Table 5. Psychological constructs: 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)
State-Trait Anxiety Inventory (STAI)[30]	State and trait anxiety.	Self-report questionnaire.	Participant completes 40 items assessing state (feelings in the current moment) and trait (feelings in general) anxiety with responses ranging from 1 = almost never to 4 = almost always. Total scores range from 20 to 80 with higher scores representing higher anxiety. The tool can be separated into 2 sub-scales: STAI-State and STAI-Trait.	Excellent internal consistency (trait sub-scale, $\alpha = .90$ - .91).[30] Very good to excellent internal consistency (state sub-scale, $\alpha = .86$ - .94).[30] Adequate test-retest reliability (trait sub-scale, $r = .71$ - .75).[30]	Statistically significant difference in state sub-scale scores between low- and high-stress conditions ($p < .05$).[126, 127]	To compare differences in stress levels of obstetric teams between in situ versus off-site simulation-based anaesthesia training.[111] To compare differences in stress levels of emergency/surgery residents between high versus low stress simulation-based trauma resuscitation scenarios.[126]

Measurement Tool	What it Measures	Type of Tool	Response Format	Reliability Evidence	Quantitative Evidence of Validity	Relevant usage example(s)
National Aeronautic and Space Administration - Task Load Index (NASA-TLX)[129]	Subjective workload.	Self-report questionnaire.	Participant completes 6 items measuring mental, physical, and temporal demand, performance, effort, and frustration, using a 20-step bipolar scale. A score of 0-100 is obtained on each scale. Raw scores can be used or a weighted score can be calculated which is the original method of evaluation.	Respectable internal consistency ($\alpha < .80$).[131] Adequate test-retest reliability ($r = .83$).[129]	Large statistically significant correlations with performance measures in expected directions (time, $r = .75$, $p < .01$, and RMSE, $r = .65$, $p < .01$).[78] Large statistically significant positive correlations with other measures of subjective workload (WP, $r = .99$, $p < .001$, and SWAT, $r = .98$, $p < .001$).[78] Sensitive to differences between high and low workload tasks.[128]	To evaluate the perceived workload of proposed teams in a simulated new clinical environment prior to the opening of a new hospital.[4] To evaluate the effect of in-situ simulations designed to identify latent safety threats and orient staff members prior to the opening of a new emergency department on staff members' perceived workload in the first two weeks of the department's opening.[6]

Measurement Tool	What it Measures	Type of Tool	Response Format	Reliability Evidence	Quantitative Evidence of Validity	Relevant usage example(s)
Workload Profile (WP)[130]	Subjective workload.	Self-report questionnaire.	Participant rates the proportion (between 0 and 1) of attentional resources used during a task on 8 dimensions of workload based on Multiple Resource Theory. The number of tasks differ depending on the context, and any task can be used.	Excellent test-retest reliability (*r* across two tasks = .92 and .94).[130]	Large statistically significant positive correlations with other measures of subjective workload (NASA-TLX, $r = .99$, $p < .001$, and SWAT, $r = .97$, $p < .001$).[78] Large and medium statistically significant correlations with performance measures in expected directions (time, $r = .73$, $p < .01$, and RMSE, $r = .30$, $p < .05$, respectively).[78] Sensitive to different types of tasks ($p < .001$).[130]	To evaluate clinicians' perceived workload when trialling potential procedural changes in a simulated scenario.
The Surgery Task Load Index (SURG-TLX)[133]	Subjective workload in surgery.	Self-report questionnaire.	Participant completes 6 items measuring mental, physical, and temporal demand, task complexity, situational stress, and distractions with a 20-step bipolar scale. A score of 0-100 is obtained on each scale. Raw scores can be used or a weighted score can be calculated which is the original method of evaluation.	None published.	Higher scores significantly predict worse technical performance ($p = .04$).[132] Scores reflect the scenario conditions of differing stressor levels ($ps < .05$).[133] Statistically significant difference in scores between training scenario and actual scenario in expected direction ($p < .01$).[132]	To evaluate the effect of surgical flow disruptions on surgeons' perceived intra-operative workload during simulated surgical scenarios.[132]

Measurement Tool	What it Measures	Type of Tool	Response Format	Reliability Evidence	Quantitative Evidence of Validity	Relevant usage example(s)
Safety Attitudes Questionnaire (SAQ)[28]	Patient safety attitudes/ safety climate/ safety culture.	Self-report questionnaire.	Participant completes 30-60 items (depending on the version) measuring teamwork climate, safety climate, perceptions of management, job satisfaction, working conditions, and stress recognition on a 5-point rating scale ranging from 1 = disagree strongly to 5 = agree strongly. There is also a section for open-ended responses.	Respectable to very good internal consistency (α across the categories = 0.71 to 0.85) except for teamwork which was minimally acceptable (α = 0.68).[134] Moderate test-retest reliability (ICC > .70 for 5 of the 7 factors).[134]	A 6-factor model fit the data well and mapped onto the 6 constructs.[28] Large statistically significant negative correlation between teamwork climate sub-scale and number of adverse events (r = -.99, p < .01) (other sub-scales had large correlation coefficients but did not reach statistical significance).[134]	To compare differences in patient safety attitudes of obstetric teams between in situ versus off-site simulation-based anaesthesia training.[111] To evaluate the effect of simulation-based patient safety training on clinical teams' patient safety attitudes.[135] To evaluate the effect of simulation-based teamwork and communication training on pediatric emergency department teams' patient safety attitudes.[136] To evaluate the effect of simulation-based non-technical skills training on perinatal teams' patient safety attitudes.[137]

Measurement Tool	What it Measures	Type of Tool	Response Format	Reliability Evidence	Quantitative Evidence of Validity	Relevant usage example(s)
TeamSTEPPS Teamwork Attitudes Questionnaire (T-TAQ)[138]	Attitudes towards teamwork in healthcare.	Self-report questionnaire.	Participant completes 30 items assessing team structure, leadership, situation monitoring, mutual support, and communication, on a 5-point rating scale ranging from 1 = strongly disagree to 5 = strongly agree.	Respectable to very good internal consistency (α across the categories = .70 - .83).[138]	Large statistically significant positive correlations between the subscales, suggesting that the constructs are unique but related (r = .53 - .63, ps < .01).[138] Statistically significant change in scores from pre- to post-training in expected direction (p < .001).[140]	To evaluate the effect of simulation-based interprofessional teamwork training on the teamwork of neonatal resuscitation teams.[140]
Trust scales[139]	Trust within organizational teams.	Self-report questionnaire.	Participant completes 21 items assessing team-level perceived trustworthiness, cooperative behaviors, propensity to trust, and monitoring behaviors on a 7-point rating scale ranging from 1 = completely disagree to 7 = completely agree.	Respectable to very good internal consistency (α across the categories = .70 - .88).[139]	Small to medium statistically significant correlations between each sub-scale and a theoretically related construct (team commitment) in expected directions (r = -.26 - .39, ps < .05).[139] A 4-factor model fit the data well and mapped onto the 4 constructs.[139]	To evaluate team trust levels of proposed new teams during simulated clinical scenarios.