doi: 10.2746/042516409X479568

Repeatability of subjective evaluation of lameness in horses

K. G. KEEGAN*, E. V. DENT, D. A. WILSON, J. JANICEK[†], J. KRAMER, A. LACARRUBBA, D. M. WALSH[‡], M. W. CASSELLS[‡], T. M. ESTHER[‡], P. SCHILTZ[§], K. E. FREES[#], C. L. WILHITE[#], J. M. CLARK[#], C. C. POLLIT[¶], R. SHAW[¥] and **T. NORRIS[·]**

Department of Veterinary Medicine and Surgery, University of Missouri, Columbia, Missouri 65211, USA; [†]Weems and Stephens Equine Hospital, Aubrey, Texas 76227, USA; [‡]Homestead Veterinary Hospital, Pacific, Missouri 63069, USA; [§]Equine Medical Services Inc., Columbia Missouri 65201, USA; [#]Wilhite & Frees Equine Veterinary Hospital, Peculiar, Missouri 64078, USA; [¶]School of Veterinary Science, University of Queensland, St. Lucia, Brisbane, Queensland 4072, Australia; [¥]Eleven Point Equine Clinic, Birch Tree, Missouri 65438, USA; and [·]All Creatures Veterinary Hospital, Mountain Home, Arkansas 72653, USA.

Keywords: horse; lameness; diagnosis; subjective

Summary

Reasons for performing study: **Previous studies have suggested that agreement between equine veterinarians subjectively evaluating lameness in horses is low. These studies were limited to small numbers of horses, evaluating movement on the treadmill or to evaluating previously-recorded videotape.**

Objectives: **To estimate agreement between equine practitioners performing lameness evaluations in horses in the live, over ground setting.**

Methods: **131 mature horses were evaluated for lameness by 2–5 clinicians (mean 3.2) with a weighted-average of 18.7 years of experience. Clinicians graded each limb using the AAEP lameness scale by first watching the horse trot in a straight line only and then after full lameness evaluation. Agreement was estimated by calculation of Fleiss' κ. Evaluators agreed if they picked the same limb as lame or not lame regardless of the severity of perceived lameness.**

Results: **After only evaluating the horse trot in a straight line clinicians agreed whether a limb was lame or not 76.6% of the time (κ = 0.44). After full lameness evaluation clinicians agreed whether a limb was lame or not 72.9% of the time (κ = 0.45). Agreement on forelimb lameness was slightly higher than on hindlimb lameness. When the mean AAEP lameness score was >1.5 clinicians agreed whether or not a limb was lame 93.1% of the time (κ = 0.86), but when the mean score was ≤1.5 they agreed 61.9% (κ = 0.23) of the time. When given the task of picking whether or not the horse was lame and picking the worst limb after full lameness evaluation, clinicians agreed 51.6% (κ = 0.37) of the time.**

Conclusions: **For horses with mild lameness subjective evaluation of lameness is not very reliable.**

Potential relevance: **A search for and the development of more objective and reliable methods of lameness evaluation is justified and should be encouraged and supported.**

Introduction

Lameness is the most economically important medical condition affecting horses (Anon 2001). Subjective evaluation of lameness by watching the horse in motion is a standard of practice.

Low agreement between equine clinicians for subjective scoring of mild to moderate lameness was first reported in a study evaluating videotapes of horses trotting on a treadmill (Keegan *et al*. 1998). More recently, other over ground studies of videotape evaluations have concluded similarly, that subjective scoring of lameness in horses is either only 'moderately reliable' (Hewetson *et al*. 2006) or 'just within acceptable limits' (Fuller *et al*. 2006). Moreover, the subjective assessment of lameness after assumption of perineural anaesthesia has been shown to be biased, with lameness severity significantly decreasing whether or not a block was actually performed (Arkell *et al*. 2006). These studies suggest that objective methods of evaluating lameness in horses would be beneficial to equine veterinarians.

In these previous studies, however, subjective evaluations were performed and scored by viewing videotapes without accompanying audio recordings of relatively small number of horses, sometimes of horses moving on a treadmill. These restrictions of artificial conditions may adversely affect the evaluators' agreement. Giving the evaluator the opportunity to observe the horse move in a natural environment, in a straight line as well at during lunging, and to utilise flexion tests, may increase interobserver agreement. There have been no published reports on studies of subjective agreement between veterinarians using the AAEP lameness or similar grading scale and performing full, live lameness evaluations on large numbers of horses moving over ground. Indeed, the proper use of the AAEP lameness scale, and a fair assessment of it, requires the evaluators to do more than just observe the horse trotting in a straight line.

The purpose of this study was to estimate the interobserver reliability of experienced equine veterinarians doing full lameness evaluations in horses moving over ground. It was hypothesised that interobserver reliability of experienced equine veterinarians performing full lameness evaluations in the natural setting is higher than these previous limited studies suggest and that subjective evaluation of lameness using the AAEP lameness scale has sufficient repeatability to be considered the gold standard for lameness detection and evaluation in the horse.

Materials and methods

In this study, 131 mature horses were evaluated for lameness. Horses were of various sizes and breeds. All horses were aged >2 years

*Author to whom correspondence should be addressed.

[Paper received for publication 02.03.09; Accepted 31.07.09]

and were either actively being trained or used for riding or had been used for riding immediately prior to a suspect lameness that prompted the veterinary evaluation. Horses used in this study were either admissions to the University of Missouri Veterinary Medical Hospital's Equine Clinic (64 cases) or boarders at 2 local College's Equestrian Program's stables (13 and 39 cases) and 5 private training stables (16 cases). A few horses were evaluated during a pre-purchase examination.

Horses were evaluated for lameness by 2–5 equine clinicians (mean 3.2). In most cases the evaluations were performed simultaneously by the different veterinarians. In a few cases, because of evaluator unavailability, evaluations were not simultaneous. However, no evaluations of 2 different veterinarians were separated by more than 30 min. In all, 16 different equine veterinarians were evaluators in the study. The weighted mean years-of-experience of the evaluators was 18.7 years. (Weighted mean years of experience = $\Sigma_{i=1}^{n} p_i \, y_i$ where p_i = proportion of all evaluations conducted by evaluator i, y_i = years of experience of evaluator i, for evaluators 1, 2, 3…..16 = *i*.) Five of the evaluators were ACVS board-certified surgeons, who were responsible for 85% of the evaluations. Thirty-two percent of the evaluations were performed by 4 veterinarians, each with over 27 years of experience. Forty-nine percent of the evaluations were carried out by University of Missouri faculty clinicians at the University of Missouri.

Each evaluator graded lameness by 2 methods. First, they graded lameness as either present or absent in each limb after observing the horse trotting up and down in a straight line only. During this method the evaluator could have requested that the horse trot up and down multiple times. Second, they graded each limb on a 0–5 scale, with a grade of 5 equivalent to nonweightbearing and 0 equivalent to no lameness, after performing further flexion tests or lunging the horse, as requested by the individual evaluator. Scores in half units (e.g. 2.5) were allowed. Each evaluator was unaware of the scores of the other evaluators until scores were turned in and no consulting between evaluators during the lameness evaluation was allowed.

Agreement between evaluators was assessed by calculation of Fleiss' kappa (κ), a statistical measure of inter-relater reliability for any number of raters.

$$\kappa = P - P_e / 1 - P_e$$

Where $P_\mathrm{e} = \Sigma_{j=1}^{k} p_j^2$, $p_j = \frac{1}{\mathrm{Nn}} \Sigma_{i=1}^{N} n_{ij}$, $P = \frac{1}{\mathrm{Nn(n\text{-}1)}} (\Sigma_{i=1}^{N} \Sigma_{j=1}^{2} n_{ij}^2 - Nn)$, and n_{ij} = number of evaluators who assigned the ith subject to the jth category. In this study N = 131, the number of horses evaluated, i = 1, 2, 3, …131; n = 3.2, the mean number of evaluations per horse; j = 1 (lame) or 2 (not lame), and k = 2 (2 categories). The factor $1 - P_e$ gives the degree of agreement that is attainable above chance and $P - P_e$ gives the agreement above chance actually attained. P_e = expected chance agreement. P = Total agreement, including chance agreement.

Agreement was calculated in 3 different ways: 1) by limb using the 'present' or 'absent' designation given by the evaluators for that limb while evaluating the horse trotting in a straight line only; 2) by limb using the AAEP score given by the evaluators for that limb after evaluating the horse trotting in a straight line and after lunging and flexion tests (hereafter referred to as full lameness evaluation); and 3) on the most affected limb in the horse using the AAEP score given by the evaluators after full lameness evaluation.

Agreement for right and left limbs were combined to obtain mean values for overall forelimb and hindlimb agreement. For the 'present' or 'absent' designation by limb, evaluators were considered in agreement if both selected that limb as lame or if both did not select that limb as lame. For the AAEP score designation by limb, evaluators were considered in agreement if both scored that limb >0 or if they both scored that limb = 0. Therefore, for either method of evaluation by limb, only 2 distinct choices were generated, lame or not lame. Chance agreement was expected to be about 50% and the highest possible agreement above chance to be 50%. For the AAEP score designation by limb, agreement was also calculated for lameness severity subgroups (mean AAEP score ≤1.5 and >1.5) using mean AAEP lameness score of all evaluators for each horse.

For the AAEP score designation on the most affected limb in the horse, evaluators were considered in agreement if both scored the same limb as the one with the most predominant lameness, regardless of the scores for any other limb, or if both scored the horse as having no lameness in any of the 4 limbs. Some evaluators scored the horse as having the most predominant lameness in more than one limb. In these cases a proportion of agreement for this evaluator to the other evaluators was calculated based on the number of total 'worst' limbs selected. Thus, for evaluation of agreement on the most affected limb, 5 potential choices were possible (right forelimb, left forelimb, right hindlimb, left hindlimb, sound). Chance agreement was expected to be about 20% (1 chance out of 5) and the highest possible agreement above chance to be 80% (100–20%).

The 95% confidence interval for a difference between any 2 lameness scores was calculated as 2.77 times the within-subject standard deviation of scores (Bland and Altman 1986). This confidence interval was calculated separately for the forelimbs and hindlimbs. Proportion of lameness scores at each level of lameness to total scores for every veterinarian in the study who evaluated at least 10 horses (n = 6) was also calculated. Possible scores for this analysis ranged from AAEP *Grade -5* for a nonweightbearing left lameness to AAEP *Grade +5* for a nonweightbearing right lameness, although no scores below -3 or greater than +3 were recorded for any limbs on any horses by any evaluator in this study.

Results

When restricted to judging lameness only by trotting the horse up and down in a straight line and selecting whether or not the horse was lame in a given limb, our evaluators were in agreement 76.6% of the time overall, 79.1% for the forelimbs and 74.1% for the hindlimbs (Tables 1 and 2). This was 18.3% above estimated chance agreement overall, 22.1% above estimated chance agreement for the forelimbs and 14.4% for the hindlimbs (κ = 0.44 overall, 0.51 for the forelimbs and 0.36 for the hindlimbs). Expected chance agreement calculated for the categorical selection of lame or not lame in a limb was slightly greater than the theoretically expected chance agreement of 50% (58.3% overall, 57.0% for the forelimbs and 59.7% for the hindlimbs). Agreement between evaluators when considering only the 64 cases presented to the University of Missouri Teaching Hospital and evaluated by University of Missouri faculty was no different than when considering all the horses evaluated in the study.

When lameness was scored after a full lameness evaluation and selecting each limb as lame or not using the AAEP lameness scale, our evaluators were in agreement 72.9% of the time overall, 76.2% for the forelimbs and 69.5% for the hindlimbs (Tables 1 and 2). This was 22.4% above estimated chance agreement overall, 26.2% above estimated chance agreement for the forelimbs and 18.6% for the hindlimbs ($\kappa = 0.45$ overall, 0.52 for the forelimbs and 0.38 for the hindlimbs). Expected chance agreement calculated for the categorical selection of lame (AAEP score >0) or not lame (AAEP score = 0) in a limb was near the theoretically expected chance agreement of 50% (50.5% overall, 50.0% for the forelimbs and 50.9% for the hindlimbs). Agreement between evaluators when considering only the 64 cases presented to the University of Missouri Teaching Hospital and evaluated by University of Missouri faculty was no different than when considering all the horses evaluated in the study.

Agreement was higher when the mean AAEP score for a particular limb was >1.5 than when the mean AAEP score was ≤1.5 (Table 3). When the mean AAEP score was >1.5, evaluators were in agreement that a given limb was lame or not 93.1% of the time overall, 94.2% of the time for the forelimbs and 92.0% for the hindlimbs. This is 44.2% above expected chance for the forelimbs and 41.6% above expected chance for the hindlimbs ($\kappa = 0.88$ and 0.84 for the forelimbs and hindlimbs, respectively). When the mean AAEP score was ≤1.5, evaluators were in agreement that a given limb was lame or not 61.9% of the time overall, 65.8% for the forelimbs and 57.9% for the hindlimbs. This is 15.8% above expected chance for the forelimbs and 7.0% above expected chance for the hindlimbs ($\kappa = 0.32$ and 0.14 for the forelimbs and hindlimbs, respectively).

When lameness was scored after a full lameness evaluation and selecting the most affected limb using the AAEP lameness scale, evaluators were in agreement 51.6% of the time (Table 1). This was 28.6% above chance ($\kappa = 0.37$). Expected chance agreement calculated for the categorical selection of right forelimb, left forelimb, right hindlimb, left hindlimb or sound was near the theoretically expected chance agreement of 20% (23.0%). Agreement between evaluators when considering only the 64 cases presented to the University of Missouri Teaching Hospital and evaluated by University of Missouri faculty was no different than when considering all the horses evaluated in the study.

Variance vs. mean of AAEP scores for each horse for forelimb and hindlimbs are shown in Figures 1 and 2. Score variance peaked around a mean AAEP score of 1.5 for the forelimbs and between 1.0 and 2.0 for the hindlimbs. The 95% confidence interval for difference in AAEP scores was 2.0 grades for the forelimb and 2.3 grades for the hindlimb. Proportion of lameness scores at each level

TABLE 1: Agreement analysis results for 3 different methods of evaluation. *Method 1:* lameness scored as either 'present' or 'absent' in each limb after watching horse trot up and down in a straight line only. *Method 2:* lameness in each limb scored using the AAEP grading scale after performing full lameness evaluations (flexion tests, lunging, etc.). *Method 3:* Worst limb selected using the AAEP grading scale after performing full lameness evaluations. % figures in parenthesis () represent agreement when only the 64 cases presented to the University of Missouri were considered

Method of evaluation	Expected chance agreement (P_e)	Total agreement (P)	Possible agreement above chance (1 - P_e)	Agreement above chance actually achieved ($P - P_e$)	κ
1. Present or absent,	58.3%	76.6%	41.7%	18.3%	0.44
trot in straight line only.	(55.8%)	(76.5%)	(44.2%)	(20.7%)	0.47
2. Present or absent	50.5%	72.9%	49.5%	22.4%	0.45
(using AAEP score),	(51.3%)	(71.5%)	(48.7%)	(20.2%)	0.41
full lameness evaluation.					
3. Most affected limb	23.0%	51.6%	77.0%	28.6%	0.37
(using AAEP score),	(23.2%)	(49.9%)	(76.8%)	(26.7%)	(0.35)
full lameness evaluation.					

TABLE 2: Agreement analysis results for *method 1* (lameness scored as either 'present' or 'absent' in each limb after watching horse trot in straight line only) and *method 2* (lameness scored using AAEP grading scale after full lameness evaluation) for lameness of the forelimb and hindlimb

Method of evaluation	Lameness location	Expected chance agreement (P_e)	Total agreement (P)	Possible agreement above chance (1- P_e)	Agreement above chance actually achieved ($P - P_e$)	κ
1. Present or absent,	Forelimb	57.0%	79.1%	43.0%	22.1%	0.51
trot in straight line only.	Hindlimb	59.7%	74.1%	40.3%	14.4%	0.36
2. Present or absent	Forelimb	50.0%	76.2%	50.0%	26.2%	0.52
(using AAEP score),						
full lameness evaluation.	Hindlimb	50.9%	69.5%	49.1%	18.6%	0.38

TABLE 3: Agreement analysis results for *method 2* (lameness scored using AAEP grading scale after full lameness evaluation) for lameness overall and for lameness of the forelimb and hindlimb when mean AAEP score was ≤1.5 and >1.5. P_e are estimated = 50.5% overall, 50.0% for forelimb and 50.9% for hindlimb scores for both ≤1.5 and >1.5 subgroups

Method of evaluation	Overall or limb involved	Total agreement (P)	Agreement above chance actually achieved ($P - P_e$)	κ
Mean AAEP score <1.5	Overall	61.9%	11.4%	0.23
	Forelimb	65.8%	15.8%	0.32
	Hindlimb	57.9%	7.0%	0.14
Mean AAEP score >1.5	Overall	93.1%	42.6%	0.86
	Forelimb	94.2%	44.2%	0.88
	Hindlimb	92.0%	41.1%	0.84

of lameness to total lameness scores for each veterinarian who evaluated at least 10 horses (n = 6) in the study are shown in Figures 3 and 4. The majority of scores from these 6 veterinarians were AAEP *Grade 0* for both forelimb and hindlimbs. However, the range of proportion of scores was also greatest for AAEP *Grade 0* than for any other AAEP grade. Evaluators used the AAEP score of 0 ranging from about 25–58% of the time for forelimb scoring and from about 34–57% of the time for hindlimb scoring.

Discussion

Results of this study indicate that, considering lameness detection in horses as a moderately difficult diagnostic skill, agreement for subjective evaluation of lameness by equine veterinarians is only marginally acceptable. For all forelimb lameness it was only 22–26% above chance. For all hindlimb lameness it was only 14–19% above chance. Therefore, veterinarians agree that a forelimb is lame about (median of 22–26% + 50% theoretical expected chance agreement) 3 out of 4 times and that a hindlimb is lame about (median of 14–19% + 50%) 2 out of 3 times.

Agreement for detection of lameness >AAEP *Grade 1.5* was higher, achieving 44% and 42% agreement above chance for forelimb and hindlimb lameness, respectively. Therefore, for lameness above *Grade 1.5*, veterinarians agree that a forelimb or hindlimb is lame (median of 42–44% + 50%) greater than 9 out of 10 times. Agreement for detection of lameness ≤ AAEP *Grade 1.5*, however, was much lower, achieving only 16% and 8% agreement above chance for forelimb and hindlimb lameness, respectively. Therefore, for lameness below *Grade 1.5* veterinarians agree that a forelimb is lame about (16% + 50%) 2 out of 3 times, and that a hindlimb is lame just over (8% + 50%) half the time.

Moreover, given the task of picking the predominant limb of all 4 limbs, agreement between veterinarians was disappointingly low. Veterinarians agree on the most affected limb only 29% above chance. This is equivalent to less than 50% of the time (29% + 20%). Additionally, for all lameness, but especially for lameness ≤AAEP *Grade 1.5*, agreement on forelimb lameness was more reliable than on hindlimb lameness and performing flexion tests and lunging the horse did not improve agreement between evaluators.

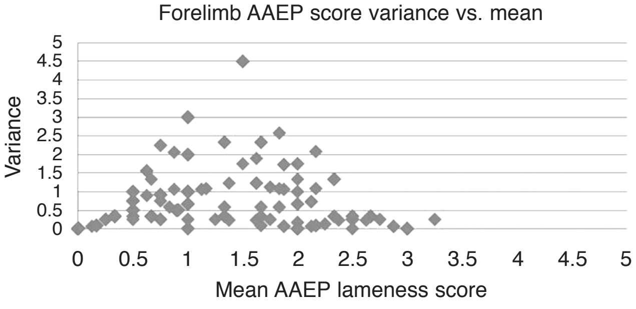

Fig 1: Variance of forelimb lameness score vs. mean forelimb lameness score for all veterinarians and horses.

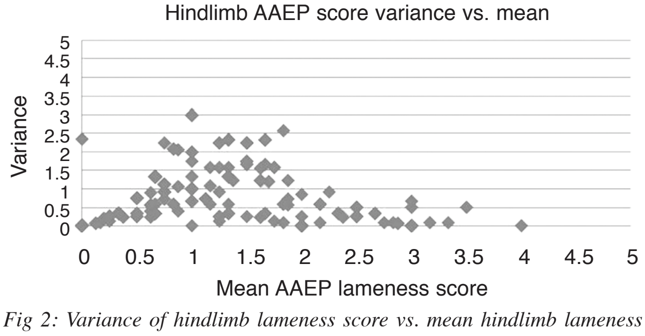

Fig 2: Variance of hindlimb lameness score vs. mean hindlimb lameness score for all veterinarians and horses.

The results of this study can be generally compared to previous studies that evaluated horses for lameness primarily by reviewing videotapes. In a previous study, using soundless videotapes of 24 horses with forelimb lameness trotting on a treadmill, inter-rater agreement was lower (κ = 0.21) (Keegan *et al*. 1998) than that measured in the current 'live' lameness evaluation study (κ ≅ 0.5 for forelimb and 0.3 for hindlimb lameness). Another study, using a verbal rating scale similar to the AAEP scale for evaluating videotapes of 20 horses, found that experienced veterinarians agreed on severity of lameness about 60% of the time (Hewetson *et al*. 2006). This is slightly lower than the present overall agreement (≅70%) but differences in the model used to assess agreement preclude exact comparison. Hewetson *et al*. (2006) assessed agreement on severity of lameness, a more conservative approach more likely to result in lower κ values. On the other hand, these authors did not assess side of lameness, a more liberal approach more likely to result in higher κ values. The present results are also similar to another study using a 10-point rating scale and videotape recordings of 20 horses with mild to moderate lameness. In that study κ = 0.41 for inter-rater agreement on severity of lameness (Fuller *et al*. 2006). The same caveats of assessing severity and not assessing side of lameness can be used to prevent strong comparisons between this additional study and

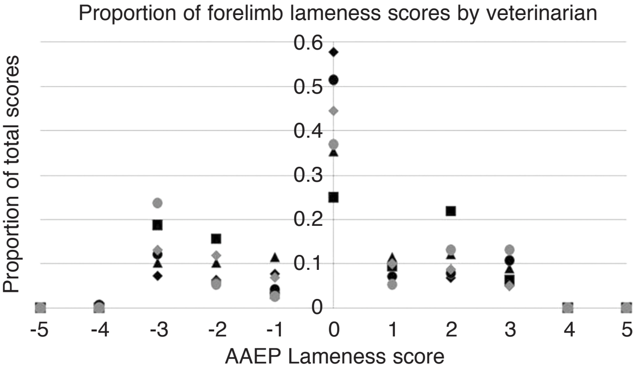

Fig 3: Proportion of forelimb lameness scores at each grade to all forelimb lameness scores for the 6 veterinarians scoring at least 10 horses.

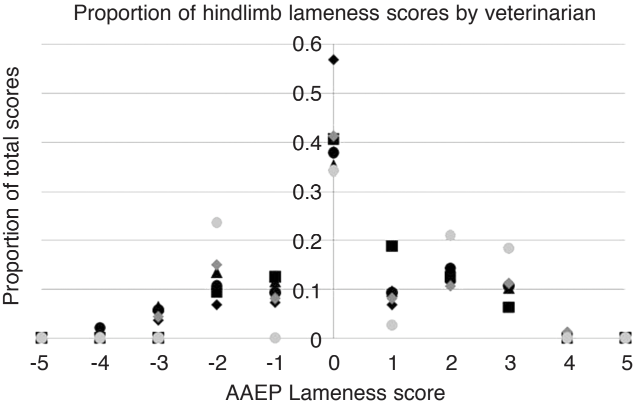

Fig 4: Proportion of hindlimb lameness scores at each grade to all hindlimb lameness scores for the 6 veterinarians scoring at least 10 horses.

the present study. Despite these limitations to direct comparison, the results of our study support that a full, 'live' lameness evaluation improves agreement between evaluators compared to viewing videotapes of horses in motion.

In the present study the interevaluator variance was estimated. Interevaluator variance is a mid-level test of reliability also referred to as reproducibility or repeatability. In this study this level of reliability was estimated using a very liberal model of categorical classification, i.e. whether or not the horse was determined to be lame in a limb or not. We did not test the reliability of the AAEP scale for estimating severity of lameness, an ordinal and more conservative model, one more likely to naturally have lower agreement. It should also be made clear that the utility of the AAEP scale was not evaluated in this study. The distinctions between scores stated in the AAEP lameness guidelines are clear, unambiguous and mutually exclusive (Anon 1991; Stashak 2002). We did not specifically train our evaluators to use the AAEP lameness scale and assumed it was a well-known and well-accepted standard of practice. Although it is likely that evaluators interpreted and utilised the AAEP lameness scale individually, we modelled our agreement independent of scale interpretation. Therefore, how the evaluators used the AAEP scale is not relevant for either accepting or rejecting the conclusions of this study.

Since the amount of agreement one would expect by chance depends on the number and relative frequencies of the categories under test, reliability for categorical classifications should be measured with a κ coefficient (Landis and Koch 1977). It is a reasonable rule of thumb for associations between 2 subjectively graded variables to require κ to be >0.8, with $0.67 \leq \kappa \leq 0.80$ allowing tentative conclusions to be drawn (Krippendorf 2003). Medical researchers have typically used less strict guidelines with $0 \leq \kappa \leq 0.20$ representing 'slight', $0.21 \leq \kappa \leq 0.40$ 'fair', $0.41 \leq \kappa \leq 0.60$ 'moderate', $0.61 \leq \kappa \leq 0.80$ 'substantial' and $0.81 \leq \kappa \leq 1$ 'near perfect' agreement. These estimates are arbitrary but useful benchmarks as long as one considers the limitations of using κ to estimate agreement. An important limitation is that calculating κ assumes a quantification of chance agreement, which is relevant only under conditions of statistical independence of the raters. We took precaution to protect this independence by requesting evaluators not to consult between themselves before scoring each horse. However, subtle cues given by the evaluators on account of the tests they requested, for example a flexion test of the right carpus, or lunging the horse in a clockwise direction, may have influenced other evaluators. The present finding, that evaluator agreement did not improve after full lameness evaluation from the initial evaluation of just observing the horse trotting in a straight line, suggests that these possible subtle clues did not have strong influence.

Another limitation of using κ to calculate agreement is that it is influenced by trait prevalence, in this case the presence of lameness, in our selection of study subjects. Because most of the horses evaluated in this study were presented specifically for lameness evaluation or may have been included in the study by a trainer because they had suspicion of lameness in that horse, the incidence of lameness (trait prevalence) in this study population was higher than in a population of randomly selected horses. When trait prevalence is high, calculation of expected chance agreement will be high. When chance agreement is high it is more difficult to achieve high agreement above chance. Our calculated chance agreements, however, were what would be expected with equal prevalence of soundness and lameness (50% by limb, either sound in that limb or lame in that limb, and 20% on the most affected limb, either sound or lame in the left forelimb, right forelimb, left hindlimb or right hindlimb). Nevertheless, these estimates for agreement should be used with caution and extrapolated only to models of lameness evaluation and cross sections of the population similar to that used in this study.

The 95% confidence interval for a difference in AAEP lameness score was about 2 grades, a large percentage (40%) of the total possible range of scores (1–5). Variance of AAEP score peaked around a mean AAEP score of 1–2 and the largest difference in proportion of AAEP scores given by the evaluators was at AAEP score of 0 (sound). Agreement between evaluators was higher when the mean AAEP score was above 1.5. These 3 observations make it clear that disagreement between evaluators occurs most often in horses with lameness of mild severity. This should be expected of any diagnostic test around its limit of detection.

It is reasonable to suspect that this limited sensitivity may be the result of deficient spatial and temporal visual acuity or resolution of the evaluators. Small asymmetries in motion between the right and left sides of the body as a result of mild lameness may be difficult to pick up with the naked eye. In one experimental study of a computer simulation modeled on the vertical movement of the *tubera coxae* in a lame horse, the threshold for detection of movement asymmetry was found to be approximately 25% difference in amplitude (Parkes *et al.* 2009). A horse trotting at 4 m/s will have a stride rate of approximately 1.5 strides/second and will move the head and pelvis up and down at twice this rate, or 3 times/s (Keegan *et al.* 2001). To prevent significant errors in measurement of amplitude, a measurement method should sample at a minimum of 5 times the rate of the event in the signal being measured (Winter 1982). The estimated sampling rate or temporal resolution of the human eye is about 15 frames/s (Sweet 1953; Näsänen 2006). Therefore, the human eye's temporal resolution just achieves this minimum sampling frequency. Small changes in movement, likely with mild lameness, especially at higher speeds (higher frequency of movement), may not be adequately picked up by the human eye. Sensitivity of lameness detection is probably also a function of the different individual method or methods that veterinarians use to detect lameness in horses, something that would theoretically improve with targeted training (Dyson 2009).

The meaning of the amplitude of the κ coefficient and whether or not it is acceptable for a given test depends on a variety of factors including the strength of the relationship being studied, in this case whether lameness actually changes the way a horse moves, and whether this can be determined visually. It also depends on the importance of the test. Is watching the horse in motion and scoring this motion an important clinical exercise? How important is the subjective evaluation in the whole scheme of lameness evaluation in horses? These are questions that are best answered individually by experienced equine practitioners. However, the authors of this study, as equine practitioners ourselves, suggest that, if evaluation of the horse in motion is important clinically for diagnosis of lameness and horses do change the way they move because of lameness, the current standard of practice of subjective evaluation of lameness is not acceptable for horses with mild lameness. Similar conclusions have been drawn about subjective evaluation of lameness in the dog and dairy cow (O'Callaghan *et al.* 2003; Quinn *et al.* 2007; Waxman *et al.* 2008; Flower and Weary 2009).

In conclusion, the results of this study suggest that: 1) because agreement between experienced evaluators grading mild lameness is low and because agreement for picking the most lame limb is low, subjective evaluation of lameness by more than one evaluator should not be used as a gold standard in evaluating other methods of lameness evaluation or in assessing lameness in clinical trials; 2) agreement between experienced evaluators for assessing lameness is not enhanced when given the opportunity of assessing the horse while lunging or after flexion tests, compared to evaluating the horse trotting in a straight line only; 3) agreement between experienced evaluators for assessing lameness in the live, over ground environment is higher than when assessing videotaped recordings of the horse's movement; and 4) a search for and the development of a more objective and reliable method of lameness evaluation for use in the field, such as body-mounted inertial sensor-based motion analysis (Barrey *et al.* 1994; Keegan *et al.* 2004; Church *et al.* 2009), is justified and should be encouraged and supported.

References

Anon (1991) *Guide for Veterinary Service and Judging of Equestrian Events,* 4th edn., American Association of Equine Practitioners, Lexington. p 19.

Anon (2001) National economic cost of equine lameness, colic, and equine protozoal myeloencephalitis in the United States. In: *USDA:APHIS:VS, National Health Monitoring System. Information Sheet*. Fort Collins. #N348.1001.

Arkell, M., Archer, R.M., Guitian, F.J. and May, S.A. (2006) Evidence of bias affecting the interpretation of the results of local anaesthetic nerve blocks when assessing lameness in horses. *Vet. Rec.* **159**, 346-349.

Barrey, E., Hermelin, M., Vaudelin, J.L., Poirel, D. and Valette, J.P. (1994) Utilisation of an accelerometric device in equine gait analysis. *Equine vet. J., Suppl.* **17**, 7-12.

Bland, J.M. and Altman, D.G. (1986) Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* **1**, 307-310.

Church, E.E., Walker, A.M., Wilson, A.M., Pfau, T., Crevier-Denoix, N., van Weeren, P.R., Chateau, H., Clayton, H.H. and Buchner, H.H.F. (2009) Evaluation of discriminant analysis based on dorsoventral symmetry indices to quantify hindlimb lameness during over ground locomotion in the horse. *Equine vet. J.* **41**, 304-308.

Dyson, S.J. (2009) The clinician's eye view of hindlimb lameness in the horse: Technology and cognitive evaluation. *Equine vet. J.* **41**, 99-100.

Flower, F.C. and Weary, D.M. (2009) Gait assessment in dairy cattle. *Animal* **3**, 87-95.

Fuller, C.J., Bladon, B.M., Driver, A.J. and Barr, A.R.S. (2006) The intra- and inter-assessor reliability of measurement of functional outcome by lameness scoring in horses. *Vet. J.* **171**, 281-286.

Hewetson, M., Christley, R.M., Hunt, I.D. and Voute, L.C. (2006) Investigations of the reliability of observational gait analysis for the assessment of lameness in horses. *Vet. Rec.* **158**, 852-858.

Keegan, K.G., Pai, P.F., Wilson, D.A. and Smith, B.K. (2001) Signal decomposition method of evaluating head movement to measure induced forelimb lameness in horses trotting on a treadmill. *Equine vet. J.* **33**, 446-451.

Keegan, K.G., Wilson, D.A., Wilson, D.J., Smith, B., Gaughan, E.M., Pleasant, R.S., Lillich, J.D., Kramer, J., Howard, R.D., Bacon-Miller, C., Davis, E.G., May, K.A., Cheramie, H.S., Valentino, W.L. and van Harrevald, P.D. (1998) Evaluation of mild lameness in horses trotting on a treadmill: Agreement by clinicians and interns or residents and correlation of their assessments with kinematic gait analysis. *Am. J. vet. Res.* **59**, 1370-1377.

Keegan, K.G., Yonezawa, Y., Pai, P.F., Wilson, D.A. and Kramer, J. (2004) Evaluation of a sensor-based system of motion analysis for detection and quantification of forelimb and hind limb lameness in horses. *Am. J. vet. Res.* **65**, 665-670.

Krippendorf, K. (2003) Reliability. In: *Content Analysis: An Introduction to Its Methodology,* Sage Publications, Thousand Oaks. p 241.

Landis, J.R. and Koch, C.G. (1977) The measurement of observer agreement for categorical data. *Biometrics* **33**, 159-174.

O'Callaghan, K.A., Cripps, P.J., Downham, D.Y. and Murray, R.D. (2003) Subjective and objective assessment of pain and discomfort due to lameness in dairy cattle. *Animal Welfare* **12**, 605-610.

Näsänen, R., Ojanpää, H., Tanskanen, T. and Päällysaho, J. (2006) Estimation of temporal resolution of object identification in human vision. *Exp Brain Res*. **172**, 464-471.

Parkes, R.S.V., Weller, R., Groth, A.M., May, S. and Pfau, T. (2009) Evidence of the development of 'domain-restricted' expertise in the recognition of asymmetric motion characteristics of hindlimb lameness in the horse. *Equine vet. J.* **41**, 112-117.

Quinn, M.M., Keuler, N.S., Lu, Y., Faria, M.L., Muir, P. and Markel, M.D. (2007) Evaluation of agreement between numerical rating scales, visual analogue scoring scales, and force plate gait analysis in dog. *Vet. Surg.* **36**, 360-367.

Stashak, T.S. (2002) Diagnosis of lameness. In: *Adams' Lameness in Horses,* Ed: T.S. Stashak, Lea & Febiger, Philadelphia. 122.

Sweet, A.L. (1953) Temporal discrimination by the human eye. *Am. J. Psych.* **66**, 185.

Waxman, A.S., Robinson, D.A., Evans, R.B., Hulse, D.A., Innes, J.F. and Conzemius, M.G. (2008) Relationship between objective and subjective assessment of limb function in normal dogs with an experimentally induced lameness. *Vet. Surg.* **37**, 241-246.

Winter, D.A. (1982) Camera speeds for normal and pathological gait analyses. *Med. Biol. Eng. Comput.* **20**, 408.

Author contributions The initiation, conception and planning for this study were by K.G.K., D.A.W., J.J. and J.K. All authors contributed to the execution, with statistics by K.G.K. The paper was written by K.G.K., E.V.D., D.A.W., J.J., J.K. and D.M.W.