Cogn Tech Work (2015) 17:63–77
DOI 10.1007/s10111-014-0306-y

ORIGINAL ARTICLE

Development of instruments for assessment of individuals' and teams' non-technical skills in healthcare: a critical review

Rikke M. H. G. Jepsen · Doris Østergaard · Peter Dieckmann

Received: 24 February 2013 / Accepted: 20 July 2013 / Published online: 5 October 2014

Abstract Focus on patient safety has increased over the last decades; however, patient safety still relies on technology and the performance of healthcare professionals. Technology improves rapidly but despite numerous reports on how to improve the performance of healthcare professionals, this improvement has been more slow than expected. The performance of healthcare professionals is a product of each individual's medical knowledge, technical and non-technical skills as well as the settings in which these skills are used. Training and the assessment of non-technical skills for individuals or teams in healthcare have shown to improve a safe and efficient performance. However, the implementation has been slow. One reason for this might be the shortage of assessment instrument to assess individuals' and teams' non-technical skills. In this paper, we review the development process of 23 instruments for assessment of non-technical skills for individuals or teams within healthcare. The instruments are mainly for use in the operating room or for teams handling emergency situations. Several of the instruments are developed based on a thorough needs analysis by a team having the same professional background as the target group. Most of the instruments consist of almost the same categories of non-technical skills and many use behavioural markers. Overall, the instruments have been validated to some extent, but there is room for improvement. However, there seems to be a lack of training of the raters. The importance of providing feedback after the assessment is emphasised. The criteria on which the behavioural markers are developed should undergo continuous changes, representing the development of the patient safety culture in healthcare organisations.

Keywords Non-technical skills · Human factors · Assessment · Healthcare · Performance · Observation · Patient safety

1 Introduction

Patient safety relies on the interplay between humans, technology and organisations. Patient safety challenges in healthcare have been described previously, for example in anaesthesia (Cooper et al. 2002; Gaba 2000). The IOM report "To err is human" (Kohn et al. 1999) and the follow-up report "Crossing the quality chasm" (Corrigan 2005) made it apparent that even though there is increased focus on patient safety, the progress is far from impressive and that patient safety problems in healthcare seem to persist (Sevdalis et al. 2012).

Organisational changes and the use of technology have improved some safety indicators but the health professionals' medical expertise, technical abilities and non-technical skills (NTS) play a major role in improving patient safety (Sevdalis et al. 2012). Traditionally, health professionals are trained in silos and are taught medical knowledge and technical skills, whereas NTS, such as communication, cooperation and leadership, are seldom taught. Yet, these skills are expected to be attained since they are indeed needed in the ill-structured world of clinical work (Rasmussen et al. 2012). NTS are important for safe and efficient teamwork especially in emergency situations, where time is an important factor. Training of NTS in healthcare was first systematically introduced with the crisis resource management (CRM) training in order to

R. M. H. G. Jepsen (✉) · D. Østergaard · P. Dieckmann
Danish Institute for Medical Simulation, Capital Region of Denmark and University of Copenhagen, Herlev Ringvej 75, 2730 Herlev, Denmark
e-mail: rikke.malene.h.g.jepsen@regionh.dk

address the human factor issues and improve patient safety by building up team skills and improving individual cognitive abilities (Gaba et al. 1998). Training and assessment of NTS for individuals or teams in healthcare have been shown to improve safe and efficient performance in the operating room (OR) (Neily et al. 2010). Specific positive effects have been reported regarding communication, teamwork and technical performance. Healthcare professionals value such training highly, they learn and change behaviour (Andersen et al. 2010; Fuhrmann et al. 2009), and they apply what they learn in practice (Morey et al. 2002).

Although there is an evident need for NTS training in healthcare to improve patient safety, the implementation of training and assessment of NTS has been slow (Flin and Patey 2011). One reason for the poor implementation of NTS training might be the lack of a shared understanding of the underlying concepts. Much can be learned from high-risk organisations that are further ahead in working with such a shared understanding, but the culture, context and organisation of work differs in different domains and organisations (Klampfer et al. 2001; Kontogiannis and Malakis 2013). Healthcare cannot simply apply the understanding from different domains (Glavin 2011; Nestel et al. 2011; Yule et al. 2009). Last but not least, there has been a tendency to view behavioural rating instruments as easy to use. Combined with financial reasons, this assumption has led to an insufficient training of the raters. Consequently, the ability to rate and provide feedback to the learners might not be optimal (Sevdalis et al. 2012).

Behavioural rating instruments are one attempt to provide a conceptual framework for NTS and make the underlying topics more accessible for healthcare professionals in educational and clinical settings. Some of these instruments include overarching categories, with a number of elements and examples of observable behaviours. The instruments are typically defined based on job analysis studies using different methods, like interviews and observations with respective stakeholders. Thus, these are based on the definition of what is seen as the non-technical expertise of the profession for which the instrument was designed.

To our knowledge, there is no overview of the studies describing the construction of individual and team behavioural rating instruments in healthcare on a detailed level. Such an overview is needed in order to interpret the results of the instruments and use these in the best interest of patients, healthcare professionals and the organisations that exist to provide the framework for this care. Different stakeholders' views might vary considerably so analysing the data basis for any instrument might help in identifying biases and in defining the scope of application.

The aim of this paper is to give an overview of the development of different behavioural rating instruments to assess NTS at individual and team level within different medical specialities and settings.

2 Methods

This paper is a critical review, identifying papers that represent certain types of instruments to assess healthcare professionals NTS at individual or team level. This review does not constitute an exhaustive list but provides an overview of different instruments for different professions and teams.

During our work with behavioural rating instruments to assess NTS (Jepsen et al. 2012; Lyk-Jensen et al. 2014; Spanager et al. 2012, 2013), we have gained insight into different aspects of these instruments for assessment of NTS. We used this knowledge to select instruments for the review. In addition, we searched online resources including PubMed Medline, EMBASE and Google Scholar for relevant research papers. Bibliographies from relevant research papers were consulted.

2.1 Selection of articles

Articles were selected if they fulfilled both of the following criteria:

- The subjects of study were physicians and/or trainee physicians alone or in combination with nurses and/or operating department practitioners and/or midwifes.
- They described instruments assessing NTS or behaviours of individuals or teams in simulated or clinical healthcare settings.

2.2 Data extraction

Two authors (DO, RMHGJ) reviewed the articles to assess their eligibility based on the selection criteria. Supported by the third author (PD), they reviewed all of the selected articles and decided which to include in order to illustrate the variability in development methods and validation procedures.

A coding sheet was developed focusing on relevant parameters that described:

- The development of the instruments; the purpose of the instrument; the profession(s) that the instrument was designed for; the other stakeholders involved in the development; psychologists; other instruments they were based on; and methods of data collection.
- The scoring system and scales.

- The validation process; the training of the raters; methods used in assessing reliability and validity.

3 Results

Tables 1 and 2 show an overview of the 23 included instruments for assessment of teams' and individuals' NTS and include 16 and 7 instruments, respectively. A large variety of behavioural rating instruments for different target groups in different specialities was found.

3.1 Purpose of the instruments

The instruments have all been designed for observation of NTS in different situations mainly for real-time, retrospective recalling or for video recordings of operations or simulations. Of all instruments, 11 are designed for assessment in the OR; five of these are for OR teams (Healey et al. 2004; Hull et al. 2011; Mishra et al. 2009; Schraagen et al. 2010; Sevdalis et al. 2008; Undre et al. 2007), three for anaesthesiologists (Crossingham et al. 2012; Fletcher et al. 2003, 2004; Jepsen et al. 2012) and three for surgeons (Parker et al. 2013; Spanager et al. 2012; Yule et al. 2006, 2008). Four instruments are developed for assessment of NTS during resuscitations (Cooper et al. 2010; Plant et al. 2011; Thomas et al. 2004; Walker et al. 2011), two for obstetric teams (Guise et al. 2008; Morgan et al. 2012), two for trauma teams NTS' (Steinemann et al. 2012; Westli et al. 2010), two for teams in acute settings (Kim et al. 2006; Malec et al. 2007) and two for critical care teams (Lambden et al. 2013; Weller et al. 2011). Six out of seven instruments for assessment of individuals are designed for the OR.

3.2 Involvement of other stakeholders in the development besides target group and psychologists

Most researchers involved in developing the instruments have the same professional background as the target group of healthcare personnel involved, and most often psychologists are part of the team. The development teams usually consist of 3–6 persons. Table 2 shows that the research groups for the individual instruments for surgeons and anaesthesiologist, Non-Technical Skills for Surgeons (NOTSS), NOTSS customised for Danish surgeons (NOTSSdk) and Anaesthesiologists Non-Technical Skills customised for Danish anaesthesiologists (ANTSdk) have included health professionals from the different professions and specialties in the OR team. Two of the team tools in Table 1 state clearly that there have been human factor experts involved in the development process (Mishra et al. 2009; Schraagen et al. 2010).

3.3 Main sources of data which the instruments are based on

At least 19 of the instruments have been inspired by aviation instruments like Non-technical Skills system for assessing pilots' CRM skills (NOTECHS) (Flin et al. 2005) and Line Operations Safety Audit (LOSA) (Klinect et al. 2003) or have been developed directly or indirectly on the basis of these instruments (Cooper et al. 2010; Fletcher et al. 2004; Guise et al. 2008; Healey et al. 2004; Jepsen et al. 2012; Lambden et al. 2013; Malec et al. 2007; Mishra et al. 2009; Plant et al. 2011; Schraagen et al. 2010; Sevdalis et al. 2008; Spanager et al. 2012; Steinemann et al. 2012; Thomas et al. 2004; Undre et al. 2007; Walker et al. 2011; Weller et al. 2011; Westli et al. 2010; Yule et al. 2006).

In Table 2, it can be seen that the Danish customised instruments for assessment of surgeons' and anaesthesiologists' NTS (Jepsen et al. 2012; Spanager et al. 2012) are developed on the basis of interviews with all members of the OR team; this is in contrast to the development of the original UK-developed instruments (Fletcher et al. 2004; Yule et al. 2006) which were developed on the basis of mono-disciplinary interviews. It is also in contrast to the other instruments for assessment of individuals in the OR (Crossingham et al. 2012; Parker et al. 2013) which were developed without involvement from other OR members than the observed physicians.

3.4 Structure

Most instruments consist of two or three levels with four to eight overarching categories/dimensions and underlying examples of skills or behaviours. The skills or behaviours can be rated after observation using a numerical scoring scale or a set of anchors, which are examples of different expressions of the NTS. The overarching categories comprise both cognitive and social skills. Examples of cognitive categories are; 'situation awareness', 'decision making', 'empathy and sensitivity'. Examples of social categories are; 'communication', 'team work', (shared) 'leadership', 'task management', 'organisation', 'working under pressure'. The same categories in different instruments can encompass different concepts and also overlap with other categories in other instruments. Four of the instruments also assess technical skills (Healey et al. 2004; Lambden et al. 2013; Schraagen et al. 2010; Undre et al. 2007).

3.5 Validation of instruments

Many of the instruments are not only based on aviation instruments but also process mapping and cognitive task

Table 1 Instruments for assessment of teams' NTS

Tool	Purpose and population	Development: Other stakeholders involved in the construction, besides involved groups and psychologists Main sources of data	Scoring system and scale	Validation procedure	Reliability	Validity
Oxford Non-Technical Skills (NOTECHS) (Mishra et al. 2009)	Intraoperative; surgical teams (surgeons, anaesthesiologists and scrub nurses)	**Other stakeholders:** 1 human factors expert and 2 aviation crew resource management trainers **Based on:** NOTECHS (aviation) Task analysis, consultation with content experts	**Scale:** 4-point **Global rating:** No	**Who:** 1 surgical trainee, 1 human factors expert and 1 expert in evaluating aviation NTS **Rater training:** Yes **Simulation:** No **Video rating:** No **Real-time observation:** Yes **How:** 65 direct observations in the OR	Inter-rater Test–retest	Predictive Concurrent Convergent
Revised Non-Technical Skills (NOTECHS) (Sevdalis et al. 2008)	Intraoperative; surgical teams (surgeons, anaesthesiologists, scrub nurses and OPDs)	**Other stakeholders:** No **Based on:** NOTECHS (aviation)	**Scale:** 6-point **Global rating:** No	**Who:** Whole OR team, trainers and trainees, 1–2 psychologists **Rater training:** Yes (in the second training series) **Simulation:** Yes **Video rating:** No **Real-time observation:** Yes **How:** 40 OR simulations involving a number of crises	Internal consistency	Construct
Objective teamwork assessment for surgery (OTAS) (Healey et al. 2004; Hull et al. 2011)	Intraoperative; surgical team (surgeons, anaesthesiologists, scrub nurses and OPDs) combined with procedural checklist	**Other stakeholders:** No **Based on:** Input-process-output model of team performance from aviation and UK health services. OR protocols, good practice guidelines and an expert consensus process. (Dickinson and McIntyre's model)	**Scale:** Team performance: 7-point. Tasks are rated by yes/no. Performance rated 3 times for each sub-team; pre-, intra- and post-operative **Global rating:** No	**Who:** 2 psychologists **Rater training:** Yes **Simulation:** No **Video rating:** No **Real-time observation:** Yes **How:** Observation of 30 operations, rating OR team with OTAS. Followed by an expert consensus process	Inter-rater	Construct Content Concurrent

Table 1 continued

Tool	Purpose and population	Development: Other stakeholders involved in the construction, besides involved groups and psychologists Main sources of data	Scoring system and scale	Validation procedure	Reliability	Validity
Objective Teamwork Assessment for Surgery (OTAS) refined for urology (Undre et al. 2007)	As OTAS	**Other stakeholders:** No **Based on:** OTAS. Observations in the OR	As OTAS	**Who:** A. A surgeon and a psychologist. B. 2 psychologists **Rater training:** Yes **Simulation:** No **Video rating:** No **Real-time observation:** Yes **How:** A. Observation of 50 urology procedures in 2 ORs for behavioural ratings. B. 6 additional cases for reliability of behavioural ratings	Inter-observer	–
Unnamed teamwork classification instrument (Schraagen et al. 2010)	Intraoperative; non-routine events and team performance effect on outcomes during paediatric cardiac surgery (PCS)	**Other stakeholders:** 2 human factor experts **Based on:** Instruments like; NOTECHS, ANTS, NOTSS and CATS. Process mapping, cognitive task analysis, Weinger and Slagle's definition of 'non-routine event' adapted from the nuclear power industry	**Scale:** Team performance 7-point. Subcategories of non-routine events were also rated **Global rating:** No	**Who:** 2 human factor experts **Rater training:** Yes **Simulation:** No **Video rating:** Yes **Real-time observation:** Yes **How:** Raters observed 10 live operations, and had several sessions of coding real videos, discrepancies in coding were discussed and settled upon using a 'gold' standard	Inter-rater	–
Assessment and global assessment of obstetrical team performance (AOTP and GAOTP) (Morgan et al. 2012)	To determine effectiveness of high-fidelity simulation team training; multidisciplinary obstetric teams (obstetricians, anaesthesiologists, registered nurses and in some cases a family doctor)	**Other stakeholders:** Unknown **Based on:** Literature, focus group interviews and simulation sessions	**Scale:** 5-point **Global rating:** Yes (GAOTP)	**Who:** 3 nurses, one midwife, 2 anaesthesiologists and 2 obstetricians **Rater training:** Yes **Simulation:** Yes **Video rating:** Yes **Real-time observation:** No **How:** 12 multidisciplinary obstetric teams. Each team participated in 4 scenarios 3 times (5–9 months after the first and 5–6 month after the second	Internal consistency Inter-scenario Inter-rater Test–retest	–

Table 1 continued

Tool	Purpose and population	Development: Other stakeholders involved in the construction, besides involved groups and psychologists Main sources of data	Scoring system and scale	Validation procedure	Reliability	Validity
Clinical Teamwork Scale (CTS) (Guise et al. 2008)	Measurement of key clinical teamwork skills; obstetric team	**Other stakeholders:** Unknown **Based on:** Crew resource management training program	**Scale:** 10-point (15 clinical skills and 5 teamwork questions). And a "target fixation item (behaviour)" evaluated with yes or no **Global rating:** Yes	**Who:** Raters were a perinatologist, a generalist obstetrician/gynaecologist and a nurse midwife **Rater training:** Not extensive because the instrument should be usable with minimum of training. Raters were previously trained in CRM principles and involved in research **Simulation:** Standardised videos were created **Video rating:** Yes **Real-time observation:** No **How:** 3 videos with 4 actors simulating different levels of teamwork in the same scenario (shoulder dystocia)	Inter-rater Generalisability coefficient	Construct Accuracy
University of Texas behavioural markers for neonatal resuscitation (UTBMNR) (Thomas et al. 2004)	Assessing teamwork behaviours for neonatal resuscitation; neonatal providers	**Other stakeholders:** No **Based on:** Line Operations Safety Audit (LOSA) and CRM from aviation. Focus groups, reviews of survey data from healthcare providers, video recording of 5 resuscitation situations of infants born by caesarean sectio—looking for observable behaviours	**Scale:** 4-point **Global rating:** 2 overall ratings (teamwork and leadership). Possible to indicate individual ratings if an individual differ significantly from the rest of the team	**Who:** The expert group of authors **Rater training:** Not mentioned, but they are experts **Simulation:** No **Video rating:** Yes **Real-time observation:** No **How:** Video recordings of resuscitation situation (real cases). Pilot tested on 20, revision and then tested on 113 additional video recordings	No (development paper)	No (development paper)
Observational skill-based clinical assessment tool for resuscitation (OSCAR) (Walker et al. 2011)	NTS and global performance; resuscitation team s (anaesthesiology, medical, nursing staff)	**Other stakeholders:** No **Based on:** OTAS, ANTS, Revised NOTECHS. Literature review	**Scale:** 7-point **Global rating:** Yes	**Who:** 2 expert clinical observers **Rater training:** No **Simulation:** Yes **Video rating:** Yes **Real-time observation:** No **How:** Rating 4 videos from simulation training suite and 4 from unannounced in situ cardiac arrest simulations	Inter-rater International consistency	Face Content Convergent

Table 1 continued

Tool	Purpose and population	Development: Other stakeholders involved in the construction, besides involved groups and psychologists Main sources of data	Scoring system and scale	Validation procedure	Reliability	Validity
Team emergency assessment measure (TEAM) (Cooper et al. 2010)	Teamwork assessment for emergency resuscitation team performance	**Other stakeholders:** No **Based on:** Review of literature, expert team (a resuscitation officer, 2 nurses, general practitioner, psychologist and medical educator), reviewed by experts (6 resuscitation experts;2 doctors and 4 nurses)	**Scale:** 5-point **Global rating:** Yes	**Who:** A. 2 expert raters, B. 3 experienced resuscitation trainers/ clinicians **Rater training:** No **Simulation:** Yes **Video rating:** Yes **Real-time observation:** No **How:** A. 56 video recorded hospital resuscitation events. B. 15 simulated resuscitations	Inter-rater International consistency Test–retest	Content Construct Concurrent
Mayo High Performance Teamwork Scale (MHPTS) (Malec et al. 2007)	CRM related NTS during training episodes in acute settings; healthcare teams (physicians and nurses)	**Other stakeholders:** Unknown **Based on:** ANTS, Anaesthesiology CRM training. Rasch analysis	**Scale:** 3-point **Global rating:** No	**Who:** 19 residents and 88 nurses **Rater training:** Part of a CRM training **Simulation:** Yes **Video rating:** No **Real-time observation:** No **How:** Participants rated team performance in 2–3 scenarios retrospectively with MHPTS directly after simulation of scenarios.	Inter-rater Internal consistency	Construct
Ottawa Crisis Resource Management Global Rating Scale (Ottawa GRS) (Kim et al. 2006)	NTS and global CRM performance; can be used for all specialities of healthcare teams (physicians and nurses) in acute settings. Developed for simulation	**Other stakeholders:** No **Based on:** Crisis Resource Management literature	**Scale:** 7-point **Global rating:** Yes (overall performance score)	**Who:** 3 raters **Rater training:** Yes **Simulation:** Yes **Video rating:** Yes **Real-time observation:** No **How:** 50 residents from first and third year participated in 2 stimulated scenarios of emergencies seen in acute care settings	Inter-rater Internal consistency	Construct

Table 1 continued

Tool	Purpose and population	Development: Other stakeholders involved in the construction, besides involved groups and psychologists Main sources of data	Scoring system and scale	Validation procedure	Reliability	Validity
The imperial paediatric emergency training toolkit (IPETT) (Lambden et al. 2013)	Assessment of technical and non-technical skills required to manage paediatric emergencies in critical care settings; paediatric (intensive care unit) teams (PICU teams)	**Other stakeholders:** No **Based on:** Non-technical component based on NOTECHS (surgery, trauma and aviation). Literature, experts, reviewed and adapted. The technical component was developed in 2 iterative stages; first a consultant paediatrician and a simulation expert. Secondly a paediatric and anaesthesiology consultant—last step consensus	**Scale:** 7-point (NTS), 6-point (technical skills) **Global rating:** Yes, both NTS and technical skills	**Who:** A PICU consultant and/or consultant anaesthesiologist trainer **Rater training:** The raters were the developers of IPETT **Simulation:** Yes in PICU or paediatrics ward **Video rating:** No **Real-time observation:** Yes **How:** 52 participants in 26 simulations with paediatric trainees and 9 with both anaesthesiology and paediatric trainees were evaluated with IPETT	Internal consistency Inter-rater	Face Content Concurrent
Unnamed instrument (Weller et al. 2011)	Measure of team behaviour; critical care teams	**Other stakeholders:** No **Based on:** Literature review	**Scale:** 7 point rating scale **Global rating:** Yes	**Who:** 3 expert raters **Rater training:** Yes (independent ratings, reconciled their ratings after each case—developed common understanding) **Simulation:** Yes **Video rating:** Yes **Real-time observation:** No **How:** Video of 40 critical care teams (one doctor, three nurses) participating in 4 different simulations	Generalisability coefficient	Construct
Trauma Non-Technical Skills (T-NOTECHS) (Steinemann et al. 2012)	Simulated or real-life trauma calls; all specialities and staff attending trauma calls (trauma/critical care surgeons, trauma/medical intensivists, trauma/critical care nurses)	**Other stakeholders:** No **Based on:** NOTECHS	**Scale:** 5-point **Global rating:** No	**Who:** Critical care nurses and research assistants **Rater training:** Yes **Simulation:** Yes **Video rating:** Yes **Real-time observation:** No **How:** Simulated and real cases. Scoring using audio response system immediately after the clinical case	Inter-rater Internal consistency	Construct

Table 1 continued

Tool	Purpose and population	Development: Other stakeholders involved in the construction, besides involved groups and psychologists Main sources of data	Scoring system and scale	Validation procedure	Reliability	Validity
Revised Anaesthetists' Non-Technical Skills (ANTS) and Anti-Air Teamwork Observation Measure (ATOM) (Westli et al. 2010)	Specific teamwork skills and a shared mental model for the effective medical management of trauma teams (surgeons, anaesthesiologists, anaesthetic nurses, emergency medical nurses and radiographers)	**Other stakeholders:** No **Based on:** ANTS and ATOM. Observations of 20 trauma teams in training simulations and 4 interviews with experienced anaesthetists and intensive care workers	**Scale:** 5-point. The scoring format of ANTS and ATOM were revised to index the moment to moment behaviour **Global rating:** No, but frequency ratings where used	**Who:** Experienced clinicians, 2 subject matter experts **Rater training:** Yes **Simulation:** Yes **Video rating:** Yes **Real-time observation:** No **How:** Based on analysis of 27 video recording of trauma teams participating in the BEST simulation-based team training program. Interviews	Inter-rater	–

ANTS Anaesthetists' Non-Technical Skills, *CATS* communication and teamwork skills, *CRM* Crisis Resource Management, *NOTSS* Non-Technical Skills for Surgeons, *NTS* non-technical skills, *OR* operating room, *ODP* operating department practitioner, *PCS* paediatric cardiac surgery

analysis. They have gone through multiple iterations and adaptations to work in healthcare. Most of the instruments are well validated for the setting that they are developed for; inter-rater/observer reliability has been tested for 17 instruments, internal consistency (nine instruments), test–retest reliability and generalisability coefficient (three instruments), construct validity (eight instruments), construct validity (six instruments), face validity (five instruments), concurrent validity (four instruments) and convergent validity (two instruments). Examples of the adaptation of instruments to other settings and cultures are also seen (Jepsen et al. 2012; Lambden et al. 2013; Spanager et al. 2012; Undre et al. 2007).

4 Discussion

In this review, the development of 23 instruments is presented illustrating different instruments used to assess NTS in different healthcare settings. Many of the teams for which these instruments were developed handle emergencies. Some of the research groups have looked beyond their own speciality and included other members in the development process (Jepsen et al. 2012; Mishra et al. 2009; Schraagen et al. 2010; Spanager et al. 2012; Yule et al. 2006). Most instruments are based on experiences learned in aviation, but are now well validated and adapted to different healthcare settings. All the instruments consist of very similar categories of NTS, with few exceptions. Similar categories have been found in many high-risk organisations, including different medical specialities (Flin et al. 2005).

In this discussion, we reflect upon the development of the instruments described and the way they are used.

4.1 The use of behavioural rating instruments

The behavioural rating instruments for individuals are primarily developed for assessment of competence in the clinical setting in order to structure learning and illustrate a development over time. The instruments can be used for assessment several times during specialist training. The ability of raters to rate and provide feedback to the learner is vital. The risk that the assessment of NTS seems deceptively simple has been discussed before (Flin and Patey 2011; Schraagen et al. 2010; Sevdalis et al. 2012), but it is still important to emphasise a focus on the training of raters. Thus, the implementation of the instruments would depend on helping people to acquire the ability to use them. This aspect, however important, was not part of our review.

One possible advantage of the NTS behavioural rating instruments on a higher level could be that they allow for

Table 2 Instruments for assessment of individuals' NTS

Tool	Purpose and population	Development: Other stakeholders involved in the construction, besides involved groups and psychologists Main sources of data	Scoring system and scale	Validation procedure	Reliability	Validity
Unnamed instrument (Crossingham et al. 2012)	Intraoperative; trainee anaesthesiologists	**Other stakeholders:** No **Based on:** A system used for recruitment of anaesthesiologists. Job analysis	**Scale:** 5-point **Global rating:** Yes	**Who:** 57 anaesthesiologist consultants and 62 OPDs **Rater training:** Yes, before 2nd round **Simulation:** No **Video rating:** Yes (for rater training) **Real-time observation:** Yes **How:** Trainees were rated during real operations by consultants and ODPs	Inter-rater	–
Anaesthetists' Non-Technical Skills (ANTS) (Fletcher et al. 2003, 2004)	Intraoperative; anaesthesiologists	**Other stakeholders:** No **Based on:** NOTECHS (aviation). Review of human factors in anaesthesiology. Critical incident interviews with 29 anaesthesiologists. Prototype system refined by multiple iterative processes; OR observations, re-coding of sample interviews, reviewing incident reports	**Scale:** 4-point **Global rating:** No	**Who:** Pilot study: 11 consultant anaesthesiologists. Validity study: 50 consultant anaesthesiologists **Rater training:** Yes **Simulation:** Yes **Video rating:** Yes **Real-time observation:** No **How:** ANTS was used to rate anaesthesiologists NTS in 8 simulated videos of OR cases	Inter-rater Internal consistency	Content Observability
ANTSdk (Jepsen et al. 2012)	As ANTS	**Other stakeholders:** A nurse anaesthetist and a surgeon **Based on:** ANTS. Observations in OR, 6 semi structured group interviews with members of the OR team (scrub nurses, nurse anaesthetists, surgeons, consultant and trainee anaesthesiologists)	**Scale:** 5-point **Global rating:** Yes	Not yet published	Not yet published	Face (Others not yet published)
Unnamed instrument (Plant et al. 2011)	Self-efficacy in CRM skills during resuscitation training; paediatric residents	**Other stakeholders:** An anaesthesiologist **Based on:** ANTS, Ottawa GRS. Conceptual analysis of construct, development of potential item pool, revision of items, draft instrument, pilot testing by content experts	**Scale:** 5-point **Global rating:** No	**Who:** 3 trained observers **Rater training:** Yes **Simulation:** Yes **Video rating:** Yes **Real-time observation:** No **How:** The 3 observers used ANTS and Ottawa GRS to rate 125 paediatric residents in simulated resuscitation scenarios. Ratings where compared to residents self-efficacy ratings using CRM instrument	Inter-item correlation Internal consistency	Content

Table 2 continued

Tool	Purpose and population	Development: Other stakeholders involved in the construction, besides involved groups and psychologists Main sources of data	Scoring system and scale	Validation procedure	Reliability	Validity
Non-Technical Skills for Surgeons (NOTSS) version 1.2 (Yule et al. 2006, 2008)	Intraoperative; surgeons	**Other stakeholders:** A consultant anaesthesiologist **Based on:** Cognitive task analyses (critical incident interviews) with 27 consultant surgeons. Attitude survey, literature review, analysis of surgical mortality reports and OR observations Revised after psychometric evaluation of the prototype	**Scale:** 4-point **Global rating:** No	**Who:** 44 surgeons **Rater training:** Yes **Simulation:** Yes **Video rating:** Yes **Real-time observation:** No **How:** NOTSS was used to rate consultant surgeons' behaviours in 6 simulated OR scenarios	Inter-rater Internal consistency	Convergent
NOTSSdk (Spanager et al. 2012, 2013)	As NOTSS	**Other stakeholders:** A consultant anaesthesiologist, a nurse anaesthetist **Based on:** NOTSS. OR observations, 8 semi structured group interviews with members of the OR team (scrub nurses, anaesthesiologists, nurse anaesthetists, consultant and trainee surgeons)	**Scale:** 5-point **Global rating:** Yes	**Who:** 15 surgeons **Rater training:** Yes **Simulation:** Yes **Video rating:** Yes **Real-time observation:** No **How:** NOTSS was used to rate consultant surgeons' behaviours in 9 simulated OR scenarios, before and after a 4 h training session on NTS, NOTSSdk and training of videorating	Inter-rater Generalisability coefficient	Face
Surgeons Leadership Inventory (SLI) (Parker et al. 2013)	Intraoperative; surgeons leadership	**Other stakeholders:** No **Based on:** OR observations, literature review, 10 single-discipline focus group interviews	**Scale:** 4-point **Global rating:** No	**Who:** 2 psychologists **Rater training:** Yes **Simulation:** No **Video rating:** Yes (live operations) **Real-time observation:** No **How:** Coding of 5 videos showing operations using SLI	Inter-rater	Face

CRM Crisis Resource Management, *NTS* non-technical skills, *OR* operating room, *ODP* operating department practitioner, Ottawa Crisis Resource Management Global Rating Scale: Ottawa GRS

aligning the definition of educational aims and objectives, the selection of contents and methods to deliver them and the analysis of the progress in the learning.

The behavioural rating instruments for teams of health professionals are mainly used to identify weaknesses (learning objectives) and to evaluate the effect of training by comparing pre- and post-values, ideally repeatedly over a longer period of time. This implies that the instruments are mainly used for formative assessment of the team. A summative assessment of teams of health professions does seldom make sense, since the team might seldom work in the same constellation again.

Our basic view is that the behavioural rating instruments described above should be integrated into the actual clinical environment. No matter, whether used in a formative or summative way, on individual basis or on team level, the instruments are used to record observations of ratings by raters. What actually happens with the recorded values and how they are interpreted is important for the usefulness of the instruments in the larger context of healthcare—for patients, their relatives and those involved in their care.

The instruments should be used in a way that is intended by their designers, but they may also be used against this intention or without considering this intention. An instrument that was designed for repeated formative feedback during training might be used in a research setting to evaluate the effect of training. While it might not have the best psychometric qualities, it will still produce results of some value.

All assessment situations should be followed by feedback to the learner. There might be differences in how helpful the feedback conversation is, depending on skills and attitudes of the rater as feedback provider, based on training, personal preferences etc. (Sevdalis et al. 2012). All these surrounding factors of integrating the instruments into context are not inherent in the instruments themselves, but need to be learned by the one using the instruments. The description of the context in which an instrument is to be used, the scope of application and warnings against using them outside of this scope are thus important (Flin and Patey 2011).

In any case, each measurement provides a snapshot only. Combining several measurements over time will increase the validity of the result. Rating performed over time by different raters and learners combining the view of different people about different people will help overcome biases and analyse the "normal variation" amongst those measured and measuring. The results from the different ratings would need to be integrated in performance assessment portfolios to allow collection of data over time.

Considerable skill is required to make observations and ratings and to provide constructive feedback to those being rated. The amount of training provided to the raters varies considerably in the papers included in this review. In addition, the methods used for training vary. In recent years, training includes rating of simulation-based scenarios, but there are no standards for how to do this. It might, however, be necessary to provide a more focused description and to, even more clearly, demand the training of raters for the whole process in which the instruments are used, before they are allowed to perform ratings. A mere training that focuses on understanding the development and the dimensions in the instrument might simply reach too short when it comes to providing useful feedback to learners (Sevdalis et al. 2012). The feedback provided based on the values achieved is what helps learners develop.

Another aspect worth considering is the challenge that there is an evident lack of gold standard for ratings of NTS (Flin and Patey 2011; Graham et al. 2010; Malec et al. 2007; Schraagen et al. 2010), this further challenge rater training and implementation. However, one might speculate if there can ever be a gold standard for NTS in healthcare as the patient safety culture develops continuously. One challenge for progress is the complex connection between NTS and technical skills in healthcare.

In the future, the instruments can be used in the simulated setting and can help in structuring debriefing and learning. The instruments might also be used for summative or "high stakes" assessments. In these situations, the training of the raters is extremely important. It is important to use the instruments in a relevant way, in realistic conditions or scenarios in which the behaviour is observed.

4.2 Criteria and their definition and ratings

The instruments aim to be based on observable behaviours—which form the criteria on which the assessment is based. The numeric values assigned to the criteria reflect how the rater mapped her impression onto the scales provided. Much is written about the mapping process and its biases. Raters who are too strict or lenient, fail to look behind the halo of specific episodes, do not agree on what they saw or how they should evaluate it, or points that are missed completely (Bested et al. 2011). There seems to be less reflection about the definition of the criteria, or in other words, the basis on which they would be called valid or not.

It has been shown that what is seen as an (in-)competent healthcare professional changes over time (Hodges 2006). Requiring such a thing as NTS is rather new, stimulated by the 2000 IOM report and the pioneering work in patient safety and simulation (Cooper et al. 2002; Gaba and DeAnda 1988; Gaba et al. 1987). The underlying skills might have been discussed for much longer, but for example, the instruments discussed in this paper help in creating terms

that can be used to describe phenomena in a more consistent way. Certain problems can now be termed 'a loss of situational awareness', others 'breakdown in communication', where formerly there might have been just 'problems'. Helping in making NTS a relevant part of the professional roles of current healthcare professionals, stimulating the discussion, and providing terms for it is a key achievement of these and other instruments. The evidence that the NTS do have an influence on the technical performance has been collected (Lingard et al. 2002; Manser 2009).

Yet, it might not be the final version of the criteria—there might not even *be* a final version in the ever-changing flow of thinking about healthcare professionals. The current versions are one view on various professional roles and their non-technical parts. As the patient safety culture in healthcare organisations develop (improve), the criteria on which the behavioural markers are developed will change. This has been seen in other high-reliability organisations. Therefore, it remains to be seen how much more refinement there will be over time.

Besides changes over time, these criteria are also subject to changes based on the frame of reference from which they are seen. One might stress health models that look beyond the absence of illness and consider a salutogenic model (Antonovsky 1996). For patients, the experience in the healthcare system might still not be "good" despite their outcome being recorded as "good", e.g. all organs functioning within reasonable limits in a highly depressive patient. The underlying question concerned is who should actually be part of forming the criteria against which both non-technical and technical performance should be evaluated? Who are the stakeholders who should be heard? How many stakeholders can be heard and in what way, so that development processes are not stalled in overwhelming complexity? Neither the instruments reviewed here nor we have an answer to these questions. Yet, one of the reflection points for their use, for new developments and further refinements should be based on this question. Can anaesthesiologists (alone) provide the criteria against which anaesthesiologists' work is evaluated in a non-technical sense? Is it enough, if psychologists help in refining those? Our group and others have extended the definition of the criteria by asking other healthcare professionals as well. Is this enough? How about the allied health professionals and, last not least, also the patients and their relatives. The current instruments do reflect a starting point and that is laudable. We think, however, that all reports about the instruments should (a) reflect in more detail who was involved in defining the criteria contained and (b) which limitations naturally follow from selecting one frame of reference.

Further, one might need to consider cultural differences (Klampfer et al. 2001), be it national cultures, but also organisational and departmental cultures, maybe even team cultures (Klampfer et al. 2001). In principle, these should play a modifying part when mapping the observations onto the scale values.

The instruments cannot just be transferred from one context to another (Flin and Patey 2011). We should evaluate how well a specific observable behaviour supported the element and category to which it belongs *in this specific team and context*. In the logic of reliability and validity testing, all these context influences are hoped for to level out as zero, as they are thought to be random errors. For an individual, however, this might mean a lot—possibly high-stake decisions based on observations. In a training context, any feedback provider would need to be trained to consider those issues. In the research context, it might be necessary to enlarge the description of the research context and to describe how it was accounted for. When adapting NOTSS to NOTSS.dk and ANTS to ANTSdk, differences were found not only on the level of the behavioural markers, but also on the level of elements (Jepsen et al. 2012; Spanager et al. 2012). This might reflect a difference in the methods used, but also differences in cultures between Scotland and Denmark. Over time, it might be necessary to adapt the rating instruments for local context—and still balancing the need for criterion-based discussion of healthcare.

4.3 Ratings on scales versus ability of people

One challenge that underlies each assessment instrument is that its values might be taken for the real thing. It is important to note that each of the instruments contains constructs that aim to describe complex socio-technical systems. They do so by reducing complexity, by levelling fine-grained differences and by forcing raters into perceiving and thinking in a standardised way, what could be perceived as the downside of bringing the topics into the discussion by using standardised terms. Bateson described a similar line of thought with his distinction between an actual territory and a map depicting this territory and points out that we should not mistake the map for the territory (Bateson and Bateson 2000). A worst-case scenario could be that the values on the assessment instrument replace the real thing, the professional capabilities of a human being. There are elements that these instruments have difficulty in capturing, such as sense of responsibility, thinking processes—all that is inside the head of the learner.

5 Conclusion

In this review, we discussed the development and the use of 23 behavioural rating instruments to assess NTS at the individual and the team level. There are several caveats to be

aware of when using the instruments. There is a need for increased knowledge of how to validate these instruments, how to train the raters and how to continuously refine these instruments in order to help health professions develop their NTS. Overall, we recommend the continuous development and implementation of these instruments in healthcare to increase the awareness of the importance of human factors, to facilitate the training and assessment of NTS and the quality of the feedback provided to the health professions, with the long-term goal to improve patient safety.

References

Andersen PO, Jensen MK, Lippert A, Østergaard D, Klausen TW (2010) Development of a formative assessment tool for measurement of performance in multi-professional resuscitation teams. Resuscitation 81(6):703–711

Antonovsky A (1996) The salutogenic model as a theory to guide health promotion. Health Promot Int 11(1):11–18

Bateson G, Bateson MC (2000) Steps to an ecology of mind (vol 1972). University of Chicago Press Chicago. http://sspa.boisestate.edu/anthropology/files/2010/08/BATESON-Experiments-in-Thinking.pdf. Accessed 20 May 2014

Bested KM, Malling B, Skjelsager K, Østergaard D, Østergaard HT, Ringsted C (2011) Rater bias in postgraduate medical education. Ugeskr Laeger 173(44):2788–2790

Cooper JB, Newbower RS, Long CD, McPeek B (2002) Preventable anesthesia mishaps: a study of human factors*. Qual Saf Health Care 11(3):277–282. doi:10.1136/qhc.11.3.277

Cooper S, Cant R, Porter J, Sellick K, Somers G, Kinsman L, Nestel D (2010) Rating medical emergency teamwork performance: development of the Team Emergency Assessment Measure (TEAM). Resuscitation 81(4):446–452

Corrigan JM (2005) Crossing the quality chasm. In: Reid PP, Compton WD, Grossman JH et al (eds) Building a better delivery system: a new engineering/health care partnership. National Academies Press, Washington, DC. Crossing the Quality Chasm. http://www.ncbi.nlm.nih.gov/books/NBK22857/. Accessed 20 Apr 2014

Crossingham GV, Sice PJA, Roberts MJ, Lam WH, Gale TCE (2012) Development of workplace-based assessments of non-technical skills in anaesthesia*. Anaesthesia 67(2):158–164. doi:10.1111/j.1365-2044.2011.06977.x

Fletcher G, Flin R, McGeorge P, Glavin R, Maran N, Patey R (2003) Anaesthetists' non-technical skills (ANTS): evaluation of a behavioural marker system. Br J Anaesth 90(5):580–588

Fletcher G, Flin R, McGeorge P, Glavin R, Maran N, Patey R (2004) Rating non-technical skills: developing a behavioural marker system for use in anaesthesia. Cognit Technol Work 6(3). doi:10.1007/s10111-004-0158-y

Flin R, Patey R (2011) Non-technical skills for anaesthetists: developing and applying ANTS. Best Pract Res Clin Anaesthesiol 25(2):215–227. doi:10.1016/j.bpa.2011.02.005

Flin R, Martin L, Goeters K-M, Hörmann H-J, Amalberti R, Valot C, Nijhuis H (2005) Development of the NOTECHS (Non-Technical Skills) system for assessing pilots' CRM skills. In: Harris D, Muir HC (eds) Contemporary issues in human factors and aviation safety. Ashgate, Aldershot, pp 133–154

Fuhrmann L, Østergaard D, Lippert A, Perner A (2009) A multi-professional full-scale simulation course in the recognition and management of deteriorating hospital patients. Resuscitation 80(6):669–673

Gaba David M (2000) Anaesthesiology as a model for patient safety in health care. BMJ 320(7237):785–788

Gaba David M, DeAnda A (1988) A comprehensive anesthesia simulation environment: re-creating the operating room for research and training. Anesthesiology 69(3):387

Gaba David M, Maxwell M, DeAnda A (1987) Anesthetic mishaps: breaking the chain of accident evolution. Anesthesiology 66(5):670

Gaba DM, Howard SK, Flanagan B, Smith BE, Fish KJ, Botney R (1998) Assessment of clinical performance during simulated crises using both technical and behavioral ratings. Anesthesiology 89(1):8

Glavin RJ (2011) Skills, training, and education. Simul Healthc 6(1):4–7. doi:10.1097/SIH.0b013e31820aa1ee

Graham J, Hocking G, Giles E (2010) Anaesthesia Non-Technical Skills: can anaesthetists be trained to reliably use this behavioural marker system in 1 day? Br J Anaesth 104(4):440–445. doi:10.1093/bja/aeq032

Guise JM, Deering SH, Kanki BG, Osterweil P, Li H, Mori M, Lowe NK (2008) Validation of a tool to measure and promote clinical teamwork. Simul Healthc 3(4):217–223

Healey A, Undre S, Vincent C (2004) Developing observational measures of performance in surgical teams. Qual Saf Health Care 13(Suppl 1):i33–i40. doi:10.1136/qshc.2004.009936

Hodges B (2006) Medical education and the maintenance of incompetence. Med Teach 28(8):690–696

Hull L, Arora S, Kassab E, Kneebone R, Sevdalis N (2011) Observational teamwork assessment for surgery: content validation and tool refinement. J Am Coll Surg 212(2):234–243.e5. doi:10.1016/j.jamcollsurg.2010.11.001

Jepsen RMHG, Lyk-Jensen HT, Spanager L, Dieckmann P, Østergaard D (2012) Anaesthetists' non-technical skills—adapting the system to another setting. In: Short communications, AMEE 2012

Kim J, Neilipovitz D, Cardinal P, Chiu M, Clinch J (2006) 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 34(8):2167–2174. doi:10.1097/01.CCM.0000229877.45125.CC

Klampfer B, Flin R, Helmreich RL, Häusler R, Sexton B, Fletcher G, Field P, Staender S, Lauche K, Dieckmann P, Amacher A (2001) Group interaction in high risk environments. Enhancing performance in high risk environments: recommendations for the use of behavioural markers. Retrieved from http://www.abdn.ac.uk/iprc/documents/ants/GIHRE21_rec_for_use_of_beh_markers.pdf. Accessed 20 May 2014

Klinect JR, Murray P, Merritt A, Helmreich R (2003) Line operations safety audit (LOSA): definition and operating characteristics. In: Proceedings of the 12th international symposium on aviation psychology, pp 663–668. Retrieved from https://www.faa.gov/about/initiatives/maintenance_hf/losa/publications/media/klinect_operatingcharacteristics2003.pdf. Accessed 20 Apr 2014

Kohn LT, Corrigan JM, Donaldson MS (1999) To err is human. Building a safer health system. National Academy Press, Washington, DC. http://books.nap.edu/openbook.php?record_id=9728. Accessed 20 Apr 2014

Kontogiannis T, Malakis S (2013) Strategies in coping with complexity: development of a behavioural marker system for air traffic controllers. Saf Sci 57:27–34

Lambden S, DeMunter C, Dowson A, Cooper M, Gautama S, Sevdalis N (2013) The imperial paediatric emergency training toolkit (IPETT) for use in paediatric emergency training: development and evaluation of feasibility and validity. Resuscitation 84(6):831–836

Lingard L, Reznick R, Espin S, Regehr G, DeVito I (2002) Team communications in the operating room: talk patterns, sites of tension, and implications for novices. Acad Med 77(3):232–237

Lyk-Jensen H, Jepsen RM, Spanager L, Dieckmann P, Østergaard D (2014) Assessing nurse anaesthetists' non-technical skills in the operating room. Acta Anaesthesiol Scand 58:794–801

Malec JF, Torsher LC, Dunn WF, Wiegmann DA, Arnold JJ, Brown DA, Phatak V (2007) The mayo high performance teamwork scale: reliability and validity for evaluating key crew resource management skills. Simul Healthc 2(1):4–10

Manser T (2009) Teamwork and patient safety in dynamic domains of healthcare: a review of the literature. Acta Anaesthesiol Scand 53(2):143–151. doi:10.1111/j.1399-6576.2008.01717.x

Mishra A, Catchpole K, McCulloch P (2009) The Oxford NOTECHS System: reliability and validity of a tool for measuring teamwork behaviour in the operating theatre. Qual Saf Health Care 18(2):104–108

Morey JC, Simon R, Jay GD, Wears RL, Salisbury M, Dukes KA, Berns SD (2002) Error reduction and performance improvement in the emergency department through formal teamwork training: evaluation results of the MedTeams project. Health Serv Res 37(6):1553–1581

Morgan PJ, Tregunno D, Pittini R, Tarshis J, Regehr G, Desousa S et al (2012) Determination of the psychometric properties of a behavioural marking system for obstetrical team training using high-fidelity simulation. BMJ Qual Saf 21(1):78–82. doi:10.1136/bmjqs-2011-000296

Neily J, Mills PD, Young-Xu Y, Carney BT, West P, Berger DH et al (2010) Association between implementation of a medical team training program and surgical mortality. JAMA 304(15):1693–1700. doi:10.1001/jama.2010.1506

Nestel D, Walker K, Simon R, Aggarwal R, Andreatta P (2011) Nontechnical skills: an inaccurate and unhelpful descriptor? Simul Healthc 6(1):2

Parker SH, Flin R, McKinley A, Yule S (2013) The Surgeons' Leadership Inventory (SLI): a taxonomy and rating system for surgeons' intraoperative leadership skills. Am J Surg 205(6):745–751

Plant JL, Van Schaik SM, Sliwka DC, Boscardin CK, O'Sullivan PS (2011) Validation of a self-efficacy instrument and its relationship to performance of crisis resource management skills. Adv Health Sci Educ 16(5):579–590

Rasmussen MB, Dieckmann P, Barry Issenberg S, Ostergaard D, Søreide E, Ringsted CV (2012) Long-term intended and unintended experiences after advanced life support training. Resuscitation. doi:10.1016/j.resuscitation.2012.07.030

Schraagen JM, Schouten T, Smit M, Haas F, van der Beek D, van de Ven J, Barach P (2010) Assessing and improving teamwork in cardiac surgery. Qual Saf Health Care 19(6):1–6. doi:10.1136/qshc.2009.040105

Sevdalis N, Davis R, Koutantji M, Undre S, Darzi A, Vincent CA (2008) Reliability of a revised NOTECHS scale for use in surgical teams. Am J Surg 196(2):184

Sevdalis N, Hull L, Birnbach DJ (2012) Improving patient safety in the operating theatre and perioperative care: obstacles, interventions, and priorities for accelerating progress. Br J Anaesth 109(Suppl 1):i3–i16. doi:10.1093/bja/aes391

Spanager L, Lyk-Jensen HT, Dieckmann P, Wettergren A, Rosenberg J, Ostergaard D (2012) Customization of a tool to assess Danish surgeons' non-technical skills in the operating room. Dan Med J 59(11):A4526

Spanager L, Beier-Holgersen R, Dieckmann P, Konge L, Rosenberg J, Oestergaard D (2013) Reliable assessment of general surgeons' non-technical skills based on video-recordings of patient simulated scenarios. Am J Surg 206(5):810–817. doi:10.1016/j.amjsurg.2013.04.002

Steinemann S, Berg B, DiTullio A, Skinner A, Terada K, Anzelon K, Ho HC (2012) Assessing teamwork in the trauma bay: introduction of a modified "NOTECHS" scale for trauma. Am J Surg 203(1):69–75. doi:10.1016/j.amjsurg.2011.08.004

Thomas EJ, Sexton JB, Helmreich RL (2004) Translating teamwork behaviours from aviation to healthcare: development of behavioural markers for neonatal resuscitation. Qual Saf Health Care 13(Suppl_1):i57–i64. doi:10.1136/qshc.2004.009811

Undre S, Sevdalis N, Healey A, Darzi A, Vincent C (2007) Observational teamwork assessment for surgery (OTAS): refinement and application in urological surgery. World J Surg 31(7):1373–1381. doi:10.1007/s00268-007-9053-z

Walker S, Brett S, McKay A, Lambden S, Vincent C, Sevdalis N (2011) Observational skill-based clinical assessment tool for resuscitation (OSCAR): development and validation. Resuscitation 82(7):835–844. doi:10.1016/j.resuscitation.2011.03.009

Weller J, Frengley R, Torrie J, Shulruf B, Jolly B, Hopley L et al (2011) Evaluation of an instrument to measure teamwork in multidisciplinary critical care teams. BMJ Qual Saf 20(3):216–222. doi:10.1136/bmjqs.2010.041913

Westli HK, Johnsen BH, Eid J, Rasten I, Brattebø G (2010) Teamwork skills, shared mental models, and performance in simulated trauma teams: an independent group design. Scand J Trauma, Resusc Emerg Med 18(1):47. doi:10.1186/1757-7241-18-47

Yule S, Flin R, Paterson-Brown S, Maran N, Rowley D (2006) Development of a rating system for surgeons' non-technical skills. Med Educ 40(11):1098–1104. doi:10.1111/j.1365-2929.2006.02610.x

Yule Steven, Flin R, Maran N, Rowley D, Youngson G, Paterson-Brown S (2008) Surgeons' non-technical skills in the operating room: reliability testing of the NOTSS behavior rating system. World J Surg 32(4):548–556. doi:10.1007/s00268-007-9320-z

Yule S, Rowley D, Flin R, Maran N, Youngson G, Duncan J, Paterson-Brown S (2009) Experience matters: comparing novice and expert ratings of non-technical skills using the NOTSS system. ANZ J Surg 79(3):154–160