Received: 17 March 2017 | Revised: 10 July 2017 | Accepted: 11 September 2017

DOI: 10.1111/vsu.12772

ORIGINAL ARTICLE

Evaluation of a method to assess digitally recorded surgical skills of novice veterinary students

Julie A. Williamson DVM, MSc, AFAMEE[1] | Robin Farrell DVM[2] | Casey Skowron BS[1] | Brigitte A. Brisson DMV, DVSc, DACVS[3] | Stacy Anderson DVM, MVSc, PhD, DACVS-LA[1] | Dawn Spangler DVM[1] | Jason Johnson DVM, MS, DACT[1]

[1] Lincoln Memorial University College of Veterinary Medicine, Harrogate, Tennessee

[2] Ross University School of Veterinary Medicine, St. Kitts, Federation of Saint Kitts and Nevis

[3] University of Guelph, Ontario Veterinary College, Guelph, Ontario, Canada

Correspondence
Julie Williamson, LMU-CVM, 6965 Cumberland Gap Parkway, Harrogate, TN 37752.
Email: julie.williamson@lmunet.edu

Abstract

Objective: To evaluate a method to assess surgical skills of veterinary students that is based on digital recording of their performance during closure of a celiotomy in canine cadavers.

Sample Population: Second year veterinary students without prior experience with live animal or simulated surgical procedure (n = 19)

Methods: Each student completed a 3-layer closure of a celiotomy on a canine cadaver. Each procedure was digitally recorded with a single small wide-angle camera mounted to the overhead surgical light. The performance was scored by 2 of 5 trained raters who were unaware of the identity of the students. Scores were based on an 8-item rubric that was created to evaluate surgical skills that are required to close a celiotomy. The reliability of scores was tested with Cronbach's α, intraclass correlation, and a generalizability study.

Results: The internal consistency of the grading rubric, as measured by α, was .76. Interrater reliability, as measured by intraclass correlation, was 0.64. The generalizability coefficient was 0.56.

Conclusion: Reliability measures of 0.60 and above have been suggested as adequate to assess low-stakes skills. The task-specific grading rubric used in this study to evaluate veterinary surgical skills captured by a single wide-angle camera mounted to an overhead surgical light produced scores with acceptable internal consistency, substantial interrater reliability, and marginal generalizability.

Impact: Evaluation of veterinary students' surgical skills by using digital recordings with a validated rubric improves flexibility when designing accurate assessments.

1 | INTRODUCTION

Closing a celiotomy is one of the most common surgical tasks expected of veterinarians in companion animal practice. This task requires surgical skills that veterinary graduates are expected to possess, such as placing secure knots with instruments and closure of subcutaneous tissue layer.[1,2] The American Veterinary Medical Association's Council on Education charges veterinary educators with observing and assessing student competencies in several areas, including basic surgical skills.[3] Educators are expected to provide documentation supporting the accuracy of their assessments.[3] Establishing the reliability and validity of an assessment method is critical in justifying decisions that are based on that assessment, such as assigning a passing or failing grade.[4,5] Reliable, valid assessments improve the evaluation

of student educational outcomes, support research in educational methods, and could eventually assist with credentialing decisions.[4-7] Demonstrating adequate reliability of scores generated by a rubric in a particular setting is an essential component of validity evidence.[5,8,9] The situational nature of measures of reliability requires that veterinary educators re-evaluate the reliability of scores generated by an assessment tool whenever any feature of the assessment is changed, such as location, raters, or student level of training.[5,10]

Methods used to assess veterinary surgical skills have previously relied on a direct observation by the rater or have been based on digital recordings, similarly to the assessments in human medical education.[11-20] The evaluation of digital recordings offers some benefits over assessing skills in real time. Raters are able to start and stop digital recordings to take breaks, make notes, or replay particular sections of the performance, which may reduce rater errors as a result of rater fatigue or skimming.[9,21] Blinding digitally recorded performances can decrease rater biases such as favoritism.[9,21] Assessments that are based on digitally recorded performances can be replayed if scores are challenged by a student, improving defensibility.[22] The ability to have multiple raters score a digitally recorded surgical performance can establish interrater reliability and offer support for generalizing the scores beyond those obtained by a single rater.[23] Digitally recorded performances were used to assess veterinary students' surgical skills, specifically ligation, in a report, but the reliability of scores was not reported.[14] No method has been described or evaluated to generate and assess digitally recorded surgical skills of novice veterinary students.

This study had a twofold objective: (1) to describe a digital recording technique of veterinary students closing a celiotomy on canine cadavers, with small wide-angle cameras mounted to the overhead surgical lights, and (2) to evaluate interrater reliability, internal consistency, and generalizability of scores derived from those digital recordings by using components of a previously published task-specific rubric. We hypothesized that the digital recordings collected by small wide-angle cameras attached to overhead surgical lights would allow consistent assessment of surgical skills that are involved in closing a celiotomy, by using the task-specific rubric.

2 | MATERIALS AND METHODS

2.1 | Population

The institutional review board of Lincoln Memorial University approved the use of veterinary students for this study. A convenience sample of second year veterinary students (n = 19) was recruited to participate during their clinical skills course and completed informed consent. None of the students had ever closed an abdominal incision on a live animal. During the preceding year, students received surgical skills training in placing secure ligatures and suturing fabric on low-fidelity models. Eight fresh mixed-breed canine cadavers, ranging in size from 15-25 kg with body condition scores of 2.5-3.5 on a 5-point scale, were positioned in dorsal recumbency on surgical tables. Each cadaver was draped with a single impermeable fenestrated drape. A ventral midline incision was created by 1 of the investigators (J.A.W.), starting at the umbilicus and extending caudally to a point halfway between the umbilicus and pubis on females or immediately cranial to the prepuce on males.

2.2 | Digital recording of the partial closure of a celiotomy by students

Study participants were instructed to view a digital video of a 3-layer abdominal closure before coming into the laboratory to partially close a celiotomy in 3 layers: 2 simple interrupted sutures in the body wall, a short section of simple continuous subcutaneous tissue closure with buried knots, and 2 simple interrupted or cruciate skin sutures. Students placed sutures in each layer directly above the sutures in the layer beneath it, closing a small section of the larger abdominal incision. Sutures were removed before the next student performed his or her 3-layer closure. After a section of the incision had been sutured and reopened, subsequent students were directed to suture an adjacent portion of the incision to prevent the tissue maceration and its potential consequences on students' performance.

Small wide-angle portable cameras (HERO; GoPro, San Mateo, California) were clipped with flexible arms (HERO; GoPro) to the handle of the moveable light above the surgical table. Cameras were positioned approximately 35 cm above the cadaver's abdomen and were directed at the center of the abdominal incision. This positioning allowed inclusion of the entire surgical field along with the student's hands and forearms (Figure 1).

Surgical lights were illuminated to half of their maximum capacity and covered with parchment paper (Reynolds Consumer Products, Lake Forest, Illinois) to diffuse the light, which prevented washout of the image while maintaining an intensity of light adequate for surgery. Students were identified by numbers on their surgical gloves to raters unaware of the student identity. To obtain quality video recordings during the study, researchers and assistants did not allow students to adjust the intensity or direction of the surgical light and discouraged them from leaning over the sterile field and obscuring the video.

2.3 | Assessment of digital recordings

From the digital recordings, students' performance scores were based on 8 checklist items relevant to the closure of a

FIGURE 1 Sample frame from a student's digital recording

celiotomy contained in the canine ovariohysterectomy grading rubric published by Read et al.[19] The rubric was chosen for this study because some validity evidence has been established for its use in evaluating veterinary student surgical skills performance. Among the 12 items contained in this portion of Read and colleagues' rubric, 8 items were relevant to the current study and were used to score students' performance. The 4 items of the rubric that were not evaluated included suture selection, starting and ending the body wall closure at an acceptable distance from each end of the incision, and burying knots in the skin layer. Anchoring descriptions such as "student performs pattern with some errors" (eg, missed bites, interlocking suture, etc) or "does not include enough tissue in closure" were added for each item in an effort to improve rater consistency. Items were scored by using a 4-point scale: zero points were awarded for unsatisfactory performance or unperformed skill, 1 point for a borderline performance, 2 points for a good performance, and 3 points for an excellent performance (Table 1). Raters were instructed to deduct 1 point from each item in which the student demonstrated traumatic tissue handling, incorrect instrument use, or poor knot quality as assessed by the number of square throws placed, the number of half hitches or granny knots created, and tension as indicated by apposition of the tissue. Students could achieve a maximum score of 24 points.

Five educators who are experienced at evaluating surgical skills were selected from 3 veterinary schools in the United States, Canada, and the Caribbean as raters. On an internet video-hosting site (YouTube; Google, San Bruno, California), raters watched a 30-minute training presentation describing how to score student performances with the rubric (https://www.youtube.com/watch?v=Ib-Woddh5fw). This 20-slide narrated presentation described each item on the rubric, reviewed examples of how to score different student performance scenarios, and discussed methods to avoid common sources of rater error. Raters were provided with a copy of the training presentation for reference purposes. Students' digital recordings were divided randomly among the raters, with 2 raters scoring each recording. Raters streamed the digital recordings from an internet file sharing site (Hightail, Campbell, California) and watched the recordings on their personal computer monitors. Raters were permitted to pause or replay the digital recordings as often as they deemed necessary.

2.4 | Data analysis

Descriptive statistics such as range, mean, and SD were used to characterize the performance scores. Data were assessed for normality by using the Shapiro-Wilk test and found to be normally distributed. Internal consistency was measured by Cronbach's α; values of .7–.9 are generally considered acceptable by medical educators.[24] Interrater reliability was evaluated by using a 1-way random single-measure intraclass correlation; values of 0.61-0.8 are considered to demonstrate substantial agreement among raters.[25] A nested generalizability analysis was used to assess variance and generate a generalizability coefficient (G-coefficient). Statistics were performed in SPSS (IBM, Armonk, New York) and R (R Foundation for Statistical Computing, Vienna, Austria) statistics software.

3 | RESULTS

Scores obtained from the rubric were combined to create a total score for the task, and all analyses were performed by using this total score. The mean performance score awarded for the 3-layer abdominal closure task was 12.76 of 24 (53%), with SD of 5.00 points. Student performance scores ranged from 4 of 24 (17%) to 22 of 24 (92%). The relatively low mean score was consistent with the second year students' limited level of previous training for the task. The wide SD suggests that students varied significantly in their baseline skill level. Raters used a very wide r of the rubric's possible scores, which improved internal consistency and the ability to differentiate student skill level.

Reliability as assessed by internal consistency was 0.76. Intraclass correlation was used to evaluate interrater reliability, or the degree to which evaluators demonstrated consistency in their ratings of student performance. The intraclass correlation for performance scores was 0.64 with a 95% CI from 0.29 to 0.84.

The G-coefficient, a robust measure of an assessment's reliability that quantifies the amount of error attributed to defined variables such as setting, rater, or occasion, was 0.56.[23] The generalizability study attributed 9.2% of the

TABLE 1 Rubric used for scoring student performance

	Task	0—Unsatisfactory, fails to perform skill at acceptable level	1—Borderline, performs skill with noteworthy or significant flaws	2—Good, performs skill with minor flaws or revisions	3—Excellent, performs skill competently on first attempt
1	Body wall: adequate bite sizes (5-10 mm)	Most or all bites were <5 mm or > 10 mm	Some bites were <5 mm or >10 mm	A single bite was <5 mm or >10 mm or multiple attempts were needed to get adequate bite size	Bites were consistently 5-10 mm on first attempt
2	Body wall: adequate bite spacing (5-10 mm)	Spacing between all sutures was <5 mm or >10 mm	Spacing between sutures was left at <5 mm or >10 mm in some cases but not in others (inconsistent spacing) or student had to replace sutures multiple times to get 5-10-mm spacing	Distance between sutures was 5-10 mm but only after the student corrected a simple spacing error	Spacing was consistently 5-10 mm the first time placed
3	Body wall: linea layer incorporates bites in ventral (external) rectus fascia only	Student includes all abdominal wall layers in all sutures	Student includes all abdominal wall layers in some sutures but not all or excessive tissue trauma in process of getting correct layer	Student includes only ventral rectus fascia in nearly all sutures; sutures may have been replaced to include only that layer	Student includes only appropriate layer on initial placement
4	SQ: correctly buries knot at start	Knot tying does not resemble a buried knot	Student understands basic concept but makes a significant error (eg, ties to a superficial end, ties over top of a superficial suture, has multiple uncorrected half hitches)	Student buries the knot but not on first attempt or student may have a single half hitch	Student buries the knot on the first attempt without half hitching or other threats to knot security
5	SQ: simple continuous pattern placed correctly	Student does not acceptably perform simple continuous pattern of superficial to deep, deep to superficial toward student, or student backhands entire pattern	Student performs pattern with some errors (eg, missed bites, interlocking suture, etc) or does not include enough tissue in closure	Student performs pattern correctly but only after correcting mistake(s)	Student performs pattern correctly on the first attempt
6	SQ: correctly ends with buried knot	Knot tying does not resemble a buried knot	Student understands basic concept but makes a significant error (eg, ties to wrong loop, ties to a superficial end, ties over top of a superficial suture, has multiple uncorrected half hitches)	Student buries the knot but not on first attempt or student may have 1 half hitch	Student buries the knot on the first attempt without half hitching or other threats to knot security
7	Skin: simple interrupted or cruciate pattern placed correctly	Pattern performed incorrectly	Student performs the right pattern but makes a significant error in placement (eg, backhands pattern, creates excessive tissue trauma)	Student places pattern correctly but requires >1 attempt	Student places pattern correctly on first attempt
8	Skin: proper tension on completion of closure—apposed but not too tight	Edges are not apposed; or suture is very tight, causing skin to bunch, evert, or overlap significantly	Some issues with tension of final sutures, too tight or too loose but not extremely so	Suture tension is appropriate but took >1 attempt	Suture tension is appropriate on first attempt

SQ, subcutaneous tissue

TABLE 2 Generalizability analysis and variance components

Factor	Variance	Variance (%)	G-coefficient
Students	2.36	9.2	0.56
Raters	2.53	9.8	. . .
Students by rater	12.14	47	. . .
Residual	8.80	34.1	. . .

. . ., not applicable

variance in scores to the students' performance of the task. The variance attributed to each rater was similar at 9.8%. Variance attributed to the student by rater facet was 47%. Unsystematic error resulting from residual factors was calculated at 34.1% (Table 2).

4 | DISCUSSION

Using small wide-angle portable cameras to record students' abdominal incision closure skills on canine cadavers allowed raters to produce scores with acceptable internal consistency, substantial interrater reliability as assessed by intraclass correlation, and marginal generalizability. These results provide evidence to support the use of this rubric in low-stakes assessment. Reliable scoring of digitally recorded surgical skills can improve the defensibility of high-stakes assessments, serve as training aids for raters, provide feedback to educators about instructional methods, encourage student reflection and self-assessment, and facilitate educational research by allowing raters to be blinded to participants' identities.[4-7]

The results of this study are consistent with a review of surgical skills assessments in human medical education, in which checklist scores are typically found reliable, valid, and sensitive.[27] However, the rubric used to evaluate abdominal closure differs from most reported checklists in veterinary and human medical education with a dichotomous choice of "performed" or "not performed." One study in humans reported on a nondichotomous task-specific rubric, the operative component rating scale (OCRS), in which each step of the procedure is scored based on a numeric scale, similarly to the abdominal closure rubric.[6] The OCRS resulted in adequate interrater reliability for 3 of 4 tested surgical tasks.[6] Clarifying the nomenclature of assessment tools would facilitate the design, interpretation, and generalization of future research related to performance evaluation. We suggest reserving the term *checklist* for a task-specific dichotomous rubric and using *operative component rating scale* to describe a task-specific rubric with nondichotomous rating.

Surgical skills development is influenced by the number of hours spent in deliberate practice.[28,29] The low scores reported here reflect the students' lack of previous hands-on experience. The wide variation in students' baseline ability may translate into large differences between students in educational time and resources that are required to acquire surgical skills. This variability supports the concept of mastery learning, a paradigm of competency-based education that requires students to reach predetermined performance standards for advancement, independent of curricular time.[30-32] The variation in baseline ability also suggests that researchers should be cautious about using only a student's raw performance score at the completion of an educational intervention to assess its effectiveness. Preintervention and postintervention testing and calculation of a net improvement score may be more appropriate for evaluating an educational method, particularly with novice participants.

In a generalizability study, educators define variables or facets that may introduce error into the assessment, including the student, rater, occasion, or task performed.[23] Unidentified facets are combined into residual or random error, which cannot be eliminated.[23] A low G-coefficient indicates a high level of random error in the assessment. What constitutes an acceptable G-coefficient is, just as with other measures of reliability, dependent on the proposed interpretation of the scores.[23] Scores for high-stakes assessments, such as progression hurdles or credentialing decisions, should be associated with more rigorous validity evidence, including reliability measures, than scores for low-stakes assessments that provide only feedback about students' progress.[22] Scores obtained by using the abdominal closure rubric resulted in a G-coefficient of 0.56, which fell just short of the generally accepted benchmark of 0.60.[26] The more detailed error analysis performed in G studies typically leads to a larger estimated standard error and lower G-coefficients than values obtained by using other reliability measures.[23] The rubric's scores reached a G-coefficient similar to other reports in veterinary skills assessments, which include G-coefficients of 0.23-0.67 for assessing feline abdominal palpation[33] and 0.42-0.86 for 4 medical and surgical skills.[18] The cadaver used by each student and the order in which it was used were not noted in the present study. This limitation prevents calculation of the variance specific to those facets; instead, those facets were included in the measurement of random error.

The intraclass correlation coefficient showed that scores from different raters were in substantial agreement. However, the high "students by rater" facet of the generalizability study indicated that inconsistencies in the raters' rank ordering of students contributed to the majority of the assessment's variance. The presence of mild variations in the distance and angle between the camera and the cadaver and the lack of standardization of the size of computer screen on which raters viewed recordings could have made tissue layers and suture spacing easier to evaluate for some students. Standardizing the camera distance, angle, and rater viewing screens

would eliminate these as potential sources of error. Instructions about how to score global flaws, such as poor instrument handling, knot quality, or tissue handling, were given in the rater training presentation; however, placing these items separately in the rubric or adding them to the rubric's anchoring descriptions may have improved rater reliability. Rank ordering could have been influenced by raters unequally scoring students with global flaws. Finally, raters were at different institutions, and may have been variably sensitive to particular errors, depending on the emphasis placed on those in their present or previous teaching institutions. Raters pose the greatest risk to reliability and validity of skills assessment.[34] Variance in scores attributed to the rater can be decreased by enhancing rater training or by increasing the number of raters contributing to each score.[23] In retrospect, performing a statistical decision study would have predicted the impact of having multiple raters score each performance on reliability. However, because skills assessments are inherently labor intensive, improving rater training is typically more feasible than adding raters. Raters in the present study were trained with 2 goals: first, making them aware of common sources of rater error and, second, familiarizing them with the performance standards for the task to be observed.[21] All raters were experienced veterinary educators; less experienced raters may have required additional rater training to reach similar reliability measures. Interrater reliability likely could be improved by collectively scoring sample performances.

Pitfalls in assessing digitally recorded surgical skills have been previously described and include surgical lights washing out recordings, low resolution equipment resulting in poor quality recordings, excessive motion when using head-mounted or hand-held cameras, obstruction of the video by members of the surgical team, and lack of sterility of the camera limiting placement options and hampering intraoperative adjustment.[35] To address these challenges, we used high-definition video cameras and set surgical lights on low power and covered them with parchment paper to diffuse light. We also mounted cameras on the surgical light, which did not move during the recording. A review article in human medical education suggested that evaluation of digitally recorded surgeries may differentiate performance levels only if those differed markedly.[15] However, many of the procedures reviewed were performed laparoscopically and are rare in general veterinary practice. In addition, all surgeries involved participants who were more experienced than veterinary students.[15] The surgical novices enrolled in our study may have had wider variations in baseline skill levels, facilitating the identification of quantifiable variations.

This study was limited by a small sample from a single veterinary school, which may reduce its generalizability to different settings. In addition, several students inadvertently sutured a portion of the same incision on the cadavers. Although the previous student's sutures had been removed, the remaining effects of these sutures on the tissue may have influenced the performance of subsequent students. If so, this impact was captured as a component of the residual error in the generalizability analysis. When the availability and cost of cadavers, storage facilities, and disposal requires having multiple students suture 1 cadaver, redraping the cadaver between procedures to expose only the current student's section of the incision would conceal any evidence of previous students' attempts.

This study considered several measures of reliability of assessment scores, which is only 1 component of validation evidence. Although creating and evaluating a full validation framework was beyond the scope of this study, content evidence supporting validation has also been provided by the cognitive task analysis process that Read and colleagues[19] used to create the parent rubric for canine ovariohysterectomy. Our study reports on the reliability of scores, but we did not measure the accuracy of raters' perceptions. For example, raters scored the suture spacing observed on the digital recording, but these scores were not compared with the actual distance measured on cadavers between sutures. Placing a measuring device on the cadaver would have allowed calibration and may have improved reliability. Also, raters assessed whether students identified and sutured the proper tissue layers, but cadavers were not examined after each student performance to verify the accuracy of raters. In-person scoring allows raters to measure suture distances and touch tissue layers if required for their evaluation, but this option is not available to raters when they are scoring by digital recordings. Future research should focus on the correlation between scores obtained from digital recordings and those obtained from in-person scoring. The application of this method to assess other surgical tasks also warrants further investigation. Finally, additional study is required to determine how variables such as the position and angle of the camera, order of students, features of the cadaver, number of sutures placed, and size of the rater viewing screen influence generalizability and rater accuracy.

In conclusion, the surgical skills required from a novice veterinary surgeon to close an abdominal incision can be assessed by using a digital recording of the procedure, obtained from a single small wide-angle camera mounted in an overhead surgical light. The task-specific rubric used to score recorded performances led to acceptable internal consistency, substantial interrater reliability, and marginal generalizability in the present setting. Standardizing the environmental arrangement of cameras and cadavers could further improve reliability. These results provide evidence to support the use of this rubric in low-stakes assessment.

CONFLICT OF INTEREST

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

REFERENCES

[1] Hill L, Smeak D, Lord L. Frequency of use and proficiency in performance of surgical skills expected of entry-level veterinarians by general practitioners. *J Am Vet Med Assoc.* 2012;240:1345-1354.

[2] Smeak D, Hill L, Lord L, Allen L. 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;41:853-861.

[3] American Veterinary Medical Association. Statement on clinical competencies outcomes. https://www.avma.org/ProfessionalDevelopment/Education/Accreditation/Colleges/Documents/coe_pp.pdf. Accessed March 10, 2017.

[4] Messick S. Meaning and values in test validation: the science and ethics of assessment. *Educ Res.* 1989;18:5-11.

[5] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association; 1999.

[6] Dath D, Regehr G, Birch D, et al. Toward reliable operative assessment: the reliability and feasibility of videotaped assessment of laparoscopic technical skills. *Surg Endosc.* 2004;18:1800-1804.

[7] Wass V, Van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. *Lancet.* 2001;357:945-949.

[8] Downing SM. Validity: on the meaningful interpretation of assessment data. *Med Educ.* 2003;37:830-837.

[9] Royal KD, Hecker KG. Understanding reliability: a review for veterinary educators. *J Vet Med Educ.* 2016;43:1-4.

[10] Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med.* 2006;119:166.e7-166.e16.

[11] Aggarwal R, Grantcharov T, Moorthy K, Milland T, Darzi A. Toward feasible, valid, and reliable video-based assessments of technical surgical skills in the operating room. *Ann Surg.* 2008;247:372-379.

[12] Chang L, Hogle NJ, Moore BB, et al. Reliable assessment of laparoscopic performance in the operating room using videotape analysis. *Surg Innov.* 2007;14:122-126.

[13] Kopta JA. An approach to the evaluation of operative skills. *Surgery.* 1971;70:297-303.

[14] Smeak DD, Beck ML, Shaffer CA, Gregg CG. Evaluation of video tape and a simulator for instruction of basic surgical skills. *Vet Surg.* 1991;20:30-36.

[15] Van Hove PD, Tuijthof GJM, Verdaasdonk EGG, Stassen LPS, Dankelman J. Objective assessment of technical surgical skills. *Br J Surg.* 2010;97:972-987.

[16] Giusto G, Comino F, Gandini M. Validation of an effective, easy-to-make hemostasis simulator. *J Med Vet Educ.* 2015;42:85-88.

[17] Griffon DJ, Cronin P, Kirby B, Cottrell DF. Evaluation of a hemostasis model for teaching ovariohysterectomy in veterinary surgery. *Vet Surg.* 2000;29:309-316.

[18] 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;37:395-402.

[19] 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;43:190-213.

[20] 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;42:488-496.

[21] Royal KD, Hecker KG. Rater errors in clinical performance assessments. *J Vet Med Educ.* 2016;43:5-8.

[22] Hecker K, Violato C. Validity, reliability, and defensibility of assessments in veterinary education. *J Vet Med Educ.* 2009;36:271-275.

[23] Kane M. The role of generalizability in validity. Paper presented at: National Council on Measurement in Education annual meeting; April 19-23, 1999; Montreal, Quebec, Canada.

[24] Tavakol M, Dennick R. Making sense of Cronbach's alpha. *Int J Med Educ.* 2011;2:53-55.

[25] Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.*1977;33:159-174.

[26] Ilgen JS, Ma IW, Hatala R, Cook DA. A systematic review of validity evidence for checklists versus global rating scales in simulation-based assessment. *Med Educ.* 2015;49:161-173.

[27] Ericsson KA, Krampe RT, Tesch-Römer C. The role of deliberate practice in the acquisition of expert performance. *Psychol Rev.* 1993;100:363-406.

[28] Ericsson KA. Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Acad Med.* 2004;79:70-81.

[29] McGaghie WC. Mastery learning: it is time for medical education to join the 21st century. *Acad Med.* 2015;90:1438-1441.

[30] Griswold-Theodorson S, Ponnuru S, Dong C, Szyld D, Reed T, McGaghie WC. Beyond the simulation laboratory: a realist synthesis review of clinical outcomes of simulation-based mastery learning. *Acad Med.* 2015;90:1553-1560.

[31] Yudkowsky R, Park YS, Lineberry M, Knox A, Ritter EM. Setting mastery learning standards. *Acad Med.* 2015;90:1495-1500.

[32] Shea JA, Fortna GS. Psychometric methods. In: Norman G, Van Der Leuten C, Newble D, eds. *International Handbook of Research in Medical Education*. New York, NY: Springer Science & Business Media; 2002:97-126.

[33] Williamson JA, Hecker K, Yvorchuk K, Artemiou E, French H, Fuentealba C. Development and validation of a feline abdominal palpation model and scoring rubric. *Vet Rec.* 2015;177:151.

[34] Linacre JM. *Many-Facet Rasch Measurement*. Chicago, IL: MESA Press; 1989.

[35] Cosman PH, Shearer CJ, Hugh TJ, Biankin AV, Merrett ND. A novel approach to high definition, high-contrast video capture in abdominal surgery. *Ann Surg.* 2007;245:533-535.

How to cite this article: Williamson JA, Farrell R, Skowron C, et al. Evaluation of a method to assess digitally recorded surgical skills of novice veterinary students. *Veterinary Surgery*. 2018;00:1-7. https://doi.org/10.1111/vsu.12772