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

Cara A Liebert, MD, FACS, Edward F Melcer, PhD, Hyrum Eddington, BS, Amber Trickey, PhD, MS, CPH, Samuel Shields, BS, Melissa Lee, BA, James R Korndorffer Jr, MD, MHPE, FACS, Abebe Bekele, MD, FACS, FCS(ECSA), Sherry M Wren, MD, FACS, FCS(ECSA), Dana T Lin, MD, FACS

BACKGROUND: To address the global need for accessible evidence-based tools for competency-based education, we developed ENTRUST, an innovative online virtual patient simulation platform to author and securely deploy case scenarios to assess surgical decision-making competence.

STUDY DESIGN: In partnership with the College of Surgeons of East, Central, and Southern Africa, ENTRUST was piloted during the Membership of the College of Surgeons (MCS) 2021 examination. Examinees (n = 110) completed the traditional 11-station oral objective structured clinical examinations (OSCEs), followed by 3 ENTRUST cases, authored to query similar clinical content of 3 corresponding OSCE cases. ENTRUST scores were analyzed for associations with MCS Examination outcome using independent sample *t* tests. Correlation of ENTRUST scores to MCS Examination Percentage and OSCE station scores was calculated with Pearson correlations. Bivariate and multivariate analyses were performed to evaluate predictors of performance.

RESULTS: ENTRUST performance was significantly higher in examinees who passed the MCS examination compared with those who failed (p < 0.001). The ENTRUST score was positively correlated with MCS Examination Percentage (p < 0.001) and combined OSCE station scores (p < 0.001). On multivariate analysis, there was a strong association between MCS Examination Percentage and ENTRUST Grand Total Score (p < 0.001), Simulation Total Score (p = 0.018), and Question Total Score (p < 0.001). Age was a negative predictor for ENTRUST Grand Total and Simulation Total Score, but not for Question Total Score. Sex, native language status, and intended specialty were not associated with performance on ENTRUST.

CONCLUSIONS: This study demonstrates feasibility and initial validity evidence for the use of ENTRUST in a high-stakes examination context for assessment of surgical decision-making. ENTRUST holds potential as an accessible learning and assessment platform for surgical trainees worldwide. (J Am Coll Surg 2023;237:117–127. © 2023 by the American College of Surgeons. Published by Wolters Kluwer Health, Inc. All rights reserved.)

From the Department of Surgery, Stanford University School of Medicine, Stanford, CA (Liebert, Lee, Korndorffer, Wren, Lin); Surgical Services, VA Palo Alto Health Care System, Department of Veterans Affairs, Palo Alto, CA (Liebert, Korndorffer, Wren); Department of Computational Media, Baskin School of Engineering, University of California- Santa Cruz, Santa Cruz, CA (Melcer, Shields); Stanford-Surgery Policy Improvement Research and Education Center, Stanford, CA (Eddington, Trickey); and School of Medicine, University of Global Health Equity, Kigali, Rwanda (Bekele).

Correspondence address: Cara A Liebert, MD, FACS, 3801 Miranda Ave (112), Palo Alto, CA 94304. email: cara.liebert@stanford.edu

Supplemental digital content is available for this article.

A discussion of this article is available at http://links.lww.com/JACS/A260.

Abbreviations and Acronyms

ABS	=	American Board of Surgery
COSECSA	=	College of Surgeons of East, Central, and Southern Africa
EPA	=	entrustable professional activity
FCS	=	Fellowship of the College of Surgeons
MCQ	=	multiple choice question
MCS	=	Membership of the College of Surgeons
OSCE	=	objective structured clinical examination

An essential component of surgical practice is the ability to make informed, safe, complex, and highly nuanced decisions across a spectrum of clinical presentations throughout the continuum of surgical care. Traditionally, certifying bodies such as the American Board of Surgery (ABS) and the College of Surgeons of East, Central, and Southern Africa (COSECSA) have evaluated applied surgical knowledge and decision-making competence of surgical trainees and recent graduates via standardized multiple choice question (MCQ) examinations and oral examinations. Standardized MCQ examinations, although efficient and easy to administer, have limited ability to comprehensively evaluate complex clinical decision-making for a patient encounter throughout the preoperative, intraoperative, and postoperative care of the patient. Oral examinations, on the other hand, have the ability to query nuanced surgical decision-making across the continuum of care, but are time- and resource-intensive and potentially subject to bias. In low- and middle-income countries, these challenges are heightened further by geographic, economic, and language fluency barriers faced by examinees.

To address the global need for efficient, cost-effective, objective, and evidence-based tools to evaluate clinical decision-making competence, our group developed the ENTRUST Assessment Platform.[1,2] ENTRUST is an innovative interactive online platform to author and securely deploy virtual patient simulation case scenarios to rigorously assess surgical decision-making across the spectrum of surgical care. ENTRUST was created to align with competency-based education[3] initially in the context of entrustable professional activities (EPAs)[4-6] for the ABS, but also more broadly for global use. The ENTRUST Authoring Portal and Assessment Platform has been previously described, with initial validity evidence for its use as an assessment of surgical decision-making for the ABS Inguinal Hernia EPA.[1,2,5]

In this study, an ENTRUST Examination containing 3 cases was developed and piloted with COSECSA during the Membership of the College of Surgeons (MCS) Examination to collect initial validity evidence for its use as a platform for high-stakes assessment and a potential alternative format to traditional oral board examinations or objective structured clinical examinations (OSCEs). This study investigated the relationship between performance on the MCS Examination and the ENTRUST Examination. We hypothesized that ENTRUST possesses validity evidence for use in the assessment of applied surgical knowledge and clinical decision-making competence of surgical trainees in COSECSA.

METHODS

ENTRUST assessment platform and case authoring

A detailed description of the ENTRUST Authoring Portal and Assessment Platform has previously been published.[1,2] ENTRUST cases start in Simulation Mode, which includes the preoperative or preadmission evaluation of a patient in the clinic or emergency department. In this mode, the examinee initiates the physical examination and workup of the patient. The examinee can order laboratory studies and imaging through a central console and must subsequently interpret these results. For patients in the emergency department setting, the patient vital signs are visible and can vary over time based on selected physiologic algorithms. The examinee can order fluids, medications, and bedside procedures as indicated, with the patient's vital signs responding dynamically based on these interventions. All actions are recorded and stored in a secure encrypted backend database. Actions are scored according to an expert consensus–derived algorithm where points are earned for appropriate studies and interventions and deducted for inappropriate, unnecessary, or harmful actions. The examinee must then select the appropriate disposition option for the patient (Clinic Evaluation: *Nonoperative Management* or *Consent for Operating Room*; Emergency Department Evaluation: *Discharge Home*, *Admit to Ward*, *Admit to Intensive Care Unit*, or *Proceed to Operating Room*). The case then proceeds to the next phase of care in Question Mode, where the examinee is tested on preoperative optimization, applied surgical knowledge, intraoperative decision-making, and postoperative care via a series of single best answer MCQs.

In partnership with the Examinations and Credentials Committee of COSECSA, 3 clinical cases were authored on the ENTRUST Assessment Platform in clinical content areas of breast, hernia, and thyroid. These cases were authored and reviewed by 2 surgical educators with content expertise to query similar clinical content, patient presentations, and objectives as 3 corresponding OSCE oral examination cases for breast, hernia, and thyroid. The case scenarios and associated scoring algorithms

were iteratively developed by expert consensus aligned with local standard-of-care practices in the COSECSA region. Cases and scoring were reviewed by 2 members of the COSECSA Examinations and Credentials Committee and all cases were beta-tested before deployment. The clinical settings for the 3 ENTRUST case scenarios were as follows: Breast (Simulation Mode: Clinic; Question Mode: Operating Room and Postoperative Ward), Hernia (Simulation Mode: Emergency Department; Question Mode: Operating Room), and Thyroid (Simulation Mode: Emergency Department; Question Mode: Intensive Care Unit and Operating Room). Representative screen shots from the ENTRUST Login Screen and the ENTRUST Cases are shown in Figure 1.

Participants and study design

The ENTRUST Breast, Hernia, and Thyroid Cases were piloted during the COSECSA MCS Examination in November 2021 with n = 110 examinees from 15 sub-Saharan countries. MCS Examination candidates have completed medical school and 2 years of surgical residency training in a COSECSA program. Trainees must achieve a passing score on the MCS Examination before continuing to 3 to 4 additional years of Fellowship of the College of
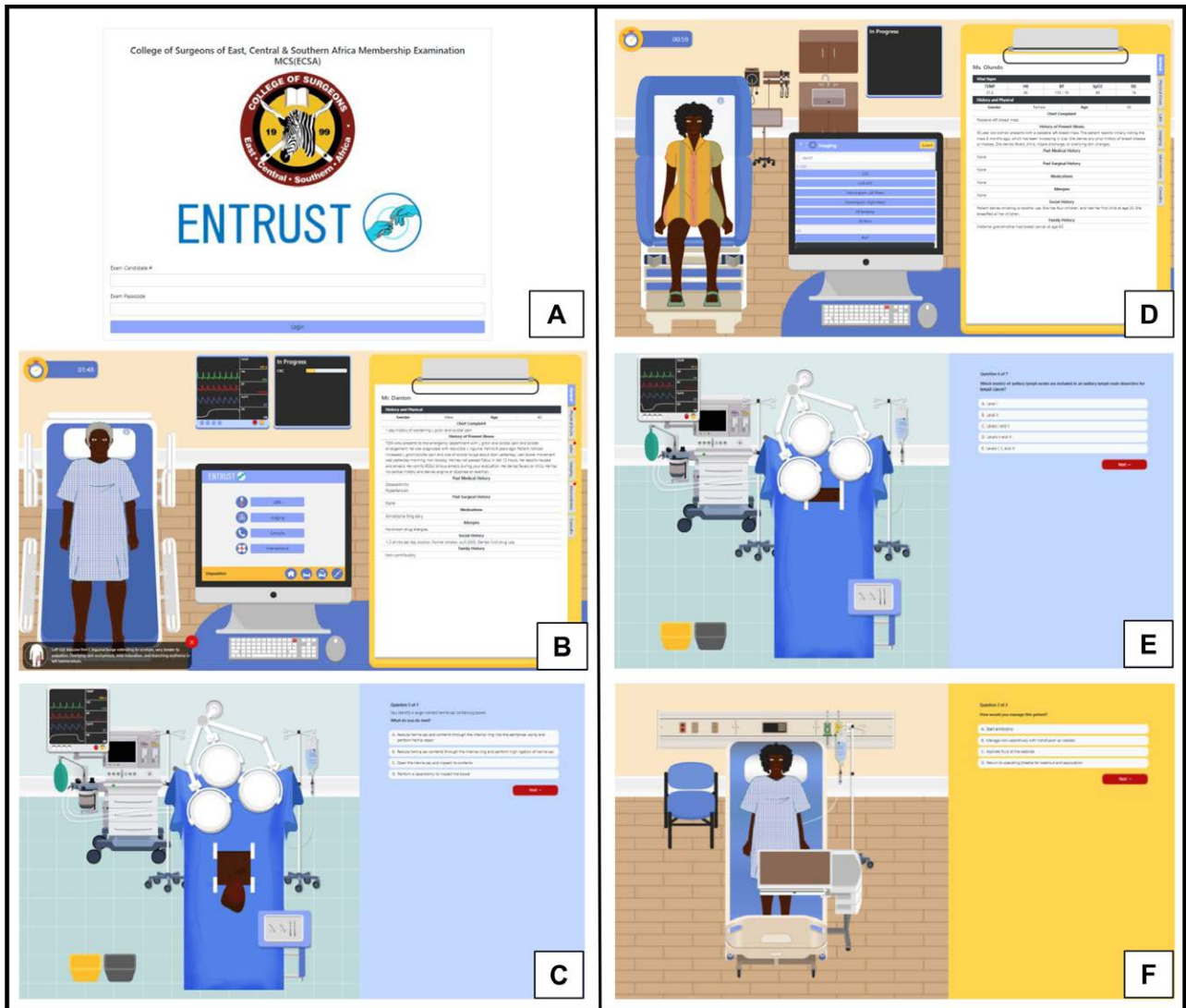

Figure 1. Representative screenshots of the ENTRUST exam login and cases. (A) Login screen, (B) ENTRUST hernia case simulation mode, (C) ENTRUST hernia case question mode, (D) ENTRUST breast case simulation mode, (E) ENTRUST breast case question mode (intraoperative), (F) ENTRUST breast question mode (postoperative).

Surgeons (FCS) training in cardiothoracic surgery, general surgery, neurosurgery, otorhinolaryngology, orthopedic surgery, pediatric surgery, plastic surgery, or urology. This study was designed in collaboration with, and approved by, the Chair of the Examination and Credentials Committee of COSECSA as an initial pilot of the ENTRUST Assessment Platform in this population. The goal of the study was to collect feasibility and validity evidence for use of ENTRUST in this setting. Therefore, examinee performance on ENTRUST was analyzed but did not impact their MCS Examination pass/fail outcome.

Participants were excluded (n = 6) from the analysis if they reported technical difficulties precluding them from completing any of the ENTRUST Examination cases (n = 3) or if they were missing a corresponding OSCE Station Score (breast, hernia, thyroid) due to technical difficulties completing the virtual MCS Examination (n = 3).

In the week before the MCS Examination, examinees were oriented to the ENTRUST Assessment Platform during the official COSECSA MCS Examination Orientation session, which included a video tutorial and live demonstration of the platform. All examinees were given access to the tutorial video and an ENTRUST practice case in the week before the examination.

On the day of the MCS Examination, examinees first completed the faculty-administered MCS 11-station oral OSCE virtually via the Zoom platform in a proctored setting at their training institution. After completion of their MCS Examination, examinees securely logged into the ENTRUST Assessment Platform where basic demographic information was collected. Examinees then watched a video tutorial and completed the ENTRUST practice case before completing 3 clinical cases on the online ENTRUST Assessment Platform (Breast, Hernia, and Thyroid). Examinees were notified that the ENTRUST portion of the examination was a pilot and that performance on this portion of the examination would not count toward their MCS Examination outcome. All ENTRUST data collection occurred after the completion of the examinees' MCS Examination. The study was reviewed and approved by the IRB at our institution.

Membership of the College of Surgeons examination scoring

MCS Examination outcome was determined by COSECSA based on the composite score of the 11 OSCE stations. Each OSCE station was scored on a scale from 0 to 20 by the MCS faculty examiner based on a provided rubric, with standard setting performed for each case by the Examinations and Credentials Committee of COSECSA. Additionally, MCS faculty examiners

provided a Global Assessment for performance on each OSCE station, including "Pass," "Borderline Pass," or "Fail." The composite 11-station OSCE raw score was used to determine the total MCS Examination Percentage correct across all OSCE stations. The Examinations and Credentials Committee of COSECSA determined the percentage passing score based on an internal standard setting, which determined the overall MCS Examination outcome of "Pass" or "Fail."

ENTRUST scoring

The scoring algorithm for ENTRUST was designed by the authors to reflect the appropriateness of actions, patient clinical status, and accuracy of MCQ responses. Diagnostic studies and interventions were categorized using the following framework: critical (+200), indicated (+100), neutral (0), not indicated but not harmful (–50), mild to moderate harm (–100), severe harm (–200), and death/cardiac arrest (–500). Questions were awarded 200 points for correct responses and deducted 200 for incorrect responses. Penalty points were additionally deducted (–200) for each instance of failure to address and correct critical vital sign abnormalities.

Case scores were derived for each of the 3 ENTRUST cases (Breast, Hernia, and Thyroid) by totaling the Simulation Mode and Question Mode scores. ENTRUST Grand Total Score was derived by adding together the ENTRUST case scores for the 3 cases. ENTRUST Simulation Total Score was the total of the Simulation Mode score for each of the 3 cases. ENTRUST Question Total Score was the total of the Question Mode score for each of the 3 cases.

Statistical analysis

Descriptive statistics were calculated for demographics and training characteristics including age, sex, country of training, native language, and intended FCS specialty. To determine whether participants who passed the MCS Examination scored higher on ENTRUST than those who failed the MCS Examination, independent sample *t* tests were performed comparing ENTRUST Grand Total Score with MCS Overall Examination pass/fail status. Similar analyses were conducted for ENTRUST Simulation Total Score and ENTRUST Question Total Score. Independent sample *t* tests were also performed comparing ENTRUST case scores with the corresponding OSCE Station Global Assessment faculty rating. Participants with an OSCE Station Global Assessment rating of "Borderline Pass" were grouped with those with a rating of "Pass."

Pearson correlations were performed between ENTRUST performance and MCS Examination Percentage for ENTRUST Grand Total Score, Simulation Total Score, and Question Total Score. Pearson correlations were performed for participants' Breast, Hernia, and Thyroid case scores on ENTRUST with the corresponding OSCE Station Score (0 to 20) to assess for associations in performance for specific cases. We also assessed the correlation between participants' ENTRUST Grand Total Score, which comprised Breast, Hernia, and Thyroid cases, with the sum of their corresponding OSCE Breast, Hernia, and Thyroid stations.

Bivariate analyses including Pearson correlations and t tests for associations between ENTRUST and MCS Examination scores with age, sex, intended FCS specialty, and native language were also completed.

To assess potential demographic biases and determine whether participants' MCS Examination Percentage was associated with ENTRUST scores after adjustment for demographics and training characteristics, we constructed 2 multivariable models using either MCS Examination Percentage or ENTRUST Grand Total Score as the primary predictor and the other score as the primary outcome. We controlled for age, sex, intended FCS specialty, and native language to examine any potential demographic sources of bias. We created 2 more multivariate models of ENTRUST Simulation Total Score and Question Total Score by MCS Examination Percentage and the same demographic covariates to assess whether predictive factors varied for different portions of the ENTRUST examination. We conducted nonparametric sensitivity analyses for all significance tests, including Wilcoxon rank-sum tests, Spearman rank correlations, and nonparametric kernel regression multivariable models (**Supplemental Digital Content 1 to 6**, http://links.lww.com/JACS/A257). All analyses were conducted in R, with the exception of nonparametric kernel regression performed in Stata v. 16.1. Statistical significance was assessed at the level of $p < 0.05$.

RESULTS
Participant demographics
Demographics and training characteristics of the participating examinees are presented in Table 1, including age, sex, country of training, native language, and intended FCS specialty. The average age was 34 ± 6, and most of the participants were male (72%) and nonnative English language speakers (83%). The most frequent intended FCS specialty was general surgery (39%). Participants represented 15 different countries of training, with the largest representation from Kenya (39%), Malawi (12%), and Uganda (11%). Of the n = 104 examinees included in

Table 1. Demographics

Demographic characteristic	Data
Age, y, mean ± SD	34 ± 6
Sex, n (%)	
Female	29 (28.8)
Male	75 (72.1)
Country of training, n (%)	
Kenya	40 (38.5)
Malawi	12 (11.5)
Uganda	11 (10.6)
Zambia	9 (8.7)
Zimbabwe	6 (5.8)
Rwanda	4 (3.8)
Tanzania	4 (3.8)
Cameroon	3 (2.9)
Ethiopia	3 (2.9)
Lesotho	3 (2.9)
Namibia	3 (2.9)
Democratic Republic of Congo	2 (1.9)
Niger	2 (1.9)
Gabon	1 (1.0)
Sudan	1 (1.0)
Native language, n (%)	
English	18 (17.3)
Other	86 (82.7)
Intended Fellowship of the College of Surgeons specialty, n (%)	
General surgery	41 (39.4)
Orthopaedic surgery	20 (19.2)
Neurosurgery	14 (13.5)
Plastic surgery	7 (6.7)
Pediatric surgery	6 (5.8)
Cardiothoracic surgery	3 (2.9)
Urologic surgery	3 (2.9)
Pediatric orthopaedic surgery	1 (1.0)
Otolaryngology	1 (1.0)
Undecided	8 (7.7)

Percentages may not add up to 100% due to rounding.

this study, 87 (84%) passed the MCS Examination and 17 (16%) failed.

ENTRUST performance by Membership of the College of Surgeons examination outcome
ENTRUST score was significantly higher in the examinees who passed the MCS Examination than in those who failed for both the mean ENTRUST Grand Total Score (3,109 vs 707, $p = 0.003$) and mean ENTRUST Question Total Score (2,740 vs 1,259, $p < 0.001$; Fig. 2). Mean
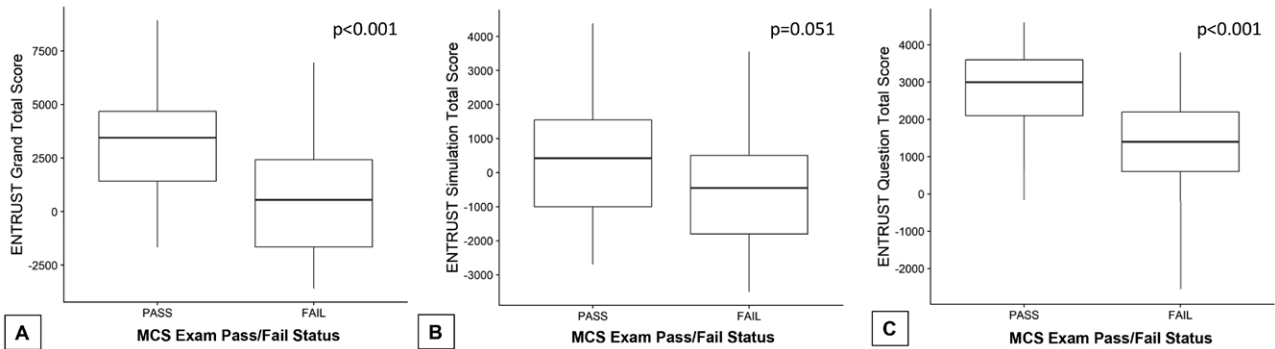
Figure 2. ENTRUST performance by Membership of the College of Surgeons (MCS) Examination outcomes for ENTRUST grand total score (A), ENTRUST simulation total score (B), and ENTRUST question total score (C). Boxes represent interquartile range from 25th to 75th percentile. Lines within boxes represent median and whiskers represent minimum and maximum.

Figure 3. ENTRUST case scores by objective structured clinical examination (OSCE) station global assessment rating for breast (A), hernia (B), and thyroid (C). Boxes represent interquartile range from 25th to 75th percentile. Lines within boxes represent median and whiskers represent minimum and maximum.
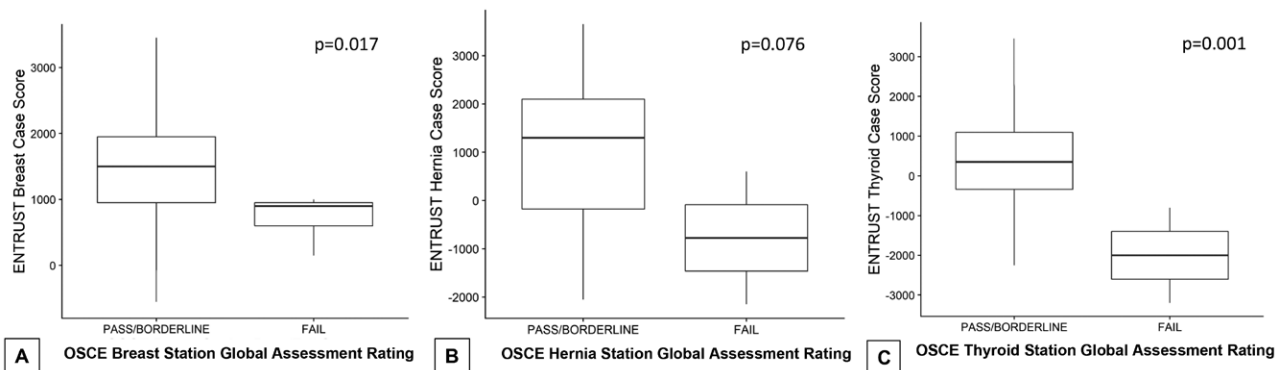
ENTRUST Simulation Total Score was higher in the examinees who passed the MCS Examination than in those who failed (368 vs –551, p = 0.051), which approached but did not reach statistical significance (Fig. 2).

When analyzed at the individual ENTRUST case level, there were significantly higher ENTRUST Breast, Hernia, and Thyroid case scores in examinees who passed the MCS Examination compared with those who failed (p = 0.005, p = 0.04, and p = 0.001, respectively).

ENTRUST case performance by objective structured clinical examination station global assessment rating

Mean ENTRUST Breast Case Score and mean ENTRUST Thyroid Case Score were significantly higher in examinees who received a faculty Global Assessment rating of "Pass" or "Borderline Pass" for the corresponding OSCE station compared with those who received a rating of "Fail" (1,450 vs 736, p = 0.017 and 368 vs –2000, p = 0.001, respectively; Fig. 3). The difference in mean ENTRUST Hernia

Case Score for those who received a faculty rating of "Pass" or "Borderline Pass" for the OSCE Hernia Station was not statistically significant compared with those who received a rating of "Fail" (1,026 vs –775, p = 0.076; Fig. 3).

Correlation of ENTRUST performance and Membership of the College of Surgeons Examination Percentage

ENTRUST Grand Total Score, Simulation Total Score, and Question Mode Total Score were positively correlated with MCS Examination Percentage (r = 0.53, p < 0.001; r = 0.35, p <0.001; r = 0.59, p < 0.001).

Correlation of ENTRUST case performance to objective structured clinical examination station performance

At the case level, there was a statistically significant correlation between performance on ENTRUST and the

respective OSCE Station for Breast and Thyroid ($r = 0.34$, $p < 0.001$ and $r = 0.46$, $p < 0.001$, respectively). Correlation between ENTRUST Hernia Case Score and OSCE Hernia Station did not reach statistical significance ($r = 0.18$, $p = 0.073$). The ENTRUST Grand Total Score, representing the Breast, Hernia, and Thyroid cases, was significantly correlated to the combined score on the corresponding OSCE stations ($r = 0.52$, $p < 0.001$; Fig. 4).

Bivariate analysis

In bivariate analysis, there was an inverse relationship between MCS Examination Percentage and age ($r = -0.35$, $p < 0.001$), as shown in Table 2. A similar inverse correlation was observed between ENTRUST Grand Total Score and age ($r = -0.40$, $p < 0.001$). There were no other statistically significant differences observed on bivariate analysis for ENTRUST Grand Total Score or MCS Examination Percentage based on sex (female/male), native language (English/other), or intended FCS specialty (general surgery/other; Table 3).

Multivariate analysis

The demographic variables used in the multivariate model included age, sex, native language, and intended FCS specialty. Country of training was not included in the model due to high variation and small group sizes in the countries represented.

On multivariate analysis of MCS Examination performance, there was a strong association between ENTRUST Grand Total Score and MCS Examination Percentage ($p < 0.001$; Table 4). In multivariate analysis, age was not a statistically significant predicator of MCS Examination performance ($p = 0.13$; Table 4). Sex, native language status, and intended FCS specialty were not significant predictors of MCS Examination Percentage.

On multivariate analysis of ENTRUST performance, there were significant associations between MCS
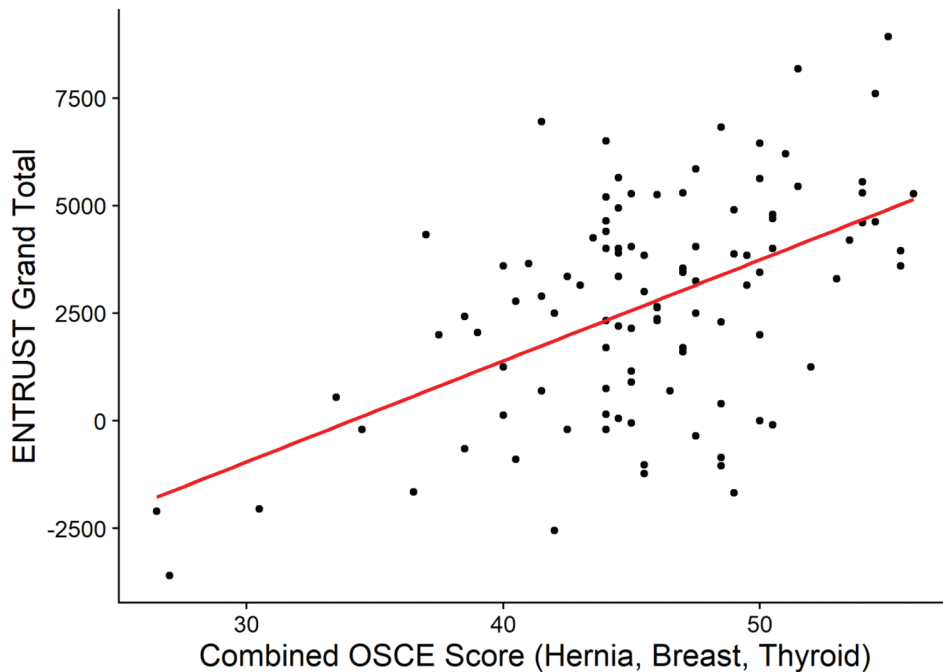
Figure 4. Correlation of ENTRUST performance to objective structured clinical examination (OSCE) stations.

Table 2. Bivariate Analysis of Continuous Model Variables

Variable	Mean ± SD	ENTRUST grand total score		MCS examination percentage	
		Pearson's correlation	p Value	Pearson's correlation	p Value
Age, y	34 ± 6	-0.40*	<0.001	-0.35*	<0.001

*Score change per unit.
MCS, Membership of the College of Surgeons.

Table 3. Bivariate Analyses of Categorical Model Variables

Variable	n (%)	ENTRUST grand total score		MCS examination percentage	
		Mean ± SD	p Value	Mean ± SD	p Value
Sex			0.12		0.07
Female	29 (28)	3,334 ± 2,459		68.0 ± 8.7	
Male	75 (72)	2,477 ± 2,541		64.5 ± 8.6	
Native language			0.15		0.84
English	18 (17)	3,294 ± 1,619		67.9 ± 9.3	
Other	86 (83)	2,595 ± 2,680		64.9 ± 8.6	
Intended FCS specialty			0.79		0.39
General surgery	41 (39)	2,799 ± 2,690		64.5 ± 9.6	
Other	63 (61)	2,662 ± 2,452		66.1 ± 8.2	

FCS, Fellowship of the College of Surgeons; MCS, Membership of the College of Surgeons.

Table 4. Multivariate Analysis of Membership of the College of Surgeons Examination Percentage

Variable	MCS examination percentage		
	Estimate	95% CI	p Value
ENTRUST grand total score*	0.0016	0.001 to 0.002	<0.001
Age*	−0.20	−0.47 to 0.061	0.13
Sex (female vs male)	1.76	−1.49 to 5.00	0.29
Native language (English vs other)	2.07	−1.82 to 5.96	0.29
Intended FCS (general surgery vs other)	−1.86	−4.88 to 1.16	0.23
Intercept	69.90	58.07 to 77.77	<0.001

*Score change per unit for continuous variables.
FCS, Fellowship of the College of Surgeons; MCS, Membership of the College of Surgeons.

Examination Percentage and ENTRUST Grand Total Score (p < 0.001), ENTRUST Simulation Total Score (p = 0.018), and ENTRUST Question Total Score (p < 0.001; Table 5). In multivariate analysis, age was a negative predictor of score for ENTRUST Grand Total Score (p = 0.006) and ENTRUST Simulation Total Score (p = 0.001), but not for ENTRUST Question Total Score (p = 0.67; Table 5). There were no other statistically significant predictors of ENTRUST performance based on sex, native language status, or intended FCS specialty.

Sensitivity analyses

All bivariate parametric tests were confirmed by nonparametric sensitivity analyses with similar relationships and statistical significance interpretations in Wilcoxon rank-sum tests and Spearman rank correlations. In multivariable nonparametric kernel regression models, all results were similar in magnitude and direction of the effects as well as statistical significance, with a single exception.

The relationship between ENTRUST Simulation Total Score and MCS Examination Percentage was not statistically significant in nonparametric kernel regression (p = 0.20). Full sensitivity methods and results are available in **Supplemental Digital Content 1 to 6**, http://links.lww.com/JACS/A257.

DISCUSSION

This study demonstrates the feasibility of use of the ENTRUST Assessment Platform for high-stakes summative assessment of clinical decision-making of surgical trainees. The online ENTRUST Examination was successfully accessed by participants in 15 sub-Saharan countries in Africa. Although there were 3 participants who were excluded from the analysis due to technical difficulties completing the ENTRUST cases, such as internet connection disruption, this was equal to the number of participants excluded due to technical difficulties during one or more of the corresponding OSCE stations administered via Zoom platform.

ENTRUST score performance was significantly higher in examinees who passed the MCS Examination than those who failed, and ENTRUST Grand Total Score and Question Mode Total Score were significantly correlated with MCS Examination Percentage. On a case level, performance on ENTRUST cases was correlated to combined performance on the corresponding OSCE stations. The results demonstrate that performance on ENTRUST Question Total Score was most closely associated with performance on the MCS Examination, which is not unexpected, because these were most similar in examination format and queried similar constructs. In contrast, the ENTRUST Simulation Total Score may represent a different construct not readily assessed in the traditional OSCE format. The Simulation Mode required examinees to respond in real time to patient vital sign deterioration,

Table 5. Multivariate Analysis of ENTRUST Score Performance

Variable	ENTRUST grand total score			ENTRUST simulation total score			ENTRUST question total score		
	Estimate	95% CI	p Value	Estimate	95% CI	p Value	Estimate	95% CI	p Value
MCS examination percentage*	129.4	78.1 to 180.5	<0.001	47.9	8.50 to 87.3	0.018	81.5	56.4 to 106.5	<0.001
Age*	−102.8	−176.1 to −29.5	0.006	−95.1	−151.5 to −38.7	0.001	−7.8	−43.7 to 28.1	0.67
Sex (female vs male)	191.7	−741.8 to 1,125	0.68	95.8	−622.6 to 814.1	0.79	95.9	−361.4 to 553.3	0.68
Native language (English vs other)	105.8	−1,013 to 1,225	0.85	−1.6	−862.8 to 859.7	0.99	107.4	−440.9 to 655.7	0.70
Intended FCS specialty (general surgery vs other)	466.7	−400.1 to 1,334	0.29	243.2	−423.9 to 910.2	0.47	223.6	−201.1 to 648.2	0.30
Intercept	−2501	−7285 to 2,283	0.49	200.2	−3481 to 3,882	0.78	−2701	−5045 to −357	0.07

*Score change per unit for continuous variables.
FCS, Fellowship of the College of Surgeons; MCS, Membership of the College of Surgeons.

with penalties received for inattention to critically abnormal vital signs or for inappropriate, unnecessary, or harmful interactions.

For the ENTRUST Hernia Case, two of the analyses approached but did not reach statistical significance. We hypothesized that this may be due to the impact of passive vital sign penalty on score performance, because this case portrayed an unstable patient presenting in septic shock. When passive penalty received for failure to address and correct critical vital sign abnormalities was excluded, the mean adjusted ENTRUST Hernia Case Score was significantly higher in examinees who received a Global Assessment rating of "Pass" or "Borderline Pass" compared with those who received "Fail" (2,005 vs 225, p = 0.009). Similarly, mean adjusted ENTRUST Hernia Case Score was significantly correlated with OSCE Hernia Station performance (r = 0.2, p = 0.037). Mean adjusted ENTRUST Simulation Total Score was also significantly higher for examinees who passed the MCS Examination compared with those who failed (2,920 vs 2,154, p = 0.027). A similar effect was observed for the ENTRUST Thyroid Case, with the adjusted analyses yielding a stronger correlation coefficient and more significant p value than the unadjusted analyses. There was no passive vital sign penalty possible for the ENTRUST Breast Case, and, therefore, adjusted analysis was not performed for this case.

On multivariate analysis, age was a predictor of performance on ENTRUST Grand Total Score and ENTRUST Simulation Total Score, but not ENTRUST Question Total Score. This suggests that usability of the ENTRUST Simulation Mode may be highest in younger examinees, perhaps due to differences in computer literacy by age group. To further investigate this, a subsequent study will evaluate computer literacy and usability of ENTRUST Assessment Platform in this population and evaluate for predictors of usability based on age or other demographic variables. Based on the findings in this study, an interactive tutorial has since been created to better orient new users to the functionality of the ENTRUST platform before completing an ENTRUST Assessment. Additionally, in subsequent MCS Examinations, examinees will be given extended access to an ENTRUST Practice Examination before the MCS Examination. Ongoing studies will continually re-evaluate the usability of ENTRUST in different populations and demographic groups and evaluate the impact of time spent on the tutorial and practice examination on usability and ENTRUST score performance. There were no significant differences in performance on ENTRUST based on sex, English native language status, or intended FCS specialty.

This study provides initial validity evidence according to Messick's unified framework of construct validity[7,8] for the use of ENTRUST as an objective measure of applied surgical knowledge and surgical decision-making for high-stakes assessment of surgical trainees. Content evidence for the case scenarios was established by authoring of the case content by content experts and surgical educators, with review of the cases and scoring by the Chair of the Examinations and Credentials Committee of COSECSA. The association and correlation of ENTRUST performance with MCS Examination outcome and OSCE Station Score provides evidence of its relationship to other variables. Data collected in this study will allow for future standard setting and consequences of cut scores for pass/fail based on ENTRUST performance.

Several limitations of this study exist, including that examinees were aware that scores on the ENTRUST cases did not contribute to their MCS Examination outcome, which may have affected their effort and ENTRUST score performance. Although all study participants were required to watch the video tutorial and complete the ENTRUST practice case on the day of the examination before proceeding to the ENTRUST Examination cases, completion of the tutorial and ENTRUST practice case before the examination was optional and not analyzed as part of this study. Thus, we are unable to determine the impact of additional time spent on the tutorial and/or practice case on ENTRUST score performance. Further studies are planned to evaluate the impact of time spent on the tutorial and practice case on usability and ENTRUST score performance. Additionally, although the sample size overall was adequate to evaluate for differences in performance, the relatively large number of countries and native languages limited the ability to evaluate country of training or specific language as a predictor of performance on multivariate analysis.

As medical education continues to move toward a competency-based paradigm,[3] the need for evidence-based tools to accurately assess the knowledge, skills, and clinical decision-making ability of surgical trainees has never been greater. This is further demonstrated by the fact that many specialties, including the ABS, are moving toward implementation and assessment of EPAs, or units of professional practice that constitute what clinicians do as daily work, to help bridge the gap between competency frameworks and clinical practice.[4,5,9] Beginning in 2015, the ABS began to explore EPAs as a foundation for competency-based education in surgical training and initial board certification, and has since announced that use of EPAs in all general surgery residency programs will begin in July 2023.[10] An ENTRUST Inguinal Hernia EPA Module was developed and piloted at our institution with

validity evidence for its use as an assessment of surgical decision-making.[1] The ENTRUST Assessment Platform holds promise as a complement to existing evidence-based tools for intraoperative assessment of technical skills and operative autonomy[11,12] in the move toward competency-based education in the US and worldwide.

Future directions include the analysis of usability of the ENTRUST Assessment Platform in surgical trainees from both the US and COSECSA training programs. We plan to collect further validity evidence for ENTRUST using Messick's framework, including analysis of response process evidence, internal structure, and consequences. In the US, we plan to correlate performance on ENTRUST with microassessments from the System for Improving and Measuring Procedural Learning[11] or other platforms for assessment of EPAs. In future studies in COSECSA training programs, we plan to investigate the relationship to other variables such as FCS examination pass rates. Based on results from this study, the ENTRUST Assessment Platform will be incorporated formally as a scored portion of the MCS Examination and is being developed as a learning platform for self-assessment for African trainees.

CONCLUSIONS

This study demonstrates the feasibility of use of the ENTRUST Assessment Platform in a formal high-stakes examination context. Surgical trainee performance on ENTRUST was strongly correlated with traditional oral examination performance, providing validity evidence for use of ENTRUST as an assessment of clinical decision-making in surgical trainees. The ENTRUST platform holds potential as an accessible learning and assessment platform for surgical trainees worldwide.

Author Contributions

Conceptualization: Liebert, Melcer, Korndorffer, Bekele, Wren, Lin

Data curation: Liebert, Melcer, Eddington, Bekele

Funding acquisition: Liebert, Korndorffer, Lin

Investigation: Liebert, Melcer, Lee, Bekele, Wren, Lin

Methodology: Liebert, Eddington, Trickey, Korndorffer, Bekele, Lin

Project administration: Liebert, Melcer, Shields, Lin

Visualization: Liebert, Eddington

Writing – original draft: Liebert, Lin

Software: Melcer, Shields

Writing – review & editing: Melcer, Eddington, Trickey, Shields, Lee, Korndorffer, Bekele, Wren

Formal analysis: Eddington, Trickey

Supervision: Trickey, Shields, Lin

REFERENCES

1. 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:e202–e212.
2. Lin DT, Melcer EF, Keehl O, et al. ENTRUST: a serious game-based virtual patient platform to assess Entrustable Professional Activities (EPAs) in graduate medical education. J Grad Med Educ 2023;15:228–236.
3. ten Cate O, Scheele F. Competency-based postgraduate training: can we bridge the gap between theory and clinical practice? Acad Med 2007;82:542–547.
4. Ten Cate O, Chen HC, Hoff RG, et al. Curriculum development for the workplace using Entrustable Professional Activities (EPAs): AMEE Guide No. 99. Med Teach 2015;37:983–1002.
5. Brasel KJ, Klingensmith ME, Englander R, et al. Entrustable Professional Activities in general surgery: development and implementation. J Surg Educ 2019;76:1174–1186.
6. ten Cate O. When I say … entrustability. Med Educ 2020;54:103–104.
7. Messick S. Standards of validity and the validity of standards in performance assessment. Educ Meas Issues Pract 1995;14:5–8.
8. Cook DA, Zendejas B, Hamstra SJ, et al. What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. Adv Health Sci Educ Theory Pract 2014;19:233–250.
9. Cate O. A primer on entrustable professional activities. Korean J Med Educ 2018;30:1–10.
10. ABS announces transition to entrustable professional activities for general surgery resident evaluation. American Board of Surgery. February 14, 2022. Available at: https://www.absurgery.org/default.jsp?news_epas0222. Accessed October 16, 2022.
11. Eaton M, Scully R, Schuller M, et al. Value and barriers to use of the SIMPL tool for resident feedback. J Surg Educ 2019;76:620–627.
12. Sandhu G, Nikolian VC, Magas CP, et al. OpTrust: validity of a tool assessing intraoperative entrustment behaviors. Ann Surg 2018;267:670–676.