

Establishing Validity for a Vaginal Hysterectomy Simulation Model for Surgical Skills Assessment

Chi Chung Grace Chen, MD, MHS, Ernest G. Lockrow, DO, Christopher C. DeStephano, MD, MPH, Mikio A. Nihira, MD, MPH, Catherine Matthews, MD, Leslie Kammire, MD, Lisa M. Landrum, MD, PhD, Bruce D. Anderson, PhD, and Douglas Miyazaki, MD

OBJECTIVE: To use the Messick validity framework for a simulation-based assessment of vaginal hysterectomy skills.

METHODS: Video recordings of physicians at different levels of training and experience performing vaginal hysterectomy on a high-fidelity vaginal surgery model were objectively assessed using a modified 10-item Vaginal Surgical Skills Index, a one-item global scale of overall performance, and a pass-fail criterion. Participants included obstetrics and gynecology trainees and faculty from five institutions. Video recordings were independently assessed by expert surgeons blinded to the identities of the study participants.

RESULTS: Fifty surgeons (11 faculty, 39 trainees) were assessed. Experience level correlated strongly with both

the modified Vaginal Surgical Skills Index and global scale score, with more experienced participants receiving higher scores (Pearson $r=0.81$, $P<0.001$; Pearson $r=0.74$, $P<0.001$). Likewise, surgical experience was also moderately correlated with the modified Vaginal Surgical Skills Index and global scale score (Pearson $r=0.55$, $P<0.001$; Pearson $r=0.58$, $P<0.001$). The internal consistency of the modified Vaginal Surgical Skills Index was excellent (Cronbach's $\alpha=0.97$). Interrater reliability of the modified Vaginal Surgical Skills Index and global scale score, as measured by the intraclass correlation coefficient, was moderate to good (0.49–0.95; 0.50–0.87). Using the receiver operating characteristic curve and the pass-fail criterion, a modified Vaginal Surgical

From the Department of Gynecology and Obstetrics, Johns Hopkins University School of Medicine, Baltimore, Maryland; the Department of Obstetrics and Gynecology, Uniformed Services University of the Health Sciences, Bethesda, Maryland; the Department of Gynecology, Mayo Clinic, Jacksonville, Florida; the Department of Obstetrics and Gynecology, University of California Riverside, Riverside, California; the Department of Obstetrics and Gynecology, Oklahoma University College of Medicine, Oklahoma City, Oklahoma; Augmented Reality Systems Inc., Windham, New Hampshire; and the Department of Obstetrics and Gynecology, Wake Forest School of Medicine, Winston-Salem, North Carolina. At the time of study, all participating institutions were members of the American College of Obstetricians and Gynecologists Simulations Working Group.

This study is funded by the Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, grant number: 2R44HD084151-02 awarded to Miyazaki Enterprises.

Presented at the Society of Gynecologic Surgeons' 45th Annual Scientific Meeting, March 31–April 3, 2019, Tucson, Arizona.

The opinions or assertions contained herein are the private views of the author and are not to be construed as the official or as the reflecting views of the Uniformed Services University or the Department of Defense.

The authors thank Cara Olsen, MS, DrPH, and Sorana Raiciulescu, MS, for statistical analysis, Eliza Burr, BA, for manuscript review, Noah Miyazaki BS and Jesse Ykimoff, MD, for data collection support, and Edgar LeClaire, MD, MS, for assistance with data interpretation and manuscript readability.

Each author has confirmed compliance with the journal's requirements for authorship.

Corresponding author: Chi Chung Grace Chen, MD, Department of Gynecology and Obstetrics, Johns Hopkins University School of Medicine, Baltimore, MD; email: cchen127@jhmi.edu.

Financial Disclosure

Douglas Miyazaki is the founder and chief executive officer of Miyazaki Enterprises (Winston-Salem, North Carolina) and speaker for Coloplast (Minneapolis, Minnesota) and Boston Scientific (Marlborough, Massachusetts). Catherine Matthews is a consultant for and has received grant funding from Boston Scientific (Marlborough, Massachusetts), has received grant funding from Neomedic (Terrasa, Barcelona), and has been an expert witness for Johnson & Johnson (New Brunswick, New Jersey). Mikio A. Nihira is a legal expert for Ethicon (Somerville, New Jersey), a consultant for Boston Scientific (Marlborough, Massachusetts), Pacira (Parsippany, New Jersey), and Hologic (Marlborough, Massachusetts), and has received honoraria from the Surgical Pain Consortium, ACOG, and ABOG. Bruce D. Anderson is shareholder of Miyazaki Enterprises (Winston-Salem, North Carolina) and owner of Augmented Reality Systems, Inc. (Windham, New Hampshire). He is a part-time employee of Miyazaki Enterprises where he worked to obtain the grant funding for this study and completed the work described in this study, specifically the study design and implementation. Augmented Reality Systems, Inc., was not involved in this study and has no financial ties with Miyazaki Enterprises. The other authors did not report any potential conflicts of interest.

© 2020 The Author(s). Published by Wolters Kluwer Health, Inc. This is an open-access article distributed under the terms of the Creative Commons Attribution-Non Commercial-No Derivatives License 4.0 (CCBY-NC-ND), where it is permissible to download and share the work provided it is properly cited. The work cannot be changed in any way or used commercially without permission from the journal.

ISSN: 0029-7844/20

Skills Index cutoff score of 27 was found to most accurately (area under the curve 0.951, 95% CI 0.917–0.983) differentiate competent from noncompetent surgeons.

CONCLUSION: We demonstrated validity evidence for using a high-fidelity vaginal surgery model with the modified Vaginal Surgical Skills Index or global scale score to assess vaginal hysterectomy skills.

(*Obstet Gynecol* 2020;136:942–9)

DOI: 10.1097/AOG.0000000000004085

Graduate medical education leaders have increasingly emphasized the importance of objective assessment tools for certification of competency in surgery, further underscoring the need for valid simulation models necessary for adequate training and assessment.^{1,2} Additionally, the need for physicians to maintain skills competency throughout their careers has been highlighted in recent years.³ Many medical boards are requiring demonstration of skills competency for board certification and maintenance of certification. The American Board of Obstetrics and Gynecology started requiring participation in a simulation course as an option to fulfill Part IV of the Maintenance of Certification “Improvement in Medical Practice” in 2016 and more recently added completion of Fundamentals of Laparoscopic Surgery as a prerequisite requirement for board certification beginning in May 2020.⁴

Simulation models for teaching and assessing surgical performance of entire procedures have the greatest potential effect when the model more closely approximates reality. The few models that have been described and studied to simulate performance of vaginal hysterectomy are low-fidelity models that do not allow for simulation of the entire procedure. The Miya Model (Fig. 1) is a high-fidelity female pelvic anatomy model developed for gynecologic skills training (including basic pelvic examination). In-office procedures and surgical techniques that can be performed include dilation and curettage, diagnostic and operative hysteroscopy, diagnostic cystoscopy, vaginal hysterectomy, slings, anterior and posterior colporrhaphies, and uterosacral ligament and sacrospinous ligament suspensions. A previous pilot study collected preliminary evidence that the model is of sufficient fidelity to be used in conjunction with global rating scales (including a modified version of a previously validated scale for evaluating vaginal surgical skills, the Vaginal Surgical Skills Index⁵), as a surgical performance assessment tool to measure vaginal hysterectomy skill among novice and expert surgeons.⁶

Before a simulation-based assessment can be considered for high-stakes end points such as an objective

measure for certification of competence, it is critical to first establish more evidence in the validation process. Surgical validation studies have traditionally been structured around classical validity frameworks demonstrating content, criterion and construct; however, contemporary validation frameworks, such as Messick’s validation standards,⁷ now view validation as a process with different levels of evidence supporting the intended construct the assessment is measuring (eg, vaginal hysterectomy skills) and the decisions based on this assessment. We chose Messick’s framework because it is advocated by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education in the 1999 *Standards for Educational Testing* and reaffirmed in the 2014 *Standards*.^{8,9} In this framework, five sources of evidence—content, response process, internal structure, relationship with other variables, and consequences—are used to support or refute the hypothesis of validity of an assessment method to measure a specific construct (eg, vaginal hysterectomy skills).^{10,11} The aim of this study was to use the Messick validation framework to evaluate sources of validity evidence for a composite, simulation-based evaluation consisting of global rating scales adapted for use with a high-fidelity vaginal surgery model to assess vaginal hysterectomy surgical skills.

METHODS

The Miya Model is a synthetic representation of the female pelvis featuring the vagina, vulva and pelvic viscera including an inflatable bladder, uterus, adnexa, rectum, and a pressurized vascular system (Fig. 1, details of the model have been previously published⁶). All the essential steps of a vaginal hysterectomy can be simulated on the model, including entry into the anterior and posterior cul-de-sac, ligation of all ligamentous and vascular pedicles, and vaginal cuff closure and suspension. In contrast to biological models such as cadavers, the features of this simulator are uniform, ensuring consistent training and assessment experiences for all learners; however, the simulator does not allow for the practice and evaluation of other aspects of the surgical process such as placement of the patient in lithotomy position. As each hysterectomy is performed, the remaining pedicles can be removed from the pelvic frame and the entire uterus-broad ligament unit replaced for the next hysterectomy. The bladder can be filled with fluid to allow for catheterization and performance of a diagnostic cystoscopy as well as to simulate bladder injury with leakage of fluid. The entire model costs \$6,700, and the disposable and nonreusable structures (vagina, uterus and

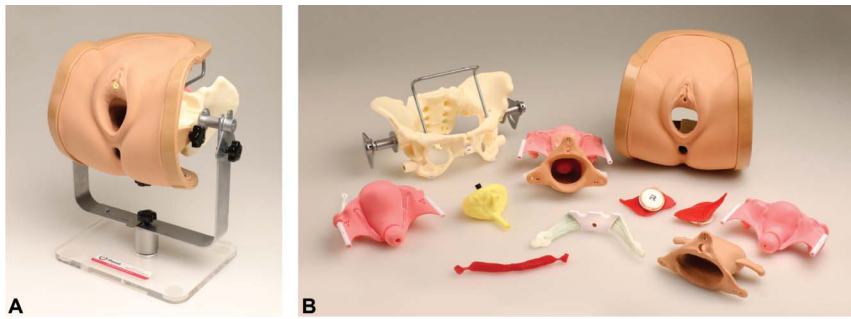


Fig. 1. The Miya Model is an injection-molded bony gynecoid pelvis that rotates 360° and is attached to a support bracket on a stand. Front and exterior view (A) and individual components (B). ©Miyazaki Enterprises. Used with permission.

Chen. Validity for a Vaginal Hysterectomy Simulation Model. Obstet Gynecol 2020.

adnexa) cost \$440, with a newer, more simplified version of uterus and vagina costing \$218 (see Appendix 1, available online at <http://links.lww.com/AOG/C44>). This study used the original, more complete, disposable and nonreusable structures.

Physicians with varied levels of training and experience were video recorded performing a vaginal hysterectomy using this model. Participants included obstetrics and gynecology residents and faculty from five different institutions (Johns Hopkins University, Mayo Clinic-Florida, Uniformed Services University of the Health Sciences, University of Oklahoma, Wake Forest University). Advertisements for trainee and faculty volunteers were administered through the residency directors at each institution to obtain a convenience sample of study participants. Specific instructions were given to participants regarding the anonymous nature of the video recordings and, for the trainees, assurance that performance on the simulator would have no effect on their residency evaluation. The research protocol was reviewed by each IRB participating in the study and found to be exempt under the provision of 32 CFR 219.101(b).¹ Before and on the day of performance, surgeons were instructed that they should perform a vaginal hysterectomy according to the procedural steps outlined by the American College of Obstetricians and Gynecologists, and they were also oriented to the simulation model.¹² Each participant was given a maximum of 60 minutes and had one nonskilled surgical assistant to complete the procedure. As there were no expert assistants, no feedback was offered during the procedure and no debrief happened after procedural completion.

Each video recorded surgical performance was independently assessed by two or three expert vaginal surgeons (coauthors on this study except B.D.A. and D.M., who have financial ties to this simulation model), using a modified 10-item Vaginal Surgical Skills Index global rating scale (each item in the Vaginal Surgical Skills Index was scored on a 4-point

anchored Likert-type scale (0–4), with higher scores indicating better performance), a one-item global scale of overall operative performance (nonanchored Likert-type scale^{1–7} with higher scores indicating better performance), and a pass–fail criterion, which was assessed as a separate item from the modified Vaginal Surgical Skills Index and global scale score. The original Vaginal Surgical Skills Index is a global rating scale with 13 items developed and validated to specifically evaluate vaginal surgical skills in live surgery (Appendix 2, available online at <http://links.lww.com/AOG/C44>).⁵ Certain metrics on the Vaginal Surgical Skills Index (initial inspection, electro-surgery, hemostasis) that were not able to be scored using this model were eliminated from use in this study. Specifically, “Initial inspection” was not evaluated because it was not often possible to determine from the video recording how well an initial inspection was being performed. It was not possible to use electrocautery on this model. The “hemostasis” metric could also not be assessed as the material used to make vascular tubing for the vascular system in this version of the model was too stiff resulting in incomplete ligation of the vessels even with proper technique. This issue has since been rectified on current versions of this model.

All evaluating surgeons were blinded to participant identities and levels of experience. Evaluating surgeons were all fellowship-trained/have received additional training, board-certified, or both, in either female pelvic medicine and reconstructive surgery, minimally invasive gynecologic surgery, or gynecologic oncology. In addition to performing and teaching vaginal hysterectomy on this simulator on other occasions, all evaluating surgeons had performed 120–900 lifetime vaginal hysterectomies, and have an average of 20 years of practice (range 9–30) with roles in resident and Fellow education, and expertise in simulation education and research. The experts in vaginal hysterectomy independently evaluated the recorded performances. To reduce the

potential for assessment bias, experts did not assess performances from their home institution. To standardize scoring, reviewers were trained to use the modified Vaginal Surgical Skills Index, global scale, and pass-fail metric on two different recorded vaginal hysterectomy performances on this model from the previous pilot study.

The primary outcome was assessment of the participants' surgical skills on this model using the modified Vaginal Surgical Skills Index and global scale score. A secondary outcome was to establish a modified Vaginal Surgical Skills Index cutoff score which differentiates participants that are competent at performing vaginal hysterectomies on this model from participants who are not competent. Additionally, participants evaluated the model as an assessment and training tool for vaginal hysterectomy using a postsimulation survey. Data were collected on demographic information and on surgical experience. Continuous variables were described as median and interquartile range and categorical variables were described using frequency and percent. Differences in median total scores for the modified Vaginal Surgical Skills Index and the global scale were assessed between experience levels using the Kruskal Wallis H test and the Mann-Whitney *U* test. Associations between continuous variables (eg, modified Vaginal Surgical Skills Index score, training and experience) were assessed using Pearson correlation coefficients and agreement for categorical variables was described using kappa statistics. Cronbach's alpha was used to describe internal consistency of the modified Vaginal Surgical Skills Index and the intraclass correlation coefficient was used to describe interrater reliability. A receiver operating characteristic (ROC) curve was used to describe the predictive accuracy of the modified Vaginal Surgical Skills Index score with the pass-fail criterion as the "gold standard." The points on the ROC curve guided the selection of a cutoff for a passing score on the modified Vaginal Surgical Skills Index that agrees well with the gold standard. Comparisons for subscales of the modified Vaginal Surgical Skills Index between competent ("pass") and noncompetent ("fail") participants were evaluated using the Wilcoxon rank-sum test. Data were analyzed using SPSS 24. A 5% two-sided significance level was used for all statistical tests.

This study is reported in accordance with the Simulation-Based Research recommendations which are specific extensions to the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines and includes the following simulation-specific elements: participant orientation,

simulator type, simulation environment, simulation event and scenario, instructional design, feedback and debriefing.¹³ We also evaluated our study methodology using the MERSQI (Median Medical Education Research Study Quality Instrument), which was developed to score the quality of medical education research using the following criteria: study design, sampling, type of data, validity of evaluation instrument, data analysis, outcomes. Each criterion is scored out of three with a potential range of five to 18 for the entire instrument.¹⁴

RESULTS

A total of 56 participants enrolled in the study with each participant performing one vaginal hysterectomy on the model. Two video-recorded performances were excluded owing to technical filming issues. Four of the participants did not complete the form on demographics and surgical experiences and so their performances were also excluded from most of the analysis unless otherwise noted, leaving a total of 50 performances (39 residents and 11 faculty). As expected, novice surgeons were significantly younger and had performed fewer hysterectomies than more experienced surgeons (Table 1).

The median total modified Vaginal Surgical Skills Index score was 25.3 (interquartile range 17.5), and the median global scale score was 4.0 (interquartile range 2.5) with the more experienced participants receiving significantly higher scores (Table 2). Experience level correlated strongly with both composite modified Vaginal Surgical Skills Index score and global scale score (Pearson $r=0.81$, $P<.001$ and Pearson $r=0.74$, $P<.001$, respectively). The number of hysterectomies performed moderately correlated with the composite modified Vaginal Surgical Skills Index scores and global scale scores (Pearson $r=0.55$, $P<.001$, and Pearson $r=0.58$, $P<.001$, respectively). The internal consistency of the modified Vaginal Surgical Skills Index employed with the model was excellent (Cronbach's alpha=0.97). Interrater reliability of the modified Vaginal Surgical Skills Index, as measured by intraclass correlation coefficient, was moderate-to-high and ranged from 0.49 to 0.95 depending on the exact pair of evaluators that were being compared. Similarly, interrater reliability of global scale scores ranged from 0.50 to 0.87. Total modified Vaginal Surgical Skills Index scores correlate highly with global scale scores (Pearson $r=0.92$, $P<.001$). The corresponding kappa statistic assessing agreement between raters was 0.76 ($P<.001$).

To determine the modified Vaginal Surgical Skills Index cutoff score that separated competent

Table 1. Participant Demographics and Experiences

Characteristic	Novice Surgeons (PGY-1-3) (n=33)	Experienced Surgeons (PGY-4 and Higher) (n=17)	P
Age category (y)			<.001
Younger than 30	26 (81)	2 (14)	
31-40	6 (19)	4 (29)	
41-49	0	3 (21)	
50 or older	0	5 (36)	
Gender			.14
Male	7 (21)	7 (41)	
Female	26 (79)	10 (59)	
Years of experience post-medical school	2 (1-2.5)	11 (5.5-19.5)	<.001
No. of all hysterectomies performed or supervised/y	4.5 (0-11.25)	50 (30-71.25)	<.001
No. of vaginal hysterectomies performed or supervised/y	2.3 (0-3)	16.4 (4-25)	<.001
No. of lifetime vaginal hysterectomies performed	1.5 (0-6.75)	200 (32.5-200)	<.001

PGY, postgraduate year.

Data are n (%) or median (range) unless otherwise specified.

from noncompetent surgeons, we were able to use performance data from the 50 study participants with complete data as well as the four study participants who did not complete the demographics and surgical experience form as we did not use their self-reported experience levels as a variable. Using a ROC and the pass-fail criterion as the gold standard, we found a composite modified Vaginal Surgical Skills Index score cutoff of 27 as the lowest passing score maximizing the sensitivity (80.3% [95% CI 69-89%]) and specificity (96.7% [95% CI 88-99.5%]) with an area under the curve of 0.951 (95% CI 0.917-0.983) (Appendix 3, available online at <http://links.lww.com/AOG/C44>). The cutoff was selected to minimize the false-positive rate (ie, 1-specificity, the percentage of participants who would have failed but passed) while maximizing the true-positive rate (ie, sensitivity, the percentage of participants who would have passed and did indeed pass). A total of 54 videos were included, 18 of which were evaluated by reviewers from group three, resulting in 126 total videos reviewed. Of 126 total expert assessments of the 54 videos, two failed under the current standard but had a composite score greater than 27, and 13 passed under the current standard but had a composite score less than 27, for 88% overall agreement.

Using the modified Vaginal Surgical Skills Index cutoff score of 27 to separate the “competent” from “noncompetent” study participants, competent participants scored significantly better on all metrics within the modified Vaginal Surgical Skills Index (Vaginal Surgical Skills Index subscales) compared with noncompetent participants (Table 3). The competent median modified Vaginal Surgical Skills Index subscale scores for most metrics was 3-3.5.

Most study participants felt that the high-fidelity vaginal surgery model use in this study was “somewhat effective,” “effective,” or “highly effective” as a simulation tool for vaginal hysterectomy training and assessment (87%, 75%, respectively). Most participants also felt that this model was a “somewhat effective,” “effective,” or “highly effective” addition to the traditional Halstedian teaching paradigm and for training before live surgery for patient safety (87%, 84%, respectively). However, only 59% of participants thought that this model was “somewhat effective,” “effective,” or “highly effective” at simulating vaginal hysterectomy in a live patient.

Using the MERSQI domains (study design, sampling, type of data, validity of evaluation instrument, data analysis, outcomes), our study scored a 14.5 out of 18. We achieved the maximum score in all

Table 2. Modified Vaginal Surgical Skills Index and Global Scale Scores By Training Level

Training Level	n	VSSI*	GS Score [†]
PGY-1	11	12.0 (6.5)	1.5 (1.5)
PGY-2 or -3	22	23.5 (9.1)	3.6 (1.5)
PGY-4 or Fellow	6	30.3 (12.9)	4.6 (2.6)
Attending	11	36.0 (6)	5.5 (1.3)

VSSI, modified Vaginal Surgical Skills Index; GS, global scale; PGY, postgraduate year.

Data are median (interquartile range).

* Range 0-40; higher score indicates better performance.

[†] One-item scale of overall performance, range 1-7; higher score indicates better performance.

Pairwise comparisons showed significant results for both the modified VSSI and the GS for PGY-1 vs PGY-4 or Fellow ($P=.013$, $.032$), PGY-1 vs attending ($P<.001$), and PGY-2 vs attending ($P<.001$).

Table 3. Modified Vaginal Surgical Skills Index Subscale Scores in Noncompetent Compared With Competent Surgeons

VSSI Subscale*	Noncompetent Surgeons (n=29) [†]	Competent Surgeons (n=25) [†]
Incision	2.0 (1.3)	3.5 (0.5)
Main visibility	1.5 (1.0)	3.0 (0.8)
Use of assistants	1.5 (1.0)	3.0 (0.8)
Knowledge of instruments	2.5 (1.5)	3.5 (0.5)
Tissue and instrument handling	1.5 (0.8)	3.0 (0.5)
Knot tying	2.0 (1.5)	3.5 (1.0)
Procedure completion	2.0 (1.0)	3.0 (1.0)
Time and motion	1.0 (1.0)	3.0 (1.0)
Flow of operation and forward planning	2.0 (1.3)	3.5 (0.8)
Knowledge of procedure	2.0 (1.5)	3.5 (1.0)
Modified total score	19.0 (10.0)	32.5 (5.8)

VSSI, Vaginal Surgical Skills Index.

Data are median (interquartile range).

* Subscales range 0–4, modified composite score range 0–40; higher score indicates better performance

[†] “Noncompetent” surgeons scored less than 27 on the modified VSSI, and “competent” surgeons scored 27 or higher, with $P < .001$ for all subscale score comparisons.

domains with points lacking in “study design” as this is most accurately described as a cross-sectional study and in “outcomes” as we did not gather data on changes in skills, behaviors or patient outcomes.

DISCUSSION

We have demonstrated validity evidence (Table 4) for a composite high-fidelity vaginal surgery model with either the modified Vaginal Surgical Skills Index or global scale to assess vaginal hysterectomy skills. Specifically, we found supportive “content” and “response process” validity evidence based on participant (novice and expert) and proctor (evaluator) surveys and moderate to high interrater reliability.⁶ Furthermore, we demonstrated support for “internal structure” evidence of validity including high internal consistency. These results are consistent with the previous pilot study using this simulation model and the modified Vaginal Surgical Skills Index as well as other

studies using both the Vaginal Surgical Skills Index and global scale in live surgery.^{5,6}

Expert surgeons were more skillful at performing vaginal hysterectomies than novice surgeons. Although expert-novice comparisons are the most frequently reported evidence used to support “relations with other variables” validity in the surgical simulation literature,¹⁵ we further demonstrated that performance on this model, as assessed using the modified Vaginal Surgical Skills Index and global scale, were moderately to highly correlated with the actual experiences of the surgeons (postgraduate years and number of vaginal hysterectomies performed). This finding is notable as it supports that this assessment method is sensitive in addition to being discriminatory. Importantly, we started the process of establishing “consequential” evidence of validity by determining cutoff scores to differentiate “competent” from “noncompetent” surgeons. In surgical simulation, consequential evidence is

Table 4. Validity Evidence

Source	Supportive Data in This Study	Other Types of Supportive Data
Content	Participant postsimulation survey	Model-development process (details not published)
Response process	Participant postsimulation survey, interrater reliability	Qualitative measures of decision making involved in performing surgery on simulator
Internal structure	Internal consistency	Factor rating method to simplify simulation model and global rating scale
Relationship with other variables	Correlation with training level and surgical experience	Correlation with operating room performance, other similar assessments
Consequences	Modified VSSI cutoff score	Implications of using the cutoff scores, eg, improved training outcomes, decreased rates of avoidable surgical error

concerned with educational and clinical effects, both intended and unintended, that result from score-based judgments. Establishing these competency cutoffs will support the development of assessment-based action implications (eg, high-stakes evaluation of surgical competency by regulatory bodies) that can be investigated for effects on training and clinical outcomes.

Many low-fidelity, “home-made” vaginal hysterectomy simulation models have previously been published in the literature.^{16–21} Although these simulation models are inexpensive to build, quantitative evidence supporting the validity of using them as simulation-based assessments in high- or low-stakes environments is limited, with most of these studies focused on participant performance differences between expert and novice surgeons. Additionally, low-fidelity synthetic surgical simulation models are often criticized for their lack of lifelike tissue appearance and feel and inability to simulate the entire procedure, which may limit their ability to simulate live surgery and their role in formative and summative assessments.^{22,23}

The strengths of this study included the use of expert surgeons to evaluate each surgical performance independently while blinded to participant identity and experience. The expert reviewers had no commercial ties to this model. Each reviewer evaluated recordings of surgical performance using previously studied global rating scales including a modified version of one specifically designed to assess vaginal surgical skills (Vaginal Surgical Skills Index).⁵ Importantly, the methodology used in this study followed a rigorous framework for validation that is widely championed by educational researchers and psychometricians, which is not commonly the case in the existing literature. The authors of a 2014 systematic review of simulation-based assessments in health care found that the majority of existing studies used “an outdated or incomplete framework to interpret validity data, if they used any framework at all.”¹¹ Additionally, using the MERSQI, our study had a total score of 14.5 with a maximum score in the validity domain, which is higher than what the authors of the MERSQI found in their review of published medical education research studies (mean total MERSQI score of 9.96 [SD 2.34, range 5–16] with the lowest scores found in the validity domain).¹⁴

A limitation of the present study was that we did not assess for test-retest reliability, as the study participants only performed one vaginal hysterectomy on the model, and we also did not evaluate the model as a teaching tool. Simulation training and

assessment are critical for any complex procedure that requires repetitive practice for skills acquisition, including vaginal hysterectomy. Ideally, in an environment that does not compromise patient safety, simulation models are most effective if incorporated as part of a residency surgical curriculum. For example, a simulation model can be used to teach and assess trainee vaginal hysterectomy skill after didactics and basic vaginal surgical skills training before allowing the trainee to be the primary surgeon in the operating room.

Under the current validation framework, validity is conceptualized as a process by which evidence from various sources is collected to ensure that assessment-based interpretations and decisions are scientifically supported and consequentially justified. In this study, we have preliminarily established cut-scores for “competent” and “noncompetent” performance, but longitudinal studies are required to determine whether such score-based judgments of competency are justified. In particular, although use of the Vaginal Surgical Skills Index has been investigated in live surgery,⁵ apart from its use in this study and in the pilot study,⁶ the modified Vaginal Surgical Skills Index and the pass-fail metric have not been studied elsewhere. Specific examples of consequential validity evidence include improved trainee performance measures in the operating room, decreases in the rates of avoidable complications, as well as improvements in feedback quality within the faculty-resident training dialogue. Obtaining data on such consequences is the ultimate goal of validation research, because they confirm there is a demonstrable benefit and, importantly, an absence of harm, educational or otherwise, when decisions are made based on the assessment score. Despite this, consequential evidence of validity is rarely sought or reported in surgical simulation research (5–20%).^{24,25}

Validation of a high-fidelity vaginal surgery model with a global rating scale (modified Vaginal Surgical Skills Index, global scale) addresses an important gap in vaginal surgery skills training and assessment. Importantly, we increased the evidentiary support for content and response process, internal structure, and relationship with other variables sources of validity for this model. We initiated the evidentiary process in “consequential” validation, which will be built on in future studies with the ultimate aim of establishing this assessment method as a tool for surgeon educators in determining surgical competence and for leadership in regulatory bodies in high-stakes endpoints such as procedural credentialing and maintenance of certification.

REFERENCES

1. The obstetrics and gynecology milestone project. *J Grad Med Educ* 2014;6(suppl 1):129–43.
2. Nasca TJ, Philibert I, Brigham T, Flynn TC. The next GME accreditation system—rationale and benefits. *N Engl J Med* 2012;366:1051–56.
3. Santen SA, Hemphill RR, Pusic M. The responsibility of physicians to maintain competency. *JAMA* 2020;323:117–8.
4. Fundamentals of Laparoscopic Surgery. ABOG announces new eligibility requirement for board certification. Available at: <https://www.flsprogram.org/news/abog-announces-new-eligibility-requirement-board-certification/>. Retrieved February 5, 2018.
5. Chen CC, Korn A, Klingele C, Barber MD, Paraiso MF, Walters MD, et al. Objective assessment of vaginal surgical skills. *Am J Obstet Gynecol* 2010;203:79.e1–8.
6. Miyazaki D, Matthews CA, Kia MV, El Haraki AS, Miyazaki N, Chen CCG. Validation of an educational simulation model for vaginal hysterectomy training: a pilot study. *Int Urogynecol J* 2019;30:1329–36.
7. Messick S. Validity. In: Linn R, editor. *Educational measurement*. 3rd ed. New York, NY,: American Council on Education and Macmillan; 1989. p. 13–103.
8. American Educational Research Association, American Psychological Association, National Council on Measurement in Education, and Joint Committee on Standards for Educational and Psychological Testing. *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association; 1999.
9. American Educational Research Association, American Psychological Association, National Council on Measurement in Education, and Joint Committee on Standards for Educational and Psychological Testing. *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association; 2014.
10. Cook DA, Hatala R. Validation of educational assessments: a primer for simulation and beyond. *Adv Simul* 2016;1:31.
11. Cook DA, Zendejas B, Hamstra SJ, Hatala R, Brydges R. What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. *Adv Health Sci Educ Theory Pract* 2014;19:233–50.
12. American College of Obstetricians and Gynecologists. Simulation: vaginal hysterectomy. Available at: <https://www.acog.org/education-and-events/simulations/scog008/simulation>. Retrieved June 21, 2020.
13. Cheng A, Kessler D, Mackinnon R, Chang TP, Nadkarni VM, Hunt EA, et al. Reporting guidelines for health care simulation research: extensions to the CONSORT and STROBE statements. *Simul Healthc* 2016;11:238–48.
14. Reed DA, Cook DA, Beckman TJ, Levine RB, Kern DE, Wright SM. Association between funding and quality of published medical education research. *JAMA* 2007;298:1002–9.
15. Cook DA. Much ado about differences: why expert-novice comparisons add little to the validity argument. *Adv Health Sci Educ Theor Pract* 2015;20:829–34.
16. Malacarne DR, Escobar CM, Lam CJ, Ferrante KL, Szyld D, Lerner VT. Teaching vaginal hysterectomy via simulation: creation and validation of the objective skills assessment tool for simulated vaginal hysterectomy on a task trainer and performance among different levels of trainees. *Female Pelvic Med Reconstr Surg* 2019;25:298–304.
17. Vaughan MH, Kim-Fine S, Hullfish KL, Smith TM, Siddiqui NY, Trowbridge ER. Validation of the simulated vaginal hysterectomy trainer. *J Minim Invasive Gynecol* 2018;25:1101–6.
18. Geoffrion R, Suen MW, Koenig NA, Yong P, Brennan E, Mehra N, et al. Teaching vaginal surgery to junior residents: initial validation of 3 novel procedure-specific low-fidelity models. *J Surg Educ* 2016;73:157–61.
19. Greer JA, Segal S, Salva CR, Arya LA. Development and validation of simulation training for vaginal hysterectomy. *J Minim Invasive Gynecol* 2014;21:74–82.
20. Barrier BF, Thompson AB, McCullough MW, Occhino JA. A novel and inexpensive vaginal hysterectomy simulator. *Simul Healthc* 2012;7:374–9.
21. Balgobin S, Owens D, Florian-Rodriguez M, Wai C, McCord EH, Hamid C. Vaginal hysterectomy suturing skills training model and curriculum. *Obstet Gynecol* 2019;134:553–8.
22. Aggarwal R, Darzi A. Technical-skills training in the 21st century. *N Engl J Med* 2006;355:2695–6.
23. Nwamaka M-A, Onwugbenu M. Vaginal hysterectomy simulation: adjunct training. *Obstet Gynecol* 2015;126:53–4S.
24. Cook DA, Lineberry M. Consequences validity evidence: evaluating the impact of educational assessments. *Acad Med* 2016;91:785–95.
25. Zendejas B, Ruparel RK, Cook DA. Validity evidence for the Fundamentals of Laparoscopic Surgery (FLS) program as an assessment tool: a systematic review. *Surg Endosc* 2016;30:512–20.

PEER REVIEW HISTORY

Received April 20, 2020. Received in revised form June 25, 2020. Accepted July 2, 2020. Peer reviews and author correspondence are available at <http://links.lww.com/AOG/C45>.