Medical Teacher

ISSN: 0142-159X (Print) 1466-187X (Online) Journal homepage: https://www.tandfonline.com/loi/imte20

Non-technical skills assessments in undergraduate medical education: A focused BEME systematic review: BEME Guide no. 54

Morris Gordon, Jeanne Farnan, Ciaran Grafton-Clarke, Ridwaan Ahmed, Dawne Gurbutt, John McLachlan & Michelle Daniel

MEDICAL TEACHER
https://doi.org/10.1080/0142159X.2018.1562166

BEME GUIDE

Non-technical skills assessments in undergraduate medical education: A focused BEME systematic review: BEME Guide no. 54

Morris Gordon[a], Jeanne Farnan[b], Ciaran Grafton-Clarke[c], Ridwaan Ahmed[d], Dawne Gurbutt[e], John McLachlan[f] and Michelle Daniel[g]

[a]Department of Evidence Synthesis and Systematic Review, University of Central Lancashire, Preston, UK; [b]Section of Hospital Medicine, The University of Chicago Pritzker School of Medicine, Chicago, IL, USA; [c]School of Life Sciences, University of Leicester, Leicester, UK; [d]Public Health Sciences, University of Central Lancashire, Preston, UK; [e]Centre for Excellence in Learning and Teaching, University of Central Lancashire, Preston, UK; [f]School of Medicine, University of Central Lancashire, Preston, UK; [g]Department of Emergency Medicine and Learning Health Sciences, University of Michigan Medical School, Ann Arbor, MI, USA

ABSTRACT

Consensus on how to assess non-technical skills is lacking. This systematic review aimed to evaluate the evidence regarding non-technical skills assessments in undergraduate medical education, to describe the tools used, learning outcomes and the validity, reliability and psychometrics of the instruments. A standardized search of online databases was conducted and consensus reached on included studies. Data extraction, quality assessment, and content analysis were conducted per Best Evidence in Medical Education guidelines. Nine papers met the inclusion criteria. Assessment methods broadly fell into three categories: simulated clinical scenarios, objective structured clinical examinations, and questionnaires or written assessments. Tools to assess non-technical skills were often developed locally, without reference to conceptual frameworks. Consequently, the tools were rarely validated, limiting dissemination and replication. There were clear themes in content and broad categories in methods of assessments employed. The quality of this evidence was poor due to lack of theoretical underpinning, with most assessments not part of normal process, but rather produced as a specific outcome measure for a teaching-based study. While the current literature forms a good starting position for educators developing materials, there is a need for future work to address these weaknesses as such tools are required across health education.

Introduction

Non-technical skills describe a set of social and cognitive abilities encompassing *situational awareness, risk assessment, clinical decision making, leadership, communication skills and teamwork* (Baldwin et al. 1999; Gordon et al. 2015a). *Situational awareness and risk assessment* involve perceiving, understanding and anticipating risks in a given environment (e.g. a physician recognizes that task-switching during a busy shift contributes to errors, thus the physician deliberately slows down and focuses on only one task when reading an ECG). *Decision making* requires the ability to diagnose situations and make judgments concerning an appropriate course of action (e.g. a nurse suspects a hemolytic reaction and immediately stops the blood transfusion when his patient develops a fever and hypotension). *Leadership* describes an ability to influence others and provide direction without imposed hierarchies (e.g. a surgeon guides the operating team through a complex case while creating an environment that encourages all members of the team to speak up to prevent error). *Communication describes the key skills needed to share information across power and professional boundaries, ensure clear messages are produced and using techniques to ensure understanding and teamwork* describe unity around shared goals, defined roles and clear information exchange.

Many may equate non-technical skills with the field of human factors. While Human Factors Ergonomics seeks to engineer entire systems to reduce errors, it recognizes that non-technical skills are a specific area which can be fostered in individuals to support safety. Indeed, as the whole field of human factors is situated in the context of enhancing safety within the many industries in which they were developed, the medical and medical education literature inescapably links the two. Lessons from other high reliability industries, such as aviation and the military, have been enthusiastically embraced within healthcare. For example,

Practice points

- Methods available in the assessment of non-technical skills include simulated clinical scenarios, OSCEs, questionnaires, and written assessments.
- The design of assessment methods should involve consideration of conceptual frameworks and theoretical models.
- Future research should primarily focus on assessment methodology, rather than non-technical skill instruction.
- Research presenting non-technical skill assessment should allow for replication through the provision of materials and tools used.
- Future research in this area should focus on the development of assessments mapped to learning outcomes and consensus descriptors of specific non-technical skills.

CONTACT Morris Gordon mgordon@uclan.ac.uk School of Medicine, University of Central Lancashire, HA340, Harrington Building, Preston, UK
Supplemental data for this article can be accessed here.

the discipline of human factors, environmental, organizational and job factors, and human and individual characteristics which influence behavior (Wet 2012), has been applied in aviation and the military to enhance the design of equipment, optimize the working environment and maximize performance (Catchpole 2013). This educational methodology has been adapted for use successfully in fields such as anesthesiology and shown to reduce error (Flin and Patey 2009).

It is interesting to ponder whether such skills should form the basis of good medical practice across the board, rather than just error avoidance. However, currently almost all works in medicine that discuss non-technical skills do so in a safety context (Gordon et al. 2012, 2015b) and as the only skills that can be cultivated in the individual to enhance to achieve such goals, this link is not surprising. What is surprising is that although two-thirds of United States (US) medical schools mention patient safety in coursework (Blumenthal 2010), only 25% describe curricula with explicit attention to safety skills-based training (Alper et al. 2009), but mention of non-technical skills conspicuous by its absence.

The World Health Organization (WHO) curriculum on patient safety for medical students (2009) aims to encourage and facilitate the teaching of skills-based patient safety topics to medical students, with a specific focus on team-based training, systems and error prevention (Walton et al. 2010), but again there is no explicit mention of non-technical skills. There are some indications that this deficit is being tackled. More recently, organizations such as the American Medical Association are championing change in undergraduate medical education (UME), with specific focus on teaching new content in health systems science, which now include mention of non-technical skills (Creating the Medical School of the Future 2017).

Non-technical skills should be delivered as part of the undergraduate core curriculum before professional attitudes are fully formed (Flin and Patey 2009). The failure to incorporate such training into UME may result in such topics being undervalued. Similarly, the absence of non-technical skills training in the postgraduate curriculum is a failed opportunity to provide repeated practice and to develop an integrated, longitudinal assessment strategy. Education and assessment of non-technical skills in UME learners may provide a pathway to achieving safer and more effective care.

A prior systematic review investigating non-technical skills educational interventions found that most were deployed with the expressed purpose to enhance patient safety. With regards to specific interventions, there was an identified a lack of scientific and theoretical rigor underpinning published teaching innovations (Gordon et al. 2012), however, synthesis allowed for some existing theoretical constructs to be identified. A psychological theory of egocentric heuristics (Chang et al. 2010) describes a tendency to overestimate how well communication has been understood. Agency theory (Cheung et al. 2010) describes how in shift based working a focus on task rather than individual patients can only be challenged by error wisdom. Finally, the theory of “coordination costs”, and theories concerning the diffusion of responsibility (Darley and Latané 1968) describe the role of systems, processes and technology in counteracting such problems as systems grow and become more complex. Further work led to the development of the SECTORS model of non-technical skills learning (Systems and technology use, Error awareness, Communication, Team-working, Observation and simulation and Risk assessment and Situational awareness) (Gordon 2013; Gordon et al. 2015a).

Educational interventions to enhance non-technical skills are necessary, but in isolation are not sufficient to advance the field within education or clinical practice. Such education must be coupled with rigorous assessments to both drive learning, and ensure competence. Unfortunately, non-technical skills are difficult to assess as they form part of a wider set of interconnected behaviors. Measuring the impact on patient outcomes also necessitates finding a way to assess the longitudinal impact of education.

A scoping review failed to reveal any systematic reviews investigating the assessment of non-technical skills within UME. Given the summative nature of many learning outcomes in the field, we feel such a review is vital to guide the inclusion of such assessments in high stakes summative examinations and to identify how they may be assessed in a methodologically robust manner. As such, we set out to systematically review the evidence regarding non-technical skills-based assessments in UME, to describe the overarching strategies utilized, learning outcomes addressed and the impact of these assessments, in terms of their validity, reliability, effect on performance and solutions to psychometric challenges. A focused review methodology has been used, defined as “a form of knowledge synthesis in which the components of the systematic process are applied to facilitate the analysis of a more focused research question” (Gordon et al. under review). The focused review still embraces the core principles of systematic methodology, as these are crucial to facilitate transparency and scholarly deployment. However, after scoping the project and identifying the close link to patient safety of such skills, it became clear that the research scope was narrow and suited this methodology.

Methods

No single research paradigm was used for this review. We embraced both positivism, through description and justification of the assessment methods used, and constructivism, through clarification of the underpinning theoretical frameworks that informed assessment choice (Gordon 2016). The manuscript was reported in accordance with the STORIES statement, publication standards for healthcare education evidence synthesis (Gordon and Gibbs 2014) and the focused review deployed in line with specific guidance (Gordon et al. under review).

Data collection

Scoping searches were performed to refine the search syntaxes and to clarify the inclusion and exclusion criteria for the review. We encountered two key problems during scoping: first, few papers described an assessment of non-technical skills as their primary focus. Second, few papers described an assessment tool at all. Thus, we broadened our initial searches and considered papers for inclusion if they either described an assessment as the main focus, or if they described an assessment as an outcome measure of an educational activity seeking to improve non-technical skills.

We embraced all study designs that targeted medical students, including when medical students participated within multidisciplinary teams and when the assessments formed a core or elective component of an undergraduate medical curriculum. Papers that described outcomes at all levels of Kirkpatrick's adapted hierarchy were eligible for inclusion. Studies from any country, published in any language were considered. Studies describing only teaching without an assessment, failed to describe outcomes or described outcomes related to teaching, but not assessment, and papers that gave opinions or reviews without the primary use of an assessment tool were excluded. We excluded studies focusing on the assessment of non-technical skills in the post-graduate populations because this landscape has already moved in many ways to integrate formative assessment of such skills in the field of simulation. We believed that the specific needs within the undergraduate landscape, in particular summative assessment of such skills, represented a distinct educational problem and context for this review (the search syntaxes, example search strategy, inclusion and exclusion criteria are summarized in Supplementary Appendix 1 and Supplementary Appendix 2).

The following online databases were searched from inception date of database up to January 2017 using a standardized search strategy: ERIC, PubMed, MEDLINE, EMBASE, CINAHL, Psychinfo, and Google Scholar. Abstracts available online from relevant education societies, including the Association for Medical Education in Europe (AMEE) and the Association for the Study of Medical Education (ASME) were also searched for the last three meetings to ensure any papers currently under review, but not fully published were included. Reference lists of the included studies were hand-searched for additional relevant studies.

Data analysis

Citations were screened independently by two authors, MG and RA. Abstracts considered potentially relevant for inclusion were independently reviewed by these same authors. Agreement was assessed using Cohen's kappa statistic. Full-text articles were then reviewed to determine whether all inclusion criteria were met. Any disputes at any stage of the data analysis process were resolved by consensus. When insufficient information on an assessment was provided to make a judgment, we attempted to contact the authors for further details. For the included studies, the full manuscripts were assessed independently using a data extraction form (Supplementary Appendix 3) by CGC and RA, with MG and DG ratifying the assessments.

The data extraction form and quality assessment tool (Supplementary Appendix 3) were produced utilizing guidance from Best Evidence Medical Education (BEME) (Hammick et al. 2010; PRISMA 2015; Reed et al. 2005.

The quality assessment of the included studies was broadly split into two main components: research methodology quality and reporting quality (Supplementary Appendix 3). The research methodology quality assessment was completed as a "yes/no" response to eight questions focusing on study objectives, study design, randomization, reporting of participant characteristics and description of the intervention. The reporting quality assessment included six items: description of underpinning theoretical models, description of the assessment process, the educational context, psychometric details, provision of materials allowing replication and the strength of the conclusions drawn. The first five of these items were scored on a three-point Likert scale, with the last item *strength of conclusions*, scored against a five-point Likert scale. The impacts of the interventions were classified in accordance with Kirkpatrick's adapted hierarchy (Bates 2004), in line with guidance provided by BEME (Hammick et al. 2010). A descriptive synthesis of all included studies was completed, summarizing key findings, with an assessment of quality indicators as listed above.

If data were provided that supported quantitative analysis, such as validity or reliability, this was completed using the Cochrane Revman software (2014). For continuous data, the standardized mean differences (SMDs) were compared. For discrete data, odds ratios (ORs) were used. For data regarding the theoretical underpinning, pedagogy and content of the assessments, *a posteriori* thematic analysis (Strauss and Corbin 1998) was planned in detail in the protocol, but as such data were ultimately not available, these details are not described.

Ethical approval was not sought for this review as it did not involve any direct participants.

Results

Search results

Initial searching of both databases and alternative sources yielded 12,180 records, leaving 10,060 citations after deduplication. After title screening, an additional 9463 citations were removed, leaving 597 abstracts to screen for eligibility. All abstracts were read by MG and RA, and 19 articles met the criteria for full-text assessment. The most frequent rationales for excluding studies at this stage included no assessment measure, a review article, letter, or editorial, or an exclusive focus on graduate or postgraduate level learners. Given the relative clarity of such judgments, these were clear from the abstract with no full text review needed for most papers to exclude at this stage. No studies were excluded on the grounds of publication in a non-English language. Agreement between the two reviewers on abstract screening was good (kappa statistic: 0.91).

Of the 19 articles undergoing independent full-text assessment for inclusion, 10 were excluded on the grounds of not meeting the full inclusion criteria (Anderson et al. 2009; Hall et al. 2010; Leung and Patil 2010; Robertson et al. 2010; Dudas et al. 2011; Stahl et al. 2011; Meier et al. 2012; Myung et al. 2012; Kiesewetter and Fischer 2015; Martinou et al. 2015). The reasons for all of the exclusions at this stage were related to not describing a non-technical skills assessment, but instead a limited outcome measure focused on verifying the education delivered within these primary studies. The key discriminating factor used to make the decision was whether the assessment had any potential utility as part of ongoing formative or summative assessment, outside of the report itself. Given the difficulty of such judgments, particularly as many of the included studies were similar, these were discussed amongst the whole team and consensus confirmed.

Nine papers were ultimately included in the qualitative and quantitative syntheses (Madigosky et al. 2006; Paxton and Rubinfeld 2010; Daud-Gallotti et al. 2011; Aboumatar et al. 2012; Müller et al. 2012; Ginsburg et al. 2014; Jansson et al. 2015; Thomas et al. 2015; Farnan et al. 2016). The search flowchart is shown in Figure 1 (2006–2016) and an overview of the included papers is shown in Table 1. Data were extracted independently by CGC and RA, who achieved concordance in 94% of quality ratings, with consensus reached on discussion (Supplementary Appendix 4).

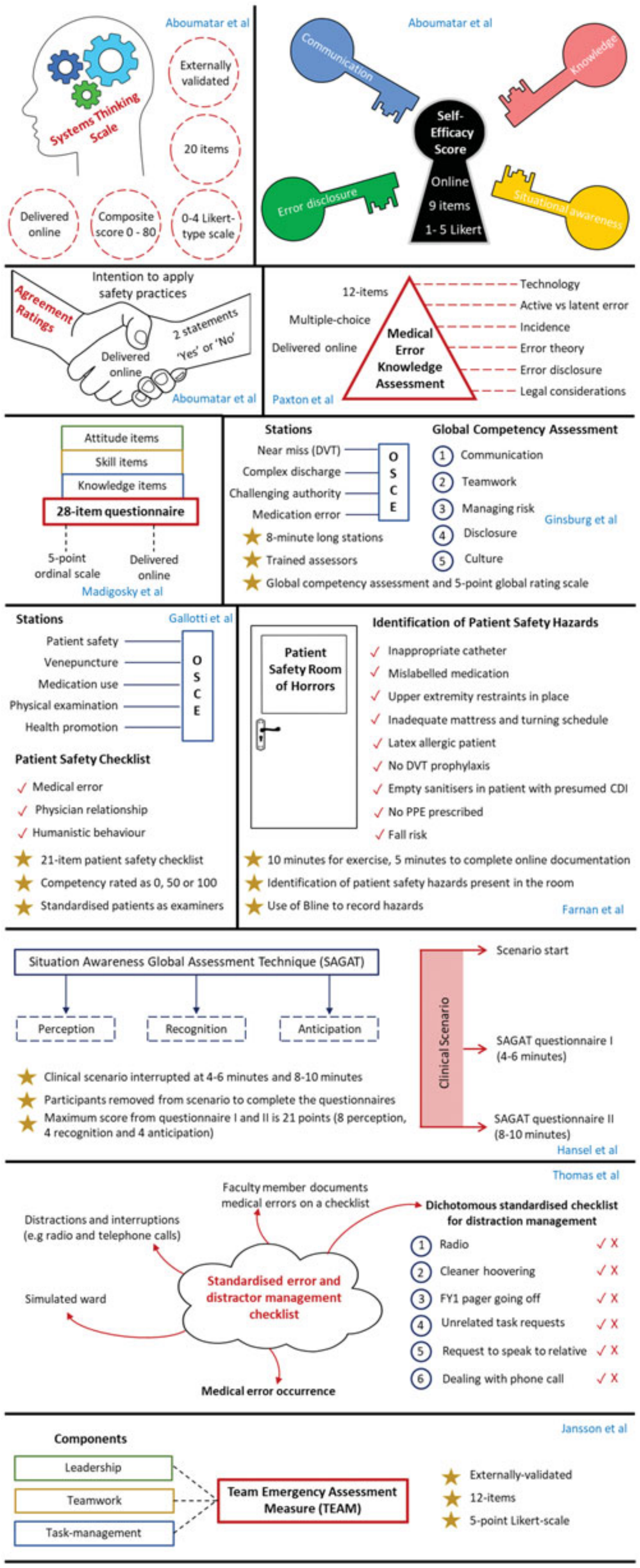

Figure 1. Analysis of the non-technical skill assessment methods and outcome measures.

Study participants

Of the nine studies, five of them were based in the US (Aboumatar et al. 2012; Madigosky et al. 2006; Paxton and Rubinfeld 2010; Jansson et al. 2015; Farnan et al. 2016), with a further one in each of the following; Canada (Ginsburg et al. 2014), Brazil (Daud-Gallotti et al. 2011), Germany (Müller et al. 2012), and the United Kingdom (Thomas et al. 2015). The average number of participants per study was 92 (range: 18–214). All but three studies focused entirely on medical students (Paxton and Rubinfeld 2010; Ginsburg et al. 2014; Farnan et al. 2016), with two studies covering second year medical students (Madigosky et al. 2006; Aboumatar et al. 2012), four studies focusing on third and fourth year medical students (Ginsburg et al. 2014; Jansson et al. 2015; Farnan et al. 2016), and the remaining three studies including fifth or sixth year medical students (Daud-Gallotti et al. 2011; Müller et al. 2012; Thomas et al. 2015). Two studies, in addition to undergraduate medical students, involved nursing students (Ginsburg et al. 2014) or physician associate students (Paxton and Rubinfeld 2010). There was minimal commonality between studies focused on participants at similar stages in medical education training, or between studies in the same country.

Quality assessment

From a methodological perspective, all studies bar one included a review of the literature (Jansson et al. 2015) and provided clearly defined objectives (Paxton and Rubinfeld 2010). All nine studies reported on and designed their study appropriately in response to the research question and provided learner characteristics. Three studies employed the use of control groups (Paxton and Rubinfeld 2010; Müller et al. 2012; Thomas et al. 2015), three utilized a form of randomization (Aboumatar et al. 2012; Müller et al. 2012; Farnan et al. 2016), and two studies described the educational intervention in enough detail as to allow for replication (Daud-Gallotti et al. 2011; Farnan et al. 2016) (Supplementary Appendix 4).

Four (Madigosky et al. 2006; Aboumatar et al. 2012; Jansson et al. 2015; Farnan et al. 2016) of the nine papers did not provide any descriptions of theoretical models or conceptual frameworks utilized for the non-technical skills assessments. A further three provided limited descriptions aligned with conceptual elements related to error wisdom and situational awareness (Paxton and Rubinfeld 2010; Müller et al. 2012; Ginsburg et al. 2014). The remaining two gave significant detail of the frameworks used, one fully aligning with theoretical principles (Daud-Gallotti et al. 2011) of non-technical skills (Ginsburg et al. 2014) and the other identifying key theories related to error origins, describing the "practice-effect" concept (Thomas et al. 2015).

Table 1. Characteristics of included studies summary.

Author	Year, location	Study type	Participants	Educational intervention	Non-technical skill assessment methods	Non-technical skills assessment outcome	Results: learner performance	Results: assessment performance	Resources	Conclusion	Average quality indicator[a] U	A	B	P	C	S	Level of outcome[b]	Strength of conclusion[c]
Farnan, Gaffney and Poston et al.	2015, USA	Observational	86 3rd year medical students and 128 clerkship year interns	Clinical scenario constructed highlighting safety hazards. For the medical students, this was a component of an OSCE, and for the interns this formed part of an introductory boot-camp.	Students (OSCE) and interns (boot camp) were tasked to review a drug chart and identify and document patient safety hazards. Students were timed and asked to identify and document as many safety hazards as possible.	Descriptive statistics (% hazards correct and mean correct hazards).	The interns correctly identified a mean of 5.1 hazards per participant. Comparing the medical students to the interns, the former correctly identified hazards more frequently in 6 of the 9 hazard groups.	Not evaluated.	Development of scenario and resources, nursing training model, construction of simulated inpatient room and proctor to supervise.	Both medical students and clerkship interns had difficulty identifying common hazards of hospitalization. Hazard identification was a feasible and effective way to introduce trainees to safety-focused content.							Level 2b	3/5
Jansson, An-Grogan and Eller et al	2015, USA	Cross-sectional	154 undergraduate 4th year medical students	Simulated-based cardiac dysrhythmia session. No further details given.	Students participated in high-fidelity simulation resuscitation scenarios with trained observers completing the Team Emergency Assessment Measure (TEAM) after viewing live/video-taped sessions.	Attitudes to Patient Safety Questionnaire (APSQ-III) Risk Management Foundation (RMF) Patient Safety Knowledge Test Team Emergency Assessment Measure (TEAM)	Student scored 8.38/14 (knowledge assessment), 5.42/10 (safety – APSQ-III), and 5.44/10 (team-work – TEAM)	Not evaluated.	Development of high-fidelity simulated resuscitation scenario, trained observers, and audio-visual recording software.	Results highlighted a lack of safety culture and experience regarding error disclosure and human factors within a cohort of fourth year medical students. Low knowledge scores and teamwork performance ratings highlight specific targets for curricular development.							Level 2b	3/5
Ginsburg, Tregunno and Smee et al	2014, Canada	Observational	18 3rd and 4th year undergraduate nursing and medical students	4 OSCE stations developed and appraised by field experts with a focus on patient safety. Each station reflected at least 4 of the competency areas in the Safety Competency Framework.	Each OSCE station was scored on 4/5 relevant patient safety competency dimensions (communication, teamwork, managing risk, recognize and response to adverse events and disclosure culture).	5-point global rating scale on patient safety competencies (0 = competence not demonstrated, 4 = above the level expected of a learner at this level).	Nursing students scored significantly lower than medical students on 3 or the 4 stations ($p < 0.05$).	Assessors exhibited excellent agreement (weighted κ scores ranged from 0.74 to 0.82 for the 4 OSCE stations)	1.5-day case writing workshop (7 staff members). Four trained assessors/facilitators and 4 evaluated OSCE stations.	Scenarios mapped to patient safety competencies were applicable to both nursing and medical students with highly reliable station scoring. Performance can be evaluated with high reliability, suggesting a single assessor per station would be sufficient.							Level 2b	4/5
Gallotti, Morinaga and Arlindo-Rodrigues et al	2011, Brazil	Observational	95 5th year undergraduate medical students	Patient safety program with a focus on human error theory, epidemiology of incidents, adverse events, and disclosure.	Students completed an OSCE with one of the stations focusing on patient safety issues including; medical error recognition/disclosure, the patient-physician relation and humanism issues.	Standardized patient safety checklist was completed by each standardized patient assessing the domains of medical error, patient physician relationship attitudes and humanistic behaviors. This was done	Global score for patient safety station was 87.59. Sub-score for medical error (77.96 ± 2.21) was significantly lower than the other 2 domains (both ∼90).	The patient safety curriculum and assessment were highly rated by students (94%), who recognized the important of learning, discussing, assessing, and receiving feedback on medical error issues.	Simulated patient safety program, standardized patient safety checklist and development/execution of an OSCE station.	An OSCE is a useful tool to evaluate patient safety competencies during medical student clerkship.							Level 4a	3/5

(*continued*)

Table 1. Continued.

Author	Year, location	Study type	Participants	Educational intervention	Non-technical skill assessment methods	Non-technical skills assessment outcome	Results: learner performance	Results: assessment performance	Resources	Conclusion	Average quality indicator[a] U	A	B	P	C	S	Level of outcome[b]	Strength of conclusion[c]
						through a 3-choice % score (with descriptive overview) and 5-point Likert-score (humanistic behavior domain only). These were converted into global and sub-domain scores.												
Madigosky, Headrick and Nelson et al	2006, USA	Cross-sectional	92 2nd year undergraduate medical students	A 10.5-hour patient safety and medical fallibility curriculum was developed aiming to address five main themes; patient safety overview, error reporting, system versus human approach, safety tools, and ethics, and disclosure.	Knowledge, skills and attitudes questionnaire pre-curriculum, post-curriculum, and 1-year post-curriculum. At 1-year students also responded about error reporting and disclosure experiences.	Mean change (95% CI) from pretest mean score to post-test mean score and mean change (95% CI) from pretest mean score to one-year post-test mean score.	8 items and knowledge score improved after the curriculum with only 7 maintained at 1 year. 7 items did not change and 5 changed in an undesired direction after the curriculum and/or after one year. At 1-year, 76% of students had observed an error, only 7% reported using electronic reporting system.	56% of students reported having used what they learned in the curriculum, with 76% stating they had observed a medical error. Of these students, 71% had disclosed an error to a fellow student, 56% had disclosed an error to a resident, 46% had disclosed an error to a faculty member and 7% had used the electronic reporting system.	10.5 hours of taught curriculum in the form of lectures, tutorials, panel-discussions, and computer-training.	The curriculum led to changes in knowledge, skills, and attitudes, but not all changes were sustained at one year.							Level 3	3/5
Aboumater, Thompson and Dawson et al	2012, USA	Before and after	119 2nd year undergraduate medical students	Patient safety curriculum utilizing case-based learning, small group activities, role play activities and simulations.	Pre-post intervention evaluation of student knowledge, awareness of safety problem, self-efficacy and system thinking. Post-intersession assessment of student intention to apply safety practices and satisfaction.	Primary outcome variables were system thinking, self-efficacy and safety knowledge. System thinking was measured using the validated System Thinking Scale (STS); 20 items on a 0–4 Likert-type scale. STS composite scores ranges from 0 to 80. Self-efficacy was measured using 9 statements scoring on a 1–5 Likert-type scale, and safety knowledge was measured using a 19-item test.	Student safety knowledge (mean score increase = 19%, $p < 0.01$), system thinking (mean score increase = 7.4, $p < 0.01$). Students had statistically significant increases in self-efficacy for all taught communication.	Intersession quality was rated as "excellent" or "very good" by 92% of responding students. The percentage of students rating sessions as "useful" or "very useful" ranged from 77% to 95% for the didactic sessions, from 74% to 100% for small group activities, and from 79% to 97% for simulation activities.	Development of a 6-hour patient safety curriculum.	The patient safety intersession resulted in increased knowledge, system-based thinking, and self-efficacy for all taught communication and safety skills.							Level 2b	3/5
Hansel, Winkelmann and Hardt et al	2012, Germany	Prospective cohort	61 6th year undergraduate medical students	Crew Resource Management (CRM) course on situational	All students took part in a pre-and post-intervention simulated	Situational Awareness Global Assessment Technique	SAGAT score increased from 10.6 ± 2.3 (pre-intervention) to	In SIM Group (20 questionnaire responses) – 12 rated the course	Simulated intensive care room, resuscitation cart, SimMan,	Neither the 1.5-day CRM course or the 1.5-day simulator training							Level 3	3/5

(*continued*)

Table 1. Continued.

Author	Year, location	Study type	Participants	Educational intervention	Non-technical skill assessment methods	Non-technical skills assessment outcome	Results: learner performance	Results: assessment performance	Resources	Conclusion	Average quality indicator[a] U	A	B	P	C	S	Level of outcome[b]	Strength of conclusion[c]
				awareness, with a focus on adopting strategies to improve situation awareness in crisis scenarios or simulation-based training on sepsis resuscitation. The control group did not obtain any training).	scenario on septic shock. The students' situation awareness and their medical performance were evaluated. During the 10-minute scenario, the students were taken from the simulation room 3 times to complete a questionnaire (10, 11, and 13 items).	(SAGAT). Level 1 questions referred to perception (13 items), level 2 questions referred to recognition (4 items) and level 3 questions referred to assessing anticipation (4 items). For each scenario, a maximum sum score of 21 points could be achieved.	11.9 ± 1.7 (post-intervention) ($p = 0.04$) in the SIM group. No significant changes were observed in the CRM intervention group or the control group. The clinical performance scores post-intervention did not differ from pre-intervention scores.	as "excellent" and 7 rated it as "good". CRM group received similar overall marks.	ventilator and defibrillator.	influenced the clinical performance. SAGAT scores were higher after simulator training, not after CRM training.	yellow	yellow	yellow	yellow	green	yellow		
Paxton and Rubinfeld	2010, USA	Prospective cohort	51 undergraduate students (46 medical students and 5 physician associate students)	Medical errors education session (45-slide presentation) covering 6 major medical errors subjects: terminology, active versus latent errors, incidence, error theory, error disclosure and legal considerations.	12-question pre- and post-intervention test. Long-term post-tests were performed 1–12 months later. The questions were in multiple-choice format with a brief question stem and 4 foils.	Mean pretest correct (%), mean post-test correct (%) and long-term mean post-test correct (%).	Pre-test mean = 29.3% ($p < 0.01$). Post-test mean = 73.7% ($p < 0.01$). Long-term post-test = 49.1% ($p < 0.01$).	Control group of 24 3rd year medical students completed the testing instrument. These students did not participate in the educational session, but were re-tested after 6 months. No statistical difference was found between the 6-moth scores and their original test scores.	45-slide presentation covering 6 major medical error subjects with a facilitator to guide small-group discussions.	This brief educational intervention led to statistically significant improved performance in general understanding of medical errors.	yellow	yellow	yellow	red	yellow	yellow	Level 2b	3/5
Thomas, Nicol and Regan et al	2014, UK	Prospective cohort	28 final year undergraduate medical students	Immediate feedback on distractor management and error as intervention following baseline ward round. The individualized feedback was targeted on distractor management and its relation to medical error.	Students participated in a post-intervention simulated ward round where a standardized checklist was used to document student performance. The checklist was completed by a trained member of the simulated ward round faculty.	Number of errors committed at each simulated ward round against a standardized checklist (baseline ward round and post-intervention ward round).	Pre-intervention errors = 72. Post-intervention errors = 17 (reduction of 76.4%). Control group baseline errors = 76. Control group simulation (4-weeks)=44 (reduction of 42.1%).	Although medical error rates fell in both groups, this intervention reduced error rates by an additional 34.3% over control (p value = 0.0035). With a sample size of 15, a significance level of 0.05, and an SD of 1.03, the study power was calculated to be over 80%.	2 × 30-minute simulated ward rounds with a focus on distraction, interruption and error.	Simulated ward round training with targeted feedback has the capability to improve patient safety behaviors significantly.	green	green	yellow	yellow	green	yellow	Level 2b	3/5

[a]Average quality indicator (U: underpinning framework; A: assessment method; B: background; P: psychometrics; S: content; C: strength of conclusions).
[b]Level of outcome (level 1: participation; level 2a: modification; level 2b: modification of knowledge/skills; level 3: behavioral change; level 4a: change in organizational practice; level 4b: benefits to patients/clients).
[c]Strength of conclusions (1: no clear conclusions can be drawn. Not significant; 2: results ambiguous, but there appears to be a trend; 3: conclusions can probably be based on the results; 4: results are clear and very likely to be true; 5, results are unequivocal).
Green or light gray means high quality, yellow or more grey means unclear quality and most grey or red means low quality.

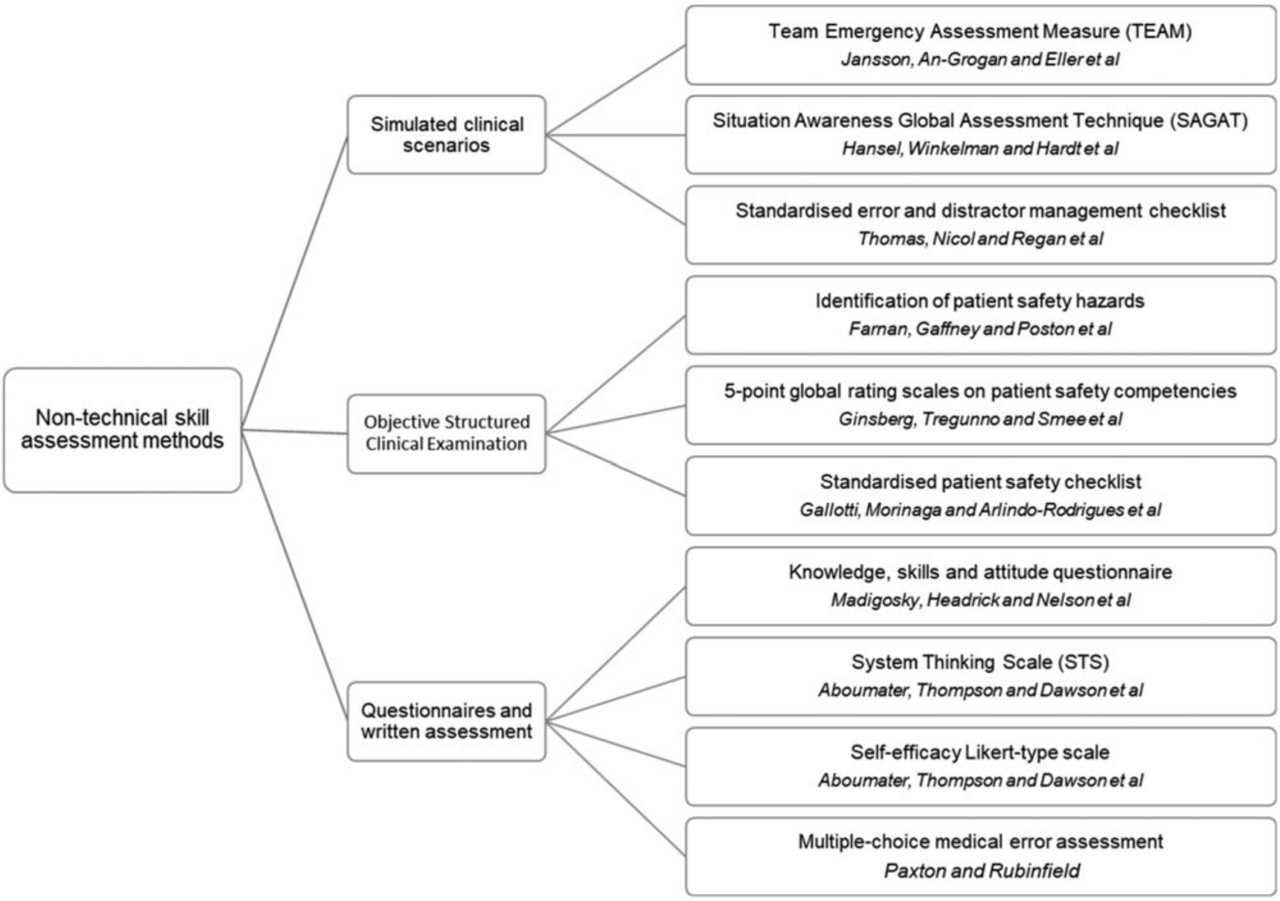

Figure 2. Assessment of non-technical skills.

Only two of the nine studies provided details of materials used, such as mark sheets, in sufficient detail as to allow replication (Müller et al. 2012; Thomas et al. 2015). The strength of conclusions estimated by employing the BEME strength of findings scale (Hammick et al. 2010), revealed eight studies scoring 3/5, suggesting that their conclusions were most likely based on the results (Madigosky et al. 2006; Paxton and Rubinfeld 2010; Daud-Gallotti et al. 2011; Aboumatar et al. 2012; Müller et al. 2012; Jansson et al. 2015; Thomas et al. 2015; Farnan et al. 2016). Only one study achieved a score of 4/5, suggesting the conclusions are clear and very likely to be true (Ginsburg et al. 2014). Over half of the studies (five of nine), provided a clear description of the process and outcomes of the assessment (Madigosky et al. 2006; Daud-Gallotti et al. 2011; Ginsburg et al. 2014; Thomas et al. 2015; Farnan et al. 2016).

Assessment tools

The wide variation of educational interventions and non-technical skills assessed meant that a diverse range of assessment methods were utilized. The specific non-technical skills assessed varied, which is understandable given the lack of explicit recognition of what constitutes such skills until recently that has been previously noted, but attempts were made to consider how such skills fit within the more recent published skill sets. Described skills included situational awareness, distraction management and managing risk (seen as a subset of risk assessment), teamwork and maintaining interprofessional relationships (which also included elements of communication), hazard identification (situational awareness), system thinking (clinical decision making), humanistic behavior, self-efficacy and workplace attitudes (all elements of leadership) (Table 2). The variation in interventions and skills assessed was from more than just nomenclature and represented a primary source of significant educational heterogeneity among the studies. Broadly, assessments fell into three categories – simulated clinical scenarios, objective structured clinical examinations (OSCEs), and questionnaires or written assessments (Figure 2). An overview of the methods used in each paper is depicted in Figure 2.

Three studies employed simulated clinical scenarios as an assessment method with a variety of outcome measures. These included the Situation Awareness Global Assessment Technique (SAGAT) questionnaire (Müller et al. 2012), the standardized error and distractor management checklist (Thomas et al. 2015), and the Team Emergency Assessment Measure (TEAM) (Jansson et al. 2015) (see Table 2 for details of simulated clinical scenario assessment methodologies). These assessment modalities tended to assess a limited number of non-technical skills, which may limit their utility in UME. However, the tools did to have good validity evidence for assessing specific skills. Thomas et al. (2015) were able to demonstrate the ability of simulated clinical scenarios to assess key behaviors through a standardized checklist, with a statistically significant reduction in medical error rates in the intervention group in practice.

Three studies utilized OSCEs. One study assessed learners' identification of patient safety hazards in a patient safety room of horrors (Farnan et al. 2015). This study aligned with error awareness as an underpinning element of non-technical skills (Gordon 2013). Another study evaluated patient safety competencies on a five-point global

Table 2. Use of simulated clinical scenarios to assess non-technical skills.

Assessment methodology	Non-technical skill assessed	Summary of assessment methodology	Example items	Use in non-technical skills assessment	Overview of scenario	Administration of assessment
Situation Awareness Global Assessment Technique (SAGAT) (Müller et al. 2012)	Situation awareness	SAGAT is an objective measure of a students' situational awareness during a simulated scenario. Each item assesses situation awareness with a focus on perception, recognition or anticipation	What is the current oxygen saturation? (perception) What is the junior doctor doing? (recognition) How old is the patient? (perception) What is the number of white blood cells in the medical examination report? (perception) How will the blood pressure change in the next minute? (anticipation) How will the central venous pressure change in the next minute? (anticipation)	SAGAT questionnaire I and II were used at 4–6 minutes and 8–10 minutes, respectively, during a 10 minute clinical scenario. Performance was assessed using the SAGAT during a pre- and post-intervention clinical scenario.	47-year-old male patient was simulated who had just been transferred to ITU. The patient was in septic shock with suspicion of valve endocarditis one-year post-valve replacement (pre-intervention scenario). 29-year-old female patient was simulated who was suffering from septic shock due to post-partum sepsis (post-intervention scenario).	The scenarios were interrupted between 4 and 6 minutes and between 8 and 10 minutes. Participants were taken out of the simulated environment and had to complete a questionnaire with 10 items (first interruption) and 11 items (second interruption).
Standardized error and distractor management checklist. (Thomas et al. 2015)	Managing distraction and interruption	Dichotomous standardized checklist measuring the number of errors made, number of distractions and interruptions managed.	Did not correctly prioritize patients and order in which to be seen. Checks antibiotic vial as satisfactory when it has actually passed its expiry date. Does not check the identity of the relative wishing to speak to them gives information pertaining to a different patient. Identifies (and deals with) radio playing in the background. Identifies (and deals with) FY1 pager going off. Deal with request to prescribe Paracetamol for a separate unrelated patient.	Standardized checklist used to document student performance at both baseline and post-intervention (targeted feedback).	Participants acted as foundation doctors in 30-minute simulated ward round. Patient 1 had sepsis. Patient 2 had a post-operative myocardial events. Patient 3 was confused. Facilitators deployed six realistic, time-critical distractions and interruptions (radio, hoovers, additional prescription tasks, phone calls and upset relatives). Each patient had a number of expected tasks, with associated potential errors (e.g. wrong drug doses, incorrect CURB-65 calculation, incorrect patient identity).	1 month between the pre-intervention and post-intervention assessment (via the standardized checklist). Checklist completed by a trained member of the simulated ward round faculty.
Team Emergency Assessment Measure (TEAM) (Jansson et al. 2015)	Teamwork	TEAM scoring is an externally validated tool for assessing teamwork, under the domains of leadership, teamwork and task management. The 12 items (11 specific and 1 global rating) are rated using a five-point scale.	The team leader let the team know what was expected of them through direction and command (leadership). The team communicated effectively (teamwork). The team acted with composure and control (teamwork). The team prioritized (task-management) The team followed approved standards and guidelines (task-management)	TEAM assessment tool was used to assess student performance following high-fidelity simulated resuscitation scenarios.	Simulated-based cardiac dysrhythmia session. No further information available.	Trained observers completed the TEAM assessment tool after viewing live or video-taped sessions.

Table 3. Use of objective structured clinical examinations (OSCEs) to assess non-technical skills.

Assessment methodology	Non-technical skill assessed	Summary of assessment methodology	Example items	Use in non-technical skills assessment	Overview of scenario	Administration of assessment
Identification of patient safety hazards (Farnan et al. 2016)	Patient safety hazard identification	9 simulated patient safety hazards. Items developed based upon discussion amongst local leaders in patient safety and strategic priorities for local hospital environments.	Inappropriate catheterization. Mislabeled medication. Restraints in place. No prophylaxis written in medication list. No personal protective equipment provided in a patient with C. diff infection. Patient with a clinical history of delirium, inappropriate bed height and catheter generating fall risk.	Participants faced the patient safety room of horrors. Participants were asked to identify safety hazards following the 10-min exercise. The participants were given 5 minutes after the simulation to complete the assessment task.	"Room of Horrors" clinical scenario which highlighted specific safety hazards and a mock patient chart were written, including clinical information about case, allergy status, medication list and a mock sign-out for the learner.	Participants given 5 min after the simulation to complete the online form documenting hazards identified. Students given a clipboard to note the hazards as they proceeded through the simulation. Students used Bline (simulation and clinical skills video and evaluation software) or ScanTron sheets to record hazards.
5-point global rating scales on patient safety competencies (Ginsburg et al. 2014)	Patient safety competencies (communication, teamwork, managing risk, disclosure and culture)	OSCE station scored on four or five patient safety competency dimensions (communication, teamwork, managing risk, disclosure and culture). A single 5-point global rating scale was developed.	No information available.	Participants were scored using the rating scales during a four-station OSCE, focused on the socio-cultural aspects of patient safety.	Station 1 required learners to uncover a deep vein thrombosis near miss and explain the factors to a spouse. Station 2 involved team dynamics and communication with a patient around a complex discharge. Station 3 required learners to persist in an interaction with a dismissive, time-pressured staff physician. Station 4 required learners to discuss an insulin overdose with the patient, including how it happened. Stations 8 min long.	Scoring in each station was done by two assessors who were both in the exam room. Each assessor was a current or retired faculty member from nursing or medicine. All assessors participated in a one-hour training session prior to the OSCE.
Patient safety checklist (Daud-Gallotti et al. 2011)	Medical error, interprofessional relationship and humanistic behavior	Checklist contains 21-items in 3 domains (medical error, patient-physician relationship, humanistic behavior). Competency in the medical error and patient-physician relationship domains were rated as follows: non-existent = 0 points, present but insufficient = 50 points and present and adequate = 100 points. Competency in the humanistic behavior domain was evaluated with 5-point Likert scales.	Did the student tell you that a preventable adverse event occurred during your admission? (medical error recognition); Did the student identify preventative actions to avoid this error? (medical error prevention); Did the student introduce self to you before the interview? (verbal communication); Did the student respect your silence? (patient-centered care); Did you feel supported in your distress? (support); Did the student respect your rights and values? (respect)	Assessor scored the students using the patient safety checklist during each of the five OSCE stations.	Station 1: A 70-yo man with renal insufficiency and lumbar pain received an anti-inflammatory prescription during his hospitalization. His renal function progressively deteriorated, and on day 4, dialysis was indicated. Task: explain to the daughter what happened to her father; Station 2: a 50-yo female with pneumonia was admitted to the ward overnight. In the same room was another person with a similar name who also had diabetes. On the following day, patient 3 received insulin instead of patient 2 and presented with confusion due to hypoglycemia. She received glucose and recovered completely.	The standardized patients completed one checklist for each student during a 3-minute break between the exit of the previous student and the entrance of the next student.

Table 4. Use of questionnaires and written assessments to assess non-technical skills.

Assessment methodology	Non-technical skill assessed	Summary of assessment methodology	Example items	Use in non-technical skills assessment	Administration of assessment
System-thinking scale (STS) (Aboumatar et al. 2012)	System thinking	Externally validated scale used to measure system thinking. It is composed on 20 items on a 0–4 Likert-type scale. The composite score ranges from 0 to 80.	Item 2: "I look beyond a specific event to determine the cause of the problem". Item 10: "I propose solutions that affect the work environment, not specific individuals" Item 14: "I think small changes can produce important results".	Used to assess system thinking before and after a six-hour patient safety curriculum.	Online
Self-efficacy Likert-type scale (Aboumatar et al. 2012)	Self-efficacy	Nine "I know how to" statements on a 1–5 Likert-type scale.	"I know how to disclose a medical error" "I know how to use personal protective equipment such as gowns, gloves and masks" "I know how to use teach back"	Used to assess self-efficacy before and after a six-hour patient safety curriculum.	Online
Safety knowledge assessment (Aboumatar et al. 2012)	Safety	19-item safety test.	No data available.	Used to assess safety knowledge before and after a six-hour patient safety curriculum.	Online
Agreement statements (Aboumatar et al. 2012)	Intention to apply safety practices	Intention to apply safety practices is assessed via agreement ratings to two statements.	Statement 1: "I will speak up about any safety concerns I have about my patients" Statement 2: "I plan to use the Teach Back method to ensure that my patients understood my instructions".	Used to assess the intention to apply safety practices following before and after a six-hour patient safety curriculum.	Online
Medical error knowledge assessment (Paxton and Rubinfeld 2010)	Medical error knowledge	12-question knowledge assessment in a multiple-choice format. Questions were under the domains of: terminology, active versus latent error, incidence, error theory, error disclosure and legal considerations.	Question 1: "define adverse event" Question 10: "define malpractice" Question 12: "identify high-risk specialities"	Used to assess medical error knowledge before and after a two-hour teaching intervention. Additionally, the assessment was used to assess long-term outcomes (up to 12 months)	Online
28-item questionnaire (Madigosky et al. 2006)	Knowledge, skills and attitudes	28-item questionnaire. 5 items assessed knowledge, 5 items measured skills, and 18 items assessed attitudes. All items were measured on a 5-point ordinal scale.	Attitude item: "Competent physicians do not make medical errors that lead to patient harm" Skill item: "Supporting and advising a peer who must decide how to respond to an error" Knowledge item: no data.	Used to assess knowledge, skill and attitudes before and after a 10.5-hour patient safety curriculum. Additionally, the assessment was used to assess long-term outcomes (12 months)	Online

rating scale (Ginsburg et al. 2014) and a third study utilized a patient safety checklist (Daud-Gallotti et al. 2011) (see Table 3 for details of OSCE assessment methodologies). OSCEs appeared better able to assess multiple non-technical skills compared to simulated clinical scenarios.

Three studies utilized questionnaires and written assessment methodologies with a variety of assessment outcomes and little overlap. These included a validated Systems-Thinking Scale (STS) (Aboumatar et al. 2012), self-efficacy Likert scales (Aboumatar et al. 2012), and multiple-choice medical error assessments (Paxton and Rubinfeld 2010) (see Table 4 for details of questionnaires and written assessments). All competency performances improved immediately after assessment when compared to the pre-intervention score across all three studies. Paxton and Rubinfeld (2010) and Madigosky et al. (2006) performed long-term post-intervention testing in addition to immediate pre- and post-intervention testing, with only Paxton and Rubinfeld demonstrating a statistically significant improvement in medical error competence after one year, when compared to the control group. Madigosky et al. (2006) reported that a number of competency scores changed in an undesired direction one year after the intervention.

Learning outcomes

The level of impact for assessments used in the majority of included studies (six of nine), sat at level 2b of Kirkpatrick's hierarchy, which correlates to the modification of knowledge and skills (Paxton and Rubinfeld 2010; Aboumatar et al. 2012; Ginsburg et al. 2014; Jansson et al. 2015; Thomas et al. 2015; Farnan et al. 2016). A further two studies impacted behavior (level 3) (Madigosky et al. 2006; Müller et al. 2012), and one study impacted at the organization level, causing change in practice (level 4a) (Daud-Gallotti et al. 2011). We could not complete a meta-analysis as there was significant assessment methodology heterogeneity and a lack of presentation of appropriate data.

Validity and reliability

No study determined the validity of the assessment tools used to assess non-technical skill performance, with the exception of the STS questionnaire (Aboumatar et al. 2012). As discussed previously, no quantitative analysis was performed relating to validity and reliability as no data was provided to allow this.

Discussion

This review identified a small number of studies that describe methods of assessing non-technical skills. As consensus on key elements that form non-technical skills has only been achieved in the past few years (Gordon and Gibbs 2014) and even more recently such key elements adopted explicitly in any policy document on UME (Creating the Medical School of the Future 2017), there was pervasive heterogeneity in the skills described that limits its synthesis or even useful comparison. The review process was further hampered as the studies universally focused on reporting non-technical skills instruction as their primary goal, and assessment was viewed as a symbiotic component of this, allowing local programs to be verified or outcomes in practice to be tracked. While included studies were only those judged to present assessment with the potential to form a continuing part of formative or summative assessment, the majority failed to provide the conceptual frameworks or theoretical models underpinning the choice of assessment. Psychometrics were almost completely absent and there were limited attempts to describe validity and reliability.

Assessment methods fell into three main categories: simulated clinical scenarios, OSCEs and questionnaire or written assessments. The methodological quality of the studies was of a reasonably good standard from a research perspective, but poor when considering the quality of reporting and evaluating the assessments themselves (Table 1). Despite many of the studies failing to evaluate the performance of the assessment method, a number of comparisons can be made across the different assessment modalities. First, studies utilizing simulated clinical scenarios as an assessment method were more likely to use validated and reliable assessment tools. For example, the TEAM tool has a substantial body of normative data confirming its validity (Cooper and Cant 2014). However, it only assesses a small component of the non-technical skill spectrum, and thus would only be called upon as an assessment measure of teamwork behaviors. The operational requirements to implement simulated clinical scenarios are significant and should be a major consideration when developing non-technical skill curricula and assessments, but similar to other areas of UME, assessment tools must be fit for purpose.

Our review leads us to conclude that there is a limited pool of non-technical skills assessment tools with an evidence-base supporting their use. The implications of this are that educators are increasingly dependent on developing their own assessment tools, with very little opportunity for reliability and validity testing. A single study demonstrated the reliability of assessing non-technical skill performance through an OSCE, suggesting a single assessor per station is sufficient, but generalization from this study is difficult. Given the unifying need for outcomes in non-technical skills to be met by learners (Flin and Patey 2009; Creating the Medical School of the Future 2017), not having similar evidence-based assessments is a major barrier to the field moving forward. There is not sufficient evidence to even suggest a specific model or conceptual framework to underpin assessment methods as given this lack of evidence, we would propose alignment with models that guide teaching of non-technical skills may be a best estimate (Gordon 2013).

In considering these findings, readers must also consider some key limitations. First, while we aligned with a consensus definition of non-technical skills (Gordon et al. 2015a, 2015b), confusion persists around these terms and how they are applied. The search went through several scoping revisions and the kappa statistic demonstrates that within the team a clear consensus understanding was reached. However, the lack of consensus in the wider literature may have led to the author team's definition not fully matching the wider research or teaching body.

Second, many of the studies primarily sought to report an intervention, with assessment devised to produce outcome measures. The inclusion of such papers is innately of lower quality from an assessment perspective. This is because researchers wishing to demonstrate effectiveness of teaching using a framework such as Kirkpatrick's hierarchy are not required to consider key issues for those designing formative and summative assessments for ongoing use (such as reliability, validity, particularly in the high stakes setting, cost and practicality, external scrutiny from university bodies and regulators). While we feel these are relevant when interpreted in that context, this does limit the utility of the findings and those wishing to employ them in a formal manner to formatively and summatively assess some outcomes would require many more aspects of the evidence to be presented to achieve this. Finally, as the goal of such outcomes is to enhance safety for patients, it is worth noting that all but one study failed to investigate patient outcomes (Daud-Gallotti et al. 2011). Subsequently, "validity" has essentially not been demonstrated. "Validity" would refer to how the assessments measure the non-technical skills in practice and their outcomes for patients.

While it is disappointing that such limited evidence has been identified, this review highlights that further work in this area is vital. First, future work must ensure that assessments align with truly ensuring outcomes that matter for care. This will allow validity in all its forms to be established, but this is challenging. Assessing the impact of education on practice is elusive, requiring as it does, the investigation of outcomes in clinical settings affected by multiple variables. This challenge is relatively common in health and medical education where the changes influencing patient outcomes may be cognitive and behavioral. It is crucial that assessment is sufficiently agile and robust to identify measurable elements which impact patients.

Current assessment modes traditionally favor models assessed close to training source and existing knowledge on more longitudinal and nuanced assessment of impact is limited. However, the nature and background of non-technical skills training provides the setting and stimulus to develop assessment modes which address these complex, problematic areas.

Second, quality elements of assessment research must be considered in detail, including clear mapping across local, national and international outcomes and psychometrics once used in practice. Educators may use the descriptive synthesis included in this review to support their current decision making, but assessments themselves must be the focus of future work, rather than a secondary component of research to implement non-technical skills education. This is key to ensure assessment has utility for both summative and formative assessment as part of ongoing UME.

Finally, consideration of the postgraduate setting is needed. This may be helpful first through a similar review piece as our scoping found no such review and then possibly through related future research. In achieving this, the issue of nomenclature is key. Using consensus terms for key skills in a consistent manner aids replication and dissemination and is key moving forward.

Conclusions

Simulated clinical scenarios, OSCEs and questionnaires have all been employed as methods to assess non-technical skills in UME. A multitude of assessment tools are used within these, including checklists, global rating scales and multiple-choice assessments and it is worth noting that the OSCE was most able to address multiple skill assessment. As studies typically report non-technical skills instruction as the primary goal, often the assessment methods described are not grounded in conceptual frameworks and key theoretical models, with further lack of clarity related to the use of variable language to describe specific skills. Educators are currently still dependent on developing their own assessment tools. This represents a major barrier to the field going forward, as such, there is a major requirement to develop and test evidence-based assessments of non-technical skills. There is an urgent need for research that focuses on developing non-technical skills assessments mapped to learning outcomes and consensus descriptions of specific skills. They must be evaluated in terms of validity and reliability, addressing psychometric properties, as our review questions related to these areas cannot be adequately answered at present. All such research must be reported in a manner that supports dissemination to advance the field.

Disclosure statement

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the article.

Notes on contributors

Morris Gordon, MBChB, PHD, MMed, is a Professor of Evidence Synthesis and Systematic Review, University of Central Lancashire, Preston, UK.

Jeanne Farnan, MD, MHPE, is an Assistant Dean Curricular Development and Evaluation, and an Associate Professor, Section of Hospital Medicine at the University of Chicago Pritzker School of Medicine in Chicago, Illinois.

Ciaran Grafton-Clarke is an Academic FY1 Doctor and Honorary Fellow at the University of Leicester, School of Life Sciences, UK.

Ridwaan Ahmed, MD, is a General Practitioner in Lancashire UK, and Lecturer in Public Health Sciences at the University of Central Lancashire, UK.

Dawne Gurbutt, PhD, is an Acting Director in the Centre for Excellence in Learning and Teaching at University of Central Lancashire, UK.

John McLachlan, PhD, is Professor of Medical Education in the School of Medicine, University of Central Lancashire, UK.

Michelle Daniel, MD, MHPE, is an Assistant Dean for Curriculum, and an Assistant Professor of Emergency Medicine and Learning Health Sciences at the University of Michigan Medical School in Ann Arbor, Michigan.

ORCID

Morris Gordon http://orcid.org/0000-0002-1216-5158
Jeanne Farnan http://orcid.org/0000-0002-1138-9416
Michelle Daniel http://orcid.org/0000-0001-8961-7119

References

Aboumatar HJ, Thompson D, Wu A, Dawson P, Colbert J, Marsteller J, Kent P, Lubomski LH, Paine L, Pronovost P. 2012. Development and evaluation of a 3-day patient safety curriculum to advance knowledge, self-efficacy and system thinking among medical students. BMJ Qual Saf. 21:416–422.

Alper E, Rosenberg EI, O'Brien KE, Fischer M, Durning SJ. 2009. Patient safety education at US and Canadian medical schools: results from the 2006 Clerkship Directors in Internal Medicine survey. Acad Med. 84:1672–1676.

Anderson E, Thorpe L, Heney D, Petersen S. 2009. Medical students benefit from learning about patient safety in an interprofessional team. Med Educ. 43:542–552.

Baldwin PJ, Paisley AM, Brown SP. 1999. Consultant surgeons' opinion of the skills required of basic surgical trainees. Br J Surg. 86: 1078–1082.

Bates R. 2004. A critical analysis of evaluation practice: the Kirkpatrick model and the principle of beneficence. Eval Prog Plann. 27: 341–347.

Blumenthal D. 2010. Patient safety: conversation to curriculum. New York Times. [accessed 2017 Oct 9]. http://www.nytimes.com/2010/01/26/health/26error.html?_r=1.

Catchpole K. 2013. Spreading human factors expertise in healthcare: untangling the knots in people and systems. BMJ Saf Qual. 22: 802–808.

Chang VY, Arora VM, Lev-Ari S, D'Arcy M, Keysar B. 2010. Interns overestimate the effectiveness of their hand-off communication. Pediatrics. 125:491–496.

Cheung DS, Kelly JJ, Beach C, Berkeley RP, Bitterman RA, Broida RI, Dalsey WC, Farley HL, Fuller DC, Garvey DJ, et al. 2010. Improving handoffs in the emergency department. Ann Emerg Med. 55: 171–180.

Cochrane Collaboration. 2014. Review manager (RevMan) [Computer program]. Version 5.3. Copenhagen: The Nordic Cochrane Centre.

Cooper S, Cant R. 2014. Measuring non-technical skills of medical emergency teams: an update on the validity and reliability of the team emergency assessment measure (TEAM). Resuscitation. 85: 31–33.

Creating the Medical School of the Future. 2017. AMA accelerating change in medical education; [accessed 2017 Oct 9]. https://www.ama-assn.org/innovations-outcomes-consortium.

Darley J, Latané B. 1968. Bystander intervention in emergencies: diffusion of responsibility. J Pers Soc Psychol. 8:377–383.

Daud-Gallotti RM, Morinaga CV, Arlindo-Rodrigues M, Velasco IT, Martins MA, Tiberio IC. 2011. A new method for the assessment of patient safety competencies during a medical school clerkship using an objective structured clinical examination. Clinics (Sao Paulo). 66:1209–1215.

Dudas R, Bundy D, Miller M, Barone M. 2011. Can teaching medical students to investigate medication errors change their attitudes towards patient safety? Qual Saf Health Care. 20:319–325.

Farnan JM, Gaffney S, Poston JT, Slawinski K, Cappaert M, Kamin B, Arora VM. 2016. Patient safety room of horrors: a novel method to assess medical students and entering residents' ability to identify hazards of hospitalisation. BMJ Qual Saf. 25:153–158.

Flin R, Patey R. 2009. Improving patient safety through training in non-technical skills. BMJ. 339:985–986.

Ginsburg L, Tregunno D, Norton P. 2014. Development and testing of an objective structured clinical exam (OSCE) to assess socio-cultural dimensions of patient safety competency. BMJ Qual Saf. 24:1–7.

Gordon M, Baker P, Catchpole K, Darbyshire D, Schocken D. 2015a. Devising a consensus definition and framework for non-technical skills in healthcare to support educational design: a modified Delphi study. Med Teach. 37:572–577.

Gordon M, Box H, Farrell M, Stewrt A. 2015b. Non-technical skills learning in healthcare through simulation education: integrating the SECTORS learning model and complexity theory. BMJ Stel. 1:67–70.

Gordon M, Darbyshir e. D, Baker P. 2012. Non-technical skills training to enhance patient safety: a systematic review. Med Educ. 46: 1042–1054.

Gordon M, Gibbs T. 2014. STORIES statement: publication standards for healthcare education evidence synthesis. BMC Med. 12:143.

Gordon M, Grafton-Clarke C, Hill E, Gurbutt D, Patricio M, Daniel M. Under review. Twelve tips for undertaking a focussed systematic review in medical education. Med Teach. doi: 10.1080/0142159X.2018.1513642.

Gordon M. 2013. Building a theoretically grounded model to support the design of effective non-technical skills training in healthcare: the SECTORS model. J Contemp Med Educ. 1:77.

Gordon M. 2016. Are we talking the same paradigm? Considering methodological choices in health education systematic review. Med Teach. 38:746–750.

Hall LW, Scott SD, Cox KR, Gosbee JW, Boshard BJ, Moylan K, Dellsperger KC. 2010. Effectiveness of patient safety training in equipping medical students to recognise safety hazards and propose robust interventions. Qual Saf Health Care. 19:3–8.

Hammick M, Dornan T, Steinert Y. 2010. Conducting a best evidence systematic review. Part 1: from idea to data coding. BEME Guide No. 13. Med Teach. 32:3–15.

Jansson PS, An-Grogan Y, Eller SG, Woods DM, Kontrick AV, Salzman DH. 2015. A needs assessment in patient safety education for fourth-year medical students. Am J Med Qual. 30:601.

Kiesewetter J, Fischer M. 2015. The teamwork assessment scale: a novel instrument to assess quality of undergraduate medical students' teamwork using the example of simulation-based ward-rounds. GMS J Med Educ. 32:1–9.

Leung G, Patil N. 2010. Patient safety in the undergraduate curriculum: medical students' perception. Hong Kong Med J. 16:101–105.

Madigosky WS, Headrick LA, Nelson K, Cox KR, Anderson T. 2006. Changing and sustaining medical students' knowledge, skills, and attitudes about patient safety and medical fallibility. Acad Med. 81: 94–101.

Martinou E, Chindambaran R, Krishnasamy G, Johnson A, Donnell JO, Vig S, Menon G. 2015. Simulation in undergraduate medical education designing a programme to improve medical students' non-technical skills. Int J Surg. 23:S102.

Meier AH, Boehler ML, McDowell CM, Schwind C, Markwell S, Roberts NK, Sanfey H. 2012. A surgical simulation curriculum for senior medical students based on TeamSTEPPS. Arch Surg. 147:761–766.

Müller M, Hänsel M, Winkelmann A. 2012. Impact of simulator training and crew resource management training on final-year medical students' performance in sepsis resuscitation: a randomized trial. Minerva Anestesiol. 78:901–909.

Myung SJ, Shin J-S, Kim JH, Roh HR, Kim Y, Kim J, Lee S-i, Lee J-H, Kim SW. 2012. The patient safety curriculum for undergraduate medical students as a first step toward improving patient safety. J Surg Educ. 69:659–664.

Paxton J, Rubinfeld I. 2010. Medical errors education: a prospective study of a new educational tool. Am J Med Qual. 25:135–142.

PRISMA. 2015. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA). [accessed 2017 Aug 29]. http://prisma-statement.org/.

Reed D, Price E, Windish D. 2005. Challenges in systematic reviews of educational intervention studies. Ann Intern Med. 142:1080–1089.

Robertson B, Kaplan B, Atallah H, Higgins M, Lewitt MJ, Ander DS. 2010. The use of simulation and a modified TeamSTEPPS curriculum for medical and nursing student team training. Simul Healthc. 5: 332–337.

Stahl K, Augenstein J, Schulman CI, Wilson K, McKenney M, Livingstone A. 2011. Assessing the impact of teaching patient safety principles to medical students during surgical clerkships. J Surg Res. 170:29–40.

Strauss AL, Corbin JM. 1998. Basics of qualitative research: procedures and techniques for generating grounded theory. 2nd ed. Thousand Oaks (CA): Sage Publications.

Thomas I, Nicol L, Regan L, Cleland J, Maliepaard D, Clark L, Walker K, Duncan J. 2015. Driven to distraction: a prospective controlled study of a simulated ward round experience to improve patient safety teaching for medical students. BMJ Qual Saf. 24:154–161.

Walton M, Woodward H, Van Staalduinen S, Lemer C, Greaves F, Noble D, Ellis B, Donaldson L, Barraclough B. 2010. The WHO patient safety curriculum guide for medical schools. Qual Saf Health Care. 19:542–546.

Wet CD. 2012. An overview of patient safety in primary care. NHS Educ Scotl. [accessed 2017 Oct 9]. http://www.nes.scot.nhs.uk/media/2075343/an-overview-of-patient-safety-in-primary-care-nov-12.pdf.