TITLE: Veterinary student competence in equine lameness recognition and assessment: a mixed methods study

AUTHORS: Sandra D Starke, Stephan A May

JOURNAL TITLE: Veterinary Record

PUBLISHER: BMJ Publishing Group

PUBLICATION DATE: 12 August 2017 (online)

DOI: 10.1136/vr.104245

Veterinary student competence in equine lameness recognition and assessment: a mixed methods study

Sandra D. Starke[1,*], MSc PhD and Stephen A. May[1], MA VetMB PhD DVR DEO FRCVS DipECVS FHEA

[1] The Royal Veterinary College, Hawkshead Lane, North Mymms, Hatfield, Hertfordshire, AL9 7TA, UK

* Corresponding author (S.D. Starke): sstarke@rvc.ac.uk

Abstract

The development of perceptual skills is an important aspect of veterinary education. In this study, we investigated veterinary student competency in lameness evaluation at two stages, prior to (3rd year) and during (4th/5th year) clinical rotations. Students evaluated horses in videos, where horses were presented on a straight line and on circles. Eye tracking data were recorded to follow student gaze. On completing the task, students filled in a structured questionnaire. Results showed that experienced veterinary students outperformed inexperienced students in classifying the correct limb as lame, although performance of both cohorts was inadequate in most scenarios. Mistakes largely arose from classifying an incorrect limb as lame. The correct classification of sound horses as non-lame was at chance level. While the experienced student cohort primarily looked at upper body movement (head and sacrum) during lameness assessment, the inexperienced cohort focused on limb movement. Student self-assessment of performance was realistic, and task difficulty was most commonly rated between 3 and 4 out of 5. The inexperienced students named a considerably greater number of visual lameness features than the experienced students. Future dedicated training based on the findings presented here may help students to develop more reliable lameness assessment skills.

Introduction

An important aspect of veterinary education is the development of perceptual skills. In equine veterinary education, the ability to correctly identify, differentiate and classify lameness is one such skill. However, even experienced veterinarians struggle with this task, especially for mild lameness (Dow and others 1991; Keegan and others 1998; Ishihara and others 2005; Orito and others 2007; Keegan and others 2010; Mc Cracken and others 2012). Currently, only a small number of studies have investigated student performance in equine gait scoring (Keegan and others 1998; Arkell and others 2006) and the nature of their visual assessments remains largely unexplored. In order to develop a systematic approach to lameness education, leading to reliable performance of our future practitioners, it would be beneficial to understand student performance better.

The experience-dependent ability to assess gait abnormalities has been investigated in numerous studies, covering not only horses (Keegan and others 1998; Arkell and others 2006) but also humans (Read and others 2003; Brunnekreef and others 2005; Maathuis and others 2005; Brown and others 2008; Ong and others 2008; Williams and others 2009), dogs (Kestin and others 1992) and pigs (Main and others 2000). Gait assessment skills improve with experience. Intra-observer agreement (the same person judging a scenario multiple times) often becomes more consistent (Keegan and others 1998; Main and others 2000; Brunnekreef and others 2005; Arkell and others 2006; Ong and others 2008), although this may simply reflect mistakes being made more consistently (Kawamura and others 2007; Waxman and others 2008). Inter-observer agreement (different people judging the same scenario) also often increases with experience (Main and others 2000; Brunnekreef and others 2005; Maathuis and others 2005; Arkell and others 2006; Ong and others 2008). For visual gait assessment in horses, a past landmark study highlighted the strong subjectivity and unreliability of gait evaluation independent of experience level. The authors hypothesised that high intra-observer variation in less experienced observers may be rooted in a lack of confidence and a less systematic assessment protocol compared to their experienced counterparts (Keegan and others 1998).

For a veterinary student, lameness assessment skills are commonly developed through a combination of taught classes and practical exposure on clinical rotations and work placements. During practical sessions, students learn from the experts of their respective institutions. Dedicated textbooks support the expansion of knowledge regarding the gait examination (Wyn-Jones 1988; Stashak and Hill 1996; Stashak 2002; Ross and Dyson 2011), although authors vary in the weighting of different lameness pointers. At the Royal Veterinary College (RVC), at the time of performing this study, the didactic aspects of equine lameness were first covered in year 3 of the BVetMed curriculum. At that point, clinical observation and decision making skills are starting to be developed. The gained knowledge is then applied during intramural clinical rotations in years 4 and 5 where students spend one week on equine orthopaedics. Here, students follow the clinician on duty when performing lameness examinations of horses admitted to the RVC's Equine Referral Hospital (ERH). Students with an interest in equine work typically also choose to perform their extramural studies (EMS) in equine practice or referral hospitals, where a larger number of lameness cases can be followed.

To date, gait assessment studies have focused largely on performance metrics as outcome variables. However, during perceptual tasks it is equally important to quantify the approach taken by participants in order to relate performance and approach. In the case of lameness assessment, this approach would be quantifiable through the pattern of visual attention allocation. Eye tracking allows quantification of where a participant is looking when making judgements, opening the door to insights into cognitive processes that were previously inaccessible. Due to the narrow field of sharp vision (1º to 2º visual angle for foveal, respectively central, vision), humans have to look at those parts in a scene where they expect to find the information that they are seeking (Findlay and Gilchrist 2003). Eye tracking has been used in thousands of studies (Tien and others). In human medicine, for example, it has helped to understand errors made when interpreting diagnostic images (Thomas and Lansdown 1963; Kundel and others 1978; Krupinski 1996; Krupinski 2005; Krupinski 2010; Balslev and others 2012).

In the present study, we investigated the approaches and abilities of a random sub-section of veterinary students from two different stages in their degree programme when evaluating horses for lameness. We used a combination of gait scoring, eye tracking and questionnaires to quantify the percentage of horses evaluated correctly, sources of judgement error, visual assessment strategies and student attitudes. The aim of this work is to provide a comprehensive baseline for the development of more effective courses on equine lameness identification.

Materials and Methods

Task

The task was to assess horses on a PC monitor. Thirty-four video recordings (25 Hz, resolution 1024 x 768 pixels) were selected to compile a database comprising an approximately equal presence of fore- and hind limb lameness in both the left and right limbs in different views (Figure 1): 15 clips showed the horses trotting away from and towards the observer on a straight line (condition ‘straight line’), 15 horses trotted on a circle on either the left or the right rein (condition ‘circle’) and four horses were shown trotting in lateral view (condition ‘lateral’). The lateral condition was only used to explore student opinion regarding this view, but not for in-depth analysis of performance. All video recordings were edited in Premiere CS4 (Adobe, USA) into clips of approx. 30 to 50 s duration. Each clip presented the horse in the chosen view without sound, the video being repeated at least once. Clips were presented on a 21 inch CRT monitor (Dell Trinitron, refresh rate: 100 Hz).

[Figure 1]

Lameness

The veterinary reports for all horses presented on straight line and circle were obtained in order to establish the lame limb and lameness score of each horse from the expert clinical examination based on visual examination and final diagnosis. Four video recordings of horses filmed in lateral view were scored by three experienced assessors who agreed on the lame leg independently and an average was taken from all three scores. For all horses classed as lame, a visually detectable lameness in the limb consistent with the expert opinion was verified by the experimenter (S D Starke) to warrant that it could be shown to students and ensure that lameness was visible to the trained eye. Fourteen out of 15 horses were deemed lame on the straight line, twelve out of 15 on the circle and three out of four in lateral view. Horses were unilaterally fore- or hind limb lame with a lameness grade of up to 5 out of 10 (UK grading scale, see Arkell and others or popular textbooks) across limbs and conditions. The lame horses on the straight line had a mean (SD) lameness score of 2.38 (1.30) out of 10 for forelimb lameness and 3.00 (1.26) out of 10 for hind limb lameness. On the circle this amounted to 3.67 (1.21) out of 10 for forelimb lameness and 3.50 (1.05) out of 10 for hind limb lameness and in lateral view to 3.67 (0.58) out of 10 for hind limb lameness.

Participants

Twenty 3rd year BVetMed students (the 'inexperienced' cohort) as well as 12 4th/5th year BVetMed students (the 'experienced' cohort, who had performed equine rotations and some participants had done at least one EMS placement in an equine hospital or practice) were recruited from the RVC student body to participate in the study. All participants volunteered to take part in the study by responding to an email invitation (inexperienced and part of the experienced cohort) or after additional encouragement from lecturing staff (part of the experienced cohort).

Eye tracking and gait scoring

Eye tracking data were collected in parallel to the scoring session using a stand-alone eye tracker (X120, Tobii Technology, Sweden). The eye tracker was positioned underneath the monitor and calibrated for each individual participant at the beginning of the assessment session using the default 5-point calibration procedure. Eye tracking data were recorded at 60 Hz, with a small subset of data recorded at 120 Hz at the beginning of the study. The viewing distance between student and eye tracker was approximately 70 cm, ranging dynamically between 50 to 80 cm according to the calibration volume of the eye tracker.

Video clips were randomised in Matlab (The MathWorks, USA) and presented as seven segments within the proprietary eye tracker software (Tobii Studio v.2). Each student was presented with the seven segments, each contaning several video clips, in random order. This provided the best trade-off at the time between software limitations and workload arising from randomisation. For the assessment of each horse, an on-screen questionnaire was shown after each video clip with the following questions and single choice answers:

1. Is the horse lame? If it is lame, which is the most affected leg?

   [Answers: Sound, Left fore, Right fore, Left hind, Right hind]

2. Which lameness score would you assign to that leg?

   [Answers: 0/10 (Sound), 1/10, 2/10, 3/10, 4/10, 5/10, 6/10, 7/10, 8/10, 9/10, 10/10 (non weight bearing)]

3. Is a further leg affected?

[Answers: No, Left fore, Right fore, Left hind, Right hind]

Questionnaires

After completion of the task, students were given a questionnaire which contained questions relating to their self-evaluation and approaches when assessing horses for lameness. Students were asked to:

- Rate the difficulty of the lameness assessment task on a scale from 1 (very simple) to 5 (very difficult).
- Estimate the percentage of horses they expected to have evaluated correctly.
- Rank the difficulty of assessing horses on the straight line, on a circle and in lateral view.
- Note down what movement features were used when assessing forelimb and hind limb lameness on the straight and circle.

The cohort of experienced students was additionally asked to estimate how many lameness cases they had seen so far.

Analysis of gait scoring

Scores for straight line and circle were analysed using custom-written software in Matlab (The MathWorks, USA) after exporting data from Tobii Studio. For each student, the response for each horse was compared to the gold standard derived from clinical records. The following metrics were calculated to quantify performance: 1) the percentage of horses correctly classified as lame and sound, 2) the percentage of horses classified lame in the correct limb and 3) the percentage of incorrect classifications. Incorrect classifications were attributed to the following five reasons: students considering a lame horse sound; students considering a sound horse lame; and students selecting the incorrect limb(s) as affected, further split into students selecting the incorrect contralateral limb as affected and students selecting an incorrect ipsilateral limb as affected. Variation in the allocated grades was quantified for each horse across all those students who had deemed the horse primarily lame in the fore- or hind limb. Score range as well as the difference between each score and the median score were calculated for each cohort. Agreement across students was calculated using free marginal kappa (κ). Inter-observer agreement was calculated using the free marginal multilayer kappa (Brennan and Prediger 1981; Randolph 2005; Warrens 2010) using an online kappa calculator (http://justusrandolph.net/kappa). Kappa was calculated for the presence or absence of lameness in general, forelimb lameness and hind limb lameness. For this metric, the side of lameness was not considered.

Statistical analysis was performed in IBM SPSS 20 (IBM Corp.), where a Shapiro-Wilk test was run to decide on the use of parametric or non-parametric tests. The percentages of lame horses correctly classed as sound or lame and correctly classed as lame in the correct limb was compared between student cohorts for straight ond circle using the Mann-Whitney U Test. Consistency in allocated grades was tested this way, too. Within student cohorts, a Wilcoxon signed-rank matched-pairs test was used to examine differences between performance for fore- and hind limb lameness and between straight line and circle. Regression analysis was performed to determine whether there was an association between the caseload seen and performance of the experienced cohort.

Analysis of visual assessment

For the assessment of the 15 horses trotting on a straight line away from / towards the observer, eye tracking data were analysed to determine the preferred regions students examine visually. The constant change in direction of horses on the circle proved prohibitive for analysis due to the manual workload. In order to assign the gaze data to discrete regions on the horse, the position of the os sacrum, head and feet in the video recordings were digitised using Tracker (www.cabrillo.edu/~dbrown/tracker). In Matlab (The MathWorks, USA), point of gaze was then calculated for each frame and assigned to an equally spaced horizontal grid, where it was assigned to one out of 18 bins (Figure 2). The grid was calculated based on the tracked os sacrum (trot away from observer) or head (trot towards observer) position with the feet as reference points. For each participant and each of the 15 horses, the percentage viewing time for each region was then calculated as the sum of frames spent looking at each region divided by the total viewing time. The total viewing time was calculated as the sum of all frames where gaze data were tracked on the monitor. For each participant, the mean viewing time was then calculated for each of the 18 discrete regions across all videos in which at least 60 % of the total video playing time was being tracked.

Statistical analysis was carried out in SPSS as described above. A Mann-Whitney U Test was performed to compare the viewing times allocated to each of the 18 regions (Figure 2 and 5) between the two student cohorts, separately for trot away from and towards the observer.

[Figure 2]

Analysis of questionnaire responses

From the answers to the questionnaires, the difficulty ratings for the task were extracted and average values and frequencies calculated. The percentage of horses students thought they had evaluated correctly (self-rating) and the objective percentage of horses evaluated correctly as described above was calculated across all 34 horses and all three views. The difference between self-assessment and objective assessment was calculated for each student and mean differences for each scenario and student cohort were calculated to examine bias in self-rating. The lameness pointers noted down by the students were extracted from the questionnaires, counted and ranked by nomination for assessment on the straight line and circle.

Statistical analysis was carried out in SPSS as described above. Differences in self-rating were compared between the inexperienced and experienced student cohort using an independent samples t-test.

Results

In this study we report results as the median [interquartile range, IQR] unless stated otherwise. For those readers unfamiliar with these metrics, they are used for data that does not follow a normal distribution (being non-parametric). The median reports the central tendency and is the point found in the middle of an ordered distribution: it is the value which separates the 50% of data points which are bigger from the 50% that are smaller. It is also called 2nd quartile: when splitting the data distribution into four equal parts from lowest to highest value, quartiles mark the thresholds between each of these four parts. The IQR reports data variation and quantifies the spread between the 1st and 3rd quartile: it is the range in which the 'middle' 50% of data points can be found, and is in meaning similar to the standard deviation in normal distributed data. Further details can be found in standard statistics textbooks.

Gait scoring

The median [IQR] percentage of lame horses correctly classed as generally lame by the inexperienced and experienced cohort for straight line and circle combined was 92 [10] % and 98 [10] %, respectively. There was no significant difference between cohorts ($P \geq 0.307$). The median [IQR] of sound horses correctly classed as sound across all views was 50 [30] % for both cohorts, not being significantly different ($P = 0.924$).

For forelimb lameness, the median [IQR] of horses classified lame in the correct limb (Figure 3) for the inexperienced and experienced cohort was 50 [38] % and 69 [25] %, respectively, on the straight as well as 83 [42] % and 100 [17] %, respectively, on the circle. For hind limb lameness, these values decreased to 33 [33] % and 67 [25] %, respectively, on the straight and 50 [33] % and 67 [33] %, respectively, on the circle. Differences in percentage correct between inexperienced and experienced cohort were significant ($P \leq 0.032$) except for forelimb lameness on the circle ($P = 0.053$). Within student cohorts, differences were significant between fore- and hind limb lameness on the circle ($P \leq 0.032$), but not on the straight ($P \geq 0.313$) and between trotting on the straight and circle for forelimb lameness ($P \leq 0.003$), but not for hind limb lameness ($P \geq 0.253$).

[Figure 3]

Incorrect classifications were mainly the result of declaring an incorrect limb as lame, with a median [IQR] of 71 [9] % and 60 [17] % for the inexperienced and experienced cohort, respectively. The incorrect contralateral and ipsilateral limbs were selected in similar proportion.

Mean grades did not differ significantly between cohorts for forelimb lameness ($P = 0.067$). However, for hind limb lameness grade differed significantly between cohorts ($P < 0.001$), where the experienced cohort's grade had a median score of one below the inexperienced cohort. The inexperienced cohort did not grade more than six grades away from the group's median grade, while the experienced student group graded no more than 3.5 grades away (Figure 4).

[Figure 4]

Within each student cohort, agreement (Table 1) was highest for the presence or absence of lameness on the circle (κ = 0.70 for inexperienced and κ = 0.89 for experienced students) and lowest for the presence or absence of hind limb lameness during evaluation away from / towards the observer (κ = 0.15 for inexperienced and κ = 0.20 for experienced students).

For the experienced student cohort, there was no association between the number of cases seen as part of their degree and the percentage of correctly evaluated horses for the nine students that provided data ($R^2 = 0.0039$).

[Table 1]

Visual assessment

During trot away from the observer, the percentage viewing time was significantly different between the two cohorts for regions 5, 6, 7, 8, 9, 13 and 14 ($P \leq 0.041$), but not for the remaining regions ($P \geq 0.114$). During trot towards the observer, the percentage viewing time was significantly different between the two cohorts for regions 6, 7, 12, 13 and 14 ($P \leq 0.047$), but not for the remaining regions ($P \geq 0.075$). Differences in distribution of gaze data are illustrated in Figure 5. There was considerable variation across students in the distribution of time across body regions for both trotting directions.

Questionnaire responses

The inexperienced cohort rated the difficulty of performing a lameness examination (Table 2) with a mean (SD) score of 4.10 (0.48) out of 5, ranging from 3 to 5. The experienced cohort rated the difficulty with a score of 3.35 (0.47) out of 5, ranging from 3 to 4. The median [IQR] number of lameness cases seen during their time at university by the experienced cohort was 24 [30] cases, ranging from 6 to 100. Assessment away from / towards the observer was judged most difficult by the majority of inexperienced students (12 / 20 students, 60%), while assessment away from / towards the observer (5 / 12 students, 42%) and in lateral view (6 / 12 students, 50%) were ranked as most difficult by a comparable number of experienced students.

[Table 2]

Compared to the calculated performance (Figure 6), students underestimated their skill by -11 (15) % (inexperienced cohort) and -13 (12) % (experienced cohort). The difference between self-rating and objective assessment of performance was not significantly different between cohorts.

[Figure 6]

Both, the inexperienced and experienced cohort, most frequently named the use of head movement for the detection of forelimb lameness on the straight line and circle. Both groups also most frequently named pelvis movement on the straight and stride length on the circle for the detection of hind limb lameness (Tables 3 and 4). The inexperienced cohort named a total of 19 and 18 features for forelimb lameness assessment on straight and circle, respectively, compared to 9 and 6 features named by the experienced cohort. The inexperienced cohort named a total of 16 and 21 features for hind limb lameness assessment on straight and circle, respectively, compared to 5 and 7 features named by the experienced cohort.

[Table 3 and 4]

Discussion

Student lameness assessment skills

This study showed that both inexperienced and experienced student cohorts, while in general reliably classifying horses as lame, were weak in correctly determining the affected limb, and only performed at the level of chance when classifying a horse as sound. The percentage of horses classified as lame in the correct limb significantly differed between the two cohorts, highlighting what we hope is the effect of learning throughout the veterinary degree. However, the variation in performance across experienced students was not explained by the amount of lame horses that they had previously observed. This finding substantiates the view that just looking at many horses does not necessarily improve skill level. Rather, the amount of time engaged in deliberate practice, the act of practising and refining a skill actively, accounts for most of the expertise level accomplished (Ericsson and others 1993). Even for the experienced cohort, performance was still not at a level which we would deem acceptable for reliable and repeatable diagnoses: there was large variation in performance across students and a median of less than 70 % correctly classified horses, with the exception of forelimb lameness on the circle. Consequently, students close to graduation would on average still incorrectly assess one out of four horses on the straight line, with individual participants assessing only 30 % of horses correctly. For the inexperienced cohort, individual performance was as low as 0% correct. If these results generalise to other veterinary students, it is clearly important to ensure more targeted teaching of lameness assessment skills, to improve student performance, especially of those choosing an equine career path. Future training should specifically target detection of mild lameness, a task difficult even for expert assessors (Keegan and others 2010; Keegan and others 2012). Also, since most mistakes resulted from classifying the incorrect limb as lame, an emphasis should be placed on how to recognise the correct limb.

We found differences in performance between assessment on straight line and circle as well as between fore- and hind limb lameness. Only on the circle did experienced students show an outstanding performance for the detection of forelimb lameness with a median of 100 %. On the circle, the mean lameness score derived from clinical records was higher (3.67) compared to the straight line (2.38). This is likely to have made lameness easier to see: with increasing lameness, the head nod converges on almost a single vertical head excursion per stride (Clayton 1987; Peloso and others 1993; Christovão and others 2007), resulting in an unambiguous nod-down pattern. This pattern is easier to see than differences in the two excursions between the two steps of a stride during milder lameness, since the eye has to follow movement at only have the step frequency and there is only a single minimum and maximum during the whole stride. In line with findings for more experienced clinicians (Keegan and others 2010), agreement amongst students was lowest for the presence or absence of hind limb lameness on both, straight line and circle. In this study, the variation in assigned scores across students compared well with previous findings, where final year students graded no more than three grades away from a horse's median grade, while more experienced assessors graded no more than two grades and experts graded no more than one grade away (Arkell and others 2006).

Reliability of self-evaluation and sources of error

While students did not perform at a competent level, their self-evaluation reflected their ‘objective’ performance, showing a realistic view of their own competence, even a slight underestimate. This finding is a positive sign, for their safety as practitioners, as they are recognising their own limitations while it also marks a positive sign for learning. The self-evaluation of experienced medical practitioners often does not bear any correlation with actual capabilities (Tracey and others 1997), a finding also observed when reviewing self-assessments of inexperienced and experienced medical personal (Gordon 1991). However, increasing self-awareness may lead to better self-evaluation (Mabe and West 1982) and identification of the need for learning. Inexperienced students, in particular, repeatedly mentioned struggling with the task due to the lack of learning opportunities and uncertainty regarding lameness features.

Errors in classifying the lame limb were rooted in what we consider to be both conceptual and perceptual mistakes. Selection of the incorrect contralateral limb (left vs. right limb) is likely to have been related to a conceptual error, where the student did not correctly remember the rules relating a movement feature to the lame limb. In contrast, the selection of the incorrect ipsilateral limb (forelimb vs. hind limb) may be related to a perceptual error due to the interaction of forelimb and hind limb lameness: especially for hind limb lameness, compensatory forelimb movements are a regular and notable occurrence (Uhlir and others 1997; Kelmer and others 2005). For forelimb lameness, compensatory hind limb movements develop only with more pronounced lameness and to a lesser visible extent (Uhlir and others 1997; Kelmer and others 2005). This linkage between the fore and hind limb movements may lead the student to mistaking compensatory movement for primary lameness, distracting them from further lameness evaluation. In future, study designs should account for this source of error by explicitly requesting observers to evaluate all limbs and if more than one appears lame, to indicate whether the lameness is assumed to be primary or “compensatory”.

Students performed at chance level for the classification of sound horses as non-lame. This finding highlights the need to not only train students in recognising lameness, but also provide training in recognising soundness and familiarising students with the spectrum of movement in sound horses. Knowing what is 'normal' and what is not is a key element for the development of discrimination skills, which in turn is an important component of expertise development (Kellman and Garrigan 2009). Research into the teaching of radiographic interpretation has shown that allowing students to compare abnormal and normal radiographs during training resulted in better performance in a subsequent detection test compared to students who were trained with abnormal radiographs only (Kok and others 2013). Similarly, facilitating 'learning by contrast' or 'comparison learning' (Kurtz and Boukrina 2005; Gentner 2010), for students learning how to analyse electrocardiograms, had a positive effect on subsequent student performance of the task (Hatala and others 2003; Ark and others 2007). The need for greater standardisation of gait analysis training and a greater emphasis on what is normal to improve inter-rater reliability has equally been highlighted for human gait analysis (Eastlack and others 1991). Hence, in future, it will be advisable to incorporate a greater number of sound horses into student training.

Visual assessment strategies

The mapped eye tracking data showed that the cohort of inexperienced students dedicated more time to the assessment of limb movement compared to the experienced cohort, both with the horse moving away from and towards the observer. The experienced students dedicated most time to the assessment of the pelvis area (trot away from the observer) and the head area (trot towards the observer). These findings matched the lameness features named by the two student cohorts: the inexperienced cohort named 16 to 21 features for assessment of fore- and hind limb lameness on straight and circle, many of them relating to movement of the limbs. In contrast, the experienced cohort named only 5 to 9 features across conditions, a large proportion referring to movement of head and pelvis. These findings indicate that as part of the development of diagnostic expertise, students may discard redundant lameness features and rely on the few features which they assume to be most reliable. However, as a possible confounding factor, experienced students may have been taught by the same veterinarian as part of their clinical lameness rotations, which could have caused the shift in feature selection. Variation in the distribution of gaze data and described lameness features between students of the same cohort as well as between cohorts highlighted inconsistency in assessment approach and evaluation protocol, which should be addressed in future training especially for students early in their careers to ensure that they acquire a more consistent skill set.

Conclusions

In this study we investigated veterinary student competency in lameness evaluation prior to and during clinical rotations. We found that the performance level of both cohorts was inadequate in most scenarios, where mistakes largely arose from classifying an incorrect limb as lame. We further found substantial differences in the assessment strategy, where the experienced student cohort primarily looked at upper body movement (head and sacrum) and named only few lameness pointers, while the inexperienced cohort focused to a large extent on limb movement and named a large variety of lameness pointers. We conclude that students require a higher level of perceptual training before commencing clinical rotations in order to clarify relevant lameness pointers and acquire the necessary discriminatory skill allowing them to differentiate sound from lame horses and correctly pinpoint the affected limb(s). We have by now released a free online training tool which aims to develop these skills, which can be found at www.lamenesstrainer.com. We hope that in future, systematic training based on the findings presented here, which takes into consideration common misconceptions and -possibly incorrect- 'intuitive' approaches, may help students to develop more reliable lameness assessment skills.

Acknowledgements

We would like to thank Jon Ward of Acuity ETS for supporting access to the eye tracker for the duration of the study, Prof. Roger K Smith for the video recordings of clinical cases, Thilo Pfau and Renate Weller for help with participant recruitment and all of our participants for their time. S. D. Starke's PhD was funded by the Mellon Trust *via* the Royal Veterinary College, UK.

References

ARK, T. K., BROOKS, L. R., EVA, K. W. (2007). The benefits of flexibility: the pedagogical value of instructions to adopt multifaceted diagnostic reasoning strategies. *Medical Education* **41,** 281-287.

ARKELL, M., ARCHER, R. M., GUITIAN, F. J., 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,** 346-348.

BALSLEV, T., JARODZKA, H., HOLMQVIST, K., DE GRAVE, W., MUIJTJENS, A. M. M., EIKA, B., VAN MERRIËNBOER, J., SCHERPBIER, A. J. J. A. (2012). Visual expertise in paediatric neurology. *European Journal of Paediatric Neurology* **16,** 161-166.

BRENNAN, R. L., PREDIGER, D. J. (1981). Coefficient Kappa - Some Uses, Misuses, and Alternatives. *Educational and Psychological Measurement* **41,** 687-699.

BROWN, C. R., HILLMAN, S. J., RICHARDSON, A. M., HERMAN, J. L., ROBB, J. E. (2008). Reliability and validity of the Visual Gait Assessment Scale for children with hemiplegic cerebral palsy when used by experienced and inexperienced observers. *Gait & Posture* **27,** 648-652.

BRUNNEKREEF, J. J., VAN UDEN, C. J., VAN MOORSEL, S., KOOLOOS, J. G. (2005). Reliability of videotaped observational gait analysis in patients with orthopedic impairments. *BMC Musculoskelet Disord* **6,** 17.

CHRISTOVÃO, F. G., BARROS, R. M. L., MARTINS, C. B., SILVA, M. A. G., FREITAS, E. V. V., LACERDA NETO, J. C., QUEIROZ-NETO, A. (2007). Three-dimensional kinematic analysis of head and limb movements of lame and non-lame colts. *Equine and Comparative Exercise Physiology* **4,** 31-36.

CLAYTON, H. M. (1987). Cinematographic analysis of the gait of lame horses III: Fracture of the third carpal bone. *Journal of Equine Veterinary Science* **7,** 130-135.

DOW, S. M., LEENDERTZ, J. A., SILVER, I. A., GOODSHIP, A. E. (1991). Identification of subclinical tendon injury from ground reaction force analysis. *Equine Veterinary Journal* **23,** 266-272.

EASTLACK, M. E., ARVIDSON, J., SNYDER-MACKLER, L., DANOFF, J. V., MCGARVEY, C. L. (1991). Interrater reliability of videotaped observational gait-analysis assessments. *PHYS THER* **71,** 465-72.

ERICSSON, K. A., KRAMPE, R. T., TESCH-RÖMER, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review* **100,** 363-406.

FINDLAY, J. M., GILCHRIST, I. D. (2003). *Active Vision - The psychology of looking and seeing,* Oxford, Oxford University Press.

GENTNER, D. (2010). Bootstrapping the mind: analogical processes and symbol systems. *Cogn Sci* **34,** 752-75.

GORDON, M. J. (1991). A review of the validity and accuracy of self-assessments in health professions training. *Academic Medicine* **66,** 762-9.

HATALA, R., BROOKS, L., NORMAN, G. (2003). Practice Makes Perfect: The Critical Role of Mixed Practice in the Acquisition of ECG Interpretation Skills. *Advances in Health Sciences Education* **8,** 17-26.

ISHIHARA, A., BERTONE, A. L., RAJALA-SCHULTZ, P. J. (2005). Association between subjective lameness grade and kinetic gait parameters in horses with experimentally induced forelimb lameness. *American Journal of Veterinary Research* **66,** 1805-1815.

KAWAMURA, C. M., DE MORAIS FILHO, M. C., BARRETO, M. M., DE PAULA ASA, S. K., JULIANO, Y., NOVO, N. F. (2007). Comparison between visual and three-dimensional gait analysis in patients with spastic diplegic cerebral palsy. *Gait & Posture* **25,** 18-24.

KEEGAN, K. G., DENT, E. V., WILSON, D. A., JANICEK, J., KRAMER, J., LACARRUBBA, A., WALSH, D. M., CASSELLS, M. W., ESTHER, T. M., SCHILTZ, P., FREES, K. E., WILHITE, C. L., CLARK, J. M., POLLITT, C. C., SHAW, R., NORRIS, T. (2010). Repeatability of subjective evaluation of lameness in horses. *Equine Vet J* **42,** 92-7.

KEEGAN, K. G., MACALLISTER, C. G., WILSON, D. A., GEDON, C. A., KRAMER, J., YONEZAWA, Y., MAKI, H., PAI, P. F. (2012). Comparison of an inertial sensor system with a stationary force plate for evaluation of horses with bilateral forelimb lameness. *American Journal of Veterinary Research* **73,** 368-74.

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., VAN HARREVELD, P. D. (1998). Evaluation of mild lameness in horses trotting on a treadmill by clinicians and interns or residents and correlation of their assessments with kinematic gait analysis. *American Journal of Veterinary Research* **59,** 1370-1377.

KELLMAN, P. J., GARRIGAN, P. (2009). Perceptual learning and human expertise. *Physics of Life Reviews* **6,** 53-84.

KELMER, G., KEEGAN, K. G., KRAMER, J., WILSON, D. A., PAI, F. P., SINGH, P. (2005). Computer-assisted kinematic evaluation of induced compensatory movements resembling lameness in horses trotting on a treadmill. *American Journal of Veterinary Research* **66,** 646-655.

KESTIN, S. C., KNOWLES, T. G., TINCH, A. E., GREGORY, N. G. (1992). Prevalence of Leg Weakness in Broiler-Chickens and Its Relationship with Genotype. *Veterinary Record* **131,** 190-194.

KOK, E. M., DE BRUIN, A. B. H., ROBBEN, S. G. F., VAN MERRIËNBOER, J. J. G. (2013). Learning radiological appearances of diseases: Does comparison help? *Learning and Instruction* **23,** 90-97.

KRUPINSKI, E. A. (1996). Visual scanning patterns of radiologists searching mammograms. *Academic Radiology* **3,** 137-144.

KRUPINSKI, E. A. (2005). Visual Search of Mammographic Images: Influence of Lesion Subtlety1. *Academic Radiology* **12,** 965-969.

KRUPINSKI, E. A. (2010). Current perspectives in medical image perception. *Attention, Perception, & Psychophysics* **72,** 1205-1217.

KUNDEL, H. L., NODINE, C. F., CARMODY, D. (1978). Visual Scanning, Pattern Recognition and Decision-making in Pulmonary Nodule Detection. *Investigative Radiology* **13,** 175-181.

KURTZ, K. J., BOUKRINA, O. (2005). Learning Relational Categories by Comparison of Paired Examples. *Proceedings of the Twenty-Sixth Annual Conference of the Cognitive Science Society*, 756-761.
MAATHUIS, K. G. B., VAN DER SCHANS, C. P., VAN IPEREN, A., RIETMAN, H. S., GEERTZEN, J. H. B. (2005). Gait in children with cerebral palsy - Observer reliability of Physician Rating Scale and Edinburgh Visual Gait Analysis Interval Testing Scale. *Journal of Pediatric Orthopaedics* **25,** 268-272.
MABE, P. A., WEST, S. G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. *Journal of Applied Psychology* **67,** 280-296.
MAIN, D. C. J., CLEGG, J., SPATZ, A., GREEN, L. E. (2000). Repeatability of a lameness scoring system for finishing pigs. *Vet Rec.* **147,** 574-576.
MC CRACKEN, M. J., KRAMER, J., KEEGAN, K. G., LOPES, M., WILSON, D. A., REED, S. K., LACARRUBBA, A., RASCH, M. (2012). Comparison of an inertial sensor system of lameness quantification with subjective lameness evaluation. *Equine Veterinary Journal* **44,** 652-656.
ONG, A. M. L., HILLMAN, S. J., ROBB, J. E. (2008). Reliability and validity of the Edinburgh Visual Gait Score for cerebral palsy when used by inexperienced observers. *Gait & Posture* **28,** 323-326.
ORITO, K., KUROZUMI, S., ISHII, I., TAKAHASHI, T., HIRAGA, A., MATSUDA, H. (2007). Identification of Equine Subclinical Lameness Induced by Pressure to the Sole of Fore- or Hindlimb. *Journal of Equine Veterinary Science* **27,** 429-434.
PELOSO, J. G., STICK, J. A., SOUTAS-LITTLE, R. W., CARON, J. C., DECAMP, C. E., LEACH, D. H. (1993). Computer-assisted three-dimensional gait analysis of amphotericin-induced carpal lameness in horses. *American Journal of Veterinary Research* **54,** 1535-43.
RANDOLPH, J. J. (2005). Free-marginal multirater kappa: An alternative to Fleiss' fixed-marginal multirater kappa. *Joensuu Learning and Instruction Symposium 2005.* Joensuu, Finland.
READ, H. S., HAZLEWOOD, M. E., HILLMAN, S. J., PRESCOTT, R. J., ROBB, J. E. (2003). Edinburgh visual gait score for use in cerebral palsy. *Journal of Pediatric Orthopaedics* **23,** 296-301.
ROSS, M. W., DYSON, S. J. (2011). *Diagnosis and Management of Lameness in Horses. ,* St. Louis, Missouri, Elsevier Saunders.
STASHAK, T. S. (2002). *Adams' Lameness in Horses*, Wiley-Blackwell.
STASHAK, T. S., HILL, C. (1996). *Practical Guide to Lameness in Horses*, Wiley-Blackwell.
THOMAS, E. L., LANSDOWN, E. L. (1963). Visual Search Patterns of Radiologists in Training. *Radiology* **81,** 288-292.
TIEN, T., PUCHER, P. H., SODERGREN, M. H., SRISKANDARAJAH, K., YANG, G.-Z., DARZI, A. Eye tracking for skills assessment and training: a systematic review. *Journal of Surgical Research* **191,** 169-178.
TRACEY, J. M., ARROLL, B., RICHMOND, D. E., BARHAM, P. M. (1997). The validity of general practitioners' self assessment of knowledge: cross sectional study. *BMJ* **315,** 1426-8.
UHLIR, C., LICKA, T., KUBBER, P., PEHAM, C., SCHEIDL, M., GIRTLER, D. (1997). Compensatory movements of horses with a stance phase lameness. *Equine Vet J Suppl 23* **29,** 102-5.
WARRENS, M. (2010). Inequalities between multi-rater kappas. *Advances in Data Analysis and Classification* **4,** 271-286.
WAXMAN, A. S., ROBINSON, D. A., EVANS, R. B., HULSE, D. A., INNES, J. F., CONZEMIUS, M. G. (2008). Relationship Between Objective and Subjective

Assessment of Limb Function in Normal Dogs with an Experimentally Induced Lameness. *Veterinary Surgery* **37,** 241-246.
WILLIAMS, G., MORRIS, M. E., SCHACHE, A., MCCRORY, P. (2009). Observational gait analysis in traumatic brain injury: Accuracy of clinical judgment. *Gait & Posture* **29,** 454-459.
WYN-JONES, G. (1988). Chapter 1. The Diagnosis of the Causes of Lameness. *In:* WYN-JONES, G. (ed.) *Equine Lameness.* Oxford: Blackwell Scientific Publications.

TABLES

Table 1. Within-observer agreement for the two student cohorts.

Free marginal kappa for the binary classification of horses as lame (yes/no), forelimb lame (yes/no) and hind limb lame (yes/no) for trotting on a straight line away from/towards the observer and in a circle.

		Inexperienced students	**Experienced students**
Lame?	**Away / towards**	0.54	0.49
	Circle	0.70	0.89
Forelimb lame?	**Away / towards**	0.30	0.64
	Circle	0.51	0.67
Hind limb lame?	**Away / towards**	0.15	0.20
	Circle	0.25	0.28

Table 2. Difficulty ranking for assessing different views.

Count of students ranking the views 'straight line', 'circle' and 'lateral' as 1 (easiest), 2 (medium) and 3 (most difficult). One experienced student perceived assessment on straight and circle to be of identical difficulty.

Difficulty ranking	**Inexperienced students**			**Experienced students**		
	Straight	***Circle***	***Lateral***	***Straight***	***Circle***	***Lateral***
Easiest (1)	5	8	7	7	4	1
Medium (2)	3	10	7	0	7	6
Most difficult (3)	12	2	6	5	1	6

Table 3. Named visual forelimb lameness pointers.

Pointers described for the determination of the presence of forelimb lameness during trot on the straight line and in a circle by respondents to the questionnaire for the cohort of inexperienced students (N = 19) and experienced students (N = 12). **: most commonly named feature, *: second most commonly named feature.

Lameness pointer	Inexperienced students		Experienced students	
	Straight	Circle	Straight	Circle
Head movement	18**	16**	12**	12**
Foot placement	6*	3	-	-
Circumduction	4	3	1	-
Stride / Step length	3	8*	4*	8*
Stance time	2	1	1	2
Toe dragging	2	1	-	-
Difference in load bearing	1	2	1	1
Lateral bending of head and neck	1	-	-	-
Shoulders	1	1	1	-
Evenness of striding / uneven gait	1	1	-	-
Dropping elbow	1	-	-	-
Hoof flight arc	1	1	-	-
Limb stiffness / general stiffness	1	1	-	-
Longer time with 1 foot in air	1	-	1	-
Hypermetria (abnormality in stepping)	1	1	-	-
Muscle atrophy	1	-	-	-
Tracking up	1	-	-	-
Fetlock extension	1	1	1	1
Loading of carpus	1	-	-	-
Looking out with affected limb on inside	-	2	-	-
Limb swing	-	1	-	-
magnitude of bending	-	1	-	-
Rhythm	-	1	-	-
Hopping sort of gait	-	1	-	-
Axillae movement asymmetry	-	-	2	-
Uncomfortable, e.g. not going forward, stopping	-	-	-	1

Table 4. Named visual hind limb lameness pointers.

Pointers described for the determination of the presence of hind limb lameness as detailed in Table 3, **: most commonly named feature, *: second most commonly named feature.

Lameness pointer	Inexperienced		Experienced	
	Straight	Circle	Straight	Circle
Hind quarters movement	15**	6*	12**	9*
Stride length	5*	9**	2*	10**
Foot placement	4	3		
Toe dragging	3	1		
Circumduction	3	2		
Stance time	2	2	1	1
Level of hocks	2			
Muscle conformation / atrophy	2	1		
Head nod	2	3		
Fetlock extension	2		2*	3
Limb stiffness	1	1		
How the legs stride through	1	1		
Less head nodding	1			
Rump moving up and down more, bouncing on one leg	1	1		
Hypermetria (abnormality in stepping)	1			
Limb (range of) movement	1			1
Tracking up		3		
Foot flight arc		2	1	
Stiffness through the back		1		
Don't know		1		
Reluctance to bend and limb position		1		
magnitude of bending		1		
Leg angles		1		1
Rhythm		1		
Freedom of movement through hip		1		
Not wanting to keep on the circle		1		
Speed of protraction				1

FIGURE LEGENDS

Figure 1. Different views in which horses were presented in the video clips: trot on a straight line away from the observer (a) and towards the observer (b), trot in a circle (c) and trot in lateral view (d).

Figure 2. Mapping of eye tracking data for all videos presenting the horse during trot away from / towards the observer. Illustration of grid overlay to map the point of gaze to discrete regions of interest, respectively bins (1-18). The grid was evenly spaced and calculated based on the location of strategic features for each frame: based on the location of the os sacrum (orange cross) and feet (red cross) for trot away from the observer (left), and the location of the head (orange cross) and feet (red cross) for trot towards the observer (right). The two points were used to create a grid of regular spacing, scaled to the height of the horse in each respective frame.

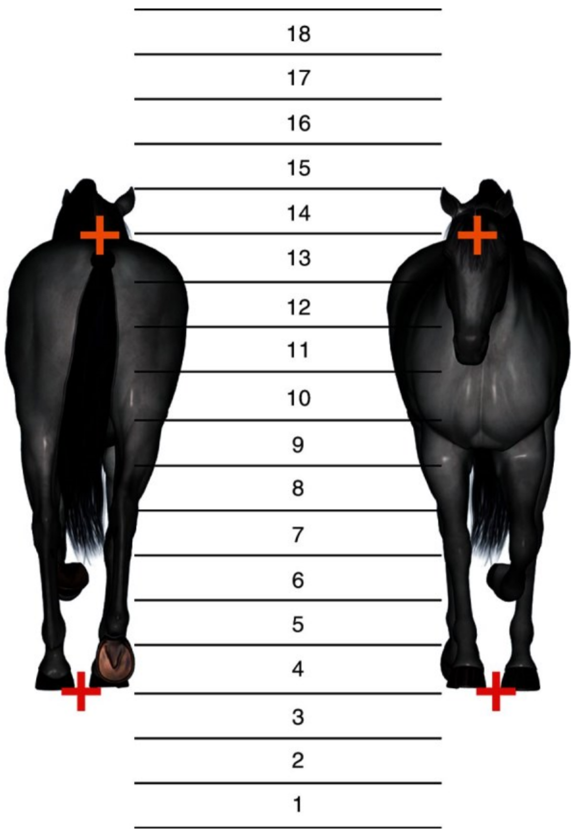

Figure 3. Boxplots of the percentage of lame horses classed as lame in the correct limb by a cohort of 20 inexperienced students ('Inexp') as well as 12 experienced students ('Exp') during assessment of horses at trot on the straight line away from/towards the observer (left) and at trot in a circle (right). Orange line: median, +: outlier.

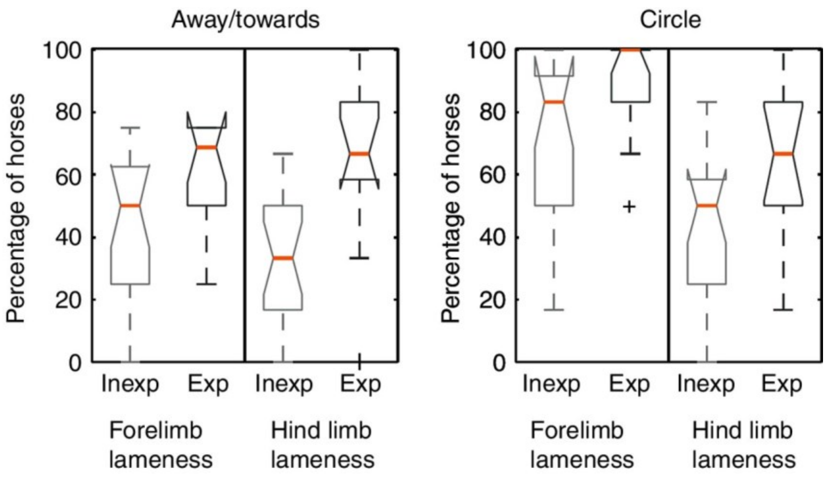

Figure 4. Boxplots of the variation in lameness grades assigned to horses by students of the inexperienced ('Inexp') and experienced ('Exp') cohort. For each cohort, the difference from the cohort's median grade as well as the total score range were calculated for those horses which each individual student considered lame with the most affected limb being a forelimb (left) or hind limb (right). Orange line: median.

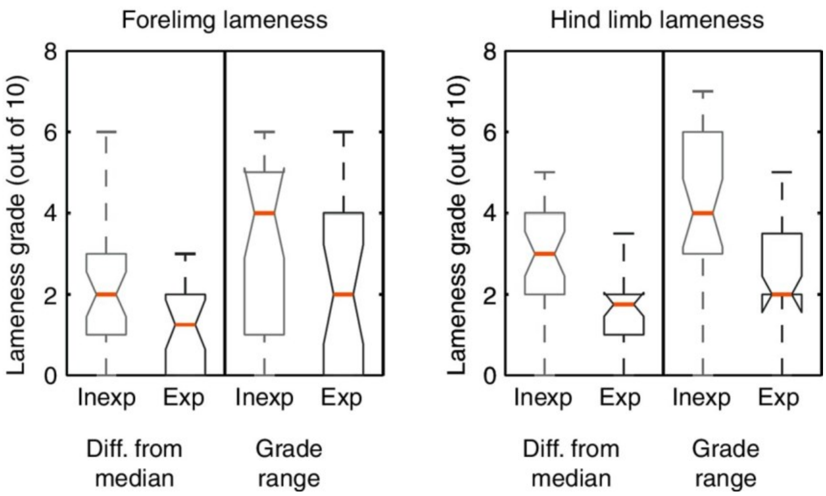

Figure 5. Boxplots for the percentage viewing time spent on each of the 18 discrete image regions based on the average value calculated for each student of the inexperienced (left) and experienced (right) cohort during trot away from the observer (top) and towards the observer (top). Orange line: median, +: outlier.

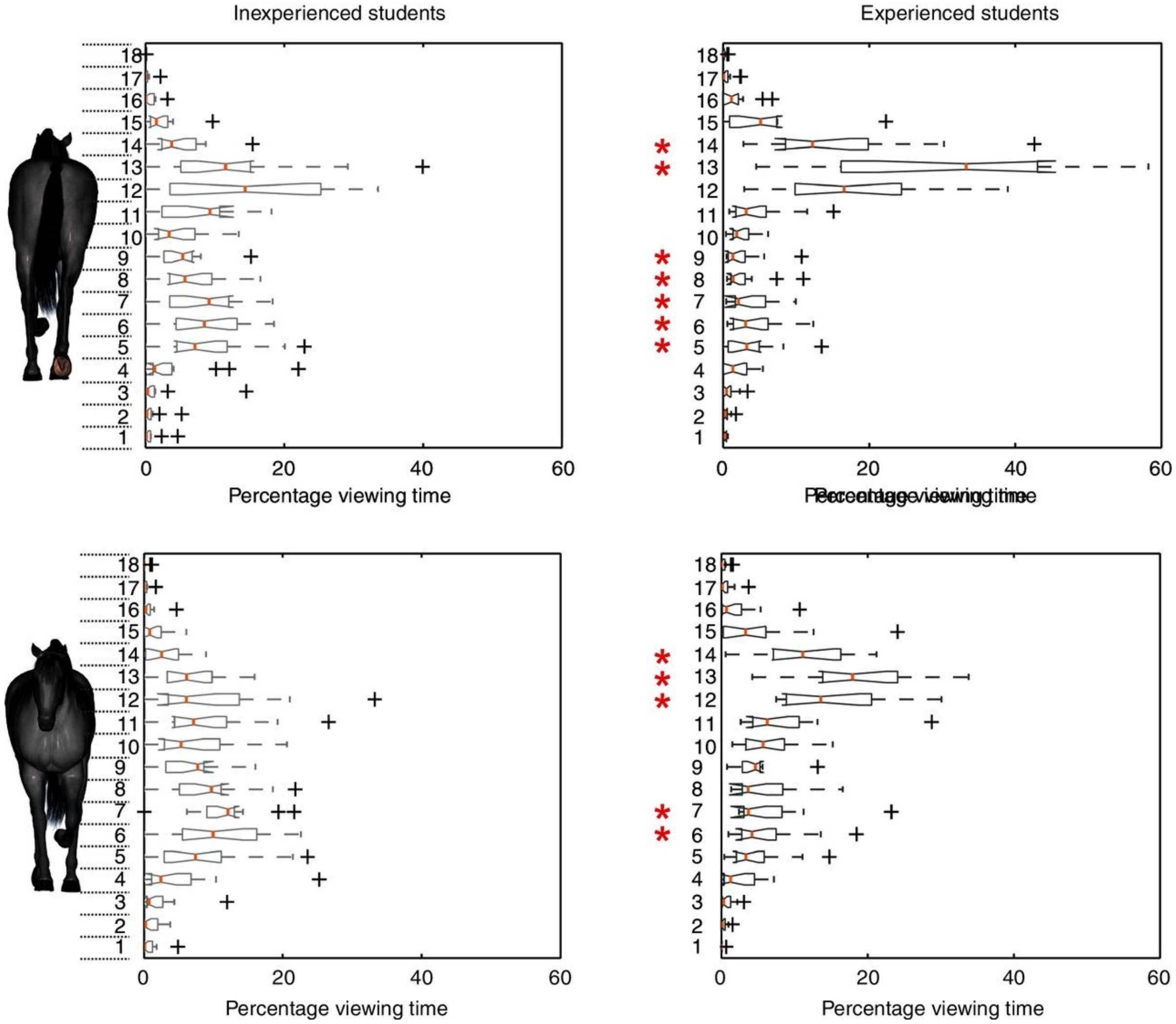

Figure 6. Left: comparison of subjective (self-rating by the student) and objective (calculated percentage correct when compared to clinical ground truth) performance evaluation. The diagonal line indicates an exact match between the the student's self evaluation and actual performance. Light grey: inexperienced students, dark grey: experienced students. Right: Difference between subjective and objective performance estimate for the inexperienced cohort ('Inexp) and the experienced cohort ('Exp). Most differences are below zero, indicating that the self-assessment was rather conservative.