Anesthesiology and Perioperative Science

ORIGINAL RESEARCH Open Access

A serious video game—EmergenCSim™—for novice anesthesia trainees to learn how to perform general anesthesia for emergency cesarean delivery: a randomized controlled trial

Allison J. Lee[1*], Stephanie Goodman[1], Beatriz Corradini[1], Sophie Cohn[2], Madhabi Chatterji[2] and Ruth Landau[1]

Abstract

Purpose We developed EmergenCSim™, a serious game (SG) with an embedded assessment, to teach and assess performing general anesthesia for cesarean delivery. We hypothesized that first-year anesthesiology trainees (CA-1) playing EmergenCSim™ would yield superior knowledge scores versus controls, and EmergenCSim™ and high-fidelity simulation (HFS) assessments would correlate.

Methods This was a single-blinded, longitudinal randomized experiment. Following a lecture (week 3), trainees took a multiple-choice question (MCQ) test (week 4) and were randomized to play EmergenCSim™ ($N=26$) or a non-content specific SG ($N=23$). Participants repeated the MCQ test (week 8). Between month 3 and 12, all repeated the MCQ test, played EmergenCSim™ and participated in HFS of an identical scenario. HFS performance was rated using a behavior checklist.

Results There was no significant change in mean MCQ scores over time between groups F (2, 94) = 0.870, $p=0.42$, and no main effect on MCQ scores, F (1, 47) = 1.110, $p=0.20$. There was significant three-way interaction between time, gender and group, F (2, 90) = 3.042, $p=0.053$, and significant two-way interaction between gender and time on MCQ scores, F (2, 94) = 107.77, $p=0.036$; outcomes improved over time among males. There was no group difference in HFS checklist and SG scores. Both instruments demonstrated good internal consistency reliability but non-significant score correlation.

Conclusions Playing EmergenCSim™ once did not improve MCQ scores; nonetheless scores slightly improved among males over time, suggesting gender may impact learning outcomes with SGs.

Keywords Serious games, Multiple-choice test, General anesthesia, Emergency cesarean delivery

1 Introduction

Anesthesia-related maternal mortality has decreased with reduced use of general anesthesia for cesarean delivery [1], however, general anesthesia -related complications in pregnancy remain high [2, 3]. This has raised concern about inadequate readiness of anesthesia trainees to manage such cases, which are usually emergencies

*Correspondence:
Allison J. Lee
al3196@cumc.columbia.edu
[1] Department of Anesthesiology, Columbia University Irving Medical Center, 622 West 168th St, PH5, New York, NY 10032, USA
[2] Columbia University Teachers College, New York, NY, USA

Springer

with increased likelihood of serious maternal comorbidities [4, 5].

Experts advocate high-fidelity simulation (HFS) as the surrogate modality for learning and maintaining knowledge and skills [5]. Although effective for teaching, HFS is time and resource intensive and its use for teaching about this clinical scenario is not mandated by the Accrediting Council for Graduate Medical Education [6, 7]. With a minimum requirement of two one-month rotations in obstetric anesthesia during a 3-year residency [8] and given the cesarean delivery general anesthesia rates of 5.8% [9] and as low as 0.5–1% [1] in some institutions, it is quite possible for residents to graduate without ever having encountered this clinical scenario, representing a significant experience gap.

Serious three-dimensional (3-D) video games (serious games) are immersive virtual platforms with a pedagogical purpose, increasingly being incorporated into healthcare education [10–12]. They produce realistic scenarios where players may conduct decision-making and actions from a "first-person shooter" perspective. Video games are enjoyed by all genders [13], and evidence of their educational effectiveness is growing, but that evidence-base is still relatively limited, comprising heterogenous and mostly low quality studies [14]. Serious games are recognized as promising educational tools due to cost-effectiveness [15, 16], flexibility, portability and scalability [14]; however, development costs are prohibitive, which has limited their use primarily to high-income settings [14].

With developers, MEDUSIMS (https://medusims.com/en/), we created EmergenCSim™, a novel 3-D English-language serious game with an embedded scoring and debriefing tool, to teach novice clinical anesthesia-year one (CA-1)'s to perform general anesthesia for emergency cesarean delivery. The purpose of this study was to evaluate the effectiveness of EmergenCSim™ as an educational and assessment tool. We hypothesized that, compared to a control condition (playing a non-content specific serious game on anaphylaxis), playing EmergenCSim™ would improve novice anesthesia trainees' knowledge regarding general anesthesia for emergency cesarean delivery, as measured by gains in knowledge scores from pre-exposure baseline to two post-exposure time points over a training year. Additionally, we examined the convergent validity and reliability of scores from the embedded EmergenCSim™ assessment against the HFS behavior checklist, both employed to evaluate residents' proficiency related to conducting general anesthesia for emergency cesarean delivery.

Finally, prompted by evidence regarding differences in male versus female perspectives of the educational value of serious games and level of interest in serious games that mimic clinical practice [17, 18], we examined a post-hoc exploratory hypothesis that there would be no variability by gender on measured knowledge outcomes over time.

2 Methods

This trial was approved by Columbia University's Institutional Review Board.

2.1 Participants

Trainees in 2 consecutive CA-1 classes provided written informed consent. Non-inclusion criteria included prior post-graduate anesthesiology training.

2.2 Educational interventions and outcome measures

The educational interventions (lecture, serious games, and serious game electronic debriefing), and outcome measures (29-item multiple-choice question (MCQ) knowledge test—correct answers receive 1 point [19], EmergenCSim™ scoring tool, and HFS behavior checklist) were developed between February 2016 and June 2017.

The lecture content and all outcome measures covered the following subdomains: (1) Physiologic Changes of Pregnancy, (2) Pharmacology, (3) Anesthetic Implications of Pregnancy, and (4) Crisis Resource Management Principles. The expected actions for EmergenCSim™ and HFS scenarios were adapted from a validated 52-item behavior checklist [20]. Additional file 1: Appendix A contains a detailed description of the development of the educational interventions and measurement instruments, including a table with the weighted scoring tools for the game and HFS scenario.

2.3 Study flow

2.3.1 Phase 1: longitudinal experiment

To examine the primary hypothesis predicting knowledge gains over time, the study was a single-blinded, two-group, longitudinal randomized experiment (Fig. 1).

- **Week 3 of CA-1 Year:**

All participants attended a 60-min lecture, delivered by A.J.L., on general anesthesia for emergency cesarean delivery.

- **Week 4 of CA-1 Year (Time 1):**

All participants took the MCQ baseline knowledge test (Additional file 2: Appendix B), completed a survey regarding prior videogame experiences, and played the serious games while being observed by investigators A.J.L. and B.C.

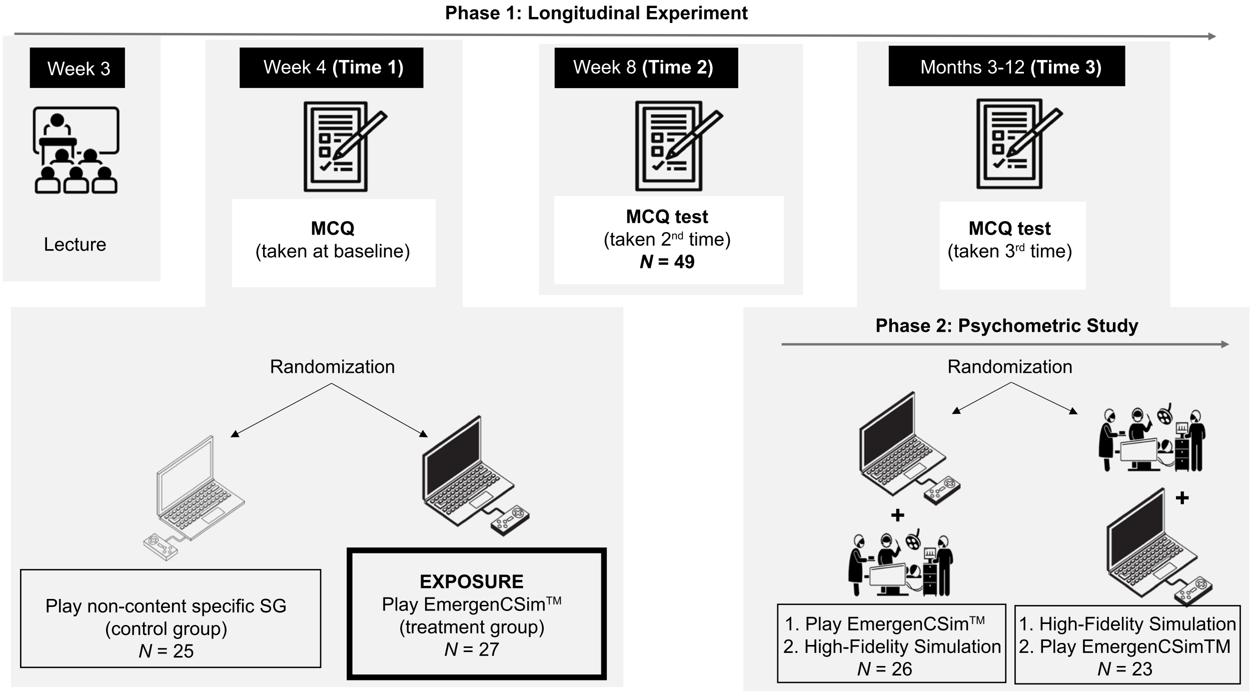


Fig. 1 Study flow diagram. Trainees received a lecture in Week 3 of clinical anesthesia-year one (CA-1), took a multiple choice question (MCQ) pre-test at baseline in Week 4 (Time 1) and were randomized to play either EmergenCSim™ or a Non-content Specific Game. In Week 8 (Time 2) they took the MCQ test a 2nd time, and between months 3–12 (Time 3) they took the same test a 3rd time and then played EmergenCSim™ and participated in a high-fidelity simulation of general anesthesia for emergency cesarean delivery; the order of playing EmergenCSim™ and simulation was randomized. SG = Serious Game

Participants were randomized using Bernoulli randomization in R (RStudio, version 3.4.0) to either:

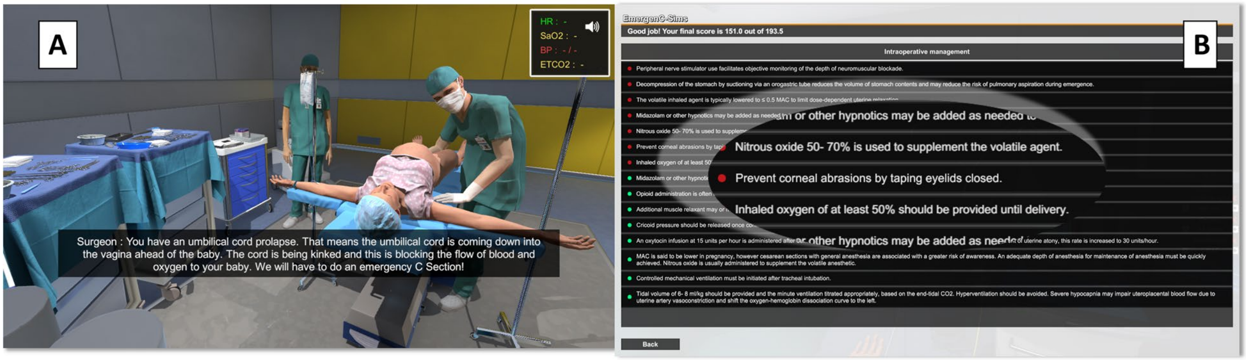
1) **EmergenCSim™ Group** (treatment, $N=27$, final $N=26$): played EmergenCSim™
2) **Non-content Specific Game Group** (control, initial $N=25$, final $N=23$): played anaphylaxis game

A video tutorial explaining how to play the games was provided beforehand. The tutorial may be viewed via this link: https://youtu.be/LP6WwQHPQ4U.

Playing EmergenCSim™ The written prebrief for EmergenCSim™ explains that the player is a CA-1 resident during their initial obstetric anesthesia rotation and that they have been paged to go emergently to the obstetric operating room. Upon entering the simulated operating room, their avatar ("first-person-shooter" perspective) is greeted by the obstetric surgeon positioned next to the patient on the operating table, with a vaginally-placed hand elevating the fetal head to relieve pressure on the umbilical cord. The obstetrician hurriedly reports that there is fetal bradycardia secondary to vaginal prolapse of the umbilical cord and the parturient needs to be immediately anesthetized for emergency cesarean delivery. The player can interact with the patient avatar, perform a focused history and physical examination, and may call on an attending anesthesiologist avatar for help in proceeding with the steps to perform general anesthesia. At the end of the game, the player's score and electronic debrief are presented (Fig. 2).

Playing the non-content specific game on anaphylaxis The player is required to evaluate a non-pregnant woman who has developed respiratory distress shortly after beginning to receive surgical antibiotic prophylaxis in the operating room, before the start of surgery. The player is expected to diagnose anaphylaxis, discontinue antibiotic administration, provide supplemental oxygen, and administer intravenous epinephrine and fluid boluses. At the end of the game, a brief electronic debrief is presented.

Week 8 of CA-1 Year (Time 2) Participants were emailed a link to the same MCQ test. Instructions were to complete the questions within 30 min, without researching answers.



Surgeon : You have an umbilical cord prolapse. That means the umbilical cord is coming down into the vagina ahead of the baby. The cord is being kinked and this is blocking the flow of blood and oxygen to your baby. We will have to do an emergency C Section!

EmergenC-Sims

Good job! Your final score is 151.0 out of 193.5

Intraoperative management

- Peripheral nerve stimulator use facilitates objective monitoring of the depth of neuromuscular blockade.
- Decompression of the stomach by suctioning via an orogastric tube reduces the volume of stomach contents and may reduce the risk of pulmonary aspiration during emergence.
- The volatile inhaled agent is typically lowered to ≤ 0.5 MAC to limit dose-dependent uterine relaxation.
- Midazolam or other hypnotics may be added as needed
- Nitrous oxide 50- 70% is used to supplement the volatile agent.
- Prevent corneal abrasions by taping eyelids closed.
- Inhaled oxygen of at least 50% should be provided until delivery.
- Midazolam or other hypnotics may be added as needed
- Opioid administration is often
- Additional muscle relaxant may or
- Cricoid pressure should be released once co
- An oxytocin infusion at 15 units per hour is administered after of uterine atony, this rate is increased to 30 units/hour.
- MAC is said to be lower in pregnancy, however cesarean sections with general anesthesia are associated with a greater risk of awareness. An adequate depth of anesthesia for maintenance of anesthesia must be quickly achieved. Nitrous oxide is usually administered to supplement the volatile anesthetic.
- Controlled mechanical ventilation must be initiated after tracheal intubation.
- Tidal volume of 6- 8 ml/kg should be provided and the minute ventilation titrated appropriately, based on the end-tidal CO2. Hyperventilation should be avoided. Severe hypocapnia may impair uteroplacental blood flow due to uterine artery vasoconstriction and shift the oxygen-hemoglobin dissociation curve to the left.

Fig. 2 **A** Screen shot of EmergenCSim™ opening scene: the obstetrician avatar, with a vaginally-placed hand explains to the parturient avatar why emergency cesarean delivery is necessary. **B** Screen shot of electronic debrief page displaying the player's final score. Red bulleted items are missed actions, and green bulleted items are correctly performed actions. One item (inset) is magnified for demonstration

- **Months 3–12 of CA-1 Year (Time 3):**

Participants (i) took the MCQ test for the 3rd time, (ii) played EmergenCSim™ (1st time for controls, 2nd time for the treatment group) and (iii) participated in HFS of a scenario identical to the game. The order of playing EmergenCSim™ and HFS was randomized to neutralize order effects; in other words, subjects at this time either played EmergenCSim™ first and then participated in simulation, or participated in simulation and then played EmergenCSim™, the order of which was randomly selected.

High-fidelity simulation The content for the mannequin-based HFS scenario was designed to be as identical as possible to the EmergenCSim™ digital simulation, with the same actions (behaviors) scored by raters of the videotaped scenario as within the game. Following a prebrief, during which A.J.L. explains that she will play the attending anesthesiologist in the scenario, the simulation starts with the learner being notified verbally by A.J.L. that they have been paged to go emergently to the obstetric operating room. On arrival, an obstetric surgeon (embedded confederate) explains that the patient has umbilical cord prolapse and requires emergent general anesthesia to facilitate cesarean delivery. Similar to the EmergenCSim™ scenario, the player is expected to call for help promptly and proceed with a focused history and physical examination and proceed with general anesthesia. Upon conclusion of the scenario, a short debrief is conducted by A.J.L.

All activities (except for week 8), were conducted in the departmental simulation center.

2.3.2 *Phase 2: pyschometric study of EmergenCSim™ assessment*

A secondary hypothesis was that EmergenCSim™ assessment scores would show internal consistency reliability, and adequate levels of convergent validity with the HFS checklist (assessed at Time 3 (months 3–12)). Both scores were expected to correlate positively and significantly.

HFS video recordings were reviewed by A.J.L. and S.R.G. (only S.R.G. was blinded to group allocation) and participants' performance was rated using the behavior checklist.

2.4 Realism survey

At Time 3 participants rated the realism of EmergenCSim™ (scale 1–10; where 1 = not realistic at all and 10 = extremely realistic) and HFS (scale 1–5; where 1 = not realistic at all and 5 = extremely realistic) via different surveys; this resulted in realism scores with different scales.

2.5 Psychometric quality of knowledge test

Based on prior item analysis [19] of the original 29-item MCQ test, four items (items #23, #26, #28 and #29) that failed to discriminate between novices and experts were removed. Those discrimination indices were 0.02, 0.00, 0.02 and −0.06, respectively. A total of 25 questions were ultimately analyzed. Internal consistency reliability estimates for the resulting 25-item test scores at Times 1, 2 and 3, yielded sub-par Cronbach's α values of 0.489, 0.674 and 0.574 respectively (cutoff for acceptability is 0.70), probably due to heterogeneous item composition.

2.6 Statistical analyses

Outcome measures included MCQ test scores (Times 1, 2 and 3), EmergenCSim™ scores (Time 1 (treatment group only) and Time 3 (both groups)), HFS behavior checklist scores (averaged over 2 independent raters, Time 3) and survey responses (Times 1 and 3).

2.7 Power analysis

We estimated that with a fixed count of 26 subjects per group, for an outcome measure yielding a maximum score of 29 with SD of 5, we would achieve 80% power to detect a 4-point difference between groups with a significance level (alpha) of 0.05 using a two-sided two-sample t test.

2.7.1 Phase 1 analysis

Statistical analysis for the longitudinal experiment was guided by our primary hypothesis that the EmergenCSim™ group would achieve a greater increase in mean MCQ scores from Time 1 to 2, and from Time 2 to 3, than the Non-content Specific Game group. analysis involved a series of repeated measures analyses of variance (ANOVAs). The within-subjects effects examined change on knowledge outcomes by group and time point. The main effects examined impact of treatment versus control condition on outcomes, and subsequently by gender, as the between-subjects factors. Two-way (treatment by time) and three-way (treatment by gender by time) interaction effects were tested. The three-way analysis was prompted by a post hoc hypothesis on gender differences and interest in exploring the potential impact of gender on learning outcomes with serious games. The Type 1 error level was set at $p < 0.05$ for statistical significance.

For post-hoc multiple comparisons analyses, the *p*-value was adjusted using the Bonferroni adjustment method (i.e., 0.05 was divided by the number of comparisons made, which was 3).

2.7.2 Testing assumptions and outliers

Prior to the ANOVAS, data was checked to ensure test assumptions had been met. There were two outliers as assessed by studentized residuals (one in the EmergenCSim™ group at Time 2 and one in the Non-content Specific Game group at Time 3, both with studentized residual values of -3.27); their data were preserved for analyses, being considered legitimate observations.

The MCQ test scores were normally distributed, as assessed by Shapiro-Wilk's test of normality for each group at each time point ($p > 0.05$). There was homogeneity of variances ($p = 0.296$ at Time 1, 0.186 at Time 2, and 0.143 at Time 3) and covariances ($p = 0.495$), as assessed by Levene's test of homogeneity of variances and Box's M test, respectively. Mauchly's test of sphericity indicated that the assumption of sphericity was met for the two-way interaction, $\chi2(2) = 0.70$, $p = 0.705$.

2.7.3 Group gender balance

The balance of male and female participants per group (based on self-identified gender within the residency program) was compared using a chi-squared test of independence. Randomized blocking by gender was not attempted due to small cohort sizes. No participants self-identified as non-binary.

2.7.4 Phase 2 data analysis

Internal consistency reliability estimates for the serious game-embedded scoring tool and HFS behavior checklist were obtained using Cronbach's α. Convergent validity evidence and inter-rater reliability were obtained using Pearson's correlations. A Mann-Whitney U test was run to analyze survey responses regarding video game experience and realism ratings of EmergenCSim™ and HFS. Analyses were performed using SPSS (IBM Corp. Released 2019. IBM SPSS Statistics for Macintosh, Version 26.0. Armonk, NY: IBM Corp.)

3 Results

Fifty-two CA-1's (30 male and 22 female) were enrolled. In the treatment group, one declined participation at Time 3. In the control group, 2 did not complete the study—one left the program after Time 1, and another declined participation at Times 1 and 2. Forty-nine trainees completed the MCQ tests at all 3 time points and were included in the analyses—26 in the treatment group and 23 in the control group. There was no difference in the proportion of males and females within groups, $p = 0.39$.

3.1 Overall time effect

Mean MCQ scores (SDs) are in Table 1. Regardless of group, the main effect of time showed a significant difference in mean scores at the different time points, F(2,94) = 7.834, $p = 0.001$, suggesting overall improvement in both groups. Post-hoc analyses with a Bonferroni adjustment revealed that scores increased significantly by a mean of 1.5 points from Time 1 to 3 (95% CI 0.51, 2.47; $p = 0.001$), but not from Time 1 to 2 (mean 0.635, 95% CI −0.32, 1.59; $p = 0.318$), nor from Time 2 to 3 (mean 0.865, 95% CI −0.02, 1.75; $p = 0.058$).

3.2 Treatment effect between groups

The change in mean scores over time were not different between groups, F (2, 49) = 0.870, $p = 0.42$ (Fig. 3). There was no significant main effect of the treatment on

Table 1 Mean knowledge test multiple choice question (MCQ) scores by group, gender and time point

Subjects	N	Time 1 (pre-test)	Time 2 (post-test #1)	Time 3 (post-test #2)
Total	49	18.2 (SD = 2.6)	18.8 (SD = 3.3)	19.7 (SD = 2.7)
Total Female	20	19.1 (SD = 1.9)	18.9 (SD = 4.0)	19.6 (SD = 3.0)
Total Male	29	17.5 (SD = 2.8)	18.8 (SD = 2.8)	19.8 (SD = 2.4)
EmergenCSim™ group	26	17.6 (SD = 2.7)	18.5 (SD = 3.7)	19.6 (SD = 3.1)
Female	9	18.3 (SD = 2.4)	17.2 (SD = 4.9)	18.2 (SD = 3.9)
Male	17	17.2 (SD = 2.8)	19.2 (SD = 2.7)	20.3 (SD = 2.4)
Non-content specific game group	23	18.8 (SD = 2.4)	19.2 (SD = 2.8)	19.8 (SD = 2.1)
Female	11	19.7 (SD = 1.7)	20.2 (SD = 2.4)	20.6 (SD = 1.4)
Male	12	18.0 (SD = 2.9)	18.3 (SD = 2.8)	19.1 (SD = 2.4)

Knowledge test was taken at 3 time points (MCQ maximum score = 29 points). Values represent mean (SD)

There was no statistically significant interaction between study group and time on knowledge test scores, F (2, 49) = 0.870, p = 0.422, partial $\eta^2 = 0.018$. There was a statistically significant three-way interaction between time, gender and group, F(2,45) = 3.042, p = 0.053, partial $\eta^2 = 0.063$

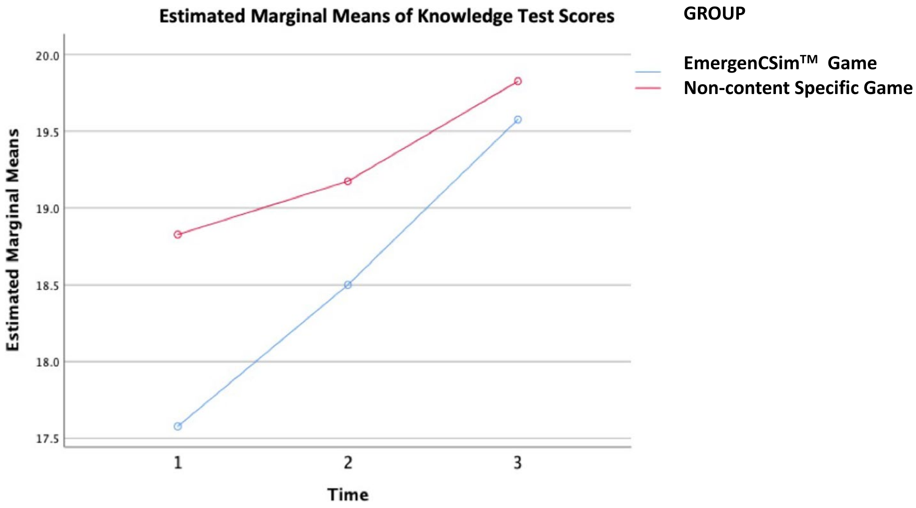


Fig. 3 Estimated marginal means of knowledge test scores over time by experimental group. Times 1, 2 and 3 refer to week 3, week 4 and months 3–12 of CA-1 year respectively. Note that the y-axis does not start at 0

MCQ score (with the treatment group playing EmergenCSim™ only once), F (1, 47) = 1.110, $p = 0.20$.

3.3 Effect of gender over time- treatment and control groups

The three-way interaction test to explore whether subgroups by gender varied significantly on mean MCQ scores in treatment versus control groups over time was marginally statistically significant, (F (2, 45) = 3.042, $p = 0.053$; Table 1). However, there was no statistically significant simple two-way interaction of gender and group at Times 1, 2 or 3 (1: F (1, 45) = 0.152, $p = 0.669$; 2: F (1, 45) = 4.293, $p = 0.044$; 3: F (1, 45) = 5.793, $p = 0.020$) using the Bonferroni corrected value.

Examining gender and time separately, two-way mixed ANOVA revealed a statistically significant interaction between gender and time on knowledge test scores, F (2, 49) = 107.77, $p = 0.036$ (Fig. 4). At Time 1, there was a statistically significant difference in test score between genders, F (1, 47) = 4.724, $p = 0.035$, with females having a mean test score 1.58 points higher than males. There was no statistically significant difference between the genders at Time 2, (F (1, 47) = 0.004, $p = 0.95$) or Time 3 (F (1, 47) = 0.098, $p = 0.76$). At Time 3, the mean (SD) scores by gender were nearly identical- for females 19.55 (3.02) and for males 19.79 (2.41).

There was a statistically significant effect of time on MCQ scores for males, F (2, 56) = 13.794, $p < 0.001$, but not for females F (2, 38) = 0.589, $p = 0.56$. Among males,

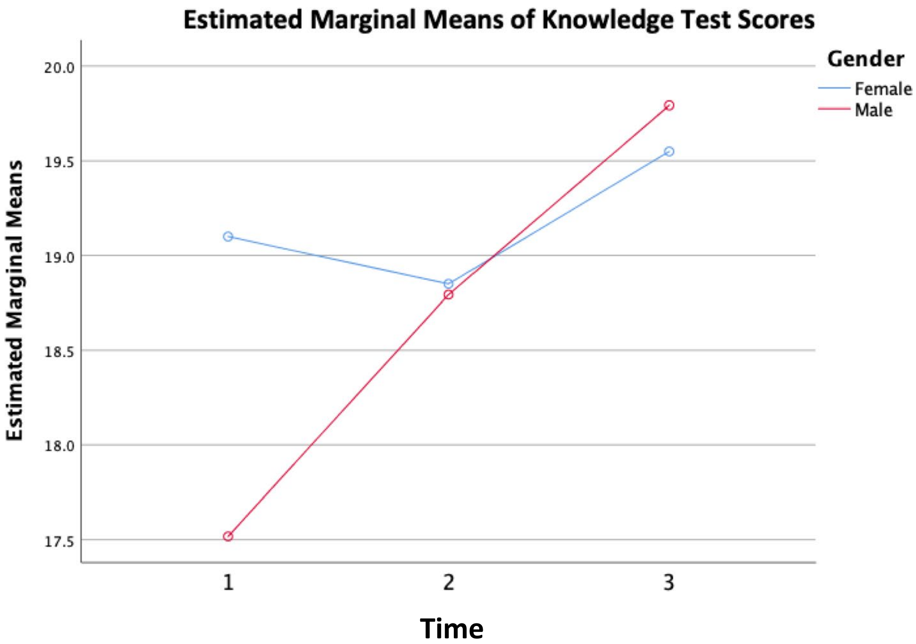


Fig. 4 Estimated marginal means of knowledge test scores over time for two-way interaction of time and gender. Times 1, 2 and 3 refer to week 3, week 4 and months 3–12 of CA-1 year respectively. Note that the y-axis does not start at 0

the mean difference in MCQ score increased from Time 1 to 2 (Mean difference = 1.28, SE = 0.45, $p = 0.025$) and from Time 2 to 3 (Mean Difference = 1.0, SE = 0.33, $p = 0.014$), and overall, from Time 1 to 3 (Mean Difference = 2.28, SE = 0.51, $p < 0.0001$). This pattern was not found among females.

3.4 Effect of treatment on serious game scores and HFS checklist

At Time 3, there was no significant difference in mean serious game scores between groups—120 (SD 16.81) and 109 (SD 28.86) in the EmergenCSim™ and control groups respectively; (t (48) = 1.731, $p = 0.09$) or by gender ($p = 0.33$). There was also no significant difference in mean HFS behavior checklist scores (average of 2 raters) between EmergenCSim™ Group (Mean 229.26, SD 41.73) vs. control group (Mean 222.08, SD 37.78), $p = 0.53$, or by gender—2017 group: F (1, 36) = 0.066, $p = 0.80$; 2018 group: F (1, 37) = 0.979, $p = 0.38$.

3.5 Phase 2

EmergenCSim™ scores demonstrated good internal consistency reliability estimate at Time 3 with a Cronbach's α of 0.806. Similarly, strong internal consistency reliability was demonstrated for HFS checklist scores – Cronbach's α 0.823 for the full sample.

Strong inter-rater reliability on HFS checklist scores was demonstrated, with a Pearson correlation coefficient of 0.877. There was no effect of the order of playing EmergenCSim™ or HFS first at Time 3 on the subsequent performance score.

While not statistically significant, there was a small positive correlation between EmergenCSim™ and HFS checklist scores ($r = 0.18$, $p = 0.195$). The EmergenCSim™ and knowledge MCQ test scores at Time 3 did not correlate ($r = 0.09$, $p = 0.50$). These assessments measure different knowledge and skill domains.

3.5.1 Survey results

Eighty percent of participants ($N = 49$) reported playing video games for leisure; electronic puzzles and brain games (e.g., Tetris) were most commonly reported. Significantly more males reported playing sports games (e.g., Madden NFL 25), first-person shooter games (e.g., Gran Turismo and Halo) ($p < 0.0001$) and massively multiplayer online games (e.g., World of Warcraft) ($p = 0.038$). Overall, the mean subject rating of the realism of EmergenCSim™ (scale 1–10) was 6.0 (SD 1.54). There was no difference in mean realism ratings between the EmergenCSim™ and Non-content Specific Game groups, U = 259, z = –0.818, $p = 0.413$; or among males and females, U = 274.5, z = –0.322, $p = 0.748$. Overall, the mean subject rating of HFS realism (scale 1—5) was 4 (SD 0.64).

4 Discussion

EmergenCSim™ is a novel serious game designed to teach performance of general anesthesia for emergent cesarean delivery. In our preliminary study aiming to evaluate its effectiveness as an educational and assessment tool, we have found that a single exposure did not yield significant differences in knowledge scores compared to controls. Unexpectedly, males had a steady improvement in knowledge scores over time whereas females' knowledge scores initially dropped slightly, but increased over time, suggesting that gender might impact learning outcomes with serious games; however, we do not believe that the small difference observed necessarily reflects a meaningful difference in learning based on gender. More robust study of this issue is necessary.

Given gender differences in neural responses to video games [21], spatial learning and memory [22], we speculate that our 3-D first-person shooter-style design might have been more engaging for males; however, the lack of gender difference in the EmergenCSim™ scores and serious game realism ratings do not actually support this notion. In order to make ongoing improvements, future surveys of player ratings of the realism of the game should explore more deeply the reasons underlying the rating given.

Although over 40% of gamers are women, game use and the popularity of different platforms vary by gender [23]. Mobile games are almost equally popular among men and women, but, whereas 48% and 37% of men play on personal computer and console games at least monthly respectively, only 35% and 23% of women do so [23]. Men and increasingly, younger women (26–30 years), are more likely to seek high-quality game experiences and spend greater time and money on gaming [13]. A third of female gamers report primarily using games as "time fillers", versus only 19% of males [13]. Correspondingly, our male trainees reported significantly higher use of 3-D first-person-shooter, sports, and massively multiplayer online games than females.

We were unable to demonstrate sufficient convergent validity of the EmergenCSim™ and HFS behavior checklist scores but found strong reliability levels. Convergent validity reflects how well scores of two instruments intended to measure a similar construct will correlate [24]. The 0.20 correlation is positive and in the expected direction, but surprisingly low since both tools were designed to measure similar concepts. The higher realism rating of HFS might be related to residents' pre-existing familiarity with simulation-based education.

We acknowledge several limitations to the (1) primary outcome measure (MCQ test), (2) serious game itself, and (3) sample size. First, MCQ tests may be suboptimal for measuring domains involving applied knowledge [25, 26]. The major limitation was insufficient difficulty of the knowledge test with less-than-acceptable score reliability levels. The test was developed in advance but subsequent validation revealed poor discrimination between experts and novices on several items [19], resulting in 31 of 49 participants (of whom 14 were male) achieving the cut score of 21/29 at baseline [19]. Building on this work, we are developing and validating parallel test forms with items having greater difficulty, to be used as pre- and post-intervention outcome measures in future studies. Stratified randomization was not considered due to small class sizes. In addition, time between the 3 tests (especially, Time 3) tapped primarily into knowledge retention; group differences may have been detected immediately afterwards. Non-standardized test conditions allowed for possible random measurement error and discussion/researching answers, which could have contaminated the results. Ideally two or more equivalent versions of the test should have been used to limit "test effect" [27, 28]. Conducting the final study phase (Time 3) over several months, remote from the initial intervention, diluted its impact and introduced threats to validity including, "history" and "maturation", explaining the lack of difference between groups in serious game and HFS performance [29]. Trainees probably discussed the activities within their peer groups, making it impossible to attain clean experimental conditions. In longitudinal studies, cases with extreme scores at earlier time points will score higher or lower in subsequent measurements due to regression towards the mean. This may have affected performance measures for females who started with extremely high scores in comparison to males.

Second, several conceptual frameworks and approaches to serious game design, development and validation exist [30, 31]. More meticulous usability testing may have provided more targeted design. We acknowledge that a single "exposure" to EmergenCSim™ was likely insufficient. Preliminarily, we could have determined the minimum "dose" of play to produce learning. Unfamiliarity with the EmergenCSim™ interface may have caused excessive cognitive load and detracted from learning [32]. We have since developed a practice game. Learner motivation would have influenced the degree to which the electronic debrief was used [33, 34].

Third, the small sample size may limit the ability to detect an effect, which we attempted to overcome by enrolling 2 CA-1 classes. A single program's experience with a novel, proprietary serious game limits the external validity of findings.

Finally, clinical performance was not assessed. Linking trainees' performance on EmergenCSim™ with actual provision of general anesthesia for emergency cesarean delivery, a rare occurrence itself, was not feasible. This is a well-recognized challenge for simulation-based training in general [35].

5 Conclusions

Serious games offer exciting avenues as scalable and flexible pedagogical tools with remote learning opportunities. We were unable to demonstrate improvement in learning outcomes following the brief serious game intervention, nor convergent validity with HFS, but this work contributes to the evidence-base. Future research into more usability testing, optimal intervention duration/dose, design tailored to player background characteristics and optimal debriefing strategies are warranted.

Abbreviations

SG	Serious game
CA-1	Clinical anesthesia-year one
HFS	High fidelity simulation
MCQ	Multiple choice question
3-D	Three dimensional
ANOVAs	Repeated measures analyses of variance

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1007/s44254-023-00016-4.

Additional file 1.

Additional file 2.

Acknowledgements

We would like to acknowledge the contribution of Professor Alexandre Mignon MD, MBA, PhD, who helped with the early design of the EmergenCSim^{TM} serious game platform during a Visiting Professorship with the Department of Anesthesiology, Columbia University, New York, NY, U.S.A.
We would like to acknowledge Woojin Shim, B.M, Simulation Center Manager, Margaret Wood Center for Simulation and Education, Department of Anesthesiology, Columbia University, New York, NY, U.S.A., who supported the technical aspects of the high-fidelity simulations for the study.
We would like to acknowledge the anesthesiology trainees and the Education Division staff of the Department of Anesthesiology, Columbia University, New York, NY, U.S.A., for their participation and support of study activities.
We would like to acknowledge the contribution of Chen-Miao Chen, PhD, Teachers College, Columbia University, New York, NY, U.S.A., who helped with preliminary data analysis.

Authors' contributions

All authors contributed to the study conception and design. Material preparation, data collection was performed by Allison Lee, Stephanie Goodman, Beatriz Corradini and Ruth Landau; data analysis was performed by Sophie Cohn and Madhabi Chatterji. The first draft of the manuscript was written by Allison Lee and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding

This study was funded by the 2016 Society for Obstetric Anesthesia and Perinatology (SOAP)/Gertie Marx Research and Education Grant.

Availability of data and materials

The datasets during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

Ethics approval was received from the Columbia University Institutional Review Board (IRB-AAQ8025) and written informed consent was received from all subjects.

Competing interests

The authors declare no competing interests.

Received: 21 January 2023 Revised: 6 April 2023 Accepted: 11 April 2023
Published online: 01 June 2023

References

1. Palanisamy A, Mitani AA, Tsen LC. General anesthesia for cesarean delivery at a tertiary care hospital from 2000 to 2005: a retrospective analysis and 10-year update. Int J Obstetr Anesthesia. 2011;20(1):10–6. https://doi.org/10.1016/j.ijoa.2010.07.002.
2. Guglielminotti J, Landau R, Li G. Adverse events and factors associated with potentially avoidable use of general anesthesia in cesarean deliveries. Anesthesiol. 2019;130(6):912–22. https://doi.org/10.1097/aln.0000000000002629.
3. D'Angelo R, Smiley RM, Riley ET, Segal S. Serious complications related to obstetric anesthesia: the serious complication repository project of the Society for Obstetric Anesthesia and Perinatology. Anesthesiol. 2014;120(6):1505–12. https://doi.org/10.1097/aln.0000000000000253.
4. Hawkins JL, Gibbs CP. General anesthesia for cesarean section: are we really prepared? Int J Obstetrc Anesthesia. 1998;7(3):145–6.
5. Lipman S, Carvalho B, Brock-Utne J. The demise of general anesthesia in obstetrics revisited: prescription for a cure. Int J Obstet Anesth. 2005;14(1):2–4. https://doi.org/10.1016/j.ijoa.2004.10.003.
6. Scavone BM, Toledo P, Higgins N, Wojciechowski K, McCarthy RJ. A randomized controlled trial of the impact of simulation-based training on resident performance during a simulated obstetric anesthesia emergency. Simul Healthc. 2010;5(6):320–4. https://doi.org/10.1097/SIH.0b013e3181e602b3.
7. Ortner CM, Richebe P, Bollag LA, Ross BK, Landau R. Repeated simulation-based training for performing general anesthesia for emergency cesarean delivery: long-term retention and recurring mistakes. Int J Obstet Anesth. 2014;23(4):341–7. https://doi.org/10.1016/j.ijoa.2014.04.008.
8. ACGME Program Requirements for Graduate Medical Education In Anesthesiology [Internet]. Chicago: Accreditation Council for Graduate Medical Education (ACGME); c2022-2023 [cited 2023 Apr 6]. Available from: https://www.acgme.org/globalassets/pfassets/programrequirements/040_anesthesiology_2022.pdf.
9. Juang J, Gabriel RA, Dutton RP, Palanisamy A, Urman RD. Choice of anesthesia for cesarean delivery: an analysis of the national anesthesia clinical outcomes registry. Anesth Analg. 2017;124(6):1914–7. https://doi.org/10.1213/ane.0000000000001677.
10. Ghoman SK, Patel SD, Cutumisu M, Hauff P, Jeffery T, Brown MRG, et al. Serious games, a game changer in teaching neonatal resuscitation? A review. Arch Dis Child Fetal Neonatal Ed. 2020;105(1):98–107. https://doi.org/10.1136/archdischild-2019-317011.
11. Nyssen AS, Larbuisson R, Janssens M, Pendeville P, Mayné A. A comparison of the training value of two types of anesthesia simulators: computer screen-based and mannequin-based simulators. Anesth Analg. 2002;94(6):1560–5.
12. Graafland M, Schraagen JM, Schijven MP. Systematic review of serious games for medical education and surgical skills training. Br J Surg. 2012;99(10):1322–30. https://doi.org/10.1002/bjs.8819.
13. Bosman S. Women account for 46% of all game enthusiasts: watching game video content and esports has changed how women and men alike engage with games. 2019 May 10 [cited 2023 Apr 6]. In Newzoo

Blog [Internet]. Amsterdam: Newzoo. c2019-2023. Available from: https://newzoo.com/insights/articles/women-account-for-46-of-all-game-enthusiasts-watching-game-video-content-and-esports-has-changed-how-women-and-men-alike-engage-with-games/.

14. Gentry SV, Gauthier A, L'Estrade Ehrstrom B, Wortley D, Lilienthal A, Tudor Car L, et al. Serious gaming and gamification education in health professions: systematic review. J Med Internet Res. 2019;21(3):e12994. https://doi.org/10.2196/12994.
15. Haerling KA. Cost-utility analysis of virtual and mannequin-based simulation. Simul Healthc. 2018;13(1):33–40. https://doi.org/10.1097/sih.0000000000000280.
16. Whitfill T, Auerbach M, Diaz MCG, Walsh B, Scherzer DJ, Gross IT, et al. Cost-effectiveness of a video game versus live simulation for disaster training. BMJ Simul Technol Enhanc Learn. 2020;6(5):268–73. https://doi.org/10.1136/bmjstel-2019-000497.
17. Kron FW, Gjerde CL, Sen A, Fetters MD. Medical student attitudes toward video games and related new media technologies in medical education. BMC Med Educ. 2010;10:50. https://doi.org/10.1186/1472-6920-10-50.
18. Chang HY, Wong LL, Yap KZ, Yap KY. Gaming preferences, motivations, and experiences of pharmacy students in Asia. Games Health J. 2016;5(1):40–9. https://doi.org/10.1089/g4h.2015.0028.
19. Lee AJ, Goodman SR, Banks SE, Lin M, Landau R. Development of a multiple-choice test for novice anesthesia residents to evaluate knowledge related to management of general anesthesia for urgent cesarean delivery. J Educ Perioper Med. 2018;20(2):E621.
20. Scavone BM, Sproviero MT, McCarthy RJ, Wong CA, Sullivan JT, Siddall VJ, et al. Development of an objective scoring system for measurement of resident performance on the human patient simulator. Anesthesiol. 2006;105(2):260–6.
21. Dong G, Wang L, Du X, Potenza MN. Gender-related differences in neural responses to gaming cues before and after gaming: implications for gender-specific vulnerabilities to internet gaming disorder. Soc Cogn Affect Neurosci. 2018;13(11):1203–14. https://doi.org/10.1093/scan/nsy084.
22. de Castell S, Larios H, Jenson J. Gender, videogames and navigation in virtual space. Acta Psychol (Amst). 2019;199:102895. https://doi.org/10.1016/j.actpsy.2019.102895.
23. Male and Female Gamers: How Their Similarities and Differences Shape the Games Market. 2017 May 3 [cited 2023 Apr 6]. In Newzoo Blog [Internet]. Amsterdam: Newzoo. c2017-2023. Available from: https://newzoo.com/resources/blog/male-and-female-gamers-how-their-similarities-and-differences-shape-the-games-market.
24. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. Am J Med. 2006;119(2):166.e7-16. https://doi.org/10.1016/j.amjmed.2005.10.036.
25. Chatterji M. Designing and using tools for educational assessment. Boston: Allyn and Bacon; 2003.
26. Sivarajan M, Miller E, Hardy C, Herr G, Liu P, Willenkin R, et al. Objective evaluation of clinical performance and correlation with knowledge. Anesth Analg. 1984;63(6):603–7.
27. Marsden E, Torgerson CJ. Single group, pre- and post-test research designs: some methodological concerns. Oxford Rev Educ. 2012;38(5):583–616. https://doi.org/10.1080/03054985.2012.731208.
28. Latimier A, Riegert A, Peyre H, Ly ST, Casati R, Ramus F. Does pre-testing promote better retention than post-testing? NPJ Sci Learn. 2019;4:15. https://doi.org/10.1038/s41539-019-0053-1.
29. Flannelly KJ, Flannelly LT, Jankowski KRB. Threats to the internal validity of experimental and quasi-experimental research in healthcare. J Health Care Chaplain. 2018;24(3):107–30. https://doi.org/10.1080/08854726.2017.1421019.
30. Tan JW, Zary N. Diagnostic markers of user experience, play, and learning for digital serious games: a conceptual framework study. JMIR Serious Games. 2019;7(3):e14620. https://doi.org/10.2196/14620.
31. Verschueren S, Buffel C, Vander Stichele G. Developing theory-driven, evidence-based serious games for health: framework based on research community insights. JMIR Serious Games. 2019;7(2):e11565. https://doi.org/10.2196/11565.
32. Fraser KL, Ayres P, Sweller J. Cognitive load theory for the design of medical simulations. Simul Healthc. 2015;10(5):295–307. https://doi.org/10.1097/sih.0000000000000097.
33. Cook DA, Artino AR Jr. Motivation to learn: an overview of contemporary theories, (in eng). Med Educ. 2016;50(10):997–1014. https://doi.org/10.1111/medu.13074.
34. Cutumisu M, Brown MRG, Fray C, Schmolzer GM. Growth mindset moderates the effect of the neonatal resuscitation program on performance in a computer-based game training simulation. Front Pediatr. 2018;6:195. https://doi.org/10.3389/fped.2018.00195.
35. Gaba DM. Simulation is a critical tool for advancing patient safety-available to everyone regardless of location or resources. Anesth Pat Safe Found Newsl. 2019;33(3):96–97.

Publisher's Note