Chapter 5
ENTRUST: Co-design and Validation of a Serious Game for Assessing Clinical Decision-Making and Readiness for Entrustment

Edward F. Melcer, Cara A. Liebert, Samuel Shields, Oleksandra G. Keehl, Jason Tsai, Fatyma Camacho, Hyrum Eddington, Amber Trickey, Melissa Lee, Sylvia Bereknyei Merrell, James R. Korndorffer Jr., and Dana T. Lin

Abstract Graduate medical education is moving toward a competency-based paradigm, predicated upon multiple real-time assessments to verify clinical and technical proficiency (i.e., readiness for entrustment of residents). This requires not only assessment of technical skills and medical knowledge but also critical clinical decision-making skills in preoperative, intraoperative, and postoperative settings. However, most medical education programs have adopted reductionist approaches, reducing assessment of readiness for entrustment to only assessing technical skill performance. As such, there is a growing need for tools that can provide more comprehensive and objective evaluations of the proficiency of residents to perform medical procedures. This chapter presents *ENTRUST*, our serious game-based online platform to assess trainees' decision-making competence across various Entrustable Professional Activity (EPA) domains. Specifically, we discuss (1) the design of *ENTRUST*; (2) insights identified and lessons learned throughout the development process that can aid collaboration between serious game developers and subject matter experts; and (3) results from a pilot study of *ENTRUST*—demonstrating the tool's capability to discriminate between levels of surgical expertise and providing initial validity evidence for its use as an objective assessment for clinical decision-making.

E. F. Melcer (✉) · S. Shields · O. G. Keehl · J. Tsai · F. Camacho
University of California, Santa Cruz, Santa Clara, CA, USA
e-mail: eddie.melcer@ucsc.edu; samshiel@ucsc.edu; okeehl@ucsc.edu; ctsai32@ucsc.edu; fcamach1@ucsc.edu

C. A. Liebert · H. Eddington · A. Trickey · M. Lee · S. B. Merrell · J. R. Korndorffer Jr. · D. T. Lin
Stanford University School of Medicine, Stanford, CA, USA
e-mail: carap@stanford.edu; hyrumedd@stanford.edu; atrickey@stanford.edu; melchlee@stanford.edu; sylviab@stanford.edu; korndorffer@stanford.edu; danalin@stanford.edu

K. M. L. Cooper, A. Bucchiarone (eds.), *Software Engineering for Games in Serious Contexts*, https://doi.org/10.1007/978-3-031-33338-5_5

Keywords Game-based assessment · Games for Health · Serious game design · Co-design · Clinical decision-making · Entrustable Professional Activities

5.1 Introduction

In recent years, medical education has moved toward a competency-based paradigm predicated upon multiple, real-time assessments to verify proficiency [43]. Within this new paradigm, 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 [44]. EPAs are effective tasks or responsibilities to be entrusted to a trainee once they have attained competence at a specific level and embody a more global integration of the Accreditation Council for Graduate Medical Education (ACGME) core competencies [43]. Notably, 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 American Board of Surgery (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 [31].

The determination of readiness for entrustment is typically predicated upon direct observation and assessment of behaviors by faculty in the clinical setting [11]. 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 [45]. Conversely, 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 [4]. In addition, standardized scenarios may be deployed to minimize implicit bias and subjectivity, reduce test anxiety, and test infrequently encountered, yet critical, clinical conditions [27, 49].

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, e.g., [6, 15, 18, 32, 37, 38, 47]. One mobile operative microassessment application, SIMPL (System for Improving and Measuring Procedural Learning) [6, 17, 18], has been widely utilized by surgical training programs to rate trainee's technical skills. While it possesses robust validity evidence for evaluating operative autonomy [6, 15, 18], 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 preoperative, intraoperative, and postoperative setting constitutes critical components of entrustment. As a result, 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 serious game-based assessment platform to deploy rigorous, case-based patient simulations for evaluation of EPAs. In this chapter, we present (1) the design of *ENTRUST*; (2) insights identified and lessons learned throughout the development process that can aid collaboration between serious game developers and subject matter experts; and (3) results from a pilot study of *ENTRUST*—demonstrating its capability to discriminate between levels of surgical expertise and providing initial validity evidence for its use as an objective assessment for clinical decision-making.

5.2 Background

5.2.1 Entrustable Professional Activities

In 2018, the ABS commenced a multi-institutional pilot to implement five general surgery EPAs, each with defined levels of entrustment from Level 0 to Level 4, in surgical residency [1, 7]. These initial five 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 healthcare providers [7]. Additionally, 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 intraoperative assessment of technical skills and operative autonomy [6, 18, 32, 37, 38, 47], they do not directly not assess clinical decision-making across the preoperative, intraoperative, and postoperative settings. The assessment of technical skills is necessary, but is not sufficient, to determine entrustment [45]. Therefore, there is a notable gap in the literature and 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 intraoperative evaluations.

5.2.2 Game-Based Assessment in the Health Domain

Educational assessment has evolved over the past decade from traditional pen-and-paper-based tests to the use of technology such as games to assess various competencies in the form of game-based assessment [48]. Notably, due to the technological enhancement of what can be measured, game-based assessment provides promising possibilities for more valid and reliable measurement of students' skills, knowledge, and attributes compared to the traditional methods of assessment such as paper-and-pencil tests or performance-based assessments [13]. Within the health domain, game-based assessment has been utilized in a variety of contexts including assessment of patient health [46], assessment of motor skills and ability to perform first aid [9], neuropsychological assessment [16], and assessment of health-related knowledge/learning [33], to name a few. However, in the context of clinical reasoning and decision-making, the predominant focus of serious games has been on training and learning, e.g., [10, 20, 21, 23, 24, 26, 29, 30]. This surprising lack of game-based assessment for clinical reasoning and decision-making highlights the notable gap in the literature and further emphasizes the need for evidence-based EPA-aligned game-based assessment tools that specifically address clinical decision-making—such as *ENTRUST*. Furthermore, such game-based assessment tools offer a number of potential benefits over traditional forms of assessment if employed correctly including reduced test anxiety [27] and more authentic contexts for assessing competency, which is crucial for acquiring more accurate assessments of skill [40].

5.3 Design of *ENTRUST*

ENTRUST is a serious game-based online virtual patient simulation platform to both train and assess medical trainees' decision-making competence within EPAs. It is therefore targeted at training and assessing the competency of the next generation of clinicians at the medical student and resident levels.

5.3.1 Co-design Process

We utilized a co-design approach for the design and development of *ENTRUST*—which is a widely used approach within the health field [41]. Co-design stems from participatory design, where the people destined to use the system play a critical role in designing it [39]. However, in a co-design process, stakeholders are treated as equal collaborators or can even take the lead in the design process rather than have limited roles [42]. In this way, co-design involves a shift in the locus of responsibility and control so that "clients" or users of services become active partners in designing, shaping, and resourcing services, rather than being passive recipients of pre-determined services [8]. For *ENTRUST*, we worked directly with medical education experts and continue to do so as co-designers (i.e., full partners in the entire design process [41]) on the project. We found this approach to be critical for the successful design and development of *ENTRUST* as the subject matter of clinical decision-making and entrustment is too complex for a serious game development team to successfully design, develop, and maintain on their own. As such, our research team utilizes the following co-design and agile development process (with a number of the steps drawn from [5]):

1. Contextual inquiry in the form of informational interviews and weekly artifact review meetings with medical education experts to identify latent needs, challenges experienced, and desired future state/artifact creation.
2. Generation of design and rapid prototyping to address identified needs and challenges. This is done through the development of new *ENTRUST* artifacts (e.g., creating an authoring platform to complement the game-based assessment tool) or incorporation of desired features into existing artifacts (e.g., adding a new vital sign algorithm to the simulation mode and authoring platform). Repeat steps 1 and 2 weekly.
3. Sharing ideas and receiving feedback through periodic presentations of design and development work on *ENTRUST* to larger subsections of the medical education community.
4. Conducting studies and data analysis to empirically validate *ENTRUST* designs.
5. Interpreting results for requirements translation, i.e., identifying action items, feasible priorities, and feeding back into steps 1 and 2.

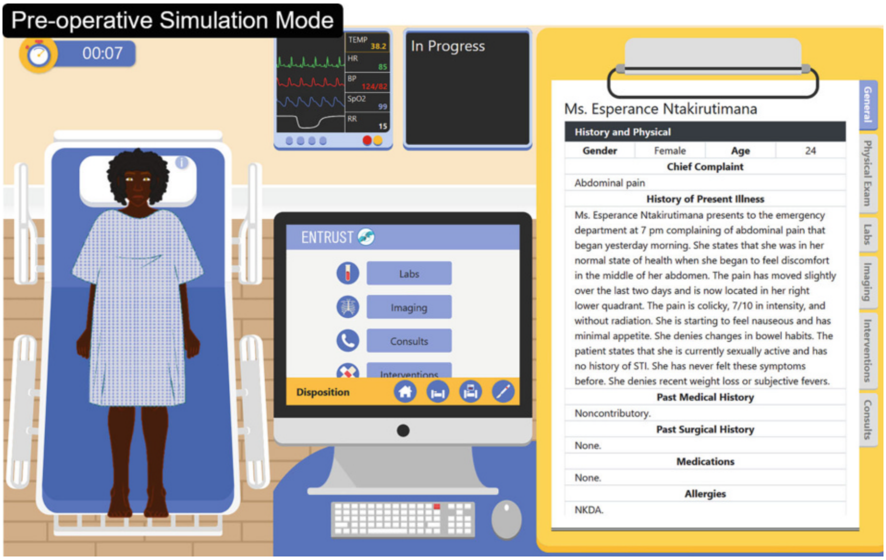
This co-design process has resulted in the current iteration of *ENTRUST* as described below.

5.3.2 Assessment Platform

The current *ENTRUST* platform includes two primary phases: simulation mode and question mode.

5.3.2.1 Simulation Mode

In simulation mode (see Fig. 5.1), the examinee engages with patient case scenarios starting from the preoperative setting. This setting can be in either the emergency department or the outpatient clinic, where the examinee initiates a physical examination and full workup of the patient. During workup, the examinee can order diagnostic tests, administer fluids and medications, perform bedside procedures, and request consultation. All actions –both player evoked (such as conducting a physical exam) and game evoked (such as changing vital signs due to deteriorating



Ms. Esperance Ntakirutimana

History and Physical

Gender Female **Age** 24

Chief Complaint

Abdominal pain

History of Present Illness

Ms. Esperance Ntakirutimana presents to the emergency department at 7 pm complaining of abdominal pain that began yesterday morning. She states that she was in her normal state of health when she began to feel discomfort in the middle of her abdomen. The pain has moved slightly over the last two days and is now located in her right lower quadrant. The pain is colicky, 7/10 in intensity, and without radiation. She is starting to feel nauseous and has minimal appetite. She denies changes in bowel habits. The patient states that she is currently sexually active and has no history of STI. She has never felt these symptoms before. She denies recent weight loss or subjective fevers.

Past Medical History

Noncontributory.

Past Surgical History

None.

Medications

None.

Allergies

NKDA.

Fig. 5.1 The simulation mode within *ENTRUST*. Enables examinees to engage with patient case scenarios starting at the preoperative setting, including physical examination and full patient workup

patient condition)– are recorded and scored on the back-end database according to an expert-consensus-derived scoring algorithm (see Sect. 5.3.4). Points are earned for ordering relevant labs and key interventions; conversely, points are lost for performing inappropriate, unnecessary, or harmful actions.

Notably, the *ENTRUST* interface in this mode consists of six key features that enable examinee input for assessment and provide feedback from the simulation:

1. **Timer** (Fig. 5.1 Top Left)—the timer displays the amount of time the examinee has been active in the preoperative setting. During play, 1 second of game time displayed on the timer equates to 1 minute of time taken in a real-world scenario.
2. **Patient/physical exam** (Fig. 5.1 Middle Left)—the virtual patient enables examinees to conduct a physical examination and see results in the medical chart. As examinees move their mouse over the virtual patient, various icons and images will appear to indicate that a physical examination can be conducted on that part of the body with a mouse click. Patient facial expressions also change depending on their health status throughout the course of the preoperative setting.
3. **Notifications** (Fig. 5.1 Bottom Left)—notifications appear in the bottom-left corner of the screen after each physical examination to report the results. This is done to remove the need to go to the right side of the screen to view a physical exam result in the medical chart before returning to continue examining the patient on the left side of the screen, i.e., to reduce extrinsic cognitive load [22, 34].

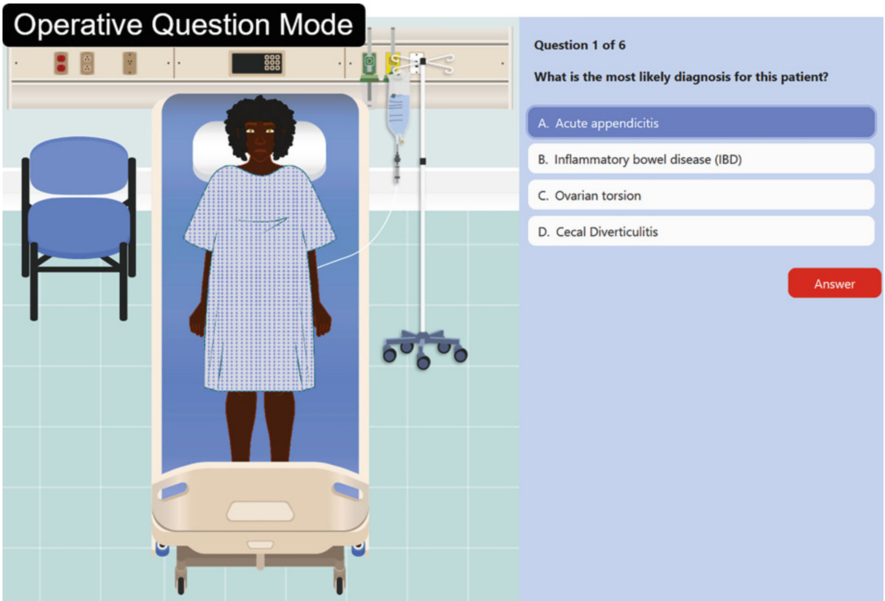
4. **Vital monitor and order progress monitor** (Fig. 5.1 Top Middle)—the vital monitor shows the virtual patient's vital signs throughout the preoperative simulation. Vitals are updated in real time (relative to game time) and can deteriorate due to lack of or improper treatment as well as improve due to performing appropriate bedside procedures or administering appropriate fluids or medications. An audible alarm (similar sounding to real-world vital machine alarms) can also be heard when patient vitals reach a dangerous level. The order progress monitor shows the time remaining for any diagnostic test, administration of fluids and medications, bedside procedures, or consultations ordered. The exact amount of seconds remaining is shown in the progress bar and mirrors typical real-world times taken for each order at a rate of one game second to one real-world minute.
5. **Order console** (Fig. 5.1 Bottom Middle)—the order console enables the examinee to order diagnostic tests, administer fluids/medications, perform bedside procedures, and request consultation. It also allows the examinee to make decisions about disposition, e.g., whether the patient should go home, to the operating room (OR), or to the intensive care unit (ICU) or proceed with nonoperative management. Selecting a disposition or causing the patient to go into cardiac arrest will proceed to the question mode of *ENTRUST*.
6. **Medical chart** (Fig. 5.1 Right)—the medical chart maintains and displays all relevant information regarding the virtual patient. This includes their medical history and initially reported health complaint as well as the results from all physical exams and orders placed. Examinees can click the tabs on the right side of the chart to toggle between this information. Whenever there is a change to the medical chart, such as when a physical exam or order is completed, the corresponding tab displays a red dot to indicate new information is available.

5.3.2.2 Question Mode

ENTRUST switches to question mode (Fig. 5.2) when the examinee opts to proceed to the operating room. In question mode, the examinee is tested on intraoperative and postoperative knowledge, decision-making, and management of complications via a series of single-best answer multiple-choice questions. Points are awarded for answering correctly and deducted for answering incorrectly.

5.3.3 Authoring Platform

ENTRUST also features an online authoring portal that is designed to be accessible for clinicians and content experts to create and deploy new case scenarios without requiring programming experience or directly modifying the game (Fig. 5.3). This portal provides user-friendly, easy creation, and customization options for a variety



Question 1 of 6

What is the most likely diagnosis for this patient?

A. Acute appendicitis

B. Inflammatory bowel disease (IBD)

C. Ovarian torsion

D. Cecal Diverticulitis

Fig. 5.2 The question mode within *ENTRUST*. Examinees are tested on intraoperative and postoperative knowledge, decision-making, and management of complications via a series of single-best answer multiple-choice questions

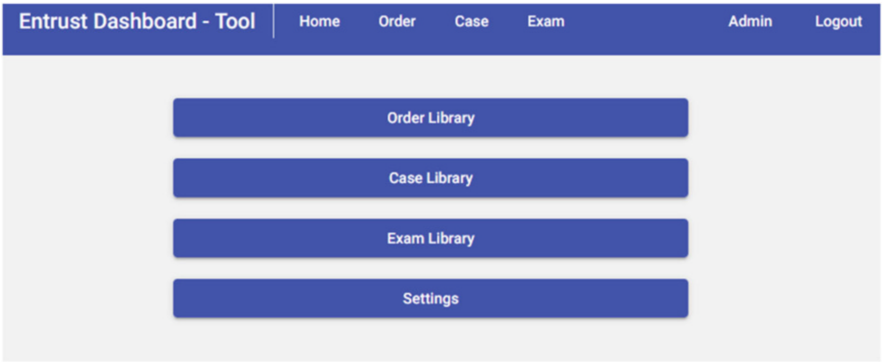


Fig. 5.3 The *ENTRUST* authoring platform. Enables clinicians and content experts to easily create and deploy new case scenarios without requiring programming experience or direct modification of the game

of aspects needed for assessment of clinical decision-making skills. Specifically, the portal provides (1) an order library for creation and management of orders that can be used in case scenarios; (2) a case library that allows for creation and management of all aspects related to a case scenario for assessment; and (3) an exam library that enables the sequencing of case scenarios to create a wide spectrum of exams. These numerous customization options allow for virtually unlimited cases and to be crafted, providing control of aspects ranging from varying patient age, appearance, and apparel via a novel patient character generating tool to specialized labs and orders on the displayed intervention menu.

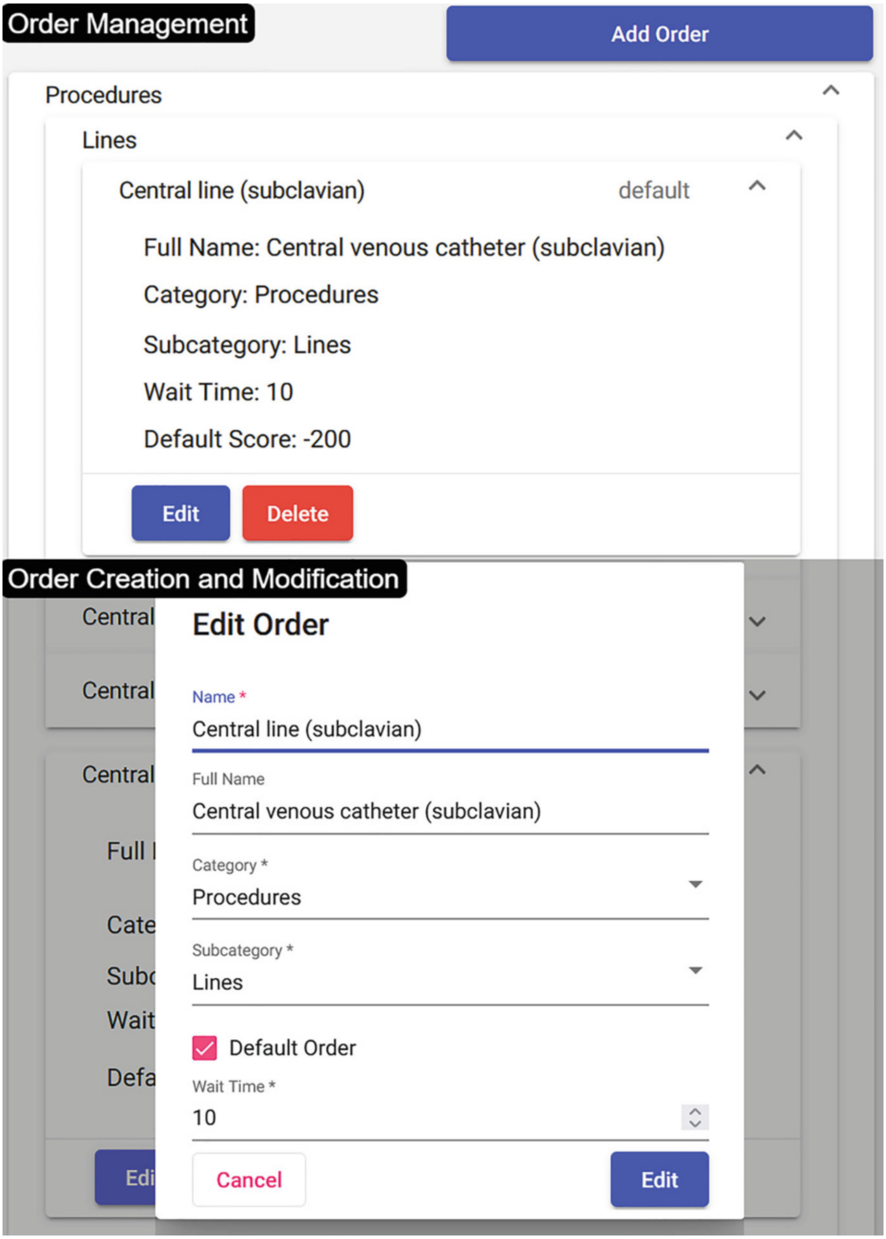
5.3.3.1 Order Library

The order library enables authors to create, manage, and modify a database of orders for use in any case scenario (see Fig. 5.4). The order library is designed to be modular and reusable, enabling authors to specify all default information necessary for a particular order to work within any case while leaving scenario specific details (such as scoring or abnormal results) to be specified in a case-by-case basis within the case library. Specifically, the order library enables authors to easily specify:

- **Order Name**
- **Order Category** (Procedure, Lab, Imaging, Medication, Transfusion, Consult)
- **Order Subcategory**, which is dependent upon what order category was selected
- **Default Order**, i.e., whether it should be included by default when creating any new case scenario in the case library)
- **Wait Time** in seconds for the order to complete during simulation
- **Default Score** when the order is made during simulation
- **Default Result** when the order is made during simulation—there are also additional options to specify if the result should randomly fall within a number range or use a default image if applicable or if there should be multiple default results provided simultaneously
- **Unit** of the default result if applicable

5.3.3.2 Case Library

The case library enables authors to create, manage, and modify a database of case scenarios for use in any examination (see Fig. 5.5). The case library is designed to enable authors to specify all core aspects of a case scenario, including designating effects of interventions on vital signs and determining the appropriateness of actions by rewarding and penalizing examinees on a tiered scoring system. Clinical vignettes and multiple-choice questions can be entered and edited with ease and flexibility as well. Additionally, media files such as photographs and radiology images can be uploaded to be interpreted by the examinee. The specific configuration options the case library provides are:



Central line (subclavian)

Full Name: Central venous catheter (subclavian)

Category: Procedures

Subcategory: Lines

Wait Time: 10

Default Score: -200

Name * Central line (subclavian)

Full Name Central venous catheter (subclavian)

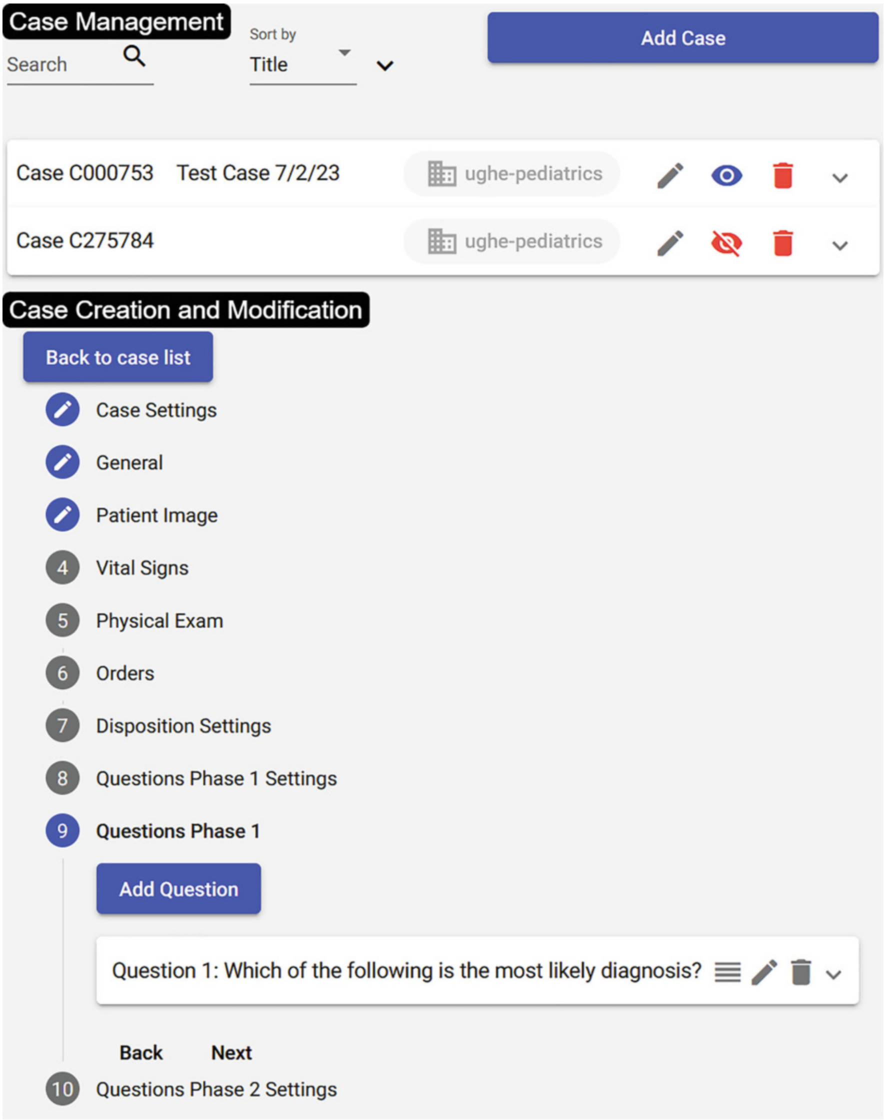
Category * Procedures

Subcategory * Lines

☑ Default Order

Wait Time * 10

Fig. 5.4 The *ENTRUST* authoring platform order library tool. Enables authors (e.g., clinicians and content experts) to easily create and manage modular orders that can be used in any case



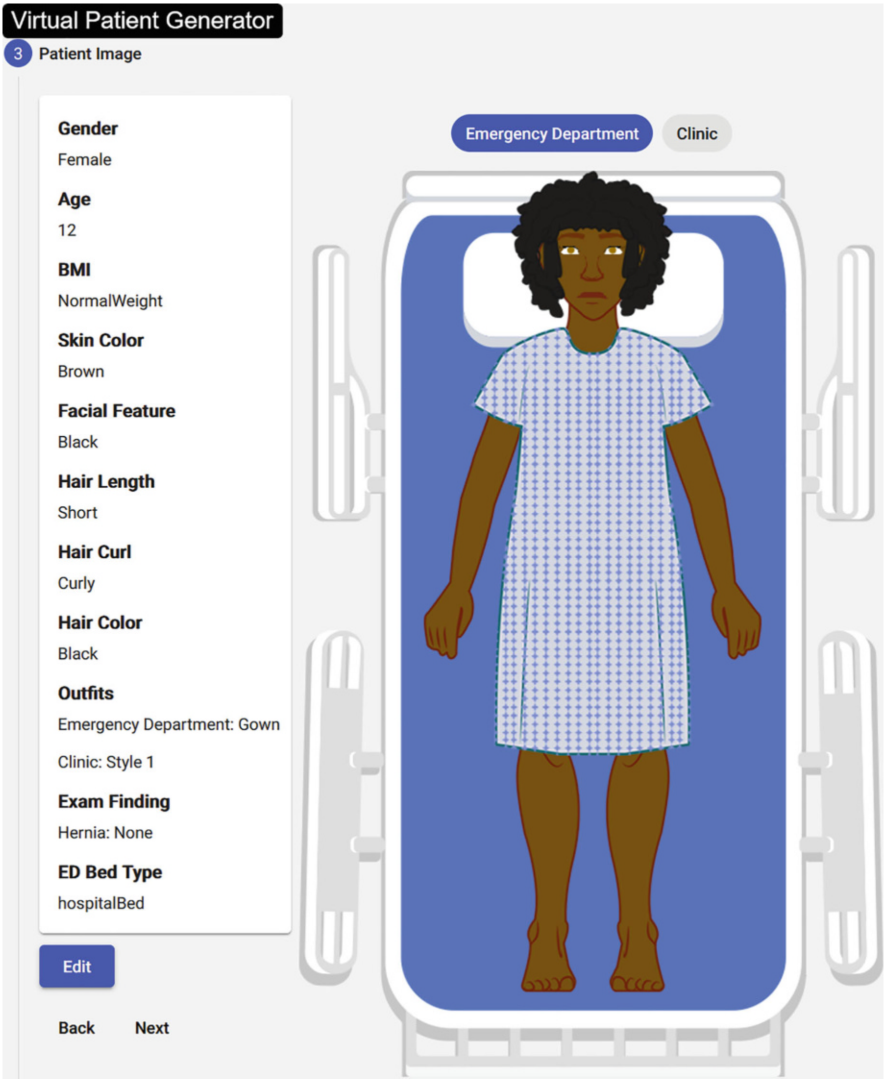
Case C000753 Test Case 7/2/23 ughe-pediatrics

Case C275784 ughe-pediatrics

Question 1: Which of the following is the most likely diagnosis?

Fig. 5.5 The *ENTRUST* authoring platform case library tool. Enables authors (e.g., clinicians and content experts) to easily create and manage case scenarios for examinations

1. **General information**—this section enables authors to specify basic information about the case scenario (such as title, summary, whether it occurs in the emergency room or clinic, and so forth) as well as general information about the virtual patient (such as patient name, their reported ailment, present illness, past medical and surgical history, medications, allergies, and so forth).
2. **Patient image**—this section provides a *virtual patient generation tool* (see Fig. 5.6) that enables authors to customize a wide range of details about the virtual patient such as their sex, age, BMI, skin color, facial features, hair, what they wear when on-screen, visible physical abnormalities during a physical exam (such as a hernia), where the incision site will be displayed during the question mode, and if they will have a C-collar or backboard for certain kinds of injuries. Notably, the broad range of customization options allows for representation of a diverse range of patients from infant to elderly, underweight to morbidly obese, and so forth (see Fig. 5.6 Right for some examples).
3. **Vital sign settings**—this section enables authors to specify the starting vitals for the virtual patient in the simulation mode, as well as specify a *vital sign update algorithm* that specifies how the patient's vitals will change throughout the simulation mode. Vital sign algorithms realistically replicate how certain vitals would change over time in the real world for certain conditions. Current options include clinic patient, stable ED patient, isolated tachycardia, hemorrhagic shock, sepsis, and septic shock.
4. **Physical exam**—this section enables authors to specify the results and score for performing various physical examinations on the virtual patient. Current physical examinations available to the examinee include general, HEENT (head, ears, eyes, nose, and throat), breast, cardiovascular, pulmonary, abdomen, left/right genitourinary, and extremities.
5. **Orders**—this section enables authors to specify what orders (i.e., procedures, labs, imaging, medications, transfusions, consults, or fluids) are available to the examinee in a specific case scenario, as well as the results and positive/negative score effect placing that order will have. By default, when a new order is added to a case scenario, it uses the default details, result(s), and score change specified in the order library. However, authors are also able to modify an order, for that specific case scenario only, to specify sophisticated result and scoring logic (see Fig. 5.7). Specifically, authors can (1) *customize results*, such as change findings or add a different image if applicable to show patient abnormalities for a case scenario; (2) *set new scoring logic* for use of an order, including setting additional penalties for extraneous, repeated use of an order when not appropriate; (3) *set pretest effects* if applicable; and (4) *set vital sign changes* that will occur upon making an order if applicable, e.g., by ordering fluids.



Virtual Patient Generator

3 Patient Image

Gender
Female

Age
12

BMI
NormalWeight

Skin Color
Brown

Facial Feature
Black

Hair Length
Short

Hair Curl
Curly

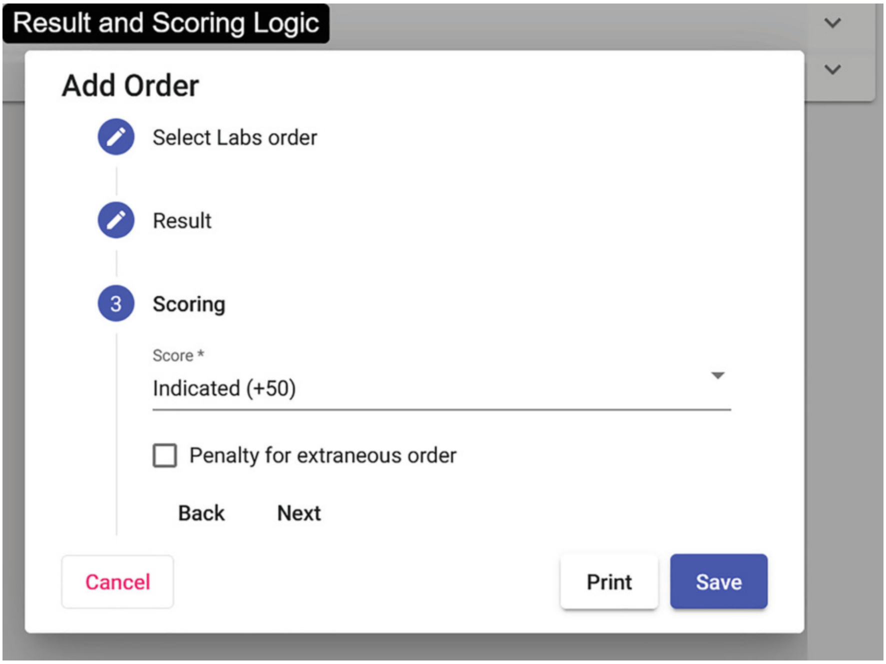
Hair Color
Black

Outfits
Emergency Department: Gown
Clinic: Style 1

Exam Finding
Hernia: None

ED Bed Type
hospitalBed

Fig. 5.6 The *ENTRUST* authoring platform virtual patient generation tool. Enables authors (e.g., clinicians and content experts) to easily define key patient details and visualizes how these will look in real time. Notably, the broad range of options allows for representation of a diverse range of patients from infant to elderly, underweight to morbidly obese, and so forth



Result and Scoring Logic

Add Order

Select Labs order

Result

3 Scoring

Score *
Indicated (+50)

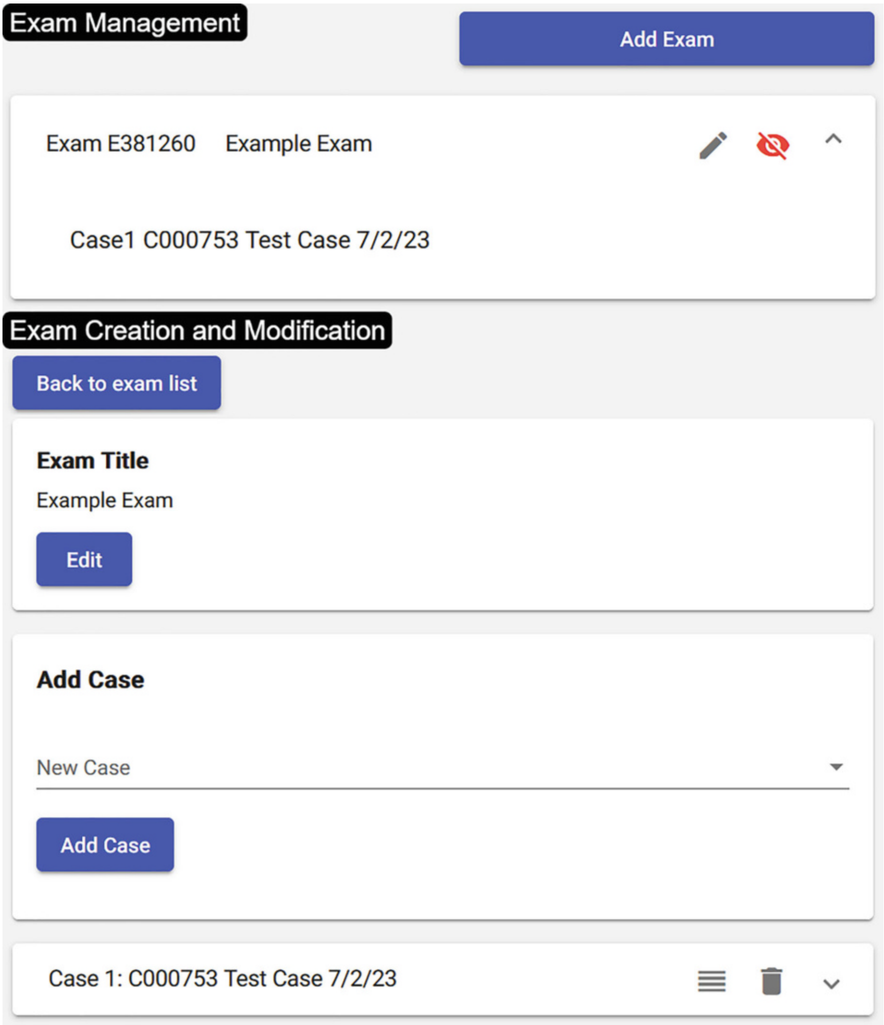
□ Penalty for extraneous order

Fig. 5.7 The customization of results and scoring logic within the *ENTRUST* authoring platform. Enables authors (e.g., clinicians and content experts) to easily define how a specific order will impact results, score, vital sign changes, and so forth for a particular case scenario

6. **Disposition settings**—this section enables authors to specify scoring for each potential disposition choice made by the examinee. Current disposition options include sending the patient home, to a ward, to the ICU, or to the OR or to proceed with nonoperative management.
7. **Intraoperative and postoperative questions**—these sections enable authors to specify single-best answer multiple-choice questions and related settings for questions that will appear in the question mode.

5.3.3.3 Exam Library

The exam library enables authors to create, manage, and modify a database of exams for use in assessment (see Fig. 5.8). Authors are able to create a new exam, select any case scenario from the case library to include in the exam, and modify the order of case scenario appearance. During play, examinees are given a prompt at completion of a case to start the next case (if applicable) upon clicking the “Next” button.



Exam E381260 Example Exam

Case1 C000753 Test Case 7/2/23

Exam Title

Example Exam

New Case

Case 1: C000753 Test Case 7/2/23

Fig. 5.8 The *ENTRUST* authoring platform exam library tool. Enables authors (e.g., clinicians and content experts) to easily create and manage a specified series of case scenarios in the form of an exam for use in assessment

5.3.4 Case Creation and Scoring Algorithm

Over a dozen cases have already been authored and iteratively refined to align with EPA standards for inguinal hernia, thyroid disease, and breast disease as articulated by the American Board of Surgery [7]. Based on feedback from an expert panel, the cases were iteratively revised with the final case scenarios reviewed and approved by the case authors.

A scoring algorithm for *ENTRUST* was also designed to reflect appropriateness of actions, patient clinical status, and accuracy of multiple-choice question responses. This scoring algorithm was vetted by two board-certified surgeons with formal training in surgical education to reflect appropriateness of clinical interventions and multiple-choice question responses. The case scenario and scoring algorithm have also been beta-tested internally by the research team prior to studies and data collection to ensure proper functionality of each case. Specifically, for diagnostic studies and interventions employed during the simulation mode, scoring was 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]
- **Death/Cardiac Arrest** [−500]

Additionally, during simulation mode, points are deducted for each instance of failure to address and correct vital sign abnormalities [−200]. During question mode, multiple-choice questions were awarded +200 points for correct responses and −200 for incorrect responses.

5.3.5 Technical Specifications and Data Collection

ENTRUST utilizes a JavaScript and P5.js front end to provide an interactive simulation interface, as well as a Google Cloud Platform backend for secure data logging and analysis of demographic data, gameplay actions, and scores during gameplay. The platform works on most modern browsers (Chrome, Firefox, and Edge) and is easily distributable to a wide range of participants through a simple Web link. *ENTRUST* requires minimal computational resources to deploy the simulations and can therefore be run on almost any modern computer. The ease of distribution through Web browsers coupled with low computational needs makes *ENTRUST* ideal for deployment in most countries around the world.

ENTRUST's secure backend database logs detailed player performance data including a time stamp of all examinee actions, changes in patient vital signs, points awarded or deducted for an action or intervention, and responses to all multiple-choice questions. The database may be queried to extract data in aggregate format for program-specific or research purposes.

5.4 Study: *ENTRUST* Inguinal Hernia EPA Assessment Pilot

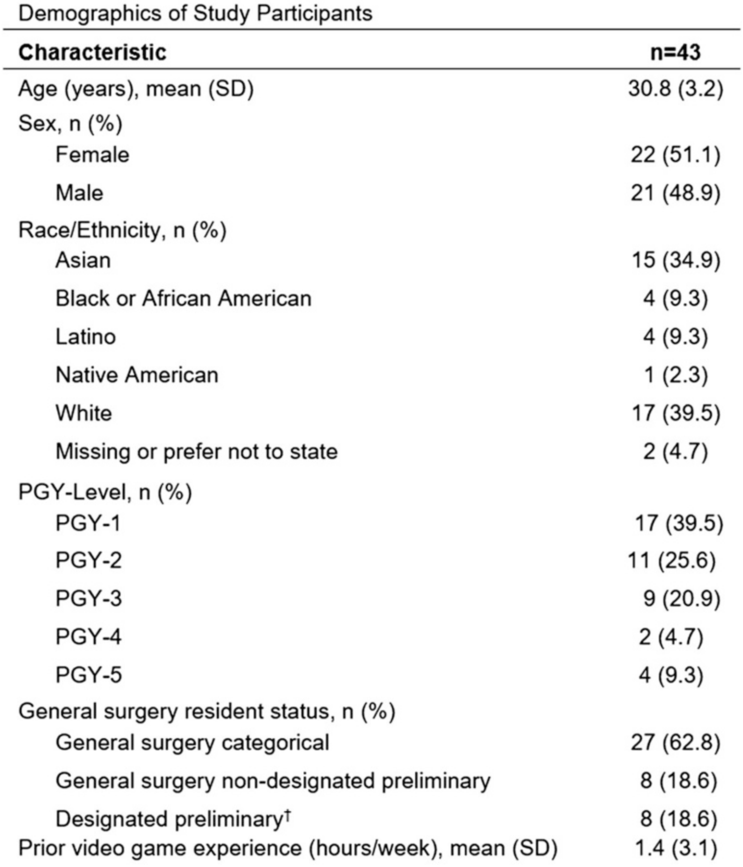
In order to provide initial validity evidence for *ENTRUST*'s capabilities as a tool for assessment of clinical decision-making skills and entrustability, we conducted an initial pilot study of an Inguinal Hernia EPA Assessment developed on *ENTRUST*—this study and results were initially reported in [25]. We hypothesized that *ENTRUST* possesses validity evidence for use in the assessment of clinical decision-making for general surgery residents. As a result, we posited the following research questions:

1. Do users of a game-based assessment tool such as *ENTRUST* need to have prior video game experience to successfully engage with the tool?
2. Does score-based performance on *ENTRUST* discriminate between levels of surgical expertise, e.g., prior operative experience or post-graduate year of training?
3. Is *ENTRUST* able to assess critical surgical decision-making performance?

5.4.1 Methodology

5.4.1.1 Participants

A total of 43 surgical residents at a US-based academic institution participated in the study. 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 post-graduate year 1 (PGY-1) or PGY-2 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 (see Fig. 5.9). Two participants preferred not to report their ethnicity. The self-reported prior video game experience of the participants ranged from 0 to 15 hours per week with mean 1.4 (SD 3.1) hours.



Demographics of Study Participants

Characteristic	n=43
Age (years), 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 non-designated preliminary	8 (18.6)
Designated preliminary†	8 (18.6)
Prior video game experience (hours/week), mean (SD)	1.4 (3.1)

Fig. 5.9 Demographics of study participants for the *ENTRUST* Inguinal Hernia EPA Assessment Pilot. Values reported as n (%) or mean (SD). *Acronymns*—Post-graduate Year (PGY) & standard deviation (SD). † 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

5.4.1.2 Measures

- **Demographic Survey**—a demographic survey was created to collect information pertaining to the age, gender, ethnicity, PGY-level, surgical specialty, self-reported inguinal hernia operative case volume, and prior video game experience of participants.
- ***ENTRUST* Inguinal Hernia EPA Assessment**—an *ENTRUST* Inguinal Hernia EPA Assessment containing four cases was developed and piloted to collect initial validity evidence using Messick's framework [12, 28]. The case scenarios consisted of (1) an outpatient elective unilateral inguinal hernia, (2) an elective bilateral inguinal hernia, (3) an acutely incarcerated inguinal hernia, and (4)

a strangulated inguinal hernia. The four case scenarios for inguinal hernia, including all multiple-choice questions, were authored and iteratively developed by a board-certified general surgeon with formal training in surgical education. Cases were also carefully created in alignment with EPA descriptions and essential functions for inguinal hernia outlined by the American Board of Surgery [7]. 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. The following scores logged by *ENTRUST* were analyzed to compare differences in performance between PGY-levels:

1. *Preoperative sub-score*—the score a participant received on just the simulation mode of *ENTRUST* for a single case scenario
2. *Preoperative total score*—the combined score for all four case scenarios that a participant received on just the simulation mode of *ENTRUST*
3. *Intraoperative sub-score*—the score a participant received on just the intraoperative questions during the question mode of *ENTRUST*
4. *Intraoperative total score*—the combined score for all four case scenarios that a participant received on just the intraoperative questions during the question mode of *ENTRUST*
5. *Postoperative sub-score*—the score a participant received on just the postoperative questions during the question mode of *ENTRUST*
6. *Postoperative total score*—the combined score for all four case scenarios that a participant received on just the postoperative questions during the question mode of *ENTRUST*
7. *Total case score*—the combined score for preoperative sub-score, intraoperative sub-score, and postoperative sub-score for a single case scenario
8. *Grand total score*—the combined total case score for all four case scenarios

5.4.1.3 Procedure

This study was conducted at a US-based academic institution in a proctored exam setting on laptop computers. Participants started by consenting to participate and then completing the demographic survey. After viewing a standardized video tutorial to orient participants to the *ENTRUST* platform, they then 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. The study protocol (#53137) was reviewed and approved by the Institutional Review Board at the authors' institution.

5.4.2 Data Analysis

Demographics are reported as mean and standard deviation for continuous variables and proportions for categorical variables. Descriptive statistics for total and sub-scores, including median and interquartile range, were calculated for each PGY-level. To assess the relationship between *ENTRUST* scores 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 sub-scores 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) [35].

5.4.3 Results

5.4.3.1 Relationship Between Performance and Prior Video Game Experience

Prior video game experience did not correlate with performance on *ENTRUST* (rho = 0.094, $p = 0.56$). This indicates that video game experience is not a prerequisite to successfully engage with *ENTRUST*.

5.4.3.2 Relationship Between Scores and Prior Operative Experience

Grand total score and intraoperative total score were correlated with self-reported prior inguinal hernia operative experience for participants (Fig. 5.10a, rho = 0.65, $p < 0.0001$, and Fig. 5.10b, rho = 0.59, $p < 0.0001$, respectively).

5.4.3.3 Relationships Between Scores and PGY-Level

ENTRUST Inguinal Hernia EPA Assessment grand total score was positively correlated with PGY-level (Fig. 5.11, rho = 0.64, $p < 0.0001$). Preoperative, intraoperative, and postoperative total scores were also positively correlated with PGY-level (preoperative, rho = 0.51, = ; intraoperative, rho = 0.50, $p = 0.0006$; postoperative, rho = 0.32, $p = 0.038$). 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. 5.12a). No statistically significant difference was found in acutely incarcerated inguinal hernia case total score by PGY-level (Fig. 5.12a, rho = 0.10, $p = 0.50$). Descriptive statistics for all *ENTRUST* Inguinal Hernia EPA Assessment scores are shown in Fig. 5.13.

For each of the four 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. 5.12b). 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,

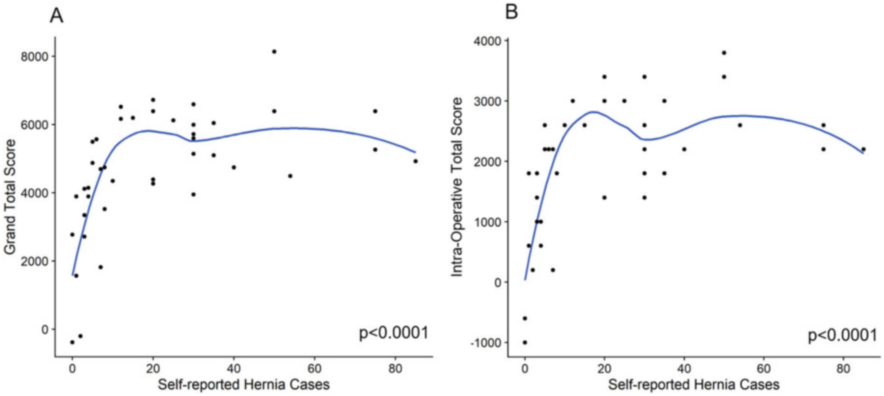


Fig. 5.10 Correlation of *ENTRUST* inguinal hernia EPA score performance to self-reported inguinal hernia operative case experience. (**a**) Grand total score. (**b**) Intraoperative total score

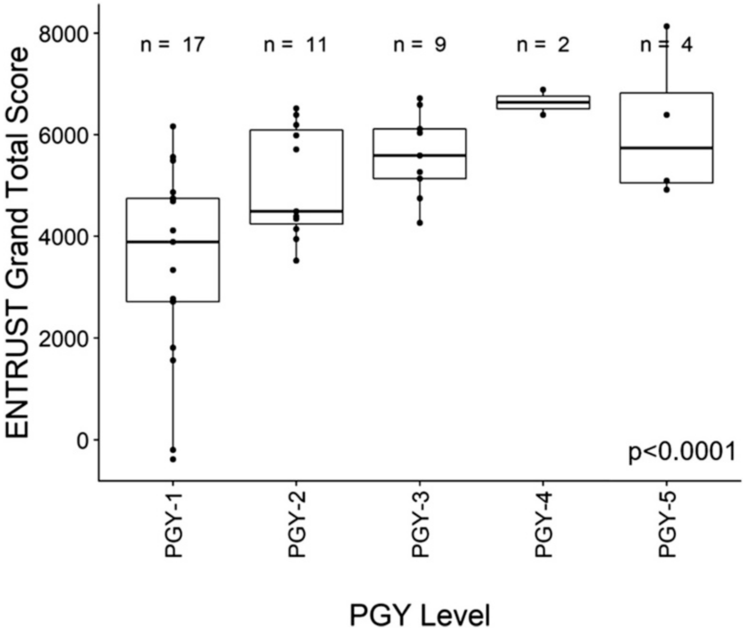


Fig. 5.11 *ENTRUST* Inguinal Hernia EPA Assessment grand total score by PGY-Level

$p = 0.0002$) case scenarios, but was not statistically significant for elective unilateral or acutely incarcerated inguinal hernia cases (Fig. 5.12c).

5.4.3.4 Categorical vs Non-categorical General Surgery Trainee Performance

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

5.4.3.5 Critical Surgical Decision-Making Performance

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. 5.14a). 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

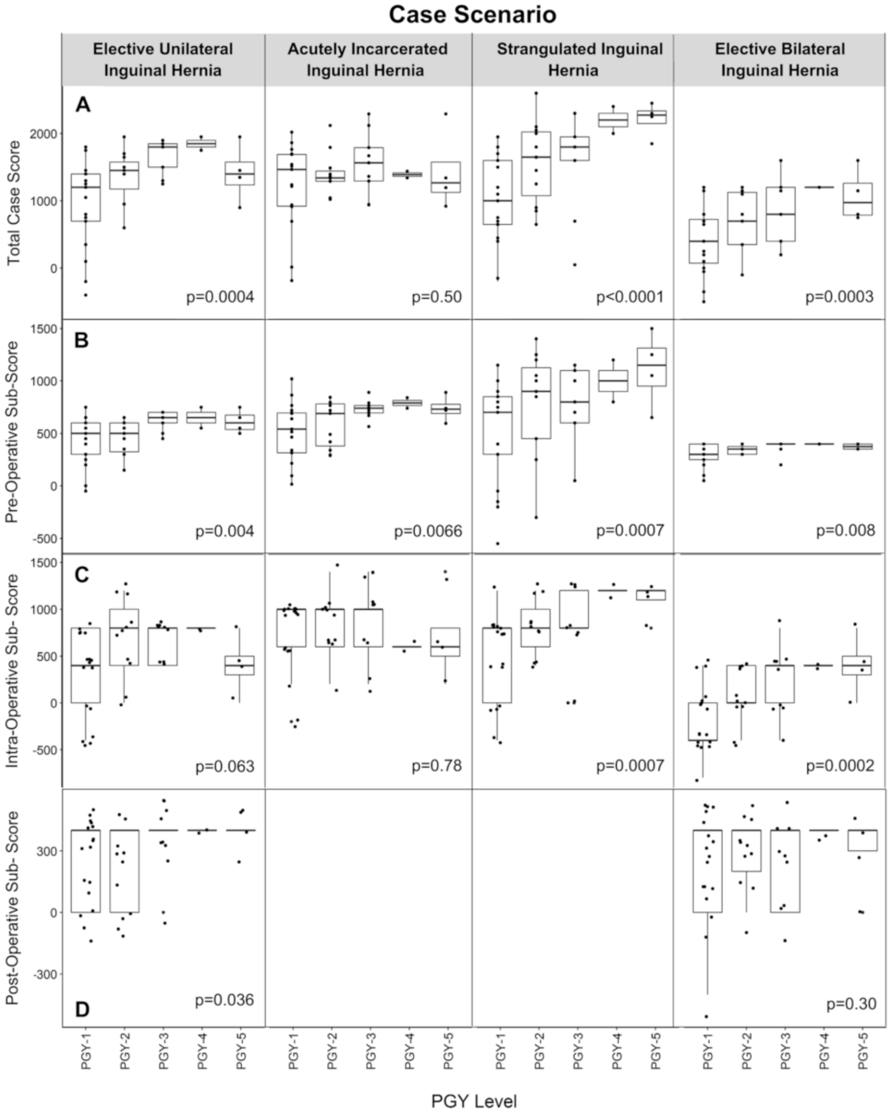
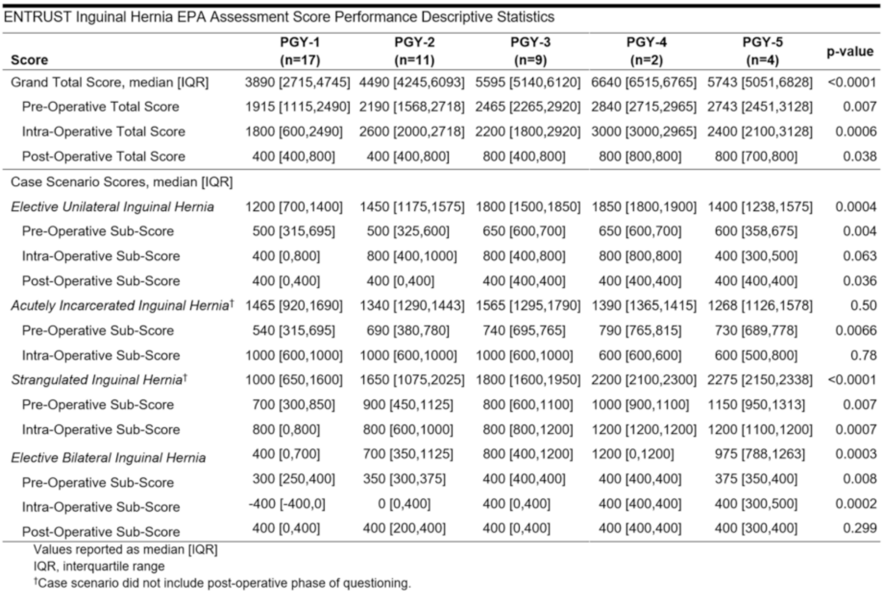


Fig. 5.12 *ENTRUST* Inguinal Hernia EPA Assessment case scenario total and sub-scores by PGY-Level. Total case score (**a**). Preoperative sub-scores (**b**). Intraoperative question sub-scores (**c**). Postoperative question sub-scores (**d**). The acutely incarcerated inguinal hernia and strangulated inguinal hernia case scenarios did not include postoperative questions



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
Pre-Operative Total Score	1915 [1115,2490]	2190 [1568,2718]	2465 [2265,2920]	2840 [2715,2965]	2743 [2451,3128]	0.007
Intra-Operative Total Score	1800 [600,2490]	2600 [2000,2718]	2200 [1800,2920]	3000 [3000,2965]	2400 [2100,3128]	0.0006
Post-Operative 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
Pre-Operative Sub-Score	500 [315,695]	500 [325,600]	650 [600,700]	650 [600,700]	600 [358,675]	0.004
Intra-Operative Sub-Score	400 [0,800]	800 [400,1000]	800 [400,800]	800 [800,800]	400 [300,500]	0.063
Post-Operative Sub-Score	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
Pre-Operative Sub-Score	540 [315,695]	690 [380,780]	740 [695,765]	790 [765,815]	730 [689,778]	0.0066
Intra-Operative Sub-Score	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
Pre-Operative Sub-Score	700 [300,850]	900 [450,1125]	800 [600,1100]	1000 [900,1100]	1150 [950,1313]	0.007
Intra-Operative Sub-Score	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
Pre-Operative Sub-Score	300 [250,400]	350 [300,375]	400 [400,400]	400 [400,400]	375 [350,400]	0.008
Intra-Operative Sub-Score	-400 [-400,0]	0 [0,400]	400 [0,400]	400 [400,400]	400 [300,500]	0.0002
Post-Operative Sub-Score	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 post-operative phase of questioning.

Fig. 5.13 *ENTRUST* Inguinal Hernia EPA Assessment score performance descriptive statistics. Values reported as median [IQR]. Acronym—interquartile range (IQR). † Case scenario did not include postoperative phase of questioning

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. 5.14b). Unadjusted total case score and 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 significant difference in preoperative sub-score remained between those who attempted reduction and those who did not attempt reduction ($p = 0.032$).

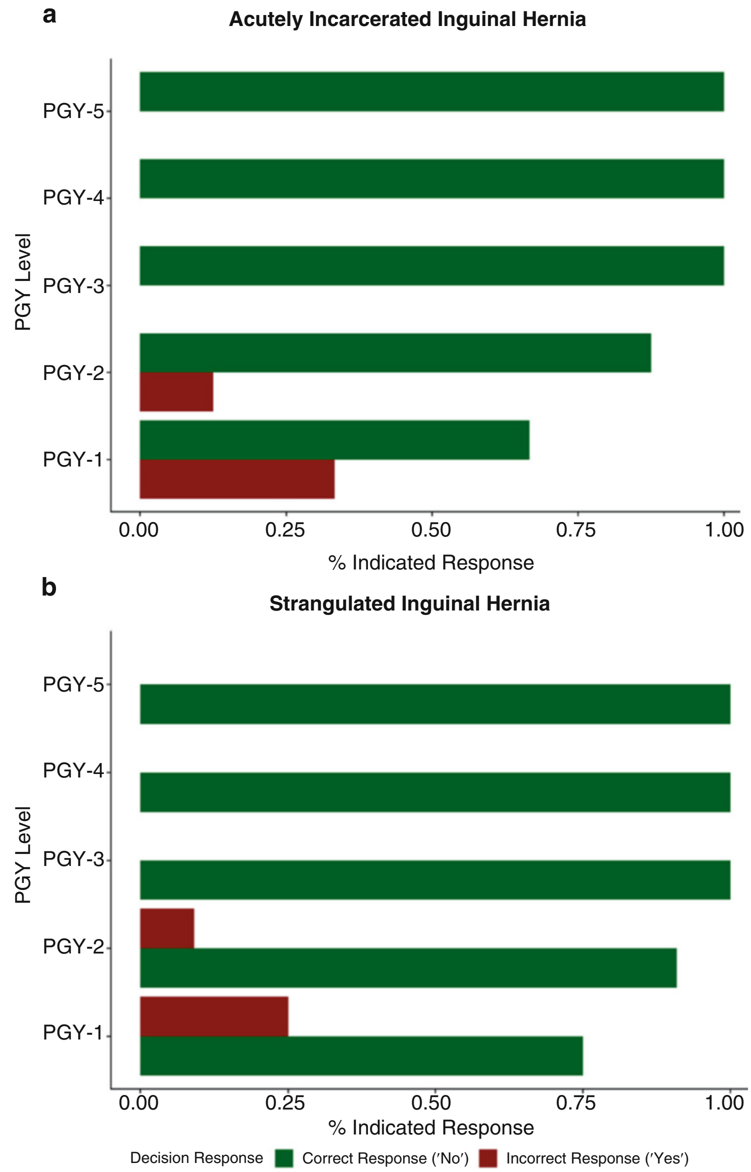


Fig. 5.14 *ENTRUST* Inguinal Hernia EPA Assessment grand total score by PGY-Level

5.5 Discussion

5.5.1 Lessons Learned from the Co-design Process

We identified a number of insights and lessons learned throughout the *ENTRUST* co-design process as follows:

- **Early development of tools to empower stakeholders**—one common technique within game development is to abstract content, design, and logic from core game engine code (e.g., through use of a level editor to create and edit levels or external script files to maintain game parameters, logic, and character dialogues). This is typically done with the intent of modularizing aspects game development as well as making that development more accessible to individuals with limited programming skills. We found this approach to be especially critical for our co-design process since stakeholders tend to have no prior programming experience, making it difficult to add or update content in the game otherwise. However, of equal importance was the creation of sophisticated tools that empowered stakeholders to easily create, edit, and view changes to the serious game in real time. For instance, during the *ENTRUST* design and development process, we initially abstracted the creation and management of case scenarios to a spreadsheet template. While this did enable stakeholders to create content for the game, it also effectively disempowered them since working with a spreadsheet was cumbersome, difficult for reusability (e.g., required reentering default orders and other repeated details for every new case scenario), and forced stakeholders to wait a substantial amount of time to view changes—as a programmer had to input spreadsheet information into the game. This process also introduced a lot of confusion and communication overhead as a by-product. These issues were not remedied until the creation of an authoring tool that enabled stakeholders to quickly and easily edit case scenario information directly in the *ENTRUST* game database. By enabling stakeholders without programming experience to easily create, edit, and view changes to *ENTRUST* in real time, we empowered them to be more directly involved with and provide input into the design and development process. This in turn greatly increased productivity, reduced errors in identifying and addressing latent needs, and ultimately improved overall development speed. Importantly, it also enabled new stakeholders (such as the College of Surgeons of East, Central and Southern Africa) to get involved with various aspects of the project far more easily. This insight also falls in line with existing research, which has highlighted the importance of empowering stakeholders for successful co-design [2].
- **Benefits of frequent review meetings with stakeholders**—another key aspect of *ENTRUST*'s successful co-design and development was the incorporation of weekly review meetings with stakeholders. Initially, *ENTRUST*'s co-design and development involved monthly review meetings with stakeholders. However, the long duration between co-design/development and stakeholder review proved

problematic as it often led to errors in identifying the appropriate items for the sprint and product backlogs. Switching to a more frequent weekly review meeting with stakeholders at the end of each sprint helped to greatly reduce such errors. While frequent review meetings are not always feasible due to time constraints for stakeholders, some form of frequent communication and review (even asynchronously) can result in similar benefits [14, 19, 36, 41].

5.5.2 Validity Evidence for Assessing Clinical Decision-Making Skills

Our pilot data indicates that *ENTRUST* score performance is correlated to PGY-level and inguinal hernia operative experience, i.e., 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.

Additionally, 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.

Ultimately, 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 [7], 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 also no difference in score performance based on prior video game experience, indicating that video game experience is not required to utilize *ENTRUST* effectively.

5.6 Limitations

There are some limitations of this work, particularly with the pilot study. This includes the single institution study design and self-reported inguinal hernia operative experience. Additionally, there were notably lower numbers of participants at higher PGY-levels (see Fig. 5.9). All of these could impact generalizability of the results to some extent.

5.7 Future Work

5.7.1 ENTRUST Development

Future development plans for *ENTRUST* include expansion of the platform to encompass all ABS general surgery EPAs, as well as creation of additional environments, assets, and functionality to accommodate higher acuity case scenarios situated in the trauma bay and ICU settings. Ultimately, this will enable *ENTRUST* to evaluate a broader spectrum of trainees' readiness for entrustment in a more diverse range of scenarios. We also plan to make *ENTRUST* more scalable for distribution by adding additional functionality and security to manage multiple organizations and allow them to maintain their own examinee assessment data and order, case, and exam libraries. Finally, we plan to extend *ENTRUST* beyond just a game-based assessment platform into a game-based learning platform as well. This will include the development of new tools to visualize player actions both individually and in aggregate to support self-regulated learning [3].

5.7.2 ENTRUST Research

Future research 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 [28]. In future studies, we intend to further investigate relationship of *ENTRUST*'s assessment capabilities/scores to other objective assessment variables such as ACGME Case Logs, 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 micro-assessments such as SIMPL or other platforms for actual clinical interactions. 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 gameplay patterns and specific key decision-making actions to EPA levels and readiness for entrustment.

5.8 Conclusion

This chapter presented the design of and preliminary validity evidence for *ENTRUST*—a virtual patient authoring and serious game-based assessment platform to deploy rigorous, case-based patient simulations for evaluation of EPAs. Our results with *ENTRUST* demonstrate feasibility and initial validity evidence for objective assessment of surgical decision-making for inguinal hernia EPA. We also discussed insights and lessons learned from the co-design and development of *ENTRUST*, as well as highlighted future directions for the game-based platform. Importantly, 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, Pichy Jumpholwong, Ananya Anand, and Sherry Wren for their contributions to the development of the ENTRUST platform.

References

1. ABS E-News – Spring 2018. In: News from the American Board of Surgery. The American Board of Surgery (2018). http://www.absurgery.org/quicklink/absnews/absupdate0518.html#epa. Accessed 12 Jan 2021
2. Ardito C., Buono P., Costabile M.F., Lanzilotti R., Piccinno A.: End users as co-designers of their own tools and products. J. Visual Lang. Comput. **23**(2), 78–90 (2012)
3. Barnard-Brak, L., Paton, V.O., Lan, W.Y.: Profiles in self-regulated learning in the online learning environment. Int. Rev. Res. Open Distrib. Learn. **11**(1), 61–80 (2010)
4. Berman, N.B., Durning, S.J., Fischer, M.R., Huwendiek, S., Triola, M.M.: The role for virtual patients in the future of medical education. Acad. Med. **91**(9), 1217–1222 (2016)
5. Bird, M., McGillion, M., Chambers, E.M., Dix, J., Fajardo, C.J., Gilmour, M., Levesque, K., Lim, A., Mierdel, S., Ouellette, C., Polanski, A.N., Reaume, S.V., Whitmore, C., Carter, N.: A generative co-design framework for healthcare innovation: development and application of an end-user engagement framework. Res. Involv. Engagem. **7**(1), 1–12 (2021)
6. Bohnen, J.D., George, B.C., Williams, R.G., Schuller, M.C., DaRosa, D.A., Torbeck, L., Mullen, J.T., Meyerson, S.L., Auyang, E.D., Chipman, J.G., Choi, J.N.: 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. **73**(6), e118–e130 (2016)
7. Brasel, K.J., Klingensmith, M.E., Englander, R., Grambau, M., Buyske, J., Sarosi, G., Minter, R.: Entrustable professional activities in general surgery: development and implementation. J. Surg. Educ. **76**(5), 1174–1186 (2019)
8. Burkett, I.: An Introduction to Co-design, vol. 12. Knode, Sydney (2012)
9. Charlier, N.: Game-based assessment of first aid and resuscitation skills. Resuscitation **82**(4), 442–446 (2011)

10. Chon, S.H., Timmermann, F., Dratsch, T., Schuelper, N., Plum, P., Berlth, F., Datta, R.R., Schramm, C., Haneder, S., Späth, M.R., Dübbers, M., Kleinert, J., Raupach, T., Bruns, C., Kleinert, R.: Serious games in surgical medical education: a virtual emergency department as a tool for teaching clinical reasoning to medical students. JMIR Serious Games **7**(1), 1–11 (2019)
11. Cianciolo, A.T., Kegg, J.A.: Behavioral specification of the entrustment process. J. Grad. Med. Educ. **5**(1), 10–12 (2013)
12. Cook, D.A., Zendejas, B., Hamstra, S.J., Hatala, R., Brydges, R.: What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. Adv. Health Sci. Educ. **19**(2), 233–250 (2014)
13. de Klerk, S., Kato, P.M.: The future value of serious games for assessment: Where do we go now?. J. Appl. Testing Technol. **18**(S1), 32–37 (2017)
14. Domecq, J.P., Prutsky, G., Elraiyah, T., Wang, Z., Nabhan, M., Shippee, N., Brito, J.P., Boehmer, K., Hasan, R., Firwana, B., Erwin, P., Eton, D., Sloan, J., Montori, V., Asi, N., Dabrh, A.M.A., Murad, M.H.: Patient engagement in research: a systematic review. BMC Health Serv. Res. **14**(1), 1–9 (2014)
15. Eaton, M., Scully, R., Schuller, M., Yang, A., Smink, D., Williams, R.G., Bohnen, J.D., George, B.C., Fryer, J.P., Meyerson, S.L.: Value and barriers to use of the SIMPL tool for resident feedback. J. Surg. Educ. **76**(3), 620–627 (2019)
16. Ferreira-Brito, F., Fialho, M., Virgolino, A., Neves, I., Miranda, A.C., Sousa-Santos, N., Caneiras, C., Carrico, L., Verdelho, A., Santos, O.: Game-based interventions for neuropsychological assessment, training and rehabilitation: which game-elements to use? A systematic review. J. Biomed. Inform. **98**, 103287 (2019)
17. George, B.C., Bohnen, J.D., Williams, R.G., Meyerson, S.L., Schuller, M.C., Clark, M.J., Meier, A.H., Torbeck, L., Mandell, S.P., Mullen, J.T., Smink, D.S.: Readiness of US general surgery residents for independent practice. Ann. Surg. **266**(4), 582–594 (2017)
18. George, B.C., Bohnen, J.D., Schuller, M.C., Fryer, J.P.: Using smartphones for trainee performance assessment: a SIMPL case study. Surgery **167**(6), 903–906 (2020)
19. Guise, J.M., O'Haire, C., McPheeters, M., Most, C., LaBrant, L., Lee, K., Cottrell, E.K.B., Graham, E.: A practice-based tool for engaging stakeholders in future research: a synthesis of current practices. J. Clin. Epidemiol. **66**(6), 666–674 (2013)
20. Hwang, G.J., Chang, C.Y. Facilitating decision-making performances in nursing treatments: a contextual digital game-based flipped learning approach. Interactive Learn. Environ. **31**(1), 1–16 (2020)
21. Johnsen, H.M., Fossum, M., Vivekananda-Schmidt, P., Fruhling, A., Slettebø, Å.: Teaching clinical reasoning and decision-making skills to nursing students: design, development, and usability evaluation of a serious game. Int. J. Med. Inform. **94**, 39–48 (2016)
22. Kalyuga, S., Plass, J.L.: Evaluating and managing cognitive load in games. In: Handbook of Research on Effective Electronic Gaming in Education, pp. 719–737. IGI Global, Pennsylvania (2009)
23. Lagro, J., van de Pol, M.H., Laan, A., Huijbregts-Verheyden, F.J., Fluit, L.C., Rikkert, M.G.O.: A randomized controlled trial on teaching geriatric medical decision making and cost consciousness with the serious game GeriatriX. J. Am. Med. Direct. Assoc. **15**(12), e1–957.e6 (2014)
24. Liebert, C.A., Mazer, L., Merrell, S.B., Lin, D.T., Lau, J.N.: Student perceptions of a simulation-based flipped classroom for the surgery clerkship: a mixed-methods study. Surgery **160**(3), 591–598 (2016)
25. Liebert, C.A., Melcer, E.F., Keehl, O., Eddington, H., Trickey, A.W., Lee, M., Tsai, J., Camacho, F., Merrell, S.B., Korndorffer, Jr. J.R., Lin, D.T.: Validity evidence for ENTRUST as an assessment of surgical decision-making for the inguinal hernia entrustable professional activity (EPA). J. Surg. Educ. **79**(6), e202–e212 (2022)
26. Lin, D.T., Park, J., Liebert, C.A., Lau, J.N.: Validity evidence for surgical improvement of clinical knowledge ops: a novel gaming platform to assess surgical decision making. Am. J. Surg. **209**(1), 79–85 (2015)

27. Mavridis, A., Tsiatsos, T.: Game-based assessment: investigating the impact on test anxiety and exam performance. J. Comput. Assist. Learn. **33**(2), 137–150 (2017)
28. Messick, S.: Standards of validity and the validity of standards in performance asessment. Educ. Measur. Issues Pract. **14**(4), 5–8 (1995)
29. Middeke, A., Anders, S., Schuelper, M., Raupach, T., Schuelper, N.: Training of clinical reasoning with a serious game versus small-group problem-based learning: a prospective study. PLoS One **13**(9), e0203851 (2018)
30. Nemirovsky, D.R., Garcia, A.J., Gupta, P., Shoen, E., Walia, N.: Evaluation of surgical improvement of clinical knowledge ops (SICKO), an interactive training platform. J. Digit. Imag. **34**(4), 1067–1071 (2021)
31. New model of surgical resident autonomy coming in 2023. In: ACS Clinical Congress News (2021). Published October 23, 2021. Accessed 21 Jan 2022. https://www.acsccnews.org/new-model-of-surgical-resident-autonomy-coming-in-2023/
32. Nikolian, V.C., Sutzko, D.C., Georgoff, P.E., Matusko, N., Boniakowski, A., Prabhu, K., Church, J.T., Thompson-Burdine, J., Minter, R.M., Sandhu, G.: Improving the feasibility and utility of OpTrust–a tool assessing intraoperative entrustment. Am. J. Surg. **216**(1), 13–18 (2018)
33. Oestreich, J.H., Guy, J.W.: Game-based learning in pharmacy education. Pharmacy **10**(1), 11 (2022)
34. Plass, J.L., Moreno, R., Brünken, R.: Cognitive Load Theory. Cambridge University Press, Cambridge (2010)
35. R Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna (2013). https://www.R-Project.org/
36. Salsberg, J., Parry, D., Pluye, P., Macridis, S., Herbert, C.P., Macaulay, A.C.: Successful strategies to engage research partners for translating evidence into action in community health: a critical review. J. Environ. Pub. Health **2015**, 1–15 (2015)
37. Sánchez, R., Rodríguez, O., Rosciano, J., Vegas, L., Bond ,V., Rojas, A., Sanchez-Ismayel, A.: Robotic surgery training: construct validity of Global Evaluative Assessment of Robotic Skills (GEARS). J. Robot. Surg. **10**(3), 227–231 (2016)
38. Sandhu, G., Nikolian, V.C., Magas, C.P., Stansfield, R.B., Sutzko, D.C., Prabhu, K., Matusko, N., Minter, R.M.: OpTrust: validity of a tool assessing intraoperative entrustment behaviors. Ann. Surg. **267**(4), 670–676 (2018)
39. Schuler, D., Namioka, A.: Participatory Design: Principles and Practices. CRC Press, Boca Raton (1993)
40. Seelow, D.: The art of assessment: using game based assessments to disrupt, innovate, reform and transform testing. J. Appl. Testing Technol. **20**(S1), 1–16 (2019)
41. Slattery, P., Saeri, A.K., Bragge, P.: Research co-design in health: a rapid overview of reviews. Health Res. Policy Syst. **18**(1), 1–13 (2020)
42. Steen, M.: Co-design as a process of joint inquiry and imagination. Des. Issues **29**(2), 16–28 (2013)
43. Ten Cate, O., Scheele, F.: Competency-based postgraduate training: can we bridge the gap between theory and clinical practice? Acad. Med. **82**(6), 542–547 (2007)

44. Ten Cate, O., Chen, H.C., Hoff, R.G., Peters, H., Bok, H., van der Schaaf, M.: Curriculum development for the workplace using entrustable professional activities (EPAs): AMEE guide no. 99. Med. Teach. **37**(11), 983–1002 (2015)
45. Ten Cate, O., Carraccio, C., Damodaran, A., Gofton, W., Hamstra, S.J., Hart, D.E., Richardson, D., Ross, S., Schultz, K., Warm, E.J., Whelan, A.J., Schumacher, D.J.: Entrustment decision making: extending Miller's pyramid. Acad. Med. **96**(2), 199–204 (2021)
46. Vallejo, V., Wyss, P., Rampa, L., Mitache, A.V., Müri, R.M., Mosimann, U.P., Nef, T.: Evaluation of a novel Serious Game based assessment tool for patients with Alzheimer's disease. PLoS One **12**(5), e0175999 (2017)
47. Vassiliou, M.C., Feldman, L.S., Andrew, C.G., Bergman, S., Leffondré, K., Stanbridge, D., Fried, G.M.: A global assessment tool for evaluation of intraoperative laparoscopic skills. Am. J. Surg. **190**(1), 107–113 (2005)
48. Verma, V., Baron, T., Bansal, A., Amresh, A.: Emerging practices in game-based assessment. In: Game-Based Assessment Revisited, pp. 327–346. Springer, Cham (2019)
49. Yeo, H.L., Dolan, P.T., Mao, J., Sosa, J.A.: Association of demographic and program factors with American Board of Surgery qualifying and certifying examinations pass rates. JAMA Surg. **155**(1), 22–30 (2020)