2023 APDS SPRING MEETING

Usability of ENTRUST as an Assessment Tool for Entrustable Professional Activities (EPAs): A Mixed Methods Analysis

Melissa C. Lee, BA,[†] Edward F. Melcer, PhD,[‡] Sylvia Bereknyei Merrell, DrPH, MS,[§] Lye-Yeng Wong, MD,[‖] Samuel Shields, MS,[‡] Hyrum Eddington, BS,[¶] Amber W. Trickey, PhD, MS, CPH,[¶] Jason Tsai, MS,[‡,††] James R. Korndorffer, Jr., MD, MHPE, FACS,[††,‡‡] Dana T. Lin, MD, FACS,[††,*] and Cara A. Liebert, MD, FACS[††,‡‡,*]

[†]Stanford University School of Medicine, Stanford, California; [‡]Department of Computational Media, University of California-Santa Cruz, Baskin School of Engineering, Santa Cruz, California; [§]Department of Pediatrics, Stanford University School of Medicine, Stanford, California; [‖]Department of Cardiothoracic Surgery, Stanford University School of Medicine, Stanford, California; [¶]Stanford-Surgery Policy Improvement Research and Education Center (S-SPIRE), Palo Alto, California; [††]Department of Surgery, Stanford University School of Medicine, Stanford, California; and [‡‡]VA Palo Alto Health Care System, Surgical Services, Palo Alto, California

OBJECTIVE: As the American Board of Surgery transitions to a competency-based model of surgical education centered upon entrustable professional activities (EPAs), there is a growing need for objective tools to determine readiness for entrustment. This study evaluates the usability of ENTRUST, an innovative virtual patient simulation platform to assess surgical trainees' decision-making skills in preoperative, intra-operative, and post-operative settings.

DESIGN: This is a mixed-methods analysis of the usability of the ENTRUST platform. Quantitative data was collected using the system usability scale (SUS) and Likert responses. Analysis was performed with descriptive statistics, bivariate analysis, and multivariable linear regression. Qualitative analysis of open-ended responses was performed using the Nielsen-Shneiderman Heuristics framework.

SETTING: This study was conducted at an academic institution in a proctored exam setting.

PARTICIPANTS: The analysis includes n = 47 (PGY 1-5) surgical residents who completed an online usability survey following the ENTRUST Inguinal Hernia EPA Assessment.

RESULTS: The ENTRUST platform had a median SUS score of 82.5. On bivariate and multivariate analyses, there were no significant differences between usability based on demographic characteristics (all $p > 0.05$), and SUS score was independent of ENTRUST performance ($r = 0.198$, $p = 0.18$). Most participants agreed that the clinical workup of the patient was engaging (91.5%) and felt realistic (85.1%). The most frequent heuristics represented in the qualitative analysis included *feedback*, *visibility*, *match*, and *control*. Additional themes of *educational value*, *enjoyment*, and *ease-of*-use highlighted participants' perspectives on the usability of ENTRUST.

CONCLUSIONS: ENTRUST demonstrates high usability in this population. Usability was independent of ENTRUST score performance and there were no differences in usability identified in this analysis based on demographic subgroups. Qualitative analysis highlighted the acceptability of ENTRUST and will inform ongoing development of the platform. The ENTRUST platform holds potential as a tool for the assessment of EPAs in surgical residency programs. (J Surg Ed 000:1−10. Published by Elsevier Inc. on behalf of Association of Program Directors in Surgery.)

ABBREVIATIONS: ABS, American Board of Surgery EPA, Entrustable Professional Activity PGY, Post-Graduate Year SUS, System Usability Scale UIM, Underrepresented in Medicine

*Co-Senior authors

Correspondence: Inquiries to Cara A. Liebert, MD, FACS, VA Palo Alto Health Care System, Surgical Services, 3801 Miranda Avenue, Palo Alto, CA 94304; e-mail: cara.liebert@stanford.edu

KEY WORDS: assessment, clinical decision-making, entrustable professional activity, simulation, serious game, usability

COMPETENCIES: Medical Knowledge, Patient Care, Interpersonal and Communication Skills

INTRODUCTION

The paradigm of surgical education and assessment of trainees is shifting towards a competency-based model.[1] This shift in framework has been accelerated by the American Board of Surgery's plan to implement entrustable professional activities (EPAs), or units of professional practice that constitute what clinicians do as daily work, as a foundation for assessment of surgical trainees.[2] As such, there is a growing need for objective, unbiased, easy-to-use, and scalable methods to assess a trainees' surgical decision-making skills and readiness for entrustment.

The implementation of EPAs in surgical education provides an opportunity to evaluate and address the limitations of traditional observation-based assessment of trainees, namely that such assessments can be difficult to scale and may introduce opportunities for implicit bias. Indeed, multiple studies have demonstrated evidence for concern regarding bias in traditional assessments in graduate medical training.[3-8] Furthermore, as EPAs expand, the required increase in resources to equitably meet the demand for assessments calls for efficient and scalable solutions to measure each trainee's competency with fidelity and to assess infrequently encountered but critical clinical scenarios.

Serious games are a growing field that addresses the aforementioned limitations of traditional assessments in surgical education. Serious games are games developed in which entertainment is not the primary goal.[9] In a similar vein to the fast-developing landscape of surgical education, the field of serious games is gaining timely acceptance as a tool for medical education, especially with increasing technical literacy among medical trainees.[9-11] With relation to medical education and assessment, it is notable in its potential to be an unbiased, scalable solution with minimal demand for time and resources from the evaluator.[12] As such, serious games as a mode of delivery for educational content is an innovative and promising approach to provide trainees with a safe environment to optimize knowledge and skills before being entrusted with real patients.[13]

In response to this evolving demand, our team developed an innovative, online, virtual patient simulation platform called ENTRUST. The ENTRUST Assessment Platform is a serious game that virtually simulates a patient encounter in the preoperative, intraoperative, and postoperative settings to objectively assess a surgical trainee's clinical decision-making competence.[14] During the Simulation Phase, examinees must complete a physical examination and preoperative workup of the patient while ordering and interpreting laboratory studies and imaging. Examinees can order fluids, medications and procedures, to which the patient's vital signs will dynamically respond. The subsequent Question Phase provides examinees with multiple choice questions that assess knowledge on pre-operative planning, intra-operative decision-making, and postoperative care.

We previously reported initial validity evidence for an ENTRUST Inguinal Hernia EPA Assessment comprised of 4 case scenarios deployed on the platform.[15] Performance was positively correlated to post-graduate year (PGY) level and prior inguinal hernia operative experience.[15] The ENTRUST Assessment Platform has also been piloted in the Membership of the College of Surgeons (MCS) Examination in the College of Surgeons of East, Central, and Southern Africa (COSECSA), which demonstrated a strong correlation of performance on ENTRUST with traditional oral objective structured clinical examinations (OSCE) in a high-stakes exam setting.[16] To further explore the acceptability and collect response process validity evidence for the ENTRUST platform, this mixed-methods study aims to assess the usability of ENTRUST Assessment Platform and evaluate for potential bias in the usability towards specific subgroups.

MATERIALS AND METHODS

Participants and Setting

In total, n = 85 surgical residents at a single institution were invited to voluntarily complete this pilot study, with n = 52 residents (61% response rate) participating. The study was completed in a 1-hour group session during residents' protected education time on a weekday morning, in a classroom setting on laptop computers, proctored by members of the study team. Participants were instructed that the purpose of the study was to pilot and evaluate the usability of the ENTRUST platform and identify opportunities for improvement. Feedback was not provided to residents on their performance. The ENTRUST Inguinal Hernia EPA assessment and online Qualtrics usability survey were completed by n = 51 surgical residents in May 2021 or May 2022; n = 1 participant completed the ENTRUST Inguinal Hernia EPA assessment but did not complete the Qualtrics usability survey and was not included in the study. All

participants were first-time users of ENTRUST and completed the study at the same time point in the academic year. A CONSORT diagram is included as Supplemental Figure 1.

Participants viewed a standardized video tutorial orientation to the ENTRUST platform, followed by a non-scored practice case that gave participants unlimited time to explore the platform and its functionalities. Participants then completed the ENTRUST Inguinal Hernia EPA Assessment, which consisted of 4 total case scenarios. Following the ENTRUST assessment, an online survey regarding the usability of the platform was administered. To ensure confidentiality, data collected in the ENTRUST assessment, demographic survey, and usability survey only contained de-identified information, connected by a unique participant ID that each participant received. The study protocol (#53137) was reviewed and approved by the Institutional Review Board.

Survey

As part of the ENTRUST assessment, participants provided demographic data, including age, sex, race/ethnicity, English language proficiency, training level, surgical subspecialty, prior video game experience, and inguinal hernia operative case volume. Following completion of the ENTRUST Inguinal Hernia Assessment, participants completed an online questionnaire querying participants on the user experience of the ENTRUST Assessment Platform which was developed in Qualtrics (Provo, UT). The survey instrument included 10 standardized questions from the System Usability Scale (SUS), an established research tool that was selected for its widespread use as a standard to measure the usability of software and hardware systems (Fig. 1).[17] The SUS score ranges from 0 to 100, with scores <50 considered "unacceptable," 50 to 70 considered "marginal," >70 considered "good," and >85 considered "excellent."[17] SUS has previously been used to evaluate multiple well-established mainstream software and hardware platforms, such as Amazon, iPhone, and Microsoft Teams, which were found to have SUS scores of 81.8, 78.5, and 77.2, respectively.[18] SUS has also previously been used to evaluate comparable serious games in medical education.[19-21]

In addition to the 10 SUS questions, the survey included 10 additional Likert scale questions developed by the study team to further query the acceptability and usability of the ENTRUST Assessment Platform, with responses ranging from 1 to 5 (1 = Strongly disagree or very difficult, 5 = Strongly agree or very easy). The survey concluded with the following 3 open-ended responses for formative feedback on the platform: *what, if any, issues/problems did you encounter while completing the ENTRUST Assessment; suggestions for improving the ENTRUST Assessment Platform*; and *additional comments or feedback*.

Quantitative Analysis

Participants were excluded from the data analysis if survey responses had no variance (n = 1), the survey was completed in less than 60 seconds (n = 2), or demographic data was missing (n = 1). Lack of variance and completion in less than 60 seconds were chosen as indications of straight-lining or poor attention, respectively. Data queries, database creation, and statistical analysis were performed by members of the study team (JT, ML, HE) with no interaction with the participants to maintain confidentiality of results and avoid the possibility of responses being linked to individuals based on identifiable demographics.

Descriptive SUS statistics were calculated for the cohort, including mean, median, and interquartile range. Descriptive statistics for the Likert scale questions were calculated for each question. Relationships between SUS and subgroup characteristics were analyzed. Race/ethnicity was collapsed into dichotomous categories, Underrepresented in Medicine (UIM) and non-UIM, as defined by the AAMC.[22] Participants who completed the study during their research years (n = 4) were instructed to select the PGY-level they completed prior to their research time. PGY-1 and PGY-2 were categorized as junior residents, while PGY-3, -4, and -5 were categorized as senior residents. Video game experience was dichotomized as "none" or "some," and inguinal hernia repair count was categorized as >5 or ≤5 cases to represent basic familiarity with the procedure.

To evaluate the distribution of continuous data, a Shapiro-Wilks test of normality was performed for the SUS scores, which was $p = 0.051$, suggesting normality. However, given this borderline value, both parametric and nonparametric bivariate analyses were performed. To further investigate whether a participants' ability to use the platform was associated with their performance, correlations between ENTRUST scores and SUS score were investigated using Pearson correlations and Spearman rank correlations. In addition to Grand Total Score, sub-scores for each individual case and each phase of the cases (Simulation Phase and Question Phase) were investigated for potential correlations. Simple linear regression models were created for bivariate exploration of SUS and each variable. Parametric and nonparametric analyses of SUS score for categorical demographic variables were performed with t-test/ANOVA and Wilcoxon/Kruskal-Wallis tests, respectively. To assess independent associations between subgroup characteristics and SUS score, we created a multivariate

		Strongly Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Strongly Agree
		1	2	3	4	5
1	I think I would like to use ENTRUST frequently.	O	O	O	O	O
2	I found ENTRUST unnecessarily complex.	O	O	O	O	O
3	I thought ENTRUST was easy to use.	O	O	O	O	O
4	I think that I would need the support of a technical person to be able to use ENTRUST.	O	O	O	O	O
5	I found the various functions in ENTRUST were well integrated.	O	O	O	O	O
6	I thought there was too much inconsistency in ENTRUST.	O	O	O	O	O
7	I would imagine that medical trainees and physicians would learn to use ENTRUST very quickly.	O	O	O	O	O
8	I found ENTRUST very cumbersome to use.	O	O	O	O	O
9	I felt very confident using ENTRUST.	O	O	O	O	O
10	I needed to learn a lot of things before I could get going with ENTRUST.	O	O	O	O	O

FIGURE 1. ENTRUST Assessment Platform System Usability Scale (SUS) 10-Item Survey.

model including the following variables: age, sex, UIM status, PGY-level, and inguinal hernia repair cases. For all comparisons, $p < 0.05$ was defined as statistically significant. Analyses were performed using Jupyter Notebook, running Python 3.6.1.[23,24]

Qualitative Analysis

To gain further insight into the user experience of the ENTRUST Assessment Platform in this population, qualitative analysis was conducted to analyze the open-ended responses that queried users regarding any issues/problems encountered while completing ENTRUST or suggestions for improving the platform. Of the participants included in the study, n = 29 (62%) provided qualitative responses for analysis. The responses were de-identified, and the unit of analysis was determined as the response-level. Codes were defined deductively from the elements of the Nielsen-Shneiderman usability heuristics framework, which features twelve heuristics used in the development of functional user interfaces[25,26]; this framework was selected for its widespread use in the evaluation of user experience of systems.[25,26] Additional codes present in the responses were developed inductively according to emerging themes from the data until consensus was reached on a final codebook. The data was coded independently by 2 members of the research team (ML and SBM) in an iterative fashion, followed by adjudication (CL, DL, EM) until 100% agreement was reached. All members of the qualitative analysis research team had training and expertise in qualitative methodology.

RESULTS

Demographics and Subgroup Characteristics

A total of n = 47 surgical residents were included in the analysis (Table 1), with all PGY levels 1 to 5 represented in the study population. The mean age was 30.6 (2.9), 57.4% identified as female, 38.3% identified as White, 36.2% as Asian, and 12.8% as Black or African American. Self-reported video game experience prior to medical training ranged from 0 to 15 hours per week, with a mean (SD) of 1.1 (2.7) hours; 66% of participants reported no experience with video games. Inguinal hernia operative case volume ranged from 0 to 85 cases, with mean (SD) 22.5 (22.8); 27.7% of participants reported 5 or fewer logged inguinal hernia cases.

System Usability Scale (SUS) Score

The mean (SD) SUS score among all participants was 80.0 (11.8), with a median of 82.5 (Fig. 2). Mean SUS scores for each subgroup are shown in Table 1. Univariate linear regressions modeling the relationship between SUS and each variable showed no statistically significant relationships based on demographic or other subgroup characteristics. Parametric and nonparametric bivariate analyses of SUS scores yielded similar results and

TABLE 1. Demographics, Subgroup Variables, and System Usability Scale (SUS) Score

	n (%)	SUS Score mean (SD)	R Squared	p-value
Age, mean (SD)	30.6 (2.9)*	-	0.053	0.12
Sex			0.009	
Male (Reference)	20 (42.6)	81.4 (12.7)		-
Female	27 (57.4)	79.2 (11.4)		0.54
PGY level			0.143	
PGY 1-2 (Reference)	29 (61.7)	80.8 (10.6)		-
PGY 3-5	18 (38.3)	79.0 (14.0)		0.63
UIM status			0.003	
Non-UIM (Reference)	42 (89.4)	79.9 (12.1)		-
UIM	5 (10.6)	82.0 (10.8)		0.71
English proficiency			0.002	
Native or bilingual proficiency (Reference)	42 (89.4)	79.9 (12.3)		-
Full professional proficiency	5 (10.6)	81.5 (9.5)		0.79
Specialty			0.037	
General Surgery Categorical (Reference)	33 (70.2)	80.4 (12.3)		-
Designated preliminary†	8 (17.0)	79.5 (9.9)		0.88
Nondesignated preliminary	6 (12.8)	81.3 (12.5)		0.86
Prior video game experience (hours/week), mean (SD)	1.1 (2.7)*	-	0.024	0.30
No experience (Reference)	31 (66.0)	80.9 (12.0)		-
Some experience	16 (34.0)	78.6 (12.0)		0.54
Inguinal hernia operative case volume, mean (SD)	22.5 (22.8)*	-	0.000	0.90
5 or fewer cases (Reference)	13 (27.7)	79.6 (8.6)		-
Greater than 5 cases	34 (72.3)	80.3 (13.1)		0.87

PGY, Post-Graduate Year; SD, standard deviation; UIM, Underrepresented in Medicine.
†Includes PGY-1 or PGY-2 orthopedic surgery, otolaryngology, plastic surgery, urology, vascular surgery.
*mean (SD).

interpretations for all analyses. The parametric results are included in Table 1, and both parametric and non-parametric analyses are included in Supplemental Table 1 and Supplemental Table 2. A multivariate model with variables of interest was created to further evaluate for bias in usability towards specific subgroups, and is summarized in Table 2.

Relationship Between Usability and ENTRUST Performance

Correlations were performed for ENTRUST Grand Total Score, Simulation Phase Total Score, Question Phase Total Score, as well as case score on each of the 4 cases. There were no statistically significant correlations

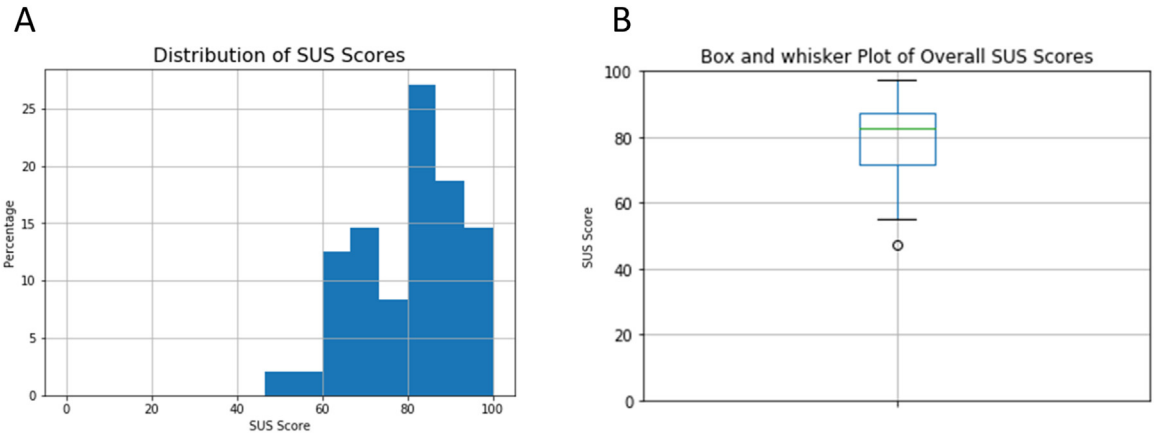

FIGURE 2. ENTRUST Assessment Platform System Usability Scale (SUS) Scores. (A) Histogram of SUS scores; (B) Box and Whisker plot of SUS scores showing the maximum (97.5), third quartile (87.5), median (82.5), first quartile (71.9), and minimum (47.5).

TABLE 2. Multivariate Analysis of ENTRUST System Usability Scale (SUS) Score

Variable	Coefficient [97.5% CI]	p-value
Age	−0.89 [−2.73, 0.95]	0.33
UIM Status (UIM)*	−3.44 [−16.40, 9.51]	0.59
Sex (Female)*	−4.59 [−11.80, 2.63]	0.21
PGY-Level (PGY 3-5)*	1.82 [−8.41, 12.05]	0.72
Inguinal Hernia Repairs Performed	0.04 [−0.17, 0.24]	0.72

*Reference group.

between participants' SUS score and their ENTRUST subscores or total score, indicating that the participants' demonstration of decision-making skills was not related to how usable they found the platform (Fig. 3).

Descriptive Usability Questions

Results of participants' Likert scale responses are summarized in Table 3. Values are reported as percentages and mean (SD). Most participants found the platform to be naturally intuitive, with 80.9% and 78.7% indicating "Very Easy" on the Likert scale for reading and understanding the cases, respectively. Regarding ordering interventions such as labs, imaging, fluids, and medications, the participants answered similarly, with >70% indicating these tasks were "Very Easy"; 74.5% of participants answered "Strongly Agree" to the statement that the video tutorial was helpful. Similarly, 85.1% and 91.5% of participants somewhat or strongly agreed with the clinical workup realism and engagement, respectively.

Qualitative Analysis

Deductive Coding

Definitions of each code in the heuristics framework are summarized in Table 4. Of these heuristics, the most frequent codes were: *feedback, visibility*, *match*, and *control. Language* and *message* heuristics were not represented in the participant responses. Example quotes of the most frequent themes from participant's suggestions for improvement of the ENTRUST platform are summarized in Table 5.

Inductive Analysis

Inductive coding was performed for additional comments or feedback not represented by the deductive codebook, and yielded 3 additional themes of *educational value*, *enjoyment*, and *ease of use*, as summarized in Table 6.

DISCUSSION

This mixed-methods study of the ENTRUST Assessment Platform demonstrated good usability and favorable user feedback, providing response process validity evidence for its use an assessment of clinical decision-making in this population. Several serious games that aim to improve medical decision-making have been developed and evaluated using the SUS scale, with scores ranging from 55 to 79.[19-21] The mean SUS score of 80.0 for the ENTRUST Assessment Platform in this study indicates high usability in this population, and places ENTRUST as having a similar or higher SUS score to comparable serious games in medical education. Participants reported ease of use in the functionality of ENTRUST, including understanding the patient's clinical state, performing a physical exam, and ordering relevant laboratory studies, imaging, fluid, and medications. The majority of the users agreed the video tutorial was helpful and felt that the clinical workup was realistic and engaging.

Bivariate analysis demonstrated no significant difference in SUS scores between key subgroups, indicating no observed bias in the usability of the ENTRUST platform with regards to age, sex, race/ethnicity, PGY-level, or surgical subspecialty in this limited sample size. There was no significant difference in usability between

FIGURE 3. Relationship between SUS score and ENTRUST score performance. Scatter plots with Spearman rank correlation coefficients (rho) and p-values representing relationships between SUS and ENTRUST Grand Total Score (A), ENTRUST Simulation Phase Total Score (B), and Question Phase Total Score (C).

TABLE 3. Participant Responses to Descriptive Usability Questions

Question	Very Difficult % (n) 1	Somewhat Difficult % (n) 2	Neither Easy nor Difficult % (n) 3	Somewhat Easy % (n) 4	Very Easy % (n) 5	Mean (SD)
I was able to read the text	0.0% (0)	0.0% (0)	4.3% (2)	14.9% (7)	80.9% (38)	4.77 (0.52)
I was able to understand the patient's clinical state	0.0% (0)	2.1% (1)	2.1% (1)	17.0% (8)	78.7% (37)	4.72 (0.62)
I was able to perform a physical exam	0.0% (0)	4.3% (2)	4.3% (2)	4.3% (2)	87.2% (41)	4.74 (0.74)
I was able to order laboratory studies	0.0% (0)	2.1% (1)	4.3% (2)	10.6% (5)	83.0% (39)	4.74 (0.64)
I was able to order imaging studies	0.0% (0)	2.1% (1)	4.3% (2)	10.6% (5)	83.0% (39)	4.74 (0.64)
I was able to order fluids	0.0% (0)	4.3% (2)	2.1% (1)	12.8% (6)	80.9% (38)	4.70 (0.72)
I was able to order medications	0.0% (0)	6.4% (3)	10.6% (5)	8.5% (4)	74.5% (35)	4.51 (0.93)
Question	**Strongly Disagree % (n)**	**Somewhat Disagree % (n)**	**Neither Agree nor Disagree % (n)**	**Somewhat Agree % (n)**	**Strongly Agree % (n)**	**Mean (SD)**
The tutorial was helpful	0.0% (0)	2.1% (1)	4.3% (2)	19.1% (9)	74.5% (35)	4.66 (0.67)
The clinical workup of the patient felt realistic	2.1% (1)	2.1% (1)	10.6% (5)	53.2% (25)	31.9% (15)	4.11 (0.84)
The clinical workup of the patient was engaging	4.2% (2)	2.1% (1)	2.1% (1)	40.4% (19)	51.1% (24)	4.32 (0.96)

English native language speakers and non-native speakers with full professional proficiency. The subsequent multivariable model similarly found no significant relationships between ENTRUST usability and specific group characteristics. In post hoc power analysis, this initial pilot study was appropriately powered to detect major differences in usability based on demographic factors such as sex or English proficiency, but was not powered to detect minor differences in SUS scores between these groups. Thus, ongoing studies with larger sample sizes are needed to continue to evaluate for bias in the usability of the platform across populations.

The findings that video game experience and hernia repair operative case volume were not associated with usability are also noteworthy. Our initial hypotheses considered that participants less comfortable with the setting of video games, or less familiar with inguinal hernia repair, could have higher cognitive strain using the platform which could impact usability. Bivariate analyses of both variables as continuous and binary demonstrated

TABLE 4. Deductive Codebook Using Nielsen-Shneiderman Heuristics Framework

Heuristic	Code Frequency n (%)	Definition
Feedback	11 (37.9)	The system provides users with appropriate and specific feedback for actions.
Visibility	10 (34.5)	Users are able to see all information and options presented on the screen, identify, and select them easily.
Match	8 (27.6)	The system's interface has intuitive design that aligned with their expectation of the given setting.
Control	6 (20.7)	Users are the initiators of actions; the system avoids surprising actions and unexpected outcomes.
Closure	5 (17.2)	Users are clear on when a certain action was started, in progress, and completed. Users are clear on when a next action was able to be selected.
Minimalist	3 (10.3)	The system's interface contains minimal extraneous information.
Flexibility	3 (10.3)	Users were able to manipulate the system to be more usable to them.
Document	2 (6.9)	The user has a method to obtain assistance in using the system.
Memory	2 (6.9)	Users were able to carry out tasks with minimal memory load.
Error	1 (3.4)	The system is designed to prevent errors and mitigate them if they occur.
Undo	1 (3.4)	Users are able to recover from errors.
Consistency	1 (3.4)	The system uses standards and conventions in product design.
Language	0 (0)	The language is presented in a form that is understandable to the users.
Message	0 (0)	The system provides users with informative and useful error messages.

TABLE 5. Suggestions for Improvement of the ENTRUST Platform

Theme	Frequency n (%)	Example Quotes
Feedback	11 (37.9)	*"It would be helpful to get some feedback on my performance at the end of each case or at the end of the assessment (good learning opportunity)."* *"Feedback after the case would be nice. At least something like 'here is the optimal way to manage this, and here is how you managed it.'"*
Visibility	10 (34.5)	*"The [video tutorial] could walk briefly through or point out some of the studies that you can order. I know there was time to look around with one practice patient but maybe the [tutorial] could point some of these specific options out."*
Match	8 (27.6)	*"When calling a pre-op consult, I called cardiology but I would have also liked the patient to be evaluated by preop anesthesia. That was not an option though. I would consider adding that."* *"Just more cases and maybe more steps to make it realistic to the hospital."*
Control	6 (20.7)	*"Was only prompted once if I wanted to reduce the hernia. Wanted to get labs before saying yes/no but was not possible."*

no associations between SUS and video game experience or hernia repair count. These results suggest that neither a trainee's familiarity with video game platforms nor their familiarity with the tested content were relevant factors in how usable the participants found ENTRUST.

The analysis found no relationship between SUS score and ENTRUST performance. This provides additional validity evidence that higher ENTRUST score performance reflects higher knowledge and surgical decision-making skills, rather than differences in usability.

While the quantitative analysis demonstrated high usability and acceptability, we performed qualitative analysis to obtain formative user feedback on the usability of the platform to guide ongoing platform development. Incorporating methodologies used in the field of user-interface design, we utilized the Nielsen-Shneiderman heuristic framework to provide insight into any issues encountered by users and suggestions for improvement of the ENTRUST platform. The most common heuristics were *feedback*, *visibility*, *match*, and *control*. Participants in this study population highly valued control and flow of clinical care within the platform to match reality as closely as possible. Our findings highlight the importance of ENTRUST to accurately mimic the progression of surgical care in an actual clinical environment and provide users with controls to seamlessly perform actions as they would in real life. Based on this feedback, we have further developed the video tutorial to an interactive tutorial and expanded the menu of potential studies, interventions, and consults.

Qualitative analysis also identified interest from trainees for ENTRUST to be further developed as a learning platform to practice surgical decision-making skills and provide feedback on their performance. In response to this, we have begun development on an ENTRUST Learning Platform with embedded feedback. The use of simulation in surgical education is a widely accepted means of technical skills acquisition, especially with regard to procedures such as laparoscopy.[27] However, as a case-based virtual patient platform, ENTRUST is distinct from traditional simulation-based education, as it simulates the diagnostic workup, stabilization, and management of patients with surgical conditions. In this way, ENTRUST emphasizes the cognitive exercise of surgical decision-making— a critical but often difficult to evaluate skill that is imperative for trainees to demonstrate as they progress to higher levels of entrustment.

Potential Impact

As a field, medical education and training are increasingly exploring and accepting the use of digital solutions

TABLE 6. Inductive Thematic Analysis of Additional Comments

Theme	Frequency n (%)	Example Quotes
Educational value	7 (24.1)	*"Would love to see an 'education' mode, with notifications of right vs wrong answers, and explanations for answers."*
Enjoyment	4 (13.8)	*"[I] think it is a great program to use for training. They should do this for all specialties because it is a great simulation."* *"Enjoyable and engaging"*
Ease of use	3 (10.3)	*"Very easy to use"* "Easy to use. Clear questions that are clinically relevant."

to meet the growing demand for standardized, equitable assessments. ENTRUST is a highly usable, objective, and scalable platform that has strong potential to meet the growing need to rigorously assess surgical trainees' clinical decision-making skills for competency-based medical education. In the context of EPAs, the ENTRUST Assessment Platform could be integrated with other microassessments to help guide entrustment decisions for trainees.

Limitations

Limitations of this study include the modest sample size of trainees with high computer literacy at a single institution. The limited sample size impacts the power of this study to detect minor biases in usability between groups. The proctors were members of the study team, and although confidentiality and deidentification were maintained during data collection, database preparation, and data analysis, it is possible that this unintentionally impacted participants' responses. Studies are ongoing to evaluate the usability of ENTRUST among users from culturally and technologically diverse backgrounds to ensure that the findings are generalizable across different populations.

CONCLUSION

Our study provides response process validity evidence on the usability of ENTRUST as an assessment tool for surgical decision-making. The ENTRUST Assessment Platform offers an objective and usable tool to meet the evolving needs of competency-based medical education.

ACKNOWLEDGMENTS

The authors would like to thank Fatyma Camacho, Oleksandra Keehl, Yulin Cai, Ruonan Chen, Ananya Anand, and Sherry Wren for their contributions to the development of the ENTRUST platform, and Kate Arnow for her valuable contributions to the revision of this manuscript.

FUNDING

This work was supported by the Mark Freidell Research Grant from the Association of Program Directors in Surgery, the Ilene B. Harris Legacy Research Fund from the Department of Medical Education at the University of Illinois at Chicago, the Stanford University School of Medicine Medical Scholars Research Program, and the Department of Surgery at Stanford University School of Medicine.

REFERENCES

1. Ten Cate O, Scheele F. Viewpoint: competency-based postgraduate training: can we bridge the gap between theory and clinical practice? *Acad Med*. 2007;82(6):542–547. https://doi.org/10.1097/ACM.0b013e31805559c7.

2. American Board of Surgery. ABS Announces Transition to Entrustable Professional Activities for General Surgery Resident Evaluation. Accessed October 12, 2022. Available at: https://www.absurgery.org/default.jsp?news_epas0222

3. Colbert CY, French JC, Herring ME, Dannefer EF. Fairness: the hidden challenge for competency-based postgraduate medical education programs. *Perspect Med Educ*. 2017;6(5):347–355. https://doi.org/10.1007/s40037-017-0359-8.

4. Lucey CR, Hauer KE, Boatright D, Fernandez A. Medical education's wicked problem: achieving equity in assessment for medical learners. *Academic Medicine*. 2020;95(12S):S98–S108. https://doi.org/10.1097/ACM.0000000000003717.

5. From the Gender Equity in Medicine (GEM) workgroup, Klein R, Julian KA, et al. Gender bias in resident assessment in graduate medical education: review of the literature. *J Gen Intern Med*. 2019;34(5):712–719. https://doi.org/10.1007/s11606-019-04884-0.

6. Yeates P, O'Neill P, Mann K, W Eva K. You're certainly relatively competent': assessor bias due to recent experiences. *Med Educ*. 2013;47(9):910–922. https://doi.org/10.1111/medu.12254.

7. Teherani A, Hauer KE, Fernandez A, King TE, Lucey C. How small differences in assessed clinical performance amplify to large differences in grades and awards: a cascade with serious consequences for students underrepresented in medicine. *Acad Med*. 2018;93(9):1286–1292. https://doi.org/10.1097/ACM.0000000000002323.

8. Mueller AS, Jenkins TM, Osborne M, Dayal A, O'Connor DM, Arora VM. Gender differences in attending physicians' feedback to residents: a qualitative analysis. *J Grad Med Educ*. 2017;9(5):577–585. https://doi.org/10.4300/JGME-D-17-00126.1.

9. Laamarti F, Eid M, El Saddik A. An overview of serious games. *Int J Comput Games Technol*. 2014: 1–15. https://doi.org/10.1155/2014/358152. 2014.

10. Kron FW, Gjerde CL, Sen A, Fetters MD. Medical student attitudes toward video games and related new

media technologies in medical education. *BMC Med Educ*. 2010;10(1):50. https://doi.org/10.1186/1472-6920-10-50.

11. Zairi I, Ben Dhiab M, Mzoughi K, Ben Mrad I, Kraiem S. Assessing medical student satisfaction and interest with serious game. *Tunis Med*. 2021;99 (11):1030–1035.

12. Gentry SV, Gauthier A, L'Estrade Ehrstrom B, et al. Serious gaming and gamification education in health professions: systematic review. *J Med Internet Res*. 2019;21(3):e12994. https://doi.org/10.2196/12994.

13. Graafland M, Schraagen JM, Schijven MP. Systematic review of serious games for medical education and surgical skills training. *Br J Surg*. 2012;99 (10):1322–1330. https://doi.org/10.1002/bjs.8819.

14. Lin DT, Melcer EF, Keehl O, et al. ENTRUST: a serious game-based virtual patient platform to assess entrustable professional activities in graduate medical education. *J Grad Med Educ*. 2023;15(2):228–236. https://doi.org/10.4300/JGME-D-22-00518.1.

15. Liebert CA, Melcer EF, Keehl O, et al. Validity evidence for ENTRUST as an assessment of surgical decision-making for the inguinal hernia entrustable professional activity (EPA). *J Surg Educ*. 2022;79(6):e202–e212. https://doi.org/10.1016/j.jsurg.2022.07.008.

16. Liebert CA, Melcer EF, Eddington H, et al. Correlation of performance on ENTRUST and traditional oral objective structured clinical examination for high-stakes assessment in the college of surgeons of East, Central, and Southern Africa. *J Am Coll Surg*. 2023;237(1):117. https://doi.org/10.1097/XCS.0000000000000740.

17. Lewis JR. The system usability scale: past, present, and future. *Int J Human–Comput Inter*. 2018;34 (7):577–590. https://doi.org/10.1080/10447318.2018.1455307.

18. Kortum PT, Bangor A. Usability ratings for everyday products measured with the system usability scale. *Int J Human–Comput Inter*. 2013;29(2):67–76. https://doi.org/10.1080/10447318.2012.681221.

19. Palee P, Wongta N, Khwanngern K, Jitmun W, Choosri N. Serious game for teaching undergraduate medical students in cleft lip and palate treatment protocol. *Int J Med Inform*. 2020;141:104166. https://doi.org/10.1016/j.ijmedinf.2020.104166.

20. Shewaga R, Uribe-Quevedo A, Kapralos B, Lee K, Alam F. A serious game for anesthesia-based crisis resource management training. *Comput Entertain*. 2018;16(2):6. https://doi.org/10.1145/3180660.

21. Lorenzini C, Faita C, Barsotti M, Carrozzino M, Tecchia F, Bergamasco M. ADITHO – a serious game for training and evaluating medical ethics skills. Chorianopoulos K, Divitini M, Baalsrud Hauge J, Jaccheri L, Malaka R, editors. *Entertainment Computing - ICEC 2015*. Lecture Notes in Computer Science, Cham, Switzerland: Springer International Publishing; 2015:59–71. https://doi.org/10.1007/978-3-319-24589-8_5.

22. Lett E, Murdock HM, Orji WU, Aysola J, Sebro R. Trends in racial/ethnic representation among US medical students. *JAMA Netw Open*. 2019;2(9): e1910490. https://doi.org/10.1001/jamanetworkopen.2019.10490.

23. Kluyver T, Ragan-Kelley B, Pérez F, et al. Jupyter Notebooks – a publishing format for reproducible computational workflows. Loizides F, Scmidt B, editors. Amsterdam: IOS Press; 2016:87–90. https://doi.org/10.3233/978-1-61499-649-1-87.

24. Van Rossum G, Drake F. Python 3 reference manual. CreateSpace; 2009.

25. Yáñez-Gómez R, Cascado-Caballero D, Sevillano JL. Academic methods for usability evaluation of serious games: a systematic review. *Multimed Tools Appl*. 2017;76(4):5755–5784. https://doi.org/10.1007/s11042-016-3845-9.

26. Walden A, Garvin L, Smerek M, Johnson C. User-centered design principles in the development of clinical research tools. *Clin Trials*. 2020;17(6):703–711. https://doi.org/10.1177/1740774520946314.

27. Bashir G. Technology and medicine: the evolution of virtual reality simulation in laparoscopic training. *Med Teach*. 2010;32(7):558–561. https://doi.org/10.3109/01421590903447708.

SUPPLEMENTARY INFORMATION

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.jsurg.2023.09.001.