BMJ Open

Reliability of instruments that measure situation awareness, team performance and task performance in a simulation setting with medical students

Magnus Hultin,[1] Karin Jonsson,[1,2] Maria Härgestam,[2] Marie Lindkvist,[3] Christine Brulin[2]

[1]Department of Surgical and Perioperative Sciences, Anesthesiology and Critical Care Medicine, Umeå University, Umeå, Sweden
[2]Department of Nursing, Umeå University, Umeå, Sweden
[3]Department of Public Health and Clinical Medicine, Umeå University, Umeå, Sweden

Correspondence to
Dr Magnus Hultin;
magnus.hultin@umu.se

ABSTRACT

Objectives The assessment of situation awareness (SA), team performance and task performance in a simulation training session requires reliable and feasible measurement techniques. The objectives of this study were to test the Airways–Breathing–Circulation–Disability–Exposure (ABCDE) checklist and the Team Emergency Assessment Measure (TEAM) for inter-rater reliability, as well as the application of Situation Awareness Global Assessment Technique (SAGAT) for feasibility and internal consistency.

Design Methodological approach.

Setting Data collection during team training using full-scale simulation at a university clinical training centre. The video-recorded scenarios were rated independently by four raters.

Participants 55 medical students aged 22–40 years in their fourth year of medical studies, during the clerkship in anaesthesiology and critical care medicine, formed 23 different teams. All students answered the SAGAT questionnaires, and of these students, 24 answered the follow-up postsimulation questionnaire (PSQ). TEAM and ABCDE were scored by four professionals.

Measures The ABCDE and TEAM were tested for inter-rater reliability. The feasibility of SAGAT was tested using PSQ. SAGAT was tested for internal consistency both at an individual level (SAGAT) and a team level (Team Situation Awareness Global Assessment Technique (TSAGAT)).

Results The intraclass correlation was 0.54/0.83 (single/average measurements) for TEAM and 0.55/0.83 for ABCDE. According to the PSQ, the items in SAGAT were rated as relevant to the scenario by 96% of the participants. Cronbach's alpha for SAGAT/TSAGAT for the two scenarios was 0.80/0.83 vs 0.62/0.76, and normed χ^2 was 1.72 vs 1.62.

Conclusion Task performance, team performance and SA could be purposefully measured, and the reliability of the measurements was good.

INTRODUCTION

Medical errors are the third leading cause of death in the USA.[1] Knowledge about the relationship between human errors and patient safety has increased in the last two decades.[2,3] Simulated environments make it possible to improve skills by employing training strategies to prevent errors while simultaneously offering an arena for reliable assessments of skills.[4] Thus, simulation training is often used by organisations to minimise adverse events and to prevent healthcare errors.[5,6] This could be accomplished by improving *task performance, team performance* or *situation awareness* (SA).[7–9]

When developing strategies for improving clinical practice, it is essential to evaluate both task performance and team performance. According to Salas *et al*[10] and Kozlowski,[11] team performance is a multilevel process that includes the inter-relation between individual-level and team-level taskwork and teamwork processes. Thus, an optimal task, as well as team performance, depends on the coordinated activities of a team of individuals.[12,13] Checklists are often used to score task performance in acute care scenarios. The lists might include adherence to resuscitation protocols, the timing of the task, as

---

Strengths and limitations of this study

► The Situation Awareness Global Assessment Technique could be used to create items in Swedish to probe situation awareness, that is, in the participants' native language.
► Team Emergency Assessment Measure (TEAM) could be used by the raters in its original language (English).
► The developed Airways–breathing–circulation–disability–exposure (ABCDE) checklist has items that are well-defined concepts, and the difficulty lies in defining the rubrics for scoring the items.
► It is a weakness that the postsimulation questionnaire was translated into Swedish only and was not retranslated back into English.
► An interprofessional set of raters with different backgrounds and experiences rated TEAM and ABCDE with similar results.

well as the time taken to complete the components.[14] The Trauma Team Evaluation Tool (TTET) was developed by Holcomb *et al* for trauma scenarios managed according to Airways–Breathing–Circulation–Disability–Exposure (ABCDE) protocols based on Advanced Trauma Life Support (ATLS) and was tested on US military resuscitation teams from community hospitals.[15] As a pilot study, psychometric data such as reliability and validity were not reported. Since TTET was developed for a specific setting, the items and criteria for scoring have to be adapted to the proficiency levels of the participants and the standard operating procedures being used.

Team performance can be measured using the Team Emergency Assessment Measure (TEAM). It measures three dimensions of team performance: leadership, teamwork and task management.[16 17] The instrument has been initially developed and validated for simulator-based team training and has been recently validated for the collection of observational ratings of non-technical skills during live resuscitations in emergency departments.[18] Using instruments to score scenarios depend on reliable interpretations of the instrument by the raters, and it might be even more important to ensure the reliability of such interpretations when using an instrument in a non-native language.

Moreover, and in addition to task and team performance, SA is a prerequisite for patient safety and the prevention of errors, particularly in acute care situations.[9 19] SA includes three levels of ability: (1) perception and attention (*What?*—what's going on), (2) comprehension (*So what?*—the ability to understand what's going on) and (3) projection (*Now what?*) in order to anticipate and plan for future events.[20] In order to measure SA in a simulation setting, the Situation Awareness Global Assessment Technique (SAGAT) has been developed and subsequently adapted for use in healthcare settings.[21 22] One feature of SAGAT is that its use requires the simulation to pause,[21] which might influence clinical understanding, both by impeding the suspension of disbelief in the simulation setting[23] and by facilitating reflection on action.[24] Gardner *et al* showed that it was feasible to use SAGAT to measure SA in the team training of surgical trainees in advanced cardiac life support.[25] Globally, SAGAT has been used to study, for example, the effect of sleep deprivation on SA in trauma team training, how SA is associated with surgical trainee team performance, as well as nurses' clinical assessment of patient deterioration.[26–28] SAGAT can be analysed on an individual level and also on a team level. A specific application of SAGAT is Team Situation Awareness Global Assessment Technique (TSAGAT), in which the different members of a team answer SAGAT questions specific to their role. TSAGAT was constructed to account for the teams' shared SA and validated in trauma team training in a Canadian setting.[29] To our knowledge, SAGAT has not been previously used in a Swedish context, and it is therefore important to evaluate both its feasibility and its internal consistency.

Table 1 Background characteristics of the medical students (N=55) participating in the study

Paramater	
Age (years), mean (SD)	25 (4.5)
Male, n (%)	28 (51)
Female, n (%)	27 (49)
Previous healthcare education	
None, n (%)	44 (80)
Assistant nurse (two years at upper secondary school), n (%)	2 (4)
Registered nurse (three years at university), n (%)	3 (5)
Short courses (Red Cross), n (%)	6 (11)
Previous team training	
Yes, n (%)	20 (36)
Previous simulation experience	
Yes, n (%)	32 (58)
Previous experience of trauma patients*	
Yes, n (%)	20 (37)

*One missing value.

To summarise, ABCDE checklists, the TEAM instrument and SAGAT have been developed in order to evaluate different aspects of teamwork. However, these instruments and questionnaires have neither been translated into Swedish nor tested for feasibility or trustworthiness in a Swedish context. Such studies are necessary to enable the evaluation of teamwork in Swedish acute care settings and simulation-based training. Thus, this study aimed to test the ABCDE checklist that we developed for inter-rater reliability, the TEAM instrument for inter-rater reliability and two Swedish SAGAT questionnaires for feasibility and internal consistency.

METHODS

This study was based on data collected during simulation-based team training sessions with medical students.

Participants and raters

From March to October 2016, all medical students (N=68) in year 4 undertaking their clerkship in anaesthesiology and critical care medicine were invited to participate in the study while receiving mandatory simulator-based team training. In total, 55 students (81%) participated in the study (table 1) and 20 of them participated in both scenario A and scenario B.

All scenarios were video-recorded and later scored by four raters to allow for calculation of inter-rater reliability. First, a registered nurse with a Master's degree (1 year) in nursing (critical care medicine), 20 years' working experiences at an intesive care unit, 9 years' experience of human patient simulator team training and also a PhD student (author KJ). Second, a physician with 20 years'

working experience, an associate professor and consultant in anesthesiology and critical care medicine, with 14 years' experience of human patient simulator team training (author MHu). Third, a registered nurse with a Master's degree (1 year) in nursing (acute care medicine), with 6 years' working experience in prehospital care and 1 year as a medical student. Fourth, a paramedic with limited experiences beyond 2 years working as a paramedic in the Israeli Army and 1 year as a medical student.

Setting

Data collection in this study was carried out during simulation-based team training at the Clinical Training Centre of the Department of Nursing, Umeå University. The briefings, scenarios and debriefings were conducted in Swedish. Two cameras mounted at an angle were used to record videos in the simulation room, and one of the views included the patient monitor.

One week before team training, the students were asked to watch a 12 min video available at the learning platform that introduced both the ABCDE concept when caring for a patient and the room and equipment to be used during simulation-based training.[30][31] At the start of the training session, the students participated in a 15 min live introduction to the simulation laboratory presented by the operator and the instructor. Each student group (four to five students/group) was trained on four scenarios that focused on assessment and treatment of severely ill patients in an emergency room. In each scenario, three to four students were active, and one to two were observers. Therefore, only three to four participated in the interactive role-play in each scenario (the actual simulation). In all, the 55 unique students made up a total of 23 teams with three to four participants in each team, and 20 students participated in both case A and case B. The students all played the role of interns, that is, physicians who are in training to become licensed, with the attending nurse currently unavailable. The assigned task involved conducting a primary survey and stabilising the patient until more senior staff arrived in 15 min. After each scenario, a 10–15 min debriefing session permitted reflection and shared learning. The first scenario was a warm-up, and the last included a training summary. Cases A and B, or in the reverse order, as determined by the toss of a coin immediately before the second scenario, were used as the second and third scenarios.

All scenarios were conducted in Swedish, were designed to last 10–15 min and were preprogrammed into a Laerdal SimMan simulator in order to support the standardisation of the simulation. In essence, the patient cases used were slightly modified versions of the scenarios used by Hogan *et al*.[22] Case A was hypovolemic shock following a traffic incident. Case B was a pneumothorax with affected vital signs following a traffic incident.

Background questionnaire

The background questionnaire included informed consent and was answered immediately before team training started. The questionnaire included questions such as year of birth, male/female gender, previous medical training, previous experiences of team training, previous experiences of human patient simulator-based training, previous experiences of crisis resource management (CRM) and previous experiences of live trauma care.

ABCDE checklist

In order to measure the completeness of critical tasks in acute care scenarios, an ABCDE checklist was used. The original TTET comprised 58 items derived from the ATLS protocol.[15] Each item in the TTET was scrutinised by the authors of this study, and the number of items was reduced to those that reflected the measures expected to be carried out by year 4 medical students in the acute care scenarios in this study. Some items were also slightly modified to reflect the current ATLS standards. The final list contained 10 items reflecting the key elements of the ATLS primary survey, that is, the core management of ABCDE: airway assessed, airway secured, saturation assessed, oxygen applied, ventilation assessed, ventilation optimised, pulses checked (radial–femoral–carotid), venous access and infusions established, neurological disabilities checked (consciousness and pupils) and full/complete exposure. The items are well described in major textbooks and were translated into Swedish in line with the nomenclature being used. Compared with the original scoring system, an additional scoring option was added: performed after a reminder from the instructor. Each item was rated on a rating scale of 0–4 (0=not initiated, 1=performed after a reminder from the instructor, 2=partially performed, 3=performed completely before the end of the simulation and 4=performed consistently during the whole simulation, and NA=not applicable). Based on all items in the ABCDE checklist, an index was constructed as a mean score ranging from 0 to 4.

TEAM instrument

In order to measure team leadership, teamwork and task management, the TEAM instrument was used unmodified, that is, in English, as developed by Cooper *et al*.[16][17] The published internal consistency (Cronbach's alpha) was 0.97, and the inter-rater reliability was 0.55, as measured by Cohen's kappa (adjusted for chance), with a mean intraclass correlation coefficient (ICC) of 0.60 for the 11 items. The instrument comprises 11 critical behaviours rated on a scale of 0–4 (0=never/hardly, 1=seldom, 2=about as often as not, 3=often, 4=always/nearly always), which are summed into a total item score ranging from 0 to 44, and finally, a global rating of the team's overall performance on a scale of 1–10. The original publication with the TEAM instrument had no descriptors for the end points. For this study, we used 1=poor and 10=excellent.

The TEAM instrument comprises three subscales: leadership (items 1 and 2), teamwork (items 3–9) and task management (items 10 and 11). Indexes were constructed

based on the items in the subscales and on all items in the instrument, ranging from 0 to 4.

Procedure for rating the TEAM instrument and the ABCDE checklist

The raters (n=4) in this study used video recordings in the rating procedure of the TEAM scale and the ABCDE checklist. The raters held two separate 2-hour long meetings to discuss the interpretation of the descriptors in the ABCDE checklist and the TEAM instrument. During the first meeting, the discussions were facilitated using a different set of videos with similar scenarios, but from other teams. A total of six videos were used for this. During the second meeting, after four of the scenarios had been rated by each participant, the raters met again to discuss the interpretation of scales.

The raters independently assessed the videos of the simulation scenarios included in this study and rated the performances using the ABCDE checklist and the TEAM instrument. Each video was viewed at least twice by each rater.

SAGAT

When developing SA items for the specific scenarios, goal-directed task analysis was used, as initially described by Endsley.[21] Briefly, for each profession, major goals were identified along with subgoals. Critical decisions were then identified, and SA requirements were defined as the dynamic information needed to achieve the major goals, as opposed to static information such as rules and guidelines. The samplings (items) were then matched against the SA requirements. The original recommendation by Endsley was 30–60 items for within-subject studies for each of the three SA levels: (1) perception and attention, (2) comprehension and (3) projection.[20] When Gardner et al validated a questionnaire based on SAGAT in a study of medical trainees, each questionnaire comprised three items for each level of SA at each freeze,[25] while Hogan et al developed a questionnaire with three items at level 1, one item at level 2 and three items at level 3.[22]

In the present study, the SA questionnaire was refined and adapted to the scenario and expected skills level of the students according to the process used by Hogan et al.[22] First, the questionnaire was translated into Swedish by the authors of this study as a basis for developing scenario-specific items in Swedish. In accordance with SAGAT, targeted learning objectives for the training scenarios were formulated, and then the specific goals for each simulation were set. An iterative process was used to reformulate the items in Swedish, using a separate group of six professionals, all registered nurses, working in both a clinical context and a teaching context. The final sets of SA items are shown in table 2 (author's translation to English) with 11 items in each freeze. To determine whether an item is essential in a specific context, Lawshe advocated the use of professional assessments such as a content validity index (CVI) defined as the fraction of professionals who rate the item as important.[32] In the

present study, the relevance was reviewed by three professionals (nurse, n=1; physician, n=2) before being used in the study. All professionals rated the questions as relevant, that is, the CVI was 1.

The answers given by the participants in the SA questionnaire were classified as incorrect (0) or correct (1). The classifications were discussed and agreed on by the two authors (KJ and MHu). For answers on a continuous scale (eg, systolic blood pressure), a 10% range around the intended correct answer was accepted as correct. One question was removed from the questionnaire since it became apparent during the classification process that the question had frequently been misinterpreted.

In order to administer the SA questionnaires, the scenarios were frozen (ie, paused) twice. The first freeze of the scenario was 5 min into the scenario, unless there was an active task activity or if the team was conducting a team re-evaluation, in which case, the freeze was briefly postponed. During the freeze, the patient monitor was switched off, and the participants turned away from the patient simulator while individually answering the questions. The second freeze took place according to the same principles after an additional 5 min. All participants were allowed to complete the questionnaire before the scenario restarted.

The length of the freezes was measured from the video recordings (table 3). The start of the freeze was defined as the beginning of the sentence 'Now we will pause the scenario so that you can answer some questions about the patient case', and the end of the freeze was defined as the end of the sentence 'Is everyone in place and ready? Now the scenario will restart'.

TSAGAT was developed by Crozier et al[29] based on SAGAT as an assessment tool for evaluating team performance. In Crozier's study, each team comprised a trauma leader, airway manager and nurse. Individual SA questionnaires were developed for the three-team roles, including both shared knowledge and complementary knowledge. TSAGAT was calculated as the sum of individual SAGAT scores, and the TSAGAT scores had a high correlation to a traditional checklist (Pearson correlation, r=0.996). However, Salas et al defined team SA as a dynamic process defined as the team's shared understanding of a situation at a specific point in time,[33] while Endsley argued that team SA involves unique activities as information sharing and coordination.[20] Thus, in the present study, in order to measure this efficiently in all team members, all participants in a case received identical SA questionnaires. In this study, to account for slight differences in the number of team members between the teams, TSAGAT was calculated as the mean SA score in each team.

Postsimulation questionnaire

In order to measure whether the items in the SA questionnaire were considered relevant to the scenarios, and whether pausing the scenarios affected the team training activity, a postsimulation questionnaire (PSQ) was used.[22] The PSQ comprises 13 statements to be rated on a 4-point

Table 2 SA items for case A and B

A. Response scale, level and proportion correct (%) (n=39)

	Question	Response scale	Level*	Correct (%)
Freeze 1				
1.	What is the patient's respiratory rate?	Free text	1	26
2.	What are the findings from the assessment of A and B?	Free text	1	51
3.	Is the patient sufficiently oxygenated?	Yes/no	2	49
4.	What did the patient state as his or her year of birth?	Free text	1	36
5.	What do you think will happen to the patient's blood pressure in the next few minutes?	Increase/decrease/no change	3	74
6.	What is wrong with the patient (preliminary diagnosis)?	Free text	2	69
7.	How many peripheral intravenous cannulas does the patient have?	Free text	1	95
	Presuming an ordinary course of development, how will vital signs change in the next 10 min?			
8.	Heart rate	Increase/decrease	3	85
9.	Blood pressure	Increase/decrease	3	82
10.	Saturation	Increase/decrease	3	59
11.	Respiratory rate	Increase/decrease	3	74
Freeze 2				
12.	What is the patient's blood pressure?	Free text	1	77
13.	What are your findings under C (circulation)?	Free text	1	82
14.	What additional examinations/tests do you think are needed?	Free text	3	67
15.	Which medications/drugs do you think are needed?	Free text	3	49
16.	What is wrong with the patient (preliminary diagnosis)?	Free text	2	95
17.	Is there an algorithm for CPR on the wall?	Yes/no	1	26
18.	What previous illnesses has the patient described?	Free text	2	21
	Presuming an ordinary course of development, how will vital signs change in the next 10 min?			
19.	Heart rate	Increase/decrease	3	74
20.	Blood pressure	Increase/decrease	3	74
21.	Saturation	Increase/decrease	3	59
22.	Respiratory rate	Increase/decrease	3	69

B. Response scale, level and proportion correct (%) (n=36)

	Question	Response scale	Level*	Correct (%)
Freeze 1				
1.	What is the patient's respiratory rate?	Free text	1	28
2.	What are the findings from the assessment of A and B?	Free text	1	64

Continued

Table 2 Continued

	Question	Response scale	Level*	Correct (%)
3.	Is the patient sufficiently oxygenated?	Yes/no	2	78
4.	What does the sign on the patient's necklace indicate?	Free text	1	36
5.	What do you believe will happen to the patient's venous return in the next few minutes?	Increase/decrease/no change	3	61
6.	What is wrong with the patient (preliminary diagnosis)?	Free text	2	58
7.	Is a suction device available?	Yes/no	1	69
	Presuming an ordinary course of development, how will vital signs change in the next 10 min?			
8.	Heart rate	Increase/decrease	3	78
9.	Blood pressure	Increase/decrease	3	58
10.	Saturation	Increase/decrease	3	78
11.	Respiratory rate	Increase/decrease	3	72
Freeze 2				
12.	What is the patient's respiratory rate?	Free text	1	33
13.	What are your findings under C (circulation)?	Free text	1	69
14.	What additional examinations/tests do you think are needed?	Free text	3	92
15.	Which medications/drugs do you think are needed?	Free text	3	22
16.	What is wrong with the patient (preliminary diagnosis)?	Free text	2	83
17.	What is the patient's blood pressure?	Free text	1	53
18.	Is a defibrillator available in the room?	Yes/no	2	50
	Presuming an ordinary course of development, how will vital signs change in the next 10 min?			
19.	Heart rate	Increase/decrease	3	72
20.	Blood pressure	Increase/decrease	3	72
21.	Saturation	Increase/decrease	3	69
22.	Respiratory rate	Increase/decrease	3	56

The questions are shown translated into English in the table, and the original questions in Swedish are available on request from the authors of the study.

*Levels of SA: 1, perception; 2, comprehension; 3, projection.

CPR, cardiopulmonary resuscitation; SA, situation awareness.

Table 3 Time (minutes) for scenario freezes to measure situation awareness with the Situation Awareness Global Assessment Technique

Case	Teams (n)	Scenario Mean (SD)	First freeze Mean (SD)	Second freeze Mean (SD)	Total time Mean (SD)
A	12	14.1 (2.0)	2.7 (0.5)	2.5 (0.6)	19.3 (2.3)
B	11	14.1 (3.4)	2.7 (0.6)	2.8 (0.6)	19.7 (3.8)

scale ranging from strongly disagree (1) to strongly agree (4), in which five statements concern the SAGAT and the effect of freezing the scenario, and eight statements concern the simulation and the scenario per se. To the best of our knowledge, no data regarding the reliability of PSQ were presented in the original study. In this study, we only use the five questions relating to SAGAT and freezing the scenario. The PSQ was translated into Swedish by one of the authors, and the translation was further refined based on iterative discussions within our research group. Next, the Swedish PSQ was sent to a professional translation agency, together with the original PSQ for verification of the translation. The PSQ was answered in an anonymous web survey during the second week after the simulation training by 24 of the 55 participating students; that is, the response rate was 44% (table 4). In this study, the results of the 4-point scale were dichotomised into disagree (strongly disagree and disagree) and agree (agree and strongly agree).

Study size

The aim of the study was to evaluate the feasibility of ABCDE, TEAM and SAGAT for use in further studies. In order to achieve this, the study sample must be large enough to allow for, with fair precision, the calculation of descriptive statistics (means and SD) and the calculation of reliabilities. In order to assess a large effect size (Cohen's d of 0.9) with t-test with a power of 80% at the 0.05 level, 16 participants per group would be needed as determined by G*Power.[34] Thus, for the aims of the present study, the inclusion of 50 individuals and 20 teams would suffice.

Statistical analysis

Statistical analysis was performed using IBM SPSS Statistics for Windows V.24. Inter-rater reliability for the ABCDE checklist and the TEAM instrument was determined by intraclass correlation (ICC) using a two-way random-effects model (ICC (2,1) type absolute).[35 36] ICC is reported both as single measures and average measures since the ICC for single measures relates to the reliability of the individual ratings, while average measures relate to the reliability of the mean values. Cronbach's alpha was used to measure the internal consistency of the SAGAT and the TSAGAT.[35] Internal consistency considered as the extent to which all items measure the same latent variable was investigated by χ^2, and as suggested by Schweizer, a normed χ^2 below 2 was taken as an indication of a good fit.[37]

Patient and public involvement

Neither the patients nor the public was involved in the design or the data collection for this study.

RESULTS
Descriptives of ABCDE checklist, TEAM instrument and SA items

Fifty-five participants participated in the study, combined into 23 teams with 3–4 participants in each team, running either case A or case B. The ABCDE mean item score was 2.6. The mean TEAM total item score was 25.3, and the mean TEAM global rating was 4.8. The mean SA score per item ranged from 0.25 to 0.95, and the mean SA score per participant was 13 (table 5).

Table 4 Student agreement about usefulness and feasibility of the prospective collection of situation awareness items

Question concerning the SAGAT questionnaires In my opinion:	Agree/strongly agree n (%)	Disagree/strongly disagree n (%)	Do not know n (%)
The introduction was adequate.	24 (100)	0 (0)	0 (0)
The questions were clear.	23 (96)	1 (4)	0 (0)
The 'freezes' in the scenario adversely affected my concentration level and my performance.	7 (29)	17 (71)	0 (0)
The questions were relevant to the way I perceived the scenario.	23 (96)	1 (4)	0 (0)
Overall, I am satisfied with SAGAT as a tool for evaluating hands-on skills during a trauma exercise.	19 (79)	1 (4)	4 (17)

Results from the postsimulation questionnaire.
SAGAT, Situation Awareness Global Assessment Technique.

Table 5 Descriptives of ABCDE checklist, TEAM instrument and SAGAT questionnaires

	Case A Mean (SD)	Case B Mean (SD)
ABCDE checklist*	2.70 (0.19)	2.58 (0.27)
TEAM instrument		
Total item score 1–11†	25.5 (5.9)	25.0 (5.1)
Mean item score 1–11*	2.32 (0.54)	2.27 (0.46)
Mean subscore, leadership (items 1 and 2)*	2.27 (0.77)	2.32 (0.65)
Mean subscore, teamwork (items 3–9)*	2.28 (0.55)	2.23 (0.51)
Mean subscore, task management (items 10 and 11)*	2.49 (0.56)	2.38 (0.50)
Global rating‡	4.90 (1.13)	4.75 (1.04)
SAGAT questionnaires§	13.95 (4.26)	12.60 (3.35)

*Maximum score of 4 per item.
†Maximum score of 44.
‡Maximum score of 11.
§Maximum score of 22.
ABCDE, Airways–Breathing–Circulation–Disability–Exposure; SAGAT, Situation Awareness Global Assessment Technique; TEAM, Team Emergency Assessment Measure.

Inter-rater reliability for the ABCDE checklist and TEAM instrument

Inter-rater reliability, as measured by intraclass correlation, was 0.55 (single measures)/0.83 (average measures) for the ABCDE checklist and 0.54/0.83 for the TEAM scale. For the TEAM subscales of leadership, teamwork, task management and global rating, the intraclass correlations were 0.36/0.70, 0.45/0.77, 0.35/0.68 and 0.38/0.72, respectively.

Feasibility and internal consistency of SAGAT

The PSQ showed that 96% of the participants considered the SA items to be relevant to the case, and 96% considered the questions to be easy to understand (table 4). About three out of four (72%) participants stated that the freeze did not negatively impact their concentration or performance during the simulation session.

The internal consistency of SAGAT measured as Cronbach's alpha was 0.80 for case A and 0.62 for case B (table 6), and normed χ^2 was 1.72 vs 1.62. For level 1 (perception), Cronbach's alpha was low, 0.06 for case 1 and 0.25 for case 2, but for level 3, it was fair, 0.89 and 0.66, respectively.

For TSAGAT, the internal consistency as measured by Cronbach's alpha was good for the SAGAT questionnaires (all levels) with 0.83 and 0.76 for cases A and B, respectively, and for level 3, it was 0.89 and 0.79, respectively (table 6).

Table 6 Internal consistency (Cronbach's alpha) of SAGAT and TSAGAT

SAGAT/TSAGAT	Case A, n/n 39/12	Case B, n/n 36/11
Level 1	0.060/0.372	0.246/0.420
Level 2	0.321/–0.018	0.332/0.332
Level 3	0.891/0.891	0.659/0.786
Total (levels 1–3)	0.800/0.827	0.620/0.759

Internal consistency calculated for the questions measuring the three levels of SA and for the questionnaire in total.
n/n, number of individuals answering the SAGAT questionnaires/number of teams; SAGAT, Situation Awareness Global Assessment Technique; TSAGAT, Team Situation Awareness Global Assessment Technique.

DISCUSSION

Research on how to maximise team performance depends on the availability of sensitive and reliable tools for measuring the impact of an intervention. This study aimed to test the usability of three instruments and techniques developed in English-speaking contexts for rating in a Swedish setting with teams performing in their native language and cultural context. The main finding of this study was that the adapted ABCDE checklist and the TEAM instrument could be used with acceptable inter-rater reliability and that it was feasible to use SAGAT to measure SA. The scores were in the middle of the scales, indicating that the scales could purposefully be used in a future effect study providing that the test group also scores in the sensitive range of these instruments. The combination of these three measurements, ABCDE, TEAM and SAGAT, could permit analysis of task performance, team performance and the relationship to situation awareness in team training.

Scaling down the comprehensive TTET[15] to a smaller ABCDE checklist could introduce unintended errors in measurements. To be useful in the intended context, the items in the scale have to be established in and calibrated to the setting. Accordingly, the items must be relevant to the task. This will increase the possibility to detect differences when evaluating potential improvements in team and task performance.[35] In the present study, the developed ABCDE checklist was perceived by professionals to be relevant to the case. Inter-rater variability was low, indicating that the checklist could be reliably used for scoring task performance. The means of the scores were in the middle of the scale, which indicates that the checklist might be sensitive for differences between low and high performers.

For the TEAM instrument, the inter-rater reliability was 0.55 for single measurements and 0.83 for average measurements, which are similar to the results reported by Cooper et al,[16] where the inter-rater reliability for the TEAM instrument was 0.55 as measured by Cohen's kappa and 0.60 as measured by ICC and also to the results reported by McKay et al,[38] where the ICC was 0.59–0.88

for the different items. According to Koo and Li, an ICC between 0.50 and 0.75 indicates moderate reliability, while an ICC between 0.75 and 0.90 indicates good reliability.[36] Thus, the averaged measurements from four raters had good reliability. In the original TEAM publication by Cooper et al,[16] the performance improved significantly from novice learners to experts, and in this study, the participants were a homogenous set of medical students, and as such to be considered as novice learners. The medical students in the present study had an average TEAM item 1–11 score of 2.3 of a maximum 4, which translates to showing the desired behaviour a bit more often than not. This is in agreement with the 2.49 score reported by Cooper et al[16] for a group of second-year medical and nursing students rated with TEAM during an interprofessional 1 day resuscitation course. The national learning objectives require systematic training in leadership and followership, which might explain the rather high score.[39]

Both the PSQ and the CVI indicate that SAGAT could be used to construct questions that were considered relevant to the case. In both cases, the internal consistency was fair (0.80 and 0.61). When analysing the subscales, levels 1, 2 and 3, the internal consistency was low for levels 1 and 2, and higher for level 3. This could indicate that the perception of the situation in the groups was diverse and not related to the total score, while the ability to project the direction in which the cases were heading was more homogeneously related to the total score. TSAGAT had a higher overall homogeneity, as might be expected when analysing the means of the group SAGAT for each question, instead of the specific SAGAT answers.

The feasibility of using SAGAT to measure situation awareness was assessed by measuring the length of the freeze (ie, pause) needed to answer questions and by asking the participants for their perception of the pauses in the scenario and how the pauses affected the training session. According to Endsley, when measuring SA using SAGAT, the scenario is frozen while the participants answer questions that probe the three levels of SA: perception, comprehension and projection.[21] Each pause in this study lasted less than 3 min. The pauses and SAGAT questions could influence SA in both a negative direction and a positive direction. In the negative direction, the flow of the simulation training might be disrupted, stress might be induced by being forced to interrupt the case, and the commitment might decrease. According to the PSQ, the majority of the participants did not perceive that the pauses adversely affected concentration and performance. Contrary, the pauses and questions might facilitate the resolution of the clinical problem in the case by triggering and allowing time for reflection, fully in line with Schön's reflection on action.[24]

Limitations of this study

The participants in the study comprised year four medical students for the video recordings of simulation-based team-training and the rating of the video material was conducted by four participants representing a wide spectrum of experience and training. It could be argued that it is a limitation that this validation was not performed on a series of critical care teams, for example. However, for testing the reliability as well as the feasibility of the checklist, techniques and instruments, the simulation-based training with medical students was a readily available series of standardised simulations.

Difficulties were encountered in the interpretation of what constitutes partially performed versus performed completely in the ABCDE checklist. This could relate to the vast differences in the raters' level of education and previous experience. To improve the accuracy and minimise the variability, the raters were trained using a separate set of video recordings. TEAM was not translated into Swedish in order to avoid inducing errors. This was possible as fluency in English is a prerequisite for academic studies in Sweden. The ratings based on the original TEAM instrument were consistent between the raters, indicating that this may have been a correct assumption.

In the present study, CVI for SAGAT was measured according to Lawshe,[32] while the development of the ABCDE checklist and the two cases relied on the authors' experiences and iterative interactions with clinicians and experts in the field. Both the cases, the ABCDE checklist and the developed SAGAT questionnaires, could have benefitted from a full formal CVI by a review panel.

The PSQ was only translated from English and was not translated back into Swedish in order to formally check the identity of the items in the final sets. However, the translation was adjusted by a professional translator before being used in the study. Thus, the results of the PSQ can be used for probing the participants in the simulation with regard to their experiences of SAGAT.

This study focused on the quantification of performance during simulation-based training. The transferability of the studied behaviours into a real-world setting is an intriguing question for further studies.

CONCLUSION

In this setting with medical students, situation awareness, team performance and task performance could be assessed with techniques that were reliable and feasible. The developed ABCDE checklist and the TEAM instrument had high inter-rater reliability. The process of using SAGAT questionnaires during simulation-based training did not negatively affect the participants' evaluation of simulation-based training, and the developed SAGAT questionnaires had a fair internal consistency. Thus, the measurement of task performance, team performance and situation awareness may be conducted in future studies in a Swedish simulation-based training setting using these techniques.

REFERENCES

1. Makary MA, Daniel M. Medical error-the third leading cause of death in the US. *BMJ* 2016;353:i2139.
2. Kohn LT, Corrigan JM, Donaldson MS. *To err is human: building a safer health system*. National Academies Press, 2000.
3. World Health Organization. Multi-professional patient safety curriculum guide. who, Geneva, 2011. Available: https://apps.who.int/iris/bitstream/handle/10665/44641/9789241501958_eng.pdf
4. Scalese RJ, Obeso VT, Issenberg SB. Simulation technology for skills training and competency assessment in medical education. *J Gen Intern Med* 2008;23:46–9.
5. Zendejas B, Brydges R, Wang AT, *et al*. Patient outcomes in simulation-based medical education: a systematic review. *J Gen Intern Med* 2013;28:1078–89.
6. Gordon M, Darbyshire D, Baker P. Non-Technical skills training to enhance patient safety: a systematic review. *Med Educ* 2012;46:1042–54.
7. Salas E, Rosen MA, Held JD, *et al*. Performance measurement in simulation-based training. *Simul Gaming* 2009;40:328–76.
8. Rosen MA, Pronovost PJ. Advancing the use of checklists for evaluating performance in health care. *Acad Med* 2014;89:963–5.
9. Brady PW, Muething S, Kotagal U, *et al*. Improving situation awareness to reduce unrecognized clinical deterioration and serious safety events. *Pediatrics* 2013;131:e298–308.
10. Salas E, Cooke NJ, Rosen MA. On teams, teamwork, and team performance: discoveries and developments. *Hum Factors* 2008;50:540–7.
11. Kozlowski SWJ. Enhancing the effectiveness of work groups and teams: a reflection. *Perspect Psychol Sci* 2018;13:205–12.
12. Converse SA, Cannon-Bowers JA, Salas E. Team member shared mental models: a theory and some methodological issues. *Proceedings of the Human Factors Society Annual Meeting* 1991;35:1417–21.
13. Mathieu JE, Heffner TS, Goodwin GF, *et al*. The influence of shared mental models on team process and performance. *J Appl Psychol* 2000;85:273–83.
14. Boyle WA, Murray DJ, Beyatte MB, *et al*. Simulation-Based Assessment of Critical Care "Front-Line" Providers*. *Crit Care Med* 2018;46:e516–22.
15. Holcomb JB, Dumire RD, Crommett JW, *et al*. Evaluation of trauma team performance using an advanced human patient simulator for resuscitation training. *J Trauma* 2002;52:1078–86.
16. Cooper S, Cant R, Porter J, *et al*. Rating medical emergency teamwork performance: development of the team emergency assessment measure (team). *Resuscitation* 2010;81:446–52.
17. Cooper SJ, Cant RP. Measuring non-technical skills of medical emergency teams: an update on the validity and reliability of the team emergency assessment measure (team). *Resuscitation* 2014;85:31–3.
18. Cooper S, Cant R, Connell C, *et al*. Measuring teamwork performance: validity testing of the team emergency assessment measure (team) with clinical resuscitation teams. *Resuscitation* 2016;101:97–101.
19. Schulz CM, Krautheim V, Hackemann A, *et al*. Situation awareness errors in anesthesia and critical care in 200 cases of a critical incident reporting system. *BMC Anesthesiol* 2015;16:4.
20. Endsley MR. Toward a theory of situation awareness in dynamic systems. *Hum Factors* 1995;37:32–64.
21. Endsley MR. Direct measurement of situation awareness: validity and use of SAGAT. in: situation awareness analysis and measuement. *Lawrence Erlbaum Associates* 2000:1–21.
22. Hogan MP, Pace DE, Hapgood J, *et al*. Use of human patient simulation and the situation awareness global assessment technique in practical trauma skills assessment. *J Trauma* 2006;61:1047–52.
23. Hamstra SJ, Brydges R, Hatala R, *et al*. Reconsidering fidelity in simulation-based training. *Academic Medicine* 2014;89:387–92.
24. Schön DA. *The reflective practitioner: how professionals think in action*. The Perseus Books Group, 1984.
25. Gardner AK, Kosemund M, Martinez J. Examining the feasibility and predictive validity of the SAGAT tool to assess situation awareness among medical trainees. *Simul Healthc* 2017;12:17–21.
26. Stratton A, Furey A, Hogan M. The effect of sleep deprivation on a Resident`s situational awareness in a trauma scenario. *J Orthop Trauma* 2014;28:e160–3.
27. Lavoie P, Cossette S, Pepin J. Testing nursing students' clinical judgment in a patient deterioration simulation scenario: development of a situation awareness instrument. *Nurse Educ Today* 2016;38:61–7.
28. McKenna L, Missen K, Cooper S, *et al*. Situation awareness in undergraduate nursing students managing simulated patient deterioration. *Nurse Educ Today* 2014;34:e27–31.
29. Crozier MS, Ting HY, Boone DC, *et al*. Use of human patient simulation and validation of the team situation awareness global assessment technique (TSAGAT): a multidisciplinary team assessment tool in trauma education. *J Surg Educ* 2015;72:156–63.
30. Thim T, Krarup G, *et al*. Initial assessment and treatment with the airway, breathing, circulation, disability, exposure (ABCDE) approach. *Int J Gen Med* 2012;117.
31. ATLS advanced trauma life support student course manual American College of Surgeons; 2018.
32. Lawshe CH. A quantitative approach to content validity. *Pers Psychol* 1975;28:563–75.
33. Salas E, Prince C, Baker DP, *et al*. Situation awareness in team performance: implications for measurement and training. *Hum Factors* 1995;37:123–36.
34. Erdfelder E, Faul F, Buchner A. GPOWER: a general power analysis program. *Behavior Research Methods, Instruments, & Computers* 1996;28:1–11.
35. Nunnally JC. *Psychometric theory*. 2nd edition. McGraw-Hill, 1978.
36. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med* 2016;15:155–63.
37. Schweizer K. Some guidelines concerning the modeling of traits and abilities in test construction. *Eur J Psychol Assess* 2010;26:1–2.
38. McKay A, Walker ST, Brett SJ, *et al*. Team performance in resuscitation teams: comparison and critique of two recently developed scoring tools. *Resuscitation* 2012;83:1478–83.
39. Lindgren S, Brännström T, Hanse E, *et al*. Medical education in Sweden. *Med Teach* 2011;33:798–803.