Table 4. Technical skills and clinical performance: 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)
Objective Structured Assessment of Technical Skills (OSATS)[38]	Technical skills during surgery.	Behavioral marker system and global rating scale.	Observer evaluates participant on operation-specific tasks with a 20-40 item checklist for each operation, as well as a global rating scale consisting of 7 performance dimensions scored on a 5-point rating scale with unique descriptions of the middle and extremes of the scale for each item. The checklist and global rating scale can be used separately.	Very good internal consistency for the global rating scale ($\alpha = .84$).[38] Respectable internal consistency for the checklist ($\alpha = .78$).[38] Very good inter-rater reliability for the global rating scale ($\alpha = .90$).[118] Excellent internal consistency for the global rating scale ($\alpha = .99$).[124] Excellent internal consistency for the checklist ($\alpha = .98$).[124] Statistically significant positive correlation between the two observers' scores on the checklist and global rating scale ($r = .99$ and .95, respectively, $p < .001$) (adequate inter-rater reliability).[124] 5 raters required for sufficiently reliable measurement across a range of different surgical procedures from 6 specialties ($G > .80$).[57]	Statistically significant improvement in scores with each year of resident training for the checklist (except between year 4 and year 5/6) and the global rating scale (*ps* < .001).[38] Statistically significant difference in scores between junior level versus middle/senior level trainees in expected direction (only global rating scale used, $p = .002$).[118] Statistically significant difference in scores between students and professors in expected direction (*ps* < .001).[124] Large statistically significant positive correlation with level of seniority ($r = .83$ for checklist and .86 for global, *ps* < .001).[124]	To evaluate the technical skills of surgical trainees during simulated surgical procedures.[118]

Measurement Tool	What it Measures	Type of Tool	Response Format	Reliability Evidence	Quantitative Evidence of Validity	Relevant usage example(s)
TeamOBS-Postpartum Hemorrhage (TeamOBS-PPH)[119]	Clinical performance of teams managing postpartum hemorrhage.	Behavioral marker system.	Observer rates participant on 19 objective checklist items with responses ranging from 0 = not done, 1 = partially or incorrectly done, and 2 = done correctly, as well as a subjective patient safety score. Items are adaptable to local clinical guidelines. Individual items are weighted differently to create a total score out of 100, with a minimum pass mark of 60.	Good inter-rater reliability (ICC = .83 in real-life scenarios and .86 in simulated scenarios).[119]	Statistically significant difference in scores between novice and expert teams in expected direction (p < .001).[119] Scores reflect the amount of patient blood loss in real-life scenarios (i.e., lower scores are associated with higher blood loss) (p = .029).[119]	To evaluate clinical performance in the management of postpartum haemorrhage during simulated scenarios.[119]
Checklist for Technical Skills[121]	Adherence to neonatal resuscitation guidelines.	Behavioral marker system.	Observer evaluates participant on 44 yes/no items that measure adherence to international guidelines for neonatal resuscitation. Correct decisions and proper procedures are given a score of 2, with selected items multiplied by 3 and penalty points subtracted, resulting in a maximum possible score of 100%.	Good inter-rater reliability (ICC = .77).[121]	10 percentage-point change in scores in expected direction from the first to second scenario after receiving feedback on performance.[121]	To evaluate the neonatal resuscitation skills of medical staff members during simulated resuscitations.[120] To evaluate the effect of simulation-based training on medical staff members' neonatal resuscitation skills.[110]

Measurement Tool	What it Measures	Type of Tool	Response Format	Reliability Evidence	Quantitative Evidence of Validity	Relevant usage example(s)
Simulation Team Assessment Tool (STAT)[122*]	Team performance during simulated pediatric resuscitations.	Behavioral marker system.	Observer rates participant on 94 elements covering basic assessment skills, airway/breathing, circulation, and human factors on a trichotomous scale (0-2 points) reflecting whether performance of each element was complete and timely (2), incomplete or untimely (1), or needed and not done (0).	Good inter-rater reliability (ICC = .81).[122]	Statistically significant difference in scores between resident and expert teams in expected direction (*p* = .02).[122]	To compare the performance of clinical teams of varying experience during simulation-based pediatric resuscitations.[122] To evaluate the impact of proposed changes in team structure on simulation-based pediatric resuscitation performance.
Clinical Performance Tool[79]	Adherence to pediatric resuscitation guidelines.	Behavioral marker system.	Observer rates participant on tasks derived from Pediatric Advanced Life Support (PALS) algorithms (number of tasks depends on the scenario). Tasks are scored on a trichotomous scale (0-2 points) reflecting whether performance of each element was complete and timely (2), incomplete or untimely (1), or needed and not done (0).	Adequate inter-rater reliability (*r* = .82).[79] Excellent inter-rater reliability (ICC = .95)[125]	Statistically significant difference in scores between first and second year residents in expected direction (*p* < .05).[79] Large statistically significant positive correlation with a clinical teamwork measure (*r* = .53, *p* < .001).[100] Statistically significant change in scores from pre- to post-training in expected direction (*p* < .001).[100, 125]	To evaluate the effect of simulation-based training on clinicians' pediatric resuscitation skills.[100] To evaluate the effect of a proposed procedural change on clinicians' adherence to pediatric resuscitation guidelines during simulated scenarios.

Measurement Tool	What it Measures	Type of Tool	Response Format	Reliability Evidence	Quantitative Evidence of Validity	Relevant usage example(s)
Structured Observation Protocol[123]	Nurses’ cardio-pulmonary resuscitation (CPR) performance.	Behavioral marker system.	Instructor evaluates participant on 12 items representing observable behaviors of First Responder CPR performance with 6 response options ranging from 1 = unable to perform to standards with verbal instruction and demonstration to 6 = independent, efficient performance with exemplary technique in application.	Very good internal consistency (α = .90).[123]	Statistically significant change in scores from pre- to post-training in expected direction (*p* < .001).[123]	To evaluate the effect of simulation-based First Responder training on nurses’ CPR performance.[123] To evaluate the effect of a proposed environmental change on nurses’ CPR performance during simulated scenarios.

* This tool also measures non-technical skill elements.