of non-technical skills in healthcare. *BMJ Qual Saf* 2013;22:586–95.
57 Brashers V, Erickson JM, Blackhall L, *et al*. Measuring the impact of clinically relevant interprofessional education on undergraduate medical and nursing student competencies: a longitudinal mixed methods approach. *J Interprof Care* 2016;30:448–57.
58 Morey JC, Simon R, Jay GD, *et al*. Error reduction and performance improvement in the emergency department through formal teamwork training: evaluation results of the MedTeams project. *Health Serv Res* 2002;37:1553–81.
59 Anderson JM, Murphy AA, Boyle KB, *et al*. Simulating extracorporeal membrane oxygenation emergencies to improve human performance. Part II: assessment of technical and behavioral skills. *Simul Healthc* 2006;1:228–32.
60 Parker-Raley J, Cerroni A, Mottet TP, *et al*. Investigating pediatric trauma team communication effectiveness phase two: achieving inter-rater reliability for the assessment of pediatric resuscitation communication team assessment. *J Commun Healthc* 2013;6:145–57.
61 Huang LC, Conley D, Lipsitz S, *et al*. The surgical safety checklist and teamwork coaching tools: a study of inter-rater reliability. *BMJ Qual Saf* 2014;23:639–50.
62 Hull L, Bicknell C, Patel K, *et al*. Content validation and evaluation of an endovascular teamwork assessment tool. *Eur J Vasc Endovasc Surg* 2016;52:11–20.
63 Flynn FM, Sandaker K, Ballangrud R. Aiming for excellence – a simulation-based study on adapting and testing an instrument for developing non-technical skills in Norwegian student nurse anaesthetists. *Nurse Educ Pract* 2017;22:37–46.
64 Weller JM, Bloch M, Young S, *et al*. Evaluation of high fidelity patient simulator in assessment of performance of anaesthetists. *Br J Anaesth* 2003;90:43–7.
65 Weller JM, Robinson BJ, Jolly B, *et al*. Psychometric characteristics of simulation-based assessment in anaesthesia and accuracy of self-assessed scores. *Anaesthesia* 2005;60:245–50.
66 Arain NA, Hogg DC, Gala RB, *et al*. Construct and face validity of the American College of Surgeons/Association of program directors in surgery laparoscopic troubleshooting team training exercise. *Am J Surg* 2012;203:54–62.
67 Spanager L, Lyk-Jensen HT, Dieckmann P, *et al*. Customization of a tool to assess Danish surgeons' non-technical skills in the operating room. *Dan Med J* 2012;59.
68 Spanager L, Beier-Holgersen R, Dieckmann P, *et al*. Reliable assessment of general surgeons' non-technical skills based on video-recordings of patient simulated scenarios. *Am J Surg* 2013;206:810–7.
69 Spanager L, Konge L, Dieckmann P, *et al*. Assessing trainee surgeons' nontechnical skills: five cases are sufficient for reliable assessments. *J Surg Educ* 2015;72:16–22.
70 Hamilton N, Freeman B, Woodhouse J, *et al*. Team behavior during trauma resuscitation: a simulation-based performance assessment. *J Grad Med Educ* 2009:253–9.
71 Hamilton NA, Kieninger AN, Woodhouse J, *et al*. Video review using a reliable evaluation metric improves team function in high-fidelity simulated trauma resuscitation. *J Surg Educ* 2012;69:428–31.
72 Andersen PO, Jensen MK, Lippert A, *et al*. Development of a formative assessment tool for measurement of performance in multi-professional resuscitation teams. *Resuscitation* 2010;81:703–11.
73 Pascual JL, Holena DN, Vella MA, *et al*. Short simulation training improves objective skills in established advanced practitioners managing emergencies on the ward and

surgical intensive care unit. *J Trauma Inj Infect Crit Care* 2011;71:330–8.
74 RW F, JM W, Torrie J, *et al*. The effect of a simulation-based training intervention on the performance of established critical care unit teams. *Crit Care Med* 2011;39:2605–11.
75 Weller J, Shulruf B, Torrie J, *et al*. Validation of a measurement tool for self-assessment of teamwork in intensive care. *Br J Anaesth* 2013;111:460–7.
76 Flowerdew L, Gaunt A, Spedding J, *et al*. A multicentre observational study to evaluate a new tool to assess emergency physicians' non-technical skills. *Emerg Med J* 2013;30:437–43.
77 DeMoor S, Abdel-Rehim S, Olmsted R, *et al*. Evaluating trauma team performance in a level I trauma center. *J Trauma Acute Care Surg* 2017;83:159–64.
78 Holcomb JB, Dumire RD, Crommett JW, *et al*. Evaluation of trauma team performance using an advanced human patient simulator for resuscitation training. *J Trauma* 2002;52:1078–85.
79 Ottestad E, Boulet JR, Lighthall GK. Evaluating the management of septic shock using patient simulation. *Crit Care Med* 2007;35:769–75.
80 Fernandez R, Kozlowski SWJ, Shapiro MJ, *et al*. Toward a definition of teamwork in emergency medicine. *Acad Emerg Med* 2008;15:1104–12.
81 Fernandez R, Pearce M, Grand JA, *et al*. Evaluation of a computer-based educational intervention to improve medical teamwork and performance during simulated patient resuscitations. *Crit Care Med* 2013;41:2551–62.
82 O'Leary KJ, Boudreau YN, Creden AJ, *et al*. Assessment of teamwork during structured interdisciplinary rounds on medical units. *J Hosp Med* 2012;7:679–83.
83 Mellanby E, Hume M, Glavin R, *et al*. Project report for: the development of a behavioural marker system for newly qualified doctors in managing acutely unwell patients. Edinburgh, 2013. Available: http://www.docs.hss.ed.ac.uk/iad/Learning_teaching/Academic_teaching/PTAS/Outputs/Mellanby_Jan2012%20award_PTAS_Final_Report.pdf [Accessed 15 Apr 2019].
84 Hull L, Birnbach D, Arora S, *et al*. Improving surgical ward care: development and psychometric properties of a global assessment toolkit. *Ann Surg* 2014;259:904–9.
85 Pugh D, Hamstra SJ, Wood TJ, *et al*. A procedural skills OSCE: assessing technical and non-technical skills of internal medicine residents. *Adv Heal Sci Educ* 2014;20:85–100.
86 Grant EC, Grant VJ, Bhanji F, *et al*. The development and assessment of an evaluation tool for pediatric resident competence in leading simulated pediatric resuscitations. *Resuscitation* 2012;83:887–93.
87 Parker-Raley J, Mottet TP, Lawson KA, *et al*. Investigating pediatric trauma team communication effectiveness phase one: the development of the assessment of pediatric resuscitation communication investigating pediatric trauma team communication effectiveness phase one: the development of the assessm. *J Commun Healthc* 2012;5:102–15.
88 Parker-Raley J, Yanez K, Cerroni A, *et al*. Assessing trauma leader communication in an ED setting. *J Commun Healthc* 2013;6:197–207.
89 Reid J, Stone K, Brown J, *et al*. The simulation team assessment tool (STAT): development, reliability and validation. *Resuscitation* 2012;83:879–86.
90 Lambden S, DeMunter C, Dowson A, *et al*. The imperial paediatric emergency training toolkit (IPETT) for use in paediatric emergency training: development and evaluation of feasibility and validity. *Resuscitation* 2013;84:831–6.
91 Daniels K, Lipman S, Harney K, *et al*. Use of simulation based team training for obstetric crises in resident education. *Simul Healthc J Soc Simul Healthc* 2008;3:154–60.
92 Bracco F, Masini M, De Tonetti G, *et al*. Adaptation of non-technical skills behavioural markers for delivery room simulation. *BMC Pregnancy Childbirth* 2017;17:1–7.
93 Von Wyl T, Zuercher M, Amsler F, *et al*. Technical and non-technical skills can be reliably assessed during paramedic simulation training. *Acta Anaesthesiol Scand* 2008;53:121–7.
94 Holly D, Swanson V, Cachia P, *et al*. Development of a behaviour rating system for rural/remote pre-hospital settings. *Appl. Ergon* 2017;58:405–13.
95 Frankel A, Gardner R, Maynard L, *et al*. Using the communication and teamwork skills (cats) assessment to measure health Care team performance. *Jt Comm J Qual Patient Saf* 2007;33:549–58.
96 King HB, Battles J, Baker DP, *et al*. TeamSTEPPSTM: team strategies and tools to enhance performance and patient safety. agency for healthcare research and quality (US), 2008. Available: http://www.ncbi.nlm.nih.gov/pubmed/21249942 [Accessed 15 Apr 2019].
97 Zhang C, Miller C, Volkman K, *et al*. Evaluation of the team performance observation tool with targeted behavioral markers in simulation-based interprofessional education. *J Interprof Care* 2015;29:202–8.
98 Carlson J, Min E, Bridges D. The impact of leadership and team behavior on standard of care delivered during human patient simulation: a pilot study for undergraduate medical students. *Teach Learn Med* 2009;21:24–32.
99 Wright MC, Phillips-Bute BG, Petrusa ER, *et al*. Assessing teamwork in medical education and practice: relating behavioural teamwork ratings and clinical performance. *Med Teach* 2009;31:30–8.
100 Sigalet E, Donnon T, Cheng A, *et al*. Development of a team performance scale to assess undergraduate health professionals. *Acad Med* 2013;88:989–96.
101 Paige JT, Garbee DD, Kozmenko V, *et al*. Getting a head start: high-fidelity, simulation-based operating room team training of interprofessional students. *J Am Coll Surg* 2014;218:140–9.
102 Emmert MC, Cai L. A pilot study to test the effectiveness of an innovative interprofessional education assessment strategy. *J Interprof Care* 2015;29:451–6.
103 Kane MT. An argument-based approach to validity. *Psychol Bull* 1992;112:527–35.
104 Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med* 2006;119:166.e7–16.
105 Weller JM, Jolly B, Robinson B. Generalisability of behavioural skills in simulated anaesthetic emergencies. *Anaesth Intensive Care* 2008;36:185–9.
106 Gale TCE, Roberts MJ, Sice PJ, *et al*. Predictive validity of a selection centre testing non-technical skills for recruitment to training in anaesthesia. *Br J Anaesth* 2010;105:603–9.
107 Morgan PJ, Tregunno D, Pittini R, *et al*. Determination of the psychometric properties of a behavioural marking system for obstetrical team training using high-fidelity simulation: table 1. *BMJ Qual Saf* 2012;21:78–82.

108 Yule S, Flin R, Maran N, *et al*. Debriefing surgeons on non-technical skills (NOTSS). *Cogn Technol Work* 2008;10:265–74.
109 Balki M, Hoppe D, Monks D, *et al*. The PETRA (perinatal emergency team response assessment) scale: a high-fidelity simulation validation study. *J Obstet Gynaecol Can* 2017;39:523–33.
110 Mitchell L, Flin R, Yule S, *et al*. Evaluation of the scrub practitioners' list of intraoperative non-technical skills system. *Int J Nurs Stud* 2012;49:201–11.
111 Olupeliyawa AM, O'Sullivan AJ, Hughes C, *et al*. The teamwork Mini-Clinical evaluation exercise (T-MEX): a workplace-based assessment focusing on collaborative competencies in health care. *Acad Med* 2014;89:359–65.
112 Thistlethwaite J, Dallest K, Moran M, *et al*. Introducing the individual teamwork observation and feedback tool (iTOFT): development and description of a new interprofessional teamwork measure. *J Interprof Care* 2016;30:526–8.
113 Jepsen RMHG, Dieckmann P, Spanager L, *et al*. Evaluating structured assessment of anaesthesiologists' non-technical skills. *Acta Anaesthesiol Scand* 2016;60:756–66.
114 Lyk-Jensen HT, Dieckmann P, Konge L, *et al*. Using a structured assessment tool to evaluate Nontechnical skills of nurse Anesthetists. *Aana J* 2016;84:122–7.
115 Yule S, Flin R, Maran N, *et al*. Surgeons' Non-technical Skills in the Operating Room: Reliability Testing of the NOTSS Behavior Rating System. *World J Surg* 2008;32:548–56.
116 Russ S, Hull L, Rout S, *et al*. Observational teamwork assessment for surgery: feasibility of clinical and nonclinical assessor calibration with short-term training. *Ann Surg* 2012;255:804–9.
117 Guise J-M, Deering SH, Kanki BG, *et al*. Validation of a tool to measure and promote clinical teamwork. *Simul Healthc* 2008;3:217–23.
118 Rosen MA, Salas E, Wilson KA, *et al*. Measuring team performance in simulation-based training: adopting best practices for healthcare. *Simulation in Healthcare: The Journal of the Society for Simulation in Healthcare* 2008;3:33–41.
119 Weaver SJ, Dy SM, Rosen MA. Team-training in healthcare: a narrative synthesis of the literature. *BMJ Qual Saf* 2014;23:359–72.
120 Flin R, O'Connor P, Crichton M. Safety at the sharp end: a guide to non-technical skills; Chapter 2. in: safety at the sharp end. *A Guide to Non-Technical Skills* 2008:17–40.
121 Mitchell L, Flin R, Yule S, *et al*. Evaluation of the Scrub Practitioners' List of Intraoperative Non-Technical Skills (SPLINTS) system. *International Journal of Nursing Studies* 2012;49:201–11.
122 Anderson A, Baxendale B, Scott L, *et al*. The National simulation development project: summary report.. Available: http://aspih.org.uk/wp-content/uploads/2017/07/national-scoping-project-summary-report.pdf [Accessed 15 Apr 2019].
123 Smith-Jentsch KA, Johnston JH, Payne SC. Measuring team-related expertise in complex environments. In: Cannon-Bowers JA, Salas E, eds. *Making decisions under stress: implications for individual and team training*. Washington: American Psychological Association, 1998: 61–87.
124 Klampfer B, Flin R, Helmreich RL, *et al*. Enhancing performance in high risk environments: recommendations for the use of behavioural markers. Zurich, 2001. Available: http://www.raes-hfg.com/reports/notechs-swiss.pdf [Accessed 15 Apr 2019].
125 Moorthy K, Munz Y, Adams S, *et al*. A human factors analysis of technical and team skills among surgical trainees during procedural simulations in a simulated operating theatre. *Ann Surg* 2005;242:631–9.
126 Schraagen JM, Schouten T, Smit M, *et al*. Assessing and improving teamwork in cardiac surgery. *Qual Saf Health Care* 2010;19:e29.
127 Sevdalis N, Lyons M, Healey AN, *et al*. Observational teamwork assessment for surgery: construct validation with expert versus novice raters. *Ann Surg* 2009;249:1047–51.
128 Hull L, Arora S, Symons NRA, *et al*. Training faculty in Nontechnical skill assessment national guidelines on program requirements. *Ann Surg* 2013;258:370–5.
129 Wisborg T, Manser T. Assessment of non-technical skills in the operating room - one assessment tool per specialty? *Acta Anaesthesiol Scand* 2014;58:773–4.
130 Yule S, Flin R, Paterson-Brown S, *et al*. Non-technical skills for surgeons in the operating room: a review of the literature. *Surgery* 2006;139:140–9.
131 Cooper S, Porter J, Peach L. Measuring situation awareness in emergency settings: a systematic review of tools and outcomes. *OAEM* 2013;6:1–7.
132 Shields A, Flin R. Paramedics' non-technical skills: a literature review. *Emerg Med J* 2013;30:350–4.
133 Watanabe Y, Bilgic E, Lebedeva E, *et al*. A systematic review of performance assessment tools for laparoscopic cholecystectomy. *Surgical Endoscopy* 2016;30:832–44.
134 Onwochei DN, Halpern S, Balki M. Teamwork assessment tools in obstetric emergencies: a systematic review. *Simul Healthc* 2017;12:165–76.
135 Fletcher G, Flin R, McGeorge P, *et al*. Rating non-technical skills: developing a behavioural marker system for use in anaesthesia. *Cogn Tech Work* 2004;6:165–71.
136 Patey R, Flin R, Fletcher G, *et al*. Developing a Taxonomy of Anesthetists' Nontechnical Skills (ANTS). *Adv Patient Saf from Res to Implement* 2005;4:325–36.
137 Healey AN, Undre S, Vincent CA. Developing observational measures of performance in surgical teams. *Qual Saf Heal Care* 2004;13(suppl_1):i33–40.
138 Undre S, Moorthy K, Sevdalis N, *et al*. Multidisciplinary crisis simulations: the way forward for training surgical teams. *World J Surg* 2007;31:1843–53.
139 Moorthy K, McCulloch P, Undre S, *et al*. Is team training in briefings for surgical teams feasible in simulation? *Cogn Technol Work* 2008;10:275–85.
140 Robertson ER, Hadi M, Morgan LJ, *et al*. Oxford NOTECHS II: a modified theatre team non-technical skills scoring system. *PLoS ONE* 2014;9:e90320.
141 Crossingham GV, Sice PJA, Roberts MJ, *et al*. Development of workplace-based assessments of non-technical skills in anaesthesia*. *Anaesthesia* 2012;67:158–64.
142 Schraagen JM, Schouten T, Smit M, *et al*. A prospective study of paediatric cardiac surgical microsystems: assessing the relationships between non-routine events, teamwork and patient outcomes. *BMJ Qual Saf* 2011;20:599–603.
143 Mitchell L, Flin R, Yule S, *et al*. Development of a behavioural marker system for scrub practitioners' non-technical skills (splints system). *J Eval Clin Pract* 2013;19:317–23.
144 Lyk-Jensen HT, Jepsen RMHG, Spanager L, *et al*. Assessing nurse anaesthetists' non-technical skills in the operating room. *Acta Anaesthesiol Scand* 2014;58:794–801.

145 Devcich DA, Weller J, Mitchell SJ, *et al*. A behaviourally anchored rating scale for evaluating the use of the WHO surgical safety checklist: development and initial evaluation of the WHOBARS. *BMJ Qual Saf* 2016;25:778–86.
146 Jepsen RMHG, Spanager L, Lyk-Jensen HT, *et al*. Customisation of an instrument to assess anaesthesiologists' non-technical skills. *Int J Med Educ* 2015;6:17–25.
147 Gjeraa K, Jepsen RMHG, Rewers M, *et al*. Exploring the relationship between anaesthesiologists' non-technical and technical skills. *Acta Anaesthesiol Scand* 2016;60:36–47.
148 Kim J, Neilipovitz D, Cardinal P, *et al*. A pilot study using high-fidelity simulation to formally evaluate performance in the resuscitation of critically ill patients: the University of Ottawa critical care medicine, high-fidelity simulation, and crisis resource management I study. *Crit Care Med* 2006;34:2167–74.
149 Kim J, Neilipovitz D, Cardinal P, *et al*. A Comparison of Global Rating Scale and Checklist Scores in the Validation of an Evaluation Tool to Assess Performance in the Resuscitation of Critically Ill Patients During Simulated Emergencies (Abbreviated as "CRM Simulator Study IB"). *Simul Healthc J Soc Simul Healthc* 2009;4:6–16.
150 Cooper SJ, Cant RP. Measuring non-technical skills of medical emergency teams: an update on the validity and reliability of the team emergency assessment measure (team). *Resuscitation* 2014;85:31–3.
151 Cooper S, Cant R, Connell C, *et al*. Measuring teamwork performance: validity testing of the team emergency assessment measure (team) with clinical resuscitation teams. *Resuscitation* 2016;101:97–101.
152 Bogossian F, Cooper S, Cant R, *et al*. Undergraduate nursing students' performance in recognising and responding to sudden patient deterioration in high psychological fidelity simulated environments: an Australian multi-centre study. *Nurse Educ Today* 2014;34:691–6.
153 Walker S, Brett S, McKay A, *et al*. Observational Skill-based clinical assessment tool for resuscitation (OSCAR): development and validation. *Resuscitation* 2011;82:835–44.
154 Steinemann S, Berg B, DiTullio A, *et al*. Assessing teamwork in the trauma bay: introduction of a modified "NOTECHS" scale for trauma. *Am J Surg* 2012;203:69–75.
155 Sutton G, Liao J, Jimmieson NL, *et al*. Measuring multidisciplinary team effectiveness in a ward-based healthcare setting: development of the team functioning assessment tool. *J Healthc Qual* 2011;33:10–24.
156 Sutton G, Liao J, Jimmieson NL, *et al*. Measuring ward-based multidisciplinary healthcare team functioning: a validation study of the team functioning assessment tool (TFAT). *J Healthc Qual* 2013;35:36–49.
157 Ahmed K, Anderson O, Jawad M, *et al*. Design and validation of the surgical ward round assessment tool: a quantitative observational study. *Am J Surg* 2015;209:682–8.
158 Thomas EJ, Sexton JB, Lasky RE, *et al*. Teamwork and quality during neonatal care in the delivery room. *J Perinatol* 2006;26:163–9.
159 Thomas EJ, Taggart B, Crandell S, *et al*. Teaching teamwork during the neonatal resuscitation program: a randomized trial. *J Perinatol* 2007;27:409–14.
160 Morgan PJ, Pittini R, Regehr G, *et al*. Evaluating teamwork in a simulated obstetric environment. *Anesthesiology* 2007;106:907–15.
161 Balki M, Hoppe D, Monks D, *et al*. Multidisciplinary Delphi development of a scale to evaluate team function in obstetric emergencies: the PETRA scale. *J Obstet Gynaecol Can* 2017;39:434–42.
162 Myers JA, Powell DMC, Psirides A, *et al*. Non-technical skills evaluation in the critical care air ambulance environment: introduction of an adapted rating instrument--an observational study. *Scand J Trauma Resusc Emerg Med* 2016;24.
163 Wright MC, Segall N, Hobbs G, *et al*. Standardized assessment for evaluation of team skills. *Simulation in Healthcare: The Journal of the Society for Simulation in Healthcare* 2013;8:292–303.