


Evaluating validity evidence for 2 instruments developed to assess students' surgical skills in a simulated environment

Robin M. Farrell BS, DVM, PGDipMED¹  | Gregory E. Gilbert EdD, MSPH, PStat^{2,3} |
 Larry Betance BS, DVM⁴ | Jennifer Huck DVM, DACVS-SA⁵ |
 Julie A. Hunt DVM, MS⁶ | James Dundas DVM, DACVS-SA⁷ |
 Eric Pope DVM, MS, DACVS⁴

¹School of Veterinary Medicine,
University College Dublin, Dublin,
Ireland

²SigmaStats Consulting, LLC, Charleston,
South Carolina, USA

³Biostatistics and Medical Writing, Real
World Evidence Strategy & Analytics,
ICON Commercialization & Outcomes
Services, North Wales, Pennsylvania, USA

⁴School of Veterinary Medicine, Ross
University, Basseterre, Saint Kitts and
Nevis

⁵School of Veterinary Medicine,
University of Pennsylvania, Philadelphia,
Pennsylvania, USA

⁶College of Veterinary Medicine, Lincoln
Memorial University, Harrogate,
Tennessee, USA

⁷Atlantic Veterinary College,
Charlottetown, Prince Edward Island,
Canada

Correspondence

Robin Farrell, UCD School of Veterinary
Medicine, Room 040, UCD, Belfield,
Dublin 4, Ireland.
Email: robin.farrell@ucd.ie

Funding information

Equipment and consumables were
supported by a Ross University Research
and Innovation in Veterinary Medical
Education intramural grant. The authors
received no stipend or salary support for
this research.

Abstract

Objective: To gather and evaluate validity evidence in the form of content and reliability of scores produced by 2 surgical skills assessment instruments, 1) a checklist, and 2) a modified form of the Objective Structured Assessment of Technical Skills (OSATS) global rating scale (GRS).

Study design: Prospective randomized blinded study.

Sample population: Veterinary surgical skills educators ($n = 10$) evaluated content validity. Scores from students in their third preclinical year of veterinary school ($n = 16$) were used to assess reliability.

Methods: Content validity was assessed using Lawshe's method to calculate the Content Validity Index (CVI) for the checklist and modified OSATS GRS. The importance and relevance of each item was determined in relation to skills needed to successfully perform supervised surgical procedures. The reliability of scores produced by both instruments was determined using generalizability (G) theory.

Results: Based on the results of the content validation, 39 of 40 checklist items were included. The 39-item checklist CVI was 0.81. One of the 6 OSATS GRS items was included. The 1-item GRS CVI was 0.80. The G-coefficients for the 40-item checklist and 6-item GRS were 0.85 and 0.79, respectively.

Conclusion: Content validity was very good for the 39-item checklist and good for the 1-item OSATS GRS. The reliability of scores from both instruments was acceptable for a moderate stakes examination.

Impact: These results provide evidence to support the use of the checklist described and a modified 1-item OSAT GRS in moderate stakes examinations when evaluating preclinical third-year veterinary students' technical surgical skills on low-fidelity models.

1 | INTRODUCTION

Veterinary graduates are expected to be competent in basic surgical skills.^{1,2} The preclinical veterinary surgical skills curriculum is continuously evolving, with educators incorporating models and new methods of clinical skills training to ensure students attain competency in core skills.^{3–18} As clinical skills training programs evolve, so do the assessment instruments used to evaluate educational interventions and students' performance on models, cadavers, and live animals. While veterinary educators have made significant progress in developing clinical skills assessments, relatively few reports including validity evidence for instruments to assess veterinary students' basic surgical skills using models have been published.^{12,19–26} The use of modified forms of the Objective Structured Assessment of Technical Skills (OSATS) global rating scale (GRS), or another validated assessment instrument adapted from medical education, has also been explored, but to date, limited validity evidence has been reported in the literature to support their use in veterinary education.^{12,14,21,23} Validity is “the degree to which evidence and theory support the interpretations of test scores entailed by the proposed uses of the tests.”²⁷

Validity evidence for assessments informs the interpretation of results and decisions that educators and curriculum committees make regarding the consequences those results have for the students and program.^{27,28} For instance, a must-pass clinical skills examination required to complete a course would be considered a high-stakes examination with significant consequences for the student and program if students did not pass or maybe even more crucial, passed despite a lack of competence. High-stakes examinations can be defined as those that either allow or prevent students from progressing in their program, such as examinations that serve as progression hurdles. Creating a strong validity argument using validity evidence to support the high-stakes nature of an examination instills confidence in both the assessor and student that the assessment and the scores produced are valid and reliable.

Objective Structured Assessment of Technical Skills, a generally accepted assessment instrument used for human surgical residents, mimics an objective structured clinical examination (OSCE), except it consists of a benchtop model simulating a procedure or task, as opposed to individual skill or part of a task.^{29,30} Learners' performance on an OSATS station is typically evaluated using a customized checklist for the task and GRS suitable for any surgical task. Hatala et al. (2015) found reasonable validity evidence supporting the use of the OSATS in a low stakes environment to provide formative feedback to physicians in surgical residency.³¹ A low stakes environment or assessment is one that does not impact students' progression through a

program and is usually formative in nature, where outcomes such as rubric scores are used to provide feedback to help students improve their performance. Literature in veterinary and medical education has reviewed and evaluated simulation-based instructional design and assessments and has suggested that better designed research projects are needed to collect data supporting the use of specific training methods, simulators, and assessments.^{32–35} To collect accurate data, validated assessment instruments must be identified or developed to facilitate larger research projects, across multiple veterinary schools that can answer questions about simulation-based teaching and assessment in veterinary education.

The aim of this study was to gather and evaluate validity evidence in the areas of content and reliability of scores produced when using a task-specific checklist assessment instrument developed at Ross University School of Veterinary Medicine (RUSVM) and an OSATS GRS assessment instrument adapted from medical education to assess preclinical third-year veterinary students' surgical technical skills performed on a low-fidelity ovariohysterectomy model.

2 | MATERIAL AND METHODS

This study was approved by the institutional review board at Ross University School of Veterinary Medicine (approval number 493) and was conducted according to the tenets of the Declaration of Helsinki.

2.1 | Context

Ross School of Veterinary Medicine (Basseterre, St. Kitts and Nevis, WI) was established in 1982. In 2008, the school began a curriculum revision, which incorporated models and simulation to enhance teaching, learning, and assessment of surgical skills. The curriculum revision expanded to include medical and professional skills and resulted in a task-based vertically integrated spiral professional and clinical skills curriculum in which students were introduced to skills training in the first year and built on those skills through exposure to simulated tasks and procedures at increasing levels of complexity.^{36–38} Students' skills development was assessed through low-stakes formative assessments during learning activities and through regularly scheduled OSCEs. Surgical skills simulation-based training was nested within the professional and clinical skills curriculum. The training program culminated with a 15-week compulsory surgery laboratory course in which students practiced basic surgical skills learned in the early curriculum, including aseptic

technique, instrument handling, ligature placement, and suturing in simulated tasks, and procedures they would encounter in general practice such as wound closure, ovariectomy, and cystotomy.

2.2 | Development of the surgical skills examination

To assess student's surgical skills gained in the simulation-based curriculum prior to students moving on to supervised live animal surgeries, a summative surgical technical skills examination was developed to be delivered within the surgical skills laboratory course.

Students were required to pass the comprehensive surgical skills examination consisting of a simulated ovariectomy (OVH) performed on a model developed at the university and evaluated in a previous study.⁷ The OVH model consisted of a wood and polyvinyl chloride (PVC) frame covered by a replaceable foam and fabric 3-layer closure pad that was rotated for reuse with a replaceable latex reproductive tract (Figure 1). The use of an OVH model for practice and assessment has subsequently been supported by research demonstrating that students benefit from presurgical skills practice on OVH models.^{3,7,39,40} The 40-item checklist used to assess the examination was developed through a collaborative process among RUSVM faculty and has gone through several revisions to improve the clarity of the criteria and feasibility of administering this time-intensive examination to a large class. Although the technical surgical skills examination has been revised based on an iterative consensus process, expert review, and student performance review, a formal evaluation of validity evidence has not previously been performed.

2.3 | Content validity evidence

A panel of expert surgical skills educators ($n = 10$) were recruited from multiple institutions according to the authors' international networks to validate the examination content. Experts were defined as veterinarians who self-reported at least 2 years' experience teaching surgical skills to veterinary students in a simulated and/or live surgical environment. Panelists received an email stating the purpose of the study and requesting their participation. If they agreed to participate, they were sent a demographic survey, data collection form, and instructions on how to perform the content review. The content review required the experts to spend approximately 1 hour to rate each item on the surgical skills checklist and modified OSAT GRS as "essential," "useful," or "not

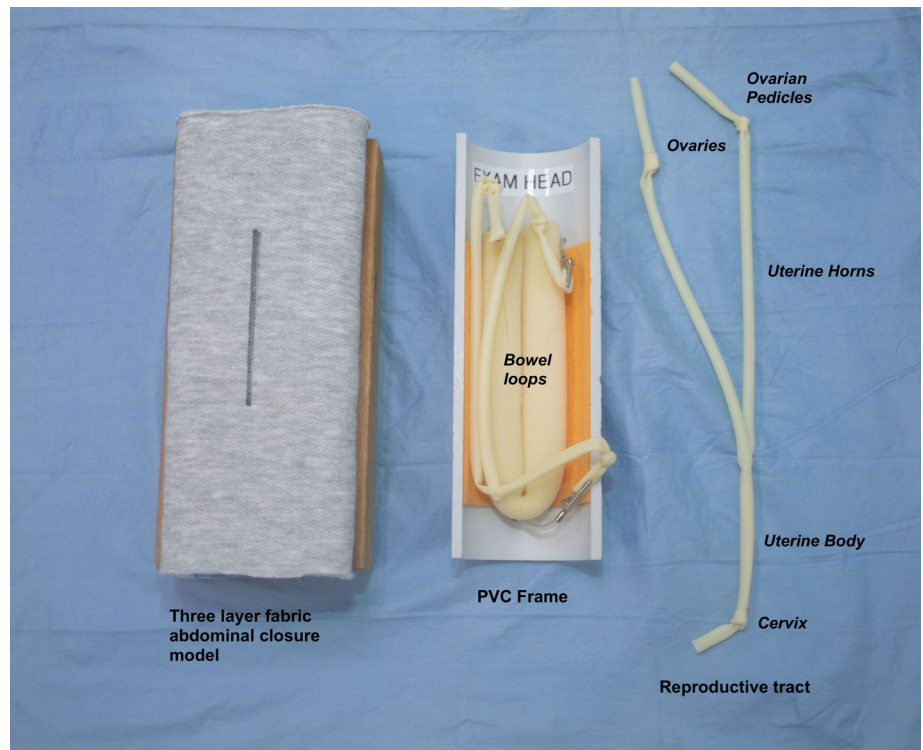
necessary." The OSATS GRS used in this study was derived from the OSATS GRS developed by Martin et al.³⁰ (Table 1). Panelists were asked to consider each item on the basis of relevance to teaching and evaluating third-year preclinical veterinary students' performance of surgical technical skills. They were not incentivized or compensated for their time.

The panelists' ratings were used to calculate the content validity ratio (CVR) for each item and the content validity index (CVI) for the overall rubric. Interrater reliability was evaluated using Gwet's AC₂. Lawshe's method using CVR and CVI is an international standard for establishing content validity, providing concrete measurements to identify rubric items for acceptance or rejection, and allowing for generalizability of findings as it requires at least 10 reviewers of varying backgrounds to participate in the content review.^{42,43} The content validity index (CVI) is a rubric-level statistic that is equal to the calculated mean CVR of all items included on the rubric.⁴⁴ The CVR values from 0 to 1 indicate that more than half of the experts considered the item(s) to be essential, and negative values mean that fewer than half of the experts considered the item(s) to be essential. Wilson et al.'s recommended CVR cut-off values were used for rubric item inclusion.^{41,42} Gwet's AC₂ was chosen over other interrater reliability statistics, such as Cohen's kappa, because Gwet's AC₂ can be used for categorical data and is more stable than kappa, being less subject to fluctuations resulting from different outcome values and marginal probability.⁴⁴⁻⁴⁶ Reliability measures for Gwet's AC₂ were interpreted using George and Mallery's guidelines stating values over 0.9 are excellent, 0.8-0.89 are good, 0.7-0.79 are acceptable, 0.6-0.69 are questionable, 0.5-0.59 are poor, and less than 0.5 are unacceptable.⁴⁷ Reviewers' data were entered into a spreadsheet. Validity and internal consistency analyses were completed using R v3.3.1 (Vienna, Austria).

2.4 | Reliability of student performance scores

Digital recording was used to allow multiple raters to rate each student's performance on the surgical skills examination. The optimal camera angle to record the examination was determined by setting up a mock surgery examination station (Figure 2). A research assistant performed a simulated examination while being digitally recorded to establish an optimal camera position and angle allowing viewers to see most of the important aspects of each skill performed without discerning the student's identity. Masking tape marked the locations of the model, instrument table, and camera to ensure

FIGURE 1 Ross ovariohysterectomy surgical simulation (ROSSie) model



standardization. A Samsung hmx-f80 camcorder positioned on a stand was used to capture the recordings.

Five raters with at least 1 year of experience teaching surgical skills and assessing student performance on the surgical skills examination were recruited from the RUSVM faculty. Raters completed a 1-hour interactive training session to review and discuss the criteria for the checklist items and global rating scales. Following the training session, raters reviewed the simulated examination digital recording to determine their ability to rate student performance on each checklist item. Raters identified 2 items that proved difficult to adequately assess from the digital recording alone; the security of the ligatures placed on the pedicles and sutures in the body wall. Based on the raters' experience assessing live examinations, it was decided ligature security on the pedicles would be determined by physically examining knot security and indentation the ligatures created on the actual cut pedicles, and body wall sutures would be evaluated by looking at apposition and suture placement through a zoomed-in view of the model provided at the end of the digital recording.

Veterinary students enrolled in the third-year surgery laboratory ($n = 136$) course participated in the surgical skills examination and were examined by an in-person rater as part of the normal delivery of the course. The students were made aware that the digital recordings of their performance would be used as part of a research study but would have no bearing on their examination

grades. Each examination was digitally recorded using the method described above. A research assistant entered the student examination roster into an Excel spreadsheet and assigned a random number to each student using the randomization function in Excel. At the start of each examination, a technician placed an index card with the student's assigned number on the model briefly for purposes of identifying the recording. At the conclusion of each examination, a technician removed the skin and subcutaneous sutures, held the model up to the camera to provide a close-up view of the body-wall sutures, and removed the ligated pedicles from the model, taping them to the numbered index card for later physical examination by the examiners. The digital recordings were uploaded to a secure password-protected network drive and labeled by number. Twenty of the 136 digital recordings were randomly selected using the randomization function in Excel and reviewed by the primary researcher for use in the generalizability study. Four of the 20 digital recordings could not be used as the recordings were incomplete.

Three months following the live assessment, raters were given access to a network drive folder holding the digital recordings for review ($n = 16$), and index cards with cut pedicles. Three months was chosen to reduce potential bias that may be introduced by raters' memory of the live assessments. A longer period of time could not be facilitated due to a risk that the absorbable sutures would degrade, compromising the ability of the

TABLE 1 Checklist items meeting Wilson's criterion for inclusion

	Item description	Item content validity ratio as assessed by expert raters
1	Ovarian pedicle #1 – clamp placement	1.00
2	Excessive force on the pedicle is avoided	0.60
3	Ligature placement	1.00
4	Absorbable suture used	0.60
5	Two secure knots placed (surgeon's knot followed by a square knot)	1.00
6	Ligatures tight	0.80
7	Appropriate spacing between ligatures (2-7 mm)	0.60
8	Pedicle severed just distal to middle forcep	0.60
9	Ovarian pedicle #2 – spacing between forceps (2-5 mm inside distance)	0.60
10	Excessive force on the pedicle is avoided	0.56
11	Ligature placement	1.00
12	Absorbable suture used	0.60
13	Two secure knots placed (miller's knot followed by a square knot)	1.00
14	Ligature tight	1.00
15	Pedicle severed just distal to middle forcep	0.60
16	Uterine body – clamp placement	1.00
17	Ligature placement	1.00
18	Absorbable suture used	0.60
19	Two secure knots placed on each ligature (surgeon's or miller's knot followed by a square knot)	1.00
20	Ligatures tight	1.00
21	Appropriate spacing between ligatures (2-7 mm)	0.60
22	Pedicle severed just distal to middle forcep	0.60
23	Body wall closure – place a minimum of 2 simple interrupted sutures ^a	1.00
24	Absorbable suture used	0.80
25	Full thickness bites of the fascia, muscle not included in suture	1.00
26	Sutures should be snug (tips of mosquito hemostats cannot easily slip underneath suture)	0.60
27	Two secure knots placed for each suture	0.80
28	Subcutaneous closure – technique of burying the knot correctly performed at beginning of pattern	1.00
29	Simple continuous pattern placed correctly (place a minimum of 3 stitches with bites 0.4-1.3 cm apart, no backhanding)	0.80
30	Only subcutaneous tissue engaged in the pattern (no fascia or skin)	0.60
31	Continuous pattern ended correctly burying the knot	1.00
32	Knots are secure	1.00
33	Skin closure – 2 secure knots placed for each suture	0.80
34	Skin edges apposed	0.80
35	Sutures not too tight (tips of hemostat can slip easily into suture loop)	0.56
36	General – holds instruments correctly; uses correct instruments	0.80
37	Refrains from grasping suture with instrument, other than tag to be discarded; does not damage suture	0.80
38	Does not engage any tissue other than the pedicle in their hemostat	1.00

TABLE 1 (Continued)

Item description	Item content validity ratio as assessed by expert raters
39 No major breaks in asepsis or multiple minor breaks in aseptic technique Note: 2 or more breaks in asepsis (not corrected properly) will result in failure of the entire examination.	1.00
Scale (entire checklist) Content Validity Index	0.81

^aSimple interrupted sutures were chosen for the body wall due to novice surgeons' potential for flaws in knot quality that may lead to dehiscence if a simple continuous pattern was used.

FIGURE 2 Examination table setup



raters to evaluate the ligatures on the cut pedicles. Each rater scored all 16 recordings using the 40-item checklist and 6-item OSATS GRS. The G-study was completed using the data collected on all 40 checklist items and six OSATS items. Reliability measures from the G-study were interpreted using George and Mallery's guidelines.⁴⁷

Generalizability (G) theory was used to assess the reliability of the scores produced.^{48–50} In this fully crossed 2-facet G theory study (participants \times raters \times items), five raters independently rated all 16 of the digitally recorded student performances, evaluating no more than four recordings a day to minimize rater fatigue. A decision study (D study) was used to determine the relationship between the number of raters and the resultant G coefficient. Decision studies help to determine how many raters must rate a single student to maintain an adequate reliability of scores. The generalizability analysis was performed using GENOVA (Iowa City, Iowa, USA).

3 | RESULTS

3.1 | Content validity evidence

Ten veterinary surgical skills educators from veterinary teaching institutions in North America, Europe, and Australia participated in the study. Thirty-nine of the 40 items on the checklist used in the G-study met Wilson's criterion for inclusion based on their CVR (Table 1). The redundancy in some items is due to the students being required to perform certain skills multiple times during the examination, such as clamping and ligating an ovarian pedicle. The CVI for the modified 39-item checklist was 0.81. Only 1 of the six items on the modified OSATS GRS, respect for tissue, met Wilson's criterion for inclusion, indicating that an inadequate number of expert reviewers deemed the other OSATS GRS items to be essential or useful. The CVI for the modified 1-item GRS was 0.8. Gwet's AC₂, a measurement for

TABLE 2 Modified OSAT global rating scale rubric

	1	2	3	4	5
Time and motion	Not efficient, many unnecessary moves	Somewhat efficient, moderate amount of unnecessary moves	Efficient time/motion but some unnecessary moves	Efficient Good economy of movement	Maximum efficiency, great economy of movement
Instrument handling	1 Novice, repeatedly makes tentative awkward moves with instruments	2 Advanced beginner, makes some tentative or awkward moves with instruments	3 Competent use of instruments although occasionally appeared stiff or awkward	4 Proficient use of instruments, fluid moves	5 Expert use of instrument, very fluid moves with instruments no awkwardness
Tissue Handling	1 Extremely rough with the tissue, repeatedly causing unnecessary trauma to the tissue	2 Moderately rough with the tissue, sometimes causing unnecessary trauma to the tissue	3 Competent tissue handling, occasionally handles it roughly	4 Proficient tissue handling, gentle use of hands and instruments	5 Expertly handled tissue with no unnecessary trauma
Knowledge of instruments	1 Frequently used the incorrect instruments for the task	2 Sometime used the incorrect instruments for the task	3 Used appropriate instruments for the task but hesitated at times	4 Used the appropriate instruments for the task	5 Obviously familiar with the instruments required
Flow of procedure and forward planning	1 Frequently stopped the procedure or hesitated to perform the next step	2 Stopped or hesitated a few times to perform the next step of the procedure	3 Demonstrated ability to progress through the task at a slow pace ^a	4 Demonstrated ability for forward planning with steady progression through the task	5 Obviously planned course of task with effortless flow from one move to the next
Knowledge of specific procedure	1 Deficient knowledge, needed specific instruction at most operative steps	2 Deficient knowledge, needed guidance at some of the operative steps	3 Knew all important aspects of the task but lacks confidence in knowledge ^b	4 Knew all important aspects of the task	5 Demonstrated familiarity with all aspects of the operation
Overall rating	1 Needs significant amount of development in basic technical skills, tissue handling and/or procedural knowledge	2 Needs moderate amount of development in basic technical skills, tissue handling and/or procedural knowledge	3 Needs minimal to moderate amount of development in basic technical skills, tissue handling and/or procedural knowledge	4 Needs minimal development in basic technical skills, tissue handling and/or procedural knowledge	5 Has mastered basic technical skills, tissue handling and has a good understanding of procedural knowledge

^aslow pace could be defined as a rate of action during parts or all of the assessment that appeared too slow for the student to meet the overall time limit placed on the assessment.

^bA lack of confidence could be inferred based on students delaying the next step of the procedure while thinking or making tentative movements.

interrater reliability was .84 (95% CI: 0.81, 0.86) for the checklist, which was good, and .77 (95% CI: 0.63, 0.92) for the GRS, which was acceptable (Table 2).

3.2 | Reliability of scores – checklist

The G study for the 40-item checklist revealed students accounted for minimal variance (5%), suggesting individual students performed similarly to one another. Raters accounted for almost no variance (0.4%), suggesting excellent interrater reliability. Minimal variance (7%) was attributable to items, suggesting individual rubric items were rated similarly. Very little variance was attributable to the student by rater interaction (0.6%), indicating the rating was fair and free of bias. Moderate variance was attributable to the student by item interaction (17%) and rater by item interaction (14%), suggesting some students and/or raters may have identified certain items as more difficult than others and vice versa. A considerable variance was attributable to student by rater by item (sri) interaction and residual or unknown factors (55%), suggesting a number of other factors not assessed in this two facet G-study contributed to the variance. The overall G-coefficient was good at 0.85 when five raters evaluated each student's performance (Table 3). The G-study was run using the original 40-item checklist and was not repeated using only the 39 items that met the CVR threshold for inclusion; however, if the study were repeated, the impact of dropping a single item would likely be minimal. The D-study results demonstrated that one rater per student would result in a G-coefficient of 0.64, which is considered to indicate questionable reliability, and two raters would generate a G-coefficient of 0.76, which is considered acceptable reliability. If 3, 4, or 5 raters were used, the G-coefficients would be 0.81, 0.83, and 0.85, respectively, which indicate good reliability.

3.3 | Reliability of scores – modified OSAT global rating scale

The G study for the OSATS GRS revealed students accounted for about one quarter (24%) of the variation in scores, suggesting student performance varied moderately. The items (4%) and raters (13%) accounted for a minimal amount of the variance. The student by rater interaction accounted for a moderate amount of variance (16%). Students scored consistently across the items as evidenced by the low percentage of variance due to the student by item interaction (0.5%). Similarly, items were ranked consistently by the raters, contributing just 8% of the variance. Thirty percent of the variance is accounted

TABLE 3 Estimated variance components and G-coefficient for the surgical skills checklist

Factor	Variance (%)	G coefficient
Students	5.4	0.85
Raters	0.4	
Item	7.4	
Students by rater	0.6	
Students by items	17.1	
Raters by item	14.2	
Students by rater by item, and residual	54.8	

TABLE 4 Estimated variance components and G-coefficient for the modified Objective Structured Assessment of Technical Skills global rating scale

Factor	Variance (%)	G coefficient
Students	24.0	0.79
Raters	13.0	
Item	4.0	
Students by rater	16.0	
Students by items	0.5	
Raters by item	8.0	
Students by rater by item, and residual	30.0	

for by the student by rater by item (sri) interaction confounded with all other sources of error. The G coefficient was .79, approaching the .80 cutoff for “good” (Table 4).

4 | DISCUSSION

We presented content validity evidence and reported the evidence of interrater reliability and generalizability of the scores produced. Each of these measures supported the use of the checklist in evaluating veterinary students' surgical skills during the third year of their preclinical studies, with surgical skills educators deeming 98% of the checklist essential and G study results demonstrating adequate interrater reliability. While the modified OSATS GRS did not have adequate content validity as assessed by the study's 10 experts, with only 1 of the six items deemed essential by the expert panels, the reliability of scores was adequate. Furthermore, the generalizability results for the OSATS GRS attributed a moderate amount of variance to students, suggesting the OSATS GRS may facilitate raters to differentiate student performance

better than the checklist which attributed minimal variance to students. This finding is similar to results reported by Read et al. which demonstrated that global rating scales in general – but not the OSATS GRS specifically – are reliable when used for scoring student performance in a clinical skills OSCE and therefore the use of a checklist in conjunction with a GRS may better differentiate student performance than a checklist alone as the GRS allows the rater to evaluate more qualitative aspects of performance.^{51,52} While the results of the content review suggest that the OSATS GRS was not suitable for assessing preclinical veterinary students' simulated surgical skills examination, outside of the item, *respect for tissue*, more research is necessary to determine at what stage of a veterinary student's education learners are experienced enough to be expected to be competent at the more qualitative OSAT GRS items of *time and motion*, *flow of procedure and forward planning*, and *knowledge of specific procedure* – items that are probably not expected of preclinical third-year students performing surgical skills on models. The OSATS GRS could be modified to a competency-based veterinary education (CBVE) assessment by defining how the 1-5 values correspond with expectations of each level of learner, similar to how milestones have been developed for some CBVE competencies.⁵³ For example, students may be expected to be at an OSATS GRS of 3 upon entering their clinical phase of training and a GRS of 4 upon graduation, with a GRS of 5 only being achieved after further post-graduation surgical experience.

Several pieces of validity evidence are necessary to determine how to interpret assessment results and set the consequences those results have for the students and program.^{27,28} Reliability evidence is a crucial piece of validity evidence for any assessment method.^{54,55} G theory is a robust measure of reliability, allowing investigators to evaluate a number of facets of variance at once, and the associated D study allows researchers to measure what impact a change will have on reliability, for example, how increasing the number of raters will impact the reliability. Therefore, a G study is useful for planning improvements to existing assessments and the way in which the scores are used.⁵⁶ The G study results indicated that both the checklist and global rating scale produced scores reliable enough for moderate stakes testing.⁵⁶ To reduce the high-stakes nature of an examination there are a number of things that can be done, such as offering in-course resits or reducing the weighting of an assessment component. In this case, while the examination is a must-pass examination, students are allowed more than 1 attempt within the course and receive detailed feedback and support to help them prepare for a resit if they are unsuccessful on their first attempt.

The D-study results suggested the current 40 item checklist would require two raters to score each student's performance in order to maintain acceptable reliability for a high-stakes examination. Increasing the number of raters for the surgical skills examination may not be feasible due to faculty workload, and it is unlikely that additional rater training would substantially improve reliability given the minimal contribution of the raters to the variance. Instead, reliability values may be further improved by investigating and standardizing subtle differences in students' examination environment and experience which were not facets included in this G study yet contributed to the residual (sri) variance.

The minimal variance attributed to student performance on the checklist scores may have been due to the small random sample of students performing similarly well, or it may reflect the overall success of the students' extended clinical skills training program at building surgical skills. Surgical skills are learned through deliberate practice,^{57–59} which may be best delivered spaced out over a longer period of time to facilitate improved retention.^{60–63} Clinical skills programs allowing students to participate in regular surgical skills practice with feedback from instructors over a period of weeks or months, as the RUSVM skills training is, are most likely to see students demonstrate consistent skills gains on assessments.

The generalizability findings also issued support for the reliability of scoring surgical skills examinations via digital recordings. This is important to consider when it is not possible to bring students together for in-person assessment, as with the recent COVID-19 pandemic, or when an inadequate number of raters are available for real-time assessment of students. Assessing digital recordings can take more time on the part of the raters as compared with live assessment of student performance.⁶⁴ Raters in this study reported they spent a considerable amount of time rating the videos and welcomed the prescribed break after rating four recordings. Similar findings were reported by Tan et al. (2020) in a study evaluating rating of digitally recorded OSCE stations.⁶⁵ Digitally recorded evaluations of surgical skills have also been assessed for evidence of reliability in other veterinary studies,^{20,23,26,52,66} so this remains a feasible option. While the digital recordings in this study were used solely to collect data for the reliability study, digital recordings can be a powerful tool to provide feedback to students on their performance to help them enhance their proficiency and will be considered for this purpose in the future.

The results of this study supported and enhanced use of the comprehensive surgical skills examination on an OVH model in the local context and provided some validity evidence to support use of the checklist instrument in other veterinary programs. Relatively few rubrics with

evidence of validity for assessing veterinary students' surgical skills have been published; however, these rubrics exist for live canine ovariohysterectomy,⁷ simulated canine ovariohysterectomy,²⁶ live canine castration,²⁰ simulated canine castration,²⁰ and celiotomy closure in a canine cadaver.²⁵ The previously existing simulated canine ovariohysterectomy rubric was an operative component rating scale, a task-specific rubric requiring raters to score each step of the procedure on a 0-3 point scale for a total of 102 points.²⁶ Our study collected validation evidence for a dichotomous checklist having 39 points (Table 1), which may be easier for a rater to use in a busy teaching environment.

This study had several limitations. Although the content review panel included experts from North America, Europe, and Australia, a more diverse panel with representation from other continents would have been preferable. Additionally, only 16 students' digital recordings were assessed due to technical errors and time constraints on the part of the raters. While specific guidelines on minimal sample size for generalizability studies have not been established, a minimum of 20 persons for a 1 facet design has been suggested.⁶⁷ Studies in veterinary and nursing education have reported successfully using fewer than 20 persons in conjunction with larger number of conditions per facet.^{52,68} The small sample size may have contributed to the observed low variance in student scores as assessed by the checklist. If the surgical skills examination on a model, scored using the checklist, is to be used as a high-stakes assessment, particularly at other institutions, further validity evidence and additional reliability data should be gathered to maintain a solid validity argument for its use. Furthermore, the use of a global rating scale in conjunction with the checklist to assess students may help differentiate student performance better than the checklist alone as it allows the rater to evaluate more qualitative aspects of the students' performance.

In conclusion, content validity was very good for the 39-item checklist and was good for the 1-item OSATS GRS, as tested here. The reliability of scores from both instruments was acceptable for a moderate stakes' examination. These results provide evidence to support the use of the checklist described over the OSATS GRS in a moderate-stakes examination when evaluating preclinical third-year veterinary students' technical surgical skills on low-fidelity models. Additional research is necessary to understand at what point in a veterinary students' education the OSATS GRS becomes suitable for assessing surgical skills.

ACKNOWLEDGMENTS

Author Contributions: Farrell RM, BSc, DVM, PGDipMed: Contributed to the study's conception and

the study design; secured human subjects' approval; drafted portions of the introduction, methods, and discussion sections; performed critical edits to the paper; formatted the final manuscript, and serves as the corresponding author. Gilbert GE, EdD, MSPH, PStat[®]: Contributed to the conception of the study and study design; drafted portions of the methods, results, and discussion sections; acted as statistician for the study, and performed critical edits to the paper. Betance L, BS, DVM: Contributed to the execution of the study and performed edits to the paper. Huck J, DVM, DACVS-SA: Contributed to the execution of the study and performed edits to the paper. Hunt JA, DVM, MS: Drafted portions of the results and discussion sections; formatted the manuscript, and performed critical edits to the paper. Dundas J, DVM, DACVS-SA: Contributed to the execution of the study; drafted portions of the results section, and performed critical edits to the paper. Pope E, DVM, MS, DACVS: Contributed to the execution of the study, drafted portions of the introduction and methods sections, and performed critical edits to the paper.

The authors are very grateful to Ms. Ermine Cotton and Ms. Grace Carr Benjamin, librarians, Ross University School of Veterinary Medicine; Ms. Nadia Poponne and Ms. Cynthia Wenham, librarians, Ross University School of Medicine, and Ms. Lisa Blackwell, director of library services, Chamberlain University for their help in locating references used in this investigation. The authors would also like to acknowledge the contributions of the expert content reviewers, Dean Hendrickson, DVM MS, DACVS Michel Heimes, DVM, Marc Dilly, DVM, PhD, MHEd, Carol Bradley, QVN, FHEA, Stephen Horvath, DVM, Tatiana Motta, DVM, MS, Cindy Shmon, DVM, DVSc, DACVS, Kay Eccleshare, BVSc, Rikke Langebaek, DVM, PhD and Mary Lummis, DVM. The authors also wish to thank RUSVM technical staff for their contributions to the completion of this project, and they are grateful to RUSVM faculty for their contributions to developing the curriculum and assessment tool used in this study. Open access funding provided by IReL.

ETHICAL APPROVAL

Approval was granted by the institutional review board at Ross University School of Veterinary Medicine, 7 July 2015 (approval no. 493).

CONFLICT OF INTEREST

The authors declare no conflicts of interest related to this report.

ORCID

Robin M. Farrell  <https://orcid.org/0000-0003-4447-4809>

REFERENCES

- Smeak DD, Hill LN, Lord LK, Allen LCV. Expected frequency of use and proficiency of Core surgical skills in entry-level veterinary practice: 2009 ACVS Core surgical skills Diplomate survey results. *Vet Surg.* 2012;41(7):853-861. doi:10.1111/j.1532-950X.2012.00978.x
- Hill LN, Smeak DD, Lord LK. Frequency of use and proficiency in performance of surgical skills expected of entry-level veterinarians by general practitioners. *J Am Vet Med Assoc.* 2012; 240(11):1345-1354. doi:10.2460/javma.240.11.1345
- Annandale A, Scheepers E, Fosgate GT. The effect of an Ovariohysterectomy model practice on surgical times for final-year veterinary Students' first live-animal Ovariohysterectomies. *J Vet Med Educ.* 2020;47(1):44-55. doi:10.3138/jvme.1217-181r1
- Au Yong JA, Kim SE, Case JB. Survey of clinician and student impressions of a synthetic canine model for gastrointestinal surgery training. *Vet Surg.* 2019;48(3):343-351. doi:10.1111/vsu.13144
- Gopinath D, McGreevy PD, Zuber RM, Klupiec C, Baguley J, Barrs VR. Developments in undergraduate teaching of small-animal soft-tissue surgical skills at the University of Sydney. *J Vet Med Educ.* 2012;39(1):21-29. doi:10.3138/jvme.0411.044R
- Aulmann M, März M, Burgener IA, Alef M, Otto S, Mülling CKW. Development and evaluation of two canine low-fidelity simulation models. *J Vet Med Educ.* May 2015. Accessed May 31, 2015. <http://jvme.utpjournals.press/doi/abs/10.3138/jvme.1114-114R>.
- Read EK, Vallevand A, Farrell RM. Evaluation of veterinary student surgical skills preparation for Ovariohysterectomy using simulators: a pilot study. *J Vet Med Educ.* 2016;43(2):190-213. doi:10.3138/jvme.0815-138R1
- Giusto G, Comino F, Gandini M. Validation of an effective, easy-to-make hemostasis simulator. *J Vet Med Educ.* 2015; 42(1):85-88. doi:10.3138/jvme.0514-050R2
- Fahie M, Cloke A, Lagman M, Levi O, Schmidt P. Training veterinary students to perform Ovariectomy using theMOOSE spay model with traditional method versus the dowling spay retractor. *J Vet Med Educ.* 2016;43(2):176-183. doi:10.3138/jvme.0915-150R
- Holmberg DL, Cockshutt JR, Basher AWP. Use of a dog abdominal surrogate for teaching surgery. *J Vet Med Educ.* 1993;20(3):61-62.
- Badman M, Tullberg M, Höglund OV, Hagman R. Veterinary student confidence after practicing with a new surgical training model for feline Ovariohysterectomy. *J Vet Med Educ.* 2016; 43(4):427-433. doi:10.3138/jvme.1015-165R2
- Caston SS, Schleining JA, Danielson JA, Kersh KD, Reinertson EL. Efficacy of teaching the Gambee suture pattern using simulated small intestine versus cadaveric small intestine. *Vet Surg.* 2016;45(8):1019-1024. doi:10.1111/vsu.12554
- Grimes JA, Wallace ML, Schmiedt CW, Parks AH. Evaluation of surgical models for training veterinary students to perform enterotomies. *Vet Surg.* 2019;48(6):985-996. doi:10.1111/vsu.13228
- MacArthur SL, Johnson MD, Colee JC. Effect of a spay simulator on student competence and anxiety. *J Vet Med Educ.* 2020; 48:115-128. doi:10.3138/jvme.0818-089r3.
- Shaver SL, Larrosa M, Hofmeister EH. Factors affecting the duration of anesthesia and surgery of canine and feline gonadectomies performed by veterinary students in a year-long pre-clinical surgery laboratory. *Vet Surg.* 2019;48(3):352-359. doi:10.1111/vsu.13163
- Griffon DJ, Cronin P, Kirby B, Cottrell DF. Evaluation of a hemostasis model for teaching ovariohysterectomy in veterinary surgery. *Vet Surg.* 2000;29(4):309-316. doi:10.1053/jvet.2000.7541
- Smeak DD, Beck ML, Shaffer CA, Gregg CG. Evaluation of video tape and a simulator for instruction of basic surgical skills. *Vet Surg.* 1991;20(1):30-36. doi:10.1111/j.1532-950X.1991.tb00302.x
- Smeak DD. Teaching surgery to the veterinary novice: the Ohio State University experience. *J Vet Med Educ.* 2007;34(5):620-627. doi:10.3138/jvme.34.5.620
- Hecker K, Read EK, Vallevand A, et al. Assessment of first-year veterinary students' clinical skills using objective structured clinical examinations. *J Vet Med Educ.* 2010;37(4):395-402. doi:10.3138/jvme.37.4.395
- Hunt JA, Heydenburg M, Kelly CK, Anderson SL, Dascanio JJ. Development and validation of a canine castration model and rubric. *J Vet Med Educ.* 2020;47(1):78-90. doi:10.3138/jvme.1117-158r1
- Chen CY, Ragle CA, Lencioni R, Fransson BA. Comparison of 2 training programs for basic laparoscopic skills and simulated surgery performance in veterinary students. *Vet Surg.* 2017; 46(8):1187-1197. doi:10.1111/vsu.12729
- Coffman JM, McConkey MJ, Colee J. Effectiveness of video-assisted, self-directed, and peer-guided learning in the acquisition of surgical skills by veterinary students. *Vet Surg.* 2020; 49(3):582-589. doi:10.1111/vsu.13368
- Hunt JA, Heydenburg M, Anderson SL, Thompson RR. Does virtual reality training improve veterinary students' first canine surgical performance? *Vet Rec.* 2020;186(17):562. doi:10.1136/vr.105749
- Schnabel LV, Maza PS, Williams KM, Irby NL, McDaniel CM, Collins BG. Use of a formal assessment instrument for evaluation of veterinary student surgical skills. *Vet Surg.* 2013;42(4): 488-496. doi:10.1111/j.1532-950X.2013.12006.x
- Williamson JA, Farrell R, Skowron C, et al. Evaluation of a method to assess digitally recorded surgical skills of novice veterinary students. *Vet Surg.* 2018;47(3):378-384. doi:10.1111/vsu.12772
- Williamson JA, Johnson JT, Anderson S, Spangler D, Stonerook M, Dascanio JJ. A randomized trial comparing freely moving and zonal instruction of veterinary surgical skills using Ovariohysterectomy models. *J Vet Med Educ.* 2019;46(2):195-204. doi:10.3138/jvme.0817-009r
- Cook D, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med.* 2006;119:7-16. doi:10.1016/j.amjmed.2005.10.036.
- Cook D, 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 heal. Sci Educ.* 2013;19(2):233-250. doi:10.1007/s10459-013-9458-4
- Reznick R, Regehr G, MacRae H, Martin J, McCulloch W. Testing technical skill via an innovative "bench station"

- examination. *Am J Surg.* 1997;173(3):226-230. doi:[10.1016/S0002-9610\(97\)89597-9](https://doi.org/10.1016/S0002-9610(97)89597-9)
30. Martin JA, Regehr G, Reznick R, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg.* 1997;84(2):273-278. doi:[10.1002/bjs.1800840237](https://doi.org/10.1002/bjs.1800840237)
 31. Hatala R, Cook DA, Brydges R, Hawkins R. Constructing a validity argument for the objective structured assessment of technical skills (OSATS): a systematic review of validity evidence. *Adv Health Sci Educ.* 2015;20(5):1149-1175. doi:[10.1007/s10459-015-9593-1](https://doi.org/10.1007/s10459-015-9593-1)
 32. Issenberg SB, McGaghie WC, Petrusa ER, Lee Gordon D, Scalese RJ. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. *Med Teach.* 2005;27(1):10-28. doi:[10.1080/01421590500046924](https://doi.org/10.1080/01421590500046924)
 33. Schaefer JJ, Vanderbilt A, Cason CL, et al. Literature review: instructional design and pedagogy science in healthcare simulation. *Simul Healthc.* 2011;6(Suppl):S30-S41. doi:[10.1097/SIH.0b013e31822237b4](https://doi.org/10.1097/SIH.0b013e31822237b4)
 34. McGaghie WC, Issenberg SB, Petrusa ER, Scalese RJ. A critical review of simulation-based medical education research: 2003-2009. *Med Educ.* 2010;44(1):50-63. doi:[10.1111/j.1365-2923.2009.03547.x](https://doi.org/10.1111/j.1365-2923.2009.03547.x)
 35. Rhind SM, Baillie S, Brown F, Hammick M, Dozier M. Assessing competence in veterinary medical education: where's the evidence? *J Vet Med Educ.* 2008;35(stage 1):407-411. doi:[10.3138/jvme.35.3.407](https://doi.org/10.3138/jvme.35.3.407)
 36. Harden RM, Stamper N. What is a spiral curriculum? *Med Teach.* 1999;21(2):141-143. doi:[10.1080/01421599979752](https://doi.org/10.1080/01421599979752)
 37. May SA, Silva-Fletcher A. Scaffolded active learning: nine pedagogical principles for building a modern veterinary curriculum. *J Vet Med Educ.* 2015;42:332-339. doi:[10.3138/jvme.0415-063R](https://doi.org/10.3138/jvme.0415-063R)
 38. Baillie, S Booth, N Catterall, A Coombes, N Crowther, E Dilly, M Farrell, R Langebaek, R O'Relilly, M Read E. A Guide to veterinary clinical skills laboratories 2015. <http://www.bris.ac.uk/vetscience/media/docs/csl-guide.pdf>.
 39. Langebæk R, Toft N, Eriksen T. The SimSpay – student perceptions of a low-cost build-it-yourself model for novice training of surgical skills in canine Ovariohysterectomy. *J Vet Med Educ.* May 2015. Accessed May 31, 2015. <http://jvme.utpjournals.press/doi/abs/10.3138/jvme.1014-105>.
 40. Yong JAA, Case JB, Kim SE, Verpaalen VD, McConkey MJ. Survey of instructor and student impressions of a high-fidelity model in canine ovariohysterectomy surgical training. *Vet Surg.* 2019;48(6):975-984. doi:[10.1111/VSU.13218](https://doi.org/10.1111/VSU.13218)
 41. Wilson FR, Pan W, Schumsky DA. Recalculation of the critical values for Lawshe's content validity ratio. *Meas Eval Couns Dev.* 2012;45(3):197-210. doi:[10.1177/0748175612440286](https://doi.org/10.1177/0748175612440286)
 42. Ayre C, Scally AJ. Critical values for Lawshe's content validity ratio: revisiting the original methods of calculation. *Meas Eval Couns Dev.* 2014;47(1):79-86. doi:[10.1177/0748175613513808](https://doi.org/10.1177/0748175613513808)
 43. Devon HA, Block ME, Moyle-Wright P, et al. A psychometric toolbox for testing validity and reliability. *J Nurs Scholarsh.* 2007;39(2):155-164. doi:[10.1111/J.1547-5069.2007.00161.X](https://doi.org/10.1111/J.1547-5069.2007.00161.X)
 44. Wongpakaran N, Wongpakaran T, Wedding D, Gwet KL. A comparison of Cohen's kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC Med Res Methodol.* 2013;13(1). doi:[10.1186/1471-2288-13-61](https://doi.org/10.1186/1471-2288-13-61)
 45. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. *Br J Math Stat Psychol.* 2008;61(Pt 1):29-48. doi:[10.1348/000711006X126600](https://doi.org/10.1348/000711006X126600)
 46. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. the problems of two paradoxes. *J Clin Epidemiol.* 1990;43(6):543-549.
 47. George D, Mallery P. *SPSS for Windows Step by Step: A Simple Guide and Reference. 11.0 Update.* 4th ed. Allyn & Bacon; 2003.
 48. Cronbach L, Rajaratnam N, Gleser G. Theory of generalizability: a liberalization of reliability theory. *Br J Stat Psychol.* 1963;16(2):137-163.
 49. Cronbach L, Gleser G, Nanda H, Rajaratnam N. *The Dependability of Behavioral Measurements.* John Wiley & Sons, Inc.; 1972.
 50. Brennan R. *Generalizability Theory.* Springer-Verlag; 2001.
 51. Walzak A, Bacchus M, Schaefer JP, et al. Diagnosing technical competence in six bedside procedures: comparing checklists and a global rating scale in the assessment of resident performance. *Acad Med.* 2015;90(8):1100-1108. doi:[10.1097/ACM.0000000000000704](https://doi.org/10.1097/ACM.0000000000000704)
 52. Read EK, Bell C, Rhind S, Hecker KG. The use of global rating scales for OSCEs in veterinary medicine. *PLoS One.* 2015;10(3):e0121000. doi:[10.1371/Journal.Pone.0121000](https://doi.org/10.1371/Journal.Pone.0121000).
 53. AAVMC Working Group on Competency-Based Veterinary Education. *Milestones Competency-Based Veterinary Education: Part 3.*; 2019.
 54. Royal KD, Hecker KG. Understanding reliability: a review for veterinary educators. *J Vet Med Educ.* 2015;1:1-4. doi:[10.3138/jvme.0315-030R](https://doi.org/10.3138/jvme.0315-030R)
 55. Brennan R. An NCME instruction module on generalizability theory. *Educ Meas Issues Pract.* 1992;11(4):27-34. doi:[10.1111/j.1745-3992.1992.tb00260.x](https://doi.org/10.1111/j.1745-3992.1992.tb00260.x)
 56. Downing SM. The metric of medical education reliability: on the reproducibility of assessment data. *Med Educ.* 2004;38:1006-1012. doi:[10.1046/j.1365-2929.2004.01932.x](https://doi.org/10.1046/j.1365-2929.2004.01932.x)
 57. Ericsson KA. Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Acad Med.* 2004;79:S70-S81. doi:[10.1097/00001888-200410001-00022](https://doi.org/10.1097/00001888-200410001-00022)
 58. Ericsson KA, Krampe RT, Tesch-Römer C. The role of deliberate practice in the acquisition of expert performance. *Psychol Rev.* 1993;100(3):363-406. doi:[10.1037/0033-295x.100.3.363](https://doi.org/10.1037/0033-295x.100.3.363)
 59. Moulton C-AE, Dubrowski A, Macrae H, Graham B, Grober E, Reznick R. Teaching surgical skills: what kind of practice makes perfect?: a randomized, controlled trial. *Ann Surg.* 2006;244(3):400-409. doi:[10.1097/01.sla.0000234808.85789.6a](https://doi.org/10.1097/01.sla.0000234808.85789.6a)
 60. Fields RD. Making memories stick. *Sci Am.* 2005;292(2):74-81. <http://www.jstor.org/stable/26060881>
 61. Cepeda NJ, Vul E, Rohrer D, Wixted JT, Pashler H. Spacing effects in learning: a temporal ridgeline of optimal retention: research article. *Psychol Sci.* 2008;19(11):1095-1102. doi:[10.1111/j.1467-9280.2008.02209.x](https://doi.org/10.1111/j.1467-9280.2008.02209.x)
 62. Van Der Vleuten CPM, Schuwirth LWT, Driessen EW, et al. A model for programmatic assessment fit for purpose. *Med Teach.* 2012;34(3):205-214. doi:[10.3109/0142159X.2012.652239](https://doi.org/10.3109/0142159X.2012.652239)

63. Mackay S, Morgan P, Datta V, Chang A, Darzi A. Practice distribution in procedural skills training. *Surg Endosc Other Intervent Tech*. 2002;16(6):957-961. doi:[10.1007/S00464-001-9132-4](https://doi.org/10.1007/S00464-001-9132-4)
64. Williamson JA, Brisson BA, Anderson SL, Farrell RM, Spangler D. Comparison of 2 canine celiotomy closure models for training novice veterinary students. *Vet Surg*. 2019;48(6):966-974. doi:[10.1111/vsu.13224](https://doi.org/10.1111/vsu.13224)
65. Tan J-Y, Ma IWY, Hunt JA, et al. Video recording in veterinary medicine OSCEs: feasibility and inter-rater agreement between live performance examiners and video recording reviewing examiners. *J Vet Med Educ*. 2020;48(4):485-491. doi:[10.3138/jvme-2019-0142](https://doi.org/10.3138/jvme-2019-0142).
66. Anderson SL, Miller L, Gibbons P, et al. Development and validation of a bovine castration model and rubric. *J Vet Med Educ*. 2020;48:e20180016. doi:[10.3138/jvme.2018-0016](https://doi.org/10.3138/jvme.2018-0016)
67. Webb NM, Rowley GL, Shavelson RJ. Using generalizability theory in counseling and development. *Meas Eval Couns Dev*. 2018;21(2):81-90. doi:[10.1080/07481756.1988.12022886](https://doi.org/10.1080/07481756.1988.12022886)
68. O'Brien J, Thompson MS, Hagler D. Using generalizability theory to inform optimal design for a nursing performance assessment. *Eval Health Prof*. 2017;42(3):297-327. doi:[10.1177/0163278717735565](https://doi.org/10.1177/0163278717735565)

How to cite this article: Farrell RM, Gilbert GE, Betance L, et al. Evaluating validity evidence for 2 instruments developed to assess students' surgical skills in a simulated environment. *Veterinary Surgery*. 2022;51(5):788-800. doi:[10.1111/vsu.13791](https://doi.org/10.1111/vsu.13791)