


ADVANCING SIMULATION PRACTICE

Open Access



Simulation-based summative assessment in healthcare: an overview of key principles for practice

Clément Buléon^{1,2,3*} , Laurent Mattatia⁴, Rebecca D. Minehart^{3,5,6}, Jenny W. Rudolph^{3,5,6}, Fernande J. Lois⁷, Erwan Guillouet^{1,2}, Anne-Laure Philippon⁸, Olivier Brissaud⁹, Antoine Lefevre-Scelles¹⁰, Dan Benhamou¹¹, François Lecomte¹², the SoFraSimS Assessment with simulation group, Anne Bellot, Isabelle Crublé, Guillaume Philippot, Thierry Vanderlinden, Sébastien Batrancourt, Claire Boithias-Guerot, Jean Bréaud, Philine de Vries, Louis Sibert, Thierry Sécheresse, Virginie Boulant, Louis Delamarre, Laurent Grillet, Marianne Jund, Christophe Mathurin, Jacques Berthod, Blaise Debien, Olivier Gacia, Guillaume Der Sahakian, Sylvain Boet, Denis Oriot and Jean-Michel Chabot

Abstract

Background: Healthcare curricula need summative assessments relevant to and representative of clinical situations to best select and train learners. Simulation provides multiple benefits with a growing literature base proving its utility for training in a formative context. Advancing to the next step, “the use of simulation for summative assessment” requires rigorous and evidence-based development because any summative assessment is high stakes for participants, trainers, and programs. The first step of this process is to identify the baseline from which we can start.

Methods: First, using a modified nominal group technique, a task force of 34 panelists defined topics to clarify the why, how, what, when, and who for using simulation-based summative assessment (SBSA). Second, each topic was explored by a group of panelists based on state-of-the-art literature reviews technique with a snowball method to identify further references. Our goal was to identify current knowledge and potential recommendations for future directions. Results were cross-checked among groups and reviewed by an independent expert committee.

Results: Seven topics were selected by the task force: “What can be assessed in simulation?”, “Assessment tools for SBSA”, “Consequences of undergoing the SBSA process”, “Scenarios for SBSA”, “Debriefing, video, and research for SBSA”, “Trainers for SBSA”, and “Implementation of SBSA in healthcare”. Together, these seven explorations provide an overview of what is known and can be done with relative certainty, and what is unknown and probably needs further investigation. Based on this work, we highlighted the trustworthiness of different summative assessment-related conclusions, the remaining important problems and questions, and their consequences for participants and institutions of how SBSA is conducted.

Conclusion: Our results identified among the seven topics one area with robust evidence in the literature (“What can be assessed in simulation?”), three areas with evidence that require guidance by expert opinion (“Assessment tools for

*Correspondence: clement.buleon@unicaen.fr

¹ Department of Anesthesiology, Intensive Care and Perioperative Medicine, Caen Normandy University Hospital, 6th Floor, Caen, France
Full list of author information is available at the end of the article



© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

SBSA, “Scenarios for SBSA”, “Implementation of SBSA in healthcare”), and three areas with weak or emerging evidence (“Consequences of undergoing the SBSA process”, “Debriefing for SBSA”, “Trainers for SBSA”). Using SBSA holds much promise, with increasing demand for this application. Due to the important stakes involved, it must be rigorously conducted and supervised. Guidelines for good practice should be formalized to help with conduct and implementation. We believe this baseline can direct future investigation and the development of guidelines.

Keywords: Medical education, Summative, Assessment, Simulation, Education, Competency-based education

Background

There is a critical need for summative assessment in healthcare education [1]. Summative assessment is high stakes, both for graduation certification and for recertification in continuing medical education [2–5]. Knowing the consequences, the decision to validate or not validate the competencies must be reliable, based on rigorous processes, and supported by data [6]. Current methods of summative assessment such as written or oral exams are imperfect and need to be improved to better benefit programs, learners, and ultimately patients [7]. A good summative assessment should sufficiently reflect clinical practice to provide a meaningful assessment of competencies [1, 8]. While some could argue that oral exams are a form of verbal simulation, hands-on simulation can be seen as a solution to complement current summative assessments and enhance their accuracy by bringing these tools closer to assessing the necessary competencies [1, 2].

Simulation is now well established in the healthcare curriculum as part of a modern, comprehensive approach to medical education (e.g., competency-based medical education) [9–11]. Rich in various modalities, simulation provides training in a wide range of technical and non-technical skills across all disciplines. Simulation adds value to the educational training process particularly with feedback and formative assessment [9]. With the widespread use of simulation in the formative setting, the next logical step is using simulation for summative assessment.

The shift from formative to summative assessment using simulation in healthcare must be thoughtful, evidence-based, and rigorous. Program directors and educators may find it challenging to move from formative to summative use of simulation. There are currently limited experiences (e.g., OSCE [12, 13]) but not established guidelines on how to proceed. The evidence needed for the feasibility, the validity, and the definition of the requirement for simulation-based summative assessment (SBSA) in healthcare education has not yet been formally gathered. With this evidence, we can hope to build a rigorous and fair pathway to SBSA.

The purpose of this work is to review current knowledge for SBSA by clarifying the guidance on why, how,

what, when, and who. We aim at identifying areas (i) with robust evidence in the literature, (ii) with evidence that requires guidance by expert opinion, and (iii) with weak or emerging evidence. This may serve as a basis for future research and guideline development for the safe and effective use of SBSA (Fig. 1).

Methods

First, we performed a modified Nominal Group Technique (NGT) to define the further questions to be explored in order to have the most comprehensive understanding of SBSA. We followed recommendations on NGT for conducting and reporting this research [14]. Second, we conducted state-of-the-art literature reviews to assess the current knowledge on the questions/topics identified by the modified NGT. This work did not require Institutional Review Board involvement.

Context

A discussion on the use of SBSA was led by executive committee members of the *Société Francophone de Simulation en Santé* (SoFraSimS) in a plenary session and involved congress participants in May 2018 at the SoFraSimS annual meeting in Strasbourg, France. Key points addressed during this meeting were the growing interest in using SBSA, its informal uses, and its inclusion in some formal healthcare curricula. The discussion identified that these important topics lacked current guidelines. To reduce knowledge gaps, the SoFraSimS executive committee assigned one of its members (FL, one of the authors) to lead and act as a NGT facilitator for a task force on SBSA. The task force’s mission was to map the current landscape of SBSA, the current knowledge and gaps; and potentially to identify experts’ recommendations.

Task force characteristics

The task force panelists were recruited among volunteer simulation healthcare trainers in French-speaking countries after a call for application by SoFraSimS in May 2019. Recruiting criteria were a minimum of 5 years of experience in simulation and a direct involvement in simulation programs development or currently running. There were 34 (12 women and 22 men) from 3 countries

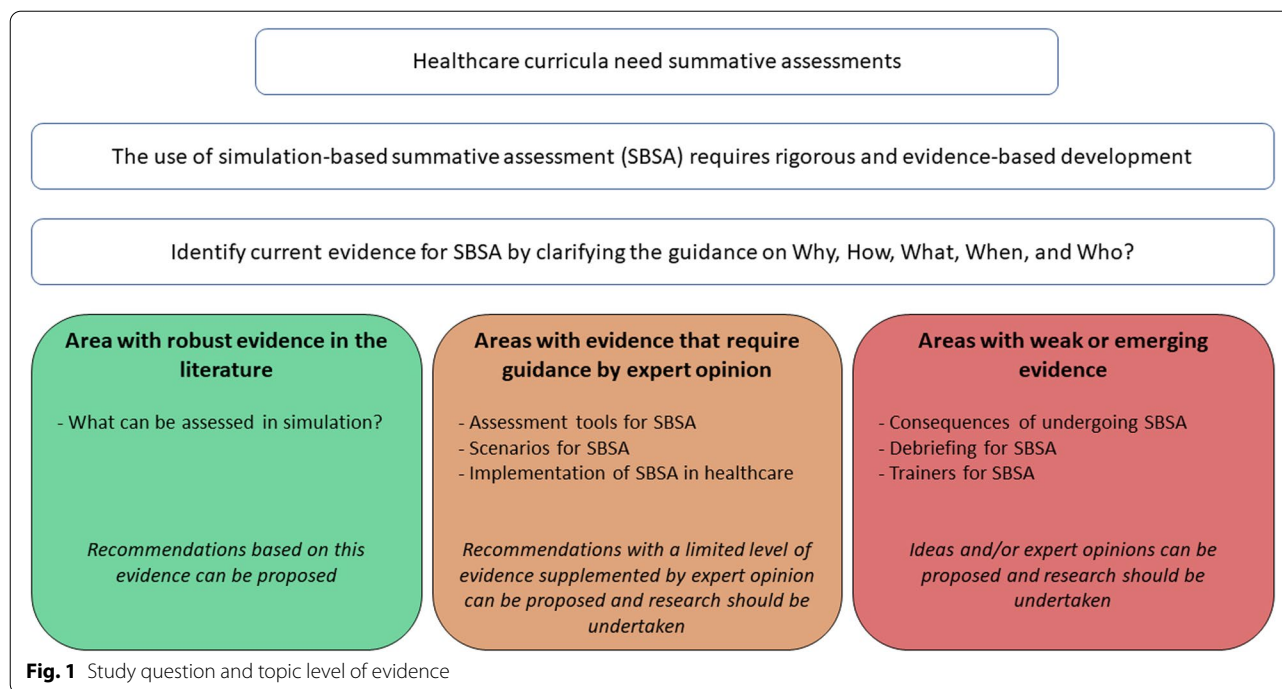


Fig. 1 Study question and topic level of evidence

(Belgium, France, Switzerland) included. Twenty-three were physicians and 11 were nurses, while 12 total had academic positions. All were experienced trainers in simulation for more than 7 years and were involved or responsible for initial training or continuing education programs with simulation. The task force leader (FL) was responsible for recruiting panelists, organizing, and coordinating the modified NGT, synthesizing responses, and writing the final report. A facilitator (CB) assisted the task force leader with the modified NGT, the synthesis of responses, and the writing of the final report. Both NGT facilitators (FL and CB) had more than 14 years of experience in simulation, had experience in research in simulation, and were responsive to simulation programs development and running.

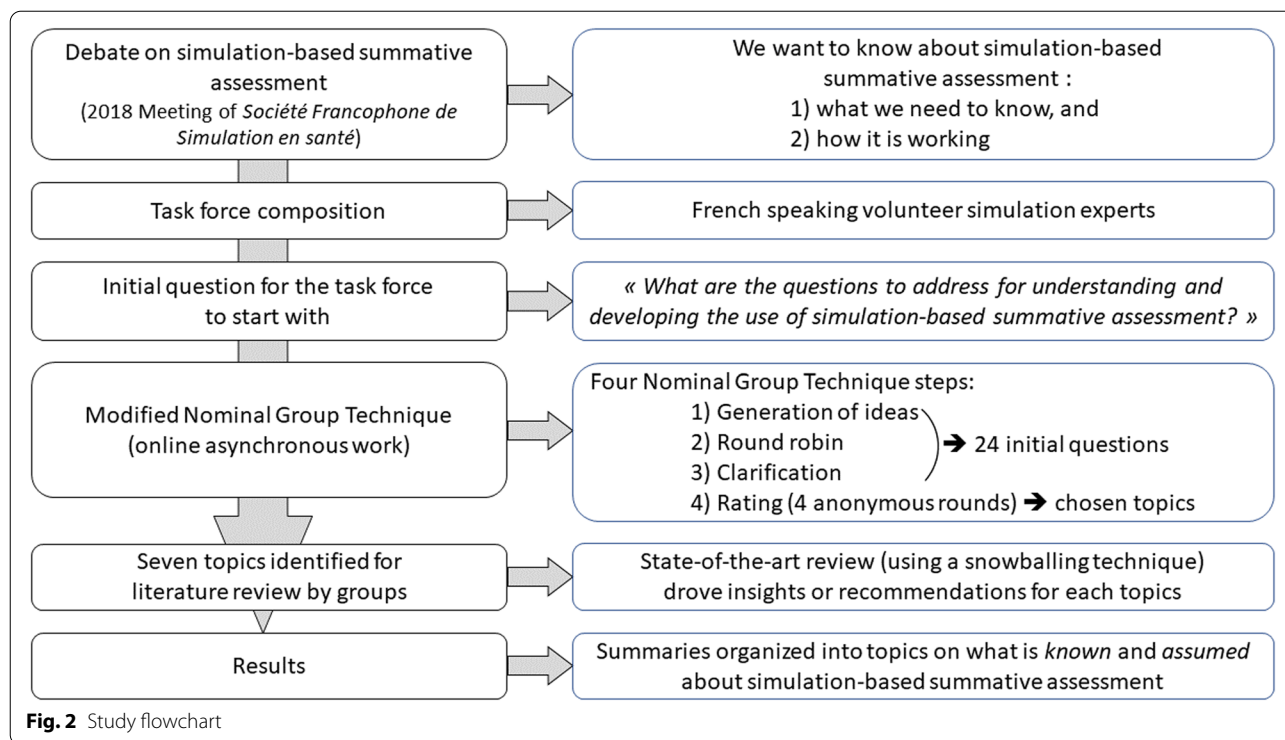
First part: initial question and modified nominal group technique (NGT)

To answer the challenging question of “What do we need to know for a safe and effective SBSA practice?”, following the French *Haute Autorité de Santé* guidelines [15], we applied a modified nominal group technique (NGT) approach [16] between September and October 2019. The goal of our modified NGT was to define the further questions to be explored to have the most comprehensive understanding of the current SBSA (Fig. 2). The modifications to NGT included interactions that were not in-person and were asynchronous for some. Those modifications were introduced as a result of the

geographic dispersion of the panelists across multiple countries and the context of the COVID-19 pandemic.

The first two steps of the NGT (generation of ideas and round robin) facilitated by the task force leader (FL) were conducted online simultaneously and asynchronously via email exchanges and online surveys over a 6-week period. For the initiation of the first step (generation of ideas), the task force leader (FL) sent an initial non-exhaustive literature review of 95 articles and proposed the initial following items for reflection: definition of assessment, educational principles of simulation, place of summative assessment and its implementation, assessment of technical and non-technical skills in initial training, continuing education, and interprofessional training. The task force leader (FL) asked the panelists to formulate topics or questions to propose for exploration in Part 2 based on their knowledge and the literature provided. Panelists independently elaborated proposals and sent them back to the task force leader (FL) who regularly synthesized them and sent the status of the questions/topics to the whole task force while preserving the anonymity of the contributors and asking them to check the accuracy of the synthesized elements (second step, as a “round robin”).

The third step of the NGT (clarification) was carried out during a 2-h video conference session. All panelists were able to discuss the proposed ideas, group the ideas into topics, and make the necessary clarifications. As a result of this step, 24 preliminary questions were



defined for the fourth step (Supplemental Digital Content 1).

The fourth step of the NGT (vote) consisted of four distinct asynchronous and anonymous online vote rounds that led to a final set of topics with related sub-questions (Supplemental Digital content 2). Panelists were asked to vote to regroup, separate, keep, or discard questions/topics. All vote rounds followed similar validation rules. We [NGT facilitators (FL and CB)] kept items (either questions or topics) with more than 70% approval ratings by panelists. We reworded and resubmitted in the next round all items with 30–70% approval. We discarded items with less than 30% approval. The task force discussed discrepancies and achieved final ratings with a complete agreement for all items. For each round, we sent reminders to reach a minimum participation rate of 80% of the panelists. Then, we split the task force into 7 groups, one for each of the 7 topics defined at the end of the vote (step 4).

Second part: literature review

From November 2019 to October 2020, the groups each identified existing literature containing the current knowledge, and potential recommendations on the topic they were to address. This identification was done based on a non-systematic review of the existing literature. To identify existing literature, the groups conducted state-of-the-art reviews [17] and expanded their reviews with

a snowballing literature review technique [18] based on the articles’ references. The selected literature search performed by each group was inserted into the task force’s common library on SBSA in healthcare as it was conducted.

For references, we searched electronic databases (MEDLINE), gray literature databases (including digital theses), simulation societies and centers’ websites, generic web searches (e.g., Google Scholar), and reference lists from articles. We selected publications related to simulation in healthcare with keywords “summative assessment,” “summative evaluation,” and also specific keywords related to topics. The search was iterative to seek all available data until saturation was achieved. Ninety-five references were initially provided to the task force by the NGT facilitator leader (FL). At the end of the work, the task force common library contained a total of 261 references.

Techniques to enhance trustworthiness from primary reports to the final report

The groups’ primary reports were reviewed and critiqued by other groups. After group cross-reviewing, primary reports were compiled by NGT facilitators (FL and CB) in a single report. This report, responding to the 7 topics, was drafted in December 2020 and submitted as a single report to an external review committee composed of 4 senior experts in education, training, and research

from 3 countries (Belgium, Canada, France) with at least 15 years of experience in simulation. NGT facilitators (FL and CB) responded directly to reviewers when possible and sought assistance from the groups when necessary. The final version of the report was approved by the SoF-rSimS executive committee in January 2021.

Results

First part: modified nominal group technique (NGT)

The first two steps of the NGT by their nature (generation of ideas and “round robin”) did not provide results. The third step (clarification phase), identified 24 preliminary questions (Supplemental digital content 1) to be submitted to the fourth step (vote). The 4 rounds of voting (step 4) resulted in 7 topics with sub-questions (Supplemental Digital content 2): (1) “What can be assessed in simulation?” (2) “Assessment tools for SBSA,” (3) “Consequences of undergoing the SBSA process,” (4) “Simulation scenarios for SBSA,” (5) “Debriefing, video, research and SBSA strategies,” (6) Trainers for SBSA,” (7) “Implementation of SBSA in healthcare.” These 7 topics and their sub-questions were the starting point for the state-of-the-art literature reviews of each group for the second part.

Second part: literature review

For each of the 7 topics, the groups highlighted what appears to be validated in the literature, the remaining important problems and questions, and their consequences for participants and institutions of how SBSA is conducted. Results in this section present the major ideas and principles from the literature review, including their nuances where necessary.

What can be assessed in simulation?

Healthcare faculty and institutions must ensure that each graduate is practice ready. Readiness to practice implies mastering certain competencies, which is dependent on learning them appropriately. The competency approach involves explicit definitions of the acquired core competencies necessary to be a “good professional.” Professional competency could be defined as the ability of a professional to use judgment, knowledge, skills, and attitudes associated with their profession to solve complex problems [19–21]. Competency is a complex “knowing how to act” based on the effective mobilization and combination of a variety of internal and external resources in a range of situations [19]. Competency is not directly observable; it is the performance in a situation that can be observed [19]. Performance can vary depending on human factors such as stress, fatigue, etc.... During simulation, competencies can be assessed by observing “key” actions using assessment tools [22]. Simulation’s limitations must

consider when defining the assessable competencies. Not all simulation methods are equivalent to assessing specific competencies [22].

Most healthcare competencies can be assessed with simulation, throughout at curriculum, if certain conditions are met. First, the competency being assessed summatively must have already been assessed formatively with simulation [23, 24]. Second, validated assessment tools must be available to conduct this summative assessment [25, 26]. These tools must be reliable, objective, reproducible, acceptable, and practical [27–30]. The small number of currently validated tools limits the use of simulation for competency certification [31]. Third, it is not necessary or desirable to certify all competencies [32]. The situations chosen must be sufficiently frequent in the student’s future professional practice (or potentially impactful for the patient) and must be hard or impossible to assess and validate in other circumstances (e.g., clinical internships) [2]. Fourth, simulation can be used for certification throughout the curriculum [33–35]. Finally, limitations for the use of simulation throughout the curriculum may be a lack of logistical resources [36]. Based on our findings in the literature, we have summarized in Table 1 the educational consideration when implementing a SBSA.

Assessment tools for simulation-based summative assessment

One of the challenges of assessing competency lies in the quality of the measurement tools [31]. A tool that allows the raters to collect data must also allow them to give meaning to their assessment, while securing that it is really measuring what it aims to [25, 37]. A tool must be valid and, capable of measuring the assessed competency with fidelity and, reliability while providing reproducible data [38]. Since a competency is not directly measurable, it will be analyzed on the basis of learning expectations, the most “concrete” and observable form of a competency [19]. Several authors have described definitions of the concept of validity and the steps to achieve it [38–41]. Despite different validation approaches, the objectives are similar: to ensure that the tool is valid, the scoring items reflect the assessed competency, and the contents are appropriated for the targeted learners and raters [20, 39, 42, 43]. A tool should have psychometric characteristics that allow users to be confident of its reproducibility, discriminatory nature, reliability, and external consistency [44]. A way to ensure that a tool has acceptable validity is to compare it to existing and validated tools that assess the same skills for the same learners. Finally, it is important to consider the consequences of the test to determine whether it best discriminates competent students from others [38, 43].

Table 1 Considerations for implementing a summative assessment with simulation

| Considerations | Elements | Items | Example adapted to cardiopulmonary resuscitation (CPR) for an emergency physician |
|----------------------------------|--|--|--|
| Competency to be assessed | Clear definition of competency | <p>Know how to act in a professional situation</p> <p>Identify <i>internal resources</i>: knowledge, skills, behavior, and reasoning</p> <p>Identify <i>external resources</i>: equipment, written or electronic resources), colleagues, and so on to mobilize</p> | <p>The practitioner is able to handle an in-hospital cardiac arrest (CA)</p> <p>ACLS algorithm, airway management, leadership, management according to the type of CA (e.g. asystole, pulseless electrical activity, ventricular fibrillation)</p> <p>e.g., defibrillator, cognitive aids (a chart, a checklist, ...), ECMO team, ...</p> |
| Assessment | <p>Number of competencies</p> <p>Measurements</p> <p>Context authenticity</p> | <p>Consider the possibility of assessing one or more competencies simultaneously</p> <p>Consider measuring performance in representative and diverse situations</p> <p>Complex problems</p> <p>Adapt the complexity to the training level</p> <p>Ensure context relevance to future or current professional practice</p> <p>Interprofessional situations (vs uniprofessional)</p> | <p>In-hospital CA alone, or CA in adult patient and/or in specific conditions (e.g. child, pregnant, ...)</p> <p>CA in a young polytrauma patient, in an elderly diabetic patient, in a pregnant woman or in a child out-of-hospital e.g., CA due to hyperkalemia in a patient with renal failure</p> <p>Complexity may be tuned for an expert with patient's chronic use of beta-blockers</p> <p>CA occurs in an ambulance or in an emergency room or in OR or in ICU</p> <p>Prefer a situation where the learner is not alone such as a member of an emergency team and not as a first responder in the street</p> |
| | Standardization | <p>Tasks and requirements known before by the participants</p> <p>Direct observation associated with a phase of student interaction (questioning)</p> <p>Rate with a checklist or a rubric</p> <p>Multiple sources and/or iteration (e.g., repeated performances of the same scenario)</p> <p>Clear and specific objectives</p> <p>Adjusted to the assessed knowledge or to the simulation</p> <p>Integration of self-assessment</p> <p>Consider only important errors</p> <p>Strategies (cognitive and metacognitive) assessed during the interaction phase</p> <p>Prior consensus on rating and definition regarding expected level of development</p> | <p>Send to the learner the assessment template prior to the assessment.</p> <p>The simulation is followed by a debriefing (feedback)</p> <p>e.g., time from the start of VF to the first external electric shock and/or compliance with ACLS steps and/or quality of external cardiac massage (visual and/or via sensors)</p> <p>Only items that have been previously decided are assessed (see above)</p> <p>It is not possible to assess the use of the defibrillator if the situation is pulseless electrical activity</p> <p>6 instead of 5 min between 2 doses of adrenaline (minor error) versus no recognition of a shockable rhythm (major error)</p> <p>Ask questions during feedback phase: "Can you remind me of the administration schedule for epinephrine in CA?" (cognition). "I have observed that you administered it every minute, but as you have just said and as I think it is every 3 to 5 min, could you explain why in the situation you administered it every minute?" (metacognition)</p> <p>Identify minor and major errors together (all instructors involved in the assessment of this competency). Define the number of acceptable minor and/or major errors to validate the acquisition or not of the competency at this level of development</p> |
| | Correction criteria | | |

Table 1 (continued)

| Considerations | Elements | Items | Example adapted to cardiopulmonary resuscitation (CPR) for an emergency physician |
|------------------------|---|--|---|
| Scenarios | Development | <p>Developing scenarios only after defining the skills and or competences to be assessed</p> <p>Ensuring the scenario reflects professional reality</p> <p>Incorporating the targeted skills into a scenario representing professional practice, rather than a task trainer, for example</p> | <p>e.g., if we want to evaluate the use of the defibrillator, we need to construct a scenario where the patient has VF or VT</p> <p>e.g., use a hyperkalemia CA scenario after a burial extraction but not when releasing a tourniquet after a knee replacement for an emergency physician</p> <p>Prefer to use a scenario with a clinical history of CA to assess CPR skills rather than performing CPR in a skill station</p> |
| | Multiple skills | <p>Several stations with short scenarios (e.g., 5–6 min) each are preferable to long scenarios (e.g., > 20 min)</p> <p>Critical situation</p> | <p>Ensure that all steps can be assessed. E.g., the use of ECMO is reserved for refractory CA and cannot be considered if the scenario lasts for 5 min and begins with the recognition of the arrest. In this case, a scenario with a CA that has already been under management for 15 min should be used</p> |
| | Test prior to use | <p>Validity, reliability, reproducibility</p> | <p>The scenarios used should be pre-tested by the teaching team including using the assessment forms</p> |
| Assessment test | Simulators (High and low-Technology) | <p>Use and difficulty level validated</p> | <p>e.g., if intubation is expected during the scenario, the chosen manikin should allow it</p> |
| | standardization (Fairness) | <p>Facilitator's role and intervention specified in advance</p> <p>Only one candidate per station</p> | <p>What can the facilitator do? E.g., can he/she guide on 4H-4 T if the learner does not think about it?</p> |
| | Practical conditions | <p>Minimum number of scenarios (8 to 15) [157]</p> <p>Incentive to verbalize after action (Reasoning, what is done or not done)</p> | <p>Scenarios in different circumstances (in and out-of-hospital), different causes (4H-4 T), different ages (child to elderly adult) To be recalled in the pre-briefing</p> |
| | Raters | <p>At least, two raters</p> <p>Ideally, a rater should be involved in the formative assessment program</p> | <p>e.g., clinical supervisor, ACLS instructor, simulation instructor who has supervised the learner during the formative sessions, ...</p> |

Like a diagnostic score, a relevant assessment tool must be specific [30, 39, 41]. It is not good or bad, but valid through a rigorous validation process [39, 41, 42]. This validation process determines whether the tool measures what it is supposed to measure and whether this measurement is reproducible at different times (test–retest) or with 2 observers simultaneously. It also determines if the tool results are correlated with another measure of the same ability or competency and if the consequences of the tool results are related to the learners’ actual competency [38].

Following Messick’s framework, which aimed to gather different sources of validity in one global concept (unified validity), Downing describes five sources of validity, which must be assessed with the validation process [38, 45, 46]. Table 2 presents an illustration of the development used in SBSA according to the unified validity

framework for a technical task [38, 45, 46]. An alternative framework using three sources of validity for teamwork’s non-technical skills are presented in Table 3.

A tool is validated in a language. Theoretically, this tool can only be used in this language, given the nuances present with interpretation [49]. In certain circumstances, a “translated” tool, but not a “translated and validated in a specific language” tool, can lead to semantic biases that can affect the meaning of the content and its representation [49–55]. For each assessment sequence, validity criteria consist of using different tools in different assessment situations and integrating them into a comprehensive program which considers all aspects of competency. The rating made with a validated tool for one situation must be combined with other assessment situations, since there is no “ideal” tool [28, 56]. A given tool can be used with different professions or with participants

Table 2 Example of the development of a tool to assess technical skill achievement in a simulated situation, based on work by Oriot et al., Downing, and Messick’s framework [38, 46, 47]

| Source of validity | Method | Judgment criteria | Results |
|---------------------------------|--|---|--|
| content | 1. Description of the checklist development by 2 experts 2. Review by 2 outside experts 3. Definitive Checklist | Relevance of items Adapted illustration of the skill Conditions of skill achievement | Obtaining a list of 12 items (after the initial proposal of 20 items) |
| Response process | Pilot study, search for error sources Adapting items Defining units of measurement | Interrater reproducibility Item content (redundant, inaccurate) Controlling the sources of measurement errors Weighing items | Fusion/removal of redundant items Minutes, degrees, centimeters justification |
| Internal structure | Internal coherence Reproducibility Discrimination of learners | Cronbach Alpha Coefficient, interrater: Cohen Kappa, ICC | Cronbach result Correlation between 2 raters |
| Comparison with other variables | Score vs success or failure of the procedure Score vs theoretical assessment Score vs previous experience/level of expertise | Correlation between procedure success or theoretical assessment and score with the tool | Time for success, score for success and rating |
| Consequences | Minimum passing score | Pass-fail score with procedure success | 14/20 |

Table 3 Example of the development of an assessment tool for the observation of teamwork in simulation [48]

| Source of validity | Method | Judgment criteria | Results |
|--------------------|---|---|--|
| Content | 1. Description of the <i>Clinical Teamwork Scale (CRM scale)</i> Development | Literature review Scale already used in another field (aeronautics) | 15 items 5 categories 1 overall skill score |
| Response process | 1. Relevance of items 2. weighting items 3. Raters’ training (moderate) | 1. Precise description of each item 2. Quantitative criteria 3. Qualitative criteria 4. CRM principles | 1. Ratings aid table 2. 0 to 10 or 0/1 Descriptive levels: not relevant/unacceptable/poor/average/good/perfect |
| Internal structure | 1. Built-in validity 2. Scale usability 3. Reproducibility | 1. Distribution of scores from the preset level 2. Number of items filled in full 3. interrater concordance, the correlation between overall score and categories (Kappa, Kendall, Pearsons, ICC) 4. Variance of each category | 1. Score tailored to each level 2. Easy-to-use scale (little loss of information) 3. correlation between raters 4. Variation in scores between scenarios sources of error |

at different levels of expertise or in different languages if it is validated for these situations [57, 58]. In a summative context, a tool must have demonstrated a high-level of validity to be used because of the high stake for the participants [56]. Finally, the use or creation of an assessment tool requires trainers to question its various aspects, from how it was created to its reproducibility and the meaning of the results generated [59, 60].

Two types of assessment tools should be distinguished: tools that can be adapted to different situations and tools that are specific to a situation [61]. Thus, technical skills may have a dedicated assessment tool (e.g., intraosseous) [47] or an assessment grid generated from a list of pre-established and validated items (e.g., TAPAS scale) [62]. Non-technical skills can be observed using scales that are not situation-specific (e.g., ANTS, NOTECHS) [63, 64] or that are situation-specific (e.g., TEAM scale for resuscitation) [57, 65]. Assessment tools should be provided to participants and should be included in the scenario framework, at least as a reference [66–69]. In the summative assessment of a procedure, structured assessment tools should probably be used, using a structured objective assessment form for technical skills [70]. The use of a scale, in the context of the assessment of a technical gesture, seems essential. As with other tools, any scale must be validated beforehand [47, 70–72].

Consequences of undergoing the simulation-based summative assessment process

Summative assessment has two notable consequences on learning strategies. First, it may drive the learner's behavior during the assessment, while it is essential to assess the competencies targeted, not the ability of the participant to adapt to the assessment tool [6]. Second, the pedagogy key concept of "pedagogical alignment" must be respected [23, 73]. It means that assessment methods must be coherent with the pedagogical activities and objectives. For this to happen, participants must have formative simulation training focusing on the assessed competencies prior to the SBSA [24].

Participants have been reported as experiencing commonly mild (e.g., appearing slightly upset, distracted, teary-eyed, quiet, or resistant to participating in the debriefing) or moderate (e.g., crying, making loud, and frustrated comments) psychological events in the simulation [74]. While voluntary recruitment for formative simulation is commonplace, all students are required to take summative assessments in training. This required participation in high-stake assessment may have a more consequential psychological impact [26, 75]. This impact can be modulated by training and assessment conditions [75]. First, the repetition of formative simulations reduces the psychological impact of SBSA on participants [76].

Second, the transparency on the objectives and methods of assessment limits detrimental psychological impact [77, 78]. Finally, detrimental psychological impacts are increased by abnormally high physiological or emotional stress such as fatigue, and stressful events in the 36 h preceding the assessment, such that students with a history of post-traumatic stress disorder or psychological disorder may be strongly and negatively impacted by the simulation [76, 79–81].

It is necessary to optimize SBSA implementation to limit its pedagogical and psychological negative impacts. Ideally, during the summative assessment, it has been proposed to take into account the formative assessment that has already been carried out [1, 20, 21]. Similarly in continuing education, the professional context of the person assessed should be considered. In the event of failure, it will be necessary to ensure sympathetic feedback and to propose a new assessment if necessary [21].

Scenarios for simulation-based summative assessment

Some authors argue that there are differences between summative and formative assessment scenarios [76, 79–81]. The development of a SBSA scenario begins with the choice of a theme, which is most often agreed upon by experts at the local level [66]. The themes are most often chosen based on the participants' competencies to be assessed and included in the competencies requirement for the initial [82] and continuing education [35, 83]. A literature review even suggests the need to choose themes covering all the competencies to be assessed [41]. These choices of themes and objectives also depend on the simulation tools technically available: "The themes were chosen if and only if the simulation tools were capable of reproducing "a realistic simulation" of the case." [84].

The main quality criterion for SBSA is that the cases selected and developed are guided by the assessment objectives [85]. It is necessary to be clear about the assessment objectives of each scenario to select the right assessment tool [86]. Scenarios should meet four main principles: predictability, programmability, standardizability, and reproducibility [25]. Scenario writing should include a specific script, cues, timing, and events to practice and assess the targeted competencies [87]. The implementation of variable scenarios remains a challenge [88]. Indeed, most authors develop only one scenario per topic and skill to be assessed [85]. There are no recommendations for setting a predictable duration for a scenario [89]. Based on our findings we suggest some key elements for structuring a SBSA scenario in Table 4. For technical skill assessment, scenarios will be short and the assessment is based on an analytical score [82, 89]. For non-technical skill assessment, scenarios will be longer

Table 4 Key element structuring a summative assessment scenario

| Elements | Recommendations |
|-----------------|---|
| Duration | 10 to 15 min Short for technical skills Longer for non-technical skills |
| Objectives | Accurate list of competencies and skills to be assessed |
| Essential items | Initial assessment Diagnostic strategy Situation management Orientation strategy |
| Script | Computerized (programed if possible) |
| Rating scale | <i>Checklist, Global Rating Scale</i> Scale (20 to 30 items) Analytic score for technical skills Analytic and holistic (e.g., ANTS) for non-technical skills |
| Validation | Pilot sessions (scenario testing and rater training) 1 or 2 cases per student during scenario testing |
| Assessment | Video rating Cohen's Kappa test for differences between raters Student's <i>t</i> test for the ability to discriminate between students |

and the assessment based on analytical and holistic scores [82, 89].

Debriefing, video, and research for simulation-based summative assessment

Studies have shown that debriefings are essential in formative assessment [90, 91]. No such studies are available for summative assessment. Good practice requires debriefing in both formative and summative simulation-based assessments [92, 93]. In SBSA, debriefing is often brief feedback given at the end of the simulation session, in groups [85, 94, 95], or individually [83]. Debriefing can also be done later with a trainer and help of video, or via written reports [96]. These debriefings make it possible to assess clinical skills for summative assessment purposes [97]. Some tools have been developed to facilitate this assessment of clinical reasoning [97].

Video can be used for four purposes: session preparation, simulation improvement, debriefing, and rating

(Table 5) [95, 98]. In SBSA sessions, video can be used during the prebriefing to provide participants with standardized and reproducible information [99]. A video can increase the realism of the situation during the simulation with ultrasound loops and laparoscopy footage. Simulation records can be reviewed either for debriefing or rating purposes [34, 71, 100, 101]. A video is very useful for the training raters (e.g., for calibration and recalibration) [102]. It enables raters to rate the participants' performance offline and to have an external review if necessary [34, 71, 101]. Despite the technical difficulties to be considered [42, 103], it can be expected that video-based automated scoring assistance will facilitate assessments in the future.

The constraints associated with the use of video rely on the participants' agreement, the compliance with local rules, and that the structure in charge of the assessment with video secures the protection of the rights of individuals and data safety, both at a national and at the higher (e.g., European GDPR) level [104, 105].

In Table 5, we list the main uses of video during simulation sessions found in the literature.

Research in SBSA can focus, as in formative assessment, on the optimization of simulation processes (programs, structures, human resources). Research can also explore the development and validation of summative assessment tools, the automation and assistance of assessment resources, and the pedagogical and clinical consequences of SBSA.

Trainers for simulation-based summative assessment

Trainers for SBSA probably need specific skills because of the high number of potential errors or biases in SBSA, despite the quest for objectivity (Table 6) [106]. The difficulty in ensuring objectivity is likely the reason why the use of self or peer assessment in the context of SBSA is not well documented and the literature does not yet support it [59, 60, 107, 108].

SBSA requires the development of specific scenarios, staged in a reproducible way, and the mastery of

Table 5 Uses of video for simulation-based formative and summative assessment

| | Formative assessment | Summative assessment |
|---|---|--|
| Prebriefing | Participant information | |
| Simulation | Increased scenario realism (e.g., coelioscopy video) Watching by observers | |
| Immediate visualization after simulation | Self-assessment Debriefing by trainers (selected sequences) | No self-assessment (in the literature) |
| Delayed visualization | Learning teamwork or skills for a formative purpose | Deferred debriefing Rater training (calibration and recalibration) Administrative evidence |

Table 6 Potential errors, effects, and bias in simulation-based summative assessment [109, 110]

| Type of error | Error description |
|-------------------------|---|
| Error of homogenization | Tendency to rate neither too good or too bad, making discrimination more difficult |
| Halo effect | Tendency to see everything right or wrong in the same performance |
| Time effect | Bias related to observations of early or late good or bad performance during sessions |
| Bias of “clemency” | Willingness not to give bad grades |
| Repository error | Judgment based on what the rater would have done and not on the assessment tool |
| Group effect | Evaluation based on the team’s performance rather than the participant’s performance |

assessment tools to avoid assessment bias [111–114]. Fulfilling those requirements calls for specific abilities to fit with the different roles of the trainer. These different roles of trainers would require specific initial and ongoing training tailored to their tasks [111, 113]. In the future, concepts of the roles and tasks of these trainers should be integrated into any “training of trainers” in simulation.

Implementation of simulation-based summative assessment in healthcare

The use of SBSA has been described by Harden in 1975 with Objective and Structured Clinical Examination (OSCE) tests for medical students [115]. The summative use of simulation has been introduced in different ways depending on the professional field and the country [116]. There is more literature on certification at the undergraduate and graduate levels than on recertification at the postgraduate level. The use of SBSA in re-certification is currently more limited [83, 117]. Participation is often mandated, and it does not provide a formal assessment of competency [83]. Some countries are defining processes for the maintenance of certification in which simulation is likely to play a role (e.g., in the USA [118] and France [116]). Recommendations regarding the development of SBSA for OSCE were issued by the AMEE (Association for Medical Education in Europe) in 2013 [12, 119]. Combined with other recommendations that address the organization of examinations on other immersive simulation modalities, in particular, full-scale sessions using complex mannequins [22, 85], they give us a solid foundation for the implementation of SBSA.

The overall process to ensure a high-quality examination by simulation is therefore defined but particularly demanding. It mobilizes many material and human resources (administrative staff, trainers, standardized patients, and healthcare professionals) and requires a long development time (several months to years depending on the stakes) [36]. We believe that the steps to overcome during the implementation of SBSA range from setting up a coordination team, to supervising the writers, the raters, and the standardized patients, as well as taking into account the logistical and practical pitfalls.

The development of a competency framework valid for an entire curriculum (e.g., medical studies) satisfies a fundamental need [7, 120]. This development allows identifying competencies to be assessed with simulation, those to be assessed by other methods, and those requiring triangulation by several assessment methods. This identification then guides scenarios’ writing and examination’s development with good content validity. Scenarios and examinations will form a bank of reproducible assessment exercises. The examination quality process, including psychometric analyses, is part of the development process from the beginning [85].

We have summarized in Table 7 the different steps in the implementation of SBSA.

Recertification Recertification programs for various healthcare domains are currently being implemented or planned in many countries (e.g., in the USA [118] and France [116]). This is a continuation of the movement to promote the maintenance of competencies. Examples can be cited in France with the creation of an agency for continuing professional development or in the USA with the Maintenance Of Certification [83, 126]. The certification of health care facilities and even teams is also being studied [116]. Simulation is regularly integrated into these processes (e.g., in the USA [118] and France [116]). Although we found some commonalities basis between the certification and recertification processes, there are many differences (Table 8).

Currently, when simulation-based training is mandatory (e.g., within the American Board of Anesthesiology’s “Maintenance Of Certification in Anesthesiology,” or MOCA 2.0® in the US), it is most often a formative process [34, 83]. SBSA has a place in the recertification process, but there are many pitfalls to avoid. In the short term, we believe that it will be easier to incorporate formative sessions as a first step. The current consensus seems to be that there should be no pass/fail recertification simulation without personalized global professional support,

Table 7 Implementation of simulation-based summative assessment step by step

| Items | Goals | Modalities |
|-------------------------------|---|---|
| Team | Identify the training staff | Structure coordination Size the team: skills, time available, stability (project over several months/years) |
| Competencies repository | Create the competencies repository to be assessed | Expert panels Define the number and type of examination needed Must be known to students |
| Curriculum | integrate summative assessment in the curriculum | Pedagogical alignment: summative part drives the formative part of the curriculum No summative assessment without pre-simulation exposure Intermediate summative assessment could be useful [121] |
| Examination | Define summative assessment modalities through simulation | Length and number of scenarios stations [122, 123] The higher the fidelity of the examination, the harder is it to set it up, the lower the feasibility |
| Scenarios | Develop a bank of scenarios and rating grids [124] | Choose the editors for the scenarios Write the scenarios Scenarios' peer-review and test Establish/choose assessment tools (Checklist or global scale) Set the minimum passing score The themes of the bank's scenarios cover the competencies of the repository |
| Training raters | Limit rating variations for a given performance | Choice of raters Raters' Training Workshop |
| Standardized Patients | Develop a standardized patient pool | Recruitment, selection, training, and standardization [125] |
| D-Day | How the examination take place | Logistics: e.g., dates, rooms, standardized patients, rights of personal portrayal, GDPR Participants' path, breaks Materials to supply, to be brought by students (e.g., stethoscope) Examination-adapted briefings Problems to anticipate: e.g., maintenance of standardization, failure or breakage of equipment, backup paper supports, dedicated staff for support to stressed participants, |
| immediately after examination | Finalize the examination | Collect and check assessment grids for early detection of inconsistencies, rating oversights, missing data Management of participants' complaints and plea |
| Quality process | Prepare future examination | Identify potential changes to do to some scenarios Removal of inappropriate scenarios: e.g., too long, misleading, source of rating inconsistency, Changes to standardized patients' training Changes in raters' training |

but which is not limited to a binary aptitude/inaptitude approach [21, 116].

Discussion

Many important issues and questions remain regarding the field of SBSA. This discussion will return to our identified 7 topics and highlight these points, their implications for the future, and some possible leads for future research and guidelines development for the safe and effective use of this tool in SBSA.

What can be assessed in simulation?

SBSA is currently mainly used in initial training in uni-professional and individual settings via standardized patients or task trainers (OSCE) [12, 13]. In the future, SBSA will also be used in continuing education for professionals who will be assessed throughout their career (re-certification) as well as in interprofessional settings [83]. When certifying competencies, it is important to keep in mind the differences between the desired competencies and the observed performances [128]. Indeed, it must be that "what is a competency" is specifically defined [6, 19, 21]. Competencies are what we wish to evaluate during the summative assessment to validate

Table 8 Commonalities and discrepancies between certification and recertification

| Items | Commonalities | Discrepancies |
|---------------------|--|---|
| Modalities | Multimodal process (course, simulation, etc.) [34, 83, 92] Field follow-up opportunities [35] | Low percentage of existing recertification [34, 83] Level of acceptability and feasibility of recertification Level of recertification: pursuing individual certification or switching with team recertification |
| Organization bodies | Accredited centers (functional specification) [34, 83] Same rigor in setting up | Can institutions (universities, schools) in charge of certification, provide recertification? |
| Objectives | Targeted level of competency | Difficulties in the efficient selection of competencies to be assessed with recertification: *Multiple constraints (time/means) *Communication/teamwork, performance gaps, frequent adverse events? Scenarios and assessment tools adapted for learning objectives [127] |
| Consequences | Possible failure of certification or recertification | The impact of a failure to recertification is major for a professional Mandatory discretion of the recertification process Opportunity for screening of professionals in difficulty (burn out...) [92, 116] |
| Funding | Funding difficulties | Many options of financing in recertification (state, professional insurance, etc.) |

or revalidate a professional for his/her practice. Performance is what can be observed during an assessment [20, 21]. In this context, we consider three unresolved issues. The first issue is that an assessment only gives access to a performance at a given moment (“Performance is a snapshot of a competency”), whereas one would like to assess a competency more generally [128]. The second issue is: How does an observed performance—especially in simulation—reveal a real competency in real life? [129] In other words, does the success or failure of a single SBSA really reflect actual real-life competency? [130] The third issue is the assessment of a team performance/competency [131–133]. Until now, SBSA has come from the academic field and has been an individual assessment (e.g., OSCE). Future SBSA could involve teams, driven by governing bodies, institutions, or insurances [134, 135]. The competency of a team is not the sum of the competencies of the individuals who compose it. How can we proceed to assess teams as a specific entity, both composed of individuals and independent of them? To make progress in answering these three issues, we believe it is probably necessary to consider the approximation between observed and assessed performance and competency as acceptable, but only by specifying the scope of validity. Research in these areas is needed to define it and answer these questions.

The consequence of undergoing SBSA has focused on the psychological aspect and have set aside the more usual consequences such as achieving (or not) the minimum passing score. Future research should embrace more global SBSA consequence field, including how reliable SBSA is at determining how someone is competent.

Assessment tools for simulation-based summative assessment

Rigor and method in the development and selection of assessment tools are paramount to the quality of the summative assessment [136]. The literature shows that is necessary that assessment tools be specific to their intended use that their intrinsic characteristics be described and that they be validated [38, 40, 41, 137]. These specific characteristics must be respected to avoid two common issues [1, 6]. The first issue is that of a poorly designed or constructed assessment tool. This tool can only give poor assessments because it will be unable to capture performance correctly and therefore to approach the skill to be assessed in a satisfactory way [56]. The second issue is related to poor or incomplete tool evaluation or inadequate tool selection. If the tool is poorly evaluated, its quality is unknown [56]. The scope of the assessment that is done with it is limited by the imprecision of the tool’s quality. If the tool is poorly selected, it will not accurately capture the performance being assessed. Again, summative assessment will be compromised. It is currently difficult to find tools that meet all the required quality and validation criteria [56]. On the one hand, this requires complex and rigorous work; on the other hand, the potential number of tools required is large. Thus, the overall volume of work to rigorously produce assessment tools is substantial. However, the literature provides the characteristics of validity (content, response process, internal structure, comparison with other variables, and consequences), and the process of developing qualitative and reliable assessment tools [38–41, 45]. It therefore

seems important to systematize the use of these guidelines for the selection, development, and validation of assessment tools [137]. Work in this area is needed and network collaboration could be a solution to move forward more quickly toward a bank of valid and validated assessment tools [39].

Consequences of undergoing the simulation-based summative assessment process

We had focused our discussion on the consequences of SBSA excluding the determining of the competencies and passing rates. Establishing and maintaining psychological safety is mandatory in simulation [138]. Considering the psychological and physiological consequences of SBSA is fundamental to control and limit negative impacts. Summative assessment has consequences for both the participants and the trainers [139]. These consequences are often ignored or underestimated. However, these consequences can have an impact on the conduct or results of the summative assessment. The consequences can be positive or negative. The “testing effect” can have a positive impact on long-term memory [139]. On the other hand, negative psychological (e.g., stress or post-traumatic stress disease), and physiological (e.g., sleep) consequences can occur or degrade a fragile state [139, 140]. These negative consequences can lead to questioning the tools used and the assessments made. These consequences must therefore be logically considered when designing and conducting the SBSA. We believe that strategies to mitigate their impact must be put in place. The trainers and the participants must be aware of these difficulties to better anticipate them. It is a real duality for the trainer: he/she has to carry out the assessment in order to determine a mark and at the same time guarantee the psychological safety of the participants. It seems fundamental to us that trainers master all aspects of SBSA as well as the concept of the safe container [138] to maximize the chances of a good experience for all. We believe that ensuring a fluid pedagogical continuum, from training to (re)certification in both initial and continuing education using modern pedagogical techniques (e.g., mastery learning, rapid cycle deliberate practice) [141–144] could help maximize the psychological and physiological safety of participants.

Scenarios for simulation-based summative assessment

The structure and use of scenarios in a summative setting are unique and therefore require specific development and skills [83, 88]. SBSA scenarios differ from formative assessment scenarios by the different educational objectives that guide their development. Summative scenarios are designed to assess a skill through observation of performance, while formative scenarios are designed to

learn and progress in mastering this same skill. Although there may be a continuum between the two, they remain distinct. SBSA scenarios must be predictable, programmable, standardizable, and reproducible [25] to ensure fairly assessed performances among participants. This is even more crucial when standardized patients are involved (e.g., OSCE) [119, 145]. In this case, a specific script with expectations and training is needed for the standardized patient. The problem is that currently there are many formative scenarios but few summative scenarios. The rigor and expertise required to develop them is time-consuming and requires expert trainer resources. We believe that a goal should be to homogenize the scenarios, along with preparing the human resources who will implement them (trainers and standardized patients) and their application. We believe one solution would be to develop a methodology for converting formative scenarios into summative ones in order to create a structuring model for summative scenarios. This would reinvest the time and expertise already used for developing = formative scenarios.

Debriefing for simulation-based summative assessment

The place of debriefing in SBSA is currently undefined and raises important questions that need exploration [77, 90, 146–148]. Debriefing for formative assessment promotes knowledge retention and helps to anchor good behaviors while correcting less ideal ones [149–151]. In general, taking an exam promotes learning of the topic [139, 152]. Formative assessment without a debriefing has been shown to be detrimental, so it is reasonable to assume that the same is true in summative assessment [91]. The ideal modalities for debriefing in SBSA are currently unknown [77, 90, 146–148]. Integrating debriefing into SBSA raises a number of organizational, pedagogical, cognitive, and ethical issues that need to be clarified. From an organizational perspective, we consider that debriefing is time and human resource-consuming. The extent of the organizational impact varies according to whether the feedback is automatized, standardized, personalized, and collective or individual. From an educational perspective, debriefing ensures pedagogical continuity and continued learning. We believe this notion is nuanced, depending on whether the debriefing is integrated into the summative assessment or instead follows the assessment while focusing on formative assessment elements. We believe that if the debriefing is part of the SBSA, it is no longer a “teaching moment.” This must be factored into the instructional strategy. How should the trainer prioritize debriefing points between those established in advance for the summative assessment and those that would emerge from any individuals’ performance? From a cognitive perspective, whether the

debriefing is integrated into the summative assessment may alter the interactions between the trainer and the participants. We believe that if the debriefing is integrated into the SBSA, the participant will sometimes be faced with the cognitive dilemma of whether to express his/her “true” opinions or instead attempt to provide the expected answers. The trainer then becomes uncertain of what he/she is actually assessing. Finally, from an ethical perspective, in the case of a mediocre or substandard clinical performance, there is a question of how the trainer resolves discrepancies between observed behavior and what the participant reveals during the debriefing. What weight should be given to the simulation and to the debriefing for the final rating? We believe there is probably no single solution to how and when the debriefing is conducted during a summative assessment but rather promote the idea of adapting debriefing approaches (e.g., group or individualized debriefing) to various conditions (e.g., success or failure in the summative assessment). These questions need to be explored to provide answers as to how debriefing should be ideally conducted in SBSA. We believe a balance must be found that is ethically and pedagogically satisfactory, does not induce a cognitive dilemma for the trainer, and is practically manageable.

Trainers for simulation-based summative assessment

The skills and training of trainers required for SBSA are crucial and must be defined [136, 153]. We consider that skills and training for SBSA closely mirror skills and training needed for formative assessment in simulation. This continuity is part of the pedagogical alignment. These different steps have common characteristics (e.g., rules in simulation, scenario flow) and specific ones (e.g., using assessment tools, validating competence). To ensure pedagogical continuity, the trainers who supervise these courses must be trained in and be masterful in simulation, adhering to pedagogical theories. We believe training for SBSA represents new skills and a potentially greater cognitive load for the trainers. It is necessary to provide solutions to both of these issues. For the new skills of trainers, we consider it necessary to adapt or complete the training of trainers by integrating knowledge and skills needed for properly conducting SBSA: good assessment practices, assessment bias mitigation, rater calibration, mastery of assessment tools, etc. [154]. To optimize the cognitive load induced by the tasks and challenges of SBSA, we suggest that it could be helpful to divide the tasks between the different trainers’ roles. We believe that conducting a SBSA therefore requires three types of trainers whose training is adapted to their specific role. First, three are the trainer-designers who are responsible for designing the assessment situation,

selecting the assessment tool(s), training the trainer-rater(s), and supervising the SBSA sessions. Second, there should be the trainer-operators responsible for running the simulation conditions that support the assessment. Third, there are the trainer-raters who conduct the assessment using the assessment tool(s) selected by the trainer-designer(s) for which these trainer-raters have been specifically trained. The high-stake nature of SBSA requires a high level of rigor and professionalism from the three levels of trainers, which implies they have a working definition of the skills and the necessary training to be up to the task.

Implementing simulation-based summative assessment in healthcare

Implementing SBSA is delicate, requires rigor, respect for each step, and must be evidence-based. While OSCEs are simulation-based, simulation is not limited to OSCEs. Summative assessment with OSCEs has been used and studied for many years [12, 13]. This literature is therefore a valuable source for learning lessons about summative assessment applied to simulation as a whole [22, 85, 155]. Knowledge from OSCE summative assessment needs to be supplemented so that simulation can perform summative assessment according to good evidence-based practices. Given the high stakes of SBSA, we believe it necessary to rigorously and methodically adapt what is already validated during implementation (e.g., scenarios, tools) and to proceed with caution for what has not yet been validated. As described above, many steps and prerequisites are necessary for optimal implementation, including (but not limited to) identifying objectives; identifying and validating assessment tools; preparing simulations scenarios, trainers, and raters; and planning a global strategy beginning from integrating the summative assessment in the curriculum to the managing the consequences of this assessment. SBSA must be conducted within a strict framework for its own sake and that of the people involved. Poor implementation would be detrimental to all participants, trainers, and the practice SBSA. This risk is greater for recertification than for certification [156], while initial training is able to accommodate SBSA easily because it is familiar (e.g., trainees engage in OSCEs at some point in their education), including SBSA in recertifying practicing professionals is not as obvious and may be context-dependent [157]. We understand that the consequences of failed recertification are potentially more impactful, both psychologically and for professional practice. We believe that solutions must be developed, tested, and validated that both fill gaps and preserve professionals and patients. Implementing SBSA therefore must be progressive, rigorous, and evidence-based to be accepted and successful [158].

Limitations

This work has some limitations that should be emphasized. First, this work covers only a limited number of issues related to SBSA. The entire topic is possibly not covered and we may not have explored other questions of interest. Nevertheless, the NGT methodology allowed this work to focus on those issues that were most relevant and challenging to the panel. Second, the literature review method (state-of-the-art literature reviews expanded with a snowball technique) does not guarantee exhaustiveness, and publications on the topic may have escaped the screening phase. However, it is likely that we have identified key articles focused on the questions explored. Potentially unidentified articles would therefore either not be important to the topic or would address questions not selected by the NGT. Third, this work was done by a French-speaking group, and a Francophone-specific approach to simulation, although not described to our knowledge, cannot be ruled out. This risk is reduced by the fact that the work is based on international literature from different specialties in different countries and that the panelists and reviewers were from different countries. Fourth, the analysis and discussion of the consequences of SBSA were focused on psychological consequences. This does not cover the full range of consequences including the impact on subsequent curricula or career pathways. Data in the literature exist on the subject and probably deserve a specific body of work. Despite these limitations, however, we believe this work is valuable because it raises questions and offers some leads toward solutions.

Conclusions

The use of SBSA is very promising with a growing demand for its application. Indeed, SBSA is a logical extension of simulation-based formative assessment and competency-based medical education development. It is probably wise to anticipate and plan for approaches to SBSA, as many important moving parts, questions, and consequences are emerging. Clearly identifying these elements and their interactions will aid in developing reliable, accurate, and reproducible models. All this requires a meticulous and rigorous approach to preparation commensurate with the challenges of certifying or recertifying the skills of healthcare professionals. We have explored the current knowledge on SBSA and have now shared an initial mapping of the topic. Among the seven topics investigated, we have identified (i) areas with robust evidence (what can be assessed with simulation?); (ii) areas with limited evidence that can be assisted by expert opinion and research (assessment tools, scenarios, and implementation); and (iii) areas with weak or emerging evidence requiring guidance by expert opinion and research (consequences,

debriefing, and trainers) (Fig. 1). We modestly hope that this work can help reflection on SBSA for future investigations and can drive guideline development for SBSA.

Abbreviations

GDPR: General data protection regulation; NGT: Nominal group technique; OSCE: Objective structured clinical examination; SBSA: Simulation-based summative assessment.

Acknowledgements

The authors thank SoFraSimS Assessment with simulation group members: Anne Bellot, Isabelle Crubl e, Guillaume Philippot, Thierry Vanderlinden, S ebastien Batrancourt, Claire Boithias-Guerot, Jean Br eaud, Philine de Vries, Louis Sibert, Thierry S echeresse, Virginie Boulant, Louis Delamarre, Laurent Grillet, Marianne Jund, Christophe Mathurin, Jacques Berthod, Blaise Debien, and Olivier Gacia who have contributed to this work. The authors thank the external experts committee members: Guillaume Der Sahakian, Sylvain Boet, Denis Oriot and Jean-Michel Chabot; and the SoFraSimS executive Committee for their review and feedback.

Author's contributions

CB helped with the study conception and design, data contribution, data analysis, data interpretation, writing, visualization, review, and editing. FL helped with the study conception and design, data contribution, data analysis, data interpretation, writing, review, and editing. RDM, JWR, and DB helped with the study writing, and review and editing. JWR and DB helped with the data interpretation, writing, and review and editing. LM, FJL, EG, ALP, OB, and ALS helped with the data contribution, data analysis, data interpretation, and review. The authors read and approved the final manuscript.

Funding

This work has been supported by the French Speaking Society for Simulation in Healthcare (SoFraSimS).

This work is a part of CB PhD which has been supported by grants from the French Society for Anesthesiology and Intensive Care (SFAR), the Arthur Sachs-Harvard Foundation, the University Hospital of Caen, the North-West University Hospitals Group (G4), and the Charles Nicolle Foundation. Funding bodies did not have any role in the design of the study, collection, analysis, and interpretation of the data and in writing the manuscript.

Availability of data and materials

All data generated or analyzed during this study are included in this published article.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Department of Anesthesiology, Intensive Care and Perioperative Medicine, Caen Normandy University Hospital, 6th Floor, Caen, France. ²Medical School, University of Caen Normandy, Caen, France. ³Center for Medical Simulation, Boston, MA, USA. ⁴Department of Anesthesiology, Intensive Care and Perioperative Medicine, Nimes University Hospital, Nimes, France. ⁵Department of Anesthesia, Critical Care and Pain Medicine, Massachusetts General Hospital, Boston, MA, USA. ⁶Harvard Medical School, Boston, MA, USA. ⁷Department of Anesthesiology, Intensive Care and Perioperative Medicine, Li ege University Hospital, Li ege, Belgique. ⁸Department of Emergency Medicine, Piti e Salp etri ere University Hospital, APHP, Paris, France. ⁹Department of Pediatric Intensive Care, Pellegrin University Hospital, Bordeaux, France. ¹⁰Department of Emergency Medicine, Rouen University Hospital, Rouen, France.

¹¹Department of Anesthesiology, Intensive Care and Perioperative Medicine, Kremlin Bicêtre University Hospital, APHP, Paris, France. ¹²Department of Emergency Medicine, Cochin University Hospital, APHP, Paris, France.

Received: 2 March 2022 Accepted: 30 November 2022

Published online: 28 December 2022

References

- van der Vleuten CPM, Schuwirth LWT. Assessment in the context of problem-based learning. *Adv Health Sci Educ Theory Pract*. 2019;24:903–14.
- Boulet JR. Summative assessment in medicine: the promise of simulation for high-stakes evaluation. *Acad Emerg Med*. 2008;15:1017–24.
- Green M, Tariq R, Green P. Improving patient safety through simulation training in anesthesiology: where are we? *Anesthesiol Res Pract*. 2016;2016:4237523.
- Krage R, Erwteman M. State-of-the-art usage of simulation in anesthesia: skills and teamwork. *Curr Opin Anaesthesiol*. 2015;28:727–34.
- Askew K, Manthey DE, Potisek NM, Hu Y, Goforth J, McDonough K, et al. Practical application of assessment principles in the development of an innovative clinical performance evaluation in the entrustable professional activity era. *Med Sci Educ*. 2020;30:499–504.
- Wass V, Van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. *Lancet*. 2001;357:945–9.
- Boulet JR, Murray D. Review article: assessment in anesthesiology education. *Can J Anaesth*. 2012;59:182–92.
- Bauer D, Lahner F-M, Schmitz FM, Guttormsen S, Huwendiek S. An overview of and approach to selecting appropriate patient representations in teaching and summative assessment in medical education. *Swiss Med Wkly*. 2020;150: w20382.
- Park CS. Simulation and quality improvement in anesthesiology. *Anesthesiol Clin*. 2011;29:13–28.
- Higham H, Baxendale B. To err is human: use of simulation to enhance training and patient safety in anaesthesia. *British Journal of Anaesthesia* [Internet]. 2017 [cited 2021 Sep 16];119:i106–14. Available from: <https://www.sciencedirect.com/science/article/pii/S0007091217541215>.
- Mann S, Truelove AH, Beesley T, Howden S, Egan R. Resident perceptions of competency-based medical education. *Can Med Educ J*. 2020;11:e31–43.
- Khan KZ3, Ramachandran S, Gaunt K, Pushkar P. The objective structured clinical examination (OSCE): AMEE Guide No. 81. Part I: an historical and theoretical perspective. *Med Teach*. 2013;35(9):e1437–1446.
- Daniels VJ, Pugh D. Twelve tips for developing an OSCE that measures what you want. *Med Teach*. 2018;40:1208–13.
- Humphrey-Murto S, Varpio L, Gonsalves C, Wood TJ. Using consensus group methods such as Delphi and Nominal Group in medical education research. *Med Teach*. 2017;39:14–9.
- Haute Autorité de Santé. Recommandations par consensus formalisé (RCF) [Internet]. Haute Autorité de Santé. 2011 [cited 2020 Oct 29]. Available from: https://www.has-sante.fr/jcms/c_272505/fr/recom-mandations-par-consensus-formalise-rcf.
- Humphrey-Murto S, Varpio L, Wood TJ, Gonsalves C, Ufholz L-A, Mascioli K, et al. The use of the delphi and other consensus group methods in medical education research: a review. *Academic Medicine* [Internet]. 2017 [cited 2021 Jul 20];92:1491–8. Available from: https://journals.lww.com/academicmedicine/Fulltext/2017/10000/The_Use_of_the_Delphi_and_Other_Consensus_Group.38.aspx.
- Booth A, Sutton A, Papaioannou D. Systematic approaches to a successful literature review [Internet]. Second edition. Los Angeles: Sage; 2016. Available from: https://uk.sagepub.com/sites/default/files/upm-assets/78595_book_item_78595.pdf.
- Morgan DL. Snowball Sampling. In: Given LM, editor. *The Sage encyclopedia of qualitative research methods* [Internet]. Los Angeles, Calif: Sage Publications; 2008. p. 815–6. Available from: <http://www.yanchukvladimir.com/docs/Library/Sage%20Encyclopedia%20of%20Qualitative%20Research%20Methods-%202008.pdf>.
- ten Cate O, Scheele F. Competency-based postgraduate training: can we bridge the gap between theory and clinical practice? *Acad Med*. 2007;82:542–7.
- Miller GE. The assessment of clinical skills/competence/performance. *Acad Med*. 1990;65:S63–67.
- Epstein RM. Assessment in medical education. *N Engl J Med*. 2007;356:387–96.
- Boulet JR, Murray DJ. Simulation-based assessment in anesthesiology: requirements for practical implementation. *Anesthesiology*. 2010;112:1041–52.
- Bédard D, Béchard JP. L'innovation pédagogique dans le supérieur: un vaste chantier. *Innovover dans l'enseignement supérieur*. Paris: Presses Universitaires de France; 2009. p. 29–43.
- Biggs J. Enhancing teaching through constructive alignment. *High Educ* [Internet]. 1996 [cited 2020 Oct 25];32:347–64. Available from: <https://doi.org/10.1007/BF00138871>.
- Wong AK. Full scale computer simulators in anesthesia training and evaluation. *Can J Anaesth*. 2004;51:455–64.
- Messick S. Evidence and ethics in the evaluation of tests. *Educational Researcher* [Internet]. 1981 [cited 2020 Mar 19];10:9–20. Available from: <http://journals.sagepub.com/doi/https://doi.org/10.3102/0013189X010009009>.
- Bould MD, Crabtree NA, Naik VN. Assessment of procedural skills in anaesthesia. *Br J Anaesth*. 2009;103:472–83.
- Schuwirth LWT, van der Vleuten CPM. Programmatic assessment and Kane's validity perspective. *Med Educ*. 2012;46:38–48.
- Brailovsky C, Charlin B, Beausoleil S, Coté S, Van der Vleuten C. Measurement of clinical reflective capacity early in training as a predictor of clinical reasoning performance at the end of residency: an experimental study on the script concordance test. *Med Educ*. 2001;35:430–6.
- van der Vleuten CPM, Schuwirth LWT. Assessing professional competence: from methods to programmes. *Med Educ*. 2005;39:309–17.
- Gordon M, Farnan J, Grafton-Clarke C, Ahmed R, Gurbutt D, McLachlan J, et al. Non-technical skills assessments in undergraduate medical education: a focused BEME systematic review: BEME Guide No. 54. *Med Teach*. 2019;41(7):732–45.
- Jouquan J. L'évaluation des apprentissages des étudiants en formation médicale initiale. *Pédagogie Médicale* [Internet]. 2002 [cited 2020 Feb 2];3:38–52. Available from: <http://www.pedagogie-medicale.org/https://doi.org/10.1051/pmed:2002006>.
- Gale TCE, Roberts MJ, Sice PJ, Langton JA, Patterson FC, Carr AS, et al. Predictive validity of a selection centre testing non-technical skills for recruitment to training in anaesthesia. *Br J Anaesth*. 2010;105:603–9.
- Gallagher CJ, Tan JM. The current status of simulation in the maintenance of certification in anaesthesia. *Int Anesthesiol Clin*. 2010;48:83–99.
- S DeMaria Jr ST, Samuelson AD, Schwartz AJ, Sim AI, Levine S. Simulation-based assessment and retraining for the anesthesiologist seeking reentry to clinical practice: a case series. *Anesthesiology* [Internet]. 2013 [cited 2021 Sep 6];119:206–17. Available from: <https://doi.org/10.1097/ALN.0b013e31829761c8>.
- Amin Z, Boulet JR, Cook DA, Ellaway R, Fahal A, Kneebone R, et al. Technology-enabled assessment of health professions education: consensus statement and recommendations from the Ottawa 2010 conference. *Medical Teacher* [Internet]. 2011 [cited 2021 Jul 7];33:364–9. Available from: <http://www.tandfonline.com/doi/full/https://doi.org/10.3109/0142159X.2011.565832>.
- Scallon G. L'évaluation des apprentissages dans une approche par compétences. *Bruxelles: De Boeck Université-Bruxelles*; 2007.
- Downing SM. Validity: on meaningful interpretation of assessment data. *Med Educ*. 2003;37:830–7.
- Cook DA, Hatala R. Validation of educational assessments: a primer for simulation and beyond. *Adv Simul* [Internet]. 2016 [cited 2021 Aug 24];1:31. Available from: <http://advancesinsimulation.biomedcentral.com/articles/https://doi.org/10.1186/s41077-016-0033-y>.
- Kane MT. Validating the interpretations and uses of test scores. *Journal of Educational Measurement* [Internet]. 2013 [cited 2020 Sep 9];50:1–73. Available from: <https://onlinelibrary.wiley.com/doi/abs/https://doi.org/10.1111/jedm.12000>.
- Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: a practical guide to Kane's framework. *Med Educ*. 2015;49:560–75.

42. DA Cook B Zendejas SJ Hamstra R Hatala R Brydges What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. *Adv in Health Sci Educ* [Internet]. 2014 [cited 2020 Feb 2];19:233–50 Available from: <https://doi.org/10.1007/s10459-013-9458-4>.
43. Cook DA, Lineberry M. Consequences validity evidence: evaluating the impact of educational assessments. *Acad Med* [Internet]. 2016 [cited 2020 Oct 24];91:785–95. Available from: <http://journals.lww.com/00001888-201606000-00018>.
44. Tavakol M, Dennick R. Post-examination analysis of objective tests. *Med Teach*. 2011;33:447–58.
45. Messick S. The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher* [Internet]. 1994 [cited 2021 Feb 15];23:13–23. Available from: <http://journals.sagepub.com/doi/https://doi.org/10.3102/0013189X023002013>.
46. Validity MS. Education measurement. 3rd ed. New York: R. L. Linn; 1989. p. 13–103.
47. Oriot D, Darrieux E, Boureau-Voultoury A, Ragot S, Scépi M. Validation of a performance assessment scale for simulated intraosseous access. *Simul Healthc*. 2012;7:171–5.
48. Guise J-M, Deering SH, Kanki BG, Osterweil P, Li H, Mori M, et al. Validation of a tool to measure and promote clinical teamwork. *Simul Healthc*. 2008;3:217–23.
49. Sousa VD, Rojjanasriat W. Translation, adaptation and validation of instruments or scales for use in cross-cultural health care research: a clear and user-friendly guideline: Validation of instruments or scales. *Journal of Evaluation in Clinical Practice*. 2011 [cited 2022 Jul 22];17:268–74. Available from: <https://onlinelibrary.wiley.com/doi/https://doi.org/10.1111/j.1365-2753.2010.01434.x>.
50. Stoyanova-Piroth G, Milanov I, Stambolieva K. Translation, adaptation and validation of the Bulgarian version of the King's Parkinson's Disease Pain Scale. *BMC Neurol* [Internet]. 2021 [cited 2022 Jul 22];21:357. Available from: <https://bmcneurol.biomedcentral.com/articles/https://doi.org/10.1186/s12883-021-02392-5>.
51. Behari M, Srivastava A, Achanti R, Nandal N, Dutta R. Pain assessment in Indian Parkinson's disease patients using King's Parkinson's disease pain scale. *Ann Indian Acad Neurol* [Internet]. 2020 [cited 2022 Jul 22];0:0. Available from: <http://www.annalsofian.org/preprintarticle.asp?id=300170?type=0>.
52. Guillemin F, Bombardier C, Beaton D. Cross-cultural adaptation of health-related quality of life measures: literature review and proposed guidelines. *Journal of Clinical Epidemiology* [Internet]. 1993 [cited 2022 Jul 22];46:1417–32. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S089543569390142N>.
53. Franc JM, Verde M, Gallardo AR, Carenzo L, Ingrassia PL. An Italian version of the Ottawa crisis resource management global rating scale: a reliable and valid tool for assessment of simulation performance. *Intern Emerg Med*. 2017;12:651–6.
54. Gosselin É, Marceau M, Vincelette C, Daneau C-O, Lavoie S, Ledoux I. French translation and validation of the Mayo High Performance Teamwork Scale for nursing students in a high-fidelity simulation context. *Clinical Simulation in Nursing* [Internet]. 2019 [cited 2022 Jul 25];30:25–33. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S1876139918301890>.
55. Sánchez-Marco M, Escribano S, Cabañero-Martínez M-J, Espinosa-Ramírez S, José Muñoz-Reig M, Juliá-Sanchis R. Cross-cultural adaptation and validation of two crisis resource management scales. *International Emergency Nursing* [Internet]. 2021 [cited 2022 Jul 25];57:101016. Available from: <https://www.sciencedirect.com/science/article/pii/S1755599X21000549>.
56. Schuwirth LWT, Van der Vleuten CPM. Programmatic assessment: from assessment of learning to assessment for learning. *Medical Teacher* [Internet]. 2011 [cited 2021 Sep 6];33:478–85. Available from: <http://www.tandfonline.com/doi/full/https://doi.org/10.3109/0142159X.2011.565828>.
57. Maignan M, Koch F-X, Chaix J, Phellouzat P, Binauld G, Collomb Muret R, et al. Team Emergency Assessment Measure (TEAM) for the assessment of non-technical skills during resuscitation: validation of the French version. *Resuscitation* [Internet]. 2016 [cited 2019 Mar 12];101:115–20. Available from: <http://www.sciencedirect.com/science/article/pii/S0300957215008989>.
58. Pires S, Monteiro S, Pereira A, Chaló D, Melo E, Rodrigues A. Non-technical skills assessment for prelicensure nursing students: an integrative review. *Nurse Educ Today*. 2017;58:19–24.
59. Khan R, Payne MWC, Chahine S. Peer assessment in the objective structured clinical examination: a scoping review. *Med Teach*. 2017;39:745–56.
60. Hegg RM, Ivan KF, Tone J, Morten A. Comparison of peer assessment and faculty assessment in an interprofessional simulation-based team training program. *Nurse Educ Pract*. 2019;42: 102666.
61. Scavone BM, Sproviero MT, McCarthy RJ, Wong CA, Sullivan JT, Siddall VJ, et al. Development of an objective scoring system for measurement of resident performance on the human patient simulator. *Anesthesiology*. 2006;105:260–6.
62. Oriot D, Bridier A, Ghazali DA. Development and assessment of an evaluation tool for team clinical performance: the Team Average Performance Assessment Scale (TAPAS). *Health Care : Current Reviews* [Internet]. 2016 [cited 2018 Jan 17];4:1–7. Available from: <https://www.omicsonline.org/open-access/development-and-assessment-of-an-evaluation-tool-for-team-clinical-performance-the-team-average-performance-assessment-scale-tapas-2375-4273-1000164.php?aid=72394>.
63. Flin R, Patey R, Glavin R, Maran N. Anaesthetists' non-technical skills. *Br J Anaesth*. 2010;105:38–44.
64. Mishra A, Catchpole K, McCulloch P. The Oxford NOTECHS System: reliability and validity of a tool for measuring teamwork behaviour in the operating theatre. *Quality and Safety in Health Care* [Internet]. 2009 [cited 2021 Jul 6];18:104–8. Available from: <https://qualitysafety.bmj.com/lookup/doi/https://doi.org/10.1136/qshc.2007.024760>.
65. Cooper S, Cant R, Porter J, Sellick K, Somers G, Kinsman L, et al. Rating medical emergency teamwork performance: development of the Team Emergency Assessment Measure (TEAM). *Resuscitation*. 2010;81:446–52.
66. Adler MD, Trainor JL, Siddall VJ, McGaghie WC. Development and evaluation of high-fidelity simulation case scenarios for pediatric resident education. *Ambul Pediatr*. 2007;7:182–6.
67. Brydges R, Hatala R, Zendejas B, Erwin PJ, Cook DA. Linking simulation-based educational assessments and patient-related outcomes: a systematic review and meta-analysis. *Acad Med*. 2015;90:246–56.
68. Cazzell M, Howe C. Using Objective Structured Clinical Evaluation for Simulation Evaluation: Checklist Considerations for Interrater Reliability. *Clinical Simulation In Nursing* [Internet]. 2012;8(6):e219–25. [cited 2019 Dec 14] Available from: [https://www.nursingsimulation.org/article/S1876-1399\(11\)00249-0/abstract](https://www.nursingsimulation.org/article/S1876-1399(11)00249-0/abstract).
69. Maignan M, Viglino D, Collomb Muret R, Vejux N, Wiel E, Jacquin L, et al. Intensity of care delivered by prehospital emergency medical service physicians to patients with deliberate self-poisoning: results from a 2-day cross-sectional study in France. *Intern Emerg Med*. 2019;14:981–8.
70. Alcaraz-Mateos E, Jiang X "Sara," Mohammed AAR, Turic I, Hernández-Sabater L, Caballero-Alemán F, et al. A novel simulator model and standardized assessment tools for fine needle aspiration cytology training. *Diagn Cytopathol* [Internet]. 2019 [cited 2020 Feb 3];47:297–301. Available from: <http://doi.wiley.com/https://doi.org/10.1002/dc.24105>.
71. I Ghaderi M Vaillancourt G Sroka PA Kaneva MC Vassiliou I Choy Evaluation of surgical performance during laparoscopic incisional hernia repair: a multicenter study. *Surg Endosc* [Internet]. et al 2011 [cited 2020 Feb 2];25:2555–63 Available from: <https://doi.org/10.1007/s00464-011-1586-4>.
72. Ijgosse WM, Leijte E, Ganni S, Luursema J-M, Francis NK, Jakimowicz JJ, et al. Competency assessment tool for laparoscopic suturing: development and reliability evaluation. *Surg Endosc*. 2020;34(7):2947–53.
73. Pelaccia T, Tardif J. In: Comment [mieux] former et évaluer les étudiants en médecine et en sciences de la santé? 1ère. Louvain-la-Neuve: De Boeck supérieur; 2016. p. 343–56. (Guides pratiques).
74. Henricksen JW, Altenburg C, Reeder RW. Operationalizing healthcare simulation psychological safety: a descriptive analysis of an intervention. *Simul Healthc*. 2017;12:289–97.
75. Gaba DM. Simulations that are challenging to the psyche of participants: how much should we worry and about what? *Simulation in Healthcare: The Journal of the Society for Simulation in Healthcare* [Internet]. 2013 [cited 2020 Mar 17];8:4–7. Available from: <http://journals.lww.com/01266021-201302000-00002>.

76. Ghazali DA, Breque C, Sosner P, Lesbordes M, Chavagnat J-J, Ragot S, et al. Stress response in the daily lives of simulation repeaters. A randomized controlled trial assessing stress evolution over one year of repetitive immersive simulations. *PLoS One*. 2019;14(7):e0220111.
77. Rudolph JW, Simon R, Raemer DB, Eppich WJ. Debriefing as formative assessment: closing performance gaps in medical education. *Acad Emerg Med*. 2008;15:1010–6.
78. Kang SJ, Min HY. Psychological safety in nursing simulation. *Nurse Educ*. 2019;44:E6–9.
79. Howard SK, Gaba DM, Smith BE, Weinger MB, Herndon C, Keshavacharya S, et al. Simulation study of rested versus sleep-deprived anesthesiologists. *Anesthesiology*. 2003;98(6):1345–55.
80. Neuschwander A, Job A, Younes A, Mignon A, Delgoulet C, Cabon P, et al. Impact of sleep deprivation on anaesthesia residents' non-technical skills: a pilot simulation-based prospective randomized trial. *Br J Anaesth*. 2017;119:125–31.
81. Eastridge BJ, Hamilton EC, O'Keefe GE, Rege RV, Valentine RJ, Jones DJ, et al. Effect of sleep deprivation on the performance of simulated laparoscopic surgical skill. *Am J Surg*. 2003;186:169–74.
82. Boulet JR, Murray D, Kras J, Woodhouse J, McAllister J, Ziv A. Reliability and validity of a simulation-based acute care skills assessment for medical students and residents. *Anesthesiology*. 2003;99:1270–80.
83. Levine AI, Flynn BC, Bryson EO, Demaria S. Simulation-based Maintenance of Certification in Anesthesiology (MOCA) course optimization: use of multi-modality educational activities. *J Clin Anesth*. 2012;24:68–74.
84. Boulet JR, Murray D, Kras J, Woodhouse J. Setting performance standards for mannequin-based acute-care scenarios: an examinee-centered approach. *Simul Healthc*. 2008;3:72–81.
85. Furman GE, Smee S, Wilson C. Quality assurance best practices for simulation-based examinations. *Simul Healthc*. 2010;5:226–31.
86. Kane MT. The assessment of professional competence. *Eval Health Prof [Internet]*. 1992 [cited 2022 Jul 22];15:163–82. Available from: <http://journals.sagepub.com/doi/https://doi.org/10.1177/016327879201500203>.
87. Blum RH, Boulet JR, Cooper JB, Muret-Wagstaff SL. Harvard Assessment of Anesthesia Resident Performance Research Group. Simulation-based assessment to identify critical gaps in safe anesthesia resident performance. *Anesthesiol*. 2014;120(1):129–41.
88. Rizzolo MA, Kardong-Edgren S, Oermann MH, Jeffries PR. The national league for nursing project to explore the use of simulation for high-stakes assessment: process, outcomes, and recommendations: nursing education perspectives. 2015 [cited 2020 Feb 3];36:299–303. Available from: <http://Insights.ovid.com/crossref?an=00024776-201509000-00006>.
89. Mudumbai SC, Gaba DM, Boulet JR, Howard SK, Davies MF. External validation of simulation-based assessments with other performance measures of third-year anesthesiology residents. *Simul Healthc*. 2012;7:73–80.
90. Fanning RM, Gaba DM. The role of debriefing in simulation-based learning. *Simul Healthc*. 2007;2:115–25.
91. Savoldelli GL, Naik VN, Park J, Joo HS, Chow R, Hamstra SJ. Value of debriefing during simulated crisis management versus video-assisted oral feedback. *Anesthesiology*. American Society of Anesthesiologists; 2006 [cited 2020 Oct 19];105:279–85. Available from: <https://pubs.asahq.org/anesthesiology/article/105/2/279/6669/Value-of-Debriefing-during-Simulated-Crisis>.
92. Haute Autorité de Santé. Guide de bonnes pratiques en simulation en santé. 2012 [cited 2020 Feb 2]. Available from: https://www.has-sante.fr/upload/docs/application/pdf/2013-01/guide_bonnes_pratiques_simulation_sante_guide.pdf.
93. INACSL Standards Committee. INACSL Standards of best practice: simulation. *Simulation design*. *Clinical Simulation In Nursing*. 2016 [cited 2020 Feb 2];12:55–12. Available from: [https://www.nursingsimulation.org/article/S1876-1399\(16\)30126-8/abstract](https://www.nursingsimulation.org/article/S1876-1399(16)30126-8/abstract).
94. Norcini J, Anderson B, Bollela V, Burch V, Costa MJ, Duvivier R, et al. Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. *Med Teach*. 2011;33:206–14.
95. Gantt LT. The effect of preparation on anxiety and performance in summative simulations. *Clinical Simulation in Nursing*. 2013 [cited 2020 Feb 2];9:e25–33. Available from: <http://www.sciencedirect.com/science/article/pii/S1876139911001277>.
96. Frey-Vogel AS, Scott-Vernaglia SE, Carter LP, Huang GC. Simulation for milestone assessment: use of a longitudinal curriculum for pediatric residents. *Simul Healthc*. 2016;11:286–92.
97. Durning SJ, Artino A, Boulet J, La Rochelle J, Van der Vleuten C, Arze B, et al. The feasibility, reliability, and validity of a post-encounter form for evaluating clinical reasoning. *Med Teach*. 2012;34:30–7.
98. Stone J. Moving interprofessional learning forward through formal assessment. *Medical Education*. 2010 [cited 2020 Feb 12];44:396–403. Available from: <http://doi.wiley.com/https://doi.org/10.1111/j.1365-2923.2009.03607.x>.
99. Manser T, Dieckmann P, Wehner T, Ralf M. Comparison of anaesthetists' activity patterns in the operating room and during simulation. *Ergonomics*. 2007;50:246–60.
100. Perrenoud P. Évaluation formative et évaluation certificative : postures contradictoires ou complémentaires ? *Formation Professionnelle suisse*. 2001 [cited 2020 Oct 29];4:25–8. Available from: https://www.unige.ch/fapse/SSE/teachers/perrenoud/php_main/php_2001/2001_13.html.
101. Atesok K, Hurwitz S, Anderson DD, Satava R, Thomas GW, Tufescu T, et al. Advancing simulation-based orthopaedic surgical skills training: an analysis of the challenges to implementation. *Adv Orthop*. 2019;2019:1–7.
102. Chiu M, Tarshis J, Antoniou A, Bosma TL, Burjorjee JE, Cowie N, et al. Simulation-based assessment of anesthesiology residents' competence: development and implementation of the Canadian National Anesthesiology Simulation Curriculum (CanNASc). *Can J Anesth/J Can Anesth*. 2016 [cited 2020 Feb 2];63:1357–63. Available from: <https://doi.org/10.1007/s12630-016-0733-8>.
103. TC Everett RJ, McKinnon E, Ng P, Kulkarni BCR, Borges M. Letal Simulation-based assessment in anesthesia: an international multicentre validation study. *Can J Anesth, J Can Anesth*. et al 2019 [cited 2020 Feb 2];66:1440–9 Available from: <https://doi.org/10.1007/s12630-019-01488-4>.
104. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). May 4, 2016. Available from: <http://data.europa.eu/eli/reg/2016/679/2016-05-04/eng>.
105. Commission Nationale de l'Informatique et des Libertés. RGPD : passer à l'action. 2021 [cited 2021 Jul 8]. Available from: <https://www.cnil.fr/fr/rgpd-passer-a-laction>.
106. Ten Cate O, Regehr G. The power of subjectivity in the assessment of medical trainees. *Acad Med*. 2019;94:333–7.
107. Weller JM, Robinson BJ, Jolly B, Watterson LM, Joseph M, Bajenov S, et al. Psychometric characteristics of simulation-based assessment in anaesthesia and accuracy of self-assessed scores. *Anaesthesia*. 2005;60:245–50.
108. Wikander L, Bouchoucha SL. Facilitating peer based learning through summative assessment - an adaptation of the objective structured clinical assessment tool for the blended learning environment. *Nurse Educ Pract*. 2018;28:40–5.
109. Gaugler BB, Rudolph AS. The influence of assessee performance variation on assessors' judgments. *Pers Psychol*. 1992;45:77–98.
110. Feldman M, Lazzara EH, Vanderbilt AA, DiazGranados D. Rater training to support high-stakes simulation-based assessments. *J Contin Educ Health Prof*. 2012 [cited 2019 Dec 14];32:279–86. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3646087/>.
111. Pelgrim E a. M, Kramer AWM, Mokkink HGA, van den Elsen L, Grol RPTM, van der Vleuten CPM. In-training assessment using direct observation of single-patient encounters: a literature review. *Adv Health Sci Educ Theory Pract*. 2011;16(1):131–42.
112. Downing SM, Tekian A, Yudkowsky R. Procedures for establishing defensible absolute passing scores on performance examinations in health professions education. *Teach Learn Med*. 2006;18:50–7.
113. Berkenstadt H, Ziv A, Gafni N, Sidi A. Incorporating simulation-based objective structured clinical examination into the Israeli National Board Examination in Anesthesiology. *Anesth Analg*. 2006;102:853–8.

114. Hedge JW, Kavanagh MJ. Improving the accuracy of performance evaluations: comparison of three methods of performance appraiser training. *J Appl Psychol*. 1988;73:68–73.
115. Harden RM, Stevenson M, Downie WW, Wilson GM. Assessment of clinical competence using objective structured examination. *Br Med J*. 1975;1:447–51.
116. Uzan S. Mission de recertification des médecins - Exercer une médecine de qualité | Vie publique.fr. Ministère des Solidarités et de la Santé - Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation; 2018 Nov. Available from: <https://www.vie-publique.fr/rapport/37741-mission-de-recertification-des-medecins-exercer-mecicine-de-qualit>.
117. Mann KV, MacDonald AC, Norcini JJ. Reliability of objective structured clinical examinations: four years of experience in a surgical clerkship. *Teaching and Learning in Medicine*. 1990 [cited 2021 May 1];2:219–24. Available from: <http://www.tandfonline.com/doi/abs/https://doi.org/10.1080/10401339009539464>.
118. Maintenance Of Certification in Anesthesiology (MOCA) 2.0. [cited 2021 Sep 18]. Available from: <https://theaba.org/about%20moca%202.0.html>.
119. Khan KZ, Gaunt K, Ramachandran S, Pushkar P. The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part II: Organisation & Administration. *Med Teach*. 2013 [cited 2020 Oct 29];35:e1447–63. Available from: <http://www.tandfonline.com/doi/full/https://doi.org/10.3109/0142159X.2013.818635>.
120. Coderre S, Woloschuk W, McLaughlin K. Twelve tips for blueprinting. *Med Teach*. 2009;31:322–4.
121. Murray DJ, Boulet JR. Anesthesiology board certification changes: a real-time example of "assessment drives learning." *Anesthesiology*. 2018;128:704–6.
122. Roberts C, Newble D, Jolly B, Reed M, Hampton K. Assuring the quality of high-stakes undergraduate assessments of clinical competence. *Med Teach*. 2006;28:535–43.
123. Newble D. Techniques for measuring clinical competence: objective structured clinical examinations. *Med Educ*. 2004;38:199–203.
124. Der Sahakian G, Lecomte F, Buléon C, Guevara F, Jaffrelot M, Alinier G. Référentiel sur l'élaboration de scénarios de simulation en immersion clinique. Paris: Société Francophone de Simulation en Santé; 2017 p. 22. Available from: <https://sofrasims.org/wp-content/uploads/2019/10/R%C3%A9f%C3%A9rentiel-Scenari-Simulation-Sofrasims.pdf>.
125. Lewis KL, Bohnert CA, Gammon WL, Hölzer H, Lyman L, Smith C, et al. The Association of Standardized Patient Educators (ASPE) Standards of Best Practice (SOBP). *Adv Simul*. 2017;2:10.
126. Board of Directors of the American Board of Medical Specialties (ABMS). Standards for the ABMS Program for Maintenance of Certification (MOC). American Board of Medical Specialties; 2014 Jan p. 13. Available from: <https://www.abms.org/media/1109/standards-for-the-abms-program-for-moc-final.pdf>.
127. Hodges B, McNaughton N, Regehr G, Tiberius R, Hanson M. The challenge of creating new OSCE measures to capture the characteristics of expertise. *Med Educ*. 2002;36:742–8.
128. Hays RB, Davies HA, Beard JD, Caldon LJM, Farmer EA, Finucane PM, et al. Selecting performance assessment methods for experienced physicians. *Med Educ*. 2002;36:910–7.
129. Ram P, Grol R, Rethans JJ, Schouten B, van der Vleuten C, Kester A. Assessment of general practitioners by video observation of communicative and medical performance in daily practice: issues of validity, reliability and feasibility. *Med Educ*. 1999;33:447–54.
130. Weersink K, Hall AK, Rich J, Szulewski A, Dagnone JD. Simulation versus real-world performance: a direct comparison of emergency medicine resident resuscitation entrustment scoring. *Adv Simul*. 2019 [cited 2020 Feb 12];4:9. Available from: <https://advancesin-simulation.biomedcentral.com/articles/https://doi.org/10.1186/s41077-019-0099-4>.
131. Buljac-Samardzic M, Doekhie KD, van Wijngaarden JDH. Interventions to improve team effectiveness within health care: a systematic review of the past decade. *Hum Resour Health*. 2020;18:2.
132. Eddy K, Jordan Z, Stephenson M. Health professionals' experience of teamwork education in acute hospital settings: a systematic review of qualitative literature. *JBI Database System Rev Implement Rep*. 2016;14:96–137.
133. Leblanc VR. Review article: simulation in anesthesia: state of the science and looking forward. *Can J Anaesth*. 2012;59:193–202.
134. Hanscom R. Medical simulation from an insurer's perspective. *Acad Emerg Med*. 2008;15:984–7.
135. McCarthy J, Cooper JB. Malpractice insurance carrier provides premium incentive for simulation-based training and believes it has made a difference. *Anesth Patient Saf Found Newsl*. 2007 [cited 2021 Sep 17];17. Available from: <https://www.apsf.org/article/malpractice-insurance-carrier-provides-premium-incentive-for-simulation-based-training-and-believes-it-has-made-a-difference/>.
136. Edler AA, Fanning RG, Chen Michael I, Claire R, Almazan D, Struyk B, et al. Patient simulation: a literary synthesis of assessment tools in anesthesiology. *J Educ Eval Health Prof*. 2009 [cited 2021 Sep 17];6:3. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2796725/>.
137. Borgersen NJ, Naur TMH, Sørensen SMD, Bjerrum F, Konge L, Subhi Y, et al. Gathering validity evidence for surgical simulation: a systematic review. *Annals of Surgery*. 2018 [cited 2022 Sep 25];267:1063–8. Available from: <https://journals.lww.com/0000658-201806000-00014>.
138. Rudolph JW, Raemer DB, Simon R. Establishing a safe container for learning in simulation: the role of the presimulation briefing. *Simul Healthc*. 2014;9:339–49.
139. Cilliers FJ, Schuwirth LW, Adendorff HJ, Herman N, van der Vleuten CP. The mechanism of impact of summative assessment on medical students' learning. *Adv Health Sci Educ Theory Pract*. 2010;15:695–715.
140. Hadi MA, Ali M, Haseeb A, Mohamed MMA, Elrggal ME, Cheema E. Impact of test anxiety on pharmacy students' performance in Objective Structured Clinical Examination: a cross-sectional survey. *Int J Pharm Pract*. 2018;26:191–4.
141. Dunn W, Dong Y, Zendejas B, Ruparel R, Farley D. Simulation, mastery learning and healthcare. *Am J Med Sci*. 2017;353:158–65.
142. McGaghie WC. Mastery learning: it is time for medical education to join the 21st century. *Acad Med*. 2015;90:1438–41.
143. Ng C, Primiani N, Orchanian-Cheff A. Rapid cycle deliberate practice in healthcare simulation: a scoping review. *Med Sci Educ*. 2021;31:2105–20.
144. Taras J, Everett T. Rapid cycle deliberate practice in medical education - a systematic review. *Cureus*. 2017;9: e1180.
145. Cleland JA, Abe K, Rethans J-J. The use of simulated patients in medical education: AMEE Guide No 42. *Med Teach*. 2009;31:477–86.
146. Garden AL, Le Fevre DM, Waddington HL, Weller JM. Debriefing after simulation-based non-technical skill training in healthcare: a systematic review of effective practice. *Anaesth Intensive Care*. 2015;43:300–8.
147. Sawyer T, Eppich W, Brett-Fleegler M, Grant V, Cheng A. More than one way to debrief: a critical review of healthcare simulation debriefing methods. *Simul Healthc*. 2016;11:209–17.
148. Rudolph JW, Simon R, Dufresne RL, Raemer DB. There's no such thing as "nonjudgmental" debriefing: a theory and method for debriefing with good judgment. *Simul Healthc*. 2006;1:49–55.
149. Levett-Jones T, Lapkin S. A systematic review of the effectiveness of simulation debriefing in health professional education. *Nurse Educ Today*. 2014;34:e58-63.
150. Palaganas JC, Fey M, Simon R. Structured debriefing in simulation-based education. *AACN Adv Crit Care*. 2016;27:78–85.
151. Rudolph JW, Foldy EG, Robinson T, Kendall S, Taylor SS, Simon R. Helping without harming: the instructor's feedback dilemma in debriefing—a case study. *Simul Healthc*. 2013;8:304–16.
152. Larsen DP, Butler AC, Roediger III HL. Test-enhanced learning in medical education. *Medical Education*. 2008 [cited 2021 Aug 25];42:959–66. Available from: <https://onlinelibrary.wiley.com/doi/https://doi.org/10.1111/j.1365-2923.2008.03124.x>.
153. Koster MA, Soffler M. Navigate the challenges of simulation for assessment: a faculty development workshop. *MedEdPORTAL*. 2021;17:11114.

154. Devitt JH, Kurrek MM, Cohen MM, Fish K, Fish P, Murphy PM, et al. Testing the raters: inter-rater reliability of standardized anaesthesia simulator performance. *Can J Anaesth*. 1997;44:924–8.
155. Kelly MA, Mitchell ML, Henderson A, Jeffrey CA, Groves M, Nulty DD, et al. OSCE best practice guidelines—applicability for nursing simulations. *Adv Simul*. 2016 [cited 2020 Feb 3];1:10. Available from: <http://advancesinsimulation.biomedcentral.com/articles/https://doi.org/10.1186/s41077-016-0014-1>.
156. Weinger MB, Banerjee A, Burden AR, Mclvor WR, Boulet J, Cooper JB, et al. Simulation-based assessment of the management of critical events by board-certified anesthesiologists. *Anesthesiology*. 2017;127:475–89.
157. Sinz E, Banerjee A, Steadman R, Shotwell MS, Slagle J, Mclvor WR, et al. Reliability of simulation-based assessment for practicing physicians: performance is context-specific. *BMC Med Educ*. 2021;21:207.
158. Ryall T, Judd BK, Gordon CJ. Simulation-based assessments in health professional education: a systematic review. *J Multidiscip Healthc*. 2016;9:69–82.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

