RESEARCH PAPER

Development of the Auburn Induction Scale for evaluating induction quality in dogs

Kathryn L Wolfe[a], **Erik H Hofmeister**[b], **Stuart C Clark-Price**[b], **Rachel Reed**[c] & **Jane Quandt**[c]

[a]College of Veterinary Medicine, Auburn University, Auburn, AL, USA

[b]Department of Clinical Sciences, College of Veterinary Medicine, Auburn University, Auburn, AL, USA

[c]Department of Small Animal Medicine and Surgery, College of Veterinary Medicine, University of Georgia, Athens, GA, USA

Correspondence: Erik Hofmeister, Department of Clinical Sciences, College of Veterinary Medicine, Auburn University, 1130 Wire Road, Auburn, AL, 36849, USA. E-mail: kaastel@gmail.com

Abstract

Objective To develop and begin establishing evidence for validity of an instrument to assess the quality of induction in dogs.

Study design Cross-sectional survey and video scoring.

Animals and population A total of 51 veterinary anesthesia personnel, four board-certified anesthesiologists and videos of induction of anesthesia in 18 dogs.

Methods In Part 1, an online survey was sent to veterinary anesthesia personnel to solicit expressions and words that they associate with induction of anesthesia. These expressions were evaluated by four anesthesiologists to create a composite scale (Auburn Induction Scale). In Part 2, 18 videos were reviewed by the same four anesthesiologists on two separate occasions. The videos were scored using the Auburn Induction Scale, a simple descriptive scale (SDS) and a visual analog scale (VAS). Intra-rater and inter-rater reliability was measured using an intraclass correlation coefficient (ICC). Significance was set at $p < 0.05$.

Results The survey yielded 51 responses that were condensed into 133 expressions. The four anesthesiologists created 18 items incorporating the 133 expressions. The mean ± standard deviation intra-rater reliability ICC was 0.81 ± 0.08 for the Auburn Induction Scale, 0.71 ± 0.02 for the SDS and 0.71 ± 0.08 for the VAS for all raters. The mean ± standard deviation inter-rater reliability ICC was 0.69 ± 0.04 for the Auburn Induction Scale, 0.61 ± 0.05 for the SDS and 0.60 ± 0.06 for the VAS.

Conclusions and clinical relevance In a research setting, widespread use of this scale may be helpful in increasing the accuracy of data and improving agreement between studies assessing induction of anesthesia in dogs. The results of this study have yielded a composite scale that is more reliable between and among raters than a unidimensional scale.

Keywords behavior observation techniques, validation study.

Introduction

Induction of anesthesia is an important step in the anesthetic process, and has been studied with a wide variety of drugs and protocols. More than 40 different induction scales have been used to assess the quality of induction (Wolfe & Hofmeister 2021). Scales have varied in complexity from two simple levels (e.g. 'good' and 'bad') to five levels of description [i.e. simple descriptive scale (SDS)] to a visual analog scale (VAS). Numerous studies using such scales have identified the limitations of their use, including subjectivity, unclear instructions, inability to compare scales and poor agreement among raters (Wolfe & Hofmeister 2021). This presents a challenge, especially to researchers, when aiming to obtain consistent and accurate data. These limitations exist largely because there is no validated scoring system developed to quantify induction quality in dogs.

Validation is the process of gathering evidence to determine if an instrument accurately measures what it was intended to measure (Raykov & Marcoulides 2011). Validity for scale development can be separated into construct, content and criterion validity (Streiner et al. 2015). A scale that is not validated has no verification that its items fully relate to what it is measuring; therefore, the data gathered from the use of that scale cannot be considered dependable. For example, the scales used to measure induction quality in recent literature cover very general behavioral signs, such as vocalization and paddling, but because these items have not been previously selected or tested with proper techniques, the capability of the

scales to measure the quality of an induction cannot be confirmed (Raykov & Marcoulides 2011; Amengual et al. 2013; Tamura et al. 2016). As instruments used to score induction gather information about unknowable phenomena (i.e. a 'good' induction), an instrument never 'is' valid. Instead, evidence can be accumulated which supports validity for that instrument. The five domains which can provide evidence in support of validity are response processes, test content, consequences of testing, relationship with other variables and internal structure (American Educational Research Association, American Psychological Association & National Council on Measurement in Education 2014).

Evidence which supports validity by response processes includes reliability. Reliability is the consistency of a measure (over time and across researchers) or the amount of error associated with a measurement (Chiang et al. 2015; Streiner et al. 2015). There are two types of reliability: intra-rater (test−retest) reliability and inter-rater reliability. Intra-rater reliability is the consistency of a measure over time. A rater using the scale should get the same results (for the same case of induction) every time they use it. The second type is inter-rater reliability, which is the extent of consistency between the results of different observers (Chiang et al. 2015). A scale that is not tested for reliability may produce inconsistent results when used among different raters or in different contexts.

Evidence which supports validity by test content (similar to content validity) includes instrument development. The domain intended to be measured − in this case, induction quality − must be identified, and expressions used in the scale should ideally be chosen by experts in the field or by the researcher after a thorough literature review of the topic (Holton et al. 2001; Boateng et al. 2018). In previous studies, careful, unbiased categorization followed by evaluation of the finished scale was also performed by groups of experts and is suggested as a first step to analyze validity (Melzack 1975; Holton et al. 2001; Boateng et al. 2018). The steps described above are often the stopping point for researchers, but at this stage of scale development, the items in the scale are too broad and still need to be analyzed for correlation to the domain (Chiang et al. 2015; Boateng et al. 2018).

The purpose of this study was to develop and begin establishing evidence for validity of an instrument to assess the quality of induction in dogs intended to be used in a research setting. Validity will be documented by response processes (reliability), test content (instrument building; content validity) and relationship to other variables (other scales which have been used for induction quality; criterion validity). The methods for this study were adapted from Holton et al. (2001), which followed the steps of a composite pain development scale from Melzack (1975).

Materials and methods

Part 1

An online survey was sent to veterinary anesthesia personnel, including Diplomates, residents and technicians, asking them to describe a perfect induction as well as to submit expressions and words that they associate with a good induction, a bad induction, and any comments to explain their submissions if needed. Participants were solicited by a recruiting email sent through the ACVA-L (American College of Veterinary Anesthesiologists) listserv. Clicking a link directed them to the survey, built using web-based survey software (Qualtrics, UT, USA). Characteristics of the respondents were not collected. The list of words and expressions was compiled and reviewed by one investigator. Items were removed according to the following criteria: similar expressions with the same meaning were combined, duplicate expressions were discarded. Items were compared with published words and expressions used to describe induction quality in dogs and, if there was an expression from the literature which was not captured by the collected items, it was added for consideration (Wolfe & Hofmeister 2021). The survey was approved by Auburn University Office of Human Research.

The list of items was reviewed and placed into categories created by four anesthesiologists (Diplomates of the American College of Veterinary Anesthesia and Analgesia) selected by the primary investigator (PI) from two universities ($n = 2$ for each institution). Each anesthesiologist was asked to categorize the items at their discretion, using their own words to name the categories. Each anesthesiologist categorized the items independently, then the PI and an undergraduate student researcher (KLW) reviewed each anesthesiologist's categorizations and determined a final set of categories. The final categories were chosen based on agreement and similarities among the anesthesiologists' submissions. Expressions requiring quantitative measurements (e.g. SpO_2 and tidal volume) were removed at this time because they would not be reliably observed or acquired in all induction scenarios. A final reduction of items was made by adjusting and combining related expressions (e.g. tachypnea and apnea combined to abnormal respiration) to make the items in each category more uniform and concise. Ultimately, expressions (e.g. gagging, vomiting) were combined into more concise items (e.g. gagging/vomiting) under categories (e.g. reflexive). The number of items per category was given no set minimum or maximum, after allocation and reduction, and each category could vary in number of items.

The same four veterinary anesthesiologists assigned a range of values to each of the items as they deemed appropriate, with 0 b-

eing absence of the behavior and higher values being increased severity/expression of the behavior (e.g. two anesthesiologists assigned vocalization as 0−1, whereas the other two assigned 0−2). The specific instructions were as follows: "We have combined all the variables you previously identified into five major categories, with each category having several variables in it. We now need to come up with a score range for each variable. For example, for 'Hypersalivation', should we score that 0/1 (not present/present) or 0/1/2 (not present/mild-mod/severe), or 0/1/2/3 (not present, mild, moderate, severe)? Should 'Excitation' be scored 0−1 or 0−3 or 0−10? So, for each variable, please put in the 'Score' cell the RANGE of values you think should be assigned for that variable".

The PI and the student researcher reviewed the four sets of scores, and a final set of scores was determined for each item. The scores from each anesthesiologist were compared alongside one another for each item, and agreements between the anesthesiologists determined the final score chosen (e.g. all four anesthesiologists gave sneezing a range of 0−1; therefore, 0−1 was chosen as the final score). If at least two anesthesiologists agreed on a score, that score was chosen as the final score for that item. In the event that all four scores were different, or there was a tie between two scores, the final score was decided upon by the PI. Instructions for using the composite scale were written in a way which was considered easy to follow and sent to the anesthesiologists for review and adjustments. This created the Auburn Induction Scale.

Part 2

A total of 39 video recordings of dogs during the process of induction were filmed at the Auburn University College of Veterinary Medicine. Client consent to use videos for research purposes was obtained for each dog. The dogs varied in breed and age and were administered varying premedication and induction medications for a variety of procedures. The videos showed the process of induction of anesthesia from the moment the induction drug was administered to the end of the process of intubation. Videos were recorded on a convenience basis when the anesthesiologists involved in the study were not on clinic duty. Out of 39 videos, 18 were selected because they contained clear visibility of the dog in the recording frame and minimal amount of personnel error (e.g. intubating with the wrong size tube). The videos were distributed to the same four anesthesiologists that evaluated expressions and allocated scores. They evaluated the quality of induction using the newly developed Auburn Induction Scale, a VAS, and a four-level SDS (0−3). The VAS and SDS scales were chosen to test alongside the composite scale as the most commonly used type of scales in current literature to measure quality of induction of anesthesia. Raters were instructed to watch the entire video once, then watch it again while scoring (Appendix A). The raters were given scoring sheets containing the three scoring systems in six different random orders (i.e. the order of the SDS, VAS and Auburn Induction scale changed for each video) so that they were not completed in the same order for every video. The scores for the Auburn Induction Scale were calculated by adding up the individual scores of each item. Raters were given written instructions on how to use the three scales prior to viewing the videos. The group of raters scored the videos again 3 months after the first scoring. The order of the videos was randomized using a computer-generated sequence (Microsoft Excel, WA, USA) before the first round of scoring and randomized again before the second round.

Statistical analysis

Normality was determined with the D'Agostino−Pearson method and examination of Q−Q plots. Intra-rater reliability was measured using a two-way random effects consistency single measure intraclass correlation coefficient (ICC). Differences between the first and second viewing were documented with a Wilcoxon signed ranks test for each rater. Inter-rater reliability was measured using a two-way mixed effects consistency single measure ICC. Absolute differences among raters were evaluated using a Friedman test. Linear regression was used to document the relationship between the Auburn Induction Scale and the SDS and between the Auburn Induction Scale and the VAS. Significance was set at $p < 0.05$.

Results

Part 1

The four-question listserv survey yielded 51 responses from ACVA-L members (Fig. 1). From the 51 responses, 326 expressions describing a perfect induction, 146 expressions describing a good induction, and 239 expressions describing a bad induction were extracted. After the duplicate terms were removed, there were 117 expressions describing a perfect induction, 69 expressions describing a good induction, and 117 expressions describing a bad induction. When like terms were combined (e.g. expressions such as 'no agitation' and 'excitement free at all times'), 133 total expressions across perfect, good, and bad remained (Appendix SA). No expressions were identified in the literature that the authors considered would add meaningfully to the list of expressions identified by the survey.

For category selection, there was general agreement among the anesthesiologists; one anesthesiologist organized the items by severity, whereas the other three organized the items by system [e.g. central nervous system (CNS)]. Chosen for the final scale were five categories: autonomic, reflexive, CNS, somatic and behavioral. Categories describing severity were not included. After the final reduction, made by combining related

Development of a scale to measure quality of induction of anesthesia in dogs

Responses to this survey will be used in the development of a scale to measure the quality of induction of anesthesia in dogs.

Describe a perfect anesthetic induction. Include descriptions of a dog's ideal behaviors at each stage of an induction:

List observable behavioral characteristics commonly seen in dogs during a good induction:

List observable behavioral characteristics commonly seen in dogs during a bad induction:

Any comments, further ideas:

Figure 1 Survey questions in an online survey distributed to the ACVA-L listserv. ACVA, American College of Veterinary Anesthesiologists.

expressions and removing expressions requiring quantitative measurements, 16 items were created incorporating the 133 expressions.

For 14 of the items, at least two anesthesiologists agreed on the scoring range (e.g. 0−2). For the two items that all four anesthesiologists disagreed on the scoring range to use (screaming and licking), the PI decided the final range for these two items (0−1 was chosen for both). The decision was made to remove two items (cardiac arrhythmias and abnormal hemodynamics) to maintain the consistency of descriptors that can be assessed without monitoring equipment.

Along with the addition of instructions for use, the final composite scale following these adjustments contained five categories (autonomic, reflexive, CNS, somatic and behavioral) and 16 items. Autonomic contained two items, reflexive contained five items, CNS contained three items, somatic contained two items and behavioral contained four items (Fig. 2).

Part 2

The mean ± standard deviation (SD) score for all videos for the Auburn Induction Scale from each rater (0−29 scoring range) ranged from 1.3 ± 1.7 to 5.1 ± 4.4. The median (interquartile range) scores from the SDSs (0−4 scoring range) ranged from 0 (0−1) to 1 (1−2). The mean ± SD score for VASs (0−100 scoring range) ranged from 8.8 ± 11 to 43 ± 28.

The ICC measuring intra-rater reliability were higher for the Auburn Induction Scale than the SDS or VAS for all raters, meaning the Auburn Induction Scale had highest agreement between the first and second rounds of scoring for each rater. The ICC for the SDS and VAS were similar for all raters except Rater 3, meaning the agreement between scores from the first to second round was similar for both the SDS and VAS (Table 1). The results of the Wilcoxon signed-ranks test showed a significant difference between the first and second round of scoring for two raters for the Auburn Induction Scale (Rater 1, $p = 0.001$; Rater 2, $p = 0.017$; Rater 3, $p = 0.17$; Rater 4, $p = 0.71$). There was no significant difference between the first and second rounds of scoring for the SDS for each rater (Rater 1, $p = 0.18$; Rater 2, $p = 0.41$; Rater 3, $p = 0.16$; Rater 4, $p = 0.66$). There was a significant difference between the first and second rounds of scoring for the VAS for one rater (Rater 1, $p = 0.73$; Rater 2, $p = 0.85$; Rater 3, $p = 0.012$; Rater 4, $p = 0.45$).

The ICC measuring inter-rater reliability showed the highest agreement between raters for the Auburn Induction Scale during both rounds of scoring, but there was a decrease in agreement from the first round to the second round. The SDS and VAS, while having lower agreement than the Auburn Induction Scale, increased in agreement from the first round to the second round (Table 2). The results of the Friedman test showed absolute differences between raters for both rounds of scoring for all three systems, with the exception of the Auburn Induction Scale at the first round (Table 3). There was a strong, significant relationship between the Auburn Induction Scale and the SDS ($p < 0.0001$, $R^2 = 0.70$) and the VAS ($p < 0.0001$, $R^2 = 0.78$).

Discussion

When conducting research, the goal is to produce the most consistent and accurate results possible. In veterinary medicine, one area that has lacked attention is how evaluation of induction of anesthesia is conducted. The simple scales in current use (SDS, VAS) are not consistent and have not been tested for validity or reliability (Wolfe & Hofmeister 2021). This study documents evidence for validity of the Auburn Induction Scale by building the instrument in an appropriate manner, documenting reliability, and testing relationships with other variables.

A substantial number of individuals ($n = 51$) contributed expressions to be considered for the instrument, and a

Write a score for all of the expressions in each category based on observed behaviors. Each expression has a different range of scores, but for each expression 0 is complete absence of the behavior, and the highest number is the worst exhibition of that behavior.

Category	Expressions		Score
Autonomic	abnormal respiration (tachypnoea, apnea, panting)	0-3	
	hypersalivation	0-1	
Reflexive	defecation/urination	0-1	
	gagging/vomiting	0-2	
	coughing/gasping	0-2	
	sneezing	0-1	
	swallowing/laryngospasm	0-2	
CNS	excitation (dysphoria, amnesia, disinhibition, seizure)	0-3	
	prolonged time to unconsciousness	0-2	
	rough transition from consciousness to unconsciousness (not smooth)	0-2	
Somatic	poor muscle posture (tonic-clonic movement, opisthotonos)	0-2	
	muscle activity (excessive shivering, fasciculation, myoclonus, muscle twitching, muscle rigidity, increased jaw tone, increased neuromotor activity, head shaking, thrashing, paddling, struggling, chewing)	0-3	
Behavioral	aggression (ears back, growling, biting/snapping, agitation)	0-1	
	vocalization	0-2	
	screaming	0-1	
	licking	0-1	

Figure 2 Composite scale created through the processes described in the study design for scoring induction of anesthesia in dogs. Total score 0−29. CNS, central nervous system.

Table 1 Intra-class correlation coefficients (ICC) and 95% confidence interval (95% CI) of four rater scores compared at two evaluations made at least 3 months apart using the Auburn Induction Scale (Scale), simple descriptive scale (SDS), and visual analog scale (VAS). This documents intra-rater reliability. ICC can be interpreted as poor (<0.50), moderate (0.50−0.75), good (0.75−0.90) or excellent (>0.90) (Koo & Li 2016)

Rater	Scale	SDS	VAS
1	0.71 (0.37−0.88)	0.69 (0.35−0.87)	0.63 (0.24−0.84)
2	0.76 (0.47−0.90)	0.70 (0.36−0.88)	0.69 (0.35−0.87)
3	0.86 (0.66−0.95)	0.69 (0.33−0.87)	0.82 (0.58−0.93)
4	0.89 (0.73−0.96)	0.74 (0.42−0.89)	0.71 (0.37−0.88)

subsequent panel agreed on the majority of items to be put into the instrument. The expressions were chosen by a methodical process that is consistent with the process for documenting validity by test content (Holton et al. 2001). Nonetheless, some

Table 2 Results of analysis of scores from four raters using the Auburn Induction Scale (Scale), simple descriptive scale (SDS) and visual analog scale (VAS) at baseline (Round 1) and at least 3 months later (Round 2) when scoring videos of 18 dogs during induction of anesthesia. Data are presented as intra-class correlation coefficient (ICC) and 95% confidence interval, representing agreement. The *p*-value represents the result of statistical testing comparing the values among raters with Friedman test. This documents inter-rater reliability. ICC can be interpreted as poor (<0.50), moderate (0.50−0.75), good (0.75−0.90) or excellent (>0.90) (Koo & Li 2016)

Scoring system	Time points			
	Round 1	*p*	Round 2	*p*
Scale	0.72 (0.53−0.87)	0.178	0.66 (0.45−0.83)	<0.0001
SDS	0.57 (0.34−0.78)	0.042	0.64 (0.42−0.82)	0.008
VAS	0.55 (0.32−0.77)	0.0004	0.64 (0.43−0.82)	<0.0001

Table 3 Results of statistical testing for differences among four raters using the Auburn Induction Scale (Scale), simple descriptive scale (SDS) and visual analog scale (VAS) at baseline (Round 1) and at least 3 months later (Round 2) when scoring videos of 18 dogs during induction of anesthesia. Differences among raters were determined using a Friedman test and are presented as *p* values

Scoring system	Round 1	Round 2
Scale	0.09	<0.0001
SDS	0.002	0.0002
VAS	0.0004	<0.0001

items on the scale may not contribute meaningfully to evaluation of induction, and some items, which may be valuable, may have been omitted. Further refinement and continuing to collect evidence for the validity of the Auburn Induction Scale is warranted.

The Auburn Induction Scale had the highest reliability for both rounds of scoring, whereas the SDS and VAS had nearly the same ICC values. This indicates that the Auburn Induction Scale was the most consistent in scores given by each rater, that is the highest intra-rater reliability. The inter-rater reliability did increase slightly for the SDS and VAS between the first and second rounds, whereas the Auburn Induction Scale's ICC decreased slightly. This increase in agreement for the SDS and VAS could be attributed to the rater's familiarity with the scales, allowing them to use the scale more efficiently, whereas the decrease in agreement for the Auburn Induction Scale may result from not following instructions, changes in the rater's understanding of the expressions or lack of attention to detail during the second round. There were statistically significant differences in scores between raters for both rounds of scoring for the SDS and VAS, but only for the second round of scoring for the Auburn Induction Scale. This indicates that, although the Auburn Induction Scale may require more training before usage, overall it performed better than both the SDS and VAS for reliability between raters. Nonetheless, the ICC were relatively similar among systems, and it is possible the Auburn Induction Scale does not provide meaningfully better reliability than the VAS and SDS.

The 95% confidence interval (CI) for all scales was relatively wide. No other study has documented ICC for induction quality scoring, so this variability may be attributable to the nature of scoring induction (i.e. it is significantly more subjective than other scoring systems). Alternatively, it is possible that video recording led to greater variability (Copeland et al. 2017) or that the individual raters were more widely variable than expected or need more thorough training.

SDS and VAS scales have been reported to have significant variability when tested for reliability in pain assessment in dogs (Holton et al. 2001). Like pain, quality of induction of anesthesia is multidimensional and, therefore, requires a multidimensional instrument to accurately measure it. Induction of anesthesia has no clearly defined starting and stopping point, and so a unidimensional scale is subjective and can yield different results from researchers watching the same induction. Scales such as SDS and VAS are widely used and familiar to researchers; however, they cannot fully capture the intended measurement (McCoach et al. 2013). Nonetheless, they are the existing scales that have been used to capture this phenomenon and the relationship that the Auburn Induction Scale has with the SDS and VAS provides further evidence of validity of the Auburn Induction Scale.

There were several limitations to this study. The scoring of induction from video recording may be inferior to scoring in person. A study evaluating the reliability of using video recordings to assess recoveries found that videos were not as reliable as live evaluation (Copeland et al. 2017). Variables such as hypersalivation and respiration cannot be seen easily on video. To account for this, the raters were allowed to view each video twice before scoring them, but there may have been behaviors that were not captured by the video at all. There was a potential for bias during scoring using the same order of scoring systems repetitively (i.e. if a rater assigns an SDS of 2, that choice might anchor their scores for the other scales). This was limited by giving the raters the three scoring systems in different orders for each video. The dogs captured in the videos all exhibited fairly smooth induction behavior, so it is unknown how the Auburn Induction Scale performs across a wide range of induction qualities. Only four raters with similar experience (i.e. board-certified anesthesiologists) were used, so it is unknown how the Auburn Induction Scale performs in the hands of inexperienced raters and across a wider range of raters. The raters were not provided with training or practice before performing scoring, which may improve the reliability. The raters were also the ones who assigned scores to the instrument initially, so it is possible they acquired familiarity with the instrument, which may have affected results. This seems unlikely, given that agreement decreased between the first and second scoring.

A multidimensional tool for evaluating the quality of induction of anesthesia has been constructed using psychometric principles to show evidence for validity, something that has not yet been presented in the literature. In a research setting, widespread use of this scale may be helpful in increasing the accuracy of data and improving agreement between studies. A validated scale would enable comparison of induction drugs and techniques.

Conclusions

The results of this study have yielded a composite scale that is more reliable between and among raters than a unidimensional scale. Future work to establish further evidence for

validity of the scale includes using it to evaluate dogs with an expected large range of induction quality, evaluating the scale in depth using qualitative analysis, and evaluating its performance when used by a variety of raters.

Author's contributions

KLW and EHH: study design, data collection, data analysis, preparation of manuscript. SCC-P, RR and JQ: data collection, data analysis, preparation of manuscript. All authors read and approved the final version of the manuscript.

Conflict of interest statement

The authors declare no conflict of interest.

References

Amengual M, Flaherty D, Auckburally A et al. (2013) An evaluation of anaesthetic induction in healthy dogs using rapid intravenous injection of propofol or alfaxalone. Vet Anaesth Analg 40, 115–123.

American Educational Research Association (2014) American Psychological Association, & National Council on Measurement in Education. In: Standards for Educational and Psychological Testing (1st edn). American Educational Research Association, USA. pp. 11–26.

Boateng GO, Neilands TB, Frongillo EA et al. (2018) Best practices for developing and validating scales for health, social, and behavioral research: a primer. Front Public Health 6, 149.

Chiang IA, Jhangiani R, Price PC (2015) Research Methods in Psychology (2nd Canadian edn). BC Campus, Canada.

Copeland JE, Hofmeister EH, Brainard BM, Quandt JE (2017) Reliability of video recordings to evaluate quality of anesthesia recovery in dogs. Vet Anaesth Analg 44, 409–416.

Holton L, Reid J, Scott EM et al. (2001) Development of a behaviour-based scale to measure acute pain in dogs. Vet Rec 148, 525–531.

Koo TK, Li MY (2016) A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 15, 155–163.

Melzack R (1975) The McGill Pain Questionnaire: major properties and scoring methods. Pain 1, 277–299.

McCoach DB, Gable RK, Madura JP (2013) Instrument Development in the Affective Domain (3rd edn). Springer, USA. pp. 33–82.

Raykov T, Marcoulides GA (2011) Introduction to Psychometric Theory. Routledge, USA.

Streiner DLN, Norman GR, Cairney J (2015) Health Measurement Scales: A Practical Guide to Their Development and Use (5th edn). Oxford University Press, UK.

Tamura J, Hatakeyama N, Ishizuka T et al. (2016) The pharmacological effects of intramuscular administration of alfaxalone combined with medetomidine and butorphanol in dogs. J Vet Med Sci 78, 929–936.

Wolfe KL, Hofmeister EH (2021) Scoping review of quality of anesthetic induction and recovery scales used for dogs. Vet Anaesth Analg 48, 823–840.

Received 15 March 2022; accepted 25 August 2022.

Available online 8 September 2022

Supporting Information.

Additional Supporting Information may be found in the online version of this article: https://doi.org/10.1016/j.vaa.2022.08.010.

Appendix SA. Descriptions of behaviors in dogs during induction of anesthesia

Appendix A. Instructions for scoring videos

Reviewers will score induction of anesthesia in 18 dogs to determine intra- and inter-observer reliability. Instructions are:

Watch each video up to two times. Complete the following three scales for each video.

SDS (simple descriptive scale)

Circle the number 0–3 according to the following scale: -

Score	Description
0	Smooth with no resistance; perfect, smooth uncomplicated induction.
1	Slight resistance but smooth.
2	Mild-moderate resistance; moderate excitation present during transition from conscious to anaesthetized.
3	Unacceptable: Marked excitation and struggling, and/or aggression.

VAS (visual analogue scale)

Mark a position on the line between 0 and 100 mm, with 0 mm being the smoothest induction possible and 100 mm being the worst induction possible.

Composite scale

Write a score for all of the expressions in each category based on observed behaviors. Each expression has a different range of scores, but for each expression 0 is complete absence of the behavior, and the highest number is the worst exhibition of that behavior.