Validating Parallel-Forms Tests for Assessing Anesthesia Resident Knowledge

Journal of Medical Education and Curricular Development
Volume 11: 1–8
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/23821205241229778

Allison J. Lee[1*], Stephanie R. Goodman[1], Melissa E. B. Bauer[2], Rebecca D. Minehart[3], Shawn Banks[4], Yi Chen[5], Ruth L. Landau[1] and Madhabi Chatterji[5]

[1]Department of Anesthesiology, Columbia University, New York, NY, USA. [2]Department of Anesthesiology, Duke University, Durham, NC, USA. [3]Department of Anesthesia, Critical Care and Pain Medicine, Harvard University, Boston, MA, USA. [4]Department of Anesthesiology, Perioperative Medicine and Pain Management, University of Miami, Miami, FL, USA. [5]Teachers College, Columbia University, New York, NY, USA.

ABSTRACT: We created a serious game to teach first year anesthesiology (CA-1) residents to perform general anesthesia for cesarean delivery. We aimed to investigate resident knowledge gains after playing the game and having received one of 2 modalities of debriefing. We report on the development and validation of scores from parallel test forms for criterion-referenced interpretations of resident knowledge. The test forms were intended for use as pre- and posttests for the experiment. Validation of instruments measuring the study's primary outcome was considered essential for adding rigor to the planned experiment, to be able to trust the study's results. Parallel, multiple-choice test forms development steps included: (1) assessment purpose and population specification; (2) content domain specification and writing/selection of items; (3) content validation by experts of paired items by topic and cognitive level; and (4) empirical validation of scores from the parallel test forms using Classical Test Theory (CTT) techniques. Field testing involved online administration of 52 shuffled items from both test forms to 24 CA-1's, 21 second-year anesthesiology (CA-2) residents, 2 fellows, 1 attending anesthesiologist, and 1 of unknown rank at 3 US institutions. Items from each form yielded near-normal score distributions, with similar medians, ranges, and standard deviations. Evaluations of CTT item difficulty (item p values) and discrimination (D) indices indicated that most items met assumptions of criterion-referenced test design, separating experienced from novice residents. Experienced residents performed better on overall domain scores than novices ($P < .05$). Kuder-Richardson Formula 20 (KR-20) reliability estimates of both test forms were above the acceptability cut of .70, and parallel forms reliability estimate was high at .86, indicating results were consistent with theoretical expectations. Total scores of parallel test forms demonstrated item-level validity, strong internal consistency and parallel forms reliability, suggesting sufficient robustness for knowledge outcomes assessments of CA-1 residents.

KEYWORDS: multiple-choice tests, parallel forms reliability, Classical Test Theory, instrument validation, general anesthesia, cesarean delivery

RECEIVED: August 3, 2023. **ACCEPTED:** January 12, 2024.

TYPE: Methodology

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was funded by a Foundation for Anesthesia Education and Research, Research in Education Grant (Schaumburg, IL), [Grant I.D.: REG-08-15-17-Lee (Allison)].

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

*Current affiliation: Department of Anesthesiology and Critical Care, University of Pennsylvania, Philadelphia, PA, USA.

CORRESPONDING AUTHOR: Allison J. Lee, Department of Anesthesiology and Critical Care, University of Pennsylvania, 3400 Spruce Street, Suite 680 Dulles Philadelphia, PA 19104, USA. Email: allison.lee@pennmedicine.upenn.edu

Introduction

A sharp decline in the use of general anesthesia (GA) for cesarean delivery (CD) has led to significant reductions in maternal morbidity and mortality in recent decades, but GA-related complications and failed intubation rates among parturients compared with nonpregnant women are still high (1:500) due to intrinsic maternal, fetal, and situational factors.[1] Physiologic changes of pregnancy (PCP), such as more vascular and edematous airway mucosa, increase the difficulty of intubation.[2] Additionally, GA is usually conducted as an emergency in obstetric units remote from assistance, and with time pressure resulting in poor preparation and performance of technical tasks.[2] Furthermore, the practice shift favoring neuraxial anesthesia over GA has not only led to a decline in airway skills[2] but has created the unintended consequence of decreased clinical exposure of anesthesiology trainees to this scenario, necessitating the development of alternative approaches to teaching the necessary knowledge and skills.[3,4]

In 2016, our research group developed a novel 3-D serious video game,[5] EmergenCSim^{TM}, designed to teach novice first-year anesthesia (CA-1) residents the knowledge and skills to perform GA for emergency CD.[6] Electronic feedback appears automatically at the end of gameplay explaining the expected actions and the underlying rationale for this clinical scenario. A detailed description of the game design and development has been previously published with a report of a single-blinded longitudinal experiment where novice CA-1 residents were randomized to play EmergenCSim^{TM} or a noncontent specific (sham) game that had no embedded electronic feedback.[6] There was no difference between experimental groups in test scores, but a post hoc exploratory analysis found a slight improvement in male residents' scores over time, suggesting that gender may impact learning outcomes with serious games.

We developed 2 parallel forms of a criterion-referenced multiple-choice test (Form A and Form B) for use in a randomized experiment to compare knowledge gains after playing EmergenCSim^{TM}. Residents would be randomized to experience, after playing the game, either the game-embedded electronic feedback alone (control group) versus electronic feedback and in-person debriefing (intervention group). Form

A was designed to be used as a pretest at baseline and Form B as a posttest after playing EmergenCSim^{TM} and having received one of the 2 latter debriefing conditions. Our experimental hypothesis was that the group receiving the additional in-person debrief would demonstrate superior improvement in knowledge and skills after playing the game, evidenced by greater improvement in mean multiple-choice test scores from baseline. Pre- and posttests with different questions but which measure the same content are a more robust approach to measuring learning gains, by avoiding measurement error due to subjects memorizing answers or learning from the pretest.[7] Validation of the instruments measuring the study's primary outcome was considered essential for adding rigor to the planned experiment, to be able to trust the study's results. Here, we describe the multiphase design process for the development, content validation, and empirical validation of scores from the parallel test forms.

Materials and Methods

The Columbia University Institutional Review Board approved this study (Protocol #: AAAR6903) in December 2017. The study activities that involved educational tests whereby the identity of the human subjects cannot readily be ascertained, directly or through identifiers to the subjects were considered to be exempt.

Test development comprised 4 phases, according to Chatterji's process model, which is a framework that takes an intentionally integrated approach to design and validation, situated in, and guided by, the specified test user contexts.[8] The process model draws on long-standing theory on assessment design and validity tied to the intended interpretations and uses of test scores and is broadly consistent with the *Standards for Educational and Psychological Testing* and recommendations of leading educational test developers.[9-13] It is unique in endorsing user-centered design principles and taking a unified validation approach that integrates evidence of validity, reliability, and utility as a whole to evaluate the quality of a test/test scores.[10,14,15] The use of the model was appropriate here as there were no existing tests for CA-1 residents in the obstetric anesthesiology domain of interest.

(1) Assessment Purpose and Population Specification: The total scores from each form were intended to permit inferences on absolute proficiency levels CA-1 residents on the tested domain, justifying their use as outcomes measures in the longitudinal experiment. The targeted population is novice CA-1 anesthesia trainees, not previously exposed to obstetric anesthesia.

(2) Specification of the Content Domains and Writing/Selection of Items by one Internal and 2 External Experts: The tested domain was resident knowledge regarding the conduct of GA for emergency CD. The expected learning outcomes for each subconstruct domain were:

(i) Physiologic Changes of Pregnancy (9 items): The physician can describe the normal PCP (eg, airway changes, pulmonary changes) which underlie the differences in management of GA in a pregnant versus a nonpregnant patient.

(ii) Pharmacology (PHA, 4 items): The physician can correctly apply understanding of pharmacokinetic and pharmacodynamic changes in pregnancy to appropriately manage medications utilized during GA for CD.

(iii) Anesthetic Implications of Pregnancy (7 items): The physician can correctly apply underlying knowledge about the PCP and drug pharmacokinetic/pharmacodynamic changes unique to pregnancy, to make appropriate clinical decisions when performing GA for CD.

(iv) Crisis Resource Management (CRM, 6 items): The physician can correctly identify crisis management, communication, and teamwork skills during emergency GA for CD, when given scenarios.

The essential skills and knowledge were based on Scavone et al's[16] content-validated weighted behavior checklist developed for this scenario. The obstetric anesthesia rotation assigned textbook was the main reference.[17] For each competency identified for the subconstructs, at least 1 item, comprising a stem, 1 correct answer, and 3 distractors was developed. To start, an initial pool of questions was generated from a previously validated and field-tested instrument.[18] Weaknesses identified were addressed by dropping poorly performing items (5 prior items—#2, #10, #20, #23, and #28) determined to be too easy and/or not aligned with content covered in the game, or by revising highly content-relevant items that failed to discriminate between novices and experts (items with D values <15). Additionally, new questions were written, ensuring there were at least 2 parallel items for each form. Question writers were instructed to aim for a "higher order thinking" cognitive level that tests applied knowledge.

(3) Design of Matched Pairs of Items for New Parallel Test Forms: For designing the parallel items, we created distinct pairs of items that measured the same specific content and cognitive skill area, while balancing the item distribution with respect to 3 levels of cognitive demand (i) concept recall and understanding, (ii) application, and (iii) higher order thinking.

A table of "Test Design Specifications" categorizing the items in each subdomain according to the 3 cognitive levels was created, to achieve equal proportions of multiple-choice items for each cell.[19]

(4) Content Validation by Experts: The annotated question bank was content validated by 3 obstetric anesthesia experts, of which 2 were external (SG, MB, and RM). Seventy-two questions were reviewed; the prior discrimination index values for retained questions and the original versions of revised items were provided.

The experts were asked to rate the content relevance of each item on a 4-point anchored scale, where 0 = not relevant, 1 = somewhat relevant, 2 = quite relevant, and 3 = highly relevant, and to provide comments/criticisms about the questions. They were also invited to propose new questions according to the specifications explained.

In each round of evaluation, an item-level content validation index (I-CVI) was calculated based on the number of experts rating relevance of an item as a 2 or 3, divided by the total number of experts (the proportion that were in agreement about relevance).[20] An I-CVI greater than 0.78 is considered excellent regardless of the number of experts.[20] Feedback given was used to revise the items and experts rated the revised items (a total of 3 rounds) until consensus was achieved regarding the design and relevance of all items. Items that performed very poorly and were not considered relevant were removed from the test. This ultimately yielded a total of 26 paired items (52 in all) to be allocated to finalized versions of each parallel test form (copies of the shuffled questions for both forms and answer keys are in Supplemental File 1).

(5) Field Testing and Empirical Validation: When the pool of 52 questions was finalized, shuffled items from both forms were uploaded to an online platform (Qualtrics^{XM}). The combined test was distributed, along with instructions, via an email-embedded link to volunteer participants from 3 institutions (Jackson Memorial Hospital/University of Miami, NewYork-Presbyterian Hospital/Columbia University, and Massachusetts General Hospital/Harvard University). The email explained that responses would be collected anonymously, and that participation was considered to serve as agreement to participate in the study.

Inclusion criteria included trainees of multiple levels of experience, with and without prior obstetric anesthesia experience and obstetric anesthesia fellowship-trained faculty. The only exclusion criterion was refusal to participate.

Residency class sizes nationally in the United States are generally small (mean of 13 (SD 7; range 3-30)).[21] Trainees at the 3 participating institutions were invited to participate. An intact cohort of CA-1 anesthesiology residents was successfully recruited from the University of Miami, along with more senior trainees from the 3 institutions. The available pool of fellows (N = 2) and obstetric anesthesia fellowship-trained experts from which volunteer participation was solicited at one of the institutions were also naturally small (total N = 7).

Survey items were included at the end of the test to collect background information from participants on institution, current training status (PGY-1 -fellow or faculty), age group (<25 years, 26-30 years, 31-35 years, 36-40 years, ≥41 years), self-reported gender, number of prior obstetric anesthesia rotations, and prior experience performing GA for either CD or nonobstetric surgery in pregnant women.

(6) Psychometric Analysis: Item analysis statistics, as p_i (item difficulty) and D (discrimination index) values, were obtained using Classical Test Theory (CTT) techniques.[22] The value of p_i, which represents the proportion of examinees who answer an item correctly, may range from 0.00 to 1.00. The discrimination index in this case indicates how well the item discriminates between the novices (juniors who have never completed an obstetric anesthesia rotation) and the experienced (have completed at least one obstetric anesthesia rotation).

Item D values should be evaluated in concert with item p_i values. Multiple-choice item p values should not be performing at the chance level. In multiple-choice items with 4 options, a chance level (suggesting random responses) is .25 (25% identified the correct response).

For enhancing validity per CTT, criterion-referenced interpretations of scores on proficiency test require that items discriminate sufficiently between experienced persons (those with prior exposure to the domain) versus novices (those who are unexposed to the same). Guidelines suggest that, assuming items meet p_i criteria, if negative D or D < 0.10, the item should be removed or examined closely; and if positive D ≥ 20%, the item is functioning well.[22-24] A D of 0 suggests no discrimination, which is not an ideal result for a criterion-referenced test, suggesting that the item is too difficult or too easy for all levels of examinees.

(7) Reliability of Scores: Internal consistency reliability estimates for subdomain and total scores were determined using KR-20 and the Parallel Forms reliability estimate with Pearson's correlation of total test scores of each test form.

(8) Validity evidence based on expected group differences.[9] We hypothesized that greater experience would be associated with higher scores on the test. The overall sample comprised 49 subjects. All analyses were performed using SPSS statistical software (version 20.0; IBM Corporation, Armonk, NY). A *P* value ≤ .05 was considered to be statistically significant in an examination of 2 group mean differences on a combined test including all 52 items.

A detailed description of the development of EmergenCSim^{TM}, and the complete list of items in the electronic feedback checklist, including the weighted scoring for each item have been previously published.[6] Congruent with the knowledge test, the checklist of expected actions within the game and the weighted scoring system were based on the content-validated weighted behavior checklist developed for this scenario by Scavone et al.[16]

For the subsequent experiment, the 10-min in-person semi-structured debriefing integrated concepts from the Promoting Excellence and Reflective Learning in Simulation (PEARLS) debriefing framework[25] and was conducted by AL. Subjects were invited to reflect on their actions taken in the game and the aspects of clinical management within the scenario using questions such as "Can you walk me through what you were thinking when you were asked to put this patient to sleep emergently?" and "Were there any aspects of the explanations given that you did not understand or need help clarifying?." Whenever gaps in knowledge or understanding of the concepts being taught were identified, direct teaching was provided. Strategies for scoring better in the game were not discussed.

Results

Field-testing occurred during July to December 2019 on a sample of CA-1 residents (N = 24), CA-2 residents (N = 21), CA-3 residents (N = 1), fellow (N = 2), and faculty anesthesiologists (N = 1), (total N = 49) from the 3 US medical institutions described above. The demographics and background characteristics of participants are described in Tables 1 and 2, respectively. Items were scored with a binary key denoting right (1 point) and wrong (0 points) answers. The 52 items

Table 1. Subject demographic variables (N = 49).

	N	%
(Q53)[a] What is your current institutional affiliation?		
University of Miami	44	92.97
Columbia University	2	5.17
Harvard University	2	5.17
Missing	1	3.08
(Q54) What is your current training status?		
PGY-1	0	0.00
CA-1 Resident	24	51.00
CA-2 Resident	21	44.75
CA-3 Resident	1	3.08
Anesthesia Fellow	1	3.08
Anesthesia Faculty	1	3.08
Missing	1	3.08
(Q55) What is your age range (years)?		
Smaller or equal to 25	0	0.00
26-30	36	76.00
31-35	9	19.75
36-40	0	0.00
Bigger or equal to 40	3	7.25
Missing	1	3.08
(Q59)[a] What is your gender?		
Female	21	44.75
Male	22	57.25
Missing	1	3.08

[a] Q53 and Q59 are nominal scale.

Table 2. Subject background characteristics (N = 49).

	f	%
(Q54) What is your current training status?		
PGY-1	0	0.00
CA-1 Resident	24	51.00
CA-2 Resident	21	44.75
CA-3 Resident	1	3.08
Anesthesia Fellow	1	3.08
Anesthesia Faculty	1	3.08
Missing	1	3.08
(Q60) How many obstetric anesthesia rotations have you completed?		
0	23	48.92
1	15	32.25
2	7	15.58
More than or equal to 3	1	3.08
I have completed residency training	2	5.17
Missing	1	3.08
(Q61) How much experience have you had performing general anesthesia for cesarean delivery?		
Never	23	48.92
1-2 times	11	23.92
3-5 times	9	19.75
5-10 times	4	9.33
More than or equal to 11	1	3.08
Missing	1	3.08
(Q62) How much experience have you had performing general anesthesia in pregnant women for nonobstetric surgery?		
Never	33	69.75
1-2 times	9	19.75
3-5 times	1	3.08
5-10 times	1	3.08
More than or equal to 11	4	9.33
Missing	1	3.08
(Q63) How often do you play video games?		
Never	8	17.67
Rarely (1 time per year or less)	13	28.08
Occasionally (1-6 times per year)	11	23.92
Often (7-12 times per year)	4	9.33
Very often (more than 1 time per month)	12	26.00
Missing	1	3.08

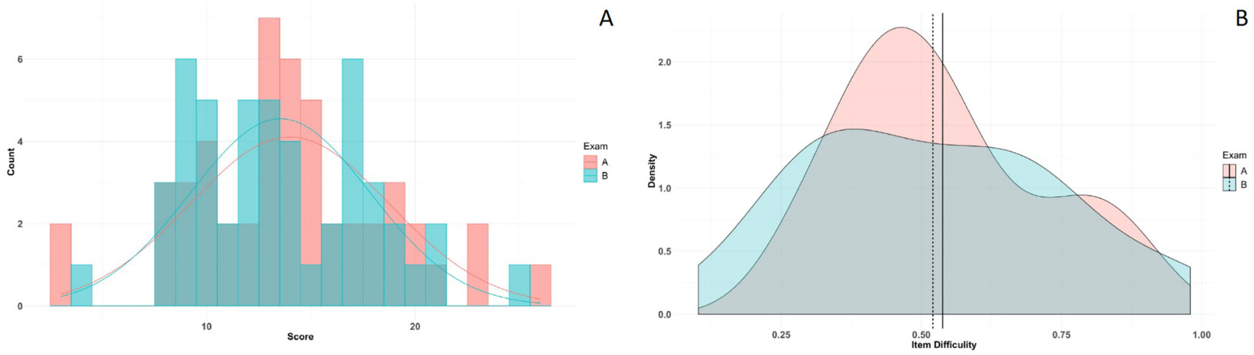

Figure 1. (A) Total score distribution for forms A and B; medians, ranges, and standard deviations were less than 2 raw score points apart—the median score (out of maximum 26) for form A was 14 and for Form B was 13. (B) Item Difficulty Distribution for Forms A and B; item p values showed a near normal distribution on Form A, but a flatter distribution with Form B.

Table 3. Descriptive statistics on the total score of parallel test forms A and B.

Exams	# item	# participants	Mean score	Median score	Min score	Max score	SD	KR-20
Form A	26	49	13.98	14	3	26	4.76	0.77
Form B	26	49	13.53	13	4	25	4.29	0.74
Parallel Forms Reliability (Pearson Correlation)				0.85				

were separated randomly by cells of the Table of Test Specifications into 2 different parallel examinations—"Form A" and "Form B." Each parallel form yielded 4 subdomain scores and a total score which were investigated for overall construct validity.

The 49th subject showed several missing values in the item response data. This subject did not answer 13 out of 52 items from Forms A and B combined. The data on all items for which valid responses had been received were retained, with the missing responses scored as incorrect. Given the relatively small sample size, this treatment allowed the retention of greater information regarding the item responses as well as avoiding the introduction of unacceptable bias into the analysis.

Measurement theory on parallel forms test design assumes similar distribution for true score and error score distributions. The "error" in measurement refers to the discrepancy in the observed score of the test taker and their "true score," which is the average score that would be observed from many repeated testings.[26]

Consistent with CTT expectations,[22] both test forms yielded similar score distributions that were near-normal (Figure 1A), with medians, ranges, and standard deviations that were less than 2 raw score points apart—the median score (out of maximum 26) for Form A was 14 and for Form B was 13 (Table 3). The KR-20[22] values calculated were well above the minimum of 0.70, with a robust parallel forms reliability of 0.86.

Joint evaluations of CTT item difficulty indicated that the majority of items performed per assumptions of criterion-referenced test design, separating experienced residents from novices.

Item Analysis Statistics

Item p values showed a near normal distribution on Form A, but a flatter distribution with Form B (Figure 1B). The item statistics (see Supplemental Files 2 and 3) were calculated for the total sample and showed one problematic item on Form B (see Supplemental File 3—item #7B). A summary of the item analysis statistics is in Table 4.

When the sample was broken down into experienced and novice groups to investigate if items functioned similarly in both groups, a few added items on Form A seemed to function in the reverse direction (negative D values) where novices performed better than experienced residents (see Supplemental File 2—items #16A, #18A, #21A, #23A, #26A).

Validity Evidence Based on Hypothesized Group Differences

Testing for hypothesized group differences on the overall construct domain measured (all 52 items) verified that greater seniority was associated with better performance on the test. Statistically significant differences were established between subgroups of the total sample broken down by year of residency,

Table 4. Summary item analysis statistics for form A and B.

Statistics	Form A	Form B
Number of items	26	26
Number of Examinees	49	49
Mean Difficulty (SD)	0.54 (0.18)	0.52 (0.23)
Mean Discrimination[a] (SD)	0.38 (0.18)	0.36 (0.18)
Easy Items (%)[b]	0	3.85
Items with Negative Discrimination Indices (%)	3.85	3.85

[a]Discrimination is calculated based on adjusted point-biserial correlations.
[b]Easy items are those with discrimination less than 0.2 and difficulty is greater than 0.8.

the number of rotations completed and more experience performing GA for CD ($P < .05$). However, according to the exploratory analysis above, there was no significant effect ($P = .05$) in groups sorted by gender, age, prior video game exposure, and experience in administering GA to pregnant women on the total test scores.

Subdomain performance (Supplemental File 4): The reliability levels of the subdomain scores on both forms were generally low, at <0.70. This was possibly due to a combination of low homogeneity of content tested and too few items in each subdomain. Hence, only total scores are recommended for use for educational evaluations with either test form.

Discussion

We developed content-validated criterion-referenced parallel test forms, designed to test CA-1 novice residents' knowledge regarding performance of GA for CD at baseline (pretest, Form A) and after playing a serious video game (posttest, Form B). The tests demonstrated item-level validity, strong internal consistency and parallel forms reliability, and validity based on expected group differences in performance of experienced versus novice residents. Together these results confirmed our hypotheses, suggesting that the scores are sufficiently valid and reliable for the purposes specified and consistent with the underlying construct theory about the measures.[22,26]

The study in which the tests were designed to be used would require novice first-year anesthesiology residents (CA-1) to take the 26-item pretest and play EmergenCSim^{TM}. They would be randomized to either the control group which experienced the game-embedded electronic feedback alone or the intervention group, which experienced the electronic feedback and an in-person debriefing. All subjects would then play the game a second time, and take the 26-item posttest. The primary outcome was to be the difference between experimental groups in the change in mean score from pretest to posttest.

The favorable empirical performance of the test forms may be attributed to our systematic item development and content validation processes.[8] The process of construct domain analysis ensured that the construct was adequately represented by the pool of items developed. Test development was an iterative process and benefitted from data derived from earlier empirical validation,[18] which prompted dropping or revising items that were performing poorly. Content validation relied on expert ratings of relevance of individual items and computation of an I-CVI, which is an index of inter-expert agreement adjusting for chance.[20]

The construction of a validity argument is based on collecting evidence to support inferences to be made from test scores.[9] Content validity indicates that the relationship between the content tested and thought processes of the test-takers and the intended construct is sound.[27] The advantage of CVI over other computed approaches, such as consistency estimates, consensus estimates, and measurement estimates, is the ease of computation, however, one drawback is a higher risk of chance agreement between experts.[20]

The emphasis of our empirical validation steps focused on estimation of a parallel forms reliability coefficient and investigating internal consistency reliability of each test form with the KR-20 formula.[22] The KR-20 compares the sum of the item score variances in the numerator with the variance of the summated total score on an instrument in the denominator and can therefore be interpreted as another measure of item homogeneity.

In future, construct validity of scores from the parallel test forms could be further investigated based on evidence of internal structure, gathered by performing exploratory and confirmatory factor analysis or a unidimensionality analysis with item response theory models.[28]

The parallel forms reliability of 0.86 is well above the acceptable minimum standard of 0.70.[22] Our finding that seniority of the physicians was associated with better performance on the test, provides added credence for the measures.

The test forms were developed for use in a randomized controlled trial exploring the educational utility of a novel serious video game. Given the results, an improvement in test scores from pretest to posttest in the forthcoming study can be interpreted meaningfully against the specified construct domain, and with precision. The 2 parallel forms were developed in order to limit a "testing effect," a potential threat to the internal validity of the forthcoming experiment whereby test takers become familiar with the items and remember the responses for later testing.[29] Strong validity and reliability of our outcome measures was important in order to trust the outcomes of our research.

A limitation of our study is that we were not able to fully control test-taking conditions of volunteers in disparate locations and to exclude factors which may have exerted nonrandom influence on scores (bias or construct-irrelevant variance) or random measurement error.[27] This, however, would have applied to only a minority of subjects—the largest group of subjects (from University of Miami) took the test in one sitting. Another limitation is that a power analysis was not performed. A pragmatic approach was taken, soliciting volunteers from multiple institutions, given the fixed small residency and fellowship class sizes and small numbers of obstetric anesthesia fellowship-trained faculty. A total sample size of N > 30 was achieved, and a retroactive power analysis indicated that there would be >80% power for a 2-group independent means comparison between test forms.

One weakness of the parallel forms was relatively low reliability of the subdomain scores on both forms (<0.70), likely due to low homogeneity of the content tested and too few items by subdomain. This result suggests that the total scores on the forms should be used rather than the subdomain scores. Another weakness found was that an item on Form A (item 23A, Supplemental File 2) seemed to function in the reverse direction with negative D values and will need to be deleted or revised for future iterations. Particular attention will be paid to items with *P* values <2 and negative discrimination data. Negatively discriminating items are least desirable, representing items missed by high-scoring examinees and answered correctly by low-scoring examinees.[22] We will continue to gather validity evidence from the use of the parallel forms on subjects of the forthcoming experiment.

Multiple-choice question-based testing is convenient and widely used for assessing knowledge in healthcare education and research. Scores must allow the intended interpretation in order to make the correct conclusions about learner knowledge and skills. Although the psychometric properties of our multiple-choice test forms are sound, this modality may be suboptimal for assessing other important behavioral domains of communication and CRM.

Conclusion

Outside of high-stakes testing situations, few validated parallel test forms exist to assess resident learning in domains related to obstetric anesthesia. Although specifically developed for assessing learning outcomes following a novel video game, we believe our parallel tests are sufficiently robust to be utilized for formative assessment of novice anesthesiology resident knowledge related to performing GA for CD, for which there is, unavoidably, diminishing exposure. Educators could use poor performance on the test to identify knowledge gaps, which could be addressed by assigning further reading, direct teaching or participation in other simulation-based teaching techniques.

Authors' contribution

All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed by Allison Lee, Stephanie Goodman, Melissa Bauer, Rebecca Minehart, Shawn Banks, Yi Chen, Ruth Landau, 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.

Authors' note

Data Availability: The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request. *Code Availability:* The code used during the analysis of the current study are available from the corresponding author on reasonable request.

Supplemental material

Supplemental material for this article is available online.

REFERENCES

1. Guglielminotti J, Landau R, Li G. Adverse events and factors associated with potentially avoidable use of general anesthesia in cesarean deliveries. *Anesthesiology*. 2019;130(6):912-922. doi:10.1097/aln.0000000000002629
2. Mushambi MC, Kinsella SM, Popat M, et al. Obstetric Anaesthetists' Association and Difficult Airway Society guidelines for the management of difficult and failed tracheal intubation in obstetrics. *Anaesthesia*. 2015;70(11):1286-1306. doi:10.1111/anae.13260
3. Hawkins JL, Gibbs CP. General anesthesia for cesarean section: are we really prepared? *Int J Obstet Anesth*. 1998;7(3):145-146.
4. 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):341-347. doi:10.1016/j.ijoa.2014.04.008
5. Maheu-Cadotte M-A, Cossette S, Dubé V, et al. Efficacy of serious games in healthcare professions education: a systematic review and meta-analysis. *Simul Healthc*. 2021;16(3):199-212. doi:10.1097/sih.0000000000000512
6. Lee AJ, Goodman S, Corradini B, Cohn S, Chatterji M, Landau R. A serious video game—EmergenCSim™—for novice anesthesia trainees to learn how to perform general anesthesia for emergency cesarean delivery: a randomized controlled trial. *Anesthesiol Perioper Sci*. 2023;1(2):14. doi:10.1007/s44254-023-00016-4
7. Willson VL, Putnam RR. A meta-analysis of pretest sensitization effects in experimental design. *Am Educ Res J*. 1982;19(2):249-258. doi:10.2307/1162568
8. Chatterji M. *Designing and Using Tools for Educational Assessment*. Chap 5. Allyn&Bacon/Pearson; 2003:105-110.
9. American Educational Research Association, American Psychological Association, National Council on Measurement in Education, eds. *Standards for Educational and Psychological Testing*. American Educational Research Association; 2014.
10. Cronbach LJ. Five perspectives on the validity argument. In: *Test Validity*. Lawrence Erlbaum Associates, Inc; 1988:3-17.
11. Kane MT. Validating the interpretations and uses of test scores. *J Educ Meas*. 2013;50(1):1-73. doi:10.1111/jedm.12000
12. Messick S. Meaning and values in test validation: the science and ethics of assessment. *Educ Res*. 1989;18(2):5-11. doi:10.2307/1175249
13. Shepard L. Validity for what purpose? *Teach Coll Rec*. 2013;115:1-12. doi:10.1177/016146811311500907
14. Giacomin J. What is human centred design? *Design J*. 2014;17(4):606-623. doi:10.2752/175630614X14056185480186
15. Still B, Crane K. *Fundamentals of User-Centered Design: A Practical Approach*. CRC Press; 2016.
16. Scavone BM, Sproviero MT, McCarthy RJ, et al. Development of an objective scoring system for measurement of resident performance on the human patient simulator. *Anesthesiology*. 2006;105(2):260-266.
17. Chestnut D, Wong C, Tsen L, et al. *Chestnut's Obstetric Anesthesia: Principles and Practice*. 6th ed. Elsevier; 2019.
18. 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.
19. Chatterji M. *Designing and Using Tools for Educational Assessment*. Allyn&Bacon/Pearson; 2003:159-161.
20. Polit DF, Beck CT, Owen SV. Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Res Nurs Health*. 2007;30(4):459-467. doi:10.1002/nur.20199

21. Huang J, Licatino LK, Long TR. Methods of orienting new anesthesiology residents to the operating room environment: a national survey of residency program directors. *J Educ Perioper Med*. 2020;22(3):E645. doi:10.46374/volxxii-issue3-Licatino
22. Crocker LM, Algina J. *Introduction to Classical and Modern Test Theory*. Wadsworth Group/Thomas Learning; 2006:311-335.
23. Chatterji M, Graham MJ, Wyer PC. Mapping cognitive overlaps between practice-based learning and improvement and evidence-based medicine: an operational definition for assessing resident physician competence. *J Grad Med Educ.* 2009;1(2):287-298. doi:10.4300/jgme-d-09-00029.1
24. Wyer PC. Designing outcome measures for the accreditation of medical education programs as an iterative process combining classical test theory and Rasch measurement. *Int J Educ Psychol Assess.* 2013;13(2):35-61.
25. Eppich W, Cheng A. Promoting Excellence and Reflective Learning in Simulation (PEARLS): development and rationale for a blended approach to health care simulation debriefing. *Simul Healthc*. 2015;10(2):106-115. doi:10.1097/sih.0000000000000072
26. Crocker LM, Algina J. *Introduction to Classical and Modern Test Theory*. 2006:101-154.
27. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med*. 2006;119(2):166.e7-166.e16. doi:10.1016/j.amjmed.2005.10.036
28. Floyd FJ, Widaman KF. Factor analysis in the development and refinement of clinical assessment instruments. *Psychol Assess.* 1995;7(3):286-299. doi:10.1037/1040-3590.7.3.286
29. Creswell JW, Creswell JD. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. 5th ed. Sage Publications; 2018.