Refinement and Revalidation of the Equine Acute Abdominal Pain Scale (EAAPS)

Sutton, G.A.* and Bar, L.

Koret School of Veterinary Medicine - Veterinary Teaching Hospital, The Hebrew University of Jerusalem, Robert H Smith Faculty of Agriculture, Food and Environmental Sciences, POB 12, Rehovot, Israel 76100.

* **Corresponding author:** Dr. Gila Sutton, DVM, M.Sc, Ph.D, Koret School of Veterinary Medicine - Veterinary Teaching Hospital, The Hebrew University of Jerusalem, Robert H Smith Faculty of Agriculture, Food and Environmental Sciences, POB 12, Rehovot, Israel 76100. Tel.: 972 3 968 8507, Fax: 972 8 946 7940, Email address: gila.sutton@mail.huji.ac.il

ABSTRACT

Assessment of pain is vital for colic treatment. The purpose of this study was to revalidate a refined version of the behaviour-based Equine Acute Abdominal Pain Scale (EAAPS). Based on an earlier study, behaviours in the scale were removed or replaced. Ten behaviours remained. For revalidation, forty films of horses with colic were presented by computer-generated random order to two randomly-assigned groups of equine veterinarians. One group (n=8) scored the severity of pain demonstrated in the films by utilizing a numerical rating scale (NRS) and one group (n=7) with the refined version of the EAAPS. Intra-rater reliabilities of the EAAPS and of the NRS were comparable based on Limits of Agreement. The inter-rater reliability of the EAAPS was significantly improved compared to the NRS (NRS; Intraclass correlation (ICC) = 0.6 (95% Confidence Interval (CI); 0.5-0.8) and EAAPS; ICC = 0.88 (95%CI: 0.8-0.9)). Face validity was 71% (95% CI; 29-96) in support of the EAAPS. The two scales showed substantial convergent validity (weighted kappa of 0.73 (95%CI; 0.58-0.88). The predictive validity of the EAAPS scale was similar to the NRS (AUC of EAAPS; 0.75 versus NRS; 0.78 for mortality; AUC of EAAPS 0.76 versus NRS of 0.83 for treatment modality) and the ability to discriminate between extreme groups of either control horses versus cases or by extreme groups defined by NRS scores of 0-2 versus 3-5 was excellent (AUC 0.99 and 0.955, respectively). In summary, revalidation of the refined EAAPS was necessary and was found to be highly reliable and comparatively valid.

Keywords: Equine; Colic; Pain Assessment; Validation; Facial Validity.

INTRODUCTION

Colic, a major cause of death in horses (1), is a condition characterised by pain but until recently, there has been no standardised pain severity scale for use in these cases. All of the earlier scales were designed for use in experimental situations, from the first scale published by Muir and Robertson in 1985 (2) through a final modification of it by Boatwright *et al.,* 1996 (3). These early scales were never validated. Beginning this century, however, there have been a number of pain scales developed for post-operative pain (4, 5, 6) as well as a scale based on equine facial expression, the "Horse Grimace Scale" (7). Some have undergone initial validation (6, 7), however, all of these more recent scales have been developed for post-operative pain and most involve complex composite scales rather than simple clinical indices (4-6, 8). In contrast, the Equine Acute Abdominal Pain Scales (EAAPS) was designed specifically for evaluation of the severity of acute colic pain for use in adult horses under clinical conditions unrelated to surgery (see Appendix 1). As a simple clinical index rather than a composite scale, it

is based entirely on behaviours, and includes only a small number of items, requiring no arithmetic calculations, in order for it to be easy to use at horse farms as well as at a referral clinic (9, 10).

Originally the EAAPS was developed as two contending behaviour-based scales; EAAPS-1 and EAAPS-2. They each included 12 identical behaviours with scores from 1 to 5 assigned to each behaviour. The scales differed in that EAAPS-1 had one score assigned to each behaviour and the EAAPS-2 had one or two scores per behaviour depending on the intensity of the behaviour demonstrated (9).

The EAAPS scales were subsequently validated (10), based on principles of clinimetrics, by demonstrating reliability and validity (11, 12). Reliability is the extent to which a scale yields the same results on repeated trials and validity conveys whether the scale measures what it aims to measure. Ideally, validity involves comparison to a gold standard but when there is no gold standard, as for the measurement of pain, alternate types of validity are required, such as face validity and various types of construct validity including convergent, discriminative and predictive validities (13).

Following the validation study comparing and contrasting the two EAAPS scales to each other and to a global pain scale, the EAAPS-1 scale was chosen due to its superior reliability (10). Then a revised version of the EAAPS was constructed by making changes based on the two studies. The number of behaviours was reduced to ten. Three behaviours; depression, weight shifting and collapse were removed due to their relatively poor performance regarding agreement of observers as to the presence of these signs in film clips (9, 10). Lip curling (Flehman) was introduced, as it had shown good agreement between observers (9) and the scores were adjusted in order to reduce the number of behaviours assigned a score of 3, since it appeared to be an overly common score for the EAAPS-1 in comparison with the other scales (9, 10).

The overall purpose of this study was to validate the revised version of the EAAPS scale for its clinimetric properties. When scales measure hypothetical constructs, such as pain, validation is an ongoing task (13). The specific objectives were to assess the inter- and intra-rater reliabilities and three general types of validities: face validity, construct validity and predictive validity of the scale.

MATERIALS AND METHODS

Study design

An evaluation group of equine practitioners viewed films of horses exhibiting signs of colic (or of control horses) via a specially designed website (see below). The films were shown in random order by the website program. Each participant viewed a total of 41 films and scored the films using one of two scales; the revised EAAPS scale or a global numerical rating scale (NRS) as a control group.

Films

Films clips were chosen purposively from films previously prepared (9), from 28 cases of colic, over 1 year of age, (cases) presented to the Koret School of Veterinary Medicine, Veterinary Teaching Hospital, Equine Department. Written approval was obtained from the Internal Ethics Review Committee of the Veterinary Teaching Hospital prior to filming. An attempt was made to select films representing the entire breadth of pain spectrum and those in which behaviours were clearly demonstrated. Six cases provided two films each giving a total of 34 digital film clips of horses with colic. Six films of control horses, not suffering from colic, who were hospitalised for fertility treatment, were added (filmed by LB) to raise the total number of films from 34 to 40. The median length of the film clips was 27 seconds (interquartile range; 19-46 seconds). One of the 34 films of colic cases was randomly selected to be shown twice for assessing intra-rater reliability.

Participants

Eighteen equine practitioners were recruited by general electronic mail sent to 59 local equine practitioners in Israel. Ten letters were returned by the postmaster. Of those not returned, 18 (36%) responded and agreed to participate. All participants were asked questions regarding themselves, their veterinary education, specialization training and experience treating horses. The assessors were assigned to one of two independent groups (n=9); a control group and a test group, by block randomization using the EXCEL program. The participants were not trained to use the EAAPS since the descriptors are commonly used in the scientific literature, the participants were equine practitioners and the descriptors were defined (Appendix 1b). The untrained participants of the test group scored the severity of pain the horses were showing in the films using the revised EAAPS scale and those of the control

group provided a global assessment of pain by using a 6-point numerical rating scale (NRS) in which '0' indicated 'no pain', '1' - 'mild pain', '2' - 'mild to moderate pain', '3' - 'moderate pain', '4' - 'moderate to severe pain', and '5' indicating 'severe pain'.

Face Validity

Face validity of the EAAPS scale was evaluated by asking the participants of the test groups, upon completion of the scoring, to what degree they agreed with the following statement: 'The EAAPS scale is a valid scale for the assessment of acute abdominal pain in the adult horse', using four response options (strongly agree, agree, disagree, strongly disagree) and by asking whether the overall impression was that the EAAPS scale made clinical sense with a yes/no response option.

Behaviour descriptor evaluation

Two new behaviour descriptors; "lip-curling"(14) ("Flehman") (15) and "crouching"(16), had not been previously evaluated for inter-rater agreement. Therefore, participants in the test group were asked to indicate whether these behaviours were demonstrated in each film. The response option was dichotomous (Yes/No).

Website Program

For the entire process, the assessors viewed, managed and scored the films using a structured, web-based course management system (Moodle) modified for this study (http://muddle.cs.huji.ac.il/mu11). Access to the website required a username and password assigned individually to each participant.

Sample size calculation

Sample size was calculated based on intra-class correlation coefficients (ICC) for the reliability analysis. At an alpha of 0.05 and power of 80%, 40 films would be needed with at least 5 observations per film in order to differentiate an expected ICC of 0.8 from an ICC of 0.67, as obtained, respectively, for the EAAPS-1 scale and the NRS scale in the earlier study (10). The calculation was made in WinPEPI as well (WinPEPI (version 2.91), PairsEtc., Sample Size S6, copyright J.H. Abramson).

Statistical analysis

Point estimates and Fisher's 95% confidence intervals (95% CI) were calculated (WinPEPI (www.brixtonhealth.com/pepi5windows.html)). Spearman's ranked correlation (rho) was used to evaluate association between age of horse and median scores of films. Non-parametric independent t-tests and Kruskall-Wallis were used to evaluate associations between sex/breed and median scores of films. Inter-rater reliability was tested by intraclass correlation (ICC), utilising the two-way random effects model, for absolute agreement and single measures (McGraw and Wong A1 Model) (17, 18) (SPSS 18, IBM, USA). Interpretation of the ICC was as follows; values above 0.75 were regarded as excellent reliability, 0.4 as good reliability (19). Intra-rater reliability was evaluated by Limits of Agreement (LOA) (20) between two scores given to identical films. Individual behaviour descriptors were evaluated for bias and for inter-rater agreement by multirater, multicategory kappa coefficient (Software by Mr. William Sears, Ontario Veterinary College, Guelph, Ontario, Canada). Interpretation of the kappa coefficient was based on Landis and Koch, 1977 (21).

Face validity was evaluated as frequency of endorsement with 95% confidence intervals for each response option (either; strongly agree, agree, disagree, strongly disagree, or; yes, no) (WinPEPI). The responses of the two questions were compared for reliability. Several constructs were used to evaluate Construct validity; convergent, discriminant and predictive. Convergent validity of the EAAPS scale in comparison to the global assessment of pain (NRS) was assessed by agreement between rounded median scores of the two scales as expressed by weighted kappas (http://www.vassarstats.net/kappa.html). Discriminant validity was expressed as the ability of the median EAAPS scores to discriminate between two types of extreme groups. First, colic versus control horses by Fisher's Exact Test (SPSS 18) and second, by ROC curve to predict severe versus mild pain as defined by the NRS scores. Severity of pain demonstrated in each film was assessed by median NRS scores of 0, 1 or 2 indicating mild pain, compared to scores of 3, 4 or 5 indicating severe pain. The cut-off value of 3 or more for the NRS scale as indicating severe pain was chosen based on the ROC curve of the mean NRS scores to predict death (area under the cure (AUC) = 0.784; 95%CI 0.633-0.934) or treatment outcome of surgery or euthanasia versus no treatment or medical treatment (AUC=0.833; 95% CI 0.705-0.962). Predictive validity of each scale was evaluated by comparing the frequency of median EAAPS scores to mortality outcome (alive or dead) and to treatment modality (controls, medically treated, sur-

gically treated or euthanatized) by Fisher's Exact and chi-square for trend tests (SPSS 18). The predictive validity of the scales was also assessed by ROC curve for treatment modality (no treatment or medical treatment versus surgical treatment or euthanasia) (SPSS 18). The frequency distribution of the rounded median scores (0 to 5) over all of the films for each scale (EAAPS and NRS) were compared by Fisher's exact test (SPSS 18).

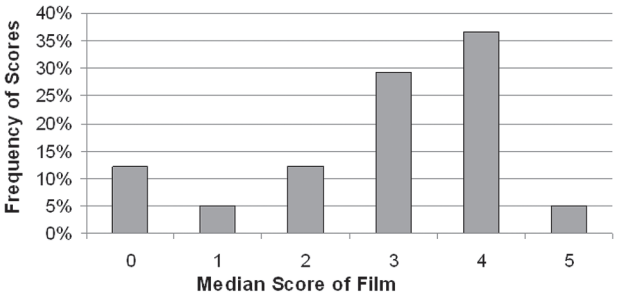

Figure 1: Frequency distribution of the severity of pain over the films (*n*=40) based on the median of the numerical rating scale (NRS) scores of each film.

RESULTS

Films

Characteristics of the horse population of the films can be found in Table 1. No statistically significant association was found between the median EAAPS or NRS scores and sex, breed or age of the horses, nor between scores and participants. The correlation between the NRS scores and age was statistically significant ($P < 0.05$) but small to moderate in strength (*Rho* = -0.364). The frequency distribution of the median NRS scores for the 40 films can be seen in Figure 1.

Participants

The response rate was 83% (15/18) (7 in the test group, 8 in the control group). The participants were general practitioners except for two; a board-certified theriogenologist and an anaesthetist. Additional characteristics of the participants appear in Table 2.

Reliability

The EAAPS scale demonstrated superior inter-rater reliability in comparison to the NRS scale ($P < 0.05$). The ICC of the EAAPS scale was 0.88 (95%CI: 0.83, 0.93) compared to the NRS; 0.60 (95%CI: 0.51, 0.79). The intra-rater reliability of the EAAPS scale was similar to the NRS based on limits of agreement (LOA) (Figure 2a & 2b). Both the EAAPS scale and the NRS scale demonstrated no difference in the scores of identical films in 5 of 7 (71%) and in 6 of 8 (75%) cases, respectively.

Table 1: Characteristics of the horse population of the films (*n*=40); median (IQR=interquartile range) or frequency (*n*) and percentage.

Characteristic		Median	IQR
Age (years)		6	(3-11)
Heart rate (beats per minute)		44	(40-60)
		Frequency (n)	**Percentage**
Sex	Mares	27	68%
	Stallions	4	10%
	Geldings	9	22%
Breed	Arabians	17	42%
	Quarterhorses	10	25%
	Grade	5	13%
	Others	8	20%
Treatment	Medical	16	40%
	Surgical	16	40%
	Euthanatised	2	5%
	Controls	6	15%
Mortality	Died	7	18%

Table 2: Demographic characteristics of participants (*n*=15); the number in each category (percentage)

Characteristic	Number (%)
Sex	
Male	9 (60)
Female	6 (40)
Clinical training level	
Internship	8 (53)
Residency	1 (7)
Neither	6 (40)
Percentage of practice equine	
>=90%	12 (80)
<90%	3 (20)
Number of years in practice	
Median 10 years (IQR*=12) Range 2-31 years	
<5 years	5 (33)
Participate in research	**6 (40)**

* IQR = interquartile range

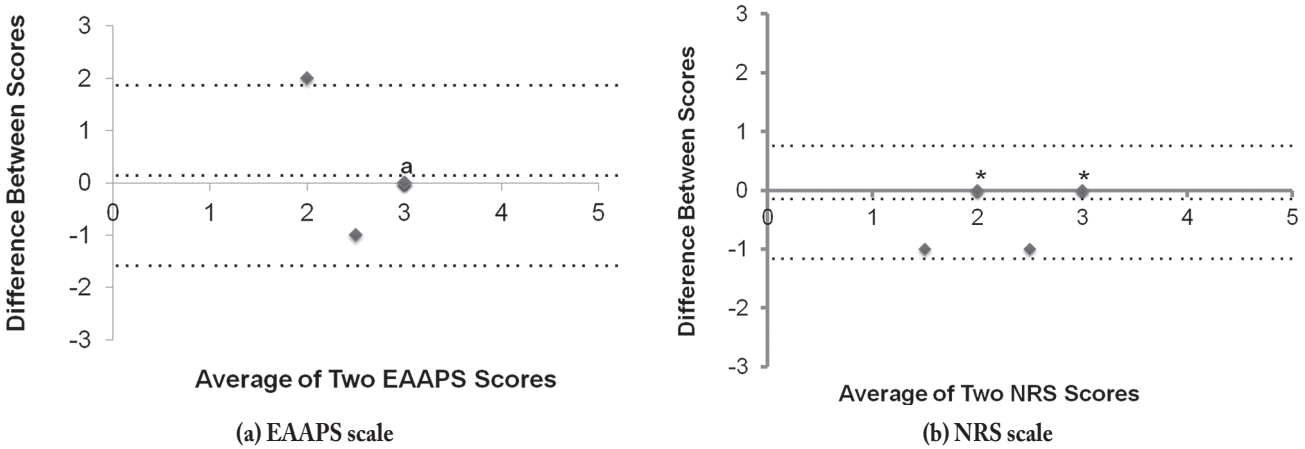

Figure 2: Intra-rater reliability for each of the scales as shown by 95% Limits of Agreement (LOA); (a) EAAPS scale, and (b) NRS scale. (a) The point marked by the letter 'a' denotes 5 points in the identical location (total n=7 points) and each asterisk in (b) denotes 3 points in the identical location (total n=8 points)

Previously Unevaluated Behaviours

Agreement between observers regarding "lip curling" demonstrated substantial agreement with a kappa coefficient of 0.77 (95%CI; 0.60-0.95) with insignificant bias. "Crouching" demonstrated fair agreement with a kappa of 0.25 (95%CI; 0.04-0.47) with significant bias (21).

Validity

Face validity was demonstrated by 5 of 7 (71%; 95%CI: 29-96) equine practitioners agreeing or strongly agreeing that the EAAPS scale was a valid scale for the assessment of acute abdominal pain in adult horses. The same practitioners replied in the affirmative that the EAAPS scale is valid in response to the second question. Convergent validity of the EAAPS scale was substantial when compared to the NRS (weighted kappa 0.73; 95%CI 0.58-0.88). As evidence of discriminant validity, the EAAPS scale highly discriminated between extreme groups, as defined either by the severity of the NRS scores or by comparing the scores of cases to controls. When compared to severity of pain demonstrated in the film, based on the NRS score, the AUC for the EAAPS scale to predict severe pain was 0.955 (95%CI: 0.87-1.0). When comparing cases to controls, 5/6 films of control horses received a median EAAPS score of 0 and 1/6 received a median score of 1, while all cases of colic received median scores greater than zero ($P < 0.00001$) (Figure 3). Predictive validity was demonstrated as a significant association between the frequency of pain scores in each treatment modality group (none, medical, surgical or euthanasia ($P = 0.001$ Fisher's Exact; $P = 0.00001$ chi-square for trend) but not between the frequency of pain scores and death ($P = 0.214$ Fisher's Exact) (Figures 4a & 4b). The AUC of the ROC curve for the outcomes of mortality or treatment modality for the EAAPS scales were comparable to the NRS scale (Table 3).

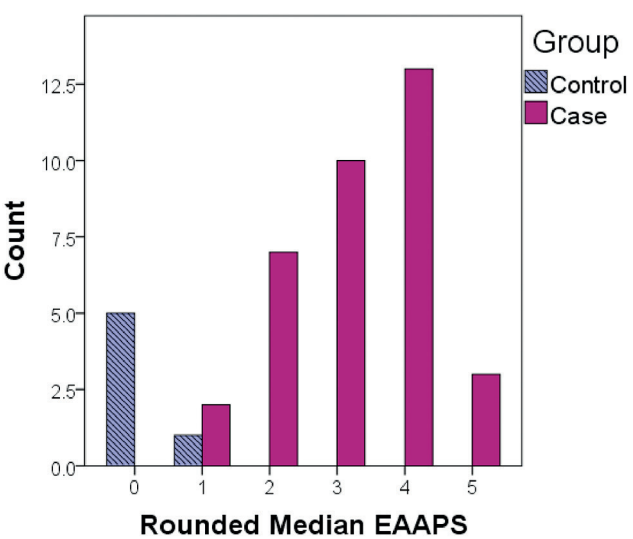

Figure 3: Discriminant reliability (extreme groups): Comparing median EAAPS scores given to cases of colic versus control horses without colic ($P < 0.00001$).

Scale comparison

The frequency distribution of the rounded median scores over all of the films varied significantly with the scale used to assess them (Fisher's Exact $P < 0.001$). The agreement was stronger at the extremes of no pain (0) and severe pain (5) and weaker over the mid-range where the EAAPS scale scored more films as 1 and 2 and less as 3 and 4 compared to the NRS scale (Figure 5).

DISCUSSION

The Equine Acute Abdominal Pain Scale (EAAPS) is the only pain scale developed and validated for acute abdominal pain prior to colic surgery. Revalidation is important since pain is not biologically measureable and validation depends on contructs or mini-theories that may not be true under differing circumstances and when changes are made (13). This validation was also carried out on a different population of observers than the previous validation as the equine practitioners who participated in this study were from Israel and in the previous study they were from Europe and the United States of America (10).

The better of two EAAPS, which had been vigorously constructed and validated, and had shown excellent reliability and adequate validity, had, nonetheless, had weaknesses identified in the previous validation study (9, 10). Alterations were made in order to remedy the weaknesses, and the revised version underwent revalidation in this study.

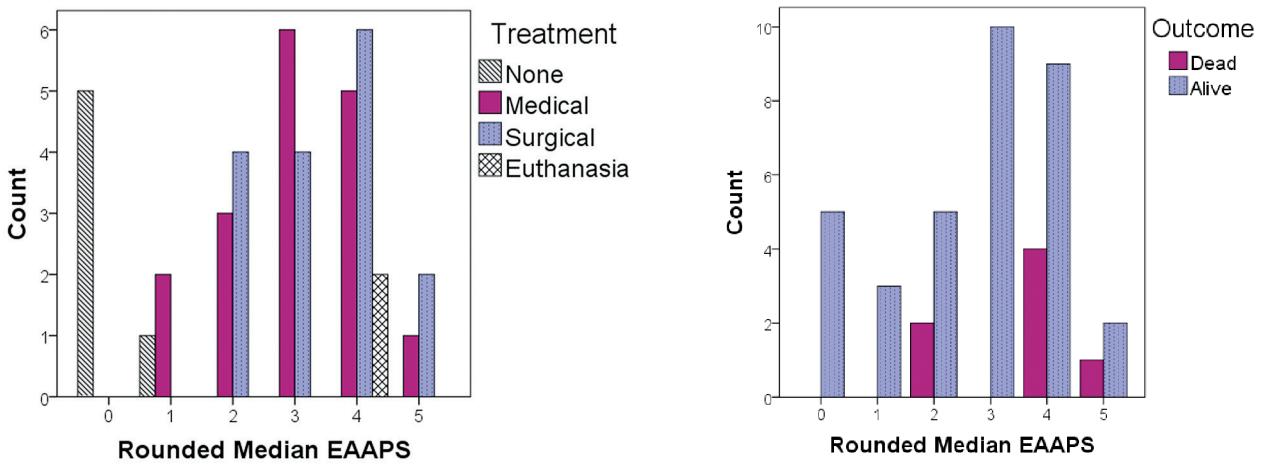

Figure 4: Predictive validity: Association of median EAAPS scores with; (a) Treatment modality, and (b) Mortality

Table 3: Predictive validity based on ROC curves (AUC (Area under the curve), 95% CI (confidence interval), cut-off values with corresponding sensitivity and specificity)

	AUC (95%CI)	Cut-off Value	Sensitivity	Specificity
Mortality				
EAAPS	0.75 (0.54-0.96)	3.93	71%	85%
NRS	0.78 (0.63-0.93)	3.36	86%	68%
Treatment				
EAAPS	0.76 (0.62-0.91)	2.92	78%	65%
NRS	0.83 (0.7-0.96)	2.64	94%	70%

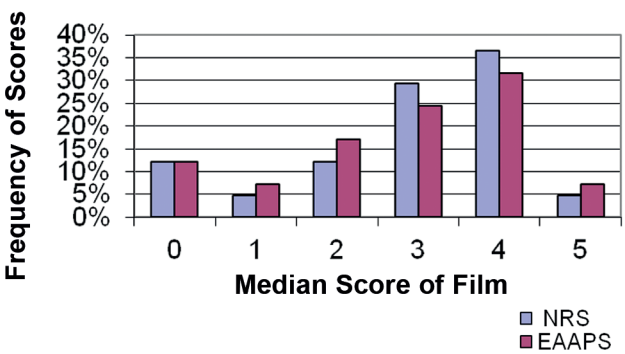

Figure 5: Comparison of the frequency (%) of films assigned each score (0-5) (median over all the observers) by each of the scales.

The revised version of the scale is an improvement because there are fewer behaviours (10 versus 12) which is preferable for clinimetric scales (22, 23), it includes a new behaviour that demonstrated substantial agreement (21) between observers (lip curling), and excluded behaviours that had demonstrated poor agreement between observers (depression, weight shifting and collapse). In this study, "crouching", a term that was introduced in order to provide an alternative descriptor for "attempting to lie down", surprisingly showed only fair agreement between observers (21). This finding justifies the removal of this term from later versions of the EAAPS since the agreement was not strong, although perhaps the observers in this study, not being native English speakers, were not clear about the meaning of the term. This theory is supported by the significant bias which may indicate that only certain people understood the word. Lip curling, which was added in a similar manner in order to explain Flehman , which itself showed excellent agreement in the earlier study (9), may be a more commonly used term (14), and therefore, may have been more universally recognised among the observers which may explain the substantial agreement and lack of bias it obtained in this study.

As in the previous validation, the inter-rater reliability of the new version was very high. It is considered excellent (19) and compares favourably to pain scales developed in young children aged 5 or younger (24) and in critically ill and cognitively impaired children (25) as well as in comparison to the earlier version of the EAAPS (EAAPS-1) (10).

In this study, the intra-rater reliability of the EAAPS scale was not better than the numerical rating scale (NRS), with a difference in one observer and in one score only. In future studies, more than one repetition of the film will be used for evaluating intra-rater reliability, particularly if the number of observers is relatively low, as in this study.

The revised version of the EAAPS scale showed adequate validity in comparison to the NRS as did the earlier EAAPS-1 scale (10). The face validity, demonstrating overall endorsement of the scale, had excellent intra-rater reliability in the responses since the two questions pertaining to face validity produced identical results. This, however, may have been influenced by the fact that the questions were placed in adjacent positions in the questionnaire. Regarding construct validity, all of the various types (convergent, predictive and extreme groups) demonstrated higher point estimates than the earlier version, however since these were different studies, they could not be statistically compared (10).

Based on the distribution of the films compared to the NRS scale (Figure 5), the scale is an improvement since the earlier version had a preponderance of films with the score of 3 (10) and in this study, the distribution of the film scores more closely paralleled those of the NRS scale, which is currently the accepted method that pain severity is assessed in cases of colic.

Limitations of this study include technical aspects of films that make it difficult at times to observe the behaviours needed to score the level of pain the horse is demonstrating. The film clips were short in order to enhance compliance of the participants, however, perhaps too short to assess severity of pain. On the other hand, pain is dynamic and perhaps longer films would encompass different levels of pain in the same film.

Future studies should evaluate the usability and feasibility (26) of the scale in a prospective, real-time study.

CONCLUSIONS

In this study, a revised version of the EAAPS scale was validated and shown to have excellent inter-rater reliability and comparable validity but appeared improved since the distribution of the scores appears closer to the distribution of the global rating scale (NRS) than the earlier version. Future studies are needed to demonstrate usefulness in the field in a prospective study.

CONFLICT OF INTEREST STATEMENT

Neither of the authors of this paper has a financial or personal relationship with other people or organizations that could inappropriately influence or bias the content of the paper.

ACKNOWLEDGEMENTS

We would like to thank Ms. Tali Bdolach-Avram, Mr. William Sears and Dr. Hillary Voet for statistical consultation, Chana Slutzkin of the System Group of the School of Computer Science and Engineering of the Hebrew University, for website management and Dvorah Weisman of the Computer Unit of the Hebrew University, The Robert H. Smith Faculty of Agriculture, Food and Environment for technical support. We would also like to thank the equine practitioners who participated in the study as well as Dr. Amir Steinman who made this project possible.

REFERENCES

1. Traub-Dargatz, J.L., Kopral, C.A., Seitzinger, A.H., Garber, L.P., Forde, K. and White, N.A.: Estimate of the national incidence of and operation-level risk factors for colic among horses in the United States, spring 1998 to spring 1999. J. Am. Vet. Med. Assoc. 219: 67-71, 2001.
2. Muir, W.W. and Robertson, J.T.: Visceral analgesia: effects of xylazine, butorphanol, meperidine, and pentazocine in horses. Am. J. Vet. Res. 46: 2081-2084, 1985.
3. Boatwright, C.E., Fubini, S.L., Grohn, Y.T. and Goossens, L.: A comparison of N-butylscopolammonium bromide and butorphanol tartrate for analgesia using a balloon model of abdominal pain in ponies. Can. J. Vet. Res. 60: 65-68, 1996.
4. Pritchett, L.C., Ulibarri, C., Roberts, C.R., Schneifer, R.K. and Sellon, D.C.: Identification of potential physiological and behavioral indicators of post operative pain in horses after exploratory celiotomy for colic. Appl. Anim. Behav. Sci. 80: 31-43, 2003.
5. Graubner, C., Gerber, V., Doherr, M. and Spadavecchia, C.: Clinical application and reliability of a post abdominal surgery pain assessment scale (PASPAS) in horses. Vet. J. 188: 178-83, 2010.
6. van Loon, J.P., Jonckheer-Sheehy, V.S., Back, W., van Weeren, P.R. and Hellebrekers, L.J.: Monitoring equine visceral pain with a composite pain scale score and correlation with survival after emergency gastrointestinal surgery. Vet. J. 200: 109-15, 2014.
7. Dalla Costa, E., Minero, M., Lebelt, D., Stucke, D., Canali, E. and Leach, M.C.: Development of the Horse Grimace Scale (HGS) as a pain assessment tool in horses undergoing routine castration. PLoS One. 9:e92281, 2014.
8. Taffarel, M.O., Luna, S.P., de Oliveira, F.A., Cardoso, G.S., Alonso, J. de M., Pantoja, J.C., Brondani, J.T., Love, E., Taylor, P., White, K. and Murrell, J.C.: Refinement and partial validation of the UNESP-Botucatu multidimensional composite pain scale for assessing postoperative pain in horses. BMC Vet. Res. 11: 83, 2015.
9. Sutton, G.A., Dahan, R., Turner, D. and Paltiel, O.: A behaviour-based pain scale for horses with acute colic: scale construction. Vet. J. 196: 394-401, 2013.
10. Sutton, G.A., Paltiel, O., Soffer, M. and Turner, D.:Validation of two behaviour-based pain scales for horses with acute colic. Vet. J. 197: 646-650, 2013.
11. de Vet, H.C., Terwee, C.B. and Bouter, L.M.: Current challenges in clinimetrics. J. Clin. Epidemiol. 56: 1137-41, 2003.
12. Galea, M.P.: Introducing Clinimetrics. Aust. J. Physiother. 51: 139-140, 2005.
13. Streiner, D.L., Norman, G.R. and Cairney, J.: Health Measurement Scales: a practical guide to their development and use. (5th Ed.), Oxford University Press, Oxford, 2015.
14. Taylor, P.M., Pascoe, P.J. and Mama, K.R.: Diagnosing and treating pain in the horse. Where are we today? Vet. Clin. North Am: Equine Pract. 18: 1-19, 2002.
15. Ashley, F.H., Waterman-Pearson, A.E. and Whay, H.R.: Behavioural assessment of pain in horses and donkeys: application to clinical practice and future studies. Equine Vet J. 37: 565-75, 2005.
16. Mair, T.S. and Smith, L.J.: Survival and complication rates in 300 horses undergoing surgical treatment of colic. Part 1: Short-term survival following a single laparotomy. Equine Vet. J. 37: 296-302, 2005.
17. McGraw, K.O. and Wong, S.P.: Forming inferences about certain intraclass correlation coefficients. Psychol. Methods. 1: 30-46, 1996.
18. Weir, J.P.: Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. J. Strength Cond. Res. 19: 231-240, 2005.
19. Shoukri, M.M. and Pause, C.A.: Models for reliability studies. In: M.M. Shoukri and C.A. Pause, (Eds): Statistical Methods for Health Sciences. CRC Press, Boca Raton. p. 27, 1999.
20. Bland, J.M. and Altman, D.G.: Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1: 307-310, 1986.
21. Landis, J.R. and Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics. 33: 159–174, 1977.
22. Wright, J.G. and Feinstein, A.R.: A comparative contrast of clinimetric and psychometric methods for constructing indexes and rating scales. J. Clin. Epidemiol. 45: 1201-1218, 1992.
23. Turner, D., Griffiths, A.M., Steinhart, A.H., Otley, A.R. and Beaton, D.E.: Mathematical weighting of a clinimetric index (Pediatric Ulcerative Colitis Activity Index) was superior to the judgmental approach. J. Clin. Epidemiol. 62: 738-44, 2009.
24. de Jong, A.E., Bremer, M., Schouten, M., Tuinebreijer, W.E. and Faber, A.W.: Reliability and validity of the pain observation scale for young children and the visual analogue scale in children with burns. Burns. 31: 198-204, 2005.
25. Voepel-Lewis, T., Zanotti, J., Dammeyer, J.A. and Merkel, S.: Reliability and validity of the face, legs, activity, cry, consolability behavioral tool in assessing acute pain in critically ill patients. Am. J. Crit. Care. 19: 55-61, 2010.
26. Johnston, C.C.: Psychometric issues in the measurement of pain. In: Finley, G.A. and McGrath, P.J. (Eds): Measurement of pain in infants and children. IASP Press, Seattle, pp. 5-20, 1998.

Appendix 1:

MILD	BEHAVIORS	SCORE
↓	FLANK WATCHING FLEHMAN or LIP CURLING	1
↓	STERNAL RECUMBENCY STRETCHING RESTLESSNESS	2
↓	KICKING ABDOMEN PAWING	3
↓	ATTEMPTING TO LIE DOWN or CROUCHING LATERAL RECUMBENCY	4
↓ SEVERE	ROLLING	5

a: The Revised Equine Acute Abdominal Pain Scale (EAAPS)

DEFINITIONS OF BEHAVIORS

FLANK WATCHING; a horse that glances at its side or flank
LIP CURLING; curling of the upper lip as in the Flehman Response usually accompanied by extension of the head and neck.
STERNAL RECUMBENCY; lying on ground but on the sternum with the forelegs tucked under the body.
STRETCHING; most commonly, taking a stance as a male horse would to urinate, but not urinating. Alternatively, may take the form of dog-sitting (forelegs straight and the hind-quarters on the ground) or bowing (forelegs straight and low to the ground with the weight pulled back onto the quarters).
RESTLESSNESS; a horse that does not stand quietly but moves, apparently aimlessly, and appears agitated. May shift weight from leg to leg. Movements may be jerky such as tail swishing and there may be wide excursions of the head.
CIRCLING; a horse demonstrating a relatively severe form of restlessness by being in constant motion. If confined to a stall, the horse will walk compulsively in a circular pattern and if being walked on a straight line will be unwilling to stand quietly.
KICKING ABDOMEN; a horse that kicks in the direction of its abdomen
PAWING; scraping the ground with a forelimb, often with the head held low and relatively close to the ground.
LATERAL RECUMBENCY; lying on ground on one side of the body and with all four legs to the same side.
CROUCHING; a horse that buckles the legs and looks like it is attempting to lie down, but does not, or lies down but gets up immediately.
ROLLING; a horse that lies down and rolls on its back, raising its hooves up in the air.

b: The Revised Equine Acute Abdominal Pain Scale (EAAPS) descriptions