Validity Evidence for ENTRUST as an Assessment of Surgical Decision-Making for the Inguinal Hernia Entrustable Professional Activity (EPA)

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

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

OBJECTIVE: As the American Board of Surgery (ABS) moves toward implementation of Entrustable Professional Activities (EPAs), there is a growing need for objective evaluation of readiness for entrustment of residents. This requires not only assessment of technical skills and knowledge, but also surgical decision-making in preoperative, intraoperative, and postoperative settings. We developed and piloted an Inguinal Hernia EPA Assessment on ENTRUST, a serious game-based online virtual patient simulation platform to assess trainees' decision-making competence.

DESIGN: This is a prospective analysis of resident performance on the ENTRUST Inguinal Hernia EPA Assessment using bivariate analyses.

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

PARTICIPANTS: Forty-three surgical residents completed the ENTRUST Inguinal Hernia EPA Assessment.

RESULTS: Four case scenarios for the Inguinal Hernia EPA and corresponding scoring algorithms were iteratively developed by expert consensus aligned with ABS EPA descriptions and functions. ENTRUST Inguinal Hernia Grand Total Score was positively correlated with PGY-level ($p < 0.0001$). Preoperative, Intraoperative, and Postoperative Total Scores were also positively correlated with PGY-level ($p = 0.001$, $p = 0.006$, and $p = 0.038$, respectively). Total Case Scores were positively correlated with PGY-level for cases representing elective unilateral inguinal hernia ($p = 0.0004$), strangulated inguinal hernia ($p < 0.0001$), and elective bilateral inguinal hernia ($p = 0.0003$). Preoperative Sub-Scores were positively correlated with PGY-level for all cases ($p < 0.01$). Intraoperative Sub-Scores were positively correlated with PGY-level for strangulated inguinal hernia and bilateral inguinal hernia ($p = 0.0007$ and $p = 0.0002$, respectively). Grand Total Score and Intraoperative Sub-Score were correlated with prior operative experience ($p < 0.0001$). Prior video game experience did not correlate with performance on ENTRUST ($p = 0.56$).

CONCLUSIONS: Performance on the ENTRUST Inguinal Hernia EPA Assessment was positively correlated to PGY-level and prior inguinal hernia operative performance, providing initial validity evidence for its use as an objective assessment for surgical decision-making. The ENTRUST platform holds potential as tool for assessment of ABS EPAs in surgical residency programs. (J

Funding: This work was supported by the Mark Freidell Research Grant from the Association of Program Directors in Surgery and through seed grant funding from the Division of General Surgery and Department of Surgery at Stanford University School of Medicine.

Correspondence: Inquiries to Cara A. Liebert, MD, FACS, VA Palo Alto Health Care System, Surgical Services, 3801 Miranda Avenue, Surgical Services 112, Palo Alto, CA 94304; e-mail: cara.liebert@stanford.edu

1931-7204/$30.00
https://doi.org/10.1016/j.jsurg.2022.07.008

Surg Ed 79:e202−e212. 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

KEY WORDS: resident assessment, surgical decision-making, general surgery, validity evidence, entrustable professional activity, inguinal hernia

COMPETENCIES: Medical Knowledge, Patient Care

INTRODUCTION

Medical education is moving increasingly towards a competency-based paradigm predicated upon multiple, real-time assessments to verify proficiency.[1] Entrustable Professional Activities (EPAs), or units of professional practice that constitute what clinicians do as daily work, were created to bridge the gap between competency frameworks and clinical practice.[2] EPAs are tasks or responsibilities to be entrusted to a trainee once they have attained competence at a specific level and are specialty-specific, observable, and measurable.[1] EPAs embody a more global integration of the Accreditation Council for Graduate Medical Education (ACGME) core competencies, and applies these competencies to a specific clinical situation or disease process.

In 2018, the American Board of Surgery (ABS) commenced a multi-institutional pilot to implement 5 general surgery EPAs, each with defined levels of entrustment from Level 0 to Level 4, in surgical residency.[3,4] These initial 5 ABS EPAs include: 1) evaluation and management of a patient with inguinal hernia, 2) evaluation and management of a patient with right lower quadrant pain, 3) evaluation and management of a patient with gallbladder disease, 4) evaluation and management of a patient with blunt/penetrating trauma, and 5) providing general surgical consultation to other health care providers.[3] The ABS has given individual residency programs the ability to determine how EPAs are piloted and assessed at their institution. While tools exist for the intra-operative assessment of technical skills and operative autonomy,[5-10] they do not directly not assess surgical decision-making across the preoperative, intraoperative, and postoperative settings. The assessment of technical skills is necessary, but is not sufficient, to determine entrustment. Therefore, there is a need for efficient, objective, evidence-based EPA-aligned tools that assess clinical decision-making across the entire course of surgical care, as a fitting complement to existing technical skill and intra-operative evaluations.

ENTRUST is an innovative serious game-based virtual patient platform developed to provide an objective, efficient, and rigorous assessment platform of surgical decision-making for EPAs. In this study, an ENTRUST Inguinal Hernia EPA Assessment containing 4 cases was developed and piloted to collect initial validity evidence using Messick's framework.[11,12] We hypothesized that ENTRUST possesses validity evidence for use in the assessment of surgical decision-making for general surgery residents.

MATERIALS AND METHODS

Participants

Surgery residents (n = 43) completed the ENTRUST Inguinal Hernia EPA Assessment at our institution in May 2021. The study was completed in a proctored exam classroom setting on laptop computers. Participants completed a demographic survey querying age, gender, ethnicity, PGY-level, surgical specialty, self-reported inguinal hernia operative case volume, and prior video game experience. After viewing a standardized video tutorial orientation to the ENTRUST platform, participants completed a non-scored practice case which enabled them to interact firsthand with ENTRUST and familiarize themselves with the platform interface and functionality. Once finished with the practice case, participants completed the ENTRUST Inguinal Hernia EPA Assessment, which included 4 case scenarios: an outpatient elective unilateral inguinal hernia, an elective bilateral inguinal hernia, an acutely incarcerated inguinal hernia, and a strangulated inguinal hernia. The study protocol (#53137) was reviewed and approved by the Institutional Review Board at our institution.

DESCRIPTION OF ENTRUST PLATFORM

Authoring Portal

ENTRUST features an online authoring platform for surgical educators to create and deploy case scenarios. Clinical vignettes and multiple-choice questions can be entered and edited. Media files such as physical exam photographs and radiology images can be uploaded to be interpreted by the examinee. Authors designate effects of diagnostic and treatment interventions on vital signs and appropriateness of actions, assigning rewards and penalties on a tiered scoring system.

Assessment Platform

The ENTRUST assessment platform includes 3 modes or phases of care: Preoperative Simulation Mode,

Intraoperative Question Mode, and Postoperative Question Mode (Fig. 1). In Preoperative Simulation Mode, case scenarios begin in either the outpatient clinic or emergency department. The examinee initiates physical examination and workup of the patient, depicted on the left side of the screen. The patient's vital signs appear on an overhead monitor and can change dynamically based on the patient's clinical status and interventions performed. The clinical vignette is located on the right side of the screen. Physical exam, laboratory, and imaging results are populated in the chart in response to the examinee's actions. A central console enables the examinee to order diagnostic tests, administer fluids and medications, perform bedside procedures, and request consultation. All actions are recorded, scored, and stored in a secure back-end database according to an expert-consensus derived scoring algorithm. Points are earned for ordering relevant labs and key interventions; conversely, points are deducted for performing inappropriate, unnecessary, or harmful actions. When the examinee proceeds to the operating room, the case scenario transitions to Intraoperative Question Mode where the examinee is tested on intra-operative decision-making via a series of single-best answer multiple choice questions. A subset of the case scenarios additionally includes a Postoperative Question Mode where the examinee is tested on diagnosis and appropriate management of postoperative complications in a similar fashion.

Technology Specifications

ENTRUST utilizes a JavaScript and P5.js front-end to provide an interactive simulation interface, and an encrypted Google Cloud database back-end for secure data logging and analysis of demographic data, player actions, and scores. The platform is accessible via online link via web browser. A secure encrypted back-end database logs detailed trainee performance data including a time stamp of all examinee actions, points awarded or deducted, and responses to all multiple-choice questions.

DATA ANALYSIS

Demographics were reported as mean and standard deviation for continuous variables and proportions for categorical variables. Descriptive statistics for total and subscores, including median and interquartile range, were calculated for each PGY-level. To assess the relationship between ENTRUST score and resident level of training, Spearman rank correlations were calculated to examine the relationship between ENTRUST scores and ordinal PGY-level (1-5). These analyses were performed for ENTRUST Grand Total Score, Preoperative Total Score, Intraoperative Total Score, and Postoperative Total Score. Additionally, Total Case Score, Preoperative Sub-Score, Intraoperative Sub-Score, and Postoperative Sub-Score were calculated for individual case scenarios. Associations of ENTRUST Grand Total Score and Intraoperative Total Score with self-reported total inguinal hernia operative cases performed and video game experience were examined using Spearman rank correlations. Correlation between score and self-reported inguinal hernia operative experience was visualized using Locally Estimated Scatterplot Smoothing (LOESS). We assessed variations in scores between categorical and non-categorical PGY-1 and PGY-2 residents using Wilcoxon rank-sum tests.

A critical clinical decision-making action relevant for entrustment, specifically, the decision to attempt to manually reduce a hernia in the emergency department, was evaluated in additional analyses for the acutely incarcerated and strangulated inguinal hernia case scenarios. For these cases, the percentage of trainees selecting the correct answer was calculated by PGY-level. Wilcoxon rank-sum tests were calculated to examine whether participants who responded correctly on this critical action had significantly higher total and preoperative subscores than those who responded incorrectly. For this analysis, the preoperative score was adjusted to remove the score reward or penalty related to this critical action to eliminate the effect of the critical action itself on participant score. For all statistical tests, a significance threshold of $p < 0.05$ was utilized. All analyses were conducted using R v.4.0.2 (Vienna, Austria).[13]

RESULTS

A total of n = 43 surgical trainees completed the study at our institution (Table 1). Participants included general surgery categorical residents, general surgery preliminary residents, and designated surgical subspecialty residents in the general surgery residency program. Designated surgical subspecialty residents were in PGY-1 or PGY-2 year of training and included residents from cardiothoracic surgery, ophthalmology, orthopedic surgery, otolaryngology, plastic surgery, urology, and vascular surgery. Participants ranged from PGY-1 though PGY-5, with representation from all PGY-levels. Participants reported their PGY-level based on number of clinical years of surgical residency training completed with research time omitted. The mean (SD) age was 30.8 (3.2) years; 51.1% of the participants were female; 2.3% identified as Native American, 9.3% as Latino, 9.3% Black or African American, 34.9% Asian, and 39.5% White (Table 1). Two participants preferred not to report their ethnicity. The self-reported prior video game experience

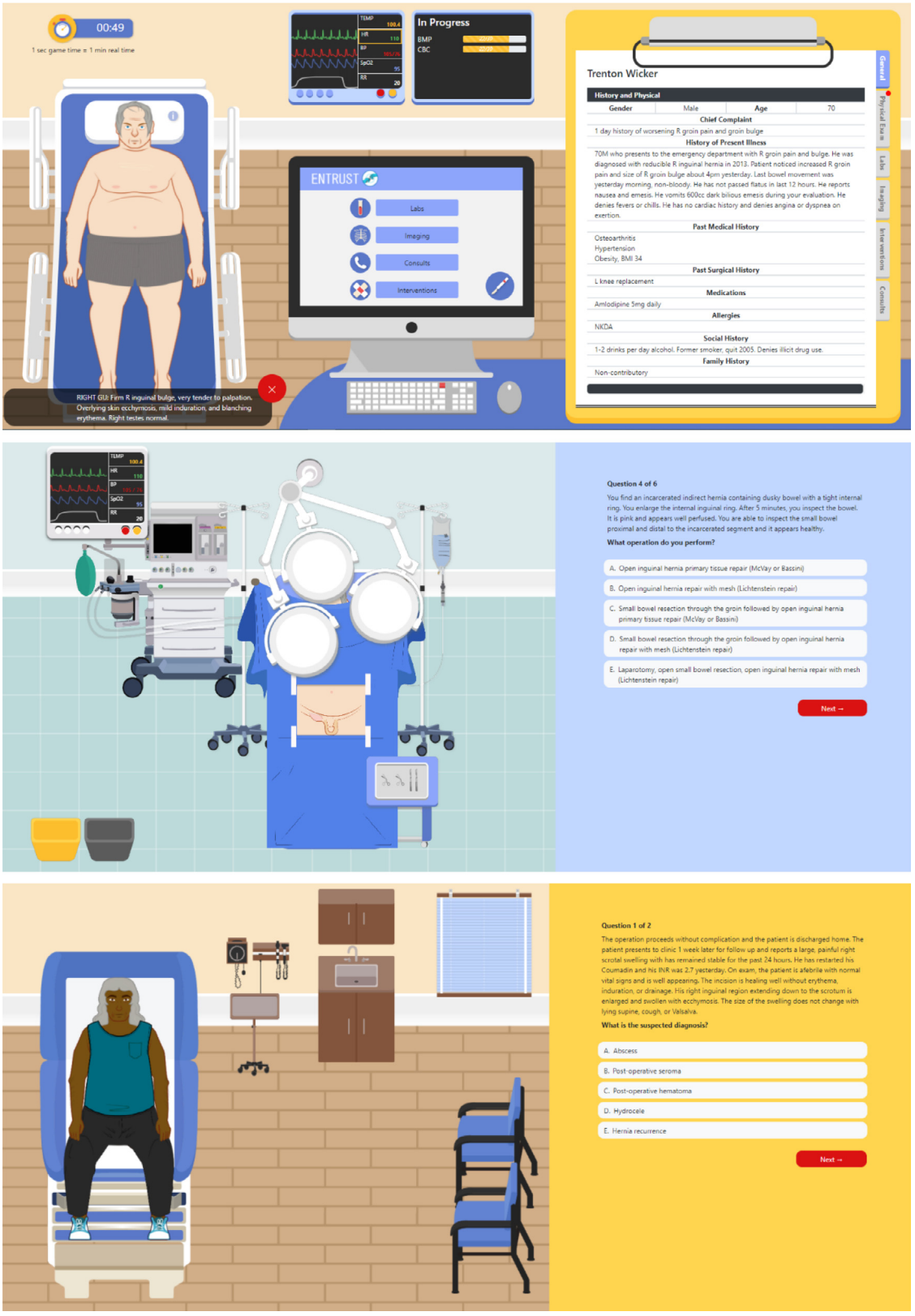

FIGURE 1. ENTRUST Assessment platform. Preoperative Simulation Mode (top panel), Intraoperative Question Mode (middle panel), and Postoperative Question Mode (lower panel). Images are taken from various case scenarios.

TABLE 1. Demographics of Study Participants

Characteristic	n = 43
Age (y), mean (SD)	30.8 (3.2)
Sex, n (%)	
Female	22 (51.1)
Male	21 (48.9)
Race/ethnicity, n (%)	
Asian	15 (34.9)
Black or African American	4 (9.3)
Latino	4 (9.3)
Native American	1 (2.3)
White	17 (39.5)
Missing or prefer not to state	2 (4.7)
PGY-Level, n (%)	
PGY-1	17 (39.5)
PGY-2	11 (25.6)
PGY-3	9 (20.9)
PGY-4	2 (4.7)
PGY-5	4 (9.3)
General surgery resident status, n (%)	
General surgery categorical	27 (62.8)
General surgery nondesignated preliminary	8 (18.6)
Designated preliminary[†]	8 (18.6)
Prior video game experience (h/wk), mean (SD)	1.4 (3.1)

Values reported as n (%) or mean (SD).
PGY, Post-Graduate Year; SD, standard deviation.
[†]Includes PGY-1 or PGY-2 cardiothoracic surgery, ophthalmology, orthopedic surgery, otolaryngology, plastic surgery, urology, and vascular surgery trainees in the general surgery residency program.

of the participants ranged from 0 to 15 hours per week with mean 1.4 (SD 3.1) hours.

CONTENT EVIDENCE

Case Creation

Four case scenarios for inguinal hernia were authored and iteratively developed by the authors, aligned with EPA descriptions and essential functions for inguinal hernia outlined by the American Board of Surgery.[3] The case, including all multiple-choice questions, was authored by a board-certified general surgeon with formal training in surgical education. The case content and multiple-choice questions were then reviewed and discussed by an expert panel (n = 5) of board-certified general surgeons representing a variety of practice settings. The case was iteratively revised based on this feedback, with the final case scenario reviewed and approved by the authors.

Scoring Algorithm

The scoring algorithm was developed by 2 board-certified surgeons with formal training in surgical education to reflect appropriateness of clinical interventions and multiple-choice question responses. Diagnostic studies and interventions were categorized using the following framework: critical [+200], indicated [+100], optional [0], not indicated but not harmful [-50], mild to moderate harm [-100], severe harm [-200], and death/cardiac arrest [-500]. Multiple choice questions were awarded +200 points for correct responses and -200 for incorrect responses. Points were additionally deducted [-200] for each instance of failure to address and correct vital sign abnormalities. All case scenarios and scoring algorithm were Beta-tested by the research team prior to data collection to ensure case functionality.

RELATIONSHIP TO OTHER VARIABLES

ENTRUST Inguinal Hernia EPA Assessment Grand Total Score was positively correlated with PGY level (Fig. 2: rho = 0.64, $p < 0.0001$). Preoperative, Intraoperative, and Postoperative Total Scores were also positively correlated with PGY-level (Preoperative: rho = 0.51, p = 0.0005, Intraoperative: rho = 0.50, p = 0.0006, Postoperative: rho: 0.32, p = 0.038) (Table 2). Total Case Scores were positively correlated with PGY-level for cases representing elective unilateral inguinal hernia (rho = 0.51, p = 0.0004), strangulated inguinal hernia (rho = 0.59, $p < 0.0001$), and elective bilateral inguinal hernia (rho = 0.52, p = 0.0003) (Fig. 3A). No statistically significant difference was found in acutely incarcerated inguinal hernia case total score by PGY-level (Fig. 3A: rho = 0.10, p = 0.50). Descriptive statistics for all ENTRUST Inguinal Hernia EPA Assessment scores are shown in Table 2.

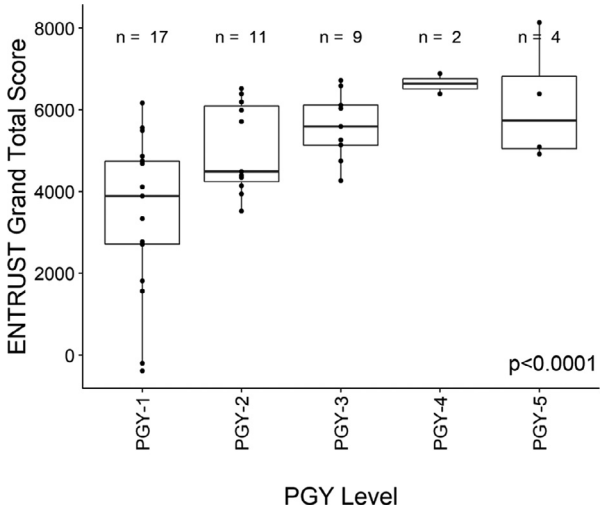

FIGURE 2. ENTRUST Inguinal Hernia EPA Assessment grand total score by PGY-Level.

TABLE 2. ENTRUST Inguinal Hernia EPA Assessment Score Performance Descriptive Statistics

Score	PGY-1 (n = 17)	PGY-2 (n = 11)	PGY-3 (n = 9)	PGY-4 (n = 2)	PGY-5 (n = 4)	p-Value
Grand total score, median [IQR]	3890 [2715,4745]	4490 [4245,6093]	5595 [5140,6120]	6640 [6515,6765]	5743 [5051,6828]	<0.0001
Preoperative total score	1915 [1115,2490]	2190 [1568,2718]	2465 [2265,2920]	2840 [2715,2965]	2743 [2451,3128]	0.007
Intraoperative total score	1800 [600,2490]	2600 [2000,2718]	2200 [1800,2920]	3000 [3000,2965]	2400 [2100,3128]	0.0006
Postoperative total score	400 [400,800]	400 [400,800]	800 [400,800]	800 [800,800]	800 [700,800]	0.038
Case scenario scores, median [IQR]						
Elective unilateral inguinal hernia	1200 [700,1400]	1450 [1175,1575]	1800 [1500,1850]	1850 [1800,1900]	1400 [1238,1575]	0.0004
Preoperative subscore	500 [315,695]	500 [325,600]	650 [600,700]	650 [600,700]	600 [358,675]	0.004
Intraoperative subscore	400 [0,800]	800 [400,1000]	800 [400,800]	800 [800,800]	400 [300,500]	0.063
Postoperative subscore	400 [0,400]	400 [0,400]	400 [400,400]	400 [400,400]	400 [400,400]	0.036
Acutely incarcerated inguinal hernia[†]	1465 [920,1690]	1340 [1290,1443]	1565 [1295,1790]	1390 [1365,1415]	1268 [1126,1578]	0.50
Preoperative subscore	540 [315,695]	690 [380,780]	740 [695,765]	790 [765,815]	730 [689,778]	0.0066
Intraoperative subscore	1000 [600,1000]	1000 [600,1000]	1000 [600,1000]	600 [600,600]	600 [500,800]	0.78
Strangulated inguinal hernia[†]	1000 [650,1600]	1650 [1075,2025]	1800 [1600,1950]	2200 [2100,2300]	2275 [2150,2338]	<0.0001
Preoperative subscore	700 [300,850]	900 [450,1125]	800 [600,1100]	1000 [900,1100]	1150 [950,1313]	0.007
Intraoperative subscore	800 [0,800]	800 [600,1000]	800 [800,1200]	1200 [1200,1200]	1200 [1100,1200]	0.0007
Elective bilateral inguinal hernia	400 [0,700]	700 [350,1125]	800 [400,1200]	1200 [0,1200]	975 [788,1263]	0.0003
Preoperative subscore	300 [250,400]	350 [300,375]	400 [400,400]	400 [400,400]	375 [350,400]	0.008
Intraoperative subscore	-400 [-400,0]	0 [0,400]	400 [0,400]	400 [400,400]	400 [300,500]	0.0002
Postoperative subscore	400 [0,400]	400 [200,400]	400 [0,400]	400 [400,400]	400 [300,400]	0.299

Values reported as median [IQR].
IQR, interquartile range.
[†]Case scenario did not include postoperative phase of questioning.

For each of the 4 case scenarios, Preoperative Sub-Score and Intraoperative Sub-Score were additionally analyzed by PGY-level. Preoperative Sub-Scores were significantly correlated with PGY-level for all cases: elective unilateral inguinal hernia (rho = 0.43, p = 0.004), acutely incarcerated inguinal hernia (rho = 0.41, p = 0.0066), strangulated inguinal hernia (rho = 0.40, p = 0.007), and elective bilateral inguinal hernia (rho = 0.40, p = 0.008) (Fig. 3B). Intraoperative Sub-Scores were significantly correlated with PGY-level for the strangulated inguinal hernia (rho = 0.50, p = 0.0007) and elective bilateral inguinal hernia (rho = 0.54, p = 0.0002) case scenarios, but was not statistically significant for elective unilateral or acutely incarcerated inguinal hernia cases (Fig. 3C).

Median Grand Total Score for PGY-1 categorical general surgery trainees was higher than PGY-1 non-categorical trainees (5190 vs 3178, p = 0.014). There was no statistically significant difference in score performance between PGY-2 categorical and noncategorical surgery trainees (6040 vs 4243, p = 0.23).

For the critical clinical decision-making choice of whether to attempt manual reduction of an acutely incarcerated inguinal hernia in the emergency department, this was performed correctly by 100% of PGY-3 through PGY-5 residents, 88% of PGY-2 residents, and 67% of PGY-1 residents (Fig. 4A). Unadjusted Total Case Score and Preoperative Sub-Score for the acutely incarcerated inguinal hernia case were both significantly higher for those trainees correctly attempting manual reduction (p = 0.007 and p < 0.0001, respectively). However, these differences in Total Case Score and Preoperative Sub-Score were not statistically significant when scores were adjusted to remove the scoring impact of the decision to manually reduce the incarcerated hernia (p = 0.11 and p = 0.17, respectively).

For the decision of whether to attempt manual reduction of a strangulated inguinal hernia, this was performed correctly by 100% of PGY-3, PGY-4, and PGY-5 residents, 91% of PGY-2 residents, and 75% of PGY-1 residents (Fig. 4B). Unadjusted Total Case Score and

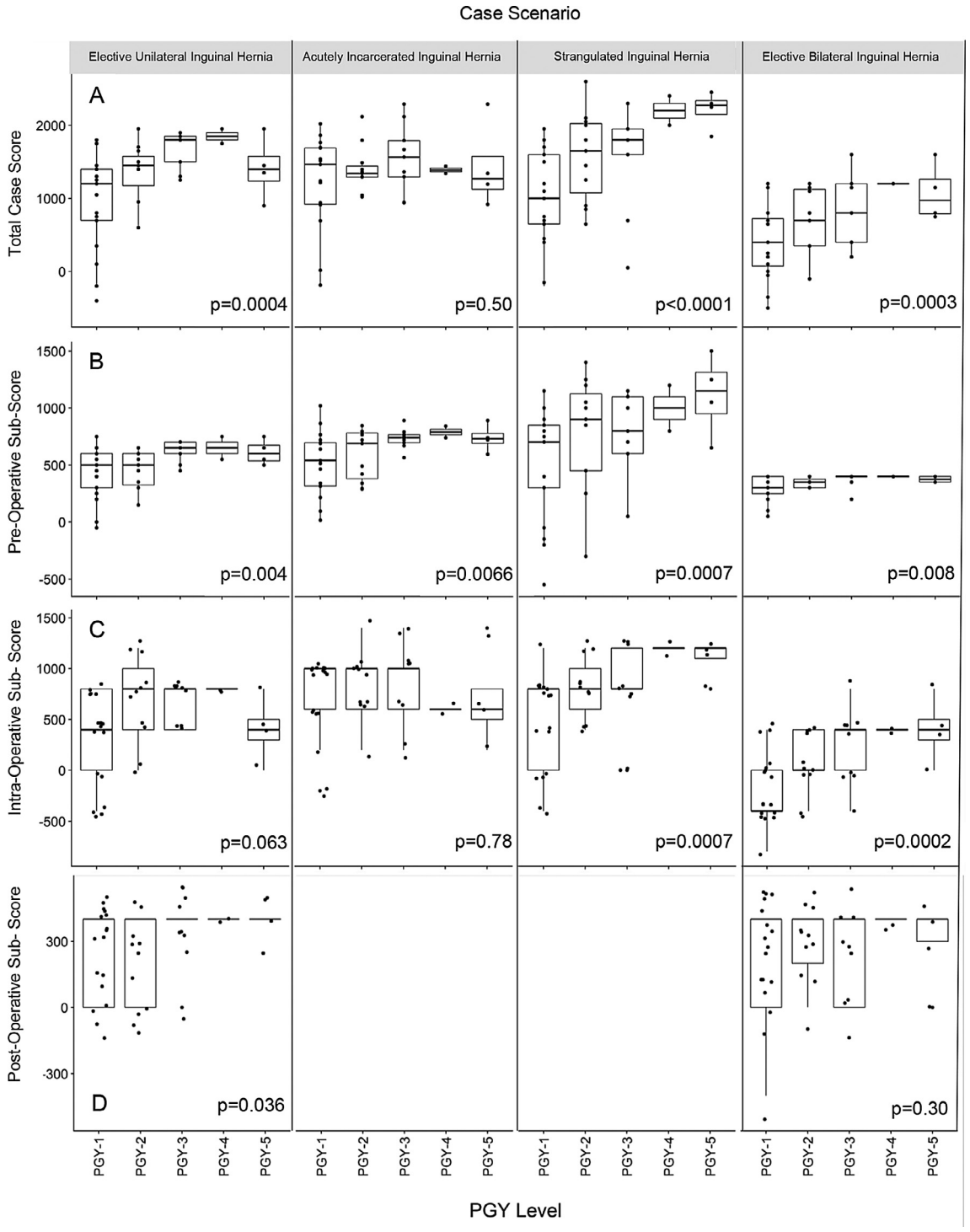

FIGURE 3. ENTRUST Inguinal Hernia EPA Assessment Case Scenario Total and Sub-Scores by PGY-Level. Total Case Score (A). Preoperative Simulation Mode Sub-Scores (B). Intraoperative Question Mode Sub-Scores (C). Postoperative Question Mode Sub-Scores (D). The Acutely Incarcerated Inguinal Hernia and Strangulated Inguinal Hernia Case Scenarios did not include a Postoperative Question Mode.

Preoperative Sub-Score for the strangulated inguinal hernia case were significantly higher for those trainees correctly deciding not to attempt manual reduction ($p = 0.009$ and $p = 0.0019$, respectively). After adjustment to remove the scoring impact of the decision to manually reduce the strangulated hernia, a statistically

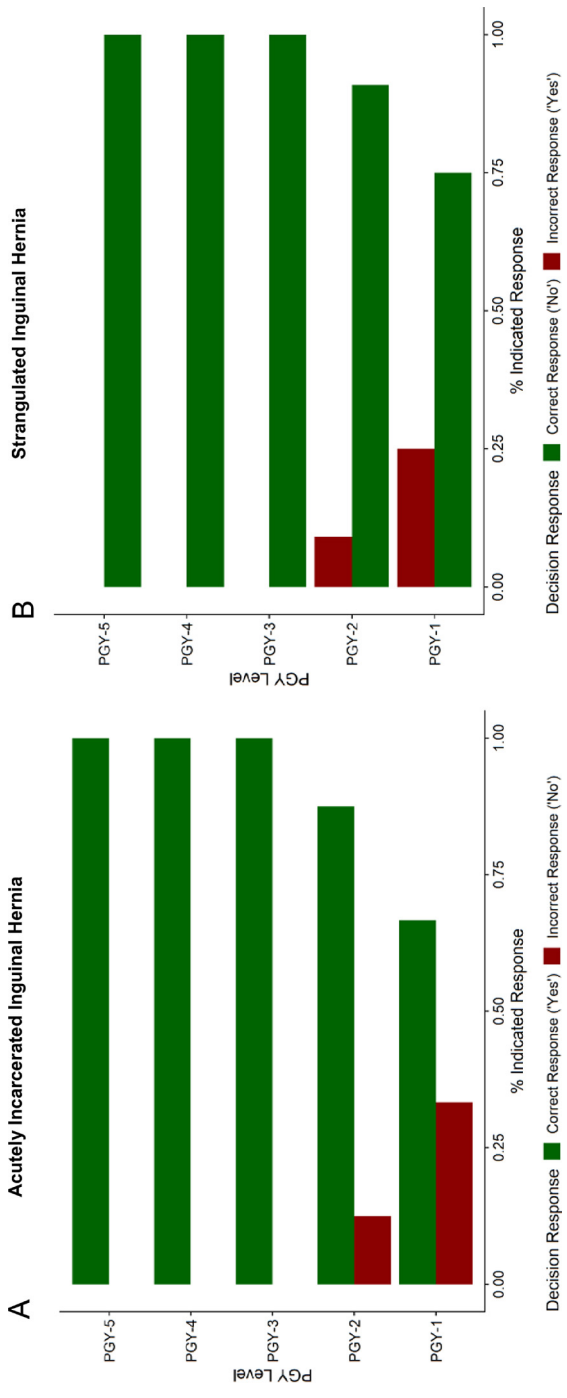

FIGURE 4. ENTRUST critical decision-making for manual hernia reduction by PGY-level. Following physical examination of the groin and review of physical findings, participants were queried during the Preoperative Simulation Mode for the Acutely Incarcerated Inguinal Hernia Case Scenario (A) and Strangulated Inguinal Hernia Case Scenario (B) whether they would attempt manual reduction of the hernia at the bedside.

significant difference in preoperative sub-score remained between those who attempted reduction and those who did not attempt reduction ($p = 0.032$).

Grand Total Score and Intraoperative Total Score were correlated with self-reported prior inguinal hernia operative experience (Fig. 5A: rho = 0.65, $p < 0.0001$ and Fig. 5B: rho = 0.59, $p < 0.0001$, respectively). Prior video game experience did not correlate with performance on ENTRUST (rho = 0.094, $p = 0.56$).

DISCUSSION

There has been a widespread initiative to adopt and incorporate EPAs in graduate medical training as a means of transitioning toward a more competency-based educational paradigm. In 2018, the ABS initiated a nationwide pilot tasking 28 general surgery programs to explore the use and implementation of 5 core general surgery EPAs, with the intention of formalizing EPAs as a requirement for all general surgery training programs by 2023.[14] The determination of readiness for entrustment is typically predicated upon direct observation and assessment of behaviors by faculty in the clinical setting. While frequent, real-time microassessments are ideal in assessment of EPAs and readiness for entrustment, this approach places a sizeable and continuous burden on faculty to regularly complete evaluations for the many individual interactions they have with multiple trainees who are to be graded across a variety of clinical skills and EPAs. In addition, there is variability in the types and severity of patient cases encountered in the real-world clinical setting, making it difficult to reliably evaluate trainees' ability to manage rare diseases or complications. Virtual patient simulations enable trainees to demonstrate their clinical and surgical decision-making in an objective, reproducible, and measurable way while decompressing the assessment burden off faculty raters. In addition, standardized scenarios may be deployed to minimize implicit bias and subjectivity, and test infrequently encountered, yet critical, clinical conditions.

Given these challenges, many pilot institutions have operationalized EPAs by adopting reductionistic approaches and focusing on assessment of operative performance only, as readily available tools exist to measure this construct. One mobile operative microassessment application, SIMPL (System for Improving and Measuring Procedural Learning),[5,6,15] has been widely utilized by surgical training programs to rate trainee's technical skills. While it possesses robust validity evidence for evaluating operative autonomy,[5,6,16] it does not assess clinical decision-making. However, based on the EPA definitions and essential functions articulated by the ABS, clinical decision-making competence in the

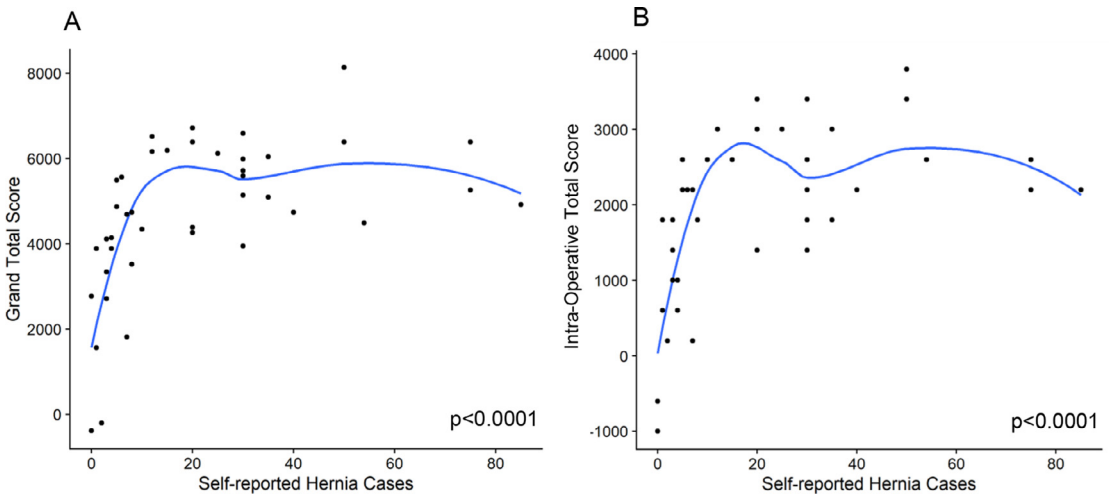

FIGURE 5. Correlation of ENTRUST inguinal hernia EPA score performance to self-reported inguinal hernia operative case experience. Grand Total Score (A). Intraoperative Total Score (B).

preoperative, intraoperative, and postoperative setting constitutes critical components of entrustment. Therefore, readiness for entrustment should include assessment of both operative autonomy and clinical decision-making. Therefore, there is a great need for evidence-based EPA-aligned tools that specifically address clinical decision-making, as a complement to existing technical skills evaluations.

To address this need for an objective, efficient, and scalable means to assess clinical and surgical decision-making, we developed ENTRUST, a virtual patient authoring and assessment platform to deploy rigorous, case-based patient simulations for evaluation of EPAs. We iteratively developed an ENTRUST Inguinal Hernia EPA Assessment, which was vetted by expert consensus for suitability and accuracy of content. The case content was uploaded by surgical educators without any background in software engineering via the ENTRUST authoring portal and deployed as an assessment. This pilot experience verifies the usability and functionality of both the ENTRUST authoring portal and assessment platform.

Our pilot data indicates that ENTRUST score performance is correlated to PGY-level and inguinal hernia operative experience. There was a statistically significant increase in total score with successively higher PGY-level. This trend was observed for Grand Total Score, Preoperative Total Score, Intraoperative Total Score, and individual Total Case Scores. However, while surgical decision-making skills tend to develop over time with increasing PGY-level, it is not a strictly time-based construct, and the variation in score within PGY-level may be explained by differences in clinical decision-making ability and readiness for entrustment. Theoretically, a junior resident with high ENTRUST performance who objectively demonstrates surgical decision-making competence may be entrusted with greater autonomy earlier than a senior resident with low ENTRUST score performance for a particular EPA domain. Thus, ENTRUST has potential to be employed as a tool to inform entrustment decisions as surgical training shifts from a time-based model toward a competency-based paradigm.

As demonstrated by the clinical decision-making surrounding whether or not to attempt manual reduction of an incarcerated or strangulated inguinal hernia, ENTRUST also holds potential to evaluate and query specific key surgical decision-making points important in determining readiness or lack of readiness for entrustment. By logging all trainee actions and querying specific decisions, ENTRUST may assist program directors and surgical educators in assigning ABS EPA Levels, independent of PGY-level. This information can be used to inform decisions on entrustment and autonomy.

This study provides initial validity evidence for use of ENTRUST as an objective measure of surgical decision-making for EPAs. Content evidence for the case scenarios was established by alignment of case content with published ABS EPA descriptions and essential functions,[3] expert review, and group consensus of case content and scoring algorithm. The ability of the ENTRUST assessment to discriminate between PGY-levels, as well as its correlation to inguinal hernia operative case experience provides evidence of its relationship to other established variables in surgical education. Importantly, there was no difference in score performance based on prior video game experience. Limitations of this pilot study include its single institution design and self-reported inguinal hernia operative experience.

Future directions include collection and analysis of additional validity evidence for ENTRUST using Messick's unified framework of construct validity, including response process evidence, internal structure, and consequences.[11] In future studies, we intend to further investigate relationship to other variables such as ACGME Case Logs, American Board of Surgery (ABS) Inservice Training Exam (ABSITE) scores, Accreditation Council for Graduate Medical Education (ACGME) Milestones, and ABS board pass rates. Additionally, we plan to correlate performance on ENTRUST to individual trainee performance on microassessments from actual clinical interactions such as SIMPL or other platforms. Results from this pilot will inform the design of future multi-institutional studies featuring a larger set of case scenarios for the Inguinal Hernia EPA to further collect validity evidence, conduct standard setting, and map game play patterns and specific key decision-making actions to EPA levels and readiness for entrustment. Future plans include expansion of the ENTRUST platform to encompass all ABS General Surgery EPAs as well as development of additional environments, assets, and functionality to further evaluate trainees' readiness for entrustment and accommodate higher acuity case scenarios situated in the trauma bay and ICU settings.

CONCLUSION

ENTRUST demonstrates feasibility and initial validity evidence for objective assessment of surgical decision-making for the inguinal hernia EPA. The ENTRUST authoring and assessment platform holds potential to inform readiness of entrustment for American Board of Surgery EPAs in the future and to support the ongoing transformation of surgical education to a competency-based paradigm.

ACKNOWLEDGMENTS

The authors would like to thank Yulin Cai, Ruonan Chen, Samuel Shields, Ananya Anand, and Sherry Wren for their contributions to the development of the ENTRUST platform.

REFERENCES

1. ten Cate O, Scheele F. Competency-based postgraduate training: can we bridge the gap between theory and clinical practice? *Acad Med J Assoc Am Med Coll*. 2007;82:542–547. https://doi.org/10.1097/ACM.0b013e31805559c7.

2. Ten Cate O, Chen HC, Hoff RG, Peters H, Bok H, van der Schaaf M. Curriculum development for the workplace using Entrustable Professional Activities (EPAs): AMEE Guide No. 99. *Med Teach*. 2015;37:983–1002. https://doi.org/10.3109/0142159X.2015.1060308.

3. Brasel KJ, Klingensmith ME, Englander R, et al. Entrustable professional activities in general surgery: development and implementation. *J Surg Educ*. 2019;76:1174–1186. https://doi.org/10.1016/j.jsurg.2019.04.003.

4. ABS E-News. Accessed January 12, 2021. http://www.absurgery.org/quicklink/absnews/absupdate0518.html#epa

5. George BC, Bohnen JD, Schuller MC, Fryer JP. Using smartphones for trainee performance assessment: a SIMPL case study. *Surgery*. 2020;167:903–906. https://doi.org/10.1016/j.surg.2019.09.011.

6. Bohnen JD, George BC, Williams RG, et al. The feasibility of real-time intraoperative performance assessment with SIMPL (System for Improving and Measuring Procedural Learning): early experience from a multi-institutional trial. *J Surg Educ*. 2016;73:e118–e130. https://doi.org/10.1016/j.jsurg.2016.08.010.

7. Sandhu G, Nikolian VC, Magas CP, et al. OpTrust: validity of a tool assessing intraoperative entrustment behaviors. *Ann Surg*. 2018;267:670–676. https://doi.org/10.1097/SLA.0000000000002235.

8. Nikolian VC, Sutzko DC, Georgoff PE, et al. Improving the feasibility and utility of OpTrust-A tool assessing intraoperative entrustment. *Am J Surg*. 2018;216:13–18. https://doi.org/10.1016/j.amjsurg.2017.10.036.

9. Vassiliou MC, Feldman LS, Andrew CG, et al. A global assessment tool for evaluation of intraoperative laparoscopic skills. *Am J Surg*. 2005;190:107–113. https://doi.org/10.1016/j.amjsurg.2005.04.004.

10. Sánchez R, Rodríguez O, Rosciano J, et al. Robotic surgery training: construct validity of Global Evaluative Assessment of Robotic Skills (GEARS). *J Robot Surg*. 2016;10:227–231. https://doi.org/10.1007/s11701-016-0572-1.

11. Messick S. Standards of validity and the validity of standards in performance assessment. *Educ Meas Issues Pract*. 1995;14:5–8. https://doi.org/10.1111/j.1745-3992.1995.tb00881.x.

12. Cook DA, Zendejas B, Hamstra SJ, Hatala R, Brydges R. 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. https://doi.org/10.1007/s10459-013-9458-4.

13. *R Core Team*. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2020. Accessed March 25, 2022. https://www.r-project.org.

14. New model of surgical resident autonomy coming in 2023. *ACS Clinical Congress News*. 2021. Published October 23. Accessed January 21, 2022. https://www.acsccnews.org/new-model-of-surgical-resident-autonomy-coming-in-2023/.

15. George BC, Bohnen JD, Williams RG, et al. Readiness of US general surgery residents for independent practice. *Ann Surg*. 2017;266:582–594. https://doi.org/10.1097/SLA.0000000000002414.

16. 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. https://doi.org/10.1016/j.jsurg.2019.01.012.