Table 3. Non-technical skills: Examples of potentially-relevant measurement tools for simulation-based healthcare improvement projects.

Measurement Tool	What it Measures	Type of Tool	Response Format	Reliability Evidence	Quantitative Evidence of Validity	Relevant usage example(s)
Oxford Non-Technical Skills Scale (NOTECHS)[80]	Non-technical skills of surgical teams and sub-teams.	Behavioral marker system.	Trained observer evaluates surgical sub-team on items assessing leadership and management, teamwork and cooperation, problem-solving and decision-making, and situation awareness. Revised versions of the scale include an additional sub-scale (communication and interaction) and multiple versions of the scale are available for different sub-teams (e.g., anaesthesiologists, surgeons, scrub nurses) with number of items varying slightly. Revised versions also use 6 response options (1 = not done, 6 = done very well). Sub-team scores can be combined for a total team score.	Adequate inter-rater reliability (r_{wg} = .99).[80] Test-retest reliability demonstrated by finding no significant differences in scores across three pre-intervention periods, and also no significant differences in scores across three post-intervention periods.[80]	Large statistically significant positive correlations with established non-technical skills measures (OTAS, r = .89, p = .046, and ANTS, r = .73, p = .01).[76, 80] Small statistically significant negative correlation with surgical errors (ρ = -.27, p = .046).[80] Statistically significant change in scores from pre- to post-training in expected direction (p = .005).[80]	To evaluate the effect of simulation-based human factors training on ophthalmic surgeons' non-technical skills.[76] To evaluate the effect of simulation-based surgical crisis management training on surgical trainees' non-technical skills.[81] To evaluate the effect of simulation-based training of high-risk clinical scenarios on surgical residents' non-technical skills.[82]

Measurement Tool	What it Measures	Type of Tool	Response Format	Reliability Evidence	Quantitative Evidence of Validity	Relevant usage example(s)
Non-Technical Skills for Surgeons (NOTSS)[85]	Non-technical skills of surgeons.	Behavioral marker system.	Trained observer evaluates participant on 14 elements assessing situation awareness, decision making, task management, leadership, and communication and teamwork with 4 response options ranging from 1 = poor to 4 = good.	Excellent inter-rater reliability (ICC across the categories = .95 - .99).[84] Strong inter-rater reliability (κ = .83).[83] Excellent inter-rater reliability (fifteen-rater α = .96-.99).[59] 2 untrained raters or 1 trained rater required for sufficiently reliable measurement across a range of elective and acute surgical procedures (G > .80).[59] 6-8 raters required for sufficiently reliable measurement across a range of different surgical procedures from 6 specialties (G > .80).[58] 5 procedures required for sufficiently reliable measurement of a single trainee surgeon (G > .80).[60]	Large statistically significant positive correlations with established non-technical skills measures (ANTS, r = .92, p < .001, and T-NOTECHS, r = .60 - .79, ps < .001).[76, 83] Medium to large statistically significant positive correlations with specialist training level and years of surgical training (rs = .36 - .57, ps < .001).[58]	To evaluate the effect of simulation-based human factors training on ophthalmic surgeons' non-technical skills.[76] To explore the relationship between non-technical skills and technical performance during simulated trauma resuscitations.[83]

Measurement Tool	What it Measures	Type of Tool	Response Format	Reliability Evidence	Quantitative Evidence of Validity	Relevant usage example(s)
Anaesthetists' Non-Technical Skills (ANTS)[86]	Non-technical skills of anaesthetists/ anaesthesi-ologists.	Behavioral marker system.	Trained observer rates participant on 15 skill elements comprised of 4 categories (task management, team-working, situation awareness, and decision making) with 3-4 elements per category. Anchored descriptors with 1 = poor (endangered patient safety) to 4 = good (outstanding, an example for others).	Respectable to very good internal consistency (α between elements in each category = .79-.86).[87] Borderline adequate inter-rater reliability for task management and team-working (r_{wg} = .65 and .65, respectively).[87] Questionable inter-rater reliability for situation awareness and decision making (r_{wg} = .56 and .61, respectively).[87]	Large statistically significant positive correlations with established non-technical skills measures (NOTECHS, r = .73, p = .01, NOTSS, r = .92, p < .001, and OTAS, r = .81, p = .03).[76] Change in scores from pre- to post-training in expected direction (one of four categories was statistically significant, p = .03).[88]	To evaluate emergency department teamwork behaviors during simulated scenarios designed to uncover latent safety threats.[5] To evaluate the effect of simulation-based human factors training on ophthalmic surgeons' non-technical skills.[76] To evaluate the effect of simulation-based continuing medical education on anaesthesiol-ogists' non-technical skills.[89]

Measurement Tool	What it Measures	Type of Tool	Response Format	Reliability Evidence	Quantitative Evidence of Validity	Relevant usage example(s)
Trauma Non-Technical Skills Scale (T-NOTECHS)[90]	Non-technical skills of trauma resuscitation teams.	Behavioral marker system.	Trained observer evaluates a team of participants on 27 exemplar behaviors assessing leadership, cooperation and resource management, communication and interaction, assessment and decision making, and situation awareness/coping with stress on a 5-point rating scale with different anchors for each behavior domain.	Moderate inter-rater reliability when evaluating videoed resuscitations (ICC = .71).[90] Poor inter-rater reliability when evaluating real-time real-life simulations (ICC = .48).[90] Poor inter-rater reliability when evaluating real-time simulated resuscitations (ICC = .44).[90]	Large statistically significant positive correlations with established non-technical skills measure (NOTSS, *r* across the categories = .60 - .79, *ps* < .001).[83] Large statistically significant positive correlation with number of completed resuscitation tasks during simulated and real-life resuscitations (*r* = .50, *p* < .01).[90] Medium statistically significant negative correlation with task completion time during simulated and real-life resuscitations (*r* = -.38, *p* < .05).[90] Statistically significant change in scores from pre- to post-training in expected direction (*p* < .001).[90]	To explore the relationship between non-technical skills and technical performance during simulated trauma resuscitations.[83] To evaluate the effect of simulation-based team training on trauma teams' non-technical skills.[91]

Measurement Tool	What it Measures	Type of Tool	Response Format	Reliability Evidence	Quantitative Evidence of Validity	Relevant usage example(s)
Human Factors Skills for Healthcare Instrument (HuFSHI)[48]	Human factors skills.	Self-report questionnaire.	Participant completes 12 items assessing unidimensional human factors skills related to healthcare: situation awareness; communication; teamwork; leadership; decision-making; and care on a 1-10 rating scale with 1 = definitely cannot do to 10 = definitely can do.	Excellent internal consistency ($\alpha = .92$).[48]	A one-factor model fit the data well.[48] Statistically significant difference in scores between new trainees versus experienced trainees in expected direction ($p < .001$).[48] Statistically significant change in scores from pre- to post-training in expected direction ($p < .001$).[48]	To evaluate the effect of simulation-based obstetric medical emergency training on medical doctors' and midwives' human factors skills.[93]
Situation Awareness Global Assessment Technique (SAGAT)[92]	Situation awareness (note: the tool has been adapted for many contexts).	Question and answer marked by examiner.	Trained observer marks participant's answers to questions during breaks in the scenario pertaining to level 1 (perception), 2 (comprehension), and 3 (projection) of situation awareness using a pre-determined rubric and/or computer records. Each item is deemed 0 = clinically unacceptable or 1 = clinically acceptable. Team scores are calculated by the proportion of correct responses logged by all team members.	Respectable internal consistency (adapted for trauma context, $\alpha = .77$).[77]	Large statistically significant positive correlation with traditional checklist assessment of each task performance (adapted for trauma context, $r = .81$, $p < .01$).[77] Statistically significant differences in scores based on level of training in expected direction (adapted for trauma context, $p < .001$).[77]	To evaluate inter-professional team situation awareness in simulated obstetrical crises (tool adapted for the obstetrics context).[94]

Measurement Tool	What it Measures	Type of Tool	Response Format	Reliability Evidence	Quantitative Evidence of Validity	Relevant usage example(s)
Ottawa Crisis Resource Management Global Rating Scale (Ottawa GRS)[97]	Crisis resource management skills during acute care emergencies.	Global rating scale.	Trained observer evaluates participant on 6 categories measuring overall performance, leadership skills, problem solving skills, situation awareness skills, resource utilization skills, and communication skills on a 7-point scale with descriptive anchors for each category.	Moderate inter-rater reliability (ICC = .59 and .61 across two scenarios).[97] No statistically significant difference in scores between first and second scenarios (test-retest reliability).[97]	Statistically significant difference in individual category and overall scores between first- and third-year postgraduate residents in expected direction ($p < .001$).[97]	To evaluate intensive care nurses' non-technical skills in a simulation-based emergency scenario.[95] To evaluate residents' crisis resource management skills during simulated emergency scenarios.[96]
Ottawa Crisis Resource Management Checklist (Ottawa CRM checklist)[96]	Crisis resource management skills during acute care emergencies.	Behavioral marker system.	Trained observer evaluates participant on the Ottawa GRS categories subdivided into 12 individual items representing important actions/behaviors with three response options: 0 = omitted or inadequately completed behavior, 1 = partially completed behavior, and 2 = successfully completed behavior. Total score is out of 30 (some items are given double weight).	Moderate inter-rater reliability (ICC = .63 and .55 across two scenarios).[96] No statistically significant difference in scores between first and second scenarios (test-retest reliability).[96]	Statistically significant difference in individual item and overall scores between first- and third-year postgraduate residents in expected direction ($p < .05$).[96]	To evaluate residents' crisis resource management skills during simulated emergency scenarios.[96]

Measurement Tool	What it Measures	Type of Tool	Response Format	Reliability Evidence	Quantitative Evidence of Validity	Relevant usage example(s)
Mayo High Performance Team Scale (MHPTS)[98]	Crew/crisis resource management (CRM) skills of teams in medical settings.	Self-report or observer-rated questionnaire (can be either).	16 items with possible total score ranging from 0 to 32 (higher scores = better adherence to CRM principles). Each item scored 0, 1, or 2 (0 = never or rarely, 1 = inconsistently, and 2 = consistently).	Very good internal consistency ($\alpha = .85$).[98] Questionable inter-rater reliability ($r = .59$).[4]	Statistically significant change in scores from pre- to post-training in expected direction ($p < .001$).[98]	To evaluate the teamwork of proposed teams in a simulated new clinical environment prior to the opening of a new hospital.[4] To compare expert- versus self-assessments of intensive care nurses' team performance in a simulation-based emergency scenario.[95]

Measurement Tool	What it Measures	Type of Tool	Response Format	Reliability Evidence	Quantitative Evidence of Validity	Relevant usage example(s)
Clinical Teamwork Scale (CTS)[52]	Teamwork during routine and emergent clinical care.	Behavioral marker system.	Observer evaluates a team of participants on 15 items assessing communication, decision making, role responsibility, situation awareness/resource management, and patient friendliness on a 10-point scale with 0 = unacceptable to 10 = perfect, except for 1 item which has a yes/no response format.	Excellent inter-rater reliability (ICC = .98).[52] Moderate inter-rater reliability (κ = .78).[52]	Large statistically significant positive correlation with a clinical performance measure (r = .53, p < .001).[100] Scores accurately reflect the standard of simulated performance (poor, average, or good).[52] Statistically significant change in scores from pre- to post-training in expected direction (p < .001).[100]	To evaluate the effect of a novel decision support technology on the teamwork of maternity teams in simulated scenarios.[101] To evaluate the effect of simulation-based training on the teamwork of resuscitation teams.[100] To evaluate the effect of in situ trauma simulation training on the teamwork and communication of trauma teams.[102] To evaluate the effect of simulation-based team training on the teamwork of obstetric teams.[99]

Measurement Tool	What it Measures	Type of Tool	Response Format	Reliability Evidence	Quantitative Evidence of Validity	Relevant usage example(s)
The Team Survey[104]	Short-term teamwork.	Self-report questionnaire.	Participant completes 44 items assessing team potency, team identification, shared mental models, and team orientation in short-term team performance on an unspecified scale (but items are worded such that a 1-5 rating scale with 1 = strongly disagree to 5 = strongly agree would be appropriate).	Respectable to excellent internal consistency (α across the categories = .73 - .93).[105]	Small statistically significant positive correlation between team orientation and a measure of team performance (r = .29, p < .001).[105] Small statistically significant negative correlation between team potency and a measure of team performance (r = -.28, p < .05).[105] A 4-factor solution fit the data well, but the factors were slightly different to what was predicted (although theoretically sensible).[105]	To evaluate the teamwork of proposed new teams during simulated clinical scenarios.
Teamwork Measurement Tool[106]	Teamwork of critical care teams.	Behavioral marker system.	Observer rates a team of participants on 23 items (adapted from the MHPTS) assessing leadership and team coordination, mutual performance monitoring, and verbalizing situational information, as well as an overall teamwork item on a 7-point scale, with descriptors of desirable and undesirable behaviors.	Very good to excellent internal consistency (α across the categories = .89 - .92).[106] Very good to excellent internal consistency (α across the factors = .89 - .96).[107]	A 3-factor model fit the data well.[106, 107] Statistically significant difference in scores between specialist and trainee teams in expected direction (p < .001).[106] Sensitive to changes over time.[106] Large statistically significant positive correlation between self (participant) and expert (observer) ratings (r = .66, p < .001).[107]	To evaluate the effect of simulation-based training on the teamwork of critical care teams.[103]

Measurement Tool	What it Measures	Type of Tool	Response Format	Reliability Evidence	Quantitative Evidence of Validity	Relevant usage example(s)
Team Emergency Assessment Measure (TEAM)[39]	Teamwork of teams during emergency resuscitations.	Behavioral marker system.	Observer evaluates a team of participants on 12 items assessing leadership, teamwork, and task management on a 5-point scale with 0 = never/hardly ever to 4 = always/nearly always, and an additional global rating item ranging from 1 to 10.	Excellent internal consistency during video-recordings (α = .97).[39] Very good internal consistency in real-time (α = .89).[39] Weak inter-rater reliability (κ = .55).[39] Weak test-retest reliability (κ = .53).[39]	Large statistically significant positive correlations between ratings on the 11 individual items and the global rating item (ρ = .75 - .95, p < .01).[39] A one-factor model fit the data well.[39]	To compare differences in teamwork performance of emergency providers in in-centre simulations, in situ simulations, and actual situations.[108] To evaluate the effect of simulation-based training on the teamwork of maternity teams.[109, 110] To compare the effect of simulation-based training conducted in situ versus off-site on the teamwork of obstetric anaesthesia teams.[111]

Measurement Tool	What it Measures	Type of Tool	Response Format	Reliability Evidence	Quantitative Evidence of Validity	Relevant usage example(s)
Observational Teamwork Assessment for Surgery (OTAS)[112]	Teamwork of surgical sub-teams.	Behavioral marker system.	Two trained observers (surgeon and a psychologist/human factors expert) evaluate sub-teams of participants on 3 teamwork tasks including patient tasks, equipment/provisions tasks, and communication tasks as well as 5 teamwork behaviours assessing communication, coordination, leadership, monitoring, and cooperation during the three stages of surgery (pre-operative, intra-operative, and post-operative). Tasks are scored by completion rates and behaviors are scored on a 7-point scale where 0 = problematic behavior and 6 = exemplary behavior.	Adequate inter-rater reliability for coordination (r = .72).[114] Questionable inter-rater reliability for all other behaviors (r = .35 - .64).[114]	Large statistically significant positive correlations with established non-technical skills measures (ANTS, r = .81, p = .03, and NOTECHS, r = .89, p = .046).[76, 80] A pair of expert raters produce more consistent scoring than a pair of expert-novice raters.[113]	To evaluate the effect of simulation-based human factors training on ophthalmic surgeons' non-technical skills.[76]

Measurement Tool	What it Measures	Type of Tool	Response Format	Reliability Evidence	Quantitative Evidence of Validity	Relevant usage example(s)
Assessment of Obstetric Team Performance (AOTP)[116]	Non-technical skills of obstetric teams.	Behavioral marker system.	Trained observer evaluates a team of participants on 6 themes (subdivided into 18 subthemes) measuring communication with patient and partner, task/case management, teamwork, situation awareness, communication with team members, and environment in the room on a 5-point rating scale with descriptive anchors from 1 = poor performance to 5 = excellent performance.	Excellent internal consistency ($\alpha = .91$).[116] Excellent internal consistency ($\alpha = .96$).[115] Excellent inter-rater reliability (ICC = .94).[116]	Large positive correlation with GAOTP ($r = .97$) (no statistical test reported).[115]	To evaluate the non-technical skills of obstetric teams performing simulated emergency obstetric scenarios.[116]
Global Assessment of Obstetric Team Performance (GAOTP)[116]	Non-technical skills of obstetric teams.	Global rating scale.	Identical to AOTP except that it evaluates only the 6 themes (i.e., no subthemes).	Very good internal consistency ($\alpha = .87$).[116] Excellent internal consistency ($\alpha = .91$).[115] Very good inter-rater reliability (eight-rater $\alpha = .81$).[115] Questionable test-retest reliability ($r = .47$).[115]	Large positive correlation with AOTP ($r = .97$) (no statistical test reported).[115]	To evaluate the effect of a novel decision support technology on the teamwork of maternity teams in simulated scenarios.[101]

Measurement Tool	What it Measures	Type of Tool	Response Format	Reliability Evidence	Quantitative Evidence of Validity	Relevant usage example(s)
Nursing Teamwork Survey (NTS)[49]	Nursing teamwork in acute care settings.	Self-report questionnaire.	Participant completes 33 items assessing trust, team orientation, backup, shared mental model, and team leadership on a 5-point rating scale with 1 = rarely and 5 = always.	Excellent internal consistency ($\alpha = .94$).[49] Adequate test-retest reliability ($r = .92$).[49]	Large statistically significant positive correlation with team satisfaction ($r = .63$, $p < .001$).[49] Large statistically significant positive correlation with established teamwork scale (SAQ teamwork sub-scale, $r = .76$, $p < .01$).[49] Survival flight nurses scored lower than inpatient nurses, as expected.[49] A 5-factor solution fit the data well and mapped onto the 5 constructs.[49]	To evaluate the effect of a virtual simulation on the teamwork of nursing staff.[117]