Research Article   **Open Access**

Development and Assessment of an Evaluation Tool for Team Clinical Performance: The Team Average Performance Assessment Scale (TAPAS)

Denis Oriot[1,2], Armelle Bridier[3] and Daniel Aiham Ghazali[2,4]*

[1]Pediatric Emergency Department, University Hospital, Poitiers, France
[2]Simulation Laboratory, Faculty of Medicine, Poitiers, France
[3]Pediatric Intensive Care Unit, University Hospital, Tours, France
[4]Emergency Department, Pitié-Salpétrière University Hospital, Paris, France

Abstract

Introduction: Team effectiveness consists in team performance (procedures, algorithms) and *team process* (Crisis Resource Management - CRM). Although CRM scales and a few team performance checklists exist, to our knowledge, no team performance assessment scale covering all critical conditions in adults and children exists.

Objective: The objective was to develop and assess a clinical team average performance assessment scale (TAPAS) evaluating clinical performance during immersive simulation of critical conditions.

Methods: Three experts selected items from PALS, EPLS, NLS, ACLS, and ATLS courses. The final TAPAS included 129 items, rated 0/1/2, with a total over 100. Items were preselected according to each scenario, making the score a reflection of the percentage of the best performance for a given scenario. Psychometric analysis was tested on 159 simulations. Scenarios were acute life-threatening conditions in neonates, children and adults (medical, trauma) (SimNewB and ALS, Laerdal*). Two independent observers among a group of 8, assessed performance, and were surveyed on TAPAS ease of use. Analysis included: comparison between observers of means and SD, linear logistic regression, coefficient correlation, discordance; Cronbach alpha (CA), intra-class coefficient (ICC), and comparison at two training times.

Results: TAPAS score was 46.6 ± 15.5 (18-83.5). Analysis showed: CA=0.745, ICC=0.862. Observers' scores were not different (means and SD), highly correlated (coefficient=0.838, p=0.0011, R^2=0.64), and with a discordance <7%. TAPAS scores increased after training (p<0.0001). TAPAS were found to be easy to use.

Conclusions: TAPAS was a valuable team performance assessment instrument, easy to use in team simulation, in different age groups (neonatal, pediatric, adult) and critical conditions (medical, trauma).

Keywords: Teamwork; Clinical performance; Evaluation scale; Simulation; ABCDE algorithm; Assessment

Introduction

Healthcare providers train as individuals, yet function as teams, thereby creating a gap between training and reality [1]. Teamwork failure is consequently a primary threat to patient safety [2]. The challenge is to distinguish team process from team performance in team-based assessment efforts and training [3]. The so-called 'global team effectiveness' relies upon two separate components: team process referring to Crisis Resource Management (CRM) principles, and team performance (technical procedures & algorithms) [4].

Simulation provides a powerful educational technique, particularly in management of emergencies. It is also used to train multi-professional teams, to identify errors and to modify behaviors in view of improving performance [5]. In fact, teamwork training constitutes a core approach to enhanced levels of quality and safety.

Team performance measurement and evaluation helps to ensure that simulation-based training for teams is systematic and effective [6]. Simulation-based training represents an opportunity to administer assessments of medical teams and to provide participants with scientifically valid and practically useful experiments [7].

Educational evaluation consists in systematic appraisal of teaching and learning quality [8]. Assessment tools are needed to assess teamwork when responding to critical events [9]. The challenge actually seems to consist largely in the fact that investigators often do not assess performance in a way that is distinct from team process. Measures often mix the two, which can be problematic. Very often, when teams are evaluated, the emphasis is put on team behavior (leadership, communication, call for help, utilization of resources, awareness of the situation) to the detriment of team clinical performance focused on which care the patient/simulator is receiving for its survival [6,10]. While several validated scales have been published to evaluate team processes – CRM principles [6,11], clinical team performance assessment in simulation-based training often relies on scenario or situation-checklists designed to assess technical procedures performed in a specific order [11-15]. Another evaluation tool for clinical performance is the global rating scale (GRS), which assesses overall performance [16-18].

To our knowledge, there exists neither gold standard nor any evaluation scale covering all life-threatening conditions and providing objective assessment of clinical team performance.

The aim of this study was to develop and to psychometrically assess a clinical evaluation tool named Team Average Performance Assessment Scale (TAPAS), usable in simulated pediatric and adult life-threatening emergencies, and covering all the procedures and algorithms required

*Corresponding author: Dr Daniel Aiham GHAZALI, Pitié-Salpétrière University Hospital, 47-83 boulevard de l'Hôpital, 75013 Paris, France, Tel: 33-(0)1-42-17-72-42; Fax: 33-(0)1-42-17-70-10; E-mail: daniel.ghazali@univ-poitiers.fr

for team performance when managing a life-threatening situation, and which could be used to assist feedback, teaching, and assessment. TAPAS may be considered as a clinical tool that can serve as a complement to teamwork evaluation. Team behavioral assessment was not in the scope of this study, but performed parallel to the use of TAPAS with a specific CRM principle assessment scale.

Methods

Study

This study was performed in the Simulation Laboratory of the Faculty of Medicine of Poitiers, France. It was reviewed and approved by the Institutional Research Board of the INSERM 1402 ('Institut National Scientifique Et de la Recherche Médicale', # 11-28, 2009-09-20) and the 'Comité de Protection des Personnes' (Committee for protection of persons) registered under the number 13.05.16. A written informed consent was obtained from all participants for research and video. All results were kept anonymous.

Creation of the instrument

Content: Three subject-matter experts selected the items. They were specialized in Emergency Medicine (pre-hospital and in-hospital, adult and pediatrics), certified PALS, EPLS, NLS, ACLS, and ATLS. Two of them were training program instructors. Items originated in the 2010 recommendations of AHA [19-21] and ERC [22-24], and the ATLS course [25]. The content creation process was designed to cover the ABCDEs (Airway, Breathing, Circulation, Disability, and Exposure), with assessment and action items in each category, and to be used for a patient (from a neonate to an adult), presenting a medical or trauma emergency. A "setting" part was added prior to the ABCDE algorithm. Most items reflected critical management, making it easier to conserve objective terminology. Many items were included in TAPAS to cover the maximum range of possible scenarios from neonatal to adult. Items were selected because of their direct impact on patient safety [7].

For convenience and rapidity of checking while the scenario was running, items were preselected prior to a given scenario, assuming what would be performed if clinical management were optimal. Chosen items were thought to be relevant to each case. Choice of items consequently differed from one scenario to another depending on etiology, age group, situation, and what was supposed to be done according to the learning objectives. Every scenario had the same printed sheet but was customized for the specific scenario by highlighting the appropriate items to be scored. Each category contained a maximum number of preselected items according to scenario and learning objectives, differing from one scenario to another.

Scoring for each item adopted the recently proposed triple item-by-item rating [17] aimed at assessing both quality and performance time, and relying upon three classes: not performed (0/2), performed but incorrectly done or delayed (1/2), correctly performed and in time (2/2). The sum of scores from each category constituted the total score. Total scoring was established by collecting among the selected items the ones having been rated. Dividing by maximum possible score (score of all preselected items) and multiplying by 100 gave a result over 100. TAPAS consequently gave a score over 100, equivalent to the percentage of optimal clinical performance for management of a given scenario. This method allowed us to use TAPAS for any scenario, whatever the age group and clinical situation, as long as a life-threatening condition was present.

TAPAS was designed to give a performance assessment score for a multi-professional team's approach to a simulated life-threatening situation. It covered medical and traumatic emergencies of neonates, children and adults. TAPAS deliberately avoided measures of good team practice in favor of items directly relevant to the patient's survival: team performance alone (technical procedures and algorithms). It was printed as a paper evaluation form, and represented a formative evaluation without any threshold. The results of this step gave the pre-scale of TAPAS.

Response process: The pre-scale was tested and modified during several simulation-based trainings with scenarios including neonatal, pediatric, and adult life-threatening emergencies. Two populations were included (February 2010-January 2013): 228 emergency physicians during the Pediatric Emergency Procedures University course, and 57 multi-professional team providers (Emergency Department, Poitiers University Hospital). The head research investigator preselected items on the evaluation form in accordance with the relevant scenarios and learning objectives. To avoid redundancy, some items were deleted or gathered, i.e., antibiotics. Difficult scoring by observers led us to redistribute items on the evaluation form. In addition, we highlighted preselected items. By the end of the response process, the pre-scale had the desired level of precision for assessment activities and was adjusted by additions/deletions to produce the TAPAS scale. TAPAS included 129 items distributed in 6 categories. As the 129 items represented varied age groups or clinical situations, they were not rated together. Different colors and/or fonts were used to facilitate rating of trauma, neonatal, and CPR items (Supplemental Digital Content – Appendix 1). The last item in each category was "miscellaneous" making it possible to add a new treatment/management for a specific scenario.

Psychometric testing

The different elements of the psychometric testing process are reported on Table 1.

	Population tested	Scenarios	Assessment	Variables	Results
1st population	48 participants=12 multi-professional teams made of 4 persons: EP, PGY, RN, AD	9 scenarios (pediatric) 72 simulation sessions Played in teams of 4 persons	144 TAPAS forms filled out by 2 independent observers randomly chosen among 8	Internal consistency: CA (tested on 3 scenarios, played 12 times each) Reliability: ICC, comparison of means, comparison of variances, R^2	CA=0.745 ICC=0.862 Means: O1 vs O2=NS Variances: O1 vs O2=NS R^2=0.64
2nd population	48 emergency physicians	15 scenarios (adult, pediatric, neonatal) 87 simulation sessions Played in pairs	174 TAPAS forms filled out by 2 independent observers randomly chosen among 8	Comparison of TAPAS score at different training times (at 4 months interval)	$p < 0.0001$
Total	96 participants	24 scenarios 159 simulation sessions	318 TAPAS forms		

AD: Ambulance Driver; CA: Cronbach Alpha Coefficient; EP: Emergency Physician; ICC: Intra-class coefficient; O1: Observer 1; O2: Observer 2; NS: Not Significant; PGY: Resident; R^2: Square of Correlation Coefficient; RN: Nurse

Table 1: Psychometric testing process.

Participants and simulation setting: In France, adult emergency teams manage out-of-hospital pediatric emergencies, and most in-hospital pediatric emergencies ones in non-university hospitals, except for the neonatal ones. As recommended for psychometric testing [26], large homogeneous populations were included.

The study of internal consistency and reliability of the scale was performed during the Sim-Stress study of which the methodology is published elsewhere [27]. Twelve multi-professional teams of 4 persons were recruited (48 care providers): senior emergency physician, resident, nurse, and ambulance driver; emergency physicians were from Emergency Department and/or pre-hospital care from Poitou-Charentes hospitals (1.8 million inhabitants); residents were in Emergency Medicine internship; emergency physicians and residents were certified following Pediatric Emergency Procedures University course; nurses and ambulance drivers were from the Emergency Medical Service of the University Hospital of Poitiers with certification for EPLS or EPILS; emergency physicians, nurses, and ambulance drivers had less than 6 years of experience. Teams were drawn by lots and remained stable throughout the sessions. A high-fidelity mannequin (SimNewB, Laerdal®) was used. Nine scenarios were used: 4 hypovolemic shocks, 2 cardiogenic shocks, hemorrhagic shock in severe trauma, anaphylactic shock, and septic shock.

The study of comparison of scores at different training times was conducted with another population of participants: 48 emergency physicians were included and evaluated during the 1st and 5th sessions of a University course. A SimNewB and an ALS Kelly mannequin (Laerdal®) were used. Different scenarios were used in neonates and children (cardiac arrest, acute asthma, purpura fulminans, severe trauma, hypovolemic shock, cardiogenic shock, severe intussusception, opioid-induced apnea, neonatal asphyxia and meconial aspiration) and in adults (cardiac arrest, STEMI, difficult intubation, severe trauma, ARDS). Scenarios were rehearsed several times before simulation sessions in order to limit variability-induced errors, or assessment errors due to the limits of realism. All sessions were videotaped and scenarios lasted 20 min on average, followed by a good-judgment debriefing [28,29].

Observers: Eight observers (1 pediatric intensivist, 3 pediatric emergency physicians, 1 anesthetist, 3 emergency physicians), were selected and received a 2-hour training in assessment with TAPAS. All were trained to simulation-based education and highly motivated in using a novel assessment tool. The assessment information given to the observers was formalized. All of 8 of them were enrolled in evaluation, but only 2 of them were randomly chosen each time among the 8, to independently assess a simulation session. As a consequence, observers 1 and 2 were always different. They did not communicate scoring to each other, and were not allowed to discuss ratings with each other. They were not instructors or research investigators. Quality of rating was assessed and a post-assessment control was established in order to ensure rating of all the preselected items by research investigators. A research assistant calculated the final score, which was unknown to observers or instructors. In parallel to TAPAS, observers rated CRM performance using the CTS – Clinical Teamwork Scale [30].

Observers were surveyed on the ease of use of TAPAS scoring with a 5-class Likert scale.

Data analysis

Analysis was carried out on SAS 9.3 software. Descriptive analysis included percentage, mean, standard deviation (SD) of every variable. Comparative analysis used paired Student t-test. Internal consistency of the scale was analyzed by the Cronbach alpha coefficient established on three scenarios played by 12 multi-professional teams and the relative weight of each category of the scale. Interobserver reproducibility was analyzed by intra-class coefficient (ICC), comparison of means, and linear regression analysis. Because several observers were included in the assessment, F-test was used to compare variance of scores obtained by observer 1 and observer 2. A p value of <0.05 was considered as significant.

Results

General findings

Three experts designed the instrument, and 8 independent observers used it. The final version of TAPAS included 129 items distributed in 6 sections. For psychometric analysis 159 simulation sessions were analyzed. Twelve multi-professional teams (48 participants) performed 72 simulation sessions (9 scenarios) over 18 months (February 2013-September 2014), meaning that 144 TAPAS evaluations were recorded (pediatric). And 48 emergency physicians carried out 87 simulations with scenarios played in pairs with 2 observations of TAPAS for each, giving 174 TAPAS evaluations (neonatal, pediatric, and adult) in February-June 2012 and in September 2012-January 2013. Mean TAPAS score was 46.6 ± 15.5, ranging from 18 to 83.5.

Survey of observers showed that TAPAS scale was found very feasible, from preselection of items, marking during simulation, to the calculation of total score (Figure 1). Furthermore, it constituted a comprehensive approach to acute life-threatening situations in neonates, children, and adults scenarios. And TAPAS was never found insufficient for assessment of an emergency scenario. Moreover, all observers were very satisfied with having used TAPAS, found its measurements to be almost "all-inclusive", and were persuaded that it simplified assessment.

Validity analysis

Internal consistency was assessed by Cronbach alpha coefficient calculated from 3 scenarios (3 out of 9, because these were the comparisons [27]) played 12 times each. Global Cronbach alpha was 0.745 - meaning a reasonable internal consistency of the scale, and the Cronbach alpha coefficients for each section of TAPAS are given on Table 2. Furthermore, TAPAS scores were found to correlate with level of training. Comparison of TAPAS scores at different times of training showed a significant difference between the 1st simulation (35 sessions analyzed) and the 5th (52 sessions analyzed); respectively: 58.7 ± 10.8 vs. 83.0 ± 9.6 (p<0.0001) (Figure 2).

Comparison of TAPAS scores with those of CTS showed a modest correlation: correlation coefficient=0.64, R^2=0.16 in linear regression (Figure 3).

Section of TAPAS	CA coefficient
Setting	0.646
Airway	0.806
Breathing	0.718
Circulation	0.703
Disability	0.722
Exposure	0.687
Total scale	**0.745**

Table 2: Values of Cronbach alpha (CA) coefficient for each TAPAS section and the total scale.
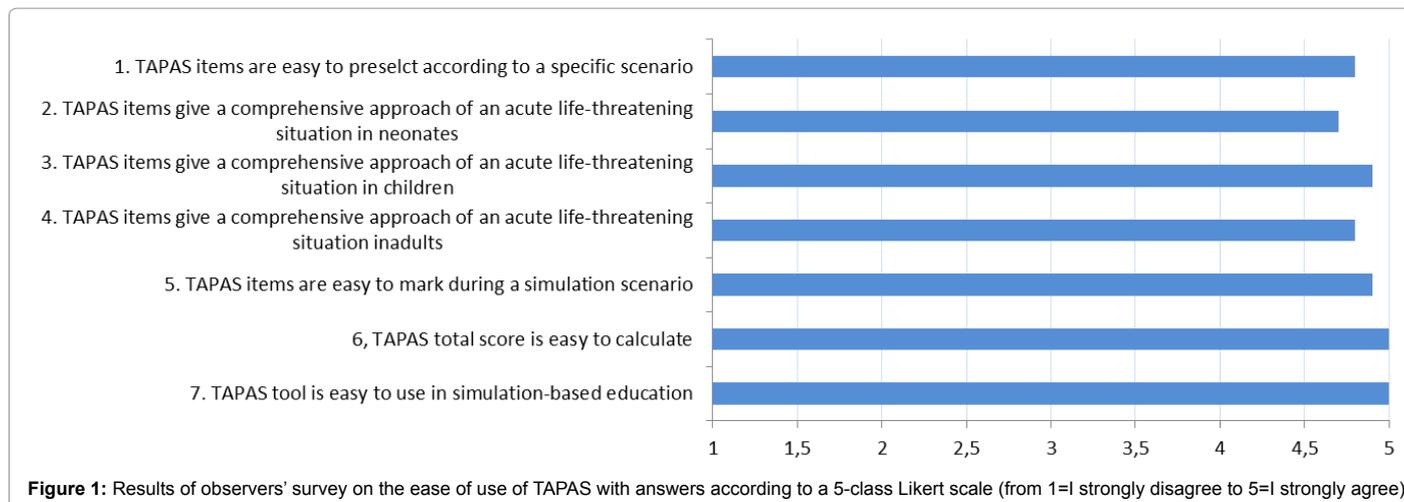
Figure 1: Results of observers' survey on the ease of use of TAPAS with answers according to a 5-class Likert scale (from 1=I strongly disagree to 5=I strongly agree).
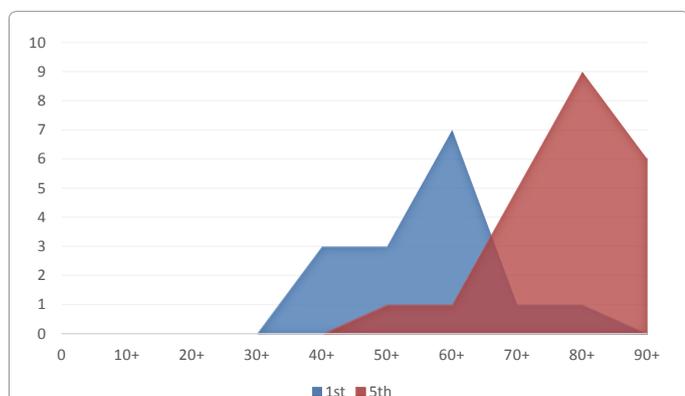

Figure 2: Distribution of mean TAPAS scores obtained at the 1st and 5th simulation sessions during the university course: 48 emergency physicians, 87 simulation sessions.
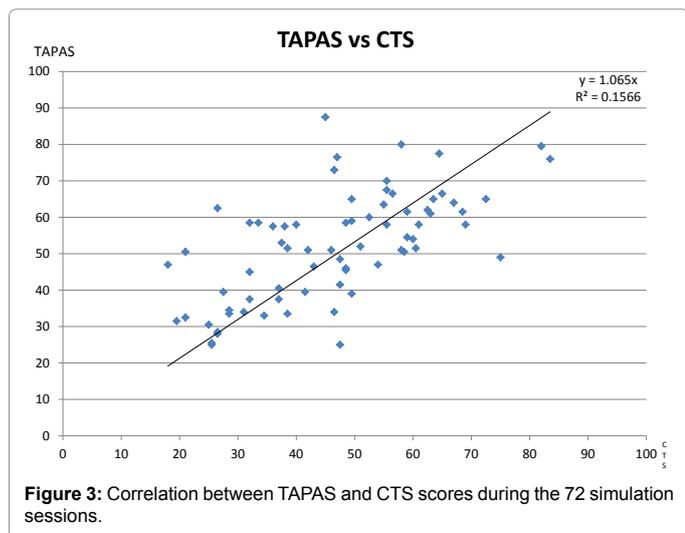

Figure 3: Correlation between TAPAS and CTS scores during the 72 simulation sessions.

Reliability analysis

Inter-rater reproducibility was assessed in different ways. There was no difference between the mean scores of observer 1 and observer 2: 46.3 ± 16.9 vs. 47.5 ± 15.8 (p=0.66). Both mean scores were highly correlated (correlation coefficient=0.838, p=0.0011, in linear regression R^2=0.64) (Figure 4). Finally, intra-class correlation coefficient was 0.862, which denotes high reproducibility.

Discordance between observers was precisely analyzed. Mean difference of scores between observers was 6.9 ± 6.4; while the overall scoring discordance between observers was inferior to 7%, it varied according to sections of the scale. This numerical discordance remained minimal, and generally consisted in a 1-point difference between ratings of 1 or 2 points (and not 0 or 1) for a given item in a selected category (Setting, A, B, C, D, or E). Because discordance could be zero with regard to the opposite variation of scores, rating discordance for each item in each section was calculated. The mean number of items rated differently between observers, according to each category of TAPAS was: Setting=0.93; A=1.70; B=2.03; C=2.76; D=0.60; E=1.86. This very good reliability between observers is represented on Figure 5.

Because several observers were used as observer 1 or 2, a comparison of variance was made. Variance of scores was 290.3 (observer 1), 252.0 (observer 2), and 241.4 (mean of observer 1+2). Comparison of variances (F-test) showed no difference between observer 1 and 2 (p=0.55). These results suggested good generalizability of TAPAS, due to its high reproducibility.
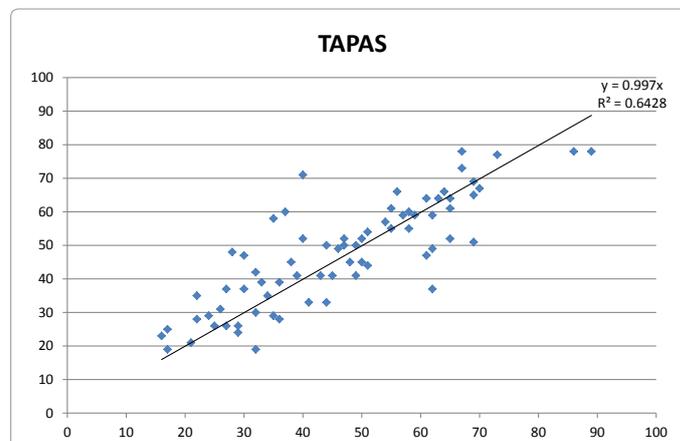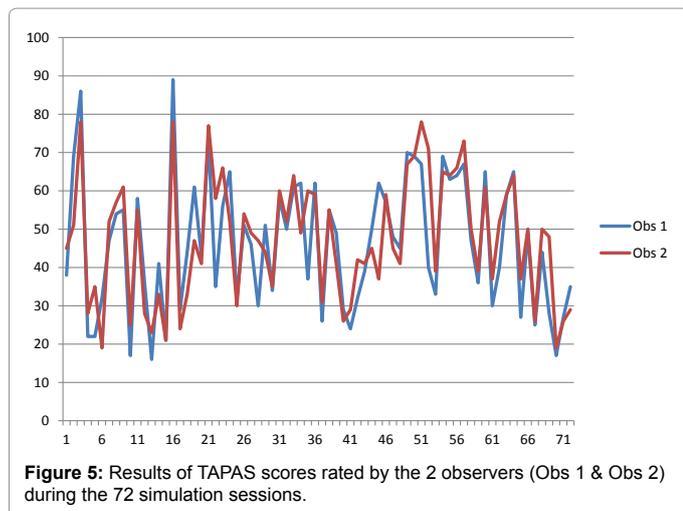

Figure 4: Relation between mean TAPAS score of observer 1 and observer 2 (72 simulation sessions).

Figure 5: Results of TAPAS scores rated by the 2 observers (Obs 1 & Obs 2) during the 72 simulation sessions.

Discussion

Main results

We designed and assessed a clinical team average performance assessment scale, named TAPAS. It was used as an evaluation instrument during simulation-based education of life-threatening conditions in different age groups ranging from neonates to adults. Made of 129 items in 6 sections, with a total score out of 100, TAPAS could cover many scenarios of critical medical or trauma-related circumstance with high reliability and good clinical relevance. Team performance assessment with TAPAS was easy to use and well-accepted.

Limitations

From its structure per se, TAPAS was designed and tailored for providers trained in using ABCDE (emergencies) or ABC (CPR) algorithms. It was not used by students not aware of such algorithms. The number of items to rate and the way of rating make it difficult to rate without previous training. One can expect that the accuracy of TAPAS would depend on the subjective opinion of whoever is running the scenario and picking the outcomes. This might be a limitation in very complex cases, but not in admitted managements like CPR, NLS or severe trauma, where what is expected to be done is rooted in international recommendations.

The scoring system adopted – 0/1/2 – for assessing both quality and performance time [17], inevitably implied some subjectivity in evaluating whether a procedure was "correctly" or "incorrectly" performed. Because TAPAS scope was to be used in team simulation sessions (when procedures are already known by participants and have been previously practiced on task-trainers), and because many checklists used this terminology, we did not detail each item for the exact description of required actions as we had previously done, for example, for the IO-access performance assessment scale [31]. Moreover, use of the same 0/1/2 scoring system for every item implied that all points were worth the same amount. This may not be true in reality. Furthermore the distance between 0, 1 and 2 may not be the same. For example, it is possible that late glucose is not particularly worse than early glucose in some cases, compared to late bag-mask ventilation in respiratory failure. On the other hand, another approach – 0/1 –, would have led to the risk of confusing not done and done lately.

Finally, current TAPAS is lacking in items for gynecology/obstetrics and anesthesiology emergencies, but given the highly adjustable nature of the TAPAS frame, items pertaining to these specialties could be added in the future.

The instrument's development and its psychometric properties (validity & reliability)

Clinical performance is most often assessed with scenario or learner-specific checklists designed to evaluate a healthcare provider [12,13,15,16,32]. Furthermore, investigators often do not assess performance in a way that is distinct from team process, and some checklists combine technical skills and behaviors in the same assessment tool mixing the two, which can be problematic [33]. As recommended, team performance was isolated from processes [4], and TAPAS dealt only with clinical performance.

In an OSCE evaluation tool – Clinical Performance Evaluation Tool, domains were too broadly defined: from data acquisition to interpersonal relations and clinical competence [34]. By contrast, TAPAS contains highly precise performance items – as recommended [7], and pre-selection ensures accuracy.

Most checklists use a 4-point Likert scale: 0) Not done; 1) Done; 2) Done well; 3) Unable to observe [35]. They may not take performance time into account [18]. Others have developed a problem-solving rating-scale with three marks 0/1/2, failing to detail the different steps of performance [36]. In contrast, a recent study proposed a checklist with the same triple item-by-item rating assessing both quality and performance time [17] that we adopted for TAPAS.

Few studies report reliability [11] or describe the validation process of assessment scales [37]. The Delphi method is commonly used for content implementation of scales [38,39] or checklists [40]. For TAPAS, it appeared simpler to ask experts to list significant items of international recommendations. Analysis of internal consistency showed significant clinical relevance. Furthermore, TAPAS had the highest inter-observer reproducibility (ICC=0.862) of the 5 reported assessment tools focusing on pediatric resuscitation in a simulated environment (although also including CRM evaluation): Ottawa Crisis Resource Management Global Rating Scale – ICC=0.61 [41], Neonatal Resuscitation Program Megacode Checklist – CA=0.70 [14], Tool For Resuscitation Assessment Using Computerized Simulation – ICC=0.80 [42], Standardized Direct Observation Tool – ICC=0.81, CA=0.95 [43], and Evaluation Tool for pediatric resident competence in leading simulated pediatric resuscitations – ICC=0.62, CA=0.81 [44].

TAPAS could not be compared to a gold standard or other validated scale covering all critical situations. And so, comparison was done at two different stages of training. A significant increase in TAPAS score after 4 months of training reflected improved clinical performance, which rendered the scale relevant for clinical performance assessment. Similarly, Brett-Fleeger found performance scores to be higher for the trainees having received the most simulation sessions [42].

Comparison of TAPAS scores with Clinical Teamwork Scale (CTS) scores (non-technical skills) gave a modest correlation. A link between behavioral teamwork and clinical performance had already been reported in medical students [45], and a recent review demonstrated that team process behaviors indeed influence clinical performance [46].

The use of the instrument

It is crucial to assess all components of team effectiveness (team process and team performance) during simulation-based trainings with multi-professional teams using reliable and accurate assessment

tools [7]. Some team-based simulation-based trainings seem to place emphasis on CRM and consequently place performance assessment at a second level. Besides evaluation of "how the team functions" (CRM principles assessment tool), TAPAS could provide an accurate complement to the evaluation by precisely assessing the medical performance produced on a simulated patient, which is crucial in high-stakes situations. In summary, in a patient-centered care evaluation, TAPAS could provide a valid and reliable tool related to patient safety, as well as complete assessment of team effectiveness.

Though not readily quantifiable, the educational effect of assessment is appreciable [26,47]. It draws sizable benefit from trainee motivation [26]. Feedback drives learning, and in team-based simulation-based trainings, facilitated debriefings are the primary method for delivering feedback [6]. One of the main goals of assessment is "to optimize the capabilities of all learners and practitioners by providing motivation and direction for future learning" [48]. During debriefing, TAPAS could pinpoint existing shortcomings and highlight areas requiring additional efforts.

At first glance, a 129-question tool would seem very unwieldy but through quick highlighting, the tool could be tailored to a large number of ACLS/ATLS/PALS/NLS cases, and be ideal for an application. Given its preselection of highly varied items, TAPAS was flexible, easy to use, and able to cover many scenarios of medical or traumatic emergency, whatever the age group. By its plasticity, TAPAS was found to be very adjustable to any life-threatening scenario with new management according to future recommendations, since one can consider adding additional "free text" rating items.

Although TAPAS was designed to assess management of life-threatening emergencies in simulation education and research, it has not yet been applied in other (non-simulation) settings but may be simple and flexible enough to be used in a clinical setting.

Conclusion

The TAPAS is a clinical team performance tool, designed and assessed with high clinical relevance and reliability. It is useable in many simulated critical situations ranging from neonatal to adult medical or traumatic scenarios. To our knowledge, there currently exists no other adjustable tool designed to assess clinical team performance.

TAPAS provides a detailed assessment of team clinical performance and during debriefing it can highlight performance shortcomings as specific issues requiring further improvement. As a faithful reflection of team performance, for us it has become a mandatory tool in any simulation-based training involving life-threatening emergencies. Future studies should focus on its use in novices as a training tool.

References

1. Hunt EA, Shilkofski NA, Stavroudis TA, Nelson KL (2007) Simulation: translation to improved team performance. Anesthesiol Clin 25: 301-319.

2. Risser DT, Rice MM, Salisbury ML, Simon R, Jay GD, et al. (1999) The potential for improved teamwork to reduce medical errors in the emergency department. The MedTeams Research Consortium. Ann Emerg Med 34: 373-383.

3. Brannick MT, Prince C (1997) An overview of team performance measurement. Team performance and measurement: theory, methods, and application. Mahwah, NJ, Erlbaum 3-16.

4. Fernandez R, Kozlowski SW, Shapiro MJ, Salas E (2008) Toward a definition of teamwork in emergency medicine. Acad Emerg Med 15: 1104-1112.

5. Overly FL, Sudikoff SN, Shapiro MJ (2007) High-fidelity medical simulation as an assessment tool for pediatric residents' airway management skills. Pediatr Emerg Care 23: 11-15.

6. Rosen MA, Weaver SJ, Lazzara EH, Salas E, Wu T, et al. (2010) Tools for evaluating team performance in simulation-based training. J Emerg Trauma Shock 3: 353-359.

7. Grand JA, Pearce M, Rench TA, Chao GT, Fernandez R, et al. (2013) Going DEEP: guidelines for building simulation-based team assessments. BMJ Qual Saf 22: 436-448.

8. Wilkes M, Bligh J (1999) Evaluating educational interventions. BMJ 318: 1269-1272.

9. Gaba DM, Howard SK, Flanagan B, Smith BE, Fish KJ, et al. (1998) Assessment of clinical performance during simulated crises using both technical and behavioral ratings. Anesthesiology 89: 8-18.

10. Shapiro MJ, Gardner R, Godwin SA, Jay GD, Lindquist DG, et al. (2008) Defining team performance for simulation-based training: methodology, metrics, and opportunities for emergency medicine. Acad Emerg Med 15: 1088-1097.

11. Kardong-Edgren S, Adamson KA, Fitzgerald C (2010) A review of currently published evaluation instruments for human patient simulation. Clin Simul Nursing 6: e25-35.

12. Murray DJ, Boulet JR, Kras JF, Woudhouse JA, Cox T, et al. (2004) Acute care skills in anesthesia practice: a simulation-based resident performance assessment. Anesthesiology 101: 1084-1095.

13. Fehr JJ, Boulet JR, Waldrop WB, Snider R, Brockel M, et al. (2011) Simulation-based assessment of pediatric anesthesia skills. Anesthesiology 115: 1308-1315.

14. Lockyer J, Singhal N, Fidler H, Weiner G, Aziz K, et al. (2006) The development and testing of a performance checklist to assess neonatal resuscitation megacode skill. Pediatrics 118: e1739-1744.

15. Halamek LP (2013) Simulation as a methodology for assessing the performance of healthcare professionals working in the delivery room. Sem Fetal Neonat Med 18: 369-372.

16. McBride ME, Waldrop WB, Fehr JJ, Boulet JR, Murray DJ (2011) Simulation in pediatrics: the reliability and validity of a multiscenario assessment. Pediatrics 128: 335-343.

17. Everett TC, Ng E, Power D, Marsh C, Tolchard S, et al. (2013) The Managing Emergencies in Paediatric Anaesthesia global rating scale is a reliable tool for simulation-based assessment in pediatric anesthesia crisis management. Paediatr Anaesth 23: 1117-1123.

18. Boulet JR, Murray DJ (2010) Simulation-based assessment in anesthesiology: requirements for practical implementation. Anesthesiology 112: 1041-1052.

19. Morrison LJ, Deakin CD, Morley PT, Callaway CW, Kerber RE, et al. (2010) Part 8: Advanced Life Support 2010 International Consensus on Cardiopulmonary Resuscitation and Emergency Cardiovascular Care Science With Treatment Recommendations. Circulation 122: S345-421.

20. Kleinman ME, de Caen AR, Chameides L, Atkins DL, Berg RA, et al. (2010) Part 10: Pediatric Basic and Advanced Life Support 2010 International Consensus on Cardiopulmonary Resuscitation and Emergency Cardiovascular Care Science With Treatment Recommendations. Circulation 122: S466-515.

21. Perlman JM, Wyllie J, Kattwinkel J, Atkins DL, Chameides L, et al. (2010) Part 11: Neonatal Resuscitation 2010 International Consensus on Cardiopulmonary Resuscitation and Emergency Cardiovascular Care Science With Treatment Recommendations. Circulation 122: S516-538.

22. Nolan JP, Soar J, Zideman DA, Biarent D, Bossaert LL, et al. (2010) European Resuscitation Council Guidelines for Resuscitation 2010 Section 1. Executive summary. Resuscitation 81: 1219-1276.

23. Biarent D, Bingham R, Eich C, López-Herce J, Maconochie I, et al. (2010) European Resuscitation Council Guidelines for Resuscitation 2010 Section 6. Paediatric life support. Resuscitation 81: 1364-1388.

24. Resuscitation Council (2015) Resuscitation Guidelines 2010. Newborn Life Support 11.

25. American College of Surgeons Committee on Trauma (2010) Advanced Trauma Life Support, Student Course Manual (9th ed.) Chicago, IL.

26. Norcini JJ, McKinley DW (2007) Assessment methods in medical education. Teaching Teach Educ 23: 239-250.

27. Ghazali A, Ragot S, Brèque C, Guechi Y, Boureau-Voultoury, et al. (2016) Randomized controlled trial of multidisciplinary team stress and performance in immersive simulation for management of infant in shock: study protocol. Scand J Trauma Resusc Emerg Med 24: 36.

28. Rudolph JW, Simon R, Dufresne RL, Raemer DB (2006) There's no such thing as « nonjudgmental » debriefing: a theory and method for debriefing with good judgment. Sim Healthcare 1: 49-55.

29. Rudolph JW, Simon R, Raemer DB, Eppich WJ (2008) Debriefing as formative assessment: closing performance gaps in medical education. Acad Emerg Med 15: 1010-1016.

30. Guise JM, Deering SH, Kanki BG, Osterweil P, Li H, et al. (2008) Validation of a tool to measure and promote clinical teamwork. Sim Healthcare 3: 217-223.

31. Oriot D, Darrieux E, Boureau-Voultoury A, Ragot S, Scepi M (2012) Validation of a performance assessment scale for simulated intraosseous access. Sim Healthcare 7: 171-175.

32. Murray DJ, Boulet JR, Avidan M, Kras JF, Henrichs B, et al. (2007) Performance of residents and anesthesiologists in a simulation-based skill assessment. Anesthesiology 107: 705-713.

33. Nishisaki A, Nguyen J, Colborn S, Watson C, Niles D, et al. (2011) Evaluation of multidisciplinary simulation training on clinical performance and team behavior during tracheal intubation procedures in a pediatric intensive care unit. Pediatr Crit Care Med 12: 406-414.

34. Gordon JA, Tancredi DN, Binder WD, Wilkerson WM, Shaffer DW (2003) Assessment of a clinical performance evaluation tool for use in a simulator-based testing environment: a pilot study. Acad Med 78: S45-S47.

35. Koch Hall A, Pickett W, Dagnone JD (2012) Development and evaluation of a simulation-based resuscitation scenario assessment tool for emergency medicine residents. Can J Emerg Med 14: 139-146.

36. Devitt JH, Kurrek MM, Cohen MM, Cleave-Hogg D (2001) The validity of performance assessments using simulation. Anesthesiology 95: 36-42.

37. Byrne AJ, Greaves JD (2001) Assessment instruments used during anaesthetic simulation: review of published studies. Br J Anaesth 86: 445-450.

38. Scavone BM, Sproviero MT, McCarthy RJ, Wong CA, Sullivan JT, et al. (2006) Development of an objective scoring system for measurement of resident performance on the human patient simulator. Anesthesiology 105: 260-266.

39. Adler MD, Vozenilek JA, Trainor JL, Eppich WJ, Wang EE, et al. (2011) Comparison of checklist and anchored global rating scales instruments for performance rating of simulated pediatric emergencies. Sim Healthcare 6: 18-24.

40. Schmutz J, Eppich WJ, Hoffmann F, Heimberg E, Manser T (2014) Five steps to develop checklists for evaluating clinical performance: an integrative approach. Acad Med 89: 996-1005.

41. Kim J, Neilipovitz D, Cardinal P, Chiu M, Clinch J (2006) A pilot study using high-fidelity simulation to formally evaluate performance in the resuscitation of critically ill patients: The University of Ottawa Critical Care Medicine, High-Fidelity Simulation, and Crisis Resource Management I Study. Crit Care Med 34: 2167-2174.

42. Brett-Fleegler MB, Vinci RJ, Weiner DL, Harris SK, Shih MC, et al. (2008) A simulator-based tool that assesses pediatric resident resuscitation competency. Pediatrics 121: e597-603.

43. Shayne P, Gallahue F, Rinnert S, Anderson CL, Hern G, et al. (2006) Reliability of a core competency checklist assessment in the emergency department: the Standardized Direct Observation Assessment Tool. Acad Emerg Med 13:727-732.

44. Grant EC, Grant VJ, Bhanji F, Duff JP, Cheng A, et al. (2012) The development and assessment of an evaluation tool for pediatric resident competence in leading simulated pediatric resuscitations. Resuscitation 83: 887-893.

45. Wright MC, Phillips-Bute BG, Petrusa ER, Griffin KL, Hobbs GW, et al. (2009) Assessing teamwork in medical education and practice: relating behavioural teamwork ratings and clinical performance. Med Teach 31: 30-38.

46. Schmutz J, Manser T (2013) Do team processes really have an effect on clinical performance? A systematic literature review. Br J Anaesth 110: 529-544.

47. van der Vleuten CPM, Schuwirth LW (2005) Assessing professional competence: From methods to programmes. Med Educ 39: 309-317.

48. Epstein RM (2007) Assessment in medical education. N Engl J Med 356: 387-396.