

Tutorial

How to Conduct Usability Testing of Medical Education Simulations: Methods Tutorial Using a Telemedicine Simulation for Maternal Care

Blake J Lesselroth^{1,2}, MD, MBI; Alexander Douglas³, DO; Helen Monkman^{1,2}, PhD, MA; Romaric Marcilly^{4,5}, PhD; Karalane Bellavia¹, BA; Mataeo Anderson⁶, BS; Jacob Van Buren⁷, MD; Charles Parsons⁸, BS; Sonakshi Mohanty⁹, BS; Juell Homco¹, PhD, MPH; Morgan Richards⁷, DO; Karen P Gold⁷, MD, MSci

¹Department of Medical Informatics, School of Community Medicine, University of Oklahoma - Tulsa, Tulsa, OK, United States

²School of Health Information Science, University of Victoria, Victoria, BC, Canada

³Department of Internal Medicine, Oklahoma State University, Tulsa, OK, United States

⁴Univ. Lille, CHU Lille, ULR 2694 - METRICS: Évaluation des technologies de santé et des pratiques médicales, F-59000, Lille, France

⁵Inserm, CIC-IT 1403, F-59000, Lille, France

⁶University of Oklahoma Polytechnic Institute, University of Oklahoma, Tulsa, OK, United States

⁷Department of Obstetrics and Gynaecology, School of Community Medicine, University of Oklahoma - Tulsa, Tulsa, OK, United States

⁸School of Community Medicine, University of Oklahoma - Tulsa, Tulsa, OK, United States

⁹Department of Computer Science and Design, University of California San Diego, San Diego, CA, United States

Corresponding Author:

Blake J Lesselroth, MD, MBI

Department of Medical Informatics

School of Community Medicine

University of Oklahoma - Tulsa

Schusterman Center

4502 E. 41st. Street

Tulsa, OK, 74135

United States

Phone: 1 9713033627

Fax: 1 9186603059

Email: Blake-Lesselroth@ou.edu

Abstract

Background: Clinical educators regularly use simulation-based education to help learners develop clinical reasoning, procedural skills, and communication strategies. However, differences between simulated and live clinical environments, such as missing case details or workflow interruptions, can reduce realism and divert learners' attention from the intended learning goals. While published design guidelines recommend pilot testing simulations, they do not explain how to systematically identify usability problems. Human factors evaluation methods can address this gap by identifying issues before piloting and providing design insights.

Objective: This tutorial describes our usability testing protocol for evaluating educational simulations before piloting with learners. Using a telemedicine maternal care simulation as an example, we show how mixed methods usability testing can identify technical, workflow, and assessment-related issues in simulation design.

Methods: We recruited obstetricians and family practitioners to test our simulation. To measure multiple usability dimensions concurrently, we used several published instruments, including an Agile Task Analysis (ATA), the Single Ease Question (SEQ), an adapted System Usability Scale (SUS), the Satisfaction With Simulation Experience Scale (SSES), and the NASA (National Aeronautics and Space Administration) Task Load Index (NASA-TLX). We also developed a competency assessment rubric to measure telemedicine competencies. We recorded observations on the ATA while testers used a think-aloud protocol. We then debriefed testers and administered the posttest questionnaires. We summarized quantitative data using descriptive statistics and analyzed qualitative data using theory-based deductive coding with published usability heuristics.

Results: Twelve testers identified 52 usability issues over 2 testing cycles; 15 issues required correction before piloting with learners. Revisions included changes to standardized patient scripts, synthetic patient data, and scoring rubric instructions. Testers assigned a median SEQ score of 6 or higher to 4 of the 6 workflow steps; most struggled with unfamiliar telemedicine tasks or completing a virtual physical examination. SUS scores ranged from 65 to 100. The median SUS score was 77.5, indicating above-average usability. The NASA-TLX subscale scores were highest for mental workload (median 60, IQR 56.3-63.8) compared with other subscales, such as physical or temporal workload (median 15.00, IQR 11.3-18.8 and median 20.00, IQR 11.3-43.8, respectively), suggesting the simulation imposed a high mental demand. The overall SSES score of 43.5 out of 50 suggested faculty testers thought the simulation offered valuable learning opportunities.

Conclusions: Following usability testing, we successfully piloted the simulation with 23 obstetrics and family medicine residents without encountering any implementation issues. Incorporating usability evaluation methods throughout the simulation development life cycle helped identify and mitigate design problems that might otherwise limit simulation fidelity and instructional effectiveness. Our tutorial showcases a modular mixed methods approach that can be adapted to study a range of simulation types and research questions related to workflow, technology, and learning objectives.

(*JMIR Med Educ* 2026;12:e102697) doi: [10.2196/102697](https://doi.org/10.2196/102697)

KEYWORDS

human factors engineering; maternal health services telehealth simulation; medical education; simulation testing; simulation training; telehealth competencies; telemedicine; usability testing; user-centered design

Introduction

Background

Health professional programs frequently use simulation-based training as an instructional strategy to introduce new concepts, enrich learning, and provide trainees with feedback [1,2]. Educational simulations are faculty-designed instructional methods that provide learners with realistic, immersive clinical situations to develop decision-making skills [1]. The simulation may include human actors representing specific roles (eg, standardized patients [SPs] and health care professionals) or technologies that mimic people and the environment (eg, augmented reality [AR], virtual reality [VR], AI, and manikins). Simulations are an important bridge between the classroom and the bedside, providing a safe yet realistic environment for trainees to learn material, test strategies, and develop reasoning skills before real-world practice [3]. Educators typically design simulations around scenarios with a clinical challenge and a normative workflow. The scenarios, however, should be flexible and responsive, allowing learners to explore, adapt, and improvise [2,4]. It is also critical that simulations include context-sensitive feedback or guided reflection informed by educational theory and learning objectives [2].

To meet instructional objectives, experts and professional societies have published best-practice recommendations for simulation development [2,5-7]. These recommendations are largely conceptual and lack the granularity needed to guide design, testing, and implementation. Conceptual approaches to design alone do not ensure simulation quality [8,9]. It can be challenging to preserve realism and consistency given the many variables affecting case fidelity and the achievement of learning objectives [4]. Simulationists must ensure the accuracy of the scenario, environment, and digital or practical assets, such as medical equipment and synthetic patient data (ie, ecological validity) [9]. Educators must evaluate learners in ways that align with learning objectives (ie, content validity) [4,10]. Finally, learners need personalized feedback addressing the ability under assessment (ie, construct validity) [2,10]. Optimizing validity

often requires educators to limit the number and types of discrepancies between the simulation and real-world clinical practice [11,12].

Discrepancies between the simulation and the real world, including design omissions (eg, incomplete scenarios, data, or objectives) or equipment issues, can limit the simulation's usability. These problems can consume cognitive resources that learners would otherwise apply to pedagogical tasks [11]. It therefore behooves developers to rigorously test the simulation from each person's perspective (eg, the learner, the instructor, and SPs) and correct usability flaws before piloting with intended learners [4]. We authored this tutorial to share our methods and demystify the testing process.

Problem Statement

Simulation development guidelines recommend rehearsing cases with representative learners and collecting feedback from content experts, colleagues, technical teams, and participants [2,4,8,9]. Yet simulation guidelines rarely include methods for testing design features or collecting performance data before piloting [2,4,8,13]. Recently, human factors researchers have published reports trying to address this methodological gap by applying usability measurement techniques [8,14,15]. The ISO defines usability as the “extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use” [16]. Utility is what a system allows users to do, whereas usability is how easily and efficiently users can access and operate these functionalities. Together, these qualities contribute to the overall user experience (ie, “users’ perceptions and responses resulting from the use and/or anticipated use of a system, product or service”).

Usability testing is an evaluation method in which users interact with a product under realistic conditions to determine how effectively, efficiently, and satisfactorily a set of tasks can be completed [17]. Da Silva and Dubrowski [8], for example, used mixed methods to improve 2 simulations over 5 iterations, demonstrating how usability testing findings can inform

revisions. Still, their approach required large numbers of learners and relied, in part, on learner self-reports—an often unreliable proxy for educational effectiveness [18]. We outline here a pragmatic strategy and toolkit to address these limitations.

Our team at the University of Oklahoma designed a series of simulations to teach telemedicine competencies to residents providing online maternal health care. We adapted published usability methods to create a simulation evaluation protocol. Our protocol had 2 aims: to identify technical and content issues requiring immediate correction and to gauge the completeness and difficulty of the clinical tasks. We focused on the following three usability issue types: (1) technology failures that could disrupt workflow; (2) missing case details, patient-script

language, or synthetic clinical data crucial for medical reasoning; and (3) faculty scoring guide instructions that might be ambiguous or difficult to apply.

Tutorial Aims and Goals

Our high-level aim with this tutorial is to provide medical educators, human factors experts, and simulationists with tools and instructions for usability testing educational simulations. Our learning objectives are to (1) explain how we adapted usability methods for this context, (2) share our data collection instruments with readers to facilitate replication ([Multimedia Appendices 1-8](#)), and (3) demonstrate our methods with a use case to illustrate how we applied our results ([Textbox 1](#)).

Textbox 1. Tutorial learning objectives.

After reading this tutorial, readers should be able to do the following:

- Describe how to apply usability testing frameworks when designing instructional medical simulations.
- Adapt published usability testing methods and instruments to test educational simulations.
- Explain how to interpret usability findings to improve simulation consistency and teaching effectiveness.

While the objectives for usability testing are straightforward, we recommend articulating a priori research questions to guide protocol design and measure selection. Readers replicating our methods will likely formulate research questions relevant to their simulation. For our study, we had the following 3 questions:

- What failure modes interfere with the function of the simulation?
- Can faculty evaluators use the scoring guide to measure learner performance?
- Is the simulation difficulty appropriate for the intended learners?

Methods

Theoretical Framework: Usability and User Experience

Usability and user experience are not intrinsic properties of a product or a simulation; instead, they are emergent properties of the work system, arising from the interaction of multiple elements: technology, learners, instructors, tasks, environment, and organizational context [12,19]. This holistic perspective emphasizes that evaluating usability requires examining the entire work system, not the interface or equipment in isolation. Consequently, usability evaluation of simulations should aim to replicate realistic conditions regarding participants, task complexity, environmental constraints, and instructor support. Regarding educational simulations, this means evaluating not only the tools and technology but also faculty roles, faculty-learner interactions, and scenario details [20].

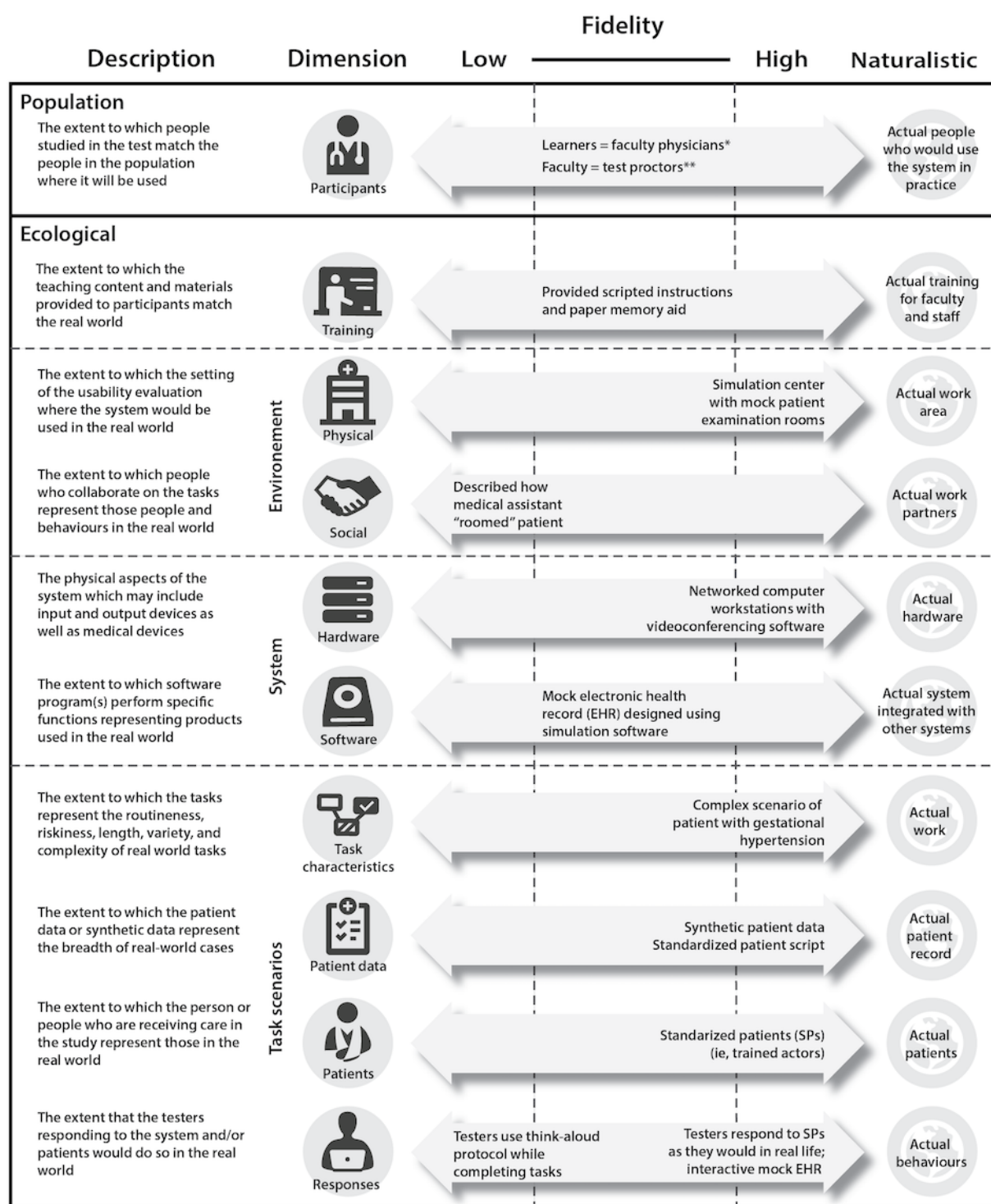
It is important to clarify how simulation usability can influence learners' performance. Human-computer interaction shows that

poor user experience increases the likelihood of errors [21-23]. Usability problems add cognitive load, forcing users to dedicate attention, working memory, and problem-solving capacity to compensate for suboptimal design. As a result, learners confronted with unintentional simulation design problems use cognitive resources to manage these problems rather than engaging with the learning objectives. Usability issues can reduce participant engagement, hinder learning, and diminish the perceived credibility of the program [12]. Hence, while simulation guidelines recommend piloting with learners to uncover problems, we believe usability issues should be addressed before piloting or implementation [2,6].

We needed to solve several unique problems in educational simulations that are atypical for usability testing of medical devices or software. First, we needed to calibrate scenario difficulty to participants' experience level (eg, residents) and expertise (eg, obstetrics and gynecology [OB/GYN], family medicine). Second, faculty observers needed to evaluate learners for specific telehealth competencies to identify skill gaps. Third, we needed to include built-in challenges for the learner. Thus, we needed faculty participants to serve as observers to provide feedback on the scoring guide while gathering tester feedback on the simulation.

The theoretical framework informing our current approach is based on a model of external validity we created for usability evaluations of health information technology ([Figure 1](#)) [20]. External validity comprises (1) population validity (ie, the extent to which the sample represents the population) and (2) ecological validity (ie, the extent to which the setting matches the real world). While both types of validity are important, the fidelity level for each dimension depends on the evaluation objectives [12].

Figure 1. External validity framework for the application of usability testing in health. We adapted the figure to illustrate fidelity levels on a continuum for each dimension of our educational simulation.



*Faculty physician-instructors served as testers to assess case accuracy and difficulty level

** Faculty physician-informaticians served as proctors or double-coded all collected data

When selecting the degree of ecological validity, we needed to consider learner attributes, presimulation didactics, the clinical environment (ie, physical and social), the technology (ie, hardware and software), and the scenario (ie, task characteristics, patient data, patients, and responses). Figure 1 [20] shows how we mapped model constructs to our use case. Poor usability arising from interactions between dimensions could manifest as unclear instructions, confusing interface layouts, incomplete

patient scripts, mismatched equipment, or poorly structured task sequences. We chose to test our simulation with faculty rather than with learners. This was an intentional choice to assess case accuracy and to determine whether faculty believed the tasks were appropriately calibrated for resident learners.

Use Case for Testing: A Telehealth Simulation for Residents Providing Maternal Care

During the COVID-19 pandemic, organizations embraced virtual care to enact social distancing precautions without limiting access [24-29]. This expansion highlighted educational gaps; practitioners were uncomfortable providing online care, and few training programs offered telemedicine instruction [30-35]. In response to this emerging need, the Association of American Medical Colleges (AAMC) published a white paper outlining a framework for telemedicine training in the United States [36]. The AAMC identified 20 competencies across six domains: (1) patient safety, (2) access and equity, (3) communication, (4) data collection, (5) technology, and (6) ethical and legal requirements. Since publication, medical schools and residencies have endeavored to teach this material using readings, didactics, and simulations [35,37-39].

At the University of Oklahoma School of Community Medicine (OUSCM), the Department of Medical Informatics partnered with the Department of Obstetrics and Gynecology to develop a telemedicine training program for maternal care [40]. Designed for OB/GYN and family and community medicine (FCM) residents, our simulations combined telemedicine instruction with vignettes managing maternal care conditions [41-43]. By training future physicians to provide online care to expectant mothers, we hoped to close a supply-demand mismatch contributing to local maternal care “deserts” [40,44-49].

Our educational program created the need—and the opportunity—to conduct usability tests of simulations before piloting with residents. The technology-forward nature of telemedicine inspired our team to adapt usability methods typically applied to health information technology. Although we have published preliminary reports describing individual tools, this is the first tutorial comprehensively describing our usability testing methods, findings, and downstream piloting experiences [41-43]. For illustrative purposes, we demonstrate our methods using a simulation of a fictional patient presenting for a telemedicine encounter to manage gestational hypertension.

Study Setting and Research Team

Our research team conducted all testing at our university’s simulation center in Tulsa, Oklahoma, United States. The team

included 2 obstetricians (MR, KB) to provide subject-matter expertise, 4 informaticians (BJL, HM, RM, and JH) and 2 data analysts to design the test protocol (MA and SM), a proctor to supervise each test and code the data (MA or CP), and at least 1 additional observer (typically another member of the research team) to double-code observations. The proctor and observers played the role of faculty during simulation testing. The team also included an SP and a simulationist (KB). The simulationist is an expert in simulation design, SP education, and script writing. The simulation center was equipped with mock clinic rooms, remote monitoring stations, simulation software, and computers with videoconferencing software. Because this was a telemedicine simulation, our SPs connected to the simulation center remotely via videoconference.

To recreate our described approach, we recommend the following resources: a dedicated space with internet access, a trained team member to serve as the SP, videoconferencing software with recording capabilities, and a proctor for the session. We also recommend recruiting team members to observe and double-code findings. We found it helpful to have trained faculty on-site to debrief testers. While the examination rooms in our center increased the ecological validity of our testing, they were not required.

Data Collection Instruments

Overview

In this section, we describe each data collection instrument and our rationale for its inclusion (Table 1). Multimedia Appendices 1-8 provide instrument examples; Multimedia Appendices 9 and 10 are examples tools used for our specific simulation. We sought validated questionnaires to gather cross-sectional data on multiple usability dimensions. In our case, a “minimally viable product” for an educational simulation needed to satisfy three requirements: (1) the learner must be able to complete the scenario without interruption, (2) the learner must be able to review synthetic patient data needed for a clinical assessment, and (3) the faculty must be able to calculate competency scores for each learner.

Table 1. List of testing instruments used in our usability assessment protocol. We adapted usability dimensions and definitions from the ISO [16], Quesenberry [50], and Davis [51]. All instruments are available in [Multimedia Appendices 1-9](#).

Instrument	Data collected	Measure type	Usability dimension	Usability dimension definition
Demographic screener (Multimedia Appendix 1)	Tester demographics, technology literacy, and attitudes and perceptions	Qualitative, binary (yes/no), and ordinal (Likert-type)	N/A ^a	N/A
ATA ^b (Multimedia Appendix 2)	Simulation step completion rates	Ordinal (trichotomous)	Effective (learner)	How completely and accurately learners can complete a task to reach their goals [50]
ATA (Multimedia Appendices 2, 7, and 8)	Usability issues	Qualitative	Error tolerance	How well a design prevents errors or helps users recover from mistakes [50]
SEQ ^c (Multimedia Appendix 3)	Task difficulty	Ordinal (Likert-type)	Easy to learn	The degree to which a person believes using a system would be free of effort [50]
Telemedicine competency rubric (Multimedia Appendix 9)	Clinical task completion rates, and competency scores	Task count and competency score	Effective (faculty)	How completely and accurately faculty could score the telemedicine rubric
Adapted SUS ^d (Multimedia Appendix 4)	User perceptions	Ordinal (Likert-type)	Ease of use	How intuitive the product is, ensuring learners can understand expectations with little to no additional instruction [51,52]
SSES ^e (Multimedia Appendix 4)	Learner perceptions	Ordinal (Likert-type)	Satisfaction	How relevant, valuable, and supportive the simulation was for acquiring skills and knowledge [53]
NASA-TLX ^f (Multimedia Appendix 5)	Cognitive load	Ordinal	Efficiency	The resources expended, including time, cognitive load, and effort in relation to the accuracy and completeness of goals [54]
Debrief interview guide (Multimedia Appendix 6)	Usability issues	Qualitative	Satisfaction	How pleasant, satisfying, and aesthetically pleasing the product is [16]
Debrief interview guide (Multimedia Appendix 6)	Usability issues	Qualitative	Fidelity	How completely and accurately the tasks align with real-world scenarios [11,55]

^aN/A: not applicable.

^bATA: Agile Task Analysis.

^cSEQ: Single Ease Question.

^dSUS: System Usability Scale.

^eSSES: Satisfaction With Simulation Experience Scale.

^fNASA-TLX: NASA (National Aeronautics and Space Administration) Task Load Index.

Demographics Questionnaire

When considering population validity, it is helpful to create a user persona and recruit testers matching the persona's characteristics. Learners are the target users for educational simulations and can provide useful feedback. While it may seem counterintuitive, it has been our experience that faculty can also provide useful feedback. They have the domain expertise to critique simulation fidelity, task complexity, the need for additional scaffolding, and alignment with learning objectives [4]. Hence, we recommend recruiting faculty testers first, followed by learners as time and resources permit. For all recruited participants, we gathered descriptive demographic data to share with stakeholders (eg, educational leadership, executive sponsors, collaborative partners, funders) when presenting findings. We therefore used a demographics questionnaire with 3 sections: (1) tester demographics, (2) a

technology literacy assessment, and (3) knowledge and attitude items about telemedicine ([Multimedia Appendix 1](#)). We confirmed that all testers provided maternal medical care. We asked whether they regularly used a computer, mobile technology, or videoconferencing software. Finally, we included a short questionnaire covering prior experiences with telemedicine. For the latter section, we based the content domains on research exploring the intersection between clinician attributes and successful telemedicine encounters [56].

Agile Task Analysis

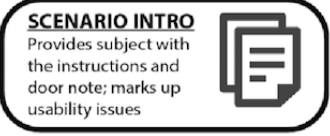
Agile Task Analysis (ATA) is a hybrid method that combines hierarchical task analysis (HTA) with qualitative observation during simulation testing ([Multimedia Appendix 2](#)) [57]. In a standard HTA, the researcher breaks complex activities into tasks and goals to understand how work is performed and what tools and information are required for each task [58].

Observations may include task lists, narrative details, workflows, and examples of tools, paper forms, or technology [59]. Borrowing from HTA, the ATA uses this framework to organize scenarios for simulation testing [57]. It also includes a data collection form that follows the scenario's normative workflow, allowing observers to record observations concurrently. Data can be quantitative or qualitative, including task completion rates, observed errors, usability issues, or user observations, such as quotes or design recommendations.

Researchers can modify the ATA based on the type of product or research questions. For example, the ATA can include screenshots of technology interfaces, informational handouts,

or blueprints of the physical environment [57]. For this study, we listed six concept-level steps in our telemedicine workflow: (1) read learner instructions, (2) start a telemedicine session, (3) complete standard telemedicine tasks, (4) collect a patient history, (5) review examination and objective data, and (6) develop an assessment and plan (Figure 2 [57]). For each step, the research team tracked simulation performance using an ordinal trichotomous scale (ie, pass, marginal, and fail). We coded a step as “pass” if the tester completed it without assistance, “marginal” if the tester required proctor assistance, or “fail” if the tester encountered a critical fault in the simulation and needed to bypass the step entirely.

Figure 2. Excerpt from the Agile Task Analysis (ATA) form we used to evaluate our telehealth maternal care education simulation.

User Experience	Task/Step (Normative workflow)	Rating 2=Pass 1=Marginal 0=Fail	Comments/Usability Problems/Subject Quotes
1. Introduction to the scenario  SCENARIO INTRO Provides subject with the instructions and door note; marks up usability issues	a. Read learner instruction ✓ b. Read case description ✓	Pass (2) Marginal (1) Fail (0) Pass (2) Marginal (1) Fail (0)	Collect comments on door note and instructions SEE PATIENT "DOOR" NOTE FOR MARK-UPS TO IMPROVE CLARITY
2. Begin telemedicine session  SIMULATION Introductions and agenda setting; completes SEQ	a. Login to session ✓ b. See and hear patient ✓	Pass (2) Marginal (1) Fail (0) Pass (2) Marginal (1) Fail (0)	Zoom session should be running. SP should be present. NO PROBLEMS
3. Personal introductions and cross check of equipment  SIMULATION Tester completes technology cross-check	a. Learner can begin the interview with SP ✓ b. Learner completed an AV cross-check ✓	Pass (2) Marginal (1) Fail (0) Pass (2) Marginal (1) Fail (0)	Reminder: simulation may function as designed even if learners skip tasks. TESTER STARTED SESSION BUT SKIPPED TELEMED COMPETENCIES! (... MISSED ADDRESS)

Telemedicine Competency Rubric

Educational simulations are often designed to challenge learners or to identify knowledge or skill gaps. A simulation step can work as designed even if the tester neglects a task or performs it incorrectly. We did not expect testers to complete every task. To differentiate simulation steps from telemedicine tasks, we included a telemedicine competency rubric (ie, faculty scoring

guide) in the ATA to record testers' performance on telemedicine tasks during the simulation (Multimedia Appendix 2) [41].

Our telemedicine competency rubric was based on the AAMC white paper and operationalized measurement for 8 of the 20 listed telehealth competencies [36]. We selected 8 competencies for our simulation: 2 focusing on patient safety, 2 on

communication skills, 1 on data collection, 2 on telehealth technologies, and 1 on ethical requirements. The AAMC white paper includes definitions for each competency but not context-specific details for clinical care or explicit measurement strategies. We were forced to create a novel rubric that organized each competency into short sets of related, contextually embedded, and directly observable tasks ([Multimedia Appendices 2 and 9](#)) [60]. We summed correctly performed tasks to calculate composite competency scores. We operationalized competency into 3 levels: entrustable, approaching entrustment, and not entrustable [60,61]. A learner was “entrustable” if faculty believed the learner could be entrusted to complete all activities within a content domain without requiring direct supervision. Faculty marked a learner as “approaching entrustment” if they believed the learner needed some supervision to complete related clinical tasks. The tester needed to complete at least half of the tasks within a domain to be approaching entrustment. A learner was “not entrustable” if faculty believed the learner must be directly supervised throughout the activity. The proctor marked a tester “not entrustable” if the tester completed fewer than 2 tasks within a telemedicine domain. As a first step toward establishing interrater reliability, we required multiple research team members to score each tester independently during data collection and then resolved discrepancies through discussion and consensus (details are provided in the Data Recording and Adjudication section).

Single Ease Question

The Single Ease Question (SEQ) is a 7-point Likert-type question administered immediately after a task, asking the tester to rate the task’s difficulty [23,62]. Tasks that score higher than 6 are considered “very easy” [63]. The SEQ offers insights that posttest questionnaires do not by dissociating individual steps from the overall user experience [64,65]. We included 1 SEQ after each of the 6 workflow tasks and 1 at the conclusion of the simulation ([Multimedia Appendix 3](#)).

Adapted System Usability Scale

The System Usability Scale (SUS) is a 10-item, 5-point Likert-type scale assessing users’ self-reported ease of use with a product [23,52,65,66]. Researchers convert respondents’ answers to a 100-point scale. Based on the distribution of scores across industries, an average score is 68 [52]. We adapted the SUS language to evaluate testers’ perceptions of a simulation rather than a technology ([Multimedia Appendix 4](#)). Prior studies of the SUS show that substituting the word “system” with a contextually appropriate product name or referent does not degrade the psychometric properties of the instrument [67].

Satisfaction With Simulation Experience Scale

The Satisfaction With Simulation Experience Scale (SSES) is an 18-item instrument used to evaluate learner satisfaction with educational simulations [53]. It measures learners’ engagement and perceptions of simulation effectiveness [53,68]. Although not a traditional usability measure, we used the SSES to capture

faculty and resident perceptions of the simulation’s educational value and treated it as a proxy for satisfaction. The SSES complements other usability measures (eg, SEQ, SUS, and think-aloud) but does not replace them. Items are grouped into 3 subscales: debriefing and reflection, clinical reasoning, and clinical learning. For this study, we adapted questions from the SSES clinical reasoning and clinical learning subscales, modifying the language for a telemedicine simulation ([Multimedia Appendix 4](#)).

NASA Task Load Index

To measure perceived simulation complexity, we administered the NASA (National Aeronautics and Space Administration) Task Load Index (NASA-TLX) after the simulation ([Multimedia Appendix 5](#)) [54]. The NASA-TLX was developed by NASA to measure perceived workload and cognitive overhead associated with a set of related tasks [69]. There are 6 subscales: mental demands, physical demands, temporal demands, perceived performance, effort, and frustration [54]. Each dimension is scored on a 100-point scale. Although there are no formal cut points, studies suggest scores ≤ 30 indicate a low workload, scores between 31 and 50 indicate a moderate workload, and scores > 50 indicate a high workload [54,69-71]. The original NASA-TLX requires participants to assign a weight to each subscale. Researchers often omit the weighting step and use raw scores [69]. We used unweighted scores in our analysis.

Debrief Semistructured Interview

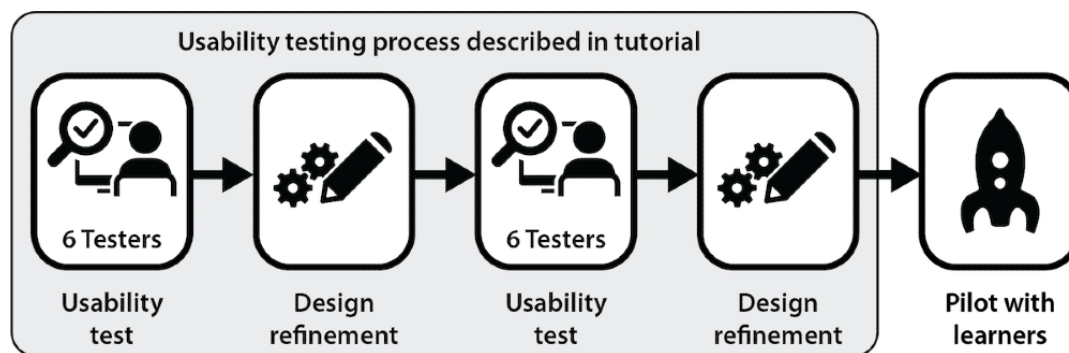
We used a semistructured debriefing guide to gather testers’ final reflections ([Multimedia Appendix 6](#)). We designed our guide based on the Rapid Usability Evaluation method by Russ et al [72] and *Rocket Surgery Made Easy* [73] by Steve Krug. In both cases, the authors recommend asking probing questions during the observation session immediately following tasks and limiting the debrief to 1 or 2 open-ended follow-up questions. The researcher can use these questions to clarify usability issues or invite the participant to suggest changes for improving the design. We used questions and probes to explore the simulation’s difficulty, the root causes of usability issues, and opportunities to improve the learning experience. We recorded responses in our usability issue log.

Testing Protocol

Overview

This section describes our testing protocol ([Figure 3](#)). While our general approach can be adapted to a range of educational simulations, the specifics of each simulation will dictate the number of testing cycles. We organized testing into sprints, gathering feedback from 6 testers before iterating on the design. We conducted 2 rounds of simulation testing and revisions (ie, sprints) to refine the simulation. We used new testers for each testing cycle to avoid data contamination. We conducted 2 refinement sessions, once at the midpoint of testing and once before piloting with target learners. We avoided making refinements to the simulation between refinement sessions.

Figure 3. Testing and refinement strategy. We used new testers for each round. The research team, which included content experts, served as faculty observers.



Recruitment

We used a purposive recruitment strategy, targeting faculty and residents to test the simulation and provide feedback on the scenario's accuracy and fidelity. Faculty and residents were eligible for inclusion if they (1) held an appointment at our institution, (2) had completed or were participating in a residency in obstetrics or family medicine, and (3) maintained an active obstetrics practice. Residents in obstetrics and family medicine are required to see obstetrics patients in a continuity clinic during training; they practice under the license of their supervising attending. We contacted every eligible practitioner at our institution and recruited the first 6 volunteers for each testing round, aiming to balance the number of obstetricians and family practitioners.

Tester Orientation

A member of the research team provided each tester with an overview of the study, including the curriculum's purpose and the high-level goals for the simulation (Figure 4). We secured

verbal and written consent from each tester during the orientation to observe and record testing and tester insights. The demographic screening questionnaire included consent language and an explanation of how the data would be used. We entered data, excluding unique identifiers, into a spreadsheet for subsequent analysis. We did not include a telemedicine didactic before testing. We did not disclose scenario details, workflow expectations, or the competencies being evaluated. A proctor sat with the tester in a mock examination room with a computer workstation, microphones, and closed-circuit cameras (Figure 5). A proctor without clinical expertise can be advantageous as a neutral observer. Before starting the simulation, the proctor furnished the tester with the screening questionnaire. We adapted the Rapid Usability Evaluation method by Russ et al [72] and the methods outlined in *Rocket Surgery Made Easy* by Steve Krug [73]. The proctor read aloud a written script explaining the purpose of testing and the think-aloud technique (ie, testers verbalize their thoughts while interacting with the system and completing tasks) [74,75].

Figure 4. Schematic for the usability testing protocol. We conducted 2 rounds of formative usability testing before piloting with target learners. NASA-TLX: NASA Task Load Index; SEQ: Single Ease Question; SP: standardized patient; SSES: Satisfaction With Simulation Experience Scale; SUS: System Usability Scale.

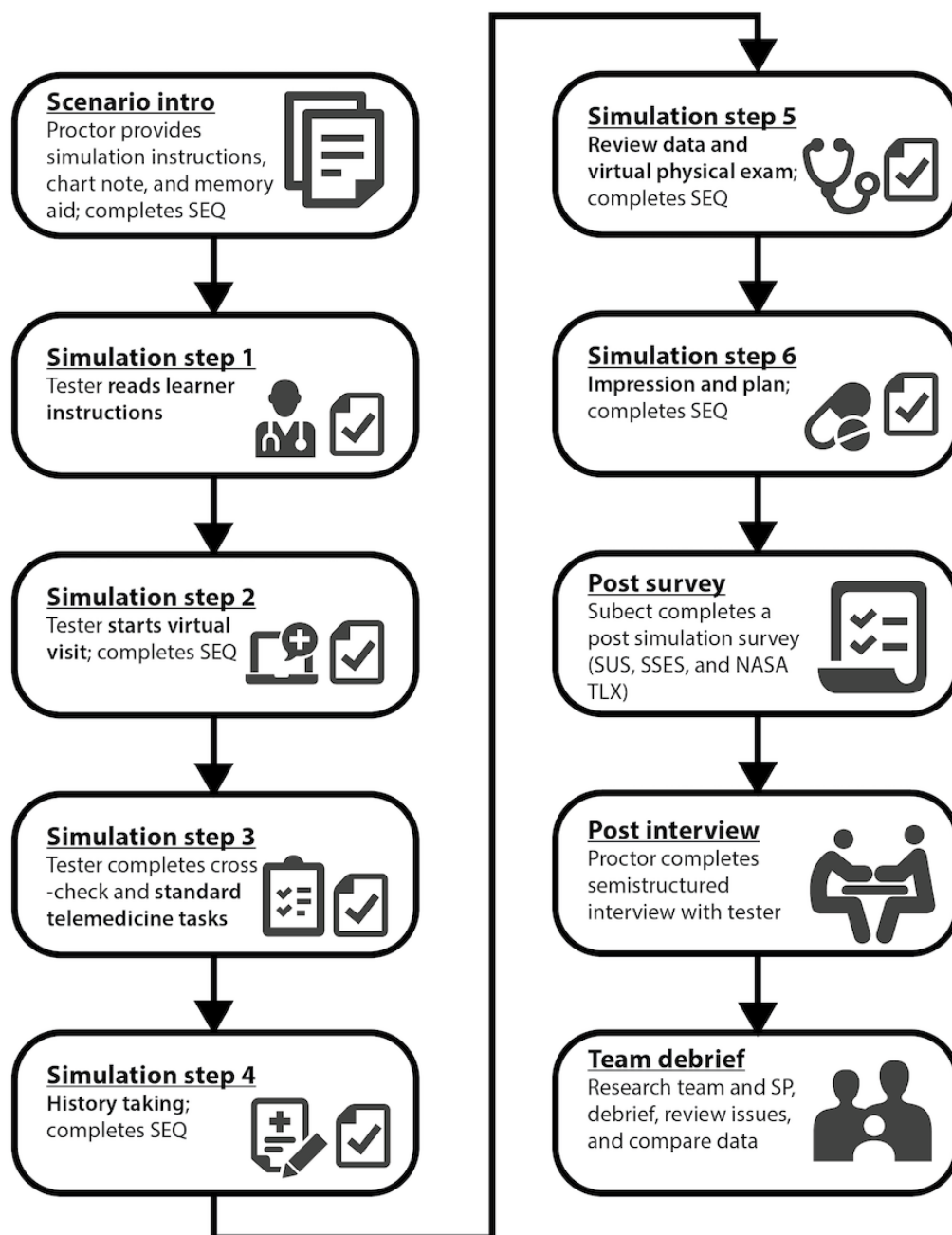
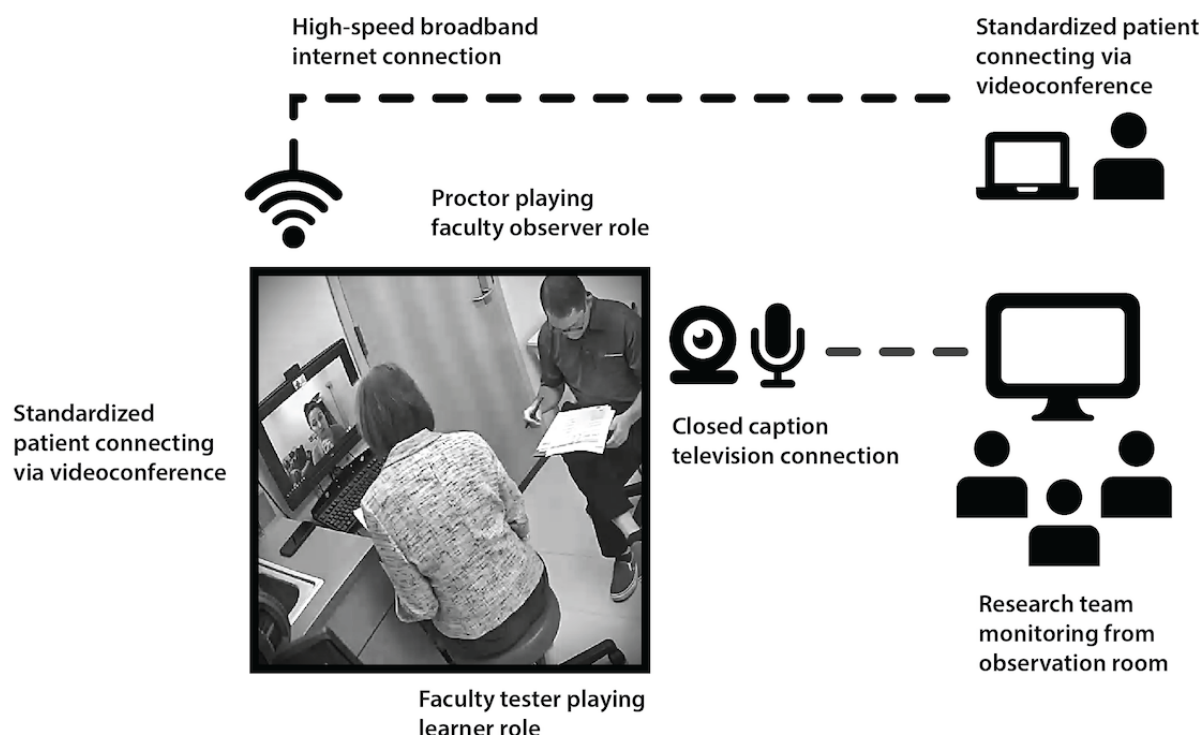


Figure 5. Setting and configuration for usability testing.

Simulation Testing

During the simulation, the tester used the computer workstation with videoconferencing software (Zoom; Zoom Communications) to complete a mock encounter with our SP [76]. We furnished the tester with a summary of the case: a pregnant woman living in a rural community and presenting at 20 weeks' gestation for a routine follow-up appointment. The patient reported recording new high blood pressure values using a home digital blood pressure cuff.

In the first round, the proctor provided the tester with written case instructions, a mock clinical note with synthetic patient data, and a printed memory aid developed by the team to help with telemedicine tasks [77]. In the second round, rather than providing a mock clinical note, the tester reviewed synthetic patient data using a mock electronic health record (EHR) we designed with LearningSpace simulation software (Elevate Healthcare) [78]. We continued to give the tester written case instructions and our memory aid. The proctor did not provide technical instructions, additional medical information, or clinical feedback to the tester during the simulation. The other members of the research team watched the simulation on closed-circuit television and recorded observations simultaneously.

We expected the tester to establish rapport with the SP, troubleshoot technology problems, gather a history, review synthetic patient data, provide a clinical assessment, and communicate a treatment plan. The learner needed to gather enough clinical data to rule out preeclampsia, an obstetric emergency. During the simulation, the proctor recorded all observations on the ATA paper form. The proctor and observers recorded (1) whether each simulation step worked, (2) whether the tester had the opportunity to perform each competency-related task, (3) whether the tester performed the

task successfully, and (4) whether the tester encountered any usability issues. Initially, the proctor asked the SEQ after each set of related tasks. However, we found this disrupted the simulation workflow and added cognitive overhead. Hence, the proctor asked all SEQs after the simulation.

Postsession Questionnaires and Semistructured Interviews

After the simulation, the proctor administered our postsession questionnaires (ie, SUS, SSES, and NASA-TLX) and conducted a semistructured interview (Multimedia Appendices 4-6). The interview provided an opportunity to explore areas of confusion, the root causes of usability issues, potential modifications to the simulation, and unmet educational needs.

Data Recording and Adjudication

For the rubric, the research team compared scores for each task and resolved discrepancies through discussion to reach consensus. When we identified vague or potentially confusing language, we revised the task description during the design and refinement phases. We did not calculate agreement statistics during testing. The Limitations section provides a more complete discussion of our rationale and next steps. We compiled these results into a "master" document. We then reviewed all qualitative data, including usability issues, observations, tester quotes, and interview responses. We summarized the qualitative findings in an issue log (Multimedia Appendices 7 and 8) for content analysis, coding, prioritization, solution ideation, and potential action.

Simulation Piloting

After we completed all usability testing and simulation refinements, we piloted the entire workshop (ie, a telemedicine didactic session, telemedicine simulation with the competency

rubric [[Multimedia Appendix 9](#)], and postsession questionnaire) with the residents. A detailed protocol for our workshop and simulation pilot falls outside the scope of this tutorial. We

included a brief vignette of the pilot herewith ([Textbox 2](#)) to illustrate the downstream value of usability testing.

Textbox 2. Workshop and simulation vignette.

A complete description of workshop piloting with residents is beyond the scope of this report. We intend to describe the implementation methods and residents' scores in a companion manuscript. In this section, we include a short vignette of our piloting experience.

We conducted 2 separate pilots on different afternoons, 1 with 6 obstetrics and gynecology residents and 1 with 17 family and community medicine residents. All residents received a brief overview of the case, expectations for the simulated encounter, and a paper copy of the telemedicine memory aid. Residents used computer workstations in our testing center to interview a standardized patient and had online access to the mock electronic health record with synthetic patient data.

All participants completed the simulation without experiencing any unanticipated technical problems. Faculty were comfortable using the supplied evaluation rubric and had no difficulty providing individualized learner feedback after the simulation. Resident feedback on the simulation was generally favorable; most indicated the simulation provided a clear, safe, and constructive learning experience.

Data Analysis

Quantitative and Statistical Analysis

We calculated descriptive statistics for quantitative data, including demographic characteristics, simulation step completion rates, competency scores, SEQ scores, SUS scores, SSES scores, and raw NASA-TLX scores. Given the small sample size and the nonparametric nature of the data, we calculated raw counts, medians, and IQRs. Given the formative nature of the analysis, we did not conduct an a priori power analysis or calculate comparative statistics.

Qualitative Analysis

While a formal qualitative analysis is not always critical for pragmatic usability testing, we found it helpful for clarifying the root causes of issues, brainstorming potential solutions, and prioritizing fixes. Using coding methods described by Saldana, we conducted a content analysis of qualitative data gathered during the simulation and tester debriefing interviews [79]. We used simultaneous coding to assign a categorical usability code, a valence (ie, "positive" or "negative"), and a priority for action. For categorical codes, we used a deductive, theory-based coding strategy based on usability heuristics published by Zhang et al

[80] ([Multimedia Appendix 7](#)) [21,81]. At least 2 members of the research team independently coded each finding. We then compared codes for concordance and resolved discrepancies through discussion. We assigned a provisional code with emergent supplementation to any findings that could not be classified according to the usability heuristics published by Zhang et al [80]. All codes and findings were recorded in a usability issue log (structural codes are provided in [Multimedia Appendix 7](#); a sample log is provided in [Multimedia Appendix 8](#)).

Ethical Considerations

The testing described in this tutorial was reviewed by the University of Oklahoma Institutional Review Board and classified as a quality improvement initiative associated with normal educational activities. Nevertheless, we collected verbal and written consent from each tester to document the testing.

Results

Tester Demographics

We recruited 12 clinicians for usability testing (first round: 4 OB/GYNs and 2 FCMs; second round: 2 OB/GYNs and 4 FCMs). [Table 2](#) provides demographic details.

Table 2. Tester demographic information and technology literacy.

Characteristic	Value	Mean (SD)	Median (IQR)
Demographic			
Age (years), range	25-60	38.9 (12.1)	37.0 (28.5-45.0)
Work experience (years), range	1-30	8.8 (11.1)	3.5 (1.0-10.5)
Specialty, n	OB/GYN ^a =6; FCM ^b =6	— ^c	—
Technology literacy, n			
Use computer once a week	Yes=12; no=0	—	—
Use portable technology	Yes=12; no=0	—	—
Use videoconferencing	Yes=12; no=0	—	—
Attitudes and perceptions, n			
Previously observed a telehealth encounter	Yes=11; no=1	—	—
Previously provided telehealth care	Yes=10; no=2	—	—
Previously received telehealth care	Yes=5; no=7	—	—
Received telemedicine instruction	Yes=4; no=8	—	—
Good understanding telemedicine practice	—	3.9 (0.70)	4.0 (4.0-4.0)
Familiar with types of technology	—	3.8 (0.8)	4.0 (3.8-4.0)
Think telemedicine is a good option	—	3.7 (1.3)	4.0 (4.0-4.0)
Will likely use telemedicine in practice	—	3.9 (0.8)	4.0 (4.0-4.0)

^aOB/GYN: obstetrics and gynecology.^bFCM: family and community medicine.^cNot applicable.

Overall Results

Table 3 summarizes all usability measures, with key findings highlighted. We include detailed results for each measure in the following sections. The ATA identified 4 problematic steps in round 1 that we resolved by round 2. Across both rounds, we identified and addressed 15 critical usability issues. The virtual physical examination received the lowest SEQ ratings. During

round 2, testers reported a median SUS score of 77.5 (IQR 73.1-80.0) the SUS and a median SSES score of 42 (IQR 41.0-46.75) out of a possible 50. The median NASA-TLX mental demand score in round 2 was 60.00 (IQR 56.3-63.8); a score >50 indicates a high workload. Proctors could not score all testers with the rubric during the first round. After modifying the rubric language during the first refinement phase, proctors successfully scored every tester.

Table 3. Summary table of quantitative results across instruments.

Usability dimension, instrument, and measured item	Frequency, n/N (%)	Median (IQR)	Maximum possible value
Effectiveness			
Agile Task Analysis			
Total number steps completed successfully	165/180 (92)	— ^a	—
Total number steps partially completed	11/180 (6)	—	—
Total number of failed steps	4/180 (2)	—	—
Error tolerance			
Agile Task Analysis			
Number of usability issues	52	—	—
Number of critical usability issues	15/52 (29)	—	—
Easy to learn			
Single Ease Question			
Reviewing materials	—	6.00 (6.00-6.25)	7.00
Starting encounter	—	6.00 (4.75-7.00)	7.00
Standardized telemedicine tasks	—	5.50 (5.00-6.25)	7.00
History collection	—	6.50 (5.00-7.00)	7.00
Virtual physical examination	—	4.50 (2.75-6.00)	7.00
Diagnosis and plan	—	6.00 (5.00-6.50)	7.00
Overall encounter	—	5.00 (5.00-6.00)	7.00
Ease of use			
Adapted System Usability Scale			
Total score round 1	—	73.80 (72.63-80.50)	100.00
Total score round 2	—	77.50 (73.13-80.00)	100.00
Total score overall	—	75.00 (71.88-81.25)	100.00
Satisfaction			
Satisfaction With Simulation Experience Scale			
Total score round 1	—	44.50 (39.5-45.0)	50.00
Total score round 2	—	42.00 (41.00-46.75)	50.00
Total score overall	—	43.50 (40.5-45.25)	50.00
Workload			
NASA^b Task Load Index			
Physical demand round 1	—	7.50 (5.00-17.50)	100.00
Temporal demand round 1	—	12.50 (6.25-22.50)	100.00
Frustration level round 1	—	22.50 (10.00-42.50)	100.00
Perceived performance round 1	—	32.50 (26.25-35.00)	100.00
Mental demand round 1	—	37.50 (21.25-57.50)	100.00
Effort (mental and physical) round 1	—	42.50 (27.50-57.50)	100.00
Physical demand round 2	—	15.00 (11.30-18.80)	100.00
Temporal demand round 2	—	20.00 (11.30-43.80)	100.00
Frustration level round 2	—	25.00 (10.00-43.80)	100.00
Perceived performance round 2	—	32.50 (21.30-43.80)	100.00
Mental demand round 2	—	60.00 (56.30-63.80)	100.00

Usability dimension, instrument, and measured item	Frequency, n/N (%)	Median (IQR)	Maximum possible value
Effort (mental and physical) round 2	—	55.00 (46.30-63.80)	100.00
Mental demand overall	—	57.50 (23.75-61.25)	100.00
Effort (mental and physical) overall	—	50.00 (32.50-60.00)	100.00
Effectiveness			
Competency rubric^c			
Incorporates telehealth	12/12 (100)	—	—
Attends to environment	11/12 (92)	—	—
Identifies and uses equipment	12/12 (100)	—	—
Develops rapport	12/12 (100)	—	—
Obtains and uses history	12/12 (100)	—	—
Troubleshoots technology	12/12 (100)	—	—
Prepares for escalating care	12/12 (100)	—	—
Adheres to professional requirements	10/12 (83)	—	—

^aNot applicable.

^bNASA: National Aeronautics and Space Administration.

^cIndicates the number of learners successfully assessed.

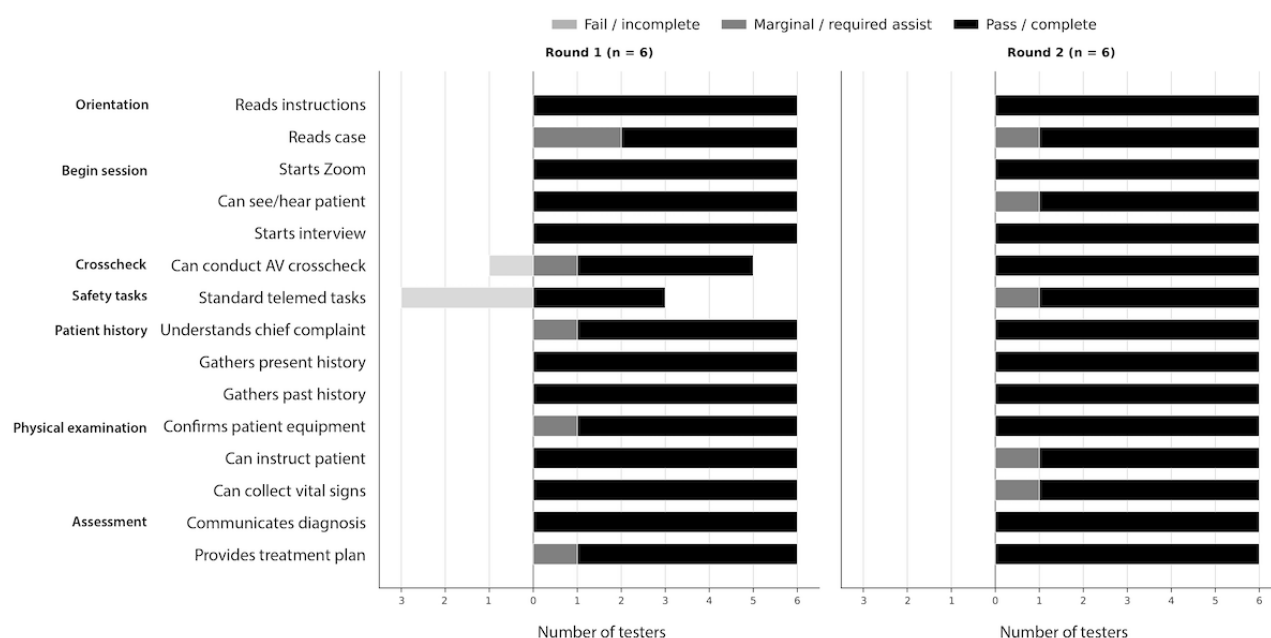
ATA Results

Simulation Step Completion Counts

Figure 6 shows simulation completion rates organized by workflow step. While most testers completed the steps without

difficulty, not every tester completed every step. In the first round, testers skipped standard telemedicine patient safety tasks (eg, confirming the patient's physical location). This issue improved during the second round after we included a mock EHR with synthetic data and visual cues (eg, patient address and phone number fields).

Figure 6. Simulation step completion counts. We show scores for each testing round, separated by one refinement session. AV: audiovisual.



Usability Issues

We logged 52 usability findings across both testing rounds; 15 required corrections before piloting (Tables 3 and 4). We identified issues with the SP script, gaps in synthetic patient

data, deviations in the simulation workflow, and problems with the competency rubric language. New issues emerged after the first refinement session due to updates to materials and assets (eg, the addition of a mock EHR).

Table 4. Excerpt from our usability issue log illustrating a sample of usability issues and remediation strategies organized according to the usability heuristics schema [80].

Test number	Finding	Code ^a	Valence	Priority	Fixed?	Intervention
7	The tester expected to see a flowsheet of blood pressure values in the mock electronic health record	Match	Negative	High	Yes	Added a flowsheet with additional blood pressure values to the LearningSpace software
8	Need to include photographs of physical findings, such as leg edema	Feedback	Negative	High	Yes	Gave the SP ^b printed photographs of physical examination findings sourced from licensed stock photography

^aCode: usability heuristic code.

^bSP: standardized patient.

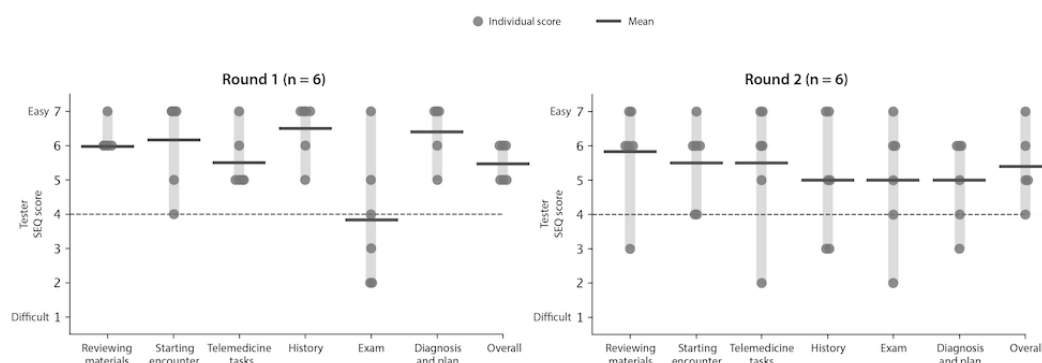
Usability issues tended to fall into the following four heuristic categories: (1) user feedback, (2) match between the system and the real world, (3) error prevention, and (4) system visibility. For feedback, testers asked to see the patient's leg to check for edema. We provided SPs with photographs of ankle edema to show on camera. To align the system with the real world, testers requested a flowsheet of blood pressure values, which we added to our mock EHR. As an example of an error-prevention issue, faculty needed a clear deadline in the task definition to score learners scheduling follow-up laboratory tests. We added "within 1 week" to the rubric. As an example of a visibility issue, faculty

could not tell when learners fabricated information to preserve the illusion of the simulation (eg, naming a fictional laboratory). We therefore added a question to the SP script that the learner would have to research (ie, the location of the laboratory nearest the patient's home).

SEQ Results

Most testers found the simulation instructions and steps easy to understand (Figure 7 [55]). Testers assigned a median SEQ score of 6 or greater to 4 of the 6 components. Testers said completing a virtual physical examination (median 4.50, IQR 2.75-6.00) was the most challenging step.

Figure 7. Single ease question results for both testing rounds. We show scores for each testing round, separated by one refinement session.

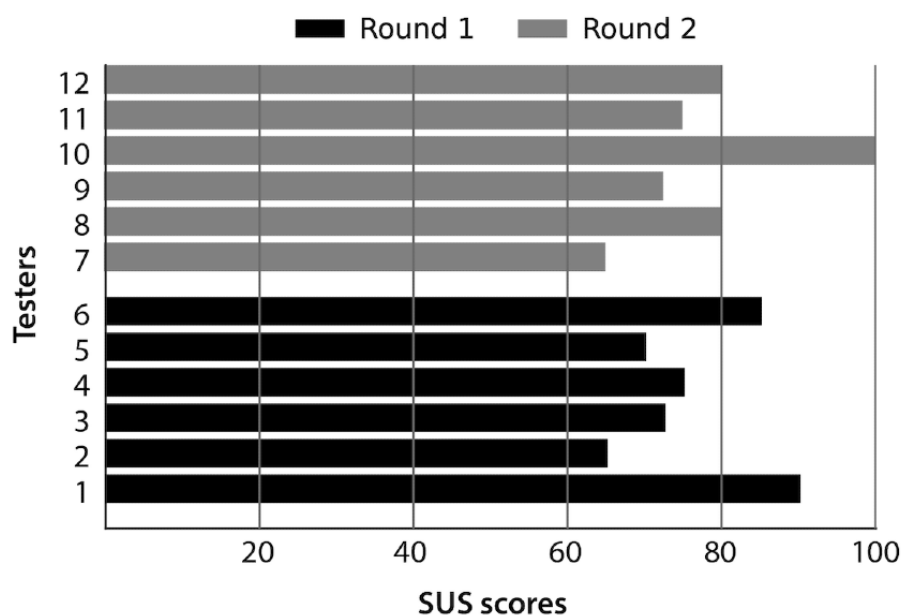


SUS Results

Testers reported some challenges with ease of use and learnability (Figure 8 [52]). SUS scores ranged from 65 to 100.

For round 1, respondents' median score was 73.80 (IQR 70.63-82.50); for round 2, the median score was 77.50 (IQR 73.13-80.00). The inclusion of a mock EHR in round 2 may account for the increase in the median SUS score.

Figure 8. System Usability Scale (SUS) scores separated by testing rounds. We show scores for each testing round, separated by one refinement session.

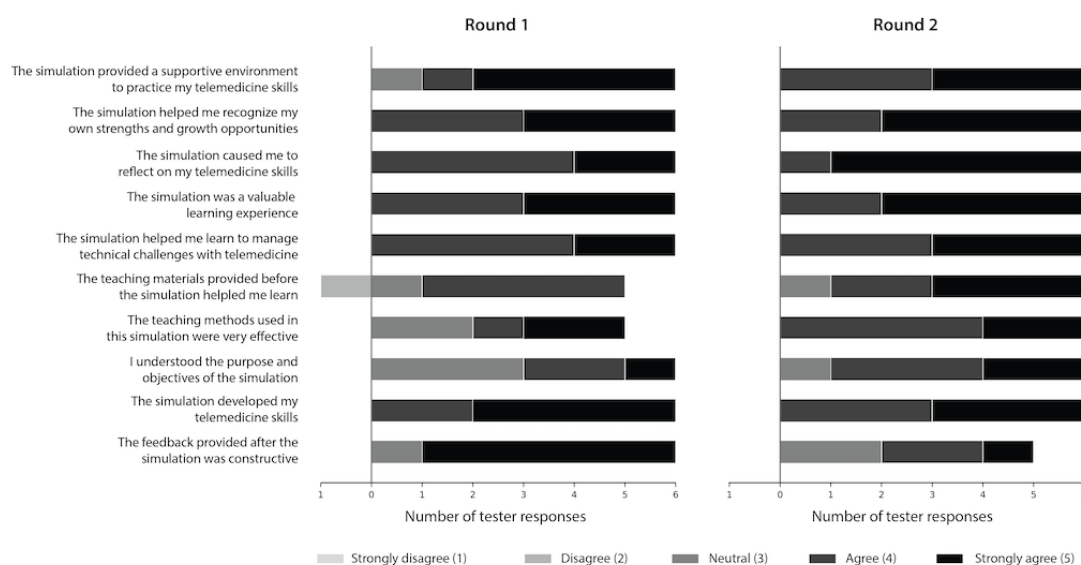


SSES Results

The SSES scores indicate that most testers considered the simulation a valuable learning experience (Figure 9 [53]). One

tester disagreed with the statement that the simulation tested their clinical abilities.

Figure 9. Satisfaction With Simulation Experience Scale scores. We show scores for each testing round, separated by one refinement session.

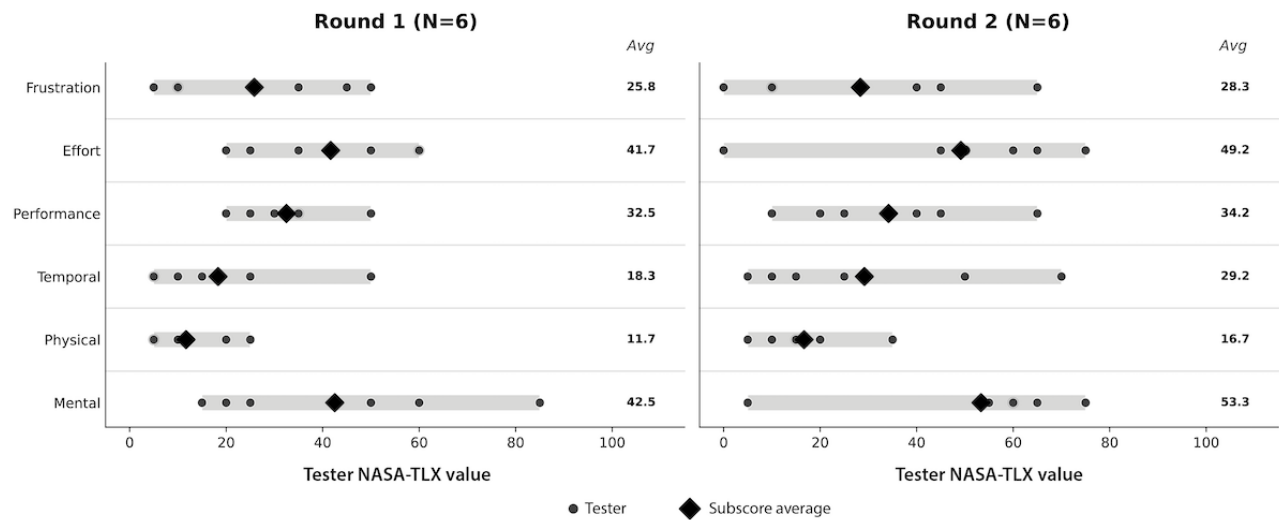


NASA-TLX Results

Testers reported considerable variation in perceived workload, depending on the subscale. Mental demand and effort were consistently the highest and increased from round 1 to round 2 (Figure 10 [69]). Mental workload was highest, with a median

score of 37.50 (IQR 21.25-57.50) for round 1 and a median score of 60.00 (IQR 56.3-63.8) for round 2. The median scores for total effort (a composite of mental and physical effort) were 42.50 (IQR 27.50-57.50) and 55.00 (IQR 46.3-63.8) for rounds 1 and 2, respectively.

Figure 10. Unweighted NASA Task Load Index (NASA-TLX) scores organized by subscales. We show scores for each testing round, separated by one refinement session.

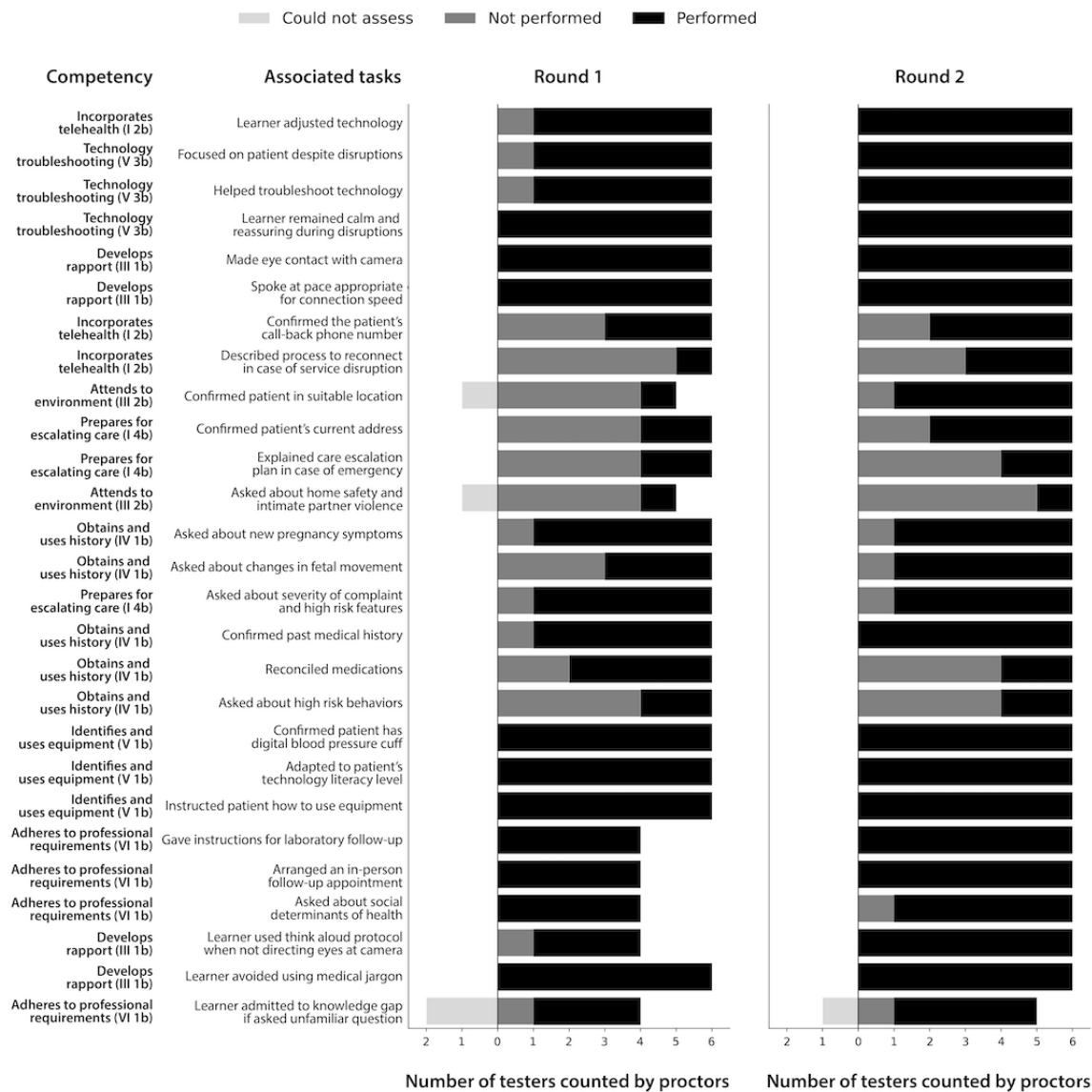


Competency Assessment

In the first round of testing, the proctor and observers could not score 3 tasks associated with 2 competencies (ie, “establishes therapeutic relationships and environments” and “supports solutions to ethical problems”). This problem was caused by

ambiguities in the scoring guide’s language (Figure 11 [36]). Revising the language improved usability. One persistent problem was identifying when a learner fabricated information (eg, improvising the name of a regional laboratory) to preserve the illusion of the simulation. We therefore modified the SP script before piloting to discourage learner improvisation.

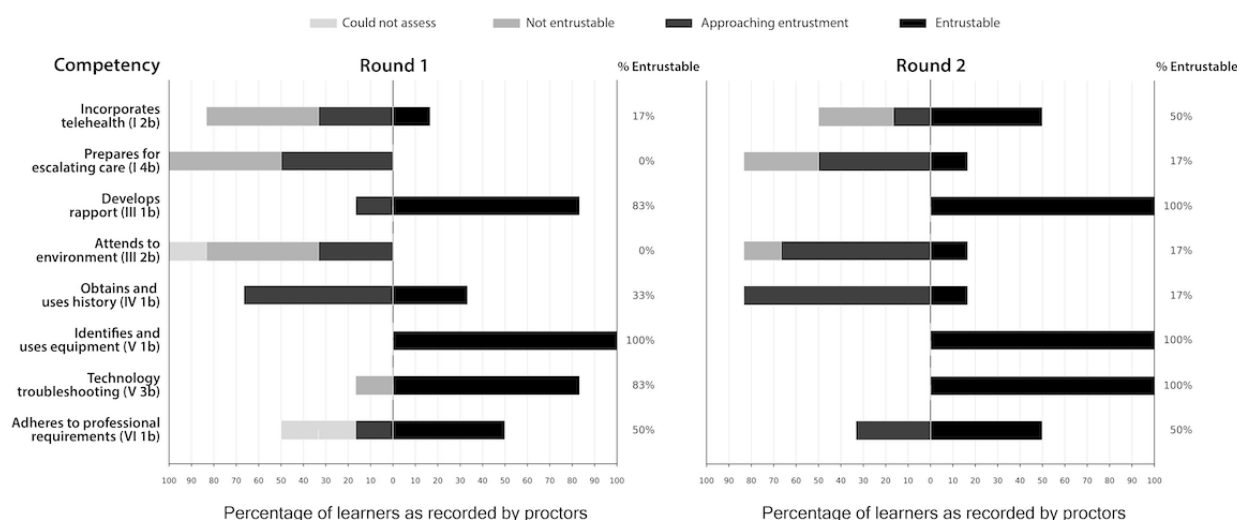
Figure 11. Individual task performance measured using a standardized rubric. Several tasks map to a single telehealth competency. We show task completion rates for each testing round, separated by one refinement session. Competencies were taken from the Association of American Medical Colleges (AAMC) white paper.



We summed proctor-scored task completion rates to calculate composite telemedicine competency scores (Figure 12 [36]). Testers tended to overlook telemedicine-specific tasks, such as gathering contact information, including the home address

(AAMC competency I 2b); describing an emergency escalation plan (AAMC competency I 4b); or checking whether the patient was in a private setting and whether other household members were present (AAMC competency III 2b) [36].

Figure 12. Telemedicine competencies measured using our scoring rubric. Each competency is a composite score of several related tasks. We calculated competencies only for testers with complete data. We show scores for each testing round, separated by one refinement session. Competencies were taken from the Association of American Medical Colleges (AAMC) white paper.



Discussion

Overview

For this tutorial discussion, we first provide an abbreviated review of the most significant usability findings. We then offer a concept-level appraisal of our testing methods and insights for readers planning similar testing protocols. We conclude by comparing with the literature, highlighting gaps, unexplored research opportunities, and future directions.

Principal Findings

Our usability evaluation offered valuable insights into the effectiveness of our educational simulation and opportunities for quality improvement. The data gathered using the ATA enabled the development team to track and communicate improvements to the research team and simulation staff. We identified 52 usability issues requiring prioritization and made 15 design revisions before piloting with learners. By the second round of testing, we removed critical chokepoints related to the audiovisual cross-check and added a mock EHR with synthetic patient data and fill-in data fields. The mock EHR provided the learner with visual cues that improved the telemedicine workflow and task completion rates. Our qualitative data clarified the root causes of usability issues and provided an efficient method for prioritizing revisions. We traced our biggest challenges to the fidelity of blood pressure data and the reliability of the evaluation rubric. We subsequently focused revision efforts on clarifying details in patient scripts, adding synthetic patient data—including a blood pressure flowsheet to the mock EHR—and modifying scoring guide language and definitions in the telemedicine rubric.

It was crucial to analyze the simulation in accordance with the workflow steps, expected tasks, and learner competencies. By measuring simulation steps and task performance independently, we could see when testers completed all scripted steps but overlooked clinical tasks. This helped distinguish undesirable technical distractors from scripted challenges. By removing

usability issues that diverted testers' attention from their tasks, testers could more effectively engage with the learning material. This also helped faculty more easily determine when simulation tasks were misaligned with learning objectives.

Testers successfully used the videoconferencing software, resolved technical issues, and established therapeutic rapport. By contrast, most testers neglected tasks necessary to optimize the telemedicine environment or prepare for a medical emergency. This is not unexpected; none of our testers received telemedicine training, whereas we gave a lecture to residents before the simulation.

For educators planning to include grading rubrics with their simulations, we recommend embedding scoring forms or debriefing guides into the testing instruments. In our example, we included the telemedicine competency rubric in the ATA. Our experience suggests a thorough appraisal of the evaluation materials by domain experts and participating faculty can improve construct validity by aligning simulation tasks and learning objectives with rubric terminology and instrument design. We also believe testing can improve instrument learnability and ease of use so that faculty can invest less time training proctors during piloting and implementation.

We used the SEQ, SUS, and NASA-TLX scores to compare our qualitative observations with usability benchmarks (ie, convergent validity). The range of SEQ scores (2 to 7) and SUS scores (65 to 100) helped identify the most challenging simulation components. The SEQ pinpointed issues in the synthetic data collected during the history and examination.

The NASA-TLX data helped estimate overall simulation difficulty. Responses indicated a relatively high score for mental workload compared with the other subscales (round 1: median score 37.50, IQR 21.25-57.50; round 2: median score 60.00, IQR 56.30-63.80). The scores may have increased due to the inclusion of a mock EHR. Considering these findings, we asked faculty whether the simulation should be modified. Obstetrics and family medicine faculty advised against changing the tasks

because they aligned with the educational goals and learning objectives. During piloting (illustrative vignette in [Textbox 2](#); key tutorial insights in [Textbox 3](#)), residents were indeed able to complete the simulation but struggled with several tasks, including developing a preemptive care escalation plan and

instructing SPs on the use of a digital blood pressure cuff. This suggests that our clinical scenario and materials provided sufficient scaffolding for learners to engage effectively with the material while gently challenging them with unfamiliar tasks.

Textbox 3. Key tutorial insights.

Key tutorial insights

- Agile Task Analysis combined with simulations using a “think aloud” protocol offers a practical strategy for identifying critical design issues.
- It is crucial to include instructional materials, scoring guides, and debriefing instruments in usability data collection instruments. This practice can help disambiguate simulation design issues from learner competency gaps.
- The System Usability Scale is a practical method for quickly benchmarking the overall perceived usability of an educational simulation.
- The addition of the Single Ease Question can help pinpoint the source of usability issues when global usability assessments raise concerns.
- The NASA (National Aeronautics and Space Administration) Task Load Index provides a “quick and dirty” method to estimate the difficulty of an educational simulation and guide future calibration.

Application of Findings to Simulation Development

This tutorial advances simulation science in several ways. First, we demonstrated how to leverage usability methods used in technology evaluations to systematically assess educational simulations before piloting with learners. Validated, theory-driven instruments can provide insights into simulation effectiveness, ease of use, and user satisfaction [22,82-84]. These methods are technology-agnostic and, with little modification, are suitable for evaluating simulation constructs such as engagement, cognitive load, skill acquisition, and competence.

While this study offers turnkey methods for formative usability testing, the choice of any single instrument should be driven by the learning objectives, the simulation’s attributes, and the phase of the development life cycle. Not all methods are necessary for every simulation study. Researchers should add or remove components in a modular fashion. For example, user satisfaction with simulation data may be incidental to the educational program objectives; studies show self-assessment is an unreliable estimate of teaching effectiveness [85,86]. We included the SSES in our protocol because we believed fellow educators could provide useful appraisals of our simulation and its educational value.

Second, our approach addressed design challenges unique to educational simulations [22,23,74,87]. We concurrently evaluated simulation design features and learner competencies to align teaching modalities with educational objectives and to balance ecological validity with simulation complexity [4,11,12,20,88,89]. Our instruments were well suited to observing technology use and should be considered for technology-forward educational initiatives that include EHRs, mobile devices, artificial intelligence, and computerized decision support. Unlike traditional usability testing, our goal was not to remove opportunities for user error [17,21,73,80,90-92]. Educational simulations must enable learning through experimentation, discovery, and even failure [4,42,93,94]. We therefore used the ATA to record whether testers had the opportunity to demonstrate their skills during the simulation;

the simulation could proceed as intended even if a tester failed a task.

Third, our methods can be used early in the product development life cycle with an array of testers [2,4,9]. While best-practice simulation guidelines recommend piloting simulations with learners, there are practical challenges [2]. Recruiting enough target learners can be impractical. Moreover, testing and piloting serve different but equally important roles. Because simulations can be challenging or stressful for learners, pilots may not be the ideal means of measuring usability dimensions such as learnability, error tolerance, and satisfaction [95-98]. Testing with experts can assess the fidelity of the case, the accuracy of the synthetic data, the appropriateness of the learning objectives, and the difficulty of tasks [41-43,99].

Fourth, we believe our methods are flexible, cost-effective, and scalable. We used low-cost tools that could be adapted to resource-constrained settings, such as smaller academic centers and schools in low- and middle-income countries [73,74,100]. Many of the instruments described herein, including the SUS, NASA-TLX, and SSES, have been validated in languages such as Chinese, Korean, Spanish, French, and Croatian [69,101-107]. These methods do not require large numbers of testers. We recommend multiple short testing rounds with as few as 5 testers; this provides sufficient data to guide refinements [108].

Comparison With Prior Work

With the expansion of advanced simulation modalities, such as AR, VR, and AI, software developers are increasingly recognizing the role of usability testing in the product development life cycle [109-114]. There are already protocols for a range of tools, including AR, VR, digital twins, and serious games [8,14,15]. While most reports focus on the technology, several have evaluated the broader simulation context. Anton and colleagues described an evaluation of 2 emergency medicine simulations that combined learner eye tracking and faculty ratings on the Non-Technical Skills for Surgeons Scale [14]. Silva et al [15] applied a more comprehensive approach by evaluating a simulation manikin for teaching cardiopulmonary resuscitation. They created a data collection instrument based on a basic life support guide and scored learners using a 16-item

Likert-type scale. The researchers recorded time on task and administered the SUS after each encounter. In this manner, they operationalized usability in terms of the dimensions of effectiveness, efficiency, and satisfaction. Together, these reports provide complementary perspectives on the evaluation of simulation-based instructional modules: Silva et al [15] focused on the physical usability of manikins, whereas Anton et al [14] focused on clinicians' cognitive skills.

Da Silva and Dubrowski [8] described one of the first attempts to incorporate usability methods early in the development life cycle. They conducted a series of design sprints to evaluate a conflict-management scenario between a nursing student and a senior nurse. The team tested 2 scenarios and recruited 20 nursing students. After each simulation, the team conducted exit interviews and administered the Simulation Design Scale (SDS)—a validated 20-item instrument measuring the quality and effectiveness of simulation-based learning experiences [115]. Unlike the SSES, which focuses on learners' satisfaction with the instructional approach, the SDS focuses on missing or underdeveloped simulation elements. Da Silva and Dubrowski [8] used the SDS as a quality metric and tracked improvement in scores across iterations. We believe this work is an important milestone toward developing a metrics-driven usability strategy.

Strengths, Limitations, and Next Steps to Improve Testing Protocols

We believe the application of usability methods to simulation-based education remains underexplored [41,42,99]. To our knowledge, our report and related preliminary work provide the most complete descriptions of mixed methods usability testing for educational simulations. Combining subjective measures (eg, user satisfaction and ease of use) with objective performance data (eg, task completion rates, error tolerance) can maximize insights into technical functionality and educational value [57,116,117].

These observations notwithstanding, several limitations may affect the generalizability of our work. We identified new usability issues after 2 refinement sessions, suggesting we had not reached saturation. Unfortunately, our team faced time and resource constraints; the number of iterations was, in part, dependent on curriculum deadlines and the number of available testers. To our knowledge, there is no universally accepted number of testers or testing rounds that guarantees researchers will find all usability issues. The usability literature, however, describes an issue-discovery curve with diminishing returns; as few as 5 testers may identify between 80% and 90% of critical issues [17]. Experts therefore recommend conducting several small testing cycles with 5 users at a time [17,118-121].

This report describes a single-site, single-SP, quality-improvement design. This raises several validity issues that usability specialists must consider. We conducted all usability testing with a single SP to control for simulation variability. However, our SP's familiarity with the case may have limited our ability to identify weak prompts, responses, or missing information. In the future, incorporating multiple SPs could reveal variations in script interpretation and behavior. We also observed variation in practice behaviors among subject-matter experts as a function of experience, specialty,

and training. This suggests that institutions may inadvertently incorporate provincial practice patterns and local health system norms into their simulation designs. This could create standardization challenges, since usability studies tend to compare testers' performance against an established workflow or an approved set of product design requirements. In our example, we compared usability findings to local workflow and learning objectives. Researchers designing future studies may elect to use an external reference standard, such as a prerecorded best-practice encounter developed by a professional society.

We did not conduct interrater reliability tests for our rubric. In general, calculating agreement statistics falls outside the scope of usability testing during the product design phase. Instead, several members of the research team independently coded each task. We then compared scores and resolved discrepancies through discussion. Task descriptions with vague or ambiguous language were updated during the refinement phases. We believe the next step in our research program is to measure the interrater reliability of our rubric by asking faculty to independently rate learners who have completed the simulation and calculating an agreement statistic (eg, Cohen κ or Fleiss κ) [122].

We qualitatively coded all usability findings using the usability heuristics for health information technology by Zhang et al [80]. While useful for categorization, the codes did not indicate priority or provide insights for remediation. We believe there are opportunities to develop and validate usability heuristics specific to educational simulations. Ideally, codes informed by educational theory would provide design guidance to educators.

Finally, we see future opportunities to test additional usability measurement tools, including the SDS, heuristic inspections, and cognitive walkthroughs [17,22,115]. This could help enrich our logic models that connect usability tools to evaluation dimensions, such as realism, memorability, completeness, and predicted future learner performance. While videoconferencing may not be required for most simulations, other teams might use this capability to link distributed resources, such as testers, researchers, SPs, and software. To support these kinds of remote testing initiatives, researchers may distribute questionnaires digitally and trial wearable technologies, such as eye-tracking lenses and wearable activity trackers, including wristbands and smartwatches.

Conclusions

We believe this tutorial illustrates a practical, reproducible, and resource-sensitive method for usability testing educational simulations before piloting with target learners. While this report describes a single-site, single-SP, quality-improvement design, we strove to include published, validated instruments with claims of generalizability and relatively low-cost testing methods to permit replication or context-appropriate adaptation. We have included appendices with our instruments and encourage readers to adapt and apply these methods to a range of simulations. While we believe this monograph provides simulationists with a turnkey toolkit to conduct similar assessments, further studies using psychometrically validated instruments are needed to clarify the best educational use cases, evaluate additional constructs relevant to teaching and learning, and, when relevant,

incorporate additional tests to measure the interrater reliability of learner assessment instruments.

Acknowledgments

The views expressed in this publication are solely those of the authors and do not necessarily reflect the official policies of the US Department of Health and Human Services or the Health Resources and Services Administration, nor does mention of the department or agency names imply endorsement by the US Government.

We thank the Simulation Center of the University of Oklahoma-Tulsa (SCOUT) for providing simulation center space, standardized patients, and consultative support for this project. We thank Billi Coil and Guimy Castor for research coordination support. We thank Nancy Melton and Trisha Cook for administrative support.

The authors declare the use of generative AI (GenAI) at specific steps in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision: code generation for Microsoft Excel formulas and spelling and grammar checking. The GenAI tools used included Grammarly, Microsoft Word 365 grammar and dictionary, and ChatGPT 5.5 (OpenAI). Responsibility for the final manuscript lies entirely with the authors. GenAI tools are not listed as authors and do not bear responsibility for the final manuscript. The authors did not use GenAI for text generation, ideation, protocol design, usability testing, data collection, final data analysis, or figure drafting.

Funding

Funding support for this project was provided in part by the University of Oklahoma School of Community Medicine (OUSCM) and the US Department of Health and Human Services, Health Resources and Services Administration (HRSA), as part of an approved federal and nonfederal award (grant number T34HP42150).

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

BJL, HM, JH, and SM developed the usability protocol. BJL, HM, and RM developed the theoretical framework and logic model. BJL, MR, KPG, and KB developed and revised the telemedicine simulation, including patient scripts, synthetic data, technology, and the evaluation rubric. BJL, KPG, MR, JVB, MA, and CP conducted all usability testing, data collection, and preliminary data analysis. BJL, HM, JH, AD, and RM conducted all final data analyses. BJL and MA generated all tables and figures. All authors contributed to the final manuscript and reviewed the submitted version.

Conflicts of Interest

BJL is the Editor-in-Chief of *JMIR Medical Education*. The authors declare no additional conflicts of interest.

Multimedia Appendix 1

Demographic questionnaire.

[\[DOCX File , 3806 KB-Multimedia Appendix 1\]](#)

Multimedia Appendix 2

Agile task analysis with embedded telemedicine competency rubric.

[\[DOCX File , 5427 KB-Multimedia Appendix 2\]](#)

Multimedia Appendix 3

List of Single Ease Questions (SEQ) for sets of related steps.

[\[DOCX File , 3805 KB-Multimedia Appendix 3\]](#)

Multimedia Appendix 4

Adapted post-simulation System Usability Scale (SUS) and Satisfaction with Simulation Experience Survey (SSES).

[\[DOCX File , 3806 KB-Multimedia Appendix 4\]](#)

Multimedia Appendix 5

Unweighted NASA Task Load Index (NASA-TLX).

[\[DOCX File , 3805 KB-Multimedia Appendix 5\]](#)

Multimedia Appendix 6

Semistructured simulation debrief guide.

[\[DOCX File , 3803 KB-Multimedia Appendix 6\]](#)

Multimedia Appendix 7

Structural codes data dictionary for usability issue coding.

[\[DOCX File , 4147 KB-Multimedia Appendix 7\]](#)

Multimedia Appendix 8

Usability issue log template with sample entries.

[\[DOCX File , 3804 KB-Multimedia Appendix 8\]](#)

Multimedia Appendix 9

Telehealth competency rubric used during resident piloting.

[\[DOCX File , 4496 KB-Multimedia Appendix 9\]](#)

Multimedia Appendix 10

Printed telemedicine memory aid.

[\[DOCX File , 5340 KB-Multimedia Appendix 10\]](#)

References

1. Elendu C, Amaechi DC, Okatta AU, Amaechi EC, Elendu TC, Ezech CP, et al. The impact of simulation-based training in medical education: a review. *Medicine (Baltimore)*. 2024;103(27):e38813. [[FREE Full text](#)] [doi: [10.1097/MD.00000000000038813](https://doi.org/10.1097/MD.00000000000038813)] [Medline: [38968472](https://pubmed.ncbi.nlm.nih.gov/38968472/)]
2. Watts PI, McDermott DS, Alinier G, Charnetski M, Ludlow J, Horsley E, et al. Healthcare simulation standards of best practice™ simulation design. *Clin Simul Nurs*. 2021;58:14-21. [doi: [10.1016/j.ecns.2021.08.009](https://doi.org/10.1016/j.ecns.2021.08.009)]
3. Aggarwal R, Mytton OT, Derbrew M, Hananel D, Heydenburg M, Issenberg B, et al. Training and simulation for patient safety. *Qual Saf Health Care*. 2010;19(Suppl 2):i34-i43. [[FREE Full text](#)] [doi: [10.1136/qshc.2009.038562](https://doi.org/10.1136/qshc.2009.038562)] [Medline: [20693215](https://pubmed.ncbi.nlm.nih.gov/20693215/)]
4. Alinier G. Developing high-fidelity health care simulation scenarios: a guide for educators and professionals. *Simul Gaming*. 2010;42(1):9-26. [doi: [10.1177/1046878109355683](https://doi.org/10.1177/1046878109355683)]
5. Gliva-McConvey G, Nicholas CF, Nicholas L, editors. *Comprehensive Healthcare Simulation: Implementing Best Practices in Standardized Patient Methodology*. Cham, Switzerland. Springer; 2020.
6. Purva M, Nicklin J. ASPIH standards for simulation-based education: process of consultation, design and implementation. *BMJ Simul Technol Enhanc Learn*. 2018;4(3):117-125. [[FREE Full text](#)] [doi: [10.1136/bmjstel-2017-000232](https://doi.org/10.1136/bmjstel-2017-000232)] [Medline: [35520464](https://pubmed.ncbi.nlm.nih.gov/35520464/)]
7. Diaz-Navarro C, Laws-Chapman C, Moneypenny M, Purva M. The ASPIH Standards – 2023: guiding simulation-based practice in health and care. *Int J Healthc Simul*. 2024. [doi: [10.54531/nyvm5886](https://doi.org/10.54531/nyvm5886)]
8. Da Silva C, Dubrowski A. Piloting simulations: a systematic refinement strategy. *Cureus*. 2019;11(12):e6434. [[FREE Full text](#)] [doi: [10.7759/cureus.6434](https://doi.org/10.7759/cureus.6434)] [Medline: [31993271](https://pubmed.ncbi.nlm.nih.gov/31993271/)]
9. 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 (Lond)*. 2017;2:10. [[FREE Full text](#)] [doi: [10.1186/s41077-017-0043-4](https://doi.org/10.1186/s41077-017-0043-4)] [Medline: [29450011](https://pubmed.ncbi.nlm.nih.gov/29450011/)]
10. Fanning RM, Gaba DM. The role of debriefing in simulation-based learning. *Simul Healthc*. 2007;2(2):115-125. [doi: [10.1097/SIH.0b013e3180315539](https://doi.org/10.1097/SIH.0b013e3180315539)] [Medline: [19088616](https://pubmed.ncbi.nlm.nih.gov/19088616/)]
11. Marcilly R, Monkman H, Pelayo S, Lesselroth B. Usability evaluation ecological validity: is more always better? *Healthcare (Basel)*. 2024;12(14):1417. [[FREE Full text](#)] [doi: [10.3390/healthcare12141417](https://doi.org/10.3390/healthcare12141417)] [Medline: [39057560](https://pubmed.ncbi.nlm.nih.gov/39057560/)]
12. Dieckmann P, Gaba D, Rall M. Deepening the theoretical foundations of patient simulation as social practice. *Simul Healthc*. 2007;2(3):183-193. [doi: [10.1097/SIH.0b013e3180f637f5](https://doi.org/10.1097/SIH.0b013e3180f637f5)] [Medline: [19088622](https://pubmed.ncbi.nlm.nih.gov/19088622/)]
13. Shin J, Lee H. Optimal usability test procedure generation for medical devices. *Healthcare (Basel)*. 2023;11(3):296. [[FREE Full text](#)] [doi: [10.3390/healthcare11030296](https://doi.org/10.3390/healthcare11030296)] [Medline: [36766871](https://pubmed.ncbi.nlm.nih.gov/36766871/)]
14. Anton NE, Cha JS, Hernandez E, Athanasiadis DI, Yang J, Zhou G, et al. Utilizing eye tracking to assess medical student non-technical performance during scenario-based simulation: results of a pilot study. *Global Surg Educ*. 2023;2(1):49. [doi: [10.1007/s44186-023-00127-3](https://doi.org/10.1007/s44186-023-00127-3)] [Medline: [38414559](https://pubmed.ncbi.nlm.nih.gov/38414559/)]

15. Silva LDB, Gonçalves FT, Arantes HP, Garcia RA, Gasparin MRFR, dos Reis MM, et al. Usability evaluation of high-fidelity simulation manikin for cardiopulmonary resuscitation training for medical students. *Res Biomed Eng*. 2024;40(1):253-264. [doi: [10.1007/s42600-024-00340-z](https://doi.org/10.1007/s42600-024-00340-z)]
16. *Ergonomics of Human-System Interaction—Part 210: Human-Centred Design for Interactive Systems*. Geneva, Switzerland. International Organization for Standardization; 2019:33.
17. Barnum CM. *Usability Testing Essentials: Ready, Set...Test!* Burlington, Massachusetts. Morgan Kaufmann; 2010.
18. Yates N, Gough S, Brazil V. Self-assessment: with all its limitations, why are we still measuring and teaching it? Lessons from a scoping review. *Med Teach*. 2022;44(11):1296-1302. [doi: [10.1080/0142159X.2022.2093704](https://doi.org/10.1080/0142159X.2022.2093704)] [Medline: [35786121](https://pubmed.ncbi.nlm.nih.gov/35786121/)]
19. Shackel B. The concept of usability. In: Bennett JL, Case D, Sandelin J, Smith M, editors. *Visual Display Terminals: Usability Issues and Health Concerns*. Englewood Cliffs, New Jersey. Prentice-Hall; 1984:4823.
20. Monkman H, Marcilly R, Lesselroth B. An integrated model of external validity usability evaluations in health care. *Stud Health Technol Inform*. 2025;326:80-84. [FREE Full text] [doi: [10.3233/SHTI250242](https://doi.org/10.3233/SHTI250242)] [Medline: [40357606](https://pubmed.ncbi.nlm.nih.gov/40357606/)]
21. Nielsen J. *Usability Engineering*. 1st Edition. San Francisco, California. Morgan Kaufmann; 1994.
22. Harrington L, Harrington C. *Usability Evaluation Handbook for Electronic Health Records*. Chicago, Illinois. HIMSS; 2014.
23. Albert W, Tullis T. *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics*. Waltham, Massachusetts. Morgan Kaufmann; 2013.
24. Bashshur R, Doarn CR, Frenk JM, Kvedar JC, Woolliscroft JO. Telemedicine and the COVID-19 pandemic, lessons for the future. *Telemed J E Health*. 2020;26(5):571-573. [doi: [10.1089/tmj.2020.29040.rb](https://doi.org/10.1089/tmj.2020.29040.rb)] [Medline: [32275485](https://pubmed.ncbi.nlm.nih.gov/32275485/)]
25. Basu A, Kuziemy C, de Araújo Novaes M, Kleber A, Sales F, Al-Shorbaji N, et al. Telehealth and the COVID-19 pandemic: international perspectives and a health systems framework for telehealth implementation to support critical response. *Yearb Med Inform*. 2021;30(1):126-133. [FREE Full text] [doi: [10.1055/s-0041-1726484](https://doi.org/10.1055/s-0041-1726484)] [Medline: [33882598](https://pubmed.ncbi.nlm.nih.gov/33882598/)]
26. Clare CA. Telehealth and the digital divide as a social determinant of health during the COVID-19 pandemic. *Netw Model Anal Health Inform Bioinform*. 2021;10(1):26. [FREE Full text] [doi: [10.1007/s13721-021-00300-y](https://doi.org/10.1007/s13721-021-00300-y)] [Medline: [33842187](https://pubmed.ncbi.nlm.nih.gov/33842187/)]
27. Eslami Jahromi M, Ayatollahi H. Utilization of telehealth to manage the COVID-19 pandemic in low- and middle-income countries: a scoping review. *J Am Med Inform Assoc*. 2023;30(4):738-751. [FREE Full text] [doi: [10.1093/jamia/ocac250](https://doi.org/10.1093/jamia/ocac250)] [Medline: [36565464](https://pubmed.ncbi.nlm.nih.gov/36565464/)]
28. Thies KM, Gonzalez M, Porto A, Ashley KL, Korman S, Lamb M. Project ECHO COVID-19: vulnerable populations and telehealth early in the pandemic. *J Prim Care Community Health*. 2021;12:21501327211019286. [FREE Full text] [doi: [10.1177/21501327211019286](https://doi.org/10.1177/21501327211019286)] [Medline: [34036832](https://pubmed.ncbi.nlm.nih.gov/34036832/)]
29. Wijesooriya NR, Mishra V, Brand PLP, Rubin BK. COVID-19 and telehealth, education, and research adaptations. *Paediatr Respir Rev*. 2020;35:38-42. [FREE Full text] [doi: [10.1016/j.prrv.2020.06.009](https://doi.org/10.1016/j.prrv.2020.06.009)] [Medline: [32653468](https://pubmed.ncbi.nlm.nih.gov/32653468/)]
30. Alcocer Alkureishi M, Lenti G, Choo Z-Y, Castaneda J, Weyer G, Oyler J, et al. Teaching telemedicine: the next frontier for medical educators. *JMIR Med Educ*. 2021;7(2):e29099. [FREE Full text] [doi: [10.2196/29099](https://doi.org/10.2196/29099)] [Medline: [33878011](https://pubmed.ncbi.nlm.nih.gov/33878011/)]
31. Belakovskiy A, Jones EK. Telehealth and medical education. *Prim Care*. 2022;49(4):575-583. [doi: [10.1016/j.pop.2022.04.003](https://doi.org/10.1016/j.pop.2022.04.003)] [Medline: [36357062](https://pubmed.ncbi.nlm.nih.gov/36357062/)]
32. Chike-Harris KE, Durham C, Logan A, Smith G, DuBose-Morris R. Integration of telehealth education into the health care provider curriculum: a review. *Telemed J E Health*. 2021;27(2):137-149. [doi: [10.1089/tmj.2019.0261](https://doi.org/10.1089/tmj.2019.0261)] [Medline: [32250196](https://pubmed.ncbi.nlm.nih.gov/32250196/)]
33. Khullar D, Mullangi S, Yu J, Weems K, Shipman SA, Caulfield M, et al. The state of telehealth education at U.S. medical schools. *Healthc (Amst)*. 2021;9(2):100522. [doi: [10.1016/j.hjdsi.2021.100522](https://doi.org/10.1016/j.hjdsi.2021.100522)] [Medline: [33548854](https://pubmed.ncbi.nlm.nih.gov/33548854/)]
34. Noronha C, Lo MC, Nikiforova T, Jones D, Nandiwada DR, Leung TI, et al. Society of General Internal Medicine (SGIM) Education Committee. Telehealth competencies in medical education: new frontiers in faculty development and learner assessments. *J Gen Intern Med*. 2022;37(12):3168-3173. [FREE Full text] [doi: [10.1007/s11606-022-07564-8](https://doi.org/10.1007/s11606-022-07564-8)] [Medline: [35474505](https://pubmed.ncbi.nlm.nih.gov/35474505/)]
35. Waseh S, Dicker AP. Telemedicine training in undergraduate medical education: mixed-methods review. *JMIR Med Educ*. 2019;5(1):e12515. [FREE Full text] [doi: [10.2196/12515](https://doi.org/10.2196/12515)] [Medline: [30958269](https://pubmed.ncbi.nlm.nih.gov/30958269/)]
36. Telehealth competencies. Association of American Medical Colleges. URL: <https://www.aamc.org/data-reports/report/telehealth-competencies> [accessed 2026-08-14]
37. Cantone RE, Palmer R, Dodson LG, Biagioli FE. Insomnia telemedicine OSCE (TeleOSCE): a simulated standardized patient video-visit case for clerkship students. *MedEdPORTAL*. 2019;15:10867. [FREE Full text] [doi: [10.15766/mep.2374-8265.10867](https://doi.org/10.15766/mep.2374-8265.10867)] [Medline: [32051850](https://pubmed.ncbi.nlm.nih.gov/32051850/)]
38. Papanagnou D, Stone D, Chandra S, Watts P, Chang A, Hollander J. Integrating telehealth emergency department follow-up visits into residency training. *Cureus*. 2018;10(4):e2433. [FREE Full text] [doi: [10.7759/cureus.2433](https://doi.org/10.7759/cureus.2433)] [Medline: [29876155](https://pubmed.ncbi.nlm.nih.gov/29876155/)]
39. Mulcare M, Naik N, Greenwald P, Schullstrom K, Gogia K, Clark S, et al. Advanced communication and examination skills in telemedicine: a structured simulation-based course for medical students. *MedEdPORTAL*. 2020;16:11047. [FREE Full text] [doi: [10.15766/mep.2374-8265.11047](https://doi.org/10.15766/mep.2374-8265.11047)] [Medline: [33365390](https://pubmed.ncbi.nlm.nih.gov/33365390/)]
40. Charron E, Castor G, Veach C, Chubb R, Lesselroth B, Sachs VESS, et al. Enhanced obstetric training to address maternity care workforce shortages in tribal, rural, and underserved communities: a case from Oklahoma. *Healthc (Amst)*. 2025;13(2):100768. [doi: [10.1016/j.hjdsi.2025.100768](https://doi.org/10.1016/j.hjdsi.2025.100768)] [Medline: [40926429](https://pubmed.ncbi.nlm.nih.gov/40926429/)]

41. Lesselroth B, Monkman H, Anderson M, Parsons C, Mohanty S, Lawson A, et al. Usability testing an educational telemedicine simulation: description of methods and preliminary findings. *Stud Health Technol Inform*. 2025;327:273-277. [doi: [10.3233/SHTI250327](https://doi.org/10.3233/SHTI250327)] [Medline: [40380438](#)]
42. Lesselroth B, Monkman H, Marcilly R, Mohanty S, Anderson M, Parsons C, et al. Hybrid usability analysis to improve an educational telemedicine simulation: task analysis and survey results. *Stud Health Technol Inform*. 2025;326:59-63. [FREE Full text] [doi: [10.3233/SHTI250237](https://doi.org/10.3233/SHTI250237)] [Medline: [40357602](#)]
43. Monkman H, Homco J, Bellavia K, Anderson M, Tanriverdi A, Lesselroth C. Iterative design of an educational telemedicine simulation: addressing usability issues to enhance the simulation. 2026. Presented at: Studies in Health Technology and Informatics; Opening the Personal Gate Between Technology and Health Care: Proceedings of MIE 2026; May 25-28, 2026:2130-2134; Genoa, Italy. [doi: [10.3233/SHTI260637](https://doi.org/10.3233/SHTI260637)]
44. Sonenberg A, Mason D. Maternity care deserts in the US. *JAMA Health Forum*. 2023;4(1):e225541. [FREE Full text] [doi: [10.1001/jamahealthforum.2022.5541](https://doi.org/10.1001/jamahealthforum.2022.5541)] [Medline: [36633853](#)]
45. Fontenot J, Brigrance C, Lucas R, Stoneburner A. Navigating geographical disparities: access to obstetric hospitals in maternity care deserts and across the United States. *BMC Pregnancy Childbirth*. 2024;24(1):350. [doi: [10.1186/s12884-024-06535-7](https://doi.org/10.1186/s12884-024-06535-7)] [Medline: [38720255](#)]
46. Atwani R, Robbins L, Saade G, Kawakita T. Association of maternity care deserts with maternal and pregnancy-related mortality. *Obstet Gynecol*. 2025;146(2):181-188. [doi: [10.1097/AOG.0000000000005976](https://doi.org/10.1097/AOG.0000000000005976)] [Medline: [40570350](#)]
47. Gleeson DE, Busch SH, Ickovics JR. State-level prevalence of maternity care deserts: association with healthcare access, utilization, and outcomes among medicaid recipients. *AJPM Focus*. 2025;4(5):100362. [FREE Full text] [doi: [10.1016/j.focus.2025.100362](https://doi.org/10.1016/j.focus.2025.100362)] [Medline: [40735223](#)]
48. Tanne JH. Nearly six million women in the US live in maternity care deserts. *BMJ*. 2023;382:1878. [doi: [10.1136/bmj.p1878](https://doi.org/10.1136/bmj.p1878)] [Medline: [37580083](#)]
49. Adashi EY, O'Mahony DP, Cohen IG. Maternity care deserts: key drivers of the national maternal health crisis. *J Am Board Fam Med*. 2025;38(1):165-167. [doi: [10.3122/jabfm.2024.240198R1](https://doi.org/10.3122/jabfm.2024.240198R1)] [Medline: [40268316](#)]
50. Quesenberry W. The five dimensions of usability. In: *Content and Complexity: Information Design in Technical Communication*. New York. Routledge; 2014:93-114.
51. Davis FD. *A Technology Acceptance Model for Empirically Testing New End-User Information Systems: Theory and Results*. Cambridge, Massachusetts. Massachusetts Institute of Technology, Sloan School of Management; 1985.
52. Sauro J. *A Practical Guide to the System Usability Scale: Background, Benchmarks & Best Practices*. Denver, Colorado. Measuring Usability LLC; 2011.
53. Levett-Jones T, McCoy M, Lapkin S, Noble D, Hoffman K, Dempsey J, et al. The development and psychometric testing of the Satisfaction with Simulation Experience Scale. *Nurse Educ Today*. 2011;31(7):705-710. [doi: [10.1016/j.nedt.2011.01.004](https://doi.org/10.1016/j.nedt.2011.01.004)] [Medline: [21288606](#)]
54. Hart SG. *NASA Task Load Index (TLX): Paper and Pencil Package, Version 1.0*. Moffett Field, California. NASA Ames Research Center; 1986.
55. Sonderegger A, Sauer J. The influence of laboratory set-up in usability tests: effects on user performance, subjective ratings and physiological measures. *Ergonomics*. 2009;52(11):1350-1361. [doi: [10.1080/00140130903067797](https://doi.org/10.1080/00140130903067797)] [Medline: [19851903](#)]
56. Crotty BH, Hyun N, Polovneff A, Dong Y, Decker MC, Mortensen N, et al. Analysis of clinician and patient factors and completion of telemedicine appointments using video. *JAMA Netw Open*. 2021;4(11):e2132917. [FREE Full text] [doi: [10.1001/jamanetworkopen.2021.32917](https://doi.org/10.1001/jamanetworkopen.2021.32917)] [Medline: [34735013](#)]
57. Lesselroth BJ, Mastarone GL, Adams K, Tallett S, Kushniruk A, Borycki EM. Hybrid usability methods: practical techniques for evaluating health information technology in an operational setting. In: *Patient-Centered Digital Healthcare Technology: Novel Applications for Next Generation Healthcare Systems*. Stevenage, England. Institution of Engineering and Technology; 2020.
58. Kirwan B, Ainsworth L, Pendlebury G. Task analysis: the state of the art in process control. In: *Contemporary Ergonomics*. London, England. Taylor & Francis; 2020:479-484.
59. Stanton NA. Systematic human error reduction and prediction approach (SHERPA). In: Stanton NA, Hedge A, Brookhuis KA, Salas E, Hendrick HW, editors. *Handbook of Human Factors and Ergonomics Methods*. London, England. CRC Press; 2004:394-403.
60. ten Cate O. An updated primer on entrustable professional activities (EPAs). *Rev Bras Educ Med*. 2019;43(1 suppl 1):712-720. [doi: [10.1590/1981-5271v43suplemento1-20190238.ING](https://doi.org/10.1590/1981-5271v43suplemento1-20190238.ING)]
61. Cate OT. A primer on entrustable professional activities. *Korean J Med Educ*. 2018;30(1):1-10. [FREE Full text] [doi: [10.3946/kjme.2018.76](https://doi.org/10.3946/kjme.2018.76)] [Medline: [29510603](#)]
62. Sauro J, Dumas J. Comparison of three one-question, post-task usability questionnaires. 2009. Presented at: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2009); April 4-9, 2009:1599-1608; Boston, Massachusetts.
63. Sauro J, Lewis J. How the SEQ correlates with other task metrics. *MeasuringU*. 2026. URL: <https://measuringu.com/how-the-seq-correlates-with-other-task-metrics/> [accessed 2026-02-05]

64. Sauro J. 10 things to know about the Single Ease Question (SEQ). MeasuringU. 2012. URL: <https://measuringu.com/seq10/> [accessed 2026-02-04]
65. Russ-Jara AL, Saleem JJ, Herout J. A practical guide to usability questionnaires that evaluate clinicians' perceptions of health information technology. *J Biomed Inform.* 2025;165:104822. [doi: [10.1016/j.jbi.2025.104822](https://doi.org/10.1016/j.jbi.2025.104822)] [Medline: [40174758](https://pubmed.ncbi.nlm.nih.gov/40174758/)]
66. Brooke J. SUS: a quick and dirty usability scale. In: Jordan PW, Thomas B, Weerdmeester BA, McClelland IL, editors. *Usability Evaluation in Industry*. London, England. Taylor & Francis; 1996:4-7.
67. Lewis JR, Sauro J. The factor structure of the System Usability Scale. 2009. Presented at: First International Conference on Human Centered Design (HCD 2009); July 19-24, 2009:94-103; San Diego, California.
68. Williams B, Dousek S. The satisfaction with simulation experience scale (SSES): a validation study. *J Nurs Educ Pract.* 2012;2(3):74. [doi: [10.5430/jnep.v2n3p74](https://doi.org/10.5430/jnep.v2n3p74)]
69. Hart SG. NASA-task load index (NASA-TLX): 20 years later. 2006. Presented at: Proceedings of the Human Factors and Ergonomics Society Annual Meeting; October 16-20, 2006; San Francisco, California. [doi: [10.1177/154193120605000909](https://doi.org/10.1177/154193120605000909)]
70. Young MS, Brookhuis KA, Wickens CD, Hancock PA. State of science: mental workload in ergonomics. *Ergonomics.* 2015;58(1):1-17. [doi: [10.1080/00140139.2014.956151](https://doi.org/10.1080/00140139.2014.956151)] [Medline: [25442818](https://pubmed.ncbi.nlm.nih.gov/25442818/)]
71. Grier RA. How high is high? A meta-analysis of NASA-TLX global workload scores. 2015. Presented at: Proceedings of the human factors and ergonomics society annual meeting; October 26-30, 2015; Los Angeles, California. [doi: [10.1177/1541931215591373](https://doi.org/10.1177/1541931215591373)]
72. Russ AL, Baker DA, Fahner WJ, Milligan BS, Cox L, Hagg HK, et al. A rapid usability evaluation (RUE) method for health information technology. *AMIA Annu Symp Proc.* 2010;2010:702-706. [FREE Full text] [Medline: [21347069](https://pubmed.ncbi.nlm.nih.gov/21347069/)]
73. Krug S. *Rocket Surgery Made Easy: The Do-It-Yourself Guide to Finding and Fixing Usability Problems*. Berkeley, California. New Riders; 2009.
74. Kushniruk AW, Patel VL. Cognitive and usability engineering methods for the evaluation of clinical information systems. *J Biomed Inform.* 2004;37(1):56-76. [FREE Full text] [doi: [10.1016/j.jbi.2004.01.003](https://doi.org/10.1016/j.jbi.2004.01.003)] [Medline: [15016386](https://pubmed.ncbi.nlm.nih.gov/15016386/)]
75. Kushniruk AW. Analysis of complex decision-making processes in health care: cognitive approaches to health informatics. *J Biomed Inform.* 2001;34(5):365-376. [FREE Full text] [doi: [10.1006/jbin.2001.1021](https://doi.org/10.1006/jbin.2001.1021)] [Medline: [12123153](https://pubmed.ncbi.nlm.nih.gov/12123153/)]
76. Zoom. Zoom Video Communications, Inc. 2013. URL: <https://zoom.us/> [accessed 2026-07-29]
77. Parsons C, Monkman H, Jones B, Wolfenbarger A, Homco J, Anderson M. Co-creating a mnemonic with learners to support telehealth competency development during simulations. 2025. Presented at: Context Sensitive Health Informatics: AI for Social Good: Proceedings of CSHI 2025; Studies in Health Technology and Informatics, volume 326; May 23-24, 2025:137-141; Bradford, United Kingdom. [doi: [10.3233/shiti250255](https://doi.org/10.3233/shiti250255)]
78. LearningSpace. CAE Healthcare. URL: <https://www.caehealthcare.com/learningspace/> [accessed 2026-08-14]
79. Saldana J. *Fundamentals of Qualitative Research*. New York. Oxford University Press; 2011.
80. Zhang J, Johnson TR, Patel VL, Paige DL, Kubose T. Using usability heuristics to evaluate patient safety of medical devices. *J Biomed Inform.* 2003;36(1-2):23-30. [FREE Full text] [doi: [10.1016/s1532-0464\(03\)00060-1](https://doi.org/10.1016/s1532-0464(03)00060-1)] [Medline: [14552844](https://pubmed.ncbi.nlm.nih.gov/14552844/)]
81. Shneiderman B. *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Bengaluru, India. Pearson Education; 2010.
82. Zhang J, Walji MF. TURF: toward a unified framework of EHR usability. *J Biomed Inform.* 2011;44(6):1056-1067. [FREE Full text] [doi: [10.1016/j.jbi.2011.08.005](https://doi.org/10.1016/j.jbi.2011.08.005)] [Medline: [21867774](https://pubmed.ncbi.nlm.nih.gov/21867774/)]
83. Kushniruk A, Nohr C, Jensen S, Borycki EM. From usability testing to clinical simulations: bringing context into the design and evaluation of usable and safe health information technologies. *Yearb Med Inform.* 2018;22(01):78-85. [doi: [10.1055/s-0038-1638836](https://doi.org/10.1055/s-0038-1638836)]
84. Lesselroth BJ, Monkman H, Adams K, Wood S, Corbett A, Homco J, et al. User experience theories, models, and frameworks: a focused review of the healthcare literature. *Stud Health Technol Inform.* 2020;270:1076-1080. [doi: [10.3233/SHTI200327](https://doi.org/10.3233/SHTI200327)] [Medline: [32570547](https://pubmed.ncbi.nlm.nih.gov/32570547/)]
85. León SP, Panadero E, García-Martínez I. How accurate are our students? A meta-analytic systematic review on self-assessment scoring accuracy. *Educ Psychol Rev.* 2023;35(4):106. [doi: [10.1007/s10648-023-09819-0](https://doi.org/10.1007/s10648-023-09819-0)]
86. Zheng B, He Q, Lei J. Informing factors and outcomes of self-assessment practices in medical education: a systematic review. *Ann Med.* 2024;56(1):2421441. [FREE Full text] [doi: [10.1080/07853890.2024.2421441](https://doi.org/10.1080/07853890.2024.2421441)] [Medline: [39460558](https://pubmed.ncbi.nlm.nih.gov/39460558/)]
87. Sauro J, Lewis JR. *Quantifying the User Experience: Practical Statistics for User Research*. Cambridge, Massachusetts. Morgan Kaufmann; 2016.
88. Issenberg SB, McGaghie WC, Petrusa ER, Lee Gordon D, Scalese RJ. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. *Med Teach.* 2005;27(1):10-28. [doi: [10.1080/01421590500046924](https://doi.org/10.1080/01421590500046924)] [Medline: [16147767](https://pubmed.ncbi.nlm.nih.gov/16147767/)]
89. Issenberg SB, Scalese RJ. Best evidence on high-fidelity simulation: what clinical teachers need to know. *Clin Teach.* 2007;4(2):73-77. [doi: [10.1111/j.1743-498X.2007.00161.x](https://doi.org/10.1111/j.1743-498X.2007.00161.x)]
90. Carayon P. *Handbook of Human Factors and Ergonomics in Health Care and Patient Safety*. Mahwah, New Jersey. CRC Press; 2006.

91. Carayon P, Wetterneck TB, Rivera-Rodriguez AJ, Hundt AS, Hoonakker P, Holden R, et al. Human factors systems approach to healthcare quality and patient safety. *Appl Ergon*. 2014;45(1):14-25. [FREE Full text] [doi: [10.1016/j.apergo.2013.04.023](https://doi.org/10.1016/j.apergo.2013.04.023)] [Medline: [23845724](https://pubmed.ncbi.nlm.nih.gov/23845724/)]
92. Kohn LT, Corrigan JM, Donladson MS. Kohn LT, Corrigan JM, Donaldson MS, editors. *To Err is Human: Building a Safer Health System*. Washington, District of Columbia. National Academies Press; 2000.
93. Monkman H, Kuziemsky C, Homco J, Liew A, Rodriguez K, Skaggs J, et al. Identifying failure modes in telemedicine: an instructional needs assessment. *Stud Health Technol Inform*. 2023;304:39-43. [doi: [10.3233/SHTI230365](https://doi.org/10.3233/SHTI230365)] [Medline: [37347566](https://pubmed.ncbi.nlm.nih.gov/37347566/)]
94. Monkman H, Palmer R, Ijams S, Kollaja L, Rodriguez KA, Liew A, et al. Using simulations to train medical students for unanticipated technology failures in telemedicine. *Stud Health Technol Inform*. 2022;294:775-779. [doi: [10.3233/SHTI220582](https://doi.org/10.3233/SHTI220582)] [Medline: [35612202](https://pubmed.ncbi.nlm.nih.gov/35612202/)]
95. Suh Y, Shim H, Lee HJ, Lee JH, Park MS, Choi S. Effects of subjective stress levels on learning effectiveness in high-fidelity simulation among healthcare workers. *Acta Psychol (Amst)*. 2025;257:105075. [FREE Full text] [doi: [10.1016/j.actpsy.2025.105075](https://doi.org/10.1016/j.actpsy.2025.105075)] [Medline: [40472585](https://pubmed.ncbi.nlm.nih.gov/40472585/)]
96. Tremblay M, Lafleur A, Leppink J, Dolmans DHJM. The simulated clinical environment: cognitive and emotional impact among undergraduates. *Med Teach*. 2017;39(2):181-187. [doi: [10.1080/0142159X.2016.1246710](https://doi.org/10.1080/0142159X.2016.1246710)] [Medline: [27832706](https://pubmed.ncbi.nlm.nih.gov/27832706/)]
97. Tremblay ML, Rethans JJ, Dolmans D. Task complexity and cognitive load in simulation-based education: a randomised trial. *Med Educ*. 2023;57(2):161-169. [doi: [10.1111/medu.14941](https://doi.org/10.1111/medu.14941)] [Medline: [36151727](https://pubmed.ncbi.nlm.nih.gov/36151727/)]
98. Al-Ghareeb AZ, Cooper SJ, McKenna LG. Anxiety and clinical performance in simulated setting in undergraduate health professionals education: an integrative review. *Clin Simul Nurs*. 2017;13(10):478-491. [doi: [10.1016/j.ecns.2017.05.015](https://doi.org/10.1016/j.ecns.2017.05.015)]
99. Lesselroth B, Monkman H, Bellavia K, Anderson M, van Buren J, Snider K, et al. Methods for usability testing an educational simulation: a case-study using a maternal telehealth scenario. *Stud Health Technol Inform*. 2026;336:2125-2129. [doi: [10.3233/SHTI260636](https://doi.org/10.3233/SHTI260636)] [Medline: [42175304](https://pubmed.ncbi.nlm.nih.gov/42175304/)]
100. Nielsen J. Discount usability: 20 years. Nielsen Norman Group. 2009. URL: <https://www.nngroup.com/articles/discount-usability-20-years/> [accessed 2009-02-06]
101. Sevilla-Gonzalez MDR, Moreno Loaeza L, Lazaro-Carrera LS, Bourguet Ramirez B, Vázquez Rodríguez A, Peralta-Pedrero ML, et al. Spanish version of the System Usability Scale for the assessment of electronic tools: development and validation. *JMIR Hum Factors*. 2020;7(4):e21161. [FREE Full text] [doi: [10.2196/21161](https://doi.org/10.2196/21161)] [Medline: [33325828](https://pubmed.ncbi.nlm.nih.gov/33325828/)]
102. Wang Y, Lei T, Liu X. Chinese system usability scale: Translation, revision, psychological measurement. *International Journal of Human-Computer Interaction*. 2019;36(10):953-963. [doi: [10.1080/10447318.2019.1700644](https://doi.org/10.1080/10447318.2019.1700644)]
103. Xiao YM, Wang ZM, Wang MZ, Lan YJ. The appraisal of reliability and validity of subjective workload assessment technique and NASA-task load index. *Zhonghua Lao Dong Wei Sheng Zhi Ye Bing Za Zhi*. 2005;23(3):178-181. [Medline: [16124892](https://pubmed.ncbi.nlm.nih.gov/16124892/)]
104. Rubio S, Díaz E, Martín J, Puente JM. Evaluation of subjective mental workload: a comparison of SWAT, NASA-TLX, and workload profile methods. *Appl Psychol*. 2004;53(1):61-86. [doi: [10.1111/j.1464-0597.2004.00161.x](https://doi.org/10.1111/j.1464-0597.2004.00161.x)]
105. Kwon H, Yoou S. Validation of a Korean version of the satisfaction with simulation experience scale for paramedic students. *Korean J Emerg Med Serv*. 2014;18(2):7-20. [FREE Full text] [doi: [10.14408/kjems.2014.18.2.007](https://doi.org/10.14408/kjems.2014.18.2.007)]
106. Smrekar M, Ledinski Fičko S, Kurtović B, Ilić B, Čukljek S, Tomac M, et al. Translation and validation of the Satisfaction with Simulation Experience scale: cross-sectional study. *Cent Eur J Nurs Midw*. 2022;13(2):633-639. [doi: [10.15452/cejnm.2021.12.0038](https://doi.org/10.15452/cejnm.2021.12.0038)]
107. Clemmensen T. Templates for cross-cultural and culturally specific usability testing: results from field studies and ethnographic interviewing in three countries. *Int J Hum Comput Interact*. 2011;27(7):634-669. [doi: [10.1080/10447318.2011.555303](https://doi.org/10.1080/10447318.2011.555303)]
108. Lewis JR, Sauro J. Ten things to know about the RITE method. MeasuringU. URL: <https://measuringu.com/rite-method/> [accessed 2026-02-06]
109. Brunzini A, Papetti A, Messi D, Germani M. A comprehensive method to design and assess mixed reality simulations. *Virtual Real*. 2022;26(4):1257-1275. [doi: [10.1007/s10055-022-00632-8](https://doi.org/10.1007/s10055-022-00632-8)]
110. Birdsall S, McAlpin E, Grandhi U. Evaluating the usability and learning effectiveness of a virtual reality patient simulation for training nursing students: A pilot study. *Teach Learn Nurs*. 2025;20(4):e1021-e1028. [doi: [10.1016/j.teln.2025.04.017](https://doi.org/10.1016/j.teln.2025.04.017)]
111. Birrenbach T, Stuber R, Müller CE, Sutter P, Hautz WE, Exadaktylos AK, et al. Virtual reality simulation to enhance advanced trauma life support trainings: a randomized controlled trial. *BMC Med Educ*. 2024;24(1):666. [FREE Full text] [doi: [10.1186/s12909-024-05645-2](https://doi.org/10.1186/s12909-024-05645-2)] [Medline: [38886688](https://pubmed.ncbi.nlm.nih.gov/38886688/)]
112. Mørk G, Bonsaksen T, Larsen OS, Kunnikoff HM, Lie SS. Virtual reality simulation in undergraduate health care education programs: usability study. *JMIR Med Educ*. 2024;10:e56844. [FREE Full text] [doi: [10.2196/56844](https://doi.org/10.2196/56844)] [Medline: [39560982](https://pubmed.ncbi.nlm.nih.gov/39560982/)]
113. Rovati L, Gary PJ, Cubro E, Dong Y, Kilickaya O, Schulte PJ, et al. Development and usability testing of a patient digital twin for critical care education: a mixed methods study. *Front Med (Lausanne)*. 2023;10:1336897. [FREE Full text] [doi: [10.3389/fmed.2023.1336897](https://doi.org/10.3389/fmed.2023.1336897)] [Medline: [38274456](https://pubmed.ncbi.nlm.nih.gov/38274456/)]

114. Lee Y, Kim SK, Eom M. Usability of mental illness simulation involving scenarios with patients with schizophrenia via immersive virtual reality: a mixed methods study. PLoS One. 2020;15(9):e0238437. [FREE Full text] [doi: [10.1371/journal.pone.0238437](https://doi.org/10.1371/journal.pone.0238437)] [Medline: [32936813](https://pubmed.ncbi.nlm.nih.gov/32936813/)]
115. Jeffries PR, Rizzolo MA. Designing and implementing models for the innovative use of simulation to teach nursing care of ill adults and children: a national, multi-site, multi-method study. NLN/Laerdal Project Summary Report. 2006. URL: <https://www.nln.org/docs/default-source/uploadedfiles/professional-development-programs/read-the-nln-laerdal-project-summary-report-pdf.pdf> [accessed 2026-08-14]
116. Keenan HL, Duke SL, Wharrad HJ, Doody GA, Patel RS. Usability: an introduction to and literature review of usability testing for educational resources in radiation oncology. Tech Innov Patient Support Radiat Oncol. 2022;24:67-72. [FREE Full text] [doi: [10.1016/j.tipsro.2022.09.001](https://doi.org/10.1016/j.tipsro.2022.09.001)] [Medline: [36262375](https://pubmed.ncbi.nlm.nih.gov/36262375/)]
117. Johnson SG, Potrebny T, Larun L, Ciliska D, Olsen NR. Usability methods and attributes reported in usability studies of mobile apps for health care education: scoping review. JMIR Med Educ. 2022;8(2):e38259. [FREE Full text] [doi: [10.2196/38259](https://doi.org/10.2196/38259)] [Medline: [35767323](https://pubmed.ncbi.nlm.nih.gov/35767323/)]
118. Nielsen J. Why you only need to test with 5 users. Nielsen Norman Group. 2000. URL: <https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/> [accessed 2026-08-14]
119. Nielsen J, Landauer TK. A mathematical model of the finding of usability problems. 1993. Presented at: INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems; April 24-29, 1993:206-213; Amsterdam, Netherlands. [doi: [10.1145/169059.169166](https://doi.org/10.1145/169059.169166)]
120. Bevan N, Barnum C, Cockton G, Nielsen J, Spool J, Wixon D. The “magic number 5”: is it enough for web testing? 2003. Presented at: CHI '03 Extended Abstracts on Human Factors in Computing Systems; April 5-10, 2003:698-699; Fort Lauderdale, Florida. [doi: [10.1145/765891.765936](https://doi.org/10.1145/765891.765936)]
121. Sauro J. A brief history of the magic number 5 in usability testing. MeasuringU. URL: <https://measuringu.com/five-history/> [accessed 2026-08-14]
122. Gwet KL. Handbook of inter-rater reliability: the definitive guide to measuring the extent of agreement among raters. Gaithersburg, Maryland. Advanced Analytics, LLC; 2014.

Abbreviations

AAMC: Association of American Medical Colleges
AR: augmented reality
ATA: Agile Task Analysis
EHR: electronic health record
FCM: family and community medicine
HTA: hierarchical task analysis
NASA: National Aeronautics and Space Administration
NASA-TLX: NASA Task Load Index
OB/GYN: obstetrics and gynecology
OUSCM: University of Oklahoma School of Community Medicine
SDS: Simulation Design Scale
SEQ: Single Ease Question
SP: standardized patient
SSES: Satisfaction With Simulation Experience Scale
SUS: System Usability Scale
VR: virtual reality

Edited by A Stone; submitted 29.May.2026; peer-reviewed by L-C Wetzlmair, K Grewal; comments to author 01.Jul.2026; revised version received 22.Jul.2026; accepted 27.Jul.2026; published 09.Sep.2026

Please cite as:

Lesselroth BJ, Douglas A, Monkman H, Marcilly R, Bellavia K, Anderson M, Van Buren J, Parsons C, Mohanty S, Homco J, Richards M, Gold KP

How to Conduct Usability Testing of Medical Education Simulations: Methods Tutorial Using a Telemedicine Simulation for Maternal Care

JMIR Med Educ 2026;12:e102697

URL: <https://mededu.jmir.org/2026/1/e102697>

doi: [10.2196/102697](https://doi.org/10.2196/102697)

PMID:

©Blake J Lesselroth, Alexander Douglas, Helen Monkman, Romaric Marcilly, Karalane Bellavia, Mataeo Anderson, Jacob Van Buren, Charles Parsons, Sonakshi Mohanty, Juell Homco, Morgan Richards, Karen P Gold. Originally published in JMIR Medical Education (<https://mededu.jmir.org>), 09.Sep.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on <https://mededu.jmir.org/>, as well as this copyright and license information must be included.