Medical Education Online

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/zmeo20

Development and validation of the QASSH scale: a tool for assessing the quality of simulation scenarios in healthcare education

Etienne Rivière, Guillaume Der Sahakian, Marie-Laurence Tremblay, Gilles Chiniara & for the QASSH working group

To cite this article: Etienne Rivière, Guillaume Der Sahakian, Marie-Laurence Tremblay, Gilles Chiniara & for the QASSH working group (2025) Development and validation of the QASSH scale: a tool for assessing the quality of simulation scenarios in healthcare education, Medical Education Online, 30:1, 2486971, DOI: 10.1080/10872981.2025.2486971

To link to this article: https://doi.org/10.1080/10872981.2025.2486971

Published online: 08 Apr 2025.

MEDICAL EDUCATION ONLINE
2025, VOL. 30, 2486971
https://doi.org/10.1080/10872981.2025.2486971

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

OPEN ACCESS Check for updates

Development and validation of the QASSH scale: a tool for assessing the quality of simulation scenarios in healthcare education

Etienne Rivière [a,b], Guillaume Der Sahakian[c], Marie-Laurence Tremblay[d], Gilles Chiniara[e] and for the QASSH working group[*]

[a]Department of Internal Medicine, Haut-Leveque Hospital, University Hospital Centre of Bordeaux, Pessac, France; [b]Medical Faculty & SimBA-S Simulation Center, Bordeaux University, Bordeaux, France; [c]Department of Emergency Medicine, Centre Hospitalier d'Orange, Orange, France; [d]Faculty of Pharmacy, Laval University, Quebec, Canada; [e]Department of Anesthesiology and Intensive Care, Laval University, Quebec, Canada

ABSTRACT

Simulation-based education has become essential for pre- and post-graduate training of healthcare professionals. However, there is no tool to help simulation educators or program managers in assessing the educational quality of simulation scenario scripts for team-based immersive simulation (IS), simulated participants (SP) and procedural simulation (PS). To that end, we developed the Quality Assessment of Simulation Scenario in Healthcare (QASSH) tool. This study aims at providing validity evidence for QASSH. We set up a francophone group of experts within the French-speaking Society for Simulation in Healthcare (SoFraSimS) network and designed this scale based on recently published best practices and our long experience in conceiving simulation scenarios. We tested it by submitting three scenarios of high, borderline and low quality for assessment to a group of experts, a third of which were involved in its development. Analysis of reliability and validity of the QASSH was done using the Standards for educational and psychological testing. Generalizability theory (GT) was used to assess the internal structure and reliability of the tool. The absolute reliability coefficients (G coefficients) calculated through GT were: 0.97 (IS), 0.96 (SP), and 0.98 (PS). G-facet analyses showed that no removal of a single item of QASSH significantly increased the G coefficient above 0.01 for any of the three variants. Cronbach's alpha coefficients were 0.94 (IS), 0.94 (SP) and 0.97 (PS). Estimating the impact of the number of raters on reliability (i.e. D-studies) showed that two raters were enough to achieve a G coefficient above 0.85. The G study shows a high generalizability coefficient (≥0.90), which demonstrates high reliability. The response process evidence for validity provides evidence that no error was associated with using the instrument and its reliability was high with two raters. The QASSH is a tool to assess the quality of healthcare simulation scenarios and will be helpful to instructors wishing to build effective IS, PS and SPs scenarios.

ARTICLE HISTORY
Received 5 December 2024
Revised 10 March 2025
Accepted 26 March 2025

KEYWORDS
Simulation training; immersive simulation; procedural simulation; simulated participants; educational measurement; formative feedback; generalizability theory; QASSH

Introduction

Simulation-based education (SBE) has become an essential educational strategy to train healthcare professionals and ultimately bring clinical benefits to patient care [1,2]. It is being increasingly used worldwide to improve patient care and safety and to overcome several challenges within healthcare [3].

As an educational method, SBE includes several modalities, mainly procedural simulation, immersive simulation, and simulated participants (simulated or standardized patients; SP) [4]. Procedural simulation aims at training healthcare professionals in achieving mastery in psychomotor competencies [5]. Immersive simulation reproduces an authentic clinical experience for learners, augmented with expert debriefing congruent to learners' needs; it is mostly used to develop competencies in crisis resource management, manage human factors in healthcare, and foster better care relationships and their inherent relational competencies (communication, empathy, professionalism) [4]. Simulated participants rely heavily on actors and confederates playing roles during simulation [6].

CONTACT Etienne Rivière etienne.riviere@u-bordeaux.fr Department of Internal Medicine, Haut-Leveque Hospital, University Hospital Centre of Bordeaux, CHU de bordeaux, 1 avenue Magelan, Pessac Cedex, 33604, France

[*]The QASSH working group: Guillaume Alinier, PhD, Christian Balmer, MD, BertrandBech, Dan Benhamou, MD, PhD, Anne Bellot, MD, Antonia Blanié, MD, PhD, SylvainBoloré, RN, CEN, PhD, Hamdi Boubaker, MD, Clément Buléon, MD, PhD, Gilles Chiniara,MD, MHPE, Pierre Clerget, MD, Guillaume Der Sahakian, MD, Nadège Dubois, FranciscoGuevara, Erwan Guillouet, PhD's, Jean-Claude Granry, MD, PhD, Morgan Jaffrelot, MD,MA(Ed), François Lecomte, MD, Fernande Lois, MD, PhD, Christophe Mathurin, MD,Mohammed Mouhaoui, MD, Julien Naud, MD, Ollivier Ortolé, MD, Méryl Paquay, RN,MSc, PhD, Mélanie Pelletier, BA, MSc, Justine Piazza, MD, Marie Pittaco, MD, PatrickPlaisance, MD, PhD, Eliane Raymond-Dufresne, MD, Etienne Rivière, MD, PhD, AurélieSan-Miguel, MD, Samuel-Lessard Tremblay, MD, Marie-Hélène Tremblay, MD, Marie-Laurence Tremblay, MSc, MHPE, BCPS, Maxime de Varennes, MD.

Simulation scenarios describe the events facing the learners during simulation and are central to any simulation learning experience. Salas and collaborators have argued that 'scenario is the curriculum' (p. 364) [7]. According to best practices in simulation, like those we recently published regarding scenario design [8], any simulation scenario in healthcare must target specific learning objectives [9] and be conceived and scripted accordingly [10]. Indeed, scenario design is a basic building block of train-the-trainer courses, and various articles have suggested necessary steps in designing appropriate simulation scenarios [10], especially for immersive simulation [11] and SPs [6]. The major goal of such frameworks is to achieve learning outcomes through a tight control over the simulated experience [7].

While effective, simulation-based education is costly in material (manikins, audiovisual equipment, accessories, infrastructure) and human resources (high instructor-to-learner ratio, technicians, embedded personnel, administrative support) [12], and tools to optimize its use are needed. It is particularly important to create tools that can assess the quality of a scenario script before resources are expended to implement the scenario. Indeed, badly crafted scenarios have the potential to hamper knowledge transfer to real-life situations or to cause negative transfer [13]. Failures in design could prevent learning outcomes from being achieved, reduce reproducibility of the learning experience, and negatively affect debriefing. Alternatively, high quality scenarios will engage learners, provide them with an opportunity to learn linked to specific learning objectives and help them develop accurate mental models [7]. Thus, we designed a new tool to assess the quality of scenario scripts for procedural simulation, immersive simulation and SPs, called QASSH for Quality Assessment of Scenarios in Simulation in Healthcare. The tool is mainly aimed at users with experience in the process of instructional design for SBE, to help in reviewing scenarios that are submitted for a new simulation activity or that are modified in a process of quality improvement. It can also be used by novices to guide them throughout the process of scenario design itself.

This article describes the development process for QASSH, and the validity evidence we collected to support its use, using the framework provided by the *Standards for educational and psychological testing*. [14] We also report the psychometric data we measured to document its reliability, based on Generalizability Theory (GT) [15,16]. GT is particularly appropriate because it allows estimating multiple sources of errors in a single analysis and estimating reliability for different conditions of use, e.g., varying the number of raters, providing important data for generalizability [17].

Methods

Design of the simulation scenarios

A subset of the study research team with a lengthy experience in simulation scenario design (the authors GC, ER and GDS) created a set of three scenarios for each of the simulation modalities: immersive simulation, procedural simulation and SP. Within each set, one scenario was designed to a clear subpar standard ('low-quality scenario'), one was designed to a high standard ('high-quality scenario'), and one was designed as a threshold scenario with the experts overall unsure on whether it would be considered of good or bad quality ('borderline scenario'). The scenarios were designed by targeting scores in specific intervals of 0–20%, 40–60% and 80–100% for low, borderline and high-quality scenarios respectively. For example, we designed these declinations with 4/35 correct items, 17/35 correct items and 32/35 correct items for the low, borderline and high-quality scenarios for simulated participants. The standards used to design the scenarios were based on a previously published method for scenario design [18] and on the recently published French guidelines for scenario design in simulation-based education [8].

The scenario for immersive simulation was the management of a gunshot wound patient in the emergency department (ED), that for procedural simulation was performing a lumbar puncture in a young febrile patient, and that for SPs was the communication of bad news (unexpected death of a relative) in the ED.

QASSH structure and content

The QASSH is a checklist with items marked dichotomously (present/absent), which evaluates the quality of a scenario script (i.e., before the scenario is actually deployed during simulation). It consists of a global section applicable to all simulation modalities and one section that is specific to each of the simulation modalities (*see the QASSH in the supplemental material*). As such, the QASSH could be considered a set of three distinct assessment instruments, each tailored to one of the targeted simulation modalities.

The global section includes 24 items split among 4 categories (alignment between learning needs and scenario design, 4 items; scenario context, 10 items; scenario writing, 7 items; quality improvement, 3 items). The specific section includes 10 (immersive simulation and procedural simulation) or 11 items (SP). Items are scored as 0 (absent) or 1 (present) for a total score of 34 or 35 depending on the specific instrument.

It should be noted that, in the case of procedural simulation, the term 'scenario' is used synonymously with the simulation or learning experience itself (i.e., it is not merely the description of a clinical scenario used to contextualize learning).

QASSH development

The draft version of the QASSH assessment instrument was based on the same literature search that generated the French guidelines for scenario design in simulation-based education, described elsewhere [8]. It was then augmented by the experience of 12 simulation experts who are members of the QASSH working group created by the French-speaking Society for Simulation in Healthcare (SoFraSimS). Development involved an extensive iterative process that generated 3 intermediary versions of the instrument before the QASSH was finalized.

The QASSH was then tested prospectively by evaluating the three simulation scenarios within each modality. A total of 35 raters (12 involved in the process of creating the QASSH, the 'experienced raters', and 23 French-speaking simulation experts within the SoFraSimS network, the 'naive raters') used the QASSH to independently assess one scenario of a different quality within each simulation modality. For example, rater 1 might have evaluated the high-quality scenario for immersive simulation, the borderline scenario for procedural simulation, and the low-quality scenario for SPs. Within each group of raters assessing a given scenario, some of them were familiar with the QASSH (experienced raters) and some were using it for the first time (naive raters). For each QASSH variant, each scenario quality was assessed using 34 items (35 for SPs), by a different group of 8 independent raters (for a total of 24 raters across the three scenarios). The passing scores for the scenarios were based on the means of scores of the borderline scenarios: 15 for immersive simulation, 20 for SPs and 15 for procedural simulation. None of the raters was involved in designing the scenarios and they were blinded to any given scenario difficulty. Naive raters were deliberately not trained in the use of the QASSH. Responses were collected over a one-month period in May 2023. As the study was conducted with the French scale, it was translated in English by an English speaker. It was then translated back into French by a separate person and compared to the original French version (double translation) by the study authors. This process ensures that both the original French version and the translated English version have item equivalence (the items are relevant in both languages) and semantic equivalence (the items have the same meaning in both languages) [19]. It is an essential part of test content evidence for validity. Figure 1 summarizes this process.

Demonstrations of validity

Using the *Standards for educational and psychological testing* as our framework, five categories of evidence were explored: test content, internal structure, relation to other variables, testing consequences and response process [20]. In this analysis, each of the three QASSH variants (corresponding to the three simulation modalities) was considered a distinct instrument.

The test content evidence for validity aimed to assess whether the instructions and item content are relevant to the purpose of the tool. We described the QASSH development, structure and content in the preceding section.

The internal structure evidence for validity could be summarized by the following questions. Are the relations between the items congruent with what is expected? Is the use of the QASSH generalizable to similar contexts? Both questions were answered with Generalizability Theory (GT) [15,16], using generalizability studies (G studies). We assessed the differences between the items by analyzing their contribution to the total variance of the score. An absolute G coefficient was used to assess the generalizability (reliability) of the results.

The response process evidence for validity could be summarized by the following questions: do the raters use the tool appropriately? Are the raters familiar with the instrument? This will be considered in the discussion, but also through analysis of interrater reliability, and through the differences, if any, between naive and expert raters. Furthermore, we performed optimization studies (D studies) to determine the reliability of the QASSH when used by different numbers of raters.

The testing consequences evidence for validity determined whether the conclusion drawn from the measurement were appropriate. For this purpose, we considered a cut-score ('pass-or-fail' decision for each scenario, i.e., whether the scenario can be considered of adequate quality) based on the means of the borderline scenario. We then determined reliability of this decision using the criterion-referenced coefficient $\Phi(\lambda)$ [21].

The relation to other variables evidence for validity, which evaluated whether the results correlate to other variables as expected, was beyond the scope of this study.

Generalizability theory (GT)

GT is a robust framework to determine the dependability (reliability or generalizability) of a measurement. It allows us to analyze reliability across multiple sources of variation and to quantify the errors that exist, or could exist, over several replications of a measurement. Stated more formally, it allows generalizing from an observed score to

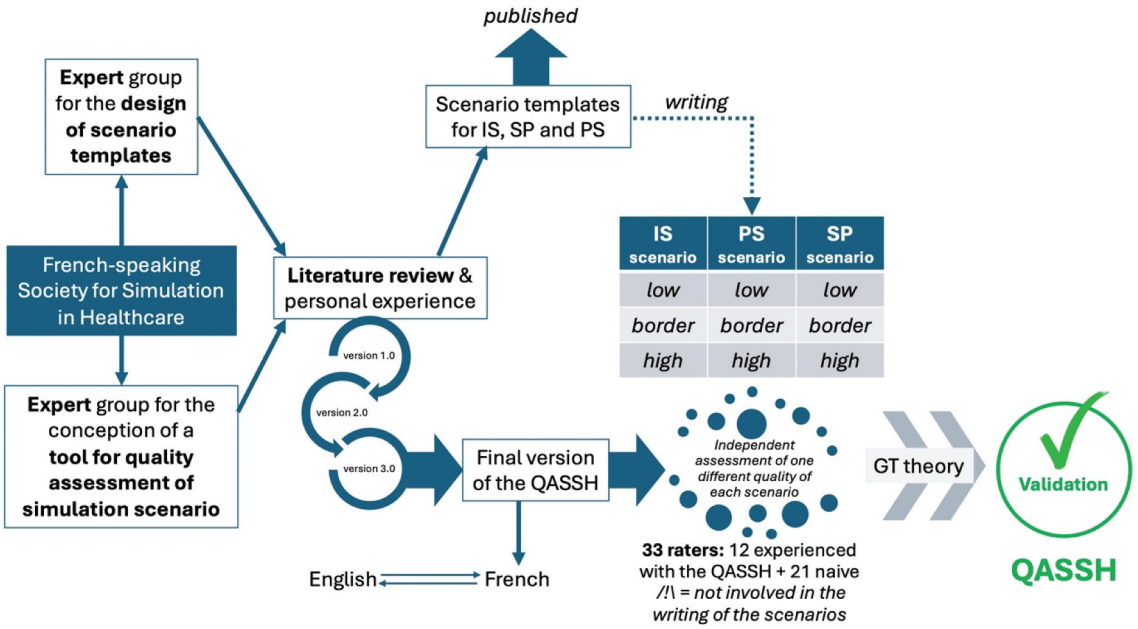

Figure 1. Methodology for the validation of the QASSH to assess the quality of scenarios for team-based (or manikin-based) immersive simulation, simulated participants or procedural simulation. *GT = Generalizability theory; is = Immersive simulation; PS = Procedural simulation; SP = Standardized patients.*.

a 'universe' of scores under different conditions of measurement. Contrary to the methods of Classical Test Theory, it is suitable for complex research designs such as the one provided here and can often untangle the different sources of error (or variance). It is considered a 'major contribution to psychometrics' [16].

GT has two components: a theoretical model and a mathematical model similar to ANOVA. Through G studies, it isolates the effects of specific sources of variance ('facets' or components) on the measurement to identify those that introduce an error in measurement (a bias) by themselves and by interacting with other facets. It also provides measures of reliability (G coefficients). Finally, through optimization studies (D studies), GT can determine the optimal test conditions, i.e., the conditions that reduce errors and increase reliability.

Three facets were taken into account in the G studies for the three variants of QASSH (Table 1). First were the scenarios (S) or, more appropriately, the scenario qualities that constituted the object of measurement, i.e., the facet that must be reliable and valid. The other two facets, which can introduce biases, were the raters (R), and the QASSH items (I). All items were considered as invariant (*fixed facet* in GT terms), and the other components were considered as extracted from a very large target population (*infinite universe* in GT terms). The design of the G studies is a mixed model 'I × R:S', meaning that all scenario qualities are assessed by all items (I × S) but raters evaluated only a subset of scenarios, i.e., the R facet was *nested in* the S facet (R:S). The reliability of the scores was evaluated relatively to the exact concordance of the scores (absolute G coefficient or Coef-G).

G-facet analyses were conducted to determine whether excluding any item (I) from analysis improved Coef-G, which would indicate a faulty item.

D studies determined the number of raters assessing a given scenario that was sufficient to ensure reliability of the QASSH (i.e., Coef-G above 0.8). The value of 0.8 was chosen because it is an adequate

Table 1. Scores and descriptive statistics of the three instruments (QASSH variants) studied.

	QASSH variant		
	Immersive simulation	Procedural simulation	Simulated participants
Number of items/total score	34	34	35
Number of raters	8	8	8
Number of scenarios/scenario quality	3	3	3
Mean score	17,4	17,8	17,7
Median score	15,5	15,5	17,0
Score for low-quality scenario (mean ± SD)	9.75 ± 3.81	7 ± 7.45	7.25 ± 4.23
Score for threshold scenario (mean ± SD)	14.25 ± 6.82	15.50 ± 5.04	19.88 ± 5.94
Score for high-quality scenario (mean ± SD)	28.13 ± 1.64	30.88 ± 2.3	26 ± 5.26

SD = Standard deviation.

level of reliability for high-stake assessments [22,23]. Analyses were done with EduG 6.1-f (Educan, Longueil, Canada).

Statistical analyses

Within each variant of QASSH and individual scenarios, the differences (or lack thereof) between the scores of experienced and naive raters were evaluated through Student's *t*-test. Interrater correlations were also calculated through intra-class coefficients (ICC). Since it is often reported, internal consistency of each variant was measured by Cronbach's alpha coefficient. All statistical measures were made with XLSTAT statistical and data analysis solution (2023).

Results

Table 2 provides the total and average scores for each of the QASSH variants. The sources of variance in the QASSH scores determined by the G studies are given in Table 3. The calculated Coef-G were: 0.97 (immersive simulation), 0.96 (SP), and 0.98 (procedural simulation). G-facet analyses showed that no removal of a single item of the QASSH significantly increases Coef-G above 0.01 for any of the three variants. Table 4 shows the result of the D studies. It reports Coef-G of the QASSH variants for different numbers of raters rating the same scenario.

The interrater reliability was 0.66 for one measure ($p < 0.0001$) and varied between 0.74 and 0.87 for all measures ($p < 0.0001$), except for the rating of the high-quality scenario in procedural simulation for which the ICC coefficient was 0.21 (95% CI 0.28–0.56). Student's *t*-test showed significant differences in mean scores ($p < 0.025$) of all scenario qualities within a given simulation modality, except for the mean scores between the high-quality and threshold scenario in SP and the threshold and low-quality scenario in immersive simulation.

Student's *t*-test for the differences between experienced and naive raters within each scenario in each variant of QASSH showed no statistical differences, except for the threshold scenario in SP (95% CI 0.431–0.125) and the high-quality scenario in immersive simulation (95% CI 0.132–0.006).

The criterion-referenced coefficients $\Phi(\lambda)$, which determine the reliability relative to the cut-score, were: 0.97 (immersive simulation), 0.96 (SP) and 0.98 (procedural simulation).

Cronbach's alpha was 0.94, 0.94 and 0.97 for the three QASSH variants, respectively, immersive simulation, SP and procedural simulation.

Discussion

Using the *Standards for educational and psychological testing* [14] as the validity framework to assess the QASSH, four categories of evidence for validity were

Table 2. Facets or sources of variance for the three QASSH variants.

Facet	Number (n)	Universe size*	Face type	Description
Scenario quality (S)	3	Infinite	Random	The object of measurement
Items (I)	34 (IS and PS) 35 (SPs)	34 (IS and PS) 35 (SPs)	Fixed	The items of the scale
Rater per scenario quality (R:S)	8	Infinite	Random	The raters or evaluators evaluating a given scenario quality

*Universe size represents the universe of admissible values.
IS = Immersive simulation; PS = Procedural simulation; SP = Simulated participants.

Table 3. Results of the G-study analysis. For each QASSH variant, the individual sources of variance (variance components) are provided. See text for further details.

Variance components	Sum of squares	*df*	Mean square	Corrected variance	%*
Immersive simulation variant					
S	43.169	2	21.585	0.077	27.4
I	37.192	33	1.127	0.034	12.2
R:S	13.114	21	0.624	0.018	6.5
I × S	18.664	66	0.283	0.019	6.7
I × R:S, *e***	91.761	693	0.132	0.132	47.2
Simulated participants					
S	41.788	2	20.894	0.072	25.7
I	38.762	34	1.140	0.034	12.0
R:S	16.182	21	0.771	0.022	7.9
I × S	21.045	68	0.309	0.023	8.1
I × R:S, *e***	92.193	714	0.129	0.129	46.3
Procedural simulation					
S	68.914	2	34.457	0.124	41.8
I	20.433	33	0.619	0.014	4.8
R:S	17.732	21	0.844	0.025	8.4
I × S	17.586	66	0.266	0.019	6.5
I × R:S, *e***	78.893	693	0.114	0.114	38.5

*Proportion of corrected variance (i.e., percentage of total variance in the score introduced by each facet).
**This source of variance includes variance linked to the interaction of all the facets (I × R:S), as well as the unidentified residual variance (*e*). See text for further details.
S = Scenario quality; I = Items; *R* = Raters; *df* = degrees of freedom.

Table 4. Results of the generalizability analyses for each QASSH variant providing the reliability (G) coefficients, the error variance and the standard errors of measurement under varying numbers of raters (R:S). The first column (G study) provides the data as measured in this study, with 8 raters.

Element	G Study	D study 1	D study 2	D study 3	D study 4
Immersive simulation variant					
R:S	8	1	2	3	4
Absolute G coefficient	0.97	0.81	0.89	0.93	0.94
Error variance	0.002	0.018	0.009	0.078	0.068
Standard error of measurement	0.048	0.136	0.096	0.078	0.068
Simulated participants (SP)					
R:S	8	1	2	3	4
Absolute G coefficient	0.96	0.765	0.867	0.907	0.929
Error variance	0.003	0.022	0.011	0.007	0.006
Standard error of measurement	0.052	0.148	0.105	0.086	0.074
Procedural simulation					
R:S	8	1	2	3	4
Absolute G coefficient	0.98	0.83	0.91	0.94	0.95
Error variance	0.003	0.025	0.012	0.008	0.006
Standard error of measurement	0.056	0.158	0.111	0.091	0.079

G = Generalizability; D = Design; R:S represents the raters facet nested within the scenario facet (see text for further details).

tested: test content, internal structure, response process, and testing consequences. The **test content** source of validity evidence provides an overview of the QASSH development process, which relies on a previously published framework as well as theoretical foundations in simulation-based education. Evidence for **internal structure** of the QASSH is mainly provided by the G studies. All three variants of the QASSH show very high reliability coefficients as attested by absolute G coefficients (Coef-G) above 0.95. The ratings also show very high internal consistency as evaluated by Cronbach's alpha.

Analysis of the percentages of variance of the total score associated with each facet and their interactions (Table 3) provides further evidence for the validity of QASSH. For all QASSH variants, and excluding the residual variance (see below), the scenario quality contributes the most to the variance of the total score (as high as 41.8% for procedural simulation). This is expected and desirable, since scenario quality is the measurement object, i.e., the construct being assessed, QASSH being an instrument aimed at discriminating between scenarios of various qualities.

The contribution to total variance of the interaction between scenario quality (S) and items (I) is low for all three variants, suggesting that the individual items in the scale behave similarly for all levels of scenario quality. It is possible that the interaction between the three facets S, I, and R – which is included in the residual variance *e*—is higher than desirable given the high contribution of the residual variance to overall variance. However, such high numbers for the residual variance are not unexpected because it also includes all other sources of variance not considered in the analysis.

Given the nature of the G-studies design, it is impossible to dissociate the variance specifically related to raters (R) from the variance related to any eventual interaction between the raters (R) and the scenario quality (S) given that the facet R is nested within S (R:S). However, R:S contributes little to overall variance (under 8.5% for all three variants), suggesting there are few errors related to the raters themselves or to the interaction between the raters and the scale (e.g., the tendency of some raters to use the scale differently for different scenario qualities).

In short, all three variants show very high generalizability (i.e., reliability), with the construct of interest (scenario quality) contributing most to overall variance after the residual variance and the other facets contributing little. This is also reflected in the low Standard Error of Measurement (SEM ≤0.056 on a scale from 0 to 1) for the QASSH score as shown in Table 4.

The **testing consequences** evidence for validity was demonstrated by determining the generalizability coefficients associated with a cut-score corresponding to the mean score of the borderline scenarios. All three values of $\Phi(\lambda)$ were above 0.95, implying that the decision to reject scenarios that score below the cut-scores seems appropriate. Care should be taken with this conclusion; however, since the number of scenarios rated (3) is fairly low, and the scenarios were specifically designed to three levels of quality. Additional testing should be done, ideally with a crossed design for GT and scenarios designed to random levels of quality, in order to establish an adequate cut-score.

The **response process** evidence for validity aims to determine that no errors were associated with the raters' use of the instrument. As previously stated, raters seem to contribute little to overall variance; this is also documented by the interrater reliability coefficients which, aside from the evaluation of the high-

quality scenario for procedural simulation, can be considered either moderate (0.5–0.75; 2 measures) or good (0.75–0.9; 6 measures) [24]. This is further demonstrated by the D studies, which show high generalizability coefficients even when the scenarios were evaluated by a single rater (Coef-G > 0.75). With two raters, Coef-G increased significantly to values above 0.85. This establishes, given the expected usage of QASSH, that a single rater can reliably evaluate a scenario's quality, but that error in measurement is significantly reduced with two raters.

The lack of differences in most scores between raters familiar with QASSH and those who were using it for the first time is reassuring and suggests that prolonged training is unnecessary. This conclusion, however, is presumptive given the significant differences in scores between expert and naive raters for the threshold scenario in SP and the high-quality scenario in immersive simulation and given the low ICC for the high-quality procedural simulation. These discrepancies might be due to greater subjectivity in rating, to varying expertise in the specific simulation modality or to bias effects such as central tendency bias. It could also be explained by the lack of training in the naive raters, which would suggest that a minimal amount of training should still be favored, as is usually the case when introducing a new assessment instrument. While more research would be needed to untangle the specific effects related to raters, the small variances associated with raters suggest their ability to distinguish between scenarios was preserved.

Two other tools to assess scenarios in SBE have previously been published : SSET [25] (US) and SSQI [26] (Saudi Arabia). Table 5 summarizes the differences between those tools and the QASSH. SSET was published first and was recently complemented by SSQI. While SSET is too broad for an extensive analysis of simulation scenarios, it allows a good overall review of most central concepts. We believe that QASSH covers more aspects of scenario engineering, with the exception of some elements linked to debriefing (place and method suggested in the SSQI [26] which we consider unrelated to scenario design. The QASSH assesses new elements, such as the triggers of observable behaviors and the natural feedback that follows from learners' actions, elements we justified in our previous work on the design of simulation scenario templates [8]. Furthermore, QASSH specifically evaluates three different (although sometimes overlapping) simulation modalities, thus providing tools that can better match the educators' or program managers' intent.

We showed that QASSH is a valid and reliable instrument for measuring scenario quality for the three simulation modalities. However, our study does have limitations. First, all raters were experts in SBE (although not necessarily in QASSH, as stated previously), so the results would not generalize to raters without such expertise (e.g., subject-matter experts with little to no expertise in simulation).

Second, the G study design is not entirely crossed, with one nested facet that reduces confidence in some of its conclusions, particularly those involving raters. We tried to minimize this impact by providing other evidence such as the ICC and the differences between naive and expert raters, but it is possible that some interactions between raters and scenario quality introduce biases that cannot be assessed; as previously stated; however, those are expected to be small, if present.

Third, while we considered the S facet as representing the quality of the scenario, it is possible that other factors related to the scenario, such as the nature of the case, may introduce unexpected errors. We tried to minimize this effect by using the same scenario with varying quality levels within each simulation modality, but it is impossible to ascertain whether other aspects related to the scenario may have played a role.

Finally, given that the QASSH was originally developed in French, linguistic and cultural idiosyncrasies may limit its use in other cultural settings. All raters were recruited from the SoFraSimS network (the French-speaking simulation society) which limits generalizability to other linguistic or cultural contexts, particularly for the English version. Further research to determine the equivalence of the English version of QASSH is warranted.

The aim of this study was not to demonstrate whether the QASSH could effectively improve scenario quality over time; a multicentric study across the SoFraSimS network will be undertaken to that end.

Nevertheless, we believe that we provided a tool capable of fully assessing the multiple elements that affect scenario design and that would allow optimal use of SBE. The QASSH is a useful tool for both experienced and novice educators for improving existing simulation activities or designing new ones. It can be used during the process of scenario design to ensure an adequate scenario quality, as a post-hoc rating tool to diagnose problematic scenarios, and in a quality-improvement process within a simulation program. Its main target users consist of experienced educators in charge of simulation facilities or educational programs in initial or continuous medical education. No tool for assessing procedural simulation scenarios or activities was previously published, a gap which the QASSH fills nicely.

Table 5. Differences between the three tools designed to assess the quality of scenarios in simulation-based education: the QASSH, the SSET and the SSQI.

Items	QASSH	SSET	SSQI
Acronym signification	Quality Assessment of Scenarios in Simulation in Healthcare	Simulation Scenario Evaluation Tool	Simulation Scenario Quality Instrument
Countries of origin	France, Canada, Belgium, Switzerland, Morocco, Qatar, Tunisia	USA	Saudi Arabia
Type	Checklist with 24 common and 10 to 11 specific items marked dichotomously (present/absent)	Global rating scale with 20 items rated on a 5-point scale and three anchors per item	Global rating scale with 44 items rated on a 3-point scale: meets Expectations = (2), needs Improvement, (1), inadequate (0) without anchor
Simulation modality(ies) assessed	Team-based (or manikin-based) immersive simulation + Simulated participants + Procedural simulation	Team-based (or manikin-based) immersive simulation + simulated participants	Team-based (or manikin-based) immersive simulation + simulated participants
Overall methodology	Literature search, then focus group of international French-speaking simulation experts to provide after 3 iterations a final version of the tool, further assessed by raters familiar with the QASSH and naive raters on three versions of increasing quality of three scenario, one for each simulation modality	Literature search then focus groups of national simulation experts in a modified two-round Delphi approach grounded on six templates published for written scenario design, with kappa agreement testing	Sequential transformative mixed-method research design: literature search then focus groups with qualitative analyses by constant comparison analysis (average content validity index)
Number of simulation experts involved	35	38	17
Number of scenarios assessed	3 declined in low, borderline and high quality	None (Delphi with experts)	125 of any quality
Statistical analyses	GT; Cronbach's alpha coefficients	Free marginal kappa, then two-way mixed intraclass correlation model for absolute agreement, ANOVA, and post-hoc HSD	Factor and confirmatory factor analysis; Cronbach's alpha coefficients
General content			
Appropriate learning needs	x	x	
Specification of prerequisite learning material		x	x
Level of difficulty tailored to learners	x		x
Focused and limited learning objectives	x	x	
Learning objectives written according to BEME	x	x	x
Learning objectives cover the professional domain and role of all target learners	x	x	x
Learning objectives are clearly detailed for each procedure (PS)	x		
Adaptation to cultural milieu			x
Scenario designer's name and contact	x		
Verification of adequate training of instructor(s)	x		x
Leading institution name/contact	x		
Description of the educational context	x		x
Scenario linked to specific content in curriculum	x		
Description of the target learner population	x	x	x
Chosen simulation modality is adequate	x		
Specification of the items for appropriate authenticity + cognitive load limitation	x		x (*distractors*)
Explanation of the level of required realism	x	x	x
Extensive details of the initial state	x		x
Planification of adequate pre-briefing	x		x
Clear scenario title	x		
Durations suggested (briefing, simulation session and debriefing)	x		x
Level of control on the learning activity clearly considered by the simulation team	x		x
Roles thoroughly described	x		x

(*Continued*)

Table 5. (Continued).

Items	QASSH	SSET	SSQI
Specification of the debriefing method			x
Specification of the debriefing site			x
Description of the evidence- or experience-based strategies to solve the clinical problem	x	x	
Suggestion of strategies to initiate learning transfer to real clinical situations	x		
Clear and detailed scenario script	x		x
Dedicated time to improve the scenario	x		
Strategies to prevent deskilling	x		
Specific content			
Clear & concise scenario briefing/ clinical vignette	x		x
Summary of the scenario	x		
Details of the initial state of the manikin/patient	x	x	x
Specific answers to learner questions and standard answers for non-pertinent questions	x		
Specification of patient's or relatives' ability to communicate and emotional state	x		
Planification of triggers for observable behaviors	x		
Planification of natural feedback to these triggered behaviors	x		x
Adjustment of the intensity of actor's emotions according to learner(s) performance	x		
Anticipation of anonymized documents/media elements	x	x	x
Planification of multiple endings of the scenario	x		
Description of the actors' role during debriefing	x		
Planification of a dry-run (piloting) to improve the scenario before it is administered to learners	x		
Planification of corrective measures to ensure that learners reach the expected learning objectives	x		x
Appropriate simulator:learner ratio (PS)	x		
Appropriate the instructor:learner ratio (PS)	x		
Possibility of hybrid simulation with simulated participants or embedded simulation personnel (PS)	x		
Variation of learning activities for each procedure (PS)	x		
Specification of adequate clinical contextualization of the procedure for authenticity (PS)	x		
Provision of an adequate assessment scale (PS)	x		x
Provision of a focused and specific feedback (PS)	x		
Possibility of deliberate practice (PS)	x		
Possibility of mental imagery practice or educational resources to further study the procedure and avoid skill decay (PS)	x		

HSD: Tukey Honestly significant Difference range test; PS: Procedural simulation.

Conclusion

In this article, we have described the 'Quality Assessment of Scenarios in Simulation in Healthcare' (QASSH), a new tool for assessing the quality of scenario scripts for three different simulation modalities: immersive simulation, SPs and procedural simulation, provided in Appendix. Using GT, we were able to provide validity and reliability evidence for its purpose, despite some limitations. Using two raters instead of one yielded more reliable scores. As with any new assessment tool,

additional studies should be conducted, especially to demonstrate the validity of the English version of the instrument.

We hope the QASSH will be useful to simulation educators as well as simulation program managers wishing to provide learners with a high-quality learning experience and to assess the quality of written scenarios before excessive resources are expended to make them a reality.

Acknowledgments

The authors amend the teams that also made great efforts for building tools to assess simulated scenarios (the SSET and SSQI teams) because we believe such works are of importance given the elevated human and financial costs of such a useful educational tool that is SBE.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The author(s) reported there is no funding associated with the work featured in this article.

Authors' contribution

ER, GDS & GC designed research. GDS and ER led the work with the group of experts in France, MLT & GC in Canada. ER, GDS & GC and built the QASSH scale. GC proceeded to GT analyses. ER & GC wrote the manuscript. All authors agreed with the final version of the manuscript.

Availability of data and materials

The original data can be retrieved via the corresponding author.

Ethics approval

All participants gave written consent to participate to the research project.

Abbreviations

ANOVA	ANalysis Of Variance test
ED	Emergency Department
GT	Generalizability Theory
ICC	Inter-Class Coefficients
QASSH	Quality Assessment of Simulation Scenario in Healthcare
SBE	Simulation-Based Education
SoFraSimS	French-speaking Society for Simulation in Healthcare
SPs	Simulated (or Standardized) Participants
SSET	Simulation Scenario Evaluation Tool
SSQI	Simulation Scenario Quality Instrument

ORCID

Etienne Rivière http://orcid.org/0000-0003-0254-3394

References

[1] Cook DA, Hatala R, Brydges R, et al. Technology-enhanced simulation for health professions education: a systematic review and meta-analysis. JAMA. 2011;306(9):978–988. doi: 10.1001/jama.2011.1234

[2] McGaghie WC, Draycott TJ, Dunn WF, et al. Evaluating the impact of simulation on translational patient outcomes. Simul Healthcare. 2011;(6 Suppl7): S42–47. doi: 10.1097/SIH.0b013e318222fde9

[3] Diaz-Navarro C, Armstrong R, Charnetski M, et al. Global consensus statement on simulation-based practice in healthcare. Adv Simul. 2024;9(1):19. doi: 10.1186/s41077-024-00288-1

[4] Chiniara G, Cole G, Brisbin K, et al. Simulation in healthcare: a taxonomy and a conceptual framework for instructional design and media selection. Med Teach. 2013;35(8):e1380–e1395. doi: 10.3109/0142159X.2012.733451

[5] Rivière E, Saucier D, Lafleur A, et al. Twelve tips for efficient procedural simulation. Med Teach. 2018;40(7):743–751. doi: 10.1080/0142159X.2017.1391375

[6] Lewis KL, Bohnert CA, Gammon WL, et al. The Association of Standardized Patient Educators (ASPE) Standards of Best Practice (SOBP). Adv Simul. 2017;2(1):10. doi: 10.1186/s41077-017-0043-4

[7] Salas E, Wilson KA, Burke CS, et al. Using simulation-based training to improve patient safety: what does it take? Joint Commission J Qual & Patient Saf. 2005;31(7):363–371. doi: 10.1016/S1553-7250(05)31049-X

[8] Der Sahakian G, de Varenne M, Buléon C, et al. The 2024 French guidelines for scenario design in simulation-based education: manikin-based immersive simulation, simulated participant-based immersive simulation and procedural simulation. Med Educ Online. 2024;29(1):2363006. doi: 10.1080/10872981.2024.2363006

[9] Lioce L, Reed CC, Lemon D, et al. Standards of best practice: simulation standard III: participant objectives. Clin Simul Nurs. 2013;9(6):S15–S18. doi: 10.1016/j.ecns.2013.04.005

[10] Der Sahakian G, Buléon C, Yall D. Chapter 24 - a pragmatic approach to scenario scripting. In: Clinical Simulation. 2nd ed. 2019. p. 337–343.

[11] Benishek LE, Lazzara EH, Gaught WL, et al. The template of events for applied and critical healthcare simulation (TEACH sim): a tool for systematic simulation scenario design. Simul Healthcare. 2015;10(1):21–30. doi: 10.1097/SIH.0000000000000058

[12] McIntosh C, Macario A, Flanagan B, et al. Simulation: what does it really cost? Simul Healthcare. 2006;1(2):109. doi: 10.1097/01266021-200600120-00041

[13] Hatala R, Norman GR, Brooks LR. Influence of a single example on subsequent electrocardiogram interpretation. Teach Learn Med. 1999;11(2):110–117. doi: 10.1207/S15328015TL110210

[14] Standards for Educational & Psychological Testing. Washington, D.C: American Educational Research Association; 2014.

[15] Cronbach LJ, Rajaratnam N, Gleser GC. Theory of generalizability: a liberalization of reliability theory†. The Br

J Stat Phychol. 1963;16(2):137–163. doi: 10.1111/j.2044-8317.1963.tb00206.x

[16] Brennan RL. Generalizability theory. (NY): Springer; 2001.

[17] Shavelson RJ, Webb NM. Generalizability theory: a primer. SAGE Publications; 1991.

[18] de Varennes M, Chiniara G, Lafleur A. Chapter 23 - a systematic approach to scenario design. In: Clinical Simulation. 2nd ed. 2019. p.315–335.

[19] Streiner DL, Norman GR, Cairney J. Health measurement scales: a practical guide to their development and use. Oxford: Oxford University Press; 2015.

[20] Downing SM. Validity: on the meaningful interpretation of assessment data. Med Educ. 2003;37 (9):830–837. doi: 10.1046/j.1365-2923.2003.01594.x

[21] Cardinet J, Johnson S, Pini G. Applying generalizability theory using EduG. (NY): Routledge; 2011. p. 216.

[22] Hamdy H, Prasad K, Williams R, et al. Reliability and validity of the Direct Observation Clinical Encounter Examination (DOCEE). Med Educ. 2003;37 (3):205–212. doi: 10.1046/j.1365-2923.2003.01438.x

[23] Burch VC, Norman GR, Schmidt HG, et al. Are specialist certification examinations a reliable measure of physician competence? Adv Health Sci Educ. 2008;13 (4):521–533. doi: 10.1007/s10459-007-9063-5

[24] Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15(2):155. doi: 10.1016/j.jcm.2016.02.012

[25] Hernandez J, Frallicciardi A, Nadir N-A, et al. Development of a Simulation Scenario Evaluation Tool (SSET): modified delphi study. BMJ Simul Technol Enhanc Learn. 2020;6(6):344. doi: 10.1136/bmjstel-2019-000521

[26] Mujlli G, Al-Ghosen A, Alrabah R, et al. Development and validation of Simulation Scenario Quality Instrument (SSQI). BMC Med Educ. 2023;23(1):972. doi: 10.1186/s12909-023-04935-5

Appendix

QASSH

Quality Assessment for Simulation Scenarios in Healthcare

Etienne RIVIERE, Guillaume DER SAHAKIAN, Marie-Laurence TREMBLAY, Gilles CHINIARA
(for the SoFraSimS French-speaking simulation network)

1- Global items		Tick if done
ALIGNMENT BETWEEN LEARNING NEEDS AND SCENARIO DESIGN	1.1 The methods used to assess the learning needs are appropriate and adequately described	
	1.2 The level of difficulty is tailored to the learners' level of professional development	
	1.3 The learning objectives are focused/limited and clearly written in accordance with best practices	
	1.4 The learning objectives cover the professional domain and role of all target learners	
CONTEXT OF THE SCENARIO	**1.5** The **scenario designer's name** and **contact information** are provided	
	1.6 The **members of the instruction team** are adequately trained to facilitate simulation-based educational activities	
	1.7 The **leading institution name** is provided	
	1.8 The **educational context** is described	
	1.9 The scenario is linked to a well-identified specific content in the **curriculum**	
	1.10 The **target learner population** is adequately described	
	1.11 The chosen **simulation modality** is adequate and mentioned	
	1.12 The scenario is designed to ensure **appropriate authenticity** and to **limit cognitive load** if need be	
	1.13 The level of **required realism** is explained	
	1.14 The **initial state of the environment** is extensively detailed (equipment included)	
WRITING OF THE SCENARIO	**1.15** An adequate **pre-briefing** is planned	
	1.16 A **clear scenario title** is present	
	1.17 The **duration** of briefing, simulation, and debriefing/feedback is adequate and suggested	
	1.18 The simulation team level of **control on the learning activity** is clearly taken into account	
	1.19 The **roles of the actors** (embedded simulation personnel) and of the patient are thoroughly described	
	1.20 Evidence- or experience-based strategies to solve the clinical problem are clearly written	
	1.21 Strategies to initiate **learning transfer to real clinical situations** are suggested	
QUALITY IMPROVEMENT	**1.22** The scenario script is **clear and detailed**	
	1.23 Time is dedicated after the simulation session to consider how to **improve the scenario**	
	1.24 Strategies to prevent **deskilling** of the learners in relation to the learning objectives are proposed	

2- Items specific to immersive simulation	Tick if done
2.1 The **scenario briefing** is clear and concise	
2.2 A **summary** of the scenario is provided	
2.3 The **initial state of the patient** is extensively detailed	
2.4 Specific answers to learner questions and standard answers for non-pertinent questions are clearly written	
2.5 Triggers are planned for important observable behaviors	
2.6 Natural feedback to these triggered behaviors is planned	
2.7 Anonymized **documents/media elements** are provided when appropriate	
2.8 Multiple **endings of the scenario** are planned	
2.9 A **dry-run (piloting)** and dedicated time to improve the scenario are planned before it is administered to learners	
2.10 Corrective measures are planned to ensure that learners reach the expected learning objectives during the simulation	

3- Items specific to simulated/standardized participants	Tick if done
3.1 An appropriate **briefing or clinical vignette** (for OSCEs) is provided to learners before simulation	
3.2 A **summary** of the scenario is provided	
3.3 The **initial state of the patient** is extensively detailed	
3.4 The patient's or relatives' ability to **communicate** and their **emotional state** are specified	

(*Continued*)

(Continued).

1- Global items	Tick if done
3.5 Specific answers to learner's questions or standard answers for non-pertinent questions are clearly written for each actor	
3.6 The **intensity of actor's emotions** is adjusted according to learner performance or the targeted population of learners	
3.7 The actors' role during **debriefing or feedback** to the learners is described	
3.8 Anonymized **documents/media elements** are provided when appropriate	
3.9 Multiple **endings of the scenario** are planned	
3.10 A **dry-run (piloting)** and dedicated time to improve the acting and the scenario are planned before it is administered to learners	
3.11 Corrective measures are planned to ensure that learners reach learning objectives	

4- Items specific to procedural simulation	Tick if done
4.1 Learning objectives are clearly detailed for each procedure	
4.2 The **simulator:learner ratio** is appropriate	
4.3 The **instructor:learner ratio** is appropriate	
4.4 hybrid simulation with simulated participants or embedded simulation personnel is adequately chosen when appropriate	
4.5 Learning activities are **varied** for each procedure	
4.6 Authenticity is achieved by an adequate clinical contextualization of the procedure	
4.7 An adequate **assessment** scale is provided	
4.8 The focus of specific **feedback** given to learners is provided	
4.9 Deliberate practice is offered to learners whenever possible	
4.10 A mental imagery practice or **educational resources** to further study the procedure and avoid skill decay are proposed to learners	