Universiteit Utrecht

Faculteit der Diergeneeskunde

Minor Research Project

Facial pain expression assessment in lame horses in locomotion using the Equine Utrecht University scale for facial assessment of pain (EQUUS-FAP)

Student
Alexandra Dwulit, BSc
5956722

Supervisors
J.P.A.M. Thijs Van Loon, DVM PhD
Filipe Serra Braganca, DVM PhD

Utrecht University
Faculty of Veterinary Medicine, Dept. of Equine Sciences

June 5, 2020

Table of Contents

I. Layman's Summary

Since pain was identified as a "fifth vital sign" by the American Pain Society in 1987, much progress has been underway in developing reliable and objective pain scales for animals, including horses (Van Loon and van Dierendonck 2018). Pain in animals is difficult to assess, as they cannot communicate with humans in a "meaningful manner" (Dalla Costa et al. 2014), but pain scales relying on behavioral and physiological measurements have been developed to bypass this limitation. Behavioral measurements can include body, posture, and gait parameters, interactive behavior, but also facial expressions.

Pain scoring scales have been developed for assessing various pain conditions in horses. In order to assess the accuracy of a pain scale, it is very important to evaluate validity of the scale, or whether the test measures what it should measure, and reliability, or how repeatable the test results are, among observers and within observer (Van Loon and van Dierendonck 2018). Composite pain scales are a combination of several single descriptive parameters that describe pain; each individual parameter is graded individually on a scale of 0-2, for example, and all scores for individual parameters are summed up to one final pain score. Composite pain scales based on whole body behavior were initially developed and validated for acute pain conditions such as acute orthopedic pain (Bussieres et al. 2008), acute colic pain (Van Loon and van Dierendonck 2015), post-castration pain (Taffarel et al. 2015), postoperative pain after emergency gastrointestinal surgery (Van Loon et al. 2014), and clinical pain (Gleerup and Lindegaard 2016). The most accurate pain readings for box-rested, or stationary, horses have used facial expression-based pain scales, which are a type of composite pain scale based on facial expression parameters, validated for post-castration pain (Dalla Costa et al. 2014), laminitis (Dalla Costa et al. 2016), and orthopedic pain in ridden horses (Mullard et al. 2017; Dyson et al. 2018a,b), among others.

In this study, we were focusing on the facial-expression based pain scale Equine Utrecht University Scale for Facial Assessment of Pain (EQUUS-FAP), which has previously been validated for box-rested horses with acute colic pain, acute and post-operative head-related pain, pain after orthopedic surgery and acute orthopedic trauma (Van Loon and van Dierendonck 2015, 2017; Van Dierendonck and van Loon 2016). Although EQUUS-FAP has been validated on live observations of stationary horses, it has not been validated on videos and photos of stationary horses, though other studies have done so with successful results for other diseases/conditions (Dalla Costa et al. 2014; Gleerup et al. 2015). EQUUS-FAP has also been used before on trotted horses with live observations (J.P.A.M. van Loon, unpublished data), though this has not yet been validated and was performed by unblinded observers.

In this study, we wanted to test the reliability and validity of the EQUUS-FAP scale in determining pain scores of trotting lame horses from videos and photos by blinded observers, and find out whether a correlation exists between pain face scores and an asymmetry index found from Qualisys gait analysis, a system that can detect lameness from asymmetry scores of trotting horses. We wanted to use EQUUS-FAP because it was a facial expression pain scale that has previously already been validated for many conditions. We were testing lameness as lameness models have already been validated previously (Merkens and Schamhardt 1988; Carregaro et al. 2014) and importantly, lameness, or an abnormal gait or stance that is a result of a dysfunction of the locomotor system, is a very important condition in horses, often caused by pain (Van Weeren et al. 2017). Figuring out whether video and photo scoring, compared to live scoring, can reliably be used to differentiate between sound and lame horses would provide an easier and potentially less biased method of assessing pain in trotting horses, and would help not only veterinarians but also owners, riders and trainers to reliably recognize pain in moving horses, which may have important welfare consequences. We hypothesized that reliability would be better for pain scoring from photos than videos, and we did not expect a greater difference between sound and induced lameness scores for photos than videos. We also expected a positive correlation between pain scores and the asymmetry index.

We used two lameness models to induce lameness in eight horses, and each horse served as its own baseline control. Lameness was induced using a special modified shoe with a screw, which when tightened, would reversibly induce lameness, and a LPS (lipopolysaccharide) intra-articular injection, which would also reversibly induce lameness. The EQUUS-FAP scale was used to assess lameness in trotting horses during baseline and after lameness induction, using videos recorded from live observations and photos taken from videos (5 clear photos taken from each video). The scale was modified slightly for video and photo scoring, as certain parameters from the original scale were not possible to determine, and others had to be added. Observers included an equine veterinary specialist in anesthesia and a master's student in Biology, both of whom were trained to use the scale. Intra- and inter-observer reliability was calculated for pain scores, and validity was determined to see if pain scores could differentiate between baseline and induced lameness horses, both for total pain scores and for pain scores of individual parameters used in the scale.

It was found that although intra- and inter-observer reliability were acceptable, the EQUUS-FAP scale was unable to differentiate between lame and sound horses for total pain scores. Thus, pain scores did not correlate with asymmetry scores from Qualisys analysis, and this study does not support the clinical application of this pain scale using video and photo coding for horses in locomotion with lameness.

Several reasons were speculated for why signs of pain were not picked up after induction of lameness in the trotting horses. Other states such as stress, fatigue, or fear could have influenced facial pain expressions, making it difficult to conclude that any observed pain scores were due to lameness; some observer bias may have influenced results, though we tried to minimize bias by randomizing and blinding the photos and videos; the EQUUS-FAP scale may have been missing subtle elements of pain expression for moving horses as it was originally validated for stationary horses; the quality of videos and photos were not good enough to detect more subtle signs of pain; coat color can influence facial pain expression reading and was not considered when choosing horses for the study; and a “cannot see” parameter was not included, which perhaps led to inaccurate coding of parameters when it was not clear what score a parameter should get.

There were several limitations to the study, including that lameness was induced and not naturally occurring, the quality of video recordings and photos was not optimal, there could have been observer influence on the horse when trotting, and coat color was not considered, for example. In the future, this study could be redone to see if the same results are acquired with better quality videos and photos, and stress should be controlled for as much as possible. Furthermore, body- and gait- related (facial) parameters could potentially be included in the scale if facial expressions cannot validly be used to differentiate between sound and lame horses during trot.

II. Abstract

The Equine Utrecht University Scale for Facial Assessment of Pain (EQUUS-FAP) was developed for box-rested horses and validated for acute colic pain, acute and post-operative head-related pain, pain after orthopedic surgery and acute orthopedic trauma. Although it has been validated on live observations of box-rested horses, it has not been validated on videos and photos of box-rested horses, though other studies have done so with successful results for horses after castration or with acute laminitis. It has also not been validated on ridden horses in locomotion, though previous studies have focused on pain expression in ridden horses. EQUUS-FAP has also been used before on trotted horses with live observations, though this has not yet been validated. In this study, the goal was to test the reliability and validity of the EQUUS-FAP scale (modified for video and photo coding) in determining pain scores of trotting lame horses from videos and photos, and to find out whether a correlation exists between pain face scores and an asymmetry index found from Qualisys gait analysis. It was hypothesized that repeatability would be better for pain scoring from photos than videos, and it was not expected there would be a greater difference between baseline and induced lameness scores for photos than videos. It was also hypothesized that there would be a positive correlation between pain face scores and the asymmetry index. Lameness was induced in eight horses using two previously validated models (special modified shoe and LPS injection), and pain scores for each horse were quantified after randomization of videos and photos during baseline and after induction of lameness by two blinded observers, an equine veterinary specialist in anesthesia and a master's student in Biology, both of whom were trained to use the scale. Results from using the modified EQUUS-FAP pain scale for video and photo scoring indicated acceptable reproducibility (intra- and inter-observer), though a significant increase in pain-related expressions during guided trot after induction of lameness compared to baseline was not found; consequently, no correlation was found between pain scores and Qualisys asymmetry scores. It was concluded that this study does not support the clinical application of this pain scale using video and photo coding for horses in locomotion with lameness. Several explanations were speculated, including effects of other states such as stress, fatigue and fear on pain scores, the scale missing subtle elements of pain expression for moving horses, and the quality of videos and photos not being adequate to detect more subtle signs of pain. Future studies will have to confirm whether facial parameters could be used to assess pain status in lame moving horses, or whether gait and body parameters, for example, should be included.

III. Introduction

1. Pain in animals

Pain was identified as a “fifth vital sign” by the American Pain Society in 1987, and since then, much progress has been made in objectively and reliably assessing pain in animals (Van Loon and van Dierendonck 2018). Pain in animals, as in humans, contains an emotional component, thus making objective pain evaluation extremely difficult (Rutherford 2002). In animals, it is especially difficult to measure their subjective experience, so pain assessment must be a “value judgment relying on behavioural and physiological indices to provide indirect evidence of mental state” (Hansen 1997; Molony and Kent 1997).

As pain expression is different in every species and depends on the type and origin of pain, an ideal pain scoring system must be “linear, weighted, sensitive to pain-type, breed- and species-specific, less dependent on the observer and closed to misinterpretation” (Ashley et al. 2005). An ideal pain scale should also be “easy to use, include parameters giving repeatable interpretation from one evaluator to another, and provide constancy in the results obtained” (Bussieres et al. 2008). To make pain evaluation in animals as “objective and consistent” as possible, Gleerup and Lindegaard (2016) mentioned several key points, including first determining which behavioral and physiological parameters from the species of focus would be associated with pain. Also important is the possibility of having instantaneous results from behavioral observations, as opposed to waiting potentially days for lab results to come back from certain physiological measurements. Second, Gleerup and Lindegaard (2016) mentioned organizing observations of these signs and third, specifying the amount of pain related to a certain type of behavior. Last, these observations would then be combined into a quantitative score. Such a score could be “useful for estimating pain intensity over time; thereby determining any potential need for analgesic treatment and detecting the effects of this treatment” (Gleerup and Lindegaard 2016). Assesssing the animal’s pain and treating it based on validated pain scales can have positive implications for the animal’s welfare.

1.1 Pain in horses

It is difficult to assess pain in animals, including horses, because they cannot communicate with humans in a “meaningful manner” (Dalla Costa et al. 2014). Prey animals such as horses have evolved to suppress any pain in the presence of a possible “predator” such as a human, making it further difficult to assess pain (Dalla Costa et al. 2014; Taylor et al. 2002). As in other animals, for horses, influences from breed, inter-individual variation, environmental characteristics and drugs on pain expression may be considerable (Wagner 2010; Flecknell 2000a).

It is important to consider whether a horse’s aggressive behavior is caused by an unpleasant situation, or due to an underlying painful condition. As a prey animal, a horse’s normal response to pain is “flight”; when confined, a horse can only respond in an “aggressive behavioural attack at the pain source or threat” (Casey 2004). Aggression is often associated with pain in horses, and can be a “genuine pain response to palpation, as a fear response in anticipation of the pain-related stimulus, or through learned association, such as linking their own offspring parturition pain” (Ashley et al. 2005; Juarbe-Diaz et al. 1998). Thus, although it can be difficult to read why a horse is displaying aggressive behavior, it is important to find the cause, as it can be in response to a painful condition.

It is also important to consider whether the horse is hungry, as food-seeking behavior, which can include pawing, aggression towards neighbors, head nodding, and mouth movements, can often be thought to be related to pain when it is really hunger (Gleerup and Lindegaard 2016). Likewise, it is important to “differentiate changes in facial expressions due to pain from changes due to stress, analgesics, anaesthetics and other interfering factors, such as influence of humans” (Love 2009; Seibert et al. 2003; Ashley et al. 2005; Gleerup et al. 2015).

Pain scales in horses have been developed for pain in general, acute colic/visceral pain, orthopedic pain (laminitis, synovitis), and post-surgical pain (post-castration, post-abdominal surgery), though all have focused on acute pain (Gleerup and Lindegaard 2016; Van Loon and van Dierendonck 2015, 2016; Mullard et al. 2017; Dalla Costa et al. 2014, 2016), which will be discussed further. These are significant acute pain conditions in horses, as acute laminitis, for example, is considered “one of the most painful conditions a horse can experience” (de Grauw and van Loon 2016). Castration is the most commonly performed surgery, and “associated with significant peri- and post-operative pain” (Love et al. 2009).

1.2 Influence of personality, stress, or coping style on pain

It has been found that horses display different facial expressions based on different emotional states. For example, in a negative emotional state such as fear, horses, as other mammals, have changes in ear posture (particularly increase in tension) and tension in chewing muscles (Boissy et al. 2011; Defensor et al. 2012). Because pain is “multifaceted,” Grandin and Mark (2002) mentioned “it is likely that other negative affective states (e.g. fear, anxiety) can be associated with it.” A review by König van Borstel et al. (2017) showed a “close relationship between behavioural indicators of stress and/or pain and/or conflict and/or anxiety” (Reid et al. 2017). This makes it difficult to assess the influence of pain on behavior, as it could lead to misinterpretation of pain for a negative affective state (Van Loon and van Dierendonck 2018).

Dalla Costa et al. (2017) completed a study on the influence of positive and negative emotional states (i.e. anticipation of food response, fear) on pain assessed using their Horse Grimace Scale (HGS), and found that affective state did not influence their HGS scale, which was pain-specific. Nonetheless, it is important to consider horse personality and coping styles, as that “may have major implications for the accurate assessment of pain in horses” (Ijichi et al. 2013). Ruling out influences of affective state on pain readings is an important consideration that must be taken when assessing pain scales for validity and reliability.

1.3 Pain-related parameters

Pain in horses has most commonly been measured using physiological, endocrine, and behavioral measures. Several studies have concluded that “physiological parameters are weakly associated with pain and that behavioural changes are often easier to evaluate and considered more pain-specific,” as well as the “most useful pain indicators for pain evaluation in horses” (Gleerup and Lindegaard 2016; Graubner et al. 2011; Raekallio et al. 1997; Price et al. 2003). Behavioral evaluations do not require expensive equipment (Gleerup and Lindegaard 2016), inflict minimal to no pain or stress, depending on how the evaluations are done, and results are available sooner than if lab tests for physiological parameters, for example, were to be taken and processed. It has also been found that behavioral changes may offer “the strongest indication of the presence, localization and severity of the pain” (De Grauw and van Loon 2016). Behavioral parameters will be focused on here, with some mention of physiological and endocrine parameters.

Behavioral changes often include “elements of demeanour, posture and gait, as well as interactive behaviour” (De Grauw and van Loon 2016). Depending on the type of pain, it can be accompanied by pain-specific behaviors and potential changes in physiology (Gleerup and Lindegaard 2016). For example, in horses with orthopedic pain, there can be decreased weightbearing on the affected limb/foot (Jones et al. 2007; Bussieres et al. 2008); in colic pain, there can be pawing, flank watching and/or rolling (Graubner et al. 2011; Sutton et al. 2012); and in general pain, there can be restlessness, depression with decreased physical activity, decreased appetite, decreased interest in socialization, standing with the head lowered at the back of the box-stall, no interest in surroundings, self-mutilation, etc., depending on the individual (Raekallio et al. 1997; Price et al. 2003; Pritchett et al. 2003; Lindegaard et al. 2010; Jones et al. 2007; McDonnell 2008).

In fact, “whenever horses display changes in attitude and/or performance, it may be that pain is the underlying cause” (Gleerup and Lindegaard 2016). For example, some horses may have “poor performance and unwillingness to work” which is due to pain, and that may “develop into aggression if pain is not diagnosed and treated in time” (McDonnell 2005). As mentioned previously, aggression can be a sign of any chronic pain, for example due to vertebral problems or hoof pain (Fureix et al. 2010). “Horses tend to be sincere in their behaviour, which means that if a certain type of behaviour is induced by a painful condition, this behaviour quickly returns to normal when pain is eliminated” (McDonnell 2005).

Unlike behavioral parameters, physiological and endocrine parameters are not as effective, sometimes require stressing the animal and can be invasive (unless using telemetric techniques like ECG), have the potential for a delayed result (due to lab tests possibly taking several hours or days to return results, for example), and sometimes have poor specificity, as they do not indicate pain but are related to the effects of bleeding, drug treatment, anesthesia, etc. (Gleerup and Lindegaard 2016). Altered physiology can also be a result of cardiovascular problems, stress, or dehydration, for example, which could explain the poor correlation of physiological parameters with pain seen in most studies (Price et al. 2003; Graubner et al. 2011; De Grauw and van Loon 2016). Nonetheless, physiological and endocrine parameters have been included in pain scales in the past as a supplement to behavioral components (Bussieres et al. 2008). These include heart rate, respiratory rate, blood pressure, serum cortisol, and β-endorphins, among others (Pritchett et al. 2003; Bussieres et al. 2008; Lindegaard et al. 2009; Gleerup et al. 2015; Raekallio et al. 1997).

2. Pain evaluation methods

Although it can be difficult to accurately read pain in horses, several evaluation methods for assessing pain have been developed, which will be further discussed here. Pain evaluation in horses has been found to be most successful when done by people familiar to the horse, in a familiar environment, while maintaining an appropriate distance from the horse or doing video observations (Gleerup and Lindegaard 2016). Horses may perceive unfamiliar humans as predators and change their behavior as a result, as "avoiding predators is a high priority behaviour overriding the awareness of pain" (Caine 1992). Sedation and anesthetic drug residues can also influence the behavior of the horse (Seibert et al. 2003), not allowing for genuine pain readings.

2.1 Analgesic testing

Although not the preferred method, it is possible to treat horses, especially those with "poor performance or riding problems where no obvious diagnosis can be established," with analgesia (NSAID or nonsteroidal anti-inflammatory drug, opioid or the like) to see if an unwanted behavior might be associated with pain (Gleerup and Lindegaard 2016). If the pain/poor performance improves after treatment and returns after the treatment is over, the problem is likely due to an undiagnosed painful condition (Gleerup and Lindegaard 2016). This type of testing has good specificity but poor sensitivity, as "pain cannot be ruled out if the analgesic testing is negative," since it may be the drug (type, dosage, duration) that is insufficient for the specific pain type in question (Gleerup and Lindegaard 2016).

2.2 Mechanical nociceptive threshold testing

A second method of pain evaluation is mechanical nociceptive threshold testing, which can be applied by touching or pressing an area on the horse to see if it reacts (Gleerup and Lindegaard 2016). Usually, these tests are looking at hyperalgesia, or lowering of the threshold to a nociceptive stimulus. This can be nicely measured in experimental conditions, where the animal's nociceptive threshold can be tested before and after induction of a painful condition. However, for clinical pain, the nociceptive threshold of an animal in pain must be compared to reference values of groups of healthy animals, which have been previously tested to determine the "normal" nociceptive thresholds of healthy animals. The level of reaction of the horse in response to palpation does not necessarily correspond to the level of pain the horse has when the area is untouched (Gleerup and Lindegaard 2016), as the pain may have multifaceted causes and the horse may be hypersensitized to pain.

These methods can be applied in addition to pain scoring scales. For example, pressure algometry, or "applying controlled pressure to a given body point" (Pelfort et al. 2015), has been used in studies with back pain (Haussler and Erb 2006). Also, von Frey filaments for analgesic testing have been used after epidural treatment in experimental settings (Redua et al. 2002). Both have been used to determine nociceptive thresholds after different procedures or analgesic treatments.

2.3 Time/activity budget analysis

Another method to measure pain is using a time/activity budget analysis, where 'activity budgets,' or time horses spend on a specific behavior, can be calculated from live observations or video recordings (Pritchett et al. 2003). Live observations can be conducted at separate time points, or done continuously. Similarly, videos can be "sufficiently long clips of film at separate time points" or continuous recordings (Pritchett et al. 2003). The previous reduces the time needed for recording and analysis, while the latter may have greater sensitivity for picking up pain-related behaviors (Price et al. 2003). Although this type of analysis is sensitive for even mild pain, the right equipment is needed and this testing cannot be done in real time in the clinic (De Grauw et al. 2006).

2.4 Pain scoring scales

Most importantly, pain scoring scales have been developed for assessing pain levels. Studies consistently show that one pain scale may not work for all pain types (e.g. visceral vs. somatic pain, acute vs. chronic pain, nociceptive vs. inflammatory vs. neuropathic pain) (De Grauw and van Loon 2016), thus at least partly explaining the variety of available pain scales. As mentioned previously, for a pain scale to work in practice, it should be easy and not take too long to use, have well-defined parameters that are easy to understand, good inter- and intra-observer agreement, good sensitivity for mild, moderate and severe pain, have a linear relation with pain severity, and be validated for the specific pain type being tested (Wagner 2010; Ashley et al. 2005; Van Loon and van Dierendonck 2018). It was found that "in most published clinical studies of pain scales, concrete and construct validity and reproducibility have been investigated," as "validity and reliability of the pain scale are very important" and "the accuracy of pain assessment scales is directly dependent on their validity and reliability" (Van Loon and van Dierendonck 2018; Dalla Costa et al. 2018). Methods for validating pain scoring tools include assessment of "internal consistency, construct validity, responsiveness, and reliability of the scale" (De Grauw and van Loon 2016). Construct validity is whether the test measures what it should measure, and includes sensitivity, specificity, and reproducibility (assessed by inter- and intra-observer reliability) (de Grauw and van Loon 2016; De Vellis 2003).

So far, the most accurate readings of pain for stationary horses are done by composite pain scales and facial expression-based pain scales when one considers practical applicability, (ease of use, criteria understanding, time for performing pain assessment using a scale, etc.), reliability, and validity of pain scales (Van Loon and van Dierendonck 2018). Pain scales for ridden horses also exist (Mullard et al. 2017), and facial expression-based pain scales for unridden horses in locomotion are only now being developed in this study. First, pain scales for stationary horses in Section 2.4 will be discussed, and later, pain scales developed for moving horses will be discussed (Section 2.5).

2.4.1 Unidimensional pain scales

In humans, unidimensional pain scales are completed by self-rating, and by nature are very subjective (Gleerup and Lindegaard 2016). When unidimensional pain scales are applied to animals or humans that lack the ability to communicate, pain scoring must be done by an observer, which leads to poor inter-observer but good intra-observer agreement (Gleerup and Lindegaard 2016). These scales include first, the Visual Analogue Scale (VAS), which consists of a 10cm horizontal line, from least pain on the left to most on the right, where pain is read off as mm from the zero end of the scale (De Grauw et al. 2006). In horses, the VAS scores depend on the time taken to observe the horse; inter-observer agreement is usually not great, especially for the middle and lower pain levels (De Grauw et al. 2006). An example of a VAS scale is shown in Figure 1A.

A second scale is the Numerical Rating Scale (NRS), which is a horizontal line with preset number tags from 0-10 at equal distances on the line (0 is no pain) (De Grauw et al. 2006). The observer must circle the level of pain they think the horse is in (De Grauw et al. 2006). This is more repeatable than the VAS as it includes a discrete and discontinuous variable, but it is less sensitive for small changes in pain (Ashley et al. 2005). An example is shown in Figure 1B.

Third, simple descriptive scales (SDS) consist of grades from 0-4 or 5, with "each grade defined as specifically as possible in order to improve objectivity and interobserver agreement" (Gleerup and Lindegaard 2016). SDS is used in clinical practice in lameness grading systems such as the AAEP (American Association of Equine Practitioners) lameness score and the Obel score for laminitis (Kester 1991; Obel 1948). Previous studies that have used SDS for specific pain in horses are shown in Table 1. When using SDS, not all components of lameness are covered, overlap between grades can occur, lameness can only be graded at trot, and inter-observer agreement is low (Lindegaard et al. 2010). This is because not all the numerous behavioral components that horses in pain display can be neatly fit into a single five grade scale (Lindegaard et al. 2010). It is best to use a composite measure pain scale, which includes multiple SDS for multiple parameters.

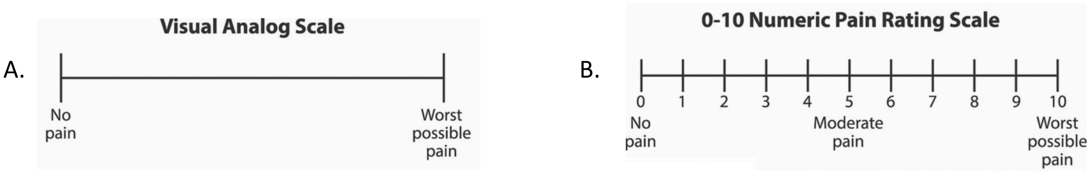
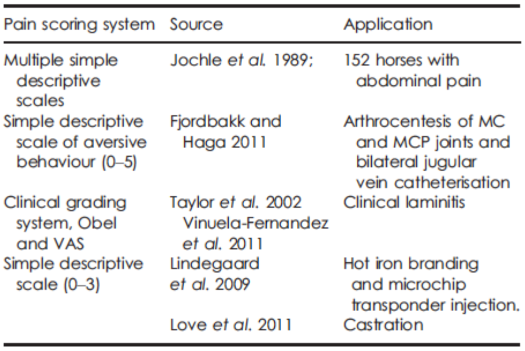


Figure 1: Unidimensional pain scales. Examples of (A) VAS and (B) NRS scales, taken from Daeninck et al. (2016).

Table 1: Simple descriptive scales. Studies that have used SDS for specific pain types in horses, taken from Gleerup and Lindegaard (2016).



Pain scoring system	Source	Application
Multiple simple descriptive scales	Jochle *et al.* 1989;	152 horses with abdominal pain
Simple descriptive scale of aversive behaviour (0–5)	Fjordbakk and Haga 2011	Arthrocentesis of MC and MCP joints and bilateral jugular vein catheterisation
Clinical grading system, Obel and VAS	Taylor *et al.* 2002 Vinuela-Fernandez *et al.* 2011	Clinical laminitis
Simple descriptive scale (0–3)	Lindegaard *et al.* 2009	Hot iron branding and microchip transponder injection.
	Love *et al.* 2011	Castration

2.4.2 Composite pain scales

Composite pain scales are a combination of several SDS, each describing one specific parameter (behavioral, physiological, etc.) in 4-5 clearly defined grades; each SDS is evaluated individually, and all scores are summed up for one final pain score (Gleerup and Lindegaard 2016; Van Loon and van Dierendonck 2018; De Grauw et al. 2006). The parameters can be nonweighted (Bussieres et al. 2008; Sutton et al. 2012; Jochle et al. 1989; Sweeting et al. 1985) or weighted according to their perceived significance (Gleerup and Lindegaard 2016). Since pain is a “complex, subjective multi-dimensional phenomenon evoking emotional, behavioural and physiological responses,” evaluation of several pain-related parameters would be expected to better quantify pain than a single parameter (Dobromylskyj et al. 2000).

The first composite pain scale (CPS) by Bussieres et al. (2008) was essentially a multifactorial SDS based on 13 parameters, including behavioral and physiological parameters, scored for 5 minutes. The scale was validated for acute orthopedic pain, which was induced via an amphotericin-B induced synovitis (lameness) model. The scale was later also validated for postoperative pain after emergency gastrointestinal surgery in horses (Van Loon et al. 2014). Video analysis was used to monitor the behavior (Bussieres et al. 2008). Of all the behavioral parameters to be included in the CPS for orthopedic pain, the most important included posture, response to palpation of painful area, pawing on the floor, head movement, and kicking at the abdomen (Bussieres et al. 2008). The CPS scale by Bussieres et al. (2008) would serve as a starting point for future scales (Van Loon and van Dierendonck 2015).

A second scale, the Equine Acute Abdominal Pain Scale (EAAPS), was later validated for acute colic pain and developed using “formal clinimetric procedures (item generation, item selection, item weighting and testing reliability and validity)” (Sutton et al. 2013a,b). This scale is a “simple ascending clinical index” (Van Loon and van Dierendonck 2018) and not a true composite pain scale; it does not include adding different scores of individual SDS for an overall pain score, making it more a singular pain scale. Equine practitioners assessed pain from film clips of clinical cases of horses exhibiting signs of acute abdominal pain (Sutton et al. 2013a). The scale included 12 behaviors; in the first version, EAAPS-1, one weight was assigned to each behavior, and to grade the severity of the pain the horse was showing, the most severe behavior would be chosen and the score for that behavior would be the pain score. In the second version, EAAPS-2, gradations of weights were assigned based on the frequency of the behavior being demonstrated (Sutton et al. 2013a,b). Sutton and Bar (2016) later presented a “refined and revalidated” version of EAAPS-1, which had good reliability and validity for horses with acute colic when observations were direct, live observations or from video playback. Unlike in CPS (Bussieres et al. 2008), physiological parameters were not included in the EAAPS scales since they are often “controversial” in pain assessment in horses (Ashley et al. 2005). For example, even though heart rate is “the most commonly cited parameter in pain investigation” (Ashley et al. 2005), it was not used because previously it was found to have only a poor to fair correlation with pain in cases of acute colic (Niinisto et al. 2010).

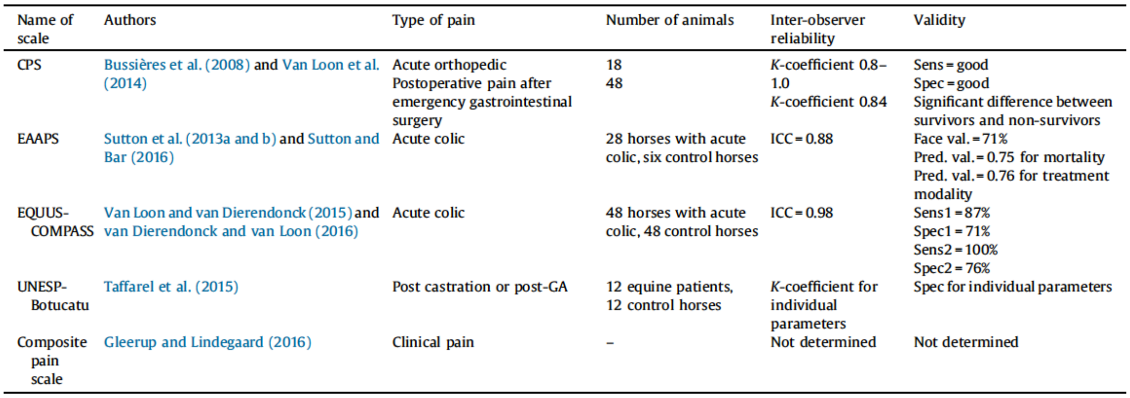
Van Loon and van Dierendonck (2015) later developed a second true composite pain scale, the Equine Utrecht University Scale for Composite Pain Assessment (EQUUS-COMPASS), based on the CPS of Bussieres et al. (2008). As EQUUS-COMPASS was validated for acute colic in horses, the original CPS was modified by “deleting parameters that are not possible to assess in horses with acute abdominal pain (e.g. appetite) and by adding parameters that are thought to be more specific for visceral pain (such as tail flicking, laying down and sounds produced as an expression of pain like teeth grinding or moaning)” (Van Loon and van Dierendonck 2015). The

EQUUS-COMPASS is a multifactorial SDS based on 14 parameters, including physiological, spontaneous behavioral and interactive (responses to stimuli) parameters, scored from 0-3 (0 being no pain) (Van Loon and van Dierendonck 2015). The most sensitive parameters included borborygmi, posture, sweating, and reaction to observer and palpation of painful flank. The scale was done by live observations and was later validated again in a follow-up study with a new cohort of horses with acute colic, controls, and new observers (Van Dierendonck and van Loon 2016). In the follow-up, physiological parameters were removed and the remaining parameters had weighting factors applied, making the scale suitable for horse owners, with high sensitivity and specificity (Van Dierendonck and van Loon 2016).

In the same year, Taffarel et al. (2015) developed the UNESP-Botucatu, a "multidimensional composite pain scale for assessing pain in horses after surgical castration." The scale's inter- and intra-observer reliability were assessed, construct validity evaluated, and the scale was refined to generate the most relevant pain behaviors. These included positioning in the stall, locomotion, locomotion when led by evaluator, response to palpation of painful area, looking at flank, kicking at abdomen, lifting hind limbs, head movement, pawing on floor, and heart rate. Although inclusion of physiological parameters was questioned, heart rate was retained due to it being the "only parameter that varied with time, it is easy to evaluate and has historical importance in the assessment of pain" (Taffarel et al. 2015; Ashley et al. 2005).

Last, Gleerup and Lindegaard (2016) created the Equine Pain Scale, a composite pain scale based on all findings to date, including the Equine Pain Face (Gleerup et al. 2015), which will be discussed later. The scale was used in horses with abdominal and orthopedic pain, but has not yet been validated for a specific type of pain, and repeatability, validity and reliability for the scale have not been analyzed yet. Since most composite pain scales included "evaluation of either subjective pain or some facial features," the Equine Pain Face was included in the scale (Gleerup and Lindegaard 2016). Physiological measures were excluded due to their "invasiveness, poor specificity and the potential for a delayed result" and the weighted behavioral measures included were gross pain behavior, activity level, position in stall, posture/demeanor, weightbearing, head position, head movement, attention towards painful area, interactive behavior and appetite (Gleerup and Lindegaard 2016). A summary of studies that used the discussed pain scales for specific types of pain is included in Table 2.

Table 2: Composite pain scales. Comparison of different composite pain scales, taken from Van Loon and van Dierendonck (2018). EAAPS is not a true composite pain scale and more a singular pain scale, but is included here nonetheless.



Name of scale	Authors	Type of pain	Number of animals	Inter-observer reliability	Validity
CPS	Bussières et al. (2008) and Van Loon et al. (2014)	Acute orthopedic Postoperative pain after emergency gastrointestinal surgery	18 48	*K*-coefficient 0.8–1.0 *K*-coefficient 0.84	Sens = good Spec = good Significant difference between survivors and non-survivors
EAAPS	Sutton et al. (2013a and b) and Sutton and Bar (2016)	Acute colic	28 horses with acute colic, six control horses	ICC = 0.88	Face val. = 71% Pred. val. = 0.75 for mortality Pred. val. = 0.76 for treatment modality
EQUUS-COMPASS	Van Loon and van Dierendonck (2015) and van Dierendonck and van Loon (2016)	Acute colic	48 horses with acute colic, 48 control horses	ICC = 0.98	Sens1 = 87% Spec1 = 71% Sens2 = 100% Spec2 = 76%
UNESP-Botucatu	Taffarel et al. (2015)	Post castration or post-GA	12 equine patients, 12 control horses	*K*-coefficient for individual parameters	Spec for individual parameters
Composite pain scale	Gleerup and Lindegaard (2016)	Clinical pain	–	Not determined	Not determined

ICC, intra class correlation coefficient; face val., face validity; pred. val., predictive validity; sens, sensitivity; spec, specificity; sens1, sensitivity for differentiation between horses with colic and healthy control horses; spec1, specificity for differentiation between horses with colic and healthy control horses; sens2, sensitivity for differentiation between conservative and surgical treatment of horses with colic; spec2, specificity for differentiation between conservative and surgical treatment of horses with colic; GA, general anesthesia.

As described above, many different groups have developed similar composite pain scales, sometimes at similar times. Some groups have validated their scales for similar painful conditions, such as Sutton et al. (2013a,b) and Van Loon and van Dierendonck (2015, 2016) for acute colic, and sometimes for different pain conditions, such as Taffarel et al. (2015) for post-castration or post-general anesthesia pain. Moving forward, it would be wise to come up with a uniform pain scale, combining elements from the pain scales developed by different groups, or use an already validated existing pain scale, and validate it on other pain conditions. There is much overlap in the already existing pain scales, and instead of focusing on developing completely new pain scales for other painful conditions, it would be more efficient to validate the already existing pain scales for other conditions, slightly altering certain parameters if necessary. Such an approach of validating already existing pain scales for other pain conditions or other ways of observing pain (video and photo vs. live observations), slightly altering the pain scales if necessary, will be taken with this study for facial expression-based pain scales, as will be discussed next.

2.4.3 Facial expression-based pain scales

In horses, facial expressions have been "described to be valid indicators of emotional states" (Hintze et al. 2016). Facial expression-based scales were proposed to evaluate more "subtle alterations in pain expressions" and to detect "mild rather than only overt pain" (de Grauw and van Loon 2016; Langford et al. 2010). They also offer benefits that traditional composite pain scales lack, including: 1) being less time-consuming, 2) observers can quickly and easily be trained and there is high reliability within and between observers, 3) "grimace scales" focus on the natural human instinct to focus on the face/head of the animal when assessing/scoring for pain, 4) grimace scales can assess pain ranging from mild to severe, and 5) they are safer for the observer, as it does not require approaching the animal or palpating a painful area (Langford et al. 2010; Keating et al. 2012; De Grauw et al. 2006; Williams 2002; Leach et al. 2011; Dalla Costa et al. 2014). As a result, facial pain scales seem to be "very promising for valid and quick pain assessment in box-rested horses with acute pain from various origins," and they can be used in daily clinical practice (Van Loon and van Dierendonck 2018). However, it is yet uncertain how effective they are in assessing pain of moving horses without also considering body or gait parameters, for example. It may be hard to consistently follow the horse's face when it is trotting, or to get accurate readings of its facial expressions. Also, although facial expression-based scales can be made to be as objective as possible, physiological or more invasive measures of pain would be useful regardless to measure pain more objectively, though due to the invasive nature of such measures they are not as feasible as looking at just behavior.

Previously, facial expressions in horses have been "based on observations by experienced horse practitioners, rather than on systematic investigations" (Taylor et al. 2002). Recently, FACS (Facial Action Coding Systems), "objective coding systems for describing facial behaviour" (Wathan et al. 2015), have been developed for horses. "Individual action units are caused by contraction or relaxation of one or more facial muscles" and FACS provides "a systematic methodology of identifying and coding facial expressions on the basis of underlying facial musculature and muscle movement" (Wathan et al. 2015); it is used to determine if a "combination of facial action units is involved in a certain type of emotion, such as expression of pain" (Wathan et al. 2015). FACS was originally developed for humans (Ekman and Friesen 1976) as "grimace scales" to assess changes in the muscles of the face, or changes in "action units" (Wathan et al. 2015). They have also been developed for chimps (Vick et al. 2007), orangutans (Caeriro et al. 2012), macaques (Parr et al. 2010), gibbons and siamangs (Waller et al. 2012), dogs (Waller et al. 2013), cats (Caeiro et al. 2013), rodents (Mouse Grimace Scale by Langford et al. 2010; Rat Grimace Scale by Sotocinal et al. 2011), rabbits (Keating et al. 2012), sheep (McLennan et al. 2016), cattle (Gleerup et al. 2015a), pigs (Viscardi et al. 2017) and relevant to this study, horses (Wathan et al. 2015).

Wathan et al. (2015) created EquiFACS (Equine Facial Action Coding System) for horses to record all "potential facial configurations" of different emotional states, and in "different social contexts," not just a single state such as pain. EquiFACS is based on 17 total "action units (AUs)," each representing "contraction or relaxation of one or more facial muscles." It provides a systematic and objective way of coding facial expressions based on facial (mimetic) musculature and muscle movement for all emotional states, and it describes the individual or groups of muscles that evoke such expressions. EquiFACS was developed by "anatomical investigation of the underlying musculature" from high quality video, and "discrete facial movements were identified and described in terms of the underlying muscle contractions" (Wathan et al. 2015). These facial "action coding systems" were found to be "very similar for various breeds of horses," indicating that different breeds of horses show similar patterns of facial expression when in pain (Van Loon and van Dierendonck 2019). The system showed high reliability for others to learn the system, including those with no previous experience with horses (Wathan et al. 2015). The study by Wathan et al. (2015) was focused on facial expressions for all emotional states, while other studies focused on using facial action units to decipher exclusively pain expression.

The first facial expression-based pain scale, the Horse Grimace Scale (HGS), was developed by Dalla Costa et al. (2014). This is a composite SDS based on 6 facial expression parameters scored 0 to 2 (0 is no pain). The facial expression parameters are "facial action units (FAUs)," or changes in muscle/observations of the face, including "stiffly backwards ears, orbital tightening, tension above the eye area, prominent strained chewing muscles, mouth strained and pronounced chin, strained nostrils and flattening of the profile." Dalla Costa et al. (2014)

developed the HGS for horses undergoing surgical castration, and found it had good reliability, validity, and correlated well with the CPS (from Bussieres et al. 2008). Pain scoring was based on photos selected from videos, taken at different times before and after castration. In a follow-up study, Dalla Costa et al. (2016) tested the HGS and Obel grade (used to determine severity of laminitis) on horses with acute laminitis, using videos, and found it had good reliability, though validity needs to be further tested. The HGS was not validated by a second dataset or from direct live observations yet, so its clinical applicability is limited (Van Loon and van Dierendonck 2018).

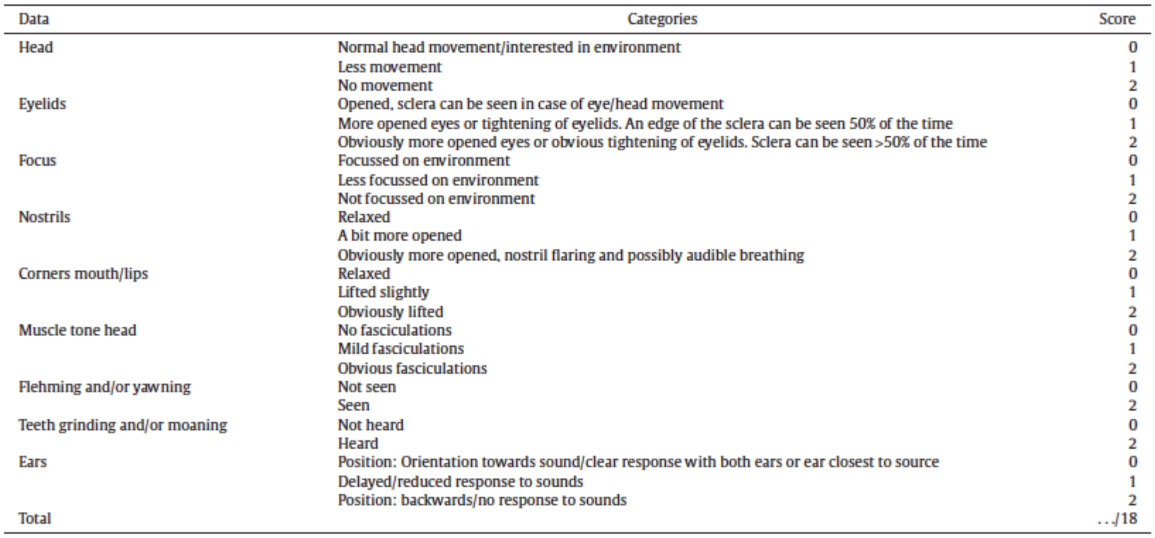
Another follow-up study (Dalla Costa et al. 2017) found that "influences of emotional states other than pain such as new environment, grooming and anticipation of food reward did not significantly change the HGS scores in horses that were not in pain," though "fear increased HGS scores slightly." In other words, negative emotions such as fear or stress only influence facial expressions in a limited way, including facial expressions of pain. Thus, facial expression of horses in pain may be more genuine to their pain state than previously thought, as the influence of fear and stress was found to be minimal.

A second facial expression-based pain scale developed solely to look at pain was developed by Gleerup et al. (2015). The Equine Pain Face, as the HGS (Dalla Costa et al. 2014), is also based on 6 facial action coding units, and was "validated with two experimentally induced pain models (a tourniquet on the antebrachium and topical application of capsaicin) in six healthy pain free animals." Both pain inductions produced an "acute, moderate and reversible pain reaction" (Gleerup et al. 2015). Parameters were assessed via live observations, videos, and photos from videos, and pain scores were compared to heart rate and CPS (Lindegaard et al. 2010). Compared to baseline pain-free conditions, during pain sessions with the observer present, horses had increased contact-seeking behavior and did not suppress changes in facial expressions much, though expressions were "less pronounced whenever the horses tried to interact with the observer" (Gleerup et al. 2015). The pain face included: 'low' and/or 'asymmetrical ears' and facing sides (outward rotation), angled appearance of eyes, withdrawn and/or tense stare, mediolaterally dilated nostrils (square-like), and tension of lips, chin, and certain facial/mimetic muscles (Gleerup et al. 2015). Although the scale seems promising, it still needs to be tested for sensitivity, intra- and inter-observer reliability, be reproduced and use a larger sample size. Van Loon and van Dierendonck (2018) also mention that the experimentally induced pain "limits the clinical applicability of this scale," as clinical pain states are not completely mimicked by the experimentally induced pain states, and that it would be best to test this scale on specific pain conditions.

At the same time, another facial expression-based pain scale, the Equine Utrecht University Scale for Facial Assessment of Pain (EQUUS-FAP), was designed for horses with acute colic pain by direct live observations (Van Loon and van Dierendonck 2015). EQUUS-FAP is a multifactorial SDS based on 9 parameters, describing different elements of facial expression. It is a "dynamic pain scale," as it comprises facial action coding units and dynamic aspects such as teeth grinding, response to sound, etc. that are "well-known head-related pain behavioural parameters from previous studies" (Van Loon and van Dierendonck 2018; Mullard et al. 2017; Dyson et al. 2017). Each of the parameters is scored from 0 to 2, leading to a total pain score from 0 (no signs of pain) to 18 (maximal pain score). The parameters include: head, eyelids, focus, nostrils, corners mouth/lips, muscle tone head, flehming and/or yawning, teeth grinding and/or moaning, and ears (shown in Table 3). The scale has been found to have good reliability and validity, and was validated with a follow-up study (Van Dierendonck and van Loon 2016) on a new cohort of patients with acute colic and new observers, showing good sensitivity and specificity. The scale discriminated significantly between painful and control horses, surgically vs. conservatively treated horses, and for monitoring over time. EQUUS-FAP can also be used by horse owners and non-veterinarians after being trained on using the scale, as it does not include physiological parameters (Van Loon and van Dierendonck 2015).

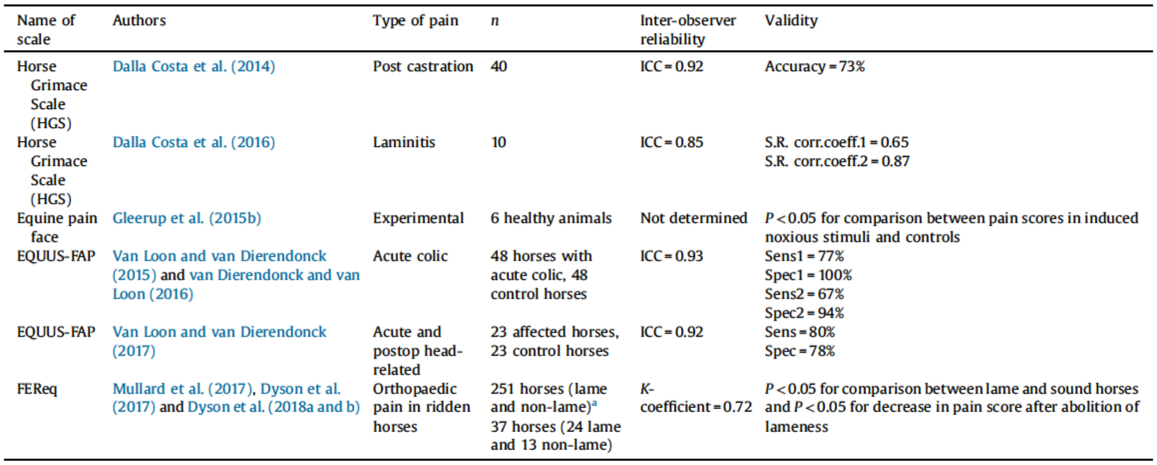
In a later study, EQUUS-FAP was also used to assess acute and postoperative head-related pain, including "dental pain, ocular pain, or trauma to the skull," using live observations (Van Loon and van Dierendonck 2017). The scale still showed good inter-observer reliability, sensitivity, and specificity. Van Loon and van Dierendonck (2019) also found that both CPS (Bussieres et al. 2008) and EQUUS-FAP "were reliable and valid for the [objective and repeatable] assessment of pain in horses after orthopaedic surgery and in horses with acute orthopaedic trauma" when using live observations. Both had high inter-observer reliability and showed significant differences between orthopedic cases and controls, independent of horse breed. Also for both scales, horses in trauma cases had "significantly higher pain scores" than postoperative cases, and both pain scores significantly decreased after administration of NSAID (Van Loon and van Dierendonck 2019). The various facial expression-based pain scales are further shown in Table 4. This excludes the EQUUS-FAP evaluated for orthopedic pain (Van Loon and van Dierendonck 2019) and includes the FEReq (Mullard et al. 2017), which will be discussed next.

Table 3: **Original EQUUS-FAP scale.** Scoring sheet for EQUUS-FAP, the Equine Utrecht University Scale for Facial Assessment of Pain, taken from Van Loon and van Dierendonck (2015).



Data	Categories	Score
Head	Normal head movement/interested in environment	0
	Less movement	1
	No movement	2
Eyelids	Opened, sclera can be seen in case of eye/head movement	0
	More opened eyes or tightening of eyelids. An edge of the sclera can be seen 50% of the time	1
	Obviously more opened eyes or obvious tightening of eyelids. Sclera can be seen >50% of the time	2
Focus	Focussed on environment	0
	Less focussed on environment	1
	Not focussed on environment	2
Nostrils	Relaxed	0
	A bit more opened	1
	Obviously more opened, nostril flaring and possibly audible breathing	2
Corners mouth/lips	Relaxed	0
	Lifted slightly	1
	Obviously lifted	2
Muscle tone head	No fasciculations	0
	Mild fasciculations	1
	Obvious fasciculations	2
Flehming and/or yawning	Not seen	0
	Seen	2
Teeth grinding and/or moaning	Not heard	0
	Heard	2
Ears	Position: Orientation towards sound/clear response with both ears or ear closest to source	0
	Delayed/reduced response to sounds	1
	Position: backwards/no response to sounds	2
Total		.../18

Table 4: **Facial expression-based pain scales.** Comparison of different studies using facial expression-based pain scales done through 2018, including the FEReq (Mullard et al. 2017) pain scale for ridden horses. Table taken from Van Loon and van Dierendonck (2018). Table does not include EQUUS-FAP evaluated for orthopedic pain (Van Loon and van Dierendonck 2019).



Name of scale	Authors	Type of pain	*n*	Inter-observer reliability	Validity
Horse Grimace Scale (HGS)	Dalla Costa et al. (2014)	Post castration	40	ICC = 0.92	Accuracy = 73%
Horse Grimace Scale (HGS)	Dalla Costa et al. (2016)	Laminitis	10	ICC = 0.85	S.R. corr.coeff.1 = 0.65 S.R. corr.coeff.2 = 0.87
Equine pain face	Gleerup et al. (2015b)	Experimental	6 healthy animals	Not determined	$P < 0.05$ for comparison between pain scores in induced noxious stimuli and controls
EQUUS-FAP	Van Loon and van Dierendonck (2015) and van Dierendonck and van Loon (2016)	Acute colic	48 horses with acute colic, 48 control horses	ICC = 0.93	Sens1 = 77% Spec1 = 100% Sens2 = 67% Spec2 = 94%
EQUUS-FAP	Van Loon and van Dierendonck (2017)	Acute and postop head-related	23 affected horses, 23 control horses	ICC = 0.92	Sens = 80% Spec = 78%
FEReq	Mullard et al. (2017), Dyson et al. (2017) and Dyson et al. (2018a and b)	Orthopaedic pain in ridden horses	251 horses (lame and non-lame)[a] 37 horses (24 lame and 13 non-lame)	*K*-coefficient = 0.72	$P < 0.05$ for comparison between lame and sound horses and $P < 0.05$ for decrease in pain score after abolition of lameness

S.R. corr.coeff.1, Spearman Rho correlation coefficient between Horse Grimace Scale and Obel lameness score; S.R. corr.coeff.2, Spearman Rho correlation coefficient between Horse Grimace Scale and pain intensity, evaluated by veterinarians; Postop, postoperative; ICC, intra class correlation coefficient; Sens, sensitivity; Spec, specificity; Sens1, sensitivity for differentiation between horses with colic and healthy control horses; Spec1, specificity for differentiation between horses with colic and healthy control horses; Sens2, sensitivity for differentiation between conservative and surgical treatment of horses with colic; Spec2, specificity for differentiation between conservative and surgical treatment of horses with colic.

[a] Lame and non-lame horses (n=150) in the study by Mullard et al. (2017); 101 horses (76 lame and 25 sound; 7 lame horses before and after diagnostic analgesia) in the study by Dyson et al. (2017); and 37 horses (24 lame and 13 non-lame horses) in the study by Dyson et al. (2018a,b).

2.5 Pain scoring scales in moving (ridden) horses

The facial expression-based pain scales previously mentioned were for box-rested horses in pain. Mullard et al. (2017) developed a scale for facial expressions of ridden horses (FEReq), the first pain scale for horses in locomotion, specifically, for ridden horses. Their goal was to develop and test an ethogram to describe facial expressions in general in ridden horses, not just focusing on pain/stress, which would be covered in future studies.

The study was restricted to analysis of still photographs from ridden horses, and observers were from different backgrounds, ranging from veterinarians to amateur horse owners. As ridden vs. nonridden horses have differences in facial expressions influenced by the "bit, a restrictive noseband, contact via reins to the rider, alteration of head posture, possibly by force, the influence of physical exertion relative to the fitness of the horse, and the skill and even weight distribution of the rider" (Manfredi et al. 2009; Casey et al. 2013; Eiserio et al. 2013; McLean and McGreevy 2010), this affects the scoring of the mouth, tongue, head posture and position of nostrils. The ethogram included features of the eyes, ears, mouth, nostrils, tongue but also muzzle and head position relative to the vertical. It was found that there was no difference in scoring facial expressions of ridden horses based on assessors' professional backgrounds, and the scoring had good consistency and repeatability (Mullard et al. 2017). Though they tested reliability, the authors concluded that future work would have to be undertaken to determine if lame horses could be differentiated from sound horses based on this ethogram.

In a follow-up study by Dyson et al. (2017), the FEReq was assessed in lame ridden horses before and after diagnostic local anesthetic blocks for lameness. The goal was to determine if based on facial expressions the FEReq could discriminate between lame and non-lame horses (thus testing validity of the scale), and whether the FEReq could be adapted to a pain scoring system for ridden horses. Still photographs of the head and neck were used from ridden horses. Images of control horses were acquired during training or warming up in trot and canter at international dressage competitions, and images of lame horses were acquired for "reasons other than this study" during lameness assessment at Animal Health Trust (AHT). Horses with at least 10 images each of "adequate quality" only were included, though the study does not cite what "adequate quality" means, or what exactly was done to take photographs with the least bias possible. Pain scores for each photo were applied for each parameter in the FEReq ethogram, based on previous studies. The results indeed showed significantly higher pain scores for lame than sound horses, and the best indicators of pain were the following: position "severely above the bit, twisting the head, asymmetrical position of the bit, ear position (both ears backward, one ear backward and one to the side, as well as one ear backward and one ear forward), and eye features (exposure of the sclera, the eye partially or completely closed, muscle tension caudal to the eye, and an intense stare)" (Dyson et al. 2017). Since the study only used photos instead of video recordings, the results do not reflect "dynamic changes in behavior" (Dyson et al. 2017). Also, "the influence of circumstances and rider-horse interaction were not randomly distributed over lame and control horses" (Van Loon and van Dierendonck 2018). Nonetheless, the study shows a "potential of changes in facial expression to detect subtle levels of lameness in ridden horses" (Van Loon and van Dierendonck 2018).

Dyson et al. (2018a) attempted to address some of these limitations by using video footage in a follow-up study to develop a pain scoring system in ridden horses. Besides developing an ethogram for whole-horse behavior, they wanted to determine whether it could be applied repeatedly by one observer, and if the pain behavior score could differentiate between lame and sound horses. The ethogram for musculoskeletal pain would include 24 whole-horse behavioral markers for ridden horses, some adapted from the FEReq ethogram (Mullard et al. 2017; Dyson et al. 2017). Behavioral markers included facial markers (from FEReq ethogram), body markers (head posture and movement, tail position and movement), and gait markers (speed, regularity of rhythm, responsiveness, bucking, rearing, and sudden stops). Body and gait related markers showed the most pronounced differences between lame and sound horses, and horses in pain were more likely to show *more* behaviors and more *severe* behaviors (Figure 2). Facial related behavioral markers also showed significantly higher pain scores for lame than sound horses, as for body and gait related behavioral markers. Behaviors that occurred more frequently for lame horses included for facial expressions, ears back, mouth opening, tongue out, and change in eye posture and expression; for body markers, head tossing and tilting head; and for gait markers, unwillingness to go, crookedness, hurrying, changing gait spontaneously, poor quality canter, resisting, stumbling, and toe dragging. The scale had good intra-observer reliability, and a pain score for the whole body (facial, body, gait) indeed showed a much larger overall difference in score between lame and sound horses than just a facial pain scale.

A follow-up study was done by Dyson et al. (2018b) to compare the results of application of the ridden-horse ethogram (Dyson et al. 2018a) by "trained and untrained assessors to horses before and after musculoskeletal pain had been substantially improved using diagnostic analgesia." Anonymized video recordings of horses ridden by professional riders in trot and canter, before and after diagnostic analgesia abolished lameness, were scored in random order using the ridden-horse ethogram by a trained assessor and 10 untrained assessors. As was found in Dyson et al. (2018a), significant differences were found in "facial, body and gait markers after diagnostic analgesia," and a significant decrease in behavior scores was seen for all assessors after diagnostic analgesia abolished lameness (Dyson et al. 2018b). Agreement between the trained and untrained assessors were moderate when the horse was lame, and nonexistent after diagnostic analgesia, indicating that "assessors find it easier to observe the presence of behaviour than its absence." It was also found that untrained assessors were "more likely, based on behavioural scores, to predict the presence of musculoskeletal pain." Thus, despite the slight differences in agreement between trained and untrained assessors, "the ethogram is a potentially valuable tool for determining the presence of musculoskeletal pain and may be useful for longitudinal monitoring of improvement in lameness" (Dyson et al. 2018b).

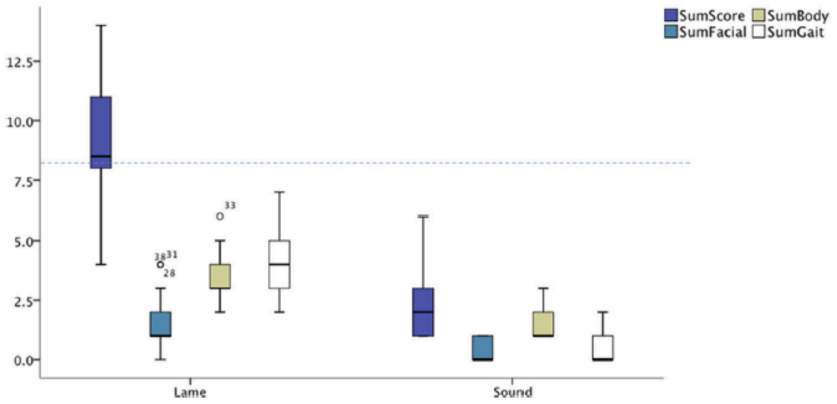


Figure 2. Whole-horse ethogram (facial, body, gait markers) validity results (from Dyson et al. 2018a). The study by Dyson et al. (2018a) found that after summing up behaviors into all behaviors scored and all behaviors into each category scored (facial, body, gait), there was a significant difference in all summed behaviors between lame and nonlame ridden horses, the smallest difference seen in facial markers. Figure shows the occurrence counts for the sum of all behaviors markers (SumScore), sum of facial markers (SumFacial), sum of body markers (SumBody), and sum of gait markers (SumGait). Dotted line is point (score of 8) above which lameness/pain is likely. For individual categories, smallest difference is found between facial markers, greatest between gait markers.

2.6 Pain assessment in orthopedic pain

As orthopedic pain is the focus in this study, the progress that has been made in assessing orthopedic pain in horses will first be summarized. Lameness, or an abnormal gait or stance that is a result of a dysfunction of the locomotor system, is often caused by pain, and is a clinical sign of an underlying orthopedic problem (Van Weeren et al. 2017). Lameness is a symptom of a whole spectrum of different types of problems, such as acute laminitis, an "equine disease characterized by intense foot pain, both acutely and chronically" (Dalla Costa et al. 2016). Associated signs include "inability or reluctance to walk, frequent weight shifting, and abnormal weight distribution on hind feet to relieve the pressure on the front feet" (Dalla Costa et al. 2016). Lameness can also be a sign of acute or chronic synovitis, an equine disease with significant synovial effusion. In severe orthopedic pain, several behavioral expressions can be present in horses, including: decreased appetite, restlessness, depression, abnormal posture, changed weightbearing, pawing, lowered head, repeated head movements, less time spent in front of box-stall, decreased social interaction, pain face, and gross pain behavior (Gleerup and Lindegaard 2016).

Common unidimensional lameness grading scales include the AAEP scale from 0-5 (Kester 1991), NRS from 0-10 (Wyn-Jones 1988), and objective gait assessment techniques (i.e. kinetic (force plate) and kinematic lameness evaluation) (Wagner 2010). The Obel scale has been particularly used for classifying the "severity of lameness due to laminitis by grade from I to IV" (Wagner 2010).

Composite and facial expression-based pain scales for monitoring orthopedic pain include first, the CPS. Bussieres et al. (2008) experimentally induced orthopedic pain in horses by using amphotericin B, which caused acute synovitis and thus, orthopedic pain. Second, the Equine Pain Scale, a composite pain scale combined with the Equine Pain Face, was used in horses with abdominal and orthopedic pain (Gleerup and Lindegaard 2016). Third, Dalla Costa et al. (2016) completed a follow-up study testing both the HGS and Obel grade on horses with acute laminitis, finding that the HGS overcomes many disadvantages that are found within the Obel grading system, including not needing to approach or move the horse. Fourth, EQUUS-FAP was validated for horses after orthopedic surgery and horses with acute orthopedic trauma (Van Loon and van Dierendonck 2019). Last, Dyson et al. (2017) found that the FEReq was able to discriminate between lame and non-lame ridden horses. Though progress is underway for assessing orthopedic pain, non-ridden moving horses must also be tested with these scales, and reliability, validity and repeatability must be measured for use of a scale in the clinic.

3. Current study

Although much progress has been made in developing pain scales for different pain conditions and assessing the reliability and validity in each case, in order to increase practical applicability of these scales in the clinic, Van Loon and van Dierendonck (2018) stress the need for more effort into validating existing pain scales for specific pain conditions. Van Loon and van Dierendonck (2019) state it is also necessary to develop pain scales for moving horses to differentiate between lame and sound horses, and not necessarily just ridden horses as was done by Mullard et al. (2017) and Dyson et al. (2017). This would facilitate “pain assessment during training and competition” (Van Loon and van Dierendonck 2019). Observations should be done by blinded observers, and the accuracy and reliability of pain scoring based on videos and photos (not live observations, as this was already done in Van Loon and van Dierendonck 2019) should be determined. This would confirm whether photos and videos are an appropriate means of determining pain related to lameness in horses, as live observations may not always be possible.

3.1 Why facial expression-based pain scale was used

As was previously mentioned, facial expression-based pain scales provide many benefits over composite pain scales, the greatest being that they are less time-consuming and based on less-extended ethograms than composite pain scales (Van Loon and van Dierendonck 2018). They also produce “valid and reproducible outcomes” among observers, among other benefits (Section 2.4.3).

Gleerup and Lindegaard (2016) mentioned that it is best to “focus all energy on the validation for one robust pain scale, rather than attempting to differentiate between pain types,” and to develop a pain-scoring system that is not disease-specific or for a specific patient group, but works for all horses. Among the facial expression-based pain scales available, it was decided the EQUUS-FAP scale would be used (Van Loon and van Dierendonck 2015) to compare the reliability and accuracy of photos and videos from moving horses with lameness, and to determine if a correlation exists between pain face scores and the asymmetry index. EQUUS-FAP has been validated for the greatest number of pain conditions compared to other existing facial expression-based pain scales, including for acute colic with two cohorts of horses and observers (Van Loon and van Dierendonck 2015, 2016), for acute and post-operative head-related pain (Van Loon and van Dierendonck 2017), and for acute orthopedic pain (Van Loon and van Dierendonck 2019), with positive results for accuracy and reliability. The focus now will be to not only look at pain scoring from live observations, but also from videos and photos to see if the pain scores can differentiate between lame and sound horses.

3.2 Why lameness was tested

Lameness is a very important condition in horses, as the “primary uses of the horse as sport and leisure animal are based on the capacity of its locomotor system” (Serra Braganca et al. 2018). Disorders of the locomotor system, mostly lameness, “are one of the main reasons for equine veterinary consultation” (Nielsen et al. 2014), and equine practitioners spend most of their working time on lameness examinations (Loomans et al. 2007). Lameness also leads to “financial loss for horse owners, days lost in training and/or competition” (Egenvall et al. 2008; 2013). Also, many horse owners are unaware that their horses are lame. It has been found that many “owner-sound” horses “present with objectively measured lameness parameters of the same magnitude or larger than horses thought to have clinically important lameness” (Rhodin et al. 2017), and that 72.5% of horses in training which were perceived free from lameness by the owner actually had “movement asymmetries above previously reported asymmetry thresholds during straight line trot” (Rhodin et al. 2017).

Importantly, induced lameness models have already been validated previously, so those induction models will be used in this study. Clinical patients could have been used instead of inducing lameness in horse subjects, but inducing lameness creates a more controlled population of horses and more homogeneity, which is better for validating the pain scales. Lameness will be induced using two models, the first of which is a sole pressure model for inducing hind and forelimb lameness, validated by Merkens and Schamhardt (1988). Similar to the “effect of a stone in a shoe,” lameness was induced using “modified shoes” where “screws could be turned onto the sole surface, resulting in lameness of adjustable severity” (Merkens and Schamhardt 1988). Removing the screws makes any pain vanish immediately, and “completely normal locomotion” is re-established (Merkens and Schamhardt 1988). This method of inducing lameness causes no injury, minimal distress to the horse, and an “acceptable uniformity in the conditions during the different recording sessions could be realised" (Merkens and Schamhardt 1988).

A second model of lameness induction involves injecting lipopolysaccharide (LPS) into either the radiocarpal joint for front limb lameness or talocrural joint for hind limb lameness, both of which result in acute synovitis (lameness) in the horse (Carregaro et al. 2014; Firth, Seuren and Wensing 1987). This also results in only short-term lameness, with an eventual complete return to normal locomotion (Merkens and Schamhardt 1988). Lameness will be induced using both models, and facial expressions will be blindly assessed during locomotion from videos and photos in lame and control horses before and during lameness induction.

3.3 Validations/reasons for study

In this study, horses had induced lameness using the two above-mentioned models, and their facial expressions were observed live during hand-led trot before and during induction of lameness, and later analyzed and scored from video playback and photos from these videos. Although hand-led trot is not a natural situation, compared to Dyson et al. (2017) who also assessed horses in locomotion, facial expressions were analyzed during trot and not riding. The EQUUS-FAP scale was used by un-blinded, possibly biased observers during real-time guided trot. To reduce the possible bias from direct live observations, an adjusted EQUUS-FAP ethogram for video scoring and photo scoring was used by both a blinded master student in Biology and a blinded equine specialist in veterinary anesthesia.

Figuring out whether video and photo scoring, compared to live scoring, can be reliability used to differentiate between lame and sound horses would first, provide a less biased method of assessing pain in lameness-induced horses during trot. Improved pain recognition, especially during locomotion of lame and sound horses, can “facilitate pain assessment during training and competition,” which can “improve training regimens for competitive horse sports” and most importantly, “benefit equine welfare” (Van Loon and van Dierendonck 2019; Mullard et al. 2017).

Second, “scoring live (i.e., in a clinical context) does not appear to be as straightforward as scoring from images,” (Dalla Costa et al. 2016) and the simpler a method for scoring pain the better, as “maintaining simplicity in pain scoring methods may improve compliance of pain evaluation, thereby potentially optimising pain management for all [horse patients]” (Gleerup and Lindegaard 2016). Scoring facial expressions from videos and photos may be much easier than from live observations, as videos and photos can be replayed, slowed down, and watched when the observer chooses without engaging the horse.

Third, many veterinarians are not trained in pain recognition or assessment of behavior, and have a limited education in identifying “low-grade lameness and recognition of musculoskeletal pain as a cause of poor performance” (Mullard et al. 2017). Similarly, many “owners, riders, and trainers fail to recognize lameness and other pain-related gait abnormalities in ridden horses,” and likely horses in locomotion (Dyson and Greve 2016). Consequently, as lameness and associated subtle expressions of pain are not easily detected, “it might be easier to educate riders and trainers to recognize changes in facial expression and behavior rather than lameness, which may have important welfare consequences” (Dyson et al. 2017). Having a facial pain recognition system for horses in trot, especially for lameness, that is reliable, accurate, easy-to-use and has the least possible bias is beneficial not only for horse welfare, but also for the owners for healthy and effective maintenance of their horse. Although “pain scores will never replace clinical decision making,” they can “aid follow-up of a patient and objectify responses to treatment” (Van Loon and van Dierendonck 2015).

The aim of this study was to see if facial expression scoring by blinded observers, using a modified form of EQUUS-FAP (Van Loon and van Dierendonck 2015) for horses in locomotion, from videos and photos of hand-led trotting horses, can be a reliable method to differentiate between horses that are sound (baseline) and those that are lame (induced via two models), by means of statistical analysis. One goal was to analyze the reliability (repeatability) and validity (accuracy) of coding facial pain scores from the videos and photographs. Another goal was to determine whether pain face scores would correlate with an asymmetry index that had been obtained by means of an objective locomotion analysis from Qualisys. It was hypothesized that repeatability (intra- and inter-observer reliability) would be better for pain scoring from photos than videos. It was not expected there would be a greater difference between baseline and induced lameness scores for photos than videos (a difference in validity between photo and video scoring was not expected). It was also hypothesized that there would be a positive correlation between pain face scores and the asymmetry index.

IV. Methods

1. Ethical considerations

The study design and experimental protocol were approved by the Ethics Committee on the Care and Use of Experimental Animals in compliance with Dutch legislation on animal experimentation (permission number: AVD108002015307WP16, date of approval 18-08-2017).

2. Animals

In this study, 8 Warmblood mares were included. Their characteristics are listed in Table 5. All horses were housed inside during the experimental period. Prior to the experiment, the horses were acclimatized to the environment for 2 weeks, during which time they were handled.

Table 5. **Animals used.** Characteristics of horses included in study (n=8). *One mare was used twice in shoe model.

	Shoe model	**LPS model**
Total number of horses (all mares)	7*	8
Mean (± SD) body weight (kg)	559.10 (±39.0)	559.30 (±36.1)
Mean (± SD) age (years)	7.71 (±2.81)	7.67 (±3.07)
Mean (± SD) height (m)	1.63 (±0.05)	1.63 (±0.04)

3. Experimental design

The practical component of this experiment was completed by other researchers and two veterinary master's students, but will be further elucidated here. Eight mares were used in a two-period randomized crossover design. For each treatment, the time of induction and allocated limbs of the horses were randomly assigned. There was a washout period of at least 7 days in between treatments.

All horses were acclimatized prior to treatment for 2 weeks, during which time they were accustomed to daily handling as well as dewormed and vaccinated if needed. Seven days before the start of the experiment, the horses had their hooves trimmed at the farrier. Three horses already had special shoes for lameness induction placed on their allocated limb, while the rest of the limbs were shoed normally with a shoe of approximately the same weight. Lameness via the special shoe was not induced yet at this point, as the screws were not yet tightened.

The day before and morning before induction of lameness, baseline measurements were taken for each horse. Gait analysis was completed by Q-Horse (Qualisys) marker setup and EquiMoves IMU system, and the gait data were processed by another student. Gait measurements on Q-Horse were collected while horses were walking and trotting in a straight line, guided by a handler, and during lunging to the left and right on hard and soft surfaces. Gait analysis was accompanied by a handheld video recording of the horse trotting in a straight line on the hard and soft surfaces guided by a handler, as well as a video recording taken by following a sensor on the head collar of the horse. These videos would be used for visual analysis of data, as well as pain scoring, which will be discussed further. Only videos taken from straight line trot were used for facial pain expression assessment. Horses were guided during trot by a horse caretaker, and an equine veterinarian stood behind the horse when necessary to urge the horse to move forward and trot.

Live pain scoring assessments, using the EQUUS-FAP scale developed by Van Loon and van Dierendonck (2015) (shown in Table 3), were completed by one of the observers during gait analysis and after lameness induction. As the live observers were not blinded to the horses' conditions, the live observations were potentially biased. These pain scores were not used in this study for analysis, as the focus was on blinded pain scoring from the recorded videos and photos acquired from these videos. After pain measurements were completed, horses were returned to their stalls, where cameras recorded their pain behavior.

3.1 Setup

Each horse served as its own control, and baseline measurements were taken before induction of both methods of lameness. Each horse was treated once with the frog pressure shoe, and once with LPS (lipopolysaccharide) to induce lameness, and the allocated limbs were chosen randomly. For induction of lameness for both the frog pressure shoe model as well as the LPS model, 4 horses had their front limb induced, 2 of which had their right front limb and the other 2 their left front limb induced. The 4 other horses had their hind limb induced, 2 of which had their right and 2 their left hind limb. A diagram of the setup is included in Figure 3.

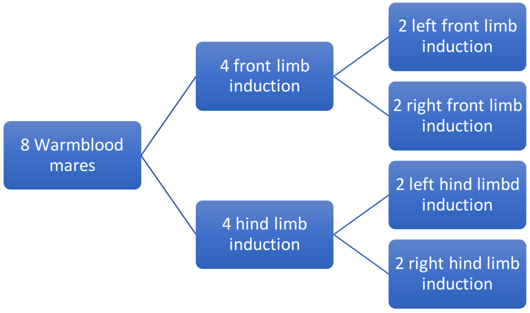

8 Warmblood mares
4 front limb induction
4 hind limb induction
2 left front limb induction
2 right front limb induction
2 left hind limbd induction
2 right hind limb induction

Figure 3. Setup. Setup of both shoe and LPS models of lameness induction.

For the frog pressure shoe model, 4 horses were treated in one hind limb, then after on the opposite side front limb, as only unilateral induced lameness was assessed. The 4 other horses were treated in one front limb, then the other front limb. One horse was not included in the shoe model due to a dangerous situation, so another horse was included twice to compensate for the lack of data from the excluded horse.

3.2 Induction of lameness

a. Special modified shoe

As mentioned previously, the sole pressure model for inducing lameness has been validated previously (Merkens and Schamhardt 1988), making it a valid model to induce lameness in this study. In this study, at least one week before induction of lameness, each horse went to the farrier to get a modified horse shoe with an iron bar and bolt hole forged into the shoe. Tightening the bolt would increase pressure on the sole, inducing lameness. The amount of tightening needed for each horse was determined by analyzing the horse's gait during straight line trot on the hard surface before and after tightening the bolt. This was tested before experimenting with the induced lameness condition.

During experimenting, pain assessment was completed after tightening the bolt on the special shoe 10 minutes after gait analysis was completed, and without tightening the bolt, 30 minutes after gait analysis was completed. Pain assessment was also done the day after induction of lameness during stable resting conditions, though the day after assessments was not included in this study.

b. LPS

Intra-articular injection of LPS has also been validated previously in inducing lameness in horses (Firth, Seuren and Wensing 1987). Before induction with LPS, horses were sedated and the skin over the arthrocentesis location was clipped and prepared for aseptic arthrocentesis. A sample of synovial fluid was collected using a 21g x 40mm needle, and 2ml of LPS solution (with dosage between 0.5-1ng/ml) was delivered into either the middle carpal joint or talocrural joint. The horses were put back into their stables with muzzles until the sedative wore off. Pain assessment was completed 2, 3-5, 6, 5-8, 5, 8, 5-10, 5, 22-24 and 44-48 hours after LPS injection. Lameness was assessed during recovery when the horses were in their stalls in the LPS lameness-induced cases, though these measurements were not included in this study.

c. Rescue analgesia

Health monitoring protocols, including general clinical examination, assessment of limb loading and degree of lameness, were performed every day, 6 times daily during the study. Rescue analgesia consisting of morphine 0.1 mg/kg IV (Centrafarm, the Netherlands) and subsequent meloxicam 0.6 mg/kg IV (Boehringer-Ingelheim, the Netherlands) if morphine did not improve lameness within 1 hour, was administered when lameness was greater than 3 out of 5 on the American Association of Equine Practitioners (AAEP) scale (Kester 1991). If lameness did not decrease to a degree of less than 3 out of 5 within 4 hours after the first meloxicam administration, the humane endpoint would be reached and the horse would need to be euthanized. If lameness did not decrease to maximally 1 out of 5 within 18 hours, another administration of meloxicam would be given. If septic arthritis developed after LPS-administration and was diagnosed by means of synovial fluid sample analysis with leucocyte count and bacterial culture and unresponsiveness to NSAID treatment, the humane endpoint would be reached as well.

3.3 Video and photo data collection

Videos were collected for each horse in each condition, and were about 2 minutes long each, though each horse varied in their stride and starting/ending points on the hard and soft surfaces. To keep the videos for analysis as homogenous as possible, all videos were cut to include a single run of the horse trotting toward the camera. Lengths of the cut videos varied based on the horse's stride, but were approximately 10-12 seconds each for a complete run. The start of the run was when the horse put one limb down for trot, and the end was when the trot was finished.

The videos were randomly numbered and saved as such so the observer would be blinded for each condition. Videos which included the horse walking, cantering or not being cooperative with the task of guided trotting were excluded from analysis. For some conditions, there were several videos per horse, while for other conditions there were no videos for some horses. There were 45 total hard surface videos to be coded, and 28 total soft surface videos to be coded. Videos were coded by a blinded to the clinical condition master's student in Biology and by a blinded equine specialist in veterinary anesthesia.

Videos were coded at 0.61x normal speed in VLC media player (Version 2.2.6) in random order, and seen three times each to gather an adequate score. The observer paid attention to the horse's face when seeing the video for the first time, and parameters were scored after the first run. The second time the video was seen, other parameters were observed and marked or edited after the run. The last time the video was run, the observer made sure all the parameters were scored adequately, both during and after seeing the video, and the scores from all parameters were tallied up for a final total score. All videos were coded two separate times (with 3 viewings of each video per coding bout) for the hard and soft surfaces by the master's student in Biology; videos were coded once by the equine specialist in veterinary anesthesia. The videos were randomized differently for each coding bout, but randomized the same for both observers. Hard surface videos were coded first, with a week between each coding bout to avoid bias, and soft surface videos were coded second, also taking a week in between each coding bout. Videos were coded using a modified EQUUS-FAP ethogram, which will be explained later (Table 7A).

Photographs were generated from each video (hard and soft surface) and also scored by the trained master's student in Biology and equine specialist in veterinary anesthesia. Photos were scored using a modified EQUUS-FAP ethogram exclusively for photos, which will be discussed later (Table 7C). To randomize photograph collection from the videos, one random photo was taken from each second frame per video. Thus, if a video was 10 seconds long, 10 photos would be extracted from this video, one photo from each second of the video. The 5 clearest photos would be collected from each video, trying to minimize bias as much as possible (photographs where the horse's face was visible, not too small, not blurry, and eyes/mouth/nose/ears were visible). Each of the 5 photos per video would later be scored. Sets of photos (per video) would be scored in random order, but within each set, the photos would be scored in the order collected from the video (from start of trot to end). Sets of photos would be scored twice by the master's student in Biology, at least one week apart, both for hard and soft surface; sets of photos were scored only once by the equine specialist in veterinary anesthesia. There were 225 photos to code for the hard surface, and 135 photos to code for the soft surface.

4. Equine Utrecht University Scale for Facial Assessment of Pain (EQUUS-FAP)

4.1 Original EQUUS-FAP scale

The original EQUUS-FAP scale, developed by Van Loon and van Dierendonck (2015) and validated further in Van Dierendonck and van Loon (2016, 2017, 2019), is shown in Table 3. As mentioned previously, the original EQUUS-FAP is a 9-parameter multifactorial SDS, describing different elements of facial expression. It is a dynamic pain scale, as it includes dynamic aspects of facial expression such as teeth grinding, moaning, etc. Each parameter is scored from 0 to 2, leading to a total pain score from 0 (no signs of pain) to a maximum of 18 (maximal pain score). This original EQUUS-FAP pain scale was used by observers during live trotting of horses in the first part of the study, but the original pain scale had to be modified for pain scoring from videos and photos, as certain parameters were particularly difficult or impossible to read, such as teeth grinding and/or moaning from photographs, and certain behaviors were observed which were likely not focused on during live trotting, such as licking and/or chewing.

4.2 Training observers to use scale

The master's student in Biology in the study was trained first using the EPWA online training app (Equine Pijn en Welzijns App Trainingssite), using the 6 parameters in the EPWA ethogram as shown in Table 6, out of a maximum score of 12. The "Training voor Paarden" feature was used with modules from 1-5, each module having 10 photos with parameters that had to be scored. When the master's student in Biology reached above 75% accuracy after the fifth module, training could continue using the EQUUS-FAP ethogram on practice videos of horses trotting.

Five random cut hard surface trot videos to be used in the experiment were used as training videos, where the master's student in Biology in training and the trained equine specialist in veterinary anesthesia both were blinded to the videos, coded the videos simultaneously, and compared inter-observer reliability (SPSS intra-class correlation coefficient, two-way, random-effects, mixed model; $p<0.05$). After coding each bout of videos, coding results between the student and veterinarian were discussed and compared, and differences in coding were speculated on and clarified. When inter-observer reliability was above 0.70 and differences were clarified, training was considered complete, and both observers could now score all the videos and photos for the study.

Table 6: EPWA training ethogram.

Data	Categories	Score
Ears	Both ears facing forward At least 1 ear directed to the side or rear Both ears facing back	0 1 2
Eyelids	Relaxed eyelids Eyelids tightened together Eyelids shut	0 1 2
Raised upper eyelid	Relaxed upper eyelid Upper eyelid partially raised Upper eyelid raised considerably	0 1 2
Eye white	Whites are not visible Whites are visible Whites are clearly visible	0 1 2
Corners of mouth	Corners of mouth relaxed Corners of mouth slightly tightened Corner of mouth and lower lip tightened	0 1 2
Nostrils	Nostrils relaxed and closed Nostrils far opened Nostrils maximally opened	0 1 2
Total		**---/12**

4.3 Development of a modified EQUUS-FAP scale

During training, it was decided that the EQUUS-FAP scale, which was originally used for box-rested horses, would have to be altered when coding facial expressions from videos, as the horses were now in locomotion. Compared to the original EQUUS-FAP scale (Van Loon and van Dierendonck 2015), the parameter 'focus' was removed, the parameter 'head' had different categories, and 'licking and/or chewing' was added. 'Focus' was removed, as this parameter makes more sense when the horse is box-rested or standing, observing its environment. When a horse is trotting, by default it will be focused on the environment as it needs to see where it is going. 'Head' originally had a score of 0 refer to normal head movement, 1 was less movement, and 2 was no movement. However, when trotting, a horse can display either greater than or less than normal head movement, so during trotting, the 'head' parameter now had a score of 0 indicating normal head movement, 1 as a moderate increase or decrease in head movement, and 2 as an obvious increase or decrease in head movement. 'Licking and/or chewing' was added, as this behavior was sometimes visible during trotting, and was not previously in the scale.

The modified EQUUS-FAP scale for videos is shown in Table 7A**.** This new scale had 9 parameters total, for a maximum pain score of 18. Three additional parameters, used in the EPWA training site scale, were also coded in the videos, though they were not included in the total pain score. These included the following: 'eye white,' 'raised upper eyelid' and 'orbital tightening.' A comparison of the scores of these parameters from videos to photos would be done later. These individual eyelid scores were not included in the total pain score for video coding due to the difficulty of coding specific features of the eye in a moving horse, and sometimes seeing certain features of the eye such as eye white clearly, while other features such as orbital tightening not clearly (depending on the quality of the video, the coat color of the horse, etc.), for example. Instead of focusing on specific eye features for the total pain score in video coding, eye coding was grouped as one parameter, 'eyelids,' as was done in the original EQUUS-FAP scale (Van Loon and van Dierendonck 2015) shown in Table 3. Each score for 'eyelids' would consist of a consideration of these three individual parameters, though each individual parameter for eyes would not be separately scored. For example, a score of 0 for the 'eyelids' parameter would include a score of 0 for each of the individual parameters 'eye white,' 'raised upper eyelid' and 'orbital tightening.' If two of these parameters had a score of 0, for example, and one of them, such as eye white, had a score of 1, this would constitute a score of 1 for the 'eyelid' parameter. Likewise, if two parameters had a score of 1, and one parameter with a score of 2, this would constitute a score of 2 for the 'eyelid' parameter.

It was also later decided that for coding the photos, an altered scale again would have to be used, as dynamic parameters such as 'head,' 'muscle tone head,' 'teeth grinding and/or moaning,' 'flehming and/or yawning,' and 'licking and/or chewing' would not make sense to code in still photographs of a horse in locomotion. The scale for pain scoring from photographs was similar to the EPWA training site ethogram, with the three main eye features being coded as the individual parameters 'orbital tightening,' 'raised upper eyelid' and 'eye white.' This was because individual eye features were easier to see in a still picture than in a video of a moving horse. The parameter 'open mouth' was also added to the photo coding scale, as that was distinctly seen in some photos. There was a total of 7 parameters, out of a maximum pain score of 14. The modified EQUUS-FAP scale for photos is shown in Table 7C.

Table 7: Modified EQUUS-FAP scales. (A) Modified scale for video coding. (B) Parameters coded from videos but not included in total pain score. (C) Modified scale for photo coding.

A.

Parameter	Categories	Score
Ears	Both ears facing forward At least 1 ear directed to the side or rear Both ears facing back	0 1 2
Eyelids	Relaxed eyelids, relaxed upper eyelid, whites are not visible (except in normal eye/head movement) Eyelids tightened together, upper eyelid partially raised (see wrinkles), whites seen around 50% of time Eyelids shut, upper eyelid raised considerably, whites seen >50% of time	0 1 2
Corners mouth/lips	Corners of mouth relaxed Corner of mouth slightly tightened Corner of mouth and lower lip tightened	0 1 2
Nostrils	Nostrils relaxed and closed Nostrils far open Nostrils maximally open, nostrils flaring and possibly audible breathing	0 1 2
Head	Normal head movement/interested in environment Moderate increase or decrease in movement Obvious increase or decrease in movement	0 1 2
Muscle tone head	No fasciculations Mild fasciculations Obvious fasciculations	0 1 2
Teeth grinding and/or moaning	Not heard Heard	0 2
Flehming and/or yawning	Not seen Seen	0 2
Licking and/or chewing	Not seen Seen	0 2
Total		**…/18**

B.

Eye white	Whites not seen (except normal eye/head movement) Whites seen around 50% of time Whites seen >50% of time	0 1 2
Raised upper eyelid	Relaxed upper eyelid Upper eyelid partially raised Upper eyelid raised considerably	0 1 2
Orbital tightening	Relaxed eyelids Eyelids tightened together Eyelids shut	0 1 2

C.

Parameter	Categories	Score
Ears	Both ears facing forward At least 1 ear directed to the side or rear Both ears facing back	0 1 2
Orbital tightening	Relaxed eyelids Eyelids tightened together Eyelids shut	0 1 2
Raised upper eyelid	Relaxed upper eyelid Upper eyelid partially raised Upper eyelid raised considerably	0 1 2
Eye white	Whites not seen (except normal eye/head movement) Whites seen around 50% of time Whites seen >50% of time	0 1 2
Corners mouth/lips	Corners of mouth relaxed Corner of mouth slightly tightened Corner of mouth and lower lip tightened	0 1 2
Nostrils	Nostrils relaxed and closed Nostrils far open Nostrils maximally open, nostrils flaring and possibly audible breathing	0 1 2
Open mouth	Mouth closed Mouth slightly opened Mouth maximally opened	0 1 2
Total		**…/14**

5. Data processing and statistical analysis

5.1 Reliability analysis

Reliability would indicate consistency of scoring, both within-observer and between-observer. Several intra-observer reliability tests were completed, as well as inter-observer reliability tests, for both total pain scores and for individual parameters, both on hard surfaces, soft surfaces, and both surfaces combined. Reliability tests were conducted for video scoring, photo scoring, and for median scores from each photo set (corresponding to a single video).

More specifically, intra-observer reliability was calculated for Scorer 1, or the trained master's student in Biology, who scored each video and photo set twice, one week apart. Intra-observer reliability was calculated using a two-way, random-effects, mixed model, average measures intra-class correlation coefficient (ICC), and was represented by Cronbach's alpha (with 95% confidence intervals, $p<0.05$). Inter-observer reliability was calculated for the first scoring attempt for videos and photos, between Scorer 1 (trained master's student in Biology) and Scorer 2 (trained equine specialist in veterinary anesthesia). The inter-observer ICC was also a two-way, random-effects, mixed model, average measures ICC represented by Cronbach's alpha. The following guidelines were used to interpret the ICC measures (Ducasse 2020):

- $\alpha \geq 0.9$ excellent
- $0.9 > \alpha \geq 0.8$ good
- $0.8 > \alpha \geq 0.7$ acceptable
- $0.7 > \alpha \geq 0.6$ questionable
- $0.6 > \alpha \geq 0.5$ poor
- $0.5 > \alpha$ unacceptable

Both intra- and inter-observer ICC were calculated first for total pain scores for videos on only hard surface, then only on soft surface, then both on hard and soft surfaces combined. The same was done for total pain scores for all photo scores (several photos per set, from a single video), and for median photo scores (median photo pain score from each set of photos). In order to assess the correlation between video and photo scores, ICC was calculated for video pain scores and photo median pain scores, as both data sets had the same sample size; this was done for each scorer, for hard surface only, soft surface only, and both surfaces combined.

Cronbach's alpha for intra- and inter-observer ICC was also later extracted for each individual parameter for videos, all photos, and photo medians on combined surfaces (hard and soft). Parameters that were used for both video and photo scoring were used to calculate ICC between video scores and photo median scores, to see how scoring of these parameters in videos and photos would compare.

The parameters that had the lowest Cronbach's alpha values as well as the least significance (highest p-values) were removed from the total pain scores. These included the parameters 'corners mouth/lips' and 'orbital tightening,' the second of which was only originally included in the total pain score for photo scoring. Upon removal of these parameters from the EQUUS-FAP scale, calculation of intra- and inter-observer ICC for videos, photo totals and photo medians was repeated for total pain scores. These Cronbach's alpha values were then compared to the original Cronbach's alpha values from the EQUUS-FAP scale that included all parameters.

5.2 Validity analysis

Testing for validity would measure the accuracy of the modified EQUUS-FAP scales in being able to differentiate between baseline and induced lameness conditions. To test validity, videos were unblinded based on horse and organized based on condition (baseline vs. induced lameness) and treatment type (shoe vs. LPS induction). Only videos for which there were Qualisys asymmetry scores were used. For some horses, videos of certain conditions were missing, which was considered when nonparametric testing was done.

To compare baseline and induced lameness conditions, for both the shoe and LPS induced lameness models, Wilcoxon signed rank tests, or nonparametric tests for two related groups (shoe baseline vs. shoe induced lameness, or LPS baseline vs. LPS induced lameness), were conducted. Depending on what data was available for each horse, each horse's baseline scores would be compared to its induced lameness scores, for both the shoe and LPS models. Analysis would be done with the compiled data from both hard and soft surfaces, and analysis was done separately for video scoring, all photo scoring (five photos per video), and medians of each set of photo scores. The test statistic would be the p-values retrieved from the Wilcoxon signed rank test. If the p-values were greater than or equal to 0.05, that would indicate there was not a significant difference between pain scores for baseline vs. induced lameness conditions. Therefore, the EQUUS-FAP scales modified for video and photo scoring would not be adequate tools to determine whether a horse is lame. Such analyses were first conducted on total pain scores.

A ROC curve analysis was also conducted on total pain scores, to ascertain via the "area under the curve" statistic, the most optimal cut-off values and to assess whether the specificity and sensitivity of the EQUUS-FAP pain scoring scale would be good enough to differentiate between baseline and induced lameness conditions (whether the pain scoring scale is clinically valid). Both baseline and induced lameness conditions would include both induced lameness models, and analysis would be done with compiled data on hard and soft surfaces, separately for video scoring, all photo scoring (five photos per video), and medians of each set of photo scores. The guidelines in this study for categorization of AUROC ("area under the ROC curve") curve values is shown in Table 8.

Table 8: Categorization of ROC curves values. AUROC indicates "area under the ROC curve."

AUROC	Category
0.9-1.0	Very good
0.8-0.9	Good
0.7-0.8	Fair
0.6-0.7	Poor
0.5-0.6	Fail

As for reliability analysis, the same parameters (those with the lowest Cronbach's alpha values as well as least significance), 'corners mouth/lips' and 'orbital tightening,' were later removed from total pain scores, to see if validity would improve upon exclusion of these parameters. The p-values, as well as AUROC values, from validity analysis of total pain scores were compared before and after exclusion of parameters.

Validity analysis was also conducted using the Wilcoxon signed rank test (p-values are statistic) for individual parameters on combined hard and soft surfaces, to find out whether any individual parameters were more accurate in differentiating between baseline and induced lameness conditions. If there would be more valid parameters, it would be necessary to look at whether pain scores for those parameters in the induced lameness condition were higher than at baseline.

5.3 Qualisys correlation analysis

The asymmetry index was obtained by objective locomotion analysis from Qualisys during live guided trot (during collection of videos). The asymmetry index scores (for each horse, each condition) were correlated to total pain scores from video coding on hard, soft, and combined surfaces, and for each scorer (trained master's student in Biology and trained equine specialist in veterinary anesthesia). A Pearson's two-tailed correlation coefficient was calculated for each scorer and surface to find if there was a correlation between the asymmetry index and total pain scores. Statistical analyses were completed in SPSS (Version 26, IBM), and statistical significance was accepted throughout the study as $p<0.05$.

V. Results

1. Reliability results

Both intra- and inter-observer agreement led to ICC's with Cronbach's alpha higher than or close to 0.70 with significant correlations ($p<0.05$) (Table 9 A, B, C). When considering inclusion of all parameters in the modified EQUUS-FAP scales for video and photo coding, all intra-observer correlations were significant ($p<0.05$) and greater than 0.70. Intra-observer correlations were higher than inter-observer correlations, for videos and photos on all surfaces. Photo scoring (considering photo totals and median photo scores) for combined surfaces had higher intra-observer reliability (photo totals Cronbach's α=0.89; photo medians Cronbach's α=0.94) than video scoring (video Cronbach's α=0.83), as well as higher inter-observer reliability (photo totals Cronbach's α=0.73; photo medians Cronbach's α=0.79) than video scoring (video Cronbach's α=0.70). Both intra- and inter-observer reliability when taking the median photo score from all 5 photos per video (median photo score) was higher than total photo score. Soft surface video coding inter-observer reliability had the lowest ICC, α=0.68, and was the only correlation considering all parameters that was below 0.70, or the threshold value below which an ICC value is not acceptable.

When inter-observer reliability was graphed (considering all parameters, for total pain scores), video and photo pain scores for Scorer 1 (trained master's student in Biology) were skewed a bit higher than for Scorer 2 (trained equine specialist in veterinary anesthesia) (Figure 4). The pain scores of photos on average were more homogenous for both scorers than for videos, shown by the data trend line being closer to a correlation of 1 (the line y=x) for photos than for videos. The data trend line for photo median scores (when considering all parameters, total pain scores) is closest to the line correlation of 1 (the line y=x), having a Cronbach's alpha of 0.79; then all photo totals scores has the next highest Cronbach's alpha at 0.73, and video total scores has the lowest Cronbach's alpha, with the greatest difference in coding between observers, at 0.70.

When individual parameters were considered on all surfaces, as for total pain scores, intra-observer reliability had higher ICC's than inter-observer reliability for most parameters (Table 10 A, B, C). 'Eye white' for video scoring was the only parameter where inter-observer reliability was higher than intra-observer reliability. The parameters with the highest ICC's for video and photo coding, considering intra- and inter-observer reliability, were the following: 'ears,' 'nostrils,' 'eye white,' and 'raised upper eyelid.' With only videos considered, the parameters 'licking and/or chewing' and 'head' had high intra- and inter-observer reliabilities; with only photos considered, 'open mouth' had high correlations.

Individual parameters with the lowest ICC values and least significance for video and photo scoring were 'corners mouth/lips' and 'orbital tightening' (Table 10 A, B, C). The parameter 'corners mouth/lips' was not significant for either video or photo scoring for inter-observer reliability (p>0.05), and for photo scoring (photo totals and medians), the ICC for intra-observer reliability was the lowest of all parameters (below 0.70 for both). 'Orbital tightening' for video scoring had the lowest intra-observer reliability (Cronbach's α=0.50) and an insignificant inter-observer reliability (Cronbach's α=0.23); for photo scoring intra-observer reliability was above α=0.70 (Cronbach's α=0.85 for photo totals; 0.86 for photo medians), but below α=0.70 for inter-observer reliability (Cronbach's α=0.51 for photo totals; 0.55 for photo medians). When white and dark horses were considered separately for the 'corners mouth/lips' and 'orbital tightening' parameters, considering all surfaces, only horses considering all photo scores for the 'orbital tightening' parameter had significant reliabilities (p<0.05). Intra-observer reliability was significant and higher for photo total scores for white than dark horses (Cronbach's α=0.862, p=0 for white; α=0.833, p=0 for dark), while inter-observer reliability was significant and higher for photo total scores for dark than white horses (α=0.443, p=0 for dark; α=0.435, p=0.024 for white).

Both 'corners mouth/lips' and 'orbital tightening' parameters were removed before analyzing total pain score ICC's again (Table 9). 'Orbital tightening' was only removed from photo scores, as it was not a parameter included in the total pain score for video coding. Upon removal of these least significant parameters, the intra- and inter- observer correlations increased, except for soft surface inter-observer ICC for video coding, and intra-observer ICC for photo medians on hard, soft and combined surfaces. All ICC's with the removed parameters were significant (p<0.05), though the soft surface video coding inter-observer reliability (Cronbach's α=0.65) was still the only ICC below α=0.70.

When total pain scores from video were compared to median photo scores per video, besides soft surface ICC's for Scorer 1 (Cronbach's α=0.77), the correlations were below α=0.70 (Table 9D). Upon removal of the parameters 'corners mouth/lips' and 'orbital tightening,' the ICC's between video and photo scores did not have a uniform trend, increasing for some modalities (i.e. hard and combined surface Scorer 1), and decreasing for others (soft surface Scorer 1 and hard, soft and combined surfaces for Scorer 2). Upon comparing video pain scores to photo median scores for individual parameters (Table 10D), only 'ears' had a good correlation ($\alpha \geq 0.80$) with significance (p<0.05) for both scorers. The parameter 'nostrils' only had an acceptable correlation ($\alpha \geq 0.70$) for Scorer 1.

Table 9. ICC for total pain scores, with all parameters and with removed parameters. Intra- and inter-observer reliability using intra-class correlation coefficient (two-way random-effects model, measured for single rating), shown by Cronbach's alpha, for total pain scores. Intra-observer reliability for Scorer 1 (Scorer 1=trained master's student in Biology, first two coding attempts after training), and inter-observer reliability between Scorer 1 and 2 (Scorer 2=trained equine specialist in veterinary anesthesia, first coding attempt). All values significant ($p<0.05$), Cronbach's alpha reported for soft, hard and combined surfaces. Cronbach's alpha ≥0.70 indicates acceptable reliability (highlighted in green). (A) Cronbach's alpha values from video scoring, (B) Cronbach's alpha values from all photos taken from videos, (C) Cronbach's alpha values from the median pain score for each photo set (from each video), (D) Comparing Cronbach's alpha for ICC of videos to medians of all photos from each video, for both Scorer 1 and Scorer 2.

A. Video

Surface	Intra-observer reliability Cronbach's alpha		Inter-observer reliability Cronbach's alpha	
	all parameters	removed parameters	all parameters	removed parameters
hard (n=45)	0.78	0.84	0.71	0.75
soft (n=27)	0.90	0.92	0.68	0.65
combined (n=72)	0.83	0.87	0.70	0.71

B. Photo totals

Surface	Intra-observer reliability Cronbach's alpha		Inter-observer reliability Cronbach's alpha	
	all parameters	removed parameters	all parameters	removed parameters
hard (n=225)	0.89	0.94	0.72	0.83
soft (n=135)	0.89	0.91	0.76	0.84
combined (n=360)	0.89	0.93	0.73	0.83

C. Photo medians

Surface	Intra-observer reliability Cronbach's alpha		Inter-observer reliability Cronbach's alpha	
	all parameters	removed parameters	all parameters	removed parameters
hard (n=45)	0.94	0.92	0.81	0.84
soft (n=27)	0.93	0.93	0.77	0.87
combined (n=72)	0.94	0.92	0.79	0.85

D. Video to Photo correlation

Surface	Scorer 1 Cronbach's alpha		Scorer 2 Cronbach's alpha	
	all parameters	removed parameters	all parameters	removed parameters
hard (n=45)	0.43	0.53	0.58	0.54
soft (n=27)	0.77	0.76	0.63	0.48
combined (n=72)	0.59	0.63	0.57	0.51

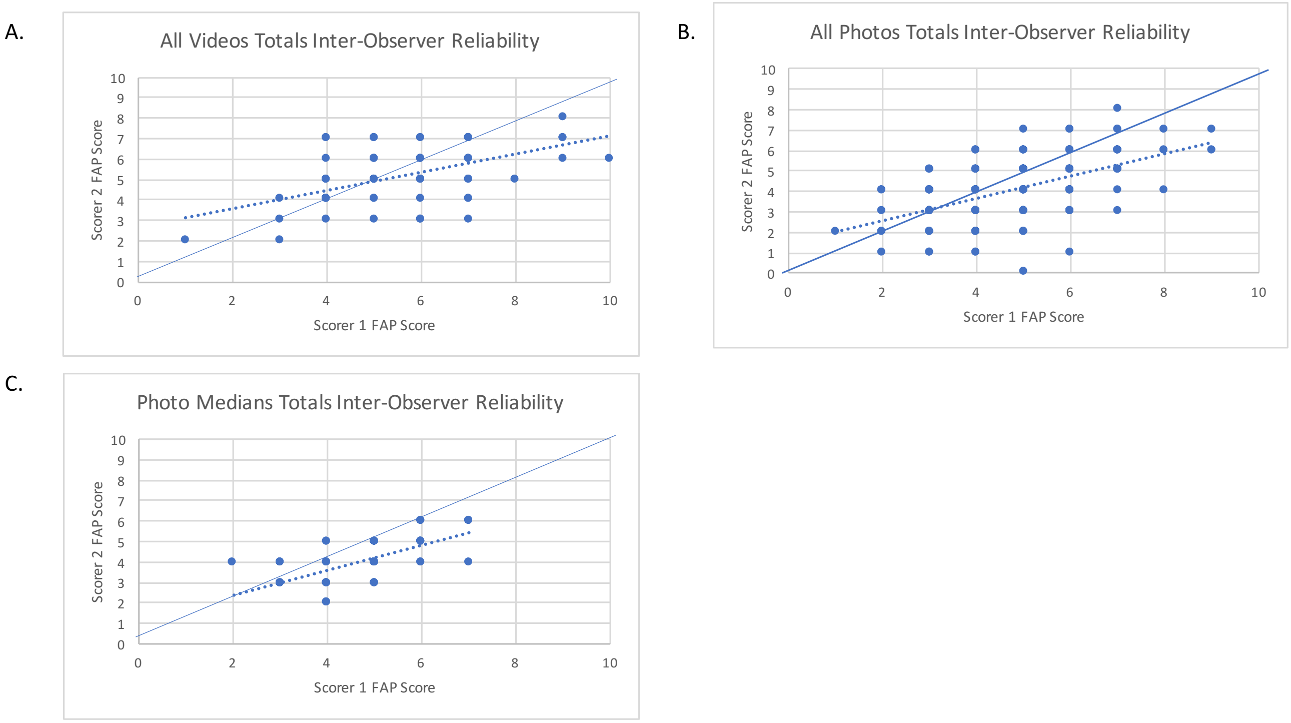


Figure 4. Graphing inter-observer agreement. Scatter plots of EQUUS-FAP (with all parameters) inter-observer agreement for video, all photo and photo medians pain score totals, considering combined surfaces (hard and soft). Linear data trend lines are included, as well as lines representing perfect correlation (y=x). Inter-observer reliability is between Scorer 1 and 2 (Scorer 1=trained master's student in Biology, first of two coding attempts after training; Scorer 2= trained equine specialist in veterinary anesthesia, first coding attempt). (A) Inter-observer ICC for video scoring (Cronbach's α=0.70, n=72), (B) Inter-observer ICC for all photo scoring (Cronbach's α=0.73, n=360), (C) Inter-observer ICC for photo medians scores (Cronbach's α=0.79, n=72).

Table 10. **ICC for individual parameters**. Intra- and inter-observer reliability using intra-class correlation coefficient (two-way random-effects model, measured for single rating), shown by Cronbach's alpha, for individual parameters for combined surfaces (hard and soft). Intra-observer reliability for Scorer 1 (Scorer 1=trained master's student in Biology, first two coding attempts after training), and inter-observer reliability between Scorer 1 and 2 (Scorer 1 first coding attempt; Scorer 2= trained equine specialist in veterinary anesthesia, first coding attempt). Cronbach's alpha ≥ 0.70 indicates acceptable reliability (highlighted in green); values highlighted in yellow are not significant (p>0.05). Cronbach's alpha values for individual parameters scored in (A) videos, (B) all 5 photos from each video, (C) photo medians, or median pain score for photos from each video, and (D) comparing videos to photo medians for Scorer 1 and 2, using only combined surfaces (hard and soft).

A. Video (n=72)

Parameter	Intra-observer reliability Cronbach's alpha	Inter-observer reliability Cronbach's alpha
Ears	0.89	0.79
Corners mouth/ lips	0.74	0.15
Nostrils	0.75	0.63
Eye white	0.71	0.73
Raised upper eyelid	0.72	0.62
Orbital tightening	0.50	0.23
Eyelids	0.69	0.44
Head	0.84	0.81
Muscle tone head	NA	NA
Teeth grinding and/or moaning	1.00	0.00
Flehming and/or yawning	NA	NA
Licking and/or chewing	1.00	0.73

B. Photo totals (n=360)

Parameter	Intra-observer reliability Cronbach's alpha	Inter-observer reliability Cronbach's alpha
Ears	0.97	0.96
Corners mouth/ lips	0.60	0.13
Nostrils	0.90	0.65
Eye white	0.92	0.44
Raised upper eyelid	0.79	0.62
Orbital tightening	0.85	0.51
Open mouth	0.90	0.84

C. Photo medians (n=72)

Parameter	Intra-observer reliability Cronbach's alpha	Inter-observer reliability Cronbach's alpha
Ears	0.98	0.95
Corners mouth/lips	0.58	0.21
Nostrils	0.94	0.68
Eye white	0.86	0.09
Raised upper eyelid	0.73	0.63
Orbital tightening	0.86	0.55
Open mouth	1.00	0.80

D. Video to photo (n=72)

Parameter	Scorer 1	Scorer 2
Ears	0.80	0.83
Corners mouth/lips	0.40	-0.09
Nostrils	0.73	0.54
Eye white	0.65	0.25
Raised upper eyelid	0.54	0.58
Orbital tightening	0.49	0.50

2. Validity results

Considering all parameters in the modified EQUUS-FAP scales, both induced lameness conditions, all surfaces, and both scorers, the differences between baseline and induced lameness conditions were not significant (p>0.05 for all), and there was no uniform pattern between p-values for videos or photos (Table 11). The area under the curve (AUC) for ROC curves values were all below 0.6 (Table 12), within the category of failure for ROC curves (Table 8). Specifically, for Scorer 1 (trained master's student in Biology), video scoring on both surfaces had a significance of p=0.17, photo scoring considering all photo scores had a significance of p=0.90, and photo scoring considering median photo scores was p=0.89; thus, video scoring for Scorer 1 had the lowest p-value for combined surfaces, followed by photos medians then photo totals. For Scorer 2 (trained equine specialist in veterinary anesthesia), video scoring on both surfaces had a significance of p=0.46, photo scoring considering all photo scores had a significance of p=0.21, and photo scoring considering median photo scores was p=0.38. Photo totals scoring for Scorer 2 had the lowest significance, followed by photo medians and then video scoring. Hard surface video scoring results for Scorer 1 were the closest to significance with a p-value of 0.06, though still insignificant. Scorer 2's video scoring had the highest AUCROC values (0.553 for EQUUS-FAP with all parameters, 0.544 with removed parameters), though they were still below 0.6, or within the category of failed ROC curve values (Table 12; Figure 7).

When this data is visually represented using boxplots (Figure 5A, 6A, 6C), for coding pain scores from videos or photos (or when considering median photo score values), there do not appear to be significant differences in pain scores between baseline and induced lameness conditions for either scorer, though there is some increase in pain scores for induced lameness compared to baseline. For video scoring, for Scorer 1, median pain scores are lower for induced lameness than baseline (6 for baseline, 5 for induced lameness), though maximum pain scores are higher for the induced lameness condition (10) compared to baseline (9). For Scorer 2, the median pain scores from video coding are the same for both conditions (5), and the IQR has a wider range with a higher maximum for induced lameness (4 to 7) than baseline (5 to 6). For photo coding, considering total pain scores for all photos, for both Scorer 1 and 2, the medians between conditions are the same (Scorer 1, median 5; Scorer 2, median 4). The maximum pain scores for Scorer 1 for the induced lameness conditions (9) are the same as for baseline, while for Scorer 2 the maximum pain scores for induced lameness are higher (8) than for baseline (7). The IQR for Scorer 1 induced lameness, however, has a wider range with a higher maximum (from 4 to 6) than for baseline (4 to 5). For photo coding considering only photo median scores from each video, maximum pain scores and medians for both scorers are the same for both conditions (Scorer 1 median 5, Scorer 2 median 4; Scorer 1 maximum 7; Scorer 2 maximum 6). Scorer 1's IQR for induced lameness has a wider range with a higher maximum (4 to 5.5) compared to baseline (4 to 5), however.

When the least significant and reliable parameters were removed from the total pain scores ('corners mouth/lips' and 'orbital tightening'), there were mixed results for video and photo scoring (Table 11). Some p-values for the difference between induced lameness and baseline decreased, others increased, while others were the same compared to p-values when all parameters were included. Similarly, for AUROC values, some increased and others decreased upon removing parameters form the EQUUS-FAP scale (Table 12), though the category for the AUROC values was still "failure" (AUROC below 0.6, Table 8). All p-values remained insignificant (p>0.05), however. Hard surface scoring for Scorer 1 remained the most significant at p=0.05, though still insignificant. Visually, boxplots show that with the removed parameters, the differences between baseline and induced lameness remain insignificant (Figures 5 and 6). For video and photo coding, the medians and maximum values for both scorers and both conditions decreased after removing parameters (Figures 5 and 6). For video coding however, for both scorers, the difference between baseline and induced lameness medians remains the same (for Scorer 1 difference in medians remains 1, for Scorer 2 difference remains 0). For photo coding, considering both total pain scores for all photos and photo medians, the difference between median pain scores between conditions for each scorer remain 0.

When individual parameters were considered for validity, certain parameters for certain scorers, such as 'head' for video coding (scorer 2, p=0.01) and 'nostrils' (scorer 2, p=0, photo totals; scorer 2, p=0, photo medians) and 'raised upper eyelid' (scorer 1, p=0.01, photo totals) for photo coding, did have p-values <0.05 for one scorer, so these parameters showed significant differences in pain scores for baseline and induced lameness conditions (Table 13). These parameters also had higher pain scores for induced lameness compared to baseline (Figure 8). Frequency charts in Figure 8 show that for each of these individual parameters in their respective coding conditions, the frequency of the lowest score (0) decreased and the frequency of either score 1, 2 or both increased upon induction of lameness, compared to baseline.

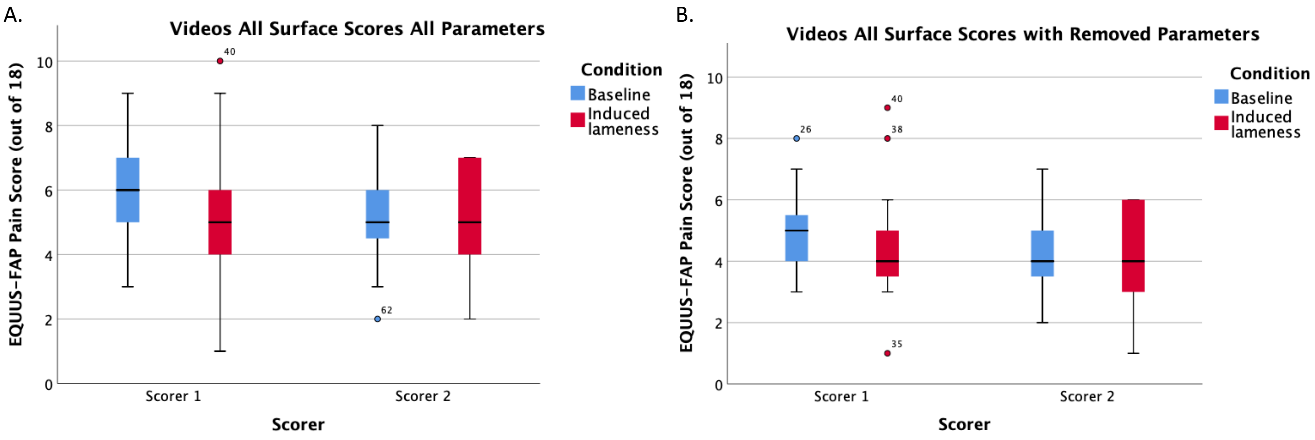


Figure 5. **All surface video pain scores, with all parameters and with removed parameters.** All surface (hard and soft) video pain scores using modified EQUUS-FAP scale for moving horses in videos. Compares baseline to induced lameness condition for each scorer (Scorer 1= trained master's student in Biology, first coding attempt after training; Scorer 2= trained equine specialist in veterinary anesthesia, first coding attempt). (A) FAP scores with all parameters included (Scorer 1: n=27, Wilcoxon signed rank test p=0.17; Scorer 2: n=27, Wilcoxon signed rank test p=0.46), (B) FAP scores with some parameters removed ('corners mouth/lips,' 'orbital tightening') (Scorer 1: n=27, Wilcoxon signed rank test p=0.26; Scorer 2: n=27, Wilcoxon signed rank test p=0.67). The numbers refer to scores that were outliers; each box represents the IQR; the bottom whisker is Q1 and the top whisker is Q4; the bold line in the box represents the median score.

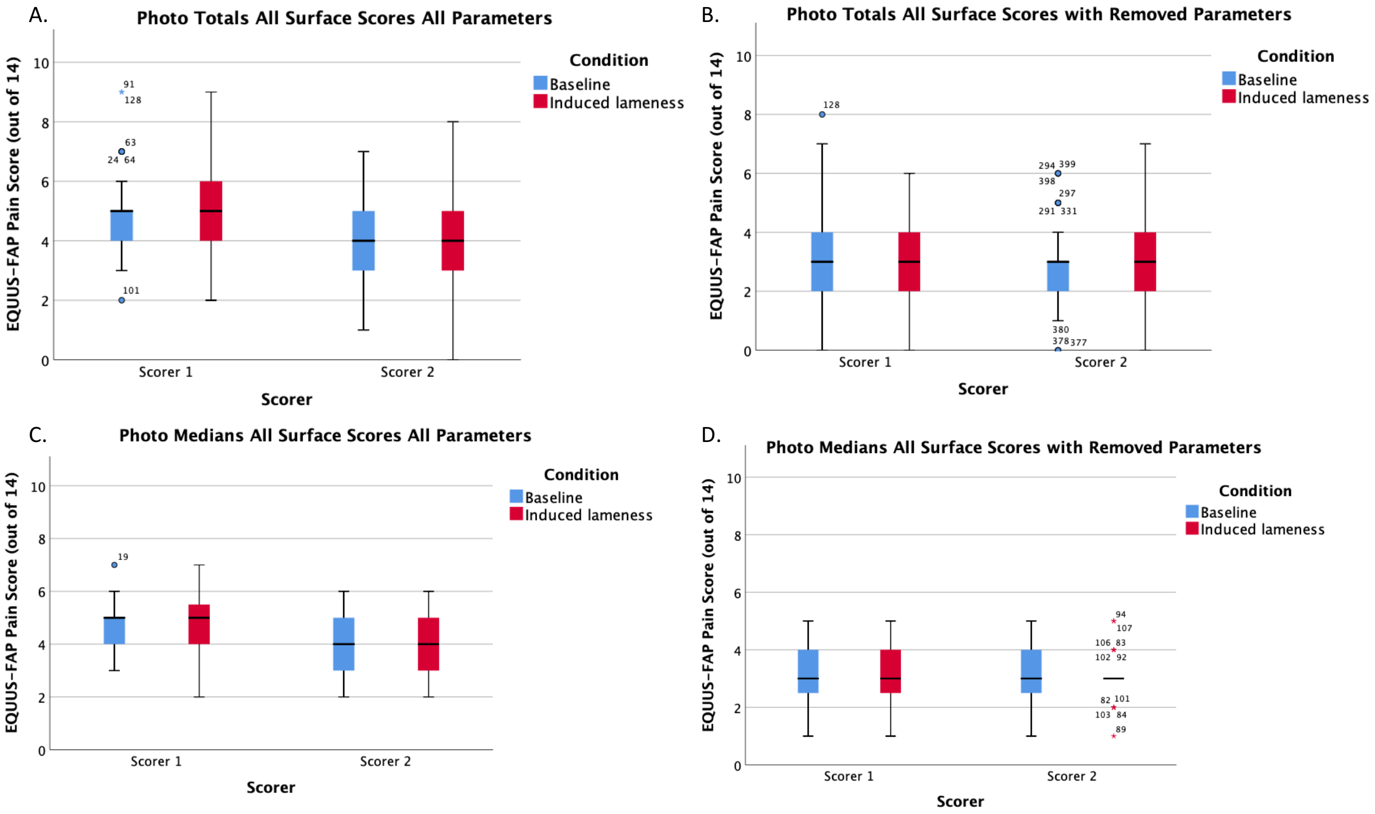


Figure 6. **All surface photo pain scores, with all parameters and with removed parameters.** All surface (hard and soft) photo pain scores using modified EQUUS-FAP scale for coding photos taken from videos of moving horses. Includes pain scores from all photos (5 photos taken from each video) and a median photo score from each video. Compares baseline to induced lameness condition for each scorer (Scorer 1= trained master's student in Biology, first coding attempt after training; Scorer 2= trained equine specialist in veterinary anesthesia, first coding attempt). (A) FAP score for all photos with all parameters (Scorer 1: n=135, Wilcoxon signed rank test p=0.90; Scorer 2: n=135, Wilcoxon signed rank test p=0.21), (B) FAP score for all photos with some parameters removed ('corners mouth/lips,' 'orbital tightening') (Scorer 1: n=135, Wilcoxon signed rank test p=0.60; Scorer 2: n=135, Wilcoxon signed rank test p=0.21), (C) FAP photo median scores for each video with all parameters (Scorer 1: n=27, Wilcoxon signed rank test p=0.89; Scorer 2: n=27, Wilcoxon signed rank test p=0.38), (D) FAP photo median scores for each video with some parameters removed ('corners mouth/lips,' 'orbital tightening') (Scorer 1: n=27, Wilcoxon signed rank test p=0.77; Scorer 2: n=27, Wilcoxon signed rank test p=0.14). The numbers refer to scores that were outliers; each box represents the IQR; the bottom whisker is Q1 and the top whisker is Q4; the bold line in the box represents the median score.

Table 11. **Validity for total pain scores from Wilcoxon signed rank test, with all parameters and with removed parameters.** Accuracy of EQUUS-FAP pain scoring and EQUUS-FAP with removed parameters (‘corners mouth/lips,’ ‘orbital tightening’) scoring in differentiating between baseline and induced lameness conditions for total pain scores on hard, soft, and combined surfaces. Accuracy found as significance (shown by p-values) of Wilcoxon signed rank test; accuracy calculated for Scorer 1 and 2 (Scorer 1= trained master’s student in Biology, first coding attempt after training; Scorer 2= trained equine specialist in veterinary anesthesia, first coding attempt). No values were significant (p<0.05) at differentiating between baseline and induced lameness conditions. (A) Indicates video scoring results, (B) all photos scoring results, and (C) photo median scores (from each video) results.

A. Video

Surface	**Scorer 1**		**Scorer 2**	
	all parameters	removed parameters	all parameters	removed parameters
hard (n=14)	0.06	0.05	0.58	0.76
soft (n=13)	0.96	0.68	0.57	0.79
hard and soft (n=27)	0.17	0.26	0.46	0.67

B. Photo totals

Surface	**Scorer 1**		**Scorer 2**	
	all parameters	removed parameters	all parameters	removed parameters
hard (n=70)	0.36	0.74	0.24	0.08
soft (n=65)	0.22	0.68	0.57	1.00
hard and soft (n=135)	0.90	0.60	0.21	0.21

C. Photo medians

Surface	**Scorer 1**		**Scorer 2**	
	all parameters	removed parameters	all parameters	removed parameters
hard (n=14)	0.60	0.86	0.56	0.26
soft (n=13)	0.36	0.79	0.51	0.33
hard and soft (n=27)	0.89	0.77	0.38	0.14

Table 12. Validity for total pain scores from AUC ROC curves, with all parameters and with removed parameters. Testing clinical validity of EQUUS-FAP pain scale with and without removed parameters (‘corners mouth/lips,’ ‘orbital tightening’) in differentiating between baseline and induced lameness conditions for total pain scores, considering combined surfaces (hard and soft). Validity found as area under the ROC curve, done separately for each scorer (Scorer 1= trained master’s student in Biology, first coding attempt after training; Scorer 2= trained equine specialist in veterinary anesthesia, first coding attempt). ROC curves were all in fail category (<0.6) for distinguishing between baseline and induced lameness.

Surface	**Scorer**	**FAP scale**	**Area under ROC curve**
video (n=27)	1	all parameters	0.417
		removed parameters	0.413
	2	all parameters	0.553
		removed parameters	0.544
photo totals (n=135)	1	all parameters	0.505
		removed parameters	0.527
	2	all parameters	0.537
		removed parameters	0.550
photo medians (n=27)	1	all parameters	0.511
		removed parameters	0.522
	2	all parameters	0.552
		removed parameters	0.453

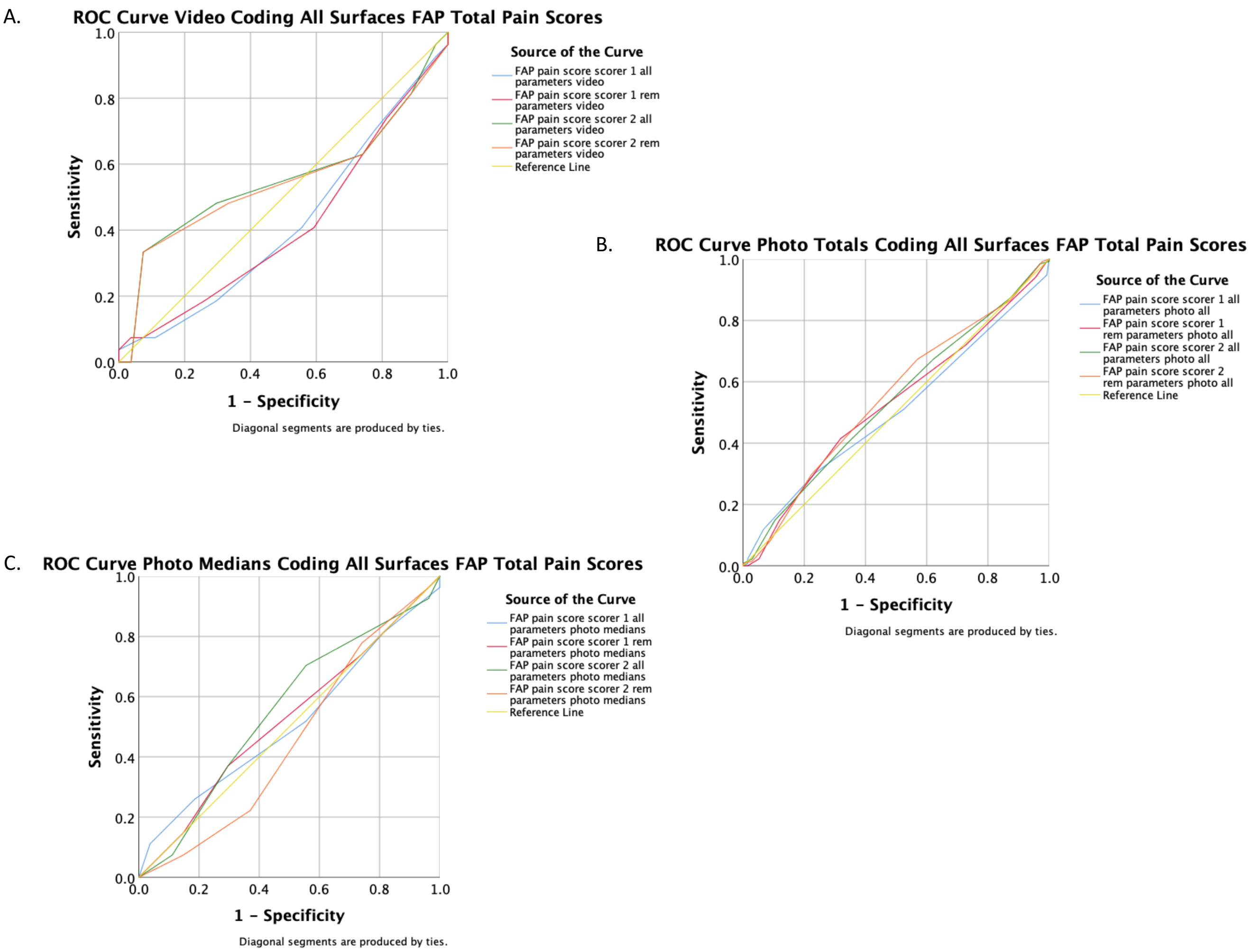


Figure 7. ROC curves showing sensitivity and specificity of pain scale. ROC curves created to find diagnostic accuracy of EQUUS-FAP pain scale for differentiating between induced lameness and baseline conditions, based on total pain scores from video scoring, all photos (5 photos taken from each video) scoring, and photo medians scores. ROC curves made for both scorers (Scorer 1= trained master's student in Biology, first coding attempt after training; Scorer 2= trained equine specialist in veterinary anesthesia, first coding attempt), for both surfaces (hard and soft) and using the EQUUS-FAP scale with all parameters and with removed parameters ('corners mouth/lips,' 'orbital tightening'). Yellow reference line is when validity of scale is at chance (0.5). False positives represented by '1-Specificity,' and true positives represented by 'Sensitivity.' (A) ROC curves from video scoring, for both scorer 1 and 2, considering all and removed parameters (n=27), (B) ROC curves for all photos scoring, for both scorer 1 and 2, considering all and removed parameters (n=135), and (C) ROC curves for photo medians scores, for both scorer 1 and 2, considering all and removed parameters (n=27).
rem parameters = removed parameters

Table 13. **Validity for individual parameters.** Accuracy of EQUUS-FAP pain scoring in differentiating between baseline and induced lameness conditions for each individual parameter on combined (hard and soft) surfaces. Accuracy found as significance (shown by p-values) of Wilcoxon signed rank test; accuracy calculated for Scorer 1 and 2 (Scorer 1= trained master's student in Biology, first coding attempt after training; Scorer 2= trained equine specialist in veterinary anesthesia, first coding attempt). Values highlighted in yellow indicate significance ($p<0.05$) of EQUUS-FAP scoring used in that modality at differentiating between baseline and induced lameness conditions. Pain scores of those parameters with a significant difference in the induced lameness condition were higher than for the baseline condition. (A) Indicates video results, (B) all photos results, and (C) photo medians (from each video) results.

A. Videos (n=27)

Parameter	**Scorer 1**	**Scorer 2**
Ears	0.17	0.56
Corners mouth/lips	0.53	0.32
Nostrils	0.16	0.37
Eye white	1.00	1.00
Raised upper eyelid	1.00	0.18
Orbital tightening	0.21	0.74
Eyelids	1.00	0.53
Head	0.64	0.01
Muscle tone head	1.00	1.00
Teeth grinding moaning	1.00	1.00
Flehming yawning	1.00	1.00
Licking chewing	0.32	0.56

B. Photo totals (n=135)

Parameter	**Scorer 1**	**Scorer 2**
Ears	0.10	0.21
Corners mouth/ lips	0.23	0.10
Nostrils	0.08	0.00
Eye white	0.37	0.21
Raised upper eyelid	0.01	0.52
Orbital tightening	0.55	0.21
Open mouth	0.59	0.41

C. Photo medians (n=27)

Parameter	**Scorer 1**	**Scorer 2**
Ears	0.80	0.80
Corners mouth/ lips	0.66	1.00
Nostrils	0.18	0.00
Eye white	0.37	0.56
Raised upper eyelid	0.16	0.74
Orbital tightening	1.00	0.66
Open mouth	0.32	0.66

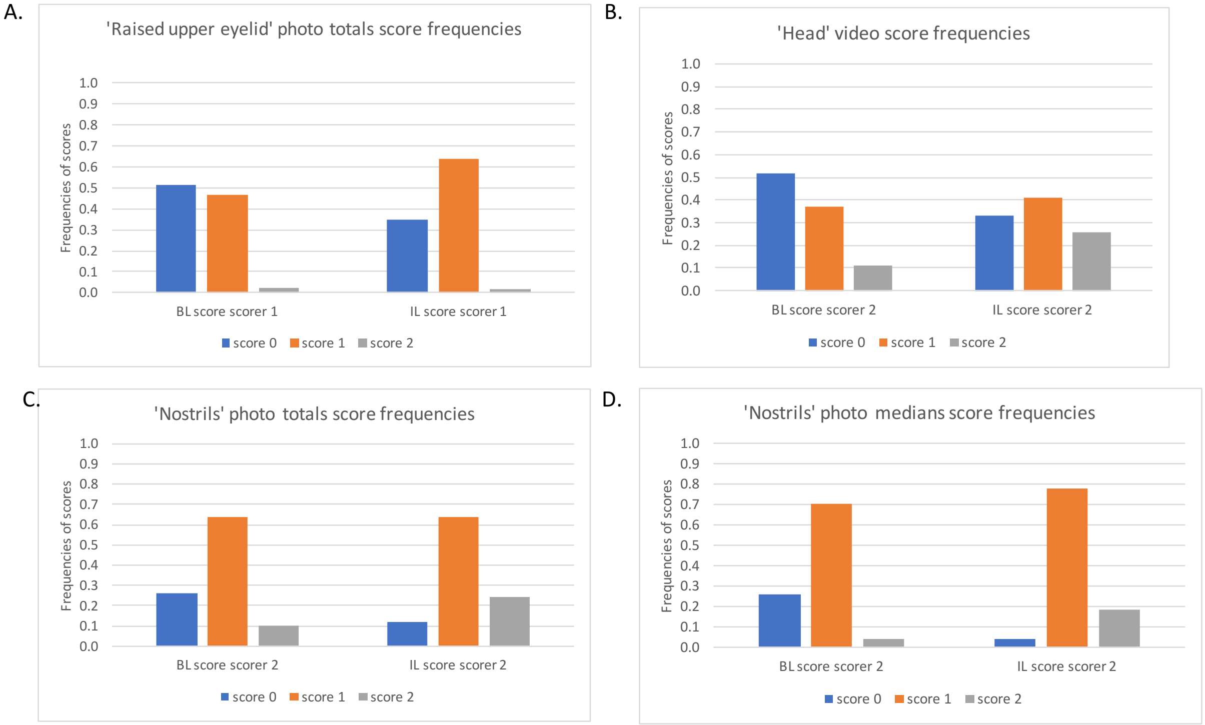


Figure 8. Validity for individual parameters that had significant differences between baseline and induced lameness. Bar graphs show differences in frequencies of particular pain scores in baseline and induced lameness conditions, scored by each observer (Scorer 1= trained master's student in Biology, first coding attempt after training; Scorer 2= trained equine specialist in veterinary anesthesia, first coding attempt), for parameters that had significant differences between baseline and induced lameness ('raised upper eyelid,' 'head,' and 'nostrils'). (A) Frequencies of 'raised upper eyelid' pain scores for photo totals scoring, both in baseline and induced lameness conditions, for scorer 1 (Wilcoxon signed rank test, p=0.01, n=135), (B) Frequencies of 'head' pain scores for video scoring, both in baseline and induced lameness conditions, for scorer 2 (Wilcoxon signed rank test, p=0.01, n=27), (C) Frequencies of 'nostrils' pain scores for photo totals scoring, both in baseline and induced lameness conditions, for scorer 2 (Wilcoxon signed rank test, p=0, n=135), and (D) Frequencies of 'nostrils' pain scores for photo medians scores (per set of 5 photos per video), both in baseline and induced lameness conditions, for scorer 2 (Wilcoxon signed rank test, p=0, n=27).

BL = baseline; IL = induced lameness

3. Qualisys correlation results

Qualisys asymmetry index scores were calculated for front and hind limb before and after induction of lameness. As seen in Figure 9, there was a significant difference in Qualisys asymmetry scores between baseline and induced lameness conditions, both for front and hind limb lameness, indicating lameness models (LPS and shoe) induced an asymmetry that was not there before lameness induction.

When a simple correlation analysis was done for the Qualisys asymmetry index scores and total pain scores from video coding for both scorers on hard, soft, and combined surfaces, the result was no significant correlation (p>0.05) for any modality (Table 14). Scorer 2 (trained equine specialist in veterinary anesthesia) had slightly higher correlations than Scorer 1 (trained master's student in Biology) between asymmetry index scores and total pain scores (for combined surfaces, r=0.227, p=0.092 for Scorer 2; r=0.046, p=0.738 for Scorer 1), though still no significant correlation (p>0.05).

When the correlations are visually assessed (Figure 10), Scorer 2 has a higher correlation between asymmetry score and pain score than Scorer 1, shown by a linear regression line with a slightly greater positive slope (Scorer 2, r=0.038; Scorer 1, r=0.0089). There is also greater variation in EQUUS-FAP pain scores when the Qualisys asymmetry score is low (below 10).

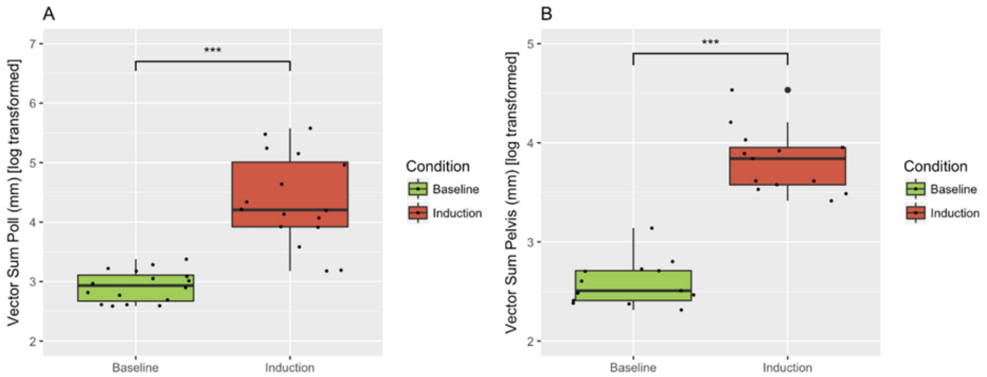


Figure 9. Asymmetry and Equine Utrecht University Scale for Facial Assessment of Pain (EQUUS-FAP) pain scores during trot with both lameness induction models (from Van Loon et al., unpublished results). (A) Vector sum poll (indicating asymmetry index for poll) in front limb lameness before and after lameness induction, (B) Vector sum pelvis (indicating asymmetry index for pelvis) in hind limb lameness before and after lameness induction. (n=31 for baseline conditions and n=54 for induced lameness). Box plots illustrate the median pain scores (bold line in box), the quartiles (box) and ranges (whiskers). The filled circles show individual pain scores. *** = p<0.001.

Table 14. Qualisys correlation analysis correlation coefficients. Correlations between Qualisys asymmetry index scores and total pain scores from video coding (using modified EQUUS-FAP for video coding) for both scorers (Scorer 1= trained master's student in Biology, first coding attempt after training; Scorer 2= trained equine specialist in veterinary anesthesia, first coding attempt) on hard, soft, and combined (hard and soft) surfaces. Correlations found by Pearson's correlation coefficient (two-tailed); no values were significant (p<0.05).

Surface	Scorer	Pearson's r coefficient	p-value	n value
Hard	1	0.114	0.554	29
	2	0.218	0.257	29
Soft	1	-0.012	0.954	27
	2	0.283	0.153	27
Hard and soft	1	0.046	0.738	56
	2	0.227	0.092	56

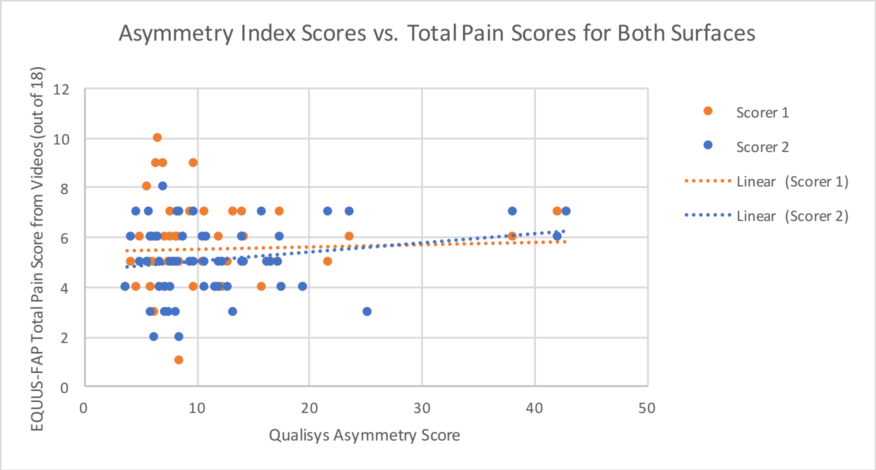


Figure 10. Qualisys asymmetry index scores vs. total pain scores from video coding for both surfaces. Scatterplot shows correlation between Qualisys asymmetry scores and total pain scores from video coding using EQUUS-FAP (modified for video coding, score out of 18), for both scorers (Scorer 1= trained master's student in Biology, first coding attempt after training; Scorer 2= trained equine specialist in veterinary anesthesia, first coding attempt) on combined (hard and soft) surfaces. Linear trend lines are included for each scorer; slope of line for scorer 2 (r=0.038) greater than for scorer 1 (r=0.0089).

VI. Discussion

In this study, it was found that the EQUUS-FAP pain scale (modified for video and photo coding) cannot reliably be used to differentiate between induced lameness and control horses using video and photo coding. Even though good reliability of pain scores was found, pain scores showed insignificant validity, even when removing the least reliable parameters from the pain scale and using that pain scale with removed parameters to code for facial pain expressions. Since the pain scale did not have significant validity in differentiating between baseline and induced lameness horses, there indeed was no significant correlation between Qualisys asymmetry index scores and EQUUS-FAP pain scores.

EQUUS-FAP has only been validated for assessment of box-rested horses in other pain states, including acute colic pain, acute and post-operative head-related pain, pain after orthopedic surgery and acute orthopedic trauma (Van Loon and van Dierendonck 2015; Van Dierendonck and van Loon 2016; Van Loon and van Dierendonck 2017). It has never been validated for video or photo coding, though both video and photo coding reliability/validity using other scales was tested before on box-rested horses with different diseases/conditions than lameness, with successful results (Dalla Costa et al. 2014; Gleerup et al. 2015). EQUUS-FAP has also been used before on trotted horses with live observations (J.P.A.M. van Loon, unpublished data). However, live observations were taken by unblinded observers, and since the researchers suspected possible bias, the scale was not validated yet from live observations. At the time, only one observer was present, so inter-observer reliability was not possible to assess. Prior to this study it was uncertain the reliability/validity of using the EQUUS-FAP scale for assessing horses in trot with lameness. As “the use of partially validated pain scales to assess animal pain could represent a serious threat to their welfare” (Dalla Costa et al. 2018), it is important to assess the validity and reliability of EQUUS-FAP for different pain states, for live observations but also video and photo observations, and for box-rested and moving horses.

1. Total pain scores

Reliability and validity were first tested for total pain scores, using the modified EQUUS-FAP scale for videos and photos. Both video and photo coding for pain scores on all surfaces had similarly good reproducibility, with only slightly higher intra- than inter-observer reliability. When intra- and inter-observer reliability of this study is compared to other studies based on video or photo analysis, there are similar findings for inter-observer reliability. For example, for box-rested horses, Dalla Costa et al. (2014) used the HGS in pain scoring photos taken from videos, at different times before and after surgical castration, and found high inter-observer reliability (ICC=0.92). In a follow-up study (Dalla Costa et al. 2016) for box-rested horses with acute laminitis, inter-observer reliability was very good (ICC=0.85) for still images, and good (ICC=0.74) for videos. These studies did not calculate intra-observer reliability, and like in this study had good inter-observer reliabilities, though higher in these studies with box-rested horses (in this study, inter-observer reliability for combined surfaces for all photos was ICC=0.73 and for videos ICC=0.70). This could be perhaps because in this study trotting, not box-rested, horses were observed from videos and photos. Other studies of horses in locomotion also found lower inter-observer reliability than in box-rested horses. For example, in Dyson et al. (2018b), based on video recordings of ridden lame horses in trot and canter, “untrained assessors were [only] in fair agreement, amongst each other [(Fleiss Kappa 0.36)] and with the trained assessor [(FK 0.49)], for evaluation of lame horses.” Perhaps inter-observer reliability is lower when observing moving horses compared to box-rested horses due to the difficulty of the task, even when observing horses via video or photo analysis.

Unlike in this study for video and photo observations, live scoring of lameness was found to have much worse inter- than intra-observer reliability, as a “lack of consistency among observers engaged in grading lameness has been previously described” (Fuller et al. 2006; Keegan et al. 2010). In fact, “when grading lameness clinically, intra-observer reproducibility tends to be good, while inter-observer agreement tends to be poor, with observers more likely to differ on the degree of subtle lameness than the degree of overt lameness” (Fuller et al. 2006; Keegan et al. 2010). When using the AAEP lameness scale, “lameness can only be graded at a trot and the inter-observer agreement is generally poor” (Lindegaard et al. 2010), unlike in this study, where pain scores did have slightly lower inter- than intra-observer reliability, but were still acceptable and not poor. This potentially shows the possible influence of bias and the limitations of observing horses under live conditions, leading to worse agreement in pain scores than when scoring pain from videos and photos. However, the AAEP lameness scale is also a unidimensional simple descriptive scale, in which subjective interpretation of observations is potentially greater compared to the multi-parameter approach used in composite pain scales, which could also be an explanation for poor inter-observer agreement.

It was found that photo pain scores (median photo scores from all 5 photos per video were higher than all photo scores) had higher reliability and were more homogeneous than video pain scores, which makes sense. First, it makes sense that photo median scores would have higher reliability than photo total scores, as photo median scores take the median score from each set of photos from each video, so they get rid of the outlying scores per photo set. Second, it makes sense photo pain scores would be more reliable and homogeneous than video scores because in theory, scoring a moving horse in a video is more difficult than in a still photo, even in a slow-motion video. It is difficult to simultaneously pay attention to all the parameters, despite the three times each video was seen when scoring, so less consistency in pain scoring would be reached with videos than photos. There was no time limit in scoring each photo, so if needed, more time could be spent on certain photos to assess all the parameters individually; this was not possible for videos, as each video was approximately 10 seconds long and it was seen only three times. In developing an initial ridden horse facial ethogram, Mullard et al. (2017) used still photographs, as whole body video recordings "would have provided insufficient detail (sharpness) for assessment of the head in isolation, and "still photographs provided an improved level of detail." Dyson et al. (2017) also concluded that "greater detail was available from still images compared with video recordings, permitting more accurate descriptions of facial expressions than could have been obtained from whole horse video recordings." In this study, the videos were not just focused on the face but included the whole body, so indeed, still photographs, even of the whole body, would likely provide more detail on facial parameters. Pilot observations from Van Loon and van Dierendonck (2017) "demonstrated that pain scoring from video recordings resulted in less detailed observation of facial expression and in lower inter-observer reliability," though video scoring was compared with live observations and not photo scoring, and they were testing for postoperative pain originating from the head, not pain associated with lameness.

Third, a similar pattern of photo pain scores for both scorers (especially photo median pain scores) being more homogeneous than video pain scores was also found in Dalla Costa et al. (2016). They found that "a comparison of the application of the Horse Grimace Scale between still images and video footage of horses with laminitis showed that there was less consistency in the analysis of the video recordings." Perhaps photos are simply easier to score than videos, and this is manifested through more reliable pain scores from photos, and more homogeneity in pain scores among both observers for photos in this study. Although they may be easier to score, there are also limitations in scoring photos. In a photo, a horse could be expressing a chance behavior based on its surroundings (for example, position of the ears is in response to a sound in the environment), which is not necessarily indicative of pain, but was just caught in the photo. In a video, it is easier to tell whether such a movement of the ear, for example, is caused by the environment or due to pain in the horse. Photos can show misleading cues, and though videos can be more difficult to score, they bypass this limitation.

The correlation between total photo median pain scores and video pain scores was also tested to compare photo to video scoring. It was found that the agreement between videos and photos was low, as videos were not significantly correlated to photo median pain scores. This makes sense, as "images may not accurately reflect the potentially changing nature of facial expressions in real time" (Miller 2015). All the data from the videos could not be represented in the photos, because it was just 5 photos taken from each video. In Dalla Costa et al. (2014), there was found to be no significant difference in HGS scores (for horses undergoing surgical castration) between still images and videos. However, they extracted their videos and photos from box-rested horses, while in this study, photos and videos were extracted from trotting horses. Even in videos of box-rested horses, horses are mostly still so it may not be so different from photo scoring, while in trotting horses, videos present a much greater challenge to score than still photos. Still photos of trotting horses may also be of lesser quality than a photo taken of a box-rested horse, for example, since they are moving more. Thus, in this study, agreement between video and photo pain scoring would not be expected, unlike in a study like Dalla Costa et al. (2014) where videos and photos were taken from box-rested horses. To bypass such a difficulty in this study, if time were not a limiting factor, more photos could have been extracted and coded from each video to give a more representative pain score based on video than can be had from just 5 photos.

Although good repeatability among video and photo scoring was found, significant differences between baseline and induced lameness conditions for total pain scores from videos or photos were not found. ROC analysis showed that values were in the "fail" range, indicating the scale has inadequate sensitivity and specificity to differentiate between baseline and induced lameness conditions. There was also no significant increase in pain scores for the induced lameness condition compared to baseline. It is clear that in this study, the induced lameness models did lead to lameness with concurrent pain, as the LPS and shoe induced lameness models were previously validated (Merkens and Schamhardt 1988; Carregaro et al. 2014) and the asymmetry in the movement patterns of the horses significantly increased after lameness induction. There can be several reasons why signs of pain were not picked up after induction of lameness; each of these reasons will be elucidated next.

One of the possibilities why signs of pain were not picked up after induction of lameness is due to pain state in horses potentially being confounded with other states, including stress, fatigue, or fear, which could also influence facial expressions. Although facial expressions can change from pain, states such as stress, fatigue, and fear can cause similar alterations in facial expressions, making it difficult to conclude that any change in facial expression is exclusively due to pain. First, stress can be a potential confound that mimics certain pain-related facial features. Stress for horses can include trotting with someone behind them, distractions from surroundings, or "influence of an observer" (Gleerup et al. 2015; Dalla Costa et al. 2016); all these factors can influence "horse facial expressions" and "modify the animals' behaviour" (Dalla Costa et al. 2016; Taffarel et al. 2015). Variation in acclimatization/exposure to the trotting area or to the person who interacts with the horses, such as the equine veterinarian, can also increase stress, and the horses' "response to stress may mimic pain behaviour" (Taffarel et al. 2015). In this study, trotting horses were exposed to experimenters/other human observers, other horses at times, and not always uniform distractions. Although conditions were made as uniform as possible, and much effort was given to acclimatize the horses to the trotting area and conditions, sometimes the experimenter had to urge the horse to trot, which could certainly induce stress for the horse. The human who guided the horse while trotting was also not uniform for all horses, the amount and identities of the observers or other horses in the vicinity were not uniform, the sounds and smells in the arena were not always uniform, and sometimes there were props in the background, such as student backpacks, that may have stressed the horse. Since stress can induce certain alterations as seen in the pain face, and even trotting with someone behind the horse in the control condition could induce stress for the horse, perhaps facial expressions of horses even in baseline could be indicative of stress, and not worsen much after induction of lameness. This would certainly make it difficult to find higher pain scores in the induced lameness condition, if due to stress facial expressions were already indicative of pain at baseline. Conditions in studies of ridden horses were also made as uniform as possible (Dyson et al. 2017, 2018a,b), and even though they did find a difference in pain scores between sound and lame horses, the effects of stress as a potential confound for influencing behavior and facial expressions cannot be excluded. Perhaps stress was a reason why facial parameters had the least pronounced differences between lame and sound horses in Dyson et al. (2018a), for example.

Fatigue can also be a potential confound in reading horse facial expressions, as it can cause similar facial expression changes as pain. Although Dalla Costa et al. (2014) and Dyson et al. (2017) claim orbital tightening (partial or complete closure of eyes) is a manifestation of pain after castration in horses, or pain due to lameness, respectively, others suggest that in the case of Dalla Costa et al. (2014), for example, this appearance of the eyes resembles "horses dozing and may represent a component of fatigue as a consequence of the surgical stress response after castration (Gleerup et al. 2015)" (Dyson et al. 2017). Dyson et al. (2017) claim that in their study, orbital tightening was a sign of pain but "may reflect learned helplessness as a response to chronic pain." It was also found previously that orbital tightening is not only present in ridden lame horses but also horses on the lunge (S. Dyson, unpublished data). An important distinction is made between pain and fatigue in small rodents, where "partial or complete closure of the eyes with tight orbital muscles is a manifestation of pain, which can be differentiated from sleep when the eyes are closed with relaxed periorbital muscles" (Langford et al. 2010; Sotocinal et al. 2011). Although this differentiates pain from fatigue in small rodents, it may or may not be extended to horses, who have different facial features than rodents.

Fear is another emotional state that can cause similar facial expressions to pain, especially for the parameter 'eye white.' It has been found that "the amount of visible white sclera is associated with the expression of fear in many animals, including humans" (Sandem and Braastad 2005; Whalen et al. 2004). Dyson et al. (2017) state that "exposure of sclera was observed more frequently in lame than sound horses and has previously been attributed to fear in both horses (von Borstel et al. 2009) and cattle (Sandem et al. 2004)." However, in the actual study by Dyson et al. (2017), they claim it is unlikely that fear was a major component, as spooky behavior was seen in the environment where lame horses were examined, but it resolved after diagnostic analgesia resolved lameness (S. Dyson, unpublished data). Thus, spookiness was likely a behavioral response to pain from lameness. In studies of horses in locomotion, it is important to consider whether eye white was visible due to movement of the horse's head or camera movement, as "these factors can also influence the amount of sclera visible" (Wathan et al. 2014).

Dalla Costa et al. (2017) claim that "if grimace expressions displayed by horses experiencing pain are pain-specific, they will not appear under different emotional states," which is not what is seen for some of the parameters in EQUUS-FAP. Horses can also be distracted by the environment, and not all identified features of the pain face would be present simultaneously at all times (Gleerup et al. 2015), especially if the horse is in locomotion. Pain could be overridden by external factors such as distractions from other people or horses, sounds, or other emotional states when a horse is in locomotion compared to box-rested, so these factors must be controlled for as much as possible to truly conclude that pain is the cause of changed facial expressions.

A second reason why signs of pain were not picked up after induction of lameness in this study is due to observer bias. For moving horses, observer bias is likely higher in live observations than video/photo observations, making it easier to find a difference between baseline and induced lameness conditions in live observations than video/photo observations. Observer bias is also likely higher for moving than box-rested horses because it is harder to score facial pain expressions of a moving horse, so the result can be less accurate scoring of pain of moving horses and increased subjective bias.

In live observations of trotting horses, it is physically very difficult and demanding to assess all the parameters at normal speed, and still difficult in slow-motion videos. This has the potential for making the effects of observer bias greater when observing trotting horses than when observing box-rested horses who are more-or-less still. In un-blinded direct live observations of trotting horses (using the original EQUUS-FAP scale), there was a difference in pain scores between baseline and induced lameness conditions (J.P.A.M. van Loon, unpublished data), while in blinded observations of videos/photos of trotting horses, a difference between baseline and induced lameness conditions was not found. This difference in outcomes is most likely due to the blinding conditions; "since the observers were aware of the presenting condition of each horse, pain scoring could have been affected by expectation bias" (Tuyttens et al. 2014) to a greater extent in the live scoring versus videos/photos of the trotting horses. However, the difference in outcomes can also partly be due to observing live trotting horses versus trotting horses in videos/photos. Since a quick decision on lameness must be made when the horse is observed trotting live (introducing much room for subjective bias), it is difficult to grasp all the facial parameters live of a moving horse. And since the horse cannot be observed in slow motion or watched multiple times like in videos/photos, this can potentially increase bias and make it easier to find a significant difference in pain scores between baseline and induced lameness conditions from live trotting than when observing videos/photos of trotting horses.

In contrast, in a study of pain scores for box-rested horses with EOTRH (Equine Odontoclastic Tooth Resorption and Hypercementosis), there was no difference between un-blinded direct live observation scores and blinded scores from video (J.P.A.M. van Loon, unpublished data). Perhaps this is because the EOTRH horses were box-rested, so technically it is easier to observe the different parameters, even with live scoring, making the possibility of bias (for knowing whether the horse was baseline or in a disease state) potentially smaller than for trotting horses. Since in this EOTRH study the live observations were also un-blinded, and the video observations were blinded, as for the trotting horses with lameness, the effect of blinding is controlled for; the difference in outcomes in this study might be due to decreased bias when observing box-rested versus moving horses.

A third reason why signs of pain were not picked up at lameness induction is that the EQUUS-FAP scale could be missing subtle elements of pain expression for moving horses, especially since the EQUUS-FAP scale was originally validated on box-rested horses (Van Loon and van Dierendonck 2015; Van Dierendonck and van Loon 2016; Van Loon and van Dierendonck 2017). In this study, the EQUUS-FAP was used on horses in locomotion with unsuccessful results for validity. The EQUUS-FAP scale might have been insufficiently sensitive in discriminating between baseline and induced lameness of moving horses from videos and photos perhaps because the scale is missing more subtle elements of pain expression for moving horses, and perhaps because certain features used in EQUUS-FAP are not actually indicative of pain in horses in locomotion, only in box-rested horses. For example, Mullard et al. (2017) state that "in a clinical situation, it is likely that the horse's facial expression may differ at rest compared with ridden exercise, allowing comparison between the 2, and may vary during ridden exercise depending on the athletic demands being placed on the horse (Christensen et al. 2014), the skill of the rider (Eiserio et al. 2013), and the environment in which it is being worked." Thus, facial expressions are likely to differ when the horse is being ridden, in locomotion, or box-rested, as certain features may be more or less indicative of pain depending on context. Mullard et al. (2017) stressed that specific ethograms are needed for ridden horses, and by extension, horses in locomotion, because "previously published pain scales (Bussieres et al. 2008; Lindegaard et al. 2010; Gleerup et al. 2015; Dalla Costa et al. 2014; Van Loon and van Dierendonck 2015) provided insufficient detail to document, in particular, alterations in ear position, mouth opening, position of the tongue, and position of the head relative to the neck," which are used in the FEReq scale. In Dyson et al. (2017), it was also found that "images of the lame horses standing still had the lowest scores, excluding the position of the head." As these horses were lame, they should have had higher total pain scores than the control horses, but they were standing still, so the FEReq did not pick up on their lameness, as it was originally developed for ridden horses. Thus, scales can be specific for a moving versus box-rested state.

Dyson et al. (2017) state the "FEReq was designed to be used in horses in motion…in contrast to previously described pain ethograms" (Dalla Costa et al. 2014; Gleerup et al. 2015; Van Loon and van Dierendonck 2015). Their study found a difference in total pain score for lame vs. non-lame ridden horses using the FEReq scale, even with blinded coding of photographs. However, the FEReq scale, though a facial expression-based pain scale, includes elements of body and gait parameters, such as the parameters "front of head vertical…" or "head erect, straight…," as it was developed for ridden trotting horses; thus, by extension, the scale does not only assess facial expressions. The EQUUS-FAP, in contrast, was developed for box-rested horses and the parameters are much more exclusively facial focused. Perhaps that is why significant differences in baseline and induced lameness from photos in Dyson et al. (2017) are found, in contrast to in this study, where differences for video/photo observations of trotting horses are not seen, as in both cases horses are moving, but in this study an exclusively facial-focused ethogram developed for box-rested horses was used.

It would be interesting to see if for only the parameters that are exclusively facial-focused in the FEReq ethogram, there is still a significant difference between baseline and lameness conditions. The "facial markers showing the greatest significant difference between lame and sound horses, with multiple likelihoods of occurrence in lame horses compared with sound horses," which were exclusively facial-focused included 'twisted nose,' 'eyes partially or fully closed,' 'tension caudal to the eye area,' 'intense stare of the eye,' and 'open mouth with exposure of teeth' (Dyson et al. 2017). Perhaps if some of these significant parameters were included in the EQUUS-FAP scale, such as 'twisted nose' or 'intense stare of the eye,' that would have enabled us to find a significant difference between baseline and induced lameness horses. As the FEReq scale was successful in finding a significant difference between sound and lame horses for ridden trotting horses, perhaps with only inclusion of facial-focused parameters it would also have good validity, and perhaps including some of their facial-focused parameters in the EQUUS-FAP scale that are not yet included could increase its validity as well.

A fourth reason why signs of pain were not picked up at lameness induction is the quality of videos and photos used in the study could have influenced the accuracy of pain scoring. The video and photo conditions were similar for both baseline and induced lameness, so in that sense, the quality of the videos and photos was controlled for, but if the quality had been better, perhaps the readings for both baseline and induced lameness conditions would have been more accurate, and subsequently, a significant difference in pain scores may have been found between conditions.

Although the quality of the videos and subsequent photos extracted from videos was not "poor," it would have been better to get more "consistent" quality videos, where there were not observers in front of the video sometimes blocking the view of the horse, the trotting horse was not blurry for a few seconds due to auto-adjust of the video, or the video was set on a tripod, for example, so the observer's movement did not influence recording and the horse was recorded from the same position each time. Also, for soft surface video recording, for example, the horse's head was frequently followed and not the whole body, while for the hard surface video recording the whole horse was followed. As statistical analysis was often completed combining both surfaces, and the recordings varied quite a bit, it can be difficult to make generalizations, which perhaps can contribute to the poor validity in this study.

If the videos were not very clear, the photos would be even less clear, as they were photos extracted from the videos. As facial features can be very subtle, such as 'raised upper eyelid,' it can be difficult to determine an accurate score from low quality images. For example, it was found that the "accuracy [or ability to reliably distinguish lame from sound horses] of the HGS (73.3%) was slightly lower than that of the other "grimace scales" (97% for the mouse grimace scale, 82% for the rat grimace scale, and 84% for the rabbit grimace scale)" (Dalla Costa et al. 2014; Langford et al. 2010; Sotocinal et al. 2011; Keating et al. 2012), mostly due to "slightly lower quality for some of the images used compared to those scored in other grimace scales" (Dalla Costa et al. 2014). Accuracy, as determined by Dalla Costa et al. (2014) was "determined by comparing the global pain and no pain judgment made by the treatment and period blind observers with actual pain state of the horse." Nonetheless, the accuracy was much higher than in this study, where ROC curve results for the specificity and sensitivity of the scale indicate failed results near or below 0.50 (Table 12; Figure 7). This is likely due to even lower quality images used in this study than in Dalla Costa et al. (2014) for example, as the images were of moving horses, which are likely to be less clear than images of box-rested horses. Dyson et al. (2017) stressed the "limitations of assessing still images of the head alone, rather than live horses or video recording." They found "no significant change in the pain scores based on the assessment of still images," which could be attributed to the horse's head being very unsteady when lame (before diagnostic analgesia) and "much more still after resolution of lameness." This would cause unclear images of lame horses, and clearer images after abolition of lameness, which of course would make it more difficult to find an accurate difference between sound and lame horses. Although it can be difficult to get clear images of moving horses, a higher quality camera can be used, or at least a higher quality video recording can be made so clearer images can be extracted from videos.

Another influence on why signs of pain were not picked up at lameness induction is that the difficulty of scoring facial pain expressions from videos and photos that are not very clear increases when dark horses are observed, as coat color can have a significant impact on accuracy of reading facial expressions. Dalla Costa et al. (2014) found that the "considerable variation in coat colour of the horses observed" could cause lower accuracy of HGS compared to other "grimace scales." They found that the "coat colour of the horse combined with the quality of some of the images meant that dark horses were often more difficult to score than those with lighter coats, especially if the background was dark," which was true throughout this study when pain scoring dark versus light colored horses. In Dalla Costa et al. (2014), four out of six of the control horses had a light coat, enabling easier scoring. Thus, the control horses not having a difference in HGS scores is a highly reliable finding. Dalla Costa et al. (2014) found a significant difference in pain scores between control horses and horses after surgical castration, by using HGS scoring on photos extracted from videos. In this study, coat color of horses was not a factor when choosing which horses to use for the study, though coat color was controlled for since the same horses were used between baseline and induced lameness conditions. Only one out of eight horses in this study had a light coat, and no difference in pain scores was found between baseline and induced lameness conditions. Perhaps having a majority of horses with dark coat color made it more difficult to accurately score facial expressions.

In Dyson et al. (2017), "images were analyzed blind without prior knowledge of sound or lame status and cropped to provide a neutral background." In the study, it was found that "lame horses had higher pain scores than nonlame horses," potentially attributed to the neutral background used, which contrasted nicely with horse coat color. It was also found in mice that the "higher the quality of the images and a contrasting background allowed the observers to more accurately score the images" (Langford et al. 2010). A contrasting background, or even neutral background, helps to score facial expressions, though it may still be more difficult to score a dark horse's facial expressions than a lighter horse's, as facial features such as muscle movement can be less distinguishable, especially for horses in locomotion. Dyson et al. (2017) also mentioned "the mane and forelock of some sound competition horses were plaited, potentially introducing bias." Thus, hair decorations and coat color should ideally be controlled for, combined with good quality videos and photos, to reduce bias and not negatively affect the accuracy of a pain scale. In this study, perhaps validity of the scale would have differed if an equal number of dark and light horses, for example, was used, as coat color can influence accuracy of pain scoring.

Continuing, signs of pain could have not been picked up at lameness induction because a "cannot see" score was not included and parameters were scored even when it was not clear what score should be given. Especially when presented with the task of scoring facial expressions of dark horses from not the best quality videos or photos, some parameters, such as 'raised upper eyelid,' were very difficult to code for, and a score had to be "guessed" in some instances. In these cases, it would have been useful to include a "cannot see" score, as stressed in Mullard et al. (2017). They claim that such a score "takes more account of human error and chance scoring" and should be used "when it [is] not possible to determine the presence or absence of a feature." Including such an option could potentially improve validity, because giving incorrect scores for certain parameters would have the potential to incorrectly influence the reading on accuracy of the pain scale, impacting whether a difference in pain scores between baseline and induced lameness conditions could be found.

Specifically, Mullard et al. (2017) tried to create an ethogram to "describe facial expressions in photographs from ridden horses." However, "certain behaviors were difficult to identify on the photographs," and 45% of behaviors were scored as "cannot see" for >25% of the observations. For the 65% of individual behaviors that were scored, the median assessor agreement was 70% for all horses, meaning these parameters could be clearly seen and allowed good inter-observer agreement. Including an option to score a parameter as “cannot see” eliminates less accurate scoring, and it would make sense that validity would thus be impacted. If still no difference were found between baseline and induced lameness conditions, as in this study, it would not be because some features were incorrectly scored when they could have been left out (as “cannot see” scores).

Last, differences in pain scores between conditions could have not been found because for horses in locomotion, it may be necessary to also score body and gait parameters, for example, and not just facial features. One of the research questions in this study was whether facial expressions only could be used to assess pain in moving horses, but perhaps the answers to this is no. Although EQUUS-FAP was validated for various conditions in box-rested horses (Van Loon and van Dierendonck 2015; Van Dierendonck and van Loon 2016; Van Loon and van Dierendonck 2017), perhaps including just the facial features in EQUUS-FAP is not enough to accurately score pain in moving horses. Normally, “aspects of behaviour that may be altered by pain include elements of demeanour, posture and gait, as well as interactive behaviour,” (De Grauw and van Loon 2016). As it may be more difficult to assess facial features in moving horses, including body and gait parameters, or facial features that incorporate elements of body and gait, such as those included in the FEReq (Mullard et al. 2017), may be helpful and important.

Dyson et al. (2018a) developed an ethogram for whole-horse behavior (body, gait, facial) of ridden horses, based on video footage, and found that “the most pronounced differences between lame and sound horses were found in body and gait related parameters,” which comprised 14 of the 24 total parameters in the scale. Their whole-horse ridden ethogram was “applied by a single trained assessor to video recordings,” while a subsequent study used video recordings of lame horses assessed “in a random order by a trained assessor and 10 untrained assessors” (Dyson et al. 2018b) with similar results. Both the original and follow-up study (Dyson et al. 2018a,b) found the whole-horse ridden ethogram was “very good at identifying changes in horses (i.e. from lame to non-lame) and significantly decreased [pain] scores were observed across facial, body and gait markers” (Dyson et al. 2018b) after diagnostic analgesia abolished lameness, though changes in facial characteristics were the smallest, compared to body and gait-related parameters.

In this study, the focus was also on horses in locomotion, so perhaps a difference in pain score between baseline and induced lameness horses could have been found had body and gait parameters been included, or had facial parameters including body and gait elements been included, as those used in the FEReq (Mullard et al. 2017). For example, in this study, there was an increase in resistance of horses to go forward and trot when lame, making it necessary for the equine veterinarian to urge the horse forward. “Unwillingness to go forward (having to be kicked repeatedly, verbally encouraged, or hit with a whip) and resistance (e.g. spontaneously stopping) were significantly associated with lameness” in the study by Dyson et al. (2018a), so if this behavior had been scored, for example, in the videos/photos, perhaps this would have increased the differences between pain scores in baseline and induced lameness conditions.

The FEReq scale for facial pain expression that Dyson et al. (2018a) included as part of their whole-horse pain scale included elements of body and gait parameters, such as parameters “front of head vertical…” or “head erect, straight…,” as previously mentioned. In the Dyson et al. (2017) study, where just the FEReq scale was used to differentiate between sound and lame horses, a significant difference was found in pain scores between sound and lame ridden horses. This is in contrast to this study, where the EQUUS-FAP scale parameters did not include body/gait related facial parameters, and no difference in pain scores between baseline and induced lameness conditions was found.

Regardless of the reasons why signs of pain were not picked up after induction of lameness, no significant difference in total pain scores between baseline and induced lameness were found, so it would be hard/would not make sense to find a correlation between pain scores and Qualisys asymmetry index scores. From the Qualisys data it is clear that baseline asymmetry was very minimal (horses were sound) and that after induction of lameness, the induced asymmetry was standardized. Although asymmetry scores do not always indicate lameness, in this study it makes sense to do so, as lameness was induced using two previously validated models (Merkens and Schamhardt 1988, Carregaro et al. 2014), and the degree of lameness induced was outside the gray area of which the asymmetry index might not correlate to lameness. If the pain scores were to be correlated with lameness, they should also be correlated with the asymmetry index scores (which is correlated with lameness). However, no significant correlation between degree of asymmetry and pain scores was found, so the EQUUS-FAP pain scale (modified for videos and photos) cannot adequately be differentiating between lame and sound horses from video and photo observations.

2. Individual pain scores

For individual pain scores, intra-observer reliability for individual parameters was good, but inter-observer reliability was much smaller than intra-observer reliability, depending on the individual parameter. The parameters with the highest intra-class correlation coefficients (considering inter- and intra-observer reliability), for video and photo coding, included 'ears,' 'nostrils,' 'eye white,' 'raised upper eyelid,' 'licking and/or chewing' (videos), 'head' (videos), and 'open mouth' (photos), most likely due to the ease with which these can be coded.

In previous studies, "putting the ears back has been a consistently reliable finding in studies of pain in nonridden horses" (Fureix et al. 2010; Gleerup et al. 2015; Dyson et al. 2018a). Dalla Costa et al. (2016) found the overall reliability for the HGS scale was good for acute laminitis and 'backwards ear position' was the best facial action unit for both still images and videos. In moving horses, importantly, Dyson et al. (2017) found that "ear position was an important differentiator between lame and sound horses and has been widely used previously for pain assessment in horses" (Dalla Costa et al. 2014, 2016; Gleerup et al. 2015; Van Loon and van Dierendonck 2015). They stated that lameness is indicated by ears back, opposing ear positions (i.e. one ear forward and one backward, or one ear to the side and one backward), and that ears forward occurs more frequently in sound horses (Dyson et al. 2017). Ear movements of horses are easy to code, as ears are "obvious and large, with little variation in the shape of the pinna (external ear), making it relatively easy to code ear movements. Often both of the ears are visible simultaneously" (Wathan et al. 2014). Thus, it is not a surprise that the 'ear' parameter has both the highest intra- and inter-observer reliability of all parameters for videos and photos.

However, the 'ear' parameter did not show a significant difference between baseline and induced lameness, as pain scores remained relatively equally high in both baseline and induced lameness conditions, for both scorers in this study on all surfaces. Perhaps the scores were relatively high because in both conditions, horses had a person behind them encouraging/forcing them to trot, which could be quite stressful for the horse. Perhaps without a person being physically behind the animals, encouraging them to trot, a difference between pain scores for both conditions could have been found, where for example in the induced lameness condition there would be higher pain scores than in baseline.

In moving horses, Dyson et al. (2017) also found that an intense stare and exposure of the sclera were significantly different between lame and nonlame horses when assessing still photographs. Although there are not other studies that specifically show good reliability for all the parameters that were determined in this study to have good reliability, future studies should be done to ascertain that the parameters in this study indeed have good reliability for assessing pain scores from videos and photos.

The parameters that were found to have the lowest intra-class correlation coefficients (considering inter- and intra-observer reliability), for video and photo coding, included 'corners mouth/lips' and 'orbital tightening.' The reason for these being the least reliable parameters the observers scored can "reflect lack of attention to detail, less ability to learn, the speed with which they performed their assessments, and genuine inability to recognize the features described in the ethogram" (Mullard et al. 2017), perhaps due to insufficient training. Perhaps these parameters were especially difficult to score or see in a video of a moving horse, leading to not only poor inter-observer, but also intra-observer reliability.

Dalla Costa et al. (2017) found the HGS demonstrated only fair reliability (ICC=0.30) for orbital tightening, the least reliable parameter. They state "it remains to be clarified why orbital tightening was less reliably assessed in the considered conditions," including in the different emotional states of new environment, grooming, anticipation of food reward, and fear. Perhaps it was just difficult to code this parameter, coat color of horses was not controlled for, or all the observers were not trained appropriately. Dalla Costa et al. (2014) found that "in profile view images, horses with dark-brown or black coats were more difficult to score than grey and light brown coat, especially for the orbital tightening." They found that 9% of 'orbital tightening' action units were "not able to score," while it was only 0% for 'stiffly backwards ears,' a parameter with a high ICC. In this study, differences in inter- and intra-observer reliability between white and dark-coated horses were found, though the differences were not significant ($p>0.05$) to make a definitive statement, likely because the sample size was n=8, and only one horse was white. Pain scores for dark and white horses were only significant when considering all photos for the 'orbital tightening' parameter, and it was found that intra-observer reliability was significant and higher for white than dark horses, and inter-observer reliability was significant but lower for white than dark horses. Thus, results are inconclusive for how light and dark horses compare for the 'orbital tightening' and 'corners mouth/lips' parameters, and a larger sample size would have to be used to confirm this.

In developing an ethogram for moving ridden horses, Mullard et al. (2017) found the least consistent observations were relating to eye and muzzle. They found "alterations in the shape of the muzzle were poorly interpreted in the present study," and "the shape of the eye and alteration in tension in the periorbital muscles were also not reliably assessed in the present study." This is similar to the findings in this study, where 'shape of the muzzle' is similar to the parameter 'corners mouth/lips' in this study, and 'shape of the eye and alteration in tension in the periorbital muscles' is similar to the 'orbital tightening' parameter in this study. Since Dalla Costa et al. (2014), for example, found acceptable reliability for 'mouth strained,' but they tested box-rested horses, and both Mullard et al. (2017) and this study found poor reliability for a similar parameter and used horses in locomotion, perhaps 'orbital tightening' and 'corners mouth/lips' or similar parameters are truly more difficult to see and code for in moving horses. Perhaps, however, the results could be different in both studies if it were considered that dark horses should be balanced with white horses; reliability may have increased with more white horses, as their facial expressions may be easier to code.

Although total pain scores did not show a significant difference between baseline and induced lameness conditions, certain individual parameters for certain scorers did have a significantly different score between baseline and induced lameness conditions (significant validity). When the scores of these significant parameters were checked, scores in the induced lameness condition were significantly higher than in the baseline condition, matching expectations. Significant validity was found for the parameters 'head' (video coding, one scorer), 'nostrils,' and 'raised upper eyelid' (photo coding, one scorer), and for each of the parameters, pain scores were higher in the induced lameness than baseline condition. These parameters not only had significant validity, but also were the most reliable individual parameters. The 'head' parameter in particular could have had significant validity and high reliability as it is somewhat of a body parameter as well, where head movement could be related to head nodding due to lameness. Thus, some parameters from the EQUUS-FAP scale for video and photo coding are indeed more reliable and valid to measure pain scores. However, since this is the case for only a few parameters, the pattern is only true for one of the scorers for each of the parameters, and only one parameter is significant for video coding and two for photo coding, the results are not consistent enough to conclude that these parameters are valid in differentiating between baseline and induced lameness conditions.

There was not one loud parameter causing a disruption in differentiating between baseline and induced lameness conditions for total pain scores. For some parameters, there is an increased pain score in induced lameness, for others in baseline, and for other parameters there is no difference between conditions. Individual parameters do not have significant differences between baseline and induced lameness conditions, except for the few significant parameters mentioned previously for one of the two scorers (head, raised upper eyelid and nostrils).

As for total pain scores, for individual parameters, the agreement between videos and photo medians was low. It makes sense, as all the data from videos is not represented in photos, since only 5 photos were used from the whole duration of each video. Only the parameter 'ears' had a good, significant correlation between video and photo median scores; 'nostrils' had an acceptable correlation for one scorer. Ears are easily visible whether in a photo or video, unlike parameters such as 'orbital tightening,' which is quite difficult to see in moving horses. Ear position changed throughout the video, but the majority position throughout the video was chosen as the 'ear' parameter score. Photos likely reflected this majority ear position, enabling a high correlation between 'ear' scores for videos and photos. Nostril dilation "is most obvious during inspiration" (Gleerup et al. 2015), which the horse is doing more of during trotting. Perhaps the nostrils with inspiration is somewhat easy to code for and see in videos and photos, and led to a decent correlation between video and photo scores.

Perhaps there was no correlation between photo median scores and videos because for certain parameters such as 'mouth strained and pronounced chin' and 'tension above eye area,' as used in Dalla Costa et al. (2016), "scoring videos poses different challenges compared to scoring still images with the expression of specific action units changing over time and complicating the assessment." Dalla Costa et al. (2016) claims that "15-sec clips were reported by the assessors to be too short to integrate the information of facial movements in a judgment for each action unit," and that 1-min video clips should have been used. Perhaps with videos, 15 seconds is too short to code for all the horse's features, especially particularly difficult features that are changing throughout a video, and having longer videos to code would help. With photos, the assessor has unlimited time in assessing facial expressions, which is very different from scoring videos, thus making it hard to have a correlation between video and photo pain scores for certain parameters, and by extension for total pain scores.

3. Removed parameters from EQUUS-FAP scale

Removing parameters that do not add to the sensitivity and specificity of a scale (such as being present more in baseline than induced lameness conditions) or that have low repeatability, can possibly increase validity and reliability for total pain scores. As De Grauw and van Loon (2016) state, “reduction in the number of parameters in each composite scale by elimination of those variables that are least sensitive and specific for the pain state under study will further improve validity and reduce the time required for repeated observations.” Consequently, the least significant and reliable parameters (‘corners mouth/lips,’ ‘orbital tightening’) were removed from the EQUUS-FAP pain scale and reliability and validity of total pain scores was recalculated. Intra- and inter-observer reliability of total pain scores remained significant and increased for most surfaces and photo and video scoring, but there were mixed results for validity of total pain scores for video and photo scoring. For some surfaces and modalities (photo or video), significance increased, while for others it decreased or remained the same. However, for all surfaces and modalities, even after exclusion of these two parameters, no significant differences were found for total pain scores between baseline and induced lameness conditions. Upon removal of these parameters, AUROC values remained in the “fail” range, still indicating poor specificity and sensitivity of the scale. Thus, after removal of the least significant and reliable parameters from the EQUUS-FAP pain scale, the resultant pain scale still was unable to differentiate between baseline and induced lameness conditions for total pain scores, despite higher reliability. Removal of parameters also did not uniformly improve the correlation of video to photo median scores.

4. Limitations

There were some limitations to this study, starting with the nature of the study. Lameness was induced and horses with naturally occurring lameness were not tested. The advantage of using a model to induce lameness is that conditions can be standardized to a certain extent, and pre- and post-induction of lameness pain can be compared. With naturally occurring lameness in horses, it is difficult to get a baseline evaluation before pain occurred. A way to bypass this is assessing pre- and post-diagnostic blocking of lameness (as in Dyson et al. 2018a), or assessing after a period of treatment of lameness. The disadvantage to inducing lameness is that it is an attempt to mimic a condition that does not naturally occur that way; lameness does not naturally occur due to injection of LPS in a joint or a screw placed under the foot.

A second limitation is the quality of the video recordings used, and consequently the quality of the photos taken from the videos. Even though only photos where the horse’s face was visible were chosen, as the videos were at times unclear and blurry, the photos retrieved from the videos would be of even lesser quality.

Third, the “influence of an observer on horse facial expressions” (Dalla Costa et al. 2016) and influence of environment was controlled for but there was much room for error. Facial expressions can change when horses react to something in their surroundings (Gleerup et al. 2015), so it is important to keep the environment as uniform as possible for all horses. At times during video recordings, experimenters/other human observers and other horses were present but not consistent throughout videos. For example, sometimes the experimenter had to urge the horse to trot, which did not happen for all horses and through all videos. Horses were also in an open arena and not their own stall, and as a result they could have experienced distress, which is “difficult to correct for” (De Grauw and van Loon 2016). Sometimes, pain can be overridden by external factors, such as distractions from the environment, other people/horses, etc., especially when the horse is moving compared to box-rested. Not all identified features of the pain face are also present simultaneously at all times (Gleerup et al. 2015), especially when the horse is in locomotion, so it is especially important to control for the environment as much as possible when testing the horse.

Continuing, coat color was not considered when choosing horses for this study and only one out of eight of the horses tested had a light coat color, which could provide unbalanced results. Previous studies found that coat color can influence the accuracy of reading facial expressions (Dalla Costa et al. 2014; Dyson et al. 2017), so it is important to consider when choosing horses. Also, in the study there was not an option to include a “cannot see” score for parameters when it was very difficult to read a certain parameter in a particular video or photo. Mullard et al. (2017) especially discusses the importance of including a “cannot see” parameter, as it “takes more account of human error and chance scoring.”

VII. Conclusion and Future Directions

In this study with the modified EQUUS-FAP pain scale for blinded video and photo scoring of trotting horses, good reproducibility was found, though a significant increase in pain-related expressions during guided trot after induction of lameness compared to baseline was not found. The EQUUS-FAP scale cannot be used for video and photo coding of trotting horses to accurately differentiate between horses with induced lameness and healthy pain-free horses, though inter- and intra-observer repeatability were acceptable.

In the future, this study should ideally be redone to see if the same results are acquired when the quality of videos and photos is more detailed and higher resolution. The study should also be redone with a larger sample size of horses, and with assessors from even more professional backgrounds, such as equine technicians, nurses, or veterinary students, for example, to see if reliability or validity improves. Observations in this study were also only done during trot, but in theory it could be useful to assess pain in canter, as it might further inform the horse's lameness condition. "Although detection of a lame limb or limbs may be easier in trot, some horses are more uncomfortable in canter than trot and show more behavioral manifestations of pain" (Dyson et al. 2017), so perhaps more robust differences in pain scores could be found between baseline and induced lameness when observing a horse cantering. However, the practicality of observing horses in canter, especially with induced lameness, would have to be worked out as it is not possible to lead a horse in canter on a rope in a straight line. Further adaptations to the study would have to be done, which could introduce further bias.

As stress was an important potential confound in the study, additional testing might be completed in the future to control for stress. For example, to see if stress during trotting has an impact on pain scores, perhaps an assessment on box-rested horses' pain face can be done before and after induction of lameness. If pain score before and after is much lower than baseline scores during trotting, perhaps stress increases the pain score even for a sound horse by the nature of the experiment. Physical activity could also increase pain scores, however, which is not necessarily stressful for the horse. This was previously done, where facial pain scores were found to increase from trotting only, but by unblinded observers, so this was not included in this study (J.P.A.M. van Loon, unpublished data). In the future it would be valuable to score facial pain expressions before and after induction of lameness in box-rested horses by blinded observers and compare these results to baseline pain scores during trotting, to verify the results found with unblinded observers. Another alternative is to look at the differences in pain face between box-rested horses and trotting horses, not considering lameness, to see what effects stress has on the horse's pain face during trotting.

Coat color should also be controlled for, where for example half the horses should be light colored and the other half dark colored. A contrasting background to the horses' coat colors, or at least a neutral background as used in Dyson et al. (2017), can allow the observers "to more accurately score the images" (Langford et al. 2010). Weighing factors can also be considered for more significant parameters, such as those that were found to be most reliable and valid, including 'head', 'nostrils,' and 'raised upper eyelid,' to potentially improve "sensitivity, specificity, and positive and negative predictive values" (Van Loon and van Dierendonck 2015).

Further, in this study observers could have potentially coded for parameters when it was not clear what score they should have. As Mullard et al. (2017) stressed, it is important to include a "cannot see" parameter, as giving a wrong score for a parameter that is difficult to code for in a particular instance can lead to inaccurate scores biasing results. As for video scoring an 'eyelids' parameter was included, and from video and photo scoring 'orbital tightening' was found to have the lowest intra-class correlation coefficients, perhaps for next time it would be best to not focus on the orbital tightening when coding the 'eyelid' parameter, and only focus on raised upper eyelid and eye white features of the horse's eye, since that may give a more accurate 'eyelid' parameter reading.

Importantly, as has been found in Dyson et al. (2018a), "the most pronounced differences between lame and sound [ridden] horses were found in body and gait related parameters," not facial related parameters, though for all groups of parameters there was a significant difference between pain scores in baseline and induced lameness. Perhaps facial features are not enough to measure pain in trotting horses, and other features for pain assessment can be included, especially since behavioral measures (not just facial expressions) "may be easier for a rider or trainer or a non-specialised veterinarian to recognise than low-grade lameness and therefore facilitate the identification of an underlying pain-related problem" (Dyson et al. 2018b). When assessing behavioral parameters that are not facial, however, there can still be influences on horse behavior from the rider or the environment, which could bias the findings, so it remains very important to keep situations for assessing pain as standardized as possible, even when non-facial related behavioral parameters are used.

Other features of pain assessment are not limited to but can include body and gait parameters as used in Dyson et al. (2018a), or non-invasive, physiological parameters, such as blink rate. It has been found that an “increased blink rate has been linked positively to the amount of dopamine in the basal ganglia of the brain (Swerdlow et al. 2003), and dopamine levels may increase as a result of pain (Raekallio et al. 1997).” A horse’s temperament should be considered, since “anxious horses may display a greater blink rate response to stress compared with more stoical horses” (Roberts et al. 2016). Eye temperature measured through thermography can also be indicative of pain state in horses (Hall et al. 2014). If gait related parameters were to be included, it could be useful to include automated gait analysis such as Qualisys, which offers an objective method for lameness detection. A pain scale would still be necessary, however, to determine pain scores, even with a method for objective lameness detection available.

Effort should continue to be devoted to validating one robust pain scale for different pain conditions, also for photos and videos, rather than using different pain scales for different conditions, as this creates some uniformity and ease in assessing pain for clinicians and horse-owners. Also, effort should continue towards creating a computerized facial pain recognition system, as has already been developed for humans. It has been found that in sheep with acute pain, “computerised technology already has shown to be able to enhance pattern recognition in facial expression” (Hutson 2017; Lu et al. 2017), and similar measures should be continued in horses.

In conclusion, this study does not support the clinical application of the modified EQUUS-FAP pain scale (for videos and photos) using video and photo coding for horses in locomotion with lameness. Perhaps facial parameters are simply not enough to assess pain status in lame moving horses, and it is necessary to include gait and body parameters, for example, in a scale for horses in locomotion, but future studies will have to confirm this. Future studies should also be done to better understand the added value of including facial expressions in pain scales.

Acknowledgements

I would like to thank my supervisor, Dr. Thijs van Loon, for teaching me much about statistical analysis, and guiding me over the past months as I have flown for interviews, deliberated many decisions, and dealt with the global pandemic, all while working on my thesis. Thank you for your patience and time, I appreciate it. I would also like to thank my other supervisor, Dr. Filipe Serra Braganca, who was busy with his own PhD but found the time to look over my results, help with statistical analysis and suggest further steps. I would like to thank the Equine department at the Utrecht University Veterinary School, previous veterinary students, veterinary assistants, and animal caretakers for collecting the data I used for my thesis, taking care of the horses, and enabling this project to happen.

References

Ashley, F.H., A.E. Waterman-Pearson and H.R. Whay. (2005). Behavioural assessment of pain in horses and donkeys: Application to clinical practice and future studies. Equine Veterinary Journal 37, 565–575.

Boissy, A., A. Aubert, L. Desire et al. (2011). Cognitive sciences to relate ear postures to emotions in sheep. Animal Welfare, 20:47–56.

Bussières, G., C. Jacques, O. Lainay, G. Beauchamp, A. Leblond, J.L. Cadoré, L.M. Desmaizières, S.G. Cuvelliez, E. Troncy. (2008). Development of a composite orthopaedic pain scale in horses. Research Veterinary Science, 85:294-306.

Caine, N. (1992). Humans as predators: observational studies and the risk of pseudohabituation. In: The Inevitable Bond: Examining Scientist-Animal Interactions, Eds: H. Davis and D. Balfour, Cambridge University Press, UK. pp 357-364.

Carregaro, A.B., G.C. Freitas, M.H. Ribeiro, N.V. Xavier, R.G. Dória. (2014). Physiological and analgesic effects of continuous-rate infusion of morphine, butorphanol, tramadol or methadone in horses with lipopolysaccharide (LPS)-induced carpal synovitis. BMC Veterinary Research 10, 966.

Casey, R.A. (2004). Clinical problems associated with the intensive management of performance horses. In: Waran, N. (Ed.), The Welfare of Horses. Kluwer Academic Publishers, Dordrecht, pp. 19–44.

Casey, V., P. McGreevy, E. O'Muiris and O. Doherty. (2013). A preliminary report on estimating the pressures exerted by a crank noseband in the horse. Journal of Veterinary Behavior: Clinical Applications Research, 8:479e484.

Daeninck, P., B. Gagnon, R.E. Gallagher et al. (2016). Canadian recommendations for the management of breakthrough cancer pain. Current Oncology, 23(2):96.

Dalla Costa, E., D. Bracci, F. Dai, D. Lebelt, M. Minero. (2017). Do different emotional states affect the horse grimace scale score? A pilot study. Journal Equine Veterinary Science, 54:114-117.

Dalla Costa, E., D. Stucke, F. Dai, M. Minero, M.C. Leach, D. Lebelt. (2016). Using the Horse Grimace Scale (HGS) to assess pain associated with acute laminitis in horses. Animals, 6:47.

Dalla Costa, E., M. Minero, D. Lebelt, D. Stucke, E. Canali, M.C. Leach. (2014). Development of the Horse Grimace Scale (HGS) as a pain assessment tool in horses undergoing routine castration. Plos One, 9:e92281.

Dalla Costa, E.D., R. Pascuzzo, M.C. Leach et al. (2018). Can grimace scales estimate thte pain status in horses and mice? A statistical approach to identify a classifier. Plos One, 13(8):e0200339.

De Grauw, J.C. and J.P.A.M. van Loon. (2016). Systemic pain assessment in horses. The Veterinary Journal, 209:14-22.

De Grauw, J.C., C.H. van de Lest, R. van Weeren, H. Brommer, P.A. Brama. (2006). Arthrogenic lameness of the fetlock: Synovial fluid markers of inflammation and cartilage turnover in relation to clinical joint pain. Equine Veterinary Journal, 38:305–311.

De Vellis, R.F. (2003). Second Ed. Scale Development, Theory and Applications, vol. 26. Sage Publications Inc., Thousand Oaks, CA.

Defensor, E.B., M.J. Corley, R.J. Blanchard, D.C. Blanchard. (2012). Facial expressions of mice in aggressive and fearful contexts. Physiology and Behavior, 107:680–5.

Dobromylskyj, P., P.A. Flecknell, B.D. Lascelles, A. Livingston, P. Taylor, A. Waterman-Pearson. (2000). Pain assessment. In: Flecknell, P.A., Waterman-Pearson, A. (Eds.), Pain Management in Animals. W.B. Saunders, London, UK, pp. 53–79.

Ducasse, S.G. (2020). Cronbach's Alpha: Simple Definition, Use and Interpretation [Statistics How To]. Retrieved from https://www.statisticshowto.com/cronbachs-alpha-spss/.

Dyson, S., J. Berger, A.D. Ellis, J. Mullard. (2018a). Development of an ethogram for a pain scoring system in ridden horses and its application to determine the presence of musculoskeletal pain. Journal of Veterinary Behaviour, 23:47-57.

Dyson, S. and J. van Dijk. (2018b). Application of a ridden horse ethogram to video recordings of 21 horses before and after diagnostic analgesia: Reduction in behaviour scores. Equine Veterinary Education, doi: 10.1111/eve.13029.

Dyson, S., J.M. Berger, A.D. Ellis, J. Mullard. (2017). Can the presence of musculoskeletal pain be determined from the facial expressions of ridden horses (FEReq)? Journal of Veterinary Behaviour, 19:78-89.

Eiseriõ, M., L. Roepstorff, M. Wesihaupt, A. Egenvall. (2013). Movements of the horse's mouth in relation to horse-rider kinematic variables. Veterinary Journal, 198:e33ee38.

Flecknell, P. (2000a). Animal pain – an introduction. In: Flecknell, P.A., Waterman-Pearson, A. (Eds.), Pain Management in Animals. W.B. Saunders Company, Philadelphia, pp. 1–7.

Fuller, C., B. Bladon, A. Driver, A. Barr. (2006). The intra- and interassessor reliability of measurement of functional outcome by lameness scoring in horses. Veterinary Journal 171, 281e286.

Fureix, C., H. Menguy. and M. Hausberger. (2010). Partners with bad temper: reject or cure? A study of chronic pain and aggression in horses. PLoS ONE, 5:e12434.

Gleerup, K.B., B. Forkman, C. Lindegaard and P.H. Andersen. (2015). An equine pain face. Veterinary Anaesthesia and Analgesia, 42:103-114.

Gleerup, K.B. and C. Lindegaard. (2016). Recognition and quantification of pain in horses: a tutorial review. Equine Veterinary Education, 28:47–57.

Grandin, T, D. Mark. (2002). Distress in animals: is it fear, pain or physical stress. Manhattan Beach (CA): American Board of Veterinary Practitioners Symposium.

Graubner, C., V. Gerber, M. Doherr, C. Spadavecchia. (2011). Clinical application and reliability of a post abdominal surgery pain assessment scale (PASPAS) in horses. The Veterinary Journal, 188:178–183.

Hall, C., R. Kay, K. Yarnell. (2014). Assessing ridden horse behavior: professional judgment and physiological measures. Journal Veterinary Behavior: Clinical Applications of Research 9, 22e29.

Hansen, B. (1997). Through a glass darkly: using behavior to assess pain. Seminars in Veterinary Medicine and Surgery, 12:61–74.

Haussler, K.K. and H.N. Erb. (2006). Mechanical nociceptive thresholds in the axial skeleton of horses. Equine Veterinary Journal, 38(1):70-5.

Hintze S, S. Smith, A. Patt, I. Bachmann, H. Würbel. (2016). Are eyes a mirror of the soul? What eye wrinkles reveal about a horse's emotional state. PLoS One, 11:e0164017.

Hutson, M. (2017). Artificial intelligence learns to spot pain in sheep. Science doi: http://dx.doi.org/10.1126/science.aan6918 June.

Ijichi, C., L.M. Collins, R.R. Elwood. (2013). Pain expression is linked to personality in horses. Applied Animal Behaviour Science, 152:35–43.

Jochle, W., J. Moore, J. Brown, G.J. Baker, J.E. Lowe, S. Fubini, M.J. Reeves, J.P. Watkins, and N.A. White. (1989). Comparison of detomidine, butorphanol, flunixin meglumine and xylazine in clinical cases of equine colic. Equine Veterinary Journal, 21:111-116.

Jones, E., I. Viñuela-Fernandez, R.A. Eager, A. Delaney, H. Anderson, A. Patel, D.C. Robertson, A. Allchorne, E.C. Sirinathsinghji, E.M. Milne, N. Macintyre, D.J. Shaw, N.K. Waran, J. Mayhew and S.M. Fleetwood-Walker. (2007). Neuropathic changes in equine laminitis pain. Pain, 132:321-331.

Juarbe-Diaz, S.V., K.A. Houpt, R. Kusunose. (1998). Prevalence and characteristics of foal rejection in Arabian mares. Equine Veterinary Journal, 30:424–428.

Keating, S.C., A.A. Thomas, P.A. Flecknell, M.C. Leach. (2012). Evaluation of EMLA cream for preventing pain during tattooing of rabbits: changes in physiological, behavioural and facial expression responses. PLoS One, 7:e44437.

Keegan, K., E. Dent, D. Wilson, et al. (2010). Repeatability of subjective evaluation of lameness in horses. Equine Veterinary Journal 42, 92e97.

Kester, W.O. (1991). Definition and Classification of Lameness. Guide for Veterinary Service and Judging of Equestrian Events. American Association of Equine Practitioners (AAEP), Lexington, Kentucky, USA, p. 19.

König van Borstel, U., E.K. Visser, C. Hall. (2017). Indicators of stress in equitation. Applied Animal Behaviour Science, 190:43–56.

Langford, D.J., A.L. Bailey, M.L. Chanda, S.E. Clarke, T.E. Drummond, S. Echols, S. Glick, J. Ingrao, T. Klassen-Ross, M.L. Lacroix-Fralish, et al. (2010). Coding of facial expressions of pain in the laboratory mouse. Nature Methods, 7:447–449.

Leach M.C., C.A. Coulter, C.A. Richardson, P.A. Flecknell. (2011). Are we looking in the wrong place? Implications for behavioural-based pain assessment in rabbits (Oryctolagus cuniculi) and beyond? PloS one, 6:e13347.

Lindegaard, C., D. Vaabengaard, M.T. Christophersen, C.T. Ekstøm, and J. Fjeldborg. (2009). Evaluation of pain and inflammation associated with hot iron branding and microchip transponder injection in horses. American Journal of Veterinary Research, 70:840-847.

Lindegaard, C., M.H. Thomsen, S. Larsen, and P.H. Andersen. (2010). Analgesic efficacy of intra-articular morphine in experimentally induced radiocarpal synovitis in horses. Veterinary Anaesthesia and Analgesia, 37:171-185.

Love, E.J. (2009). Assessment and management of pain in horses. Equine Veterinary Education, 21:46–48.

Love, E.J., P.M. Taylor, C. Clark, H.R. Whay, J. Murrell. (2009). Analgesic effect of butorphanol in ponies following castration. Equine Veterinary Journal, 41:552–556.

Lu, Y., M. Mahmoud, P. Robinson. (2017). Estimating sheep pain level using facial action unit detection. 12th IEEE International Conference on Automatic Face & Gesture Recognition doi:http://dx.doi.org/10.1109/FG. 2017.56.

Manfredi, J., D. Rosenstein, J. Lanovaz, S. Nauwelaerts, and H. Clayton. (2009). Fluoroscopic study of oral behaviors in response to the presence of a bit and the effects of rein tension. Comparative Exercise Physiology, 6:143e148.

McDonnell, S. (2005). Is it psychological, physical, or both? Proceedings of the American Association of Equine Practitioners, 51:231-238.

McDonnell, S.M. (2008). Practical review of self-mutilation in horses. Animal Reproductive Sciences, 107:219-228.

McLean, A. and P. McGreevy. (2010). Horse-training techniques that may defy the principles of learning theory and compromise welfare. Journal of Veterinary Behavior: Clinical Applications Research, 5:187e195.

Merkens, H.W. and H.C. Schamhardt. (1988). Evaluation of equine locomotion during different degrees of experimentally induced lameness I: Lameness model and quantification of ground reaction force patterns of the limbs. Equine Orthopaedics, 6:99-106.

Miller A.L., M.C. Leach. (2015). The mouse grimace scale: A clinically useful tool? PloS one 10: 1±10.

Molony, V. and J.E. Kent. (1997). Assessment of acute pain in farm animals using behavioural and physiological measurements. Journal of Animal Science, 75:266-272.

Mullard, J., J.M. Berger, A.D. Ellis, S. Dyson. (2017). Development of an ethogram to describe facial expressions in ridden horses (FEReq). Journal of Veterinary Behavior, 18:7–12.

Niinisto, K.E., R.V. Korolainen, M.R. Raekallio et al. (2010). Plasma levels of heat shock protein 72 (hsp72) and beta-endorphin as indicators of stress, pain and prognosis in horses with colic. The Veterinary Journal, 184:100-4.

Obel, N. (1948). Studies on the Histopathology of Acute Laminitis. PhD Dissertation. Swedish University of Agricultural Sciences. Almqvist and Wiksells, Uppsala, Sweden.

Pelfort, X., R. Torres-Claramunt, J.F. Sánchez-Soler, P. Hinarejos et al. (2015). Pressure algometry is a useful tool to quantify pain in the medial part of the knee: An intra- and inter-reliability study in healthy subjects. Orthopaedics & Traumatology: Surgery & Research, 101(5):559-63.

Price, J., S. Catriona, E.M. Welsh and N.K. Waran. (2003). Preliminary evaluation of a behaviour-based system for assessment of postoperative pain in horses following arthroscopic surgery. Veterinary Anaesthesia and Analgesia, 30:124-137.

Pritchett, L.C., C. Ulibarri, M.C. Roberts, R.K. Schneider and D. Sellon. (2003). Identification of potential physiological and behavioral indicators of postoperative pain in horses after exploratory celiotomy for colic. Applied Animal Behavior Sciences, 80:31-43.

Raekallio, M., P.M. Taylor and R.C. Bennett. (1997). Preliminary investigations of pain and analgesia assessment in horses administered phenylbutazone or placebo after arthroscopic surgery. Veterinary Surgery, 26:150-155.

Rédua, M.A., C.A. Valadão, J.C. Duque and L.T. Balestrero. (2002). The pre-emptive effect of epidural ketamine on wound sensitivity in horses tested by using von Frey filaments. Veterinary Anaesthesia and Analgesia, 29(4):200-6.

Reid, K., C.W. Rogers, G. Gronqvist, E.K. Gee, and C.F. Bolwell. (2017). Anxiety and pain in horses measured by heart rate variability and behavior. Journal of Veterinary Behavior: Clinical Applications and Research, 22:1–6.

Roberts, K., A.J. Hemmings, M. Moore-Colyer, M.O. Parker, S.D. Mcbride. (2016). Neural modulators of temperament: A multivariate approach to personality trait identification in the horse. Physiological Behavior 167, 125e131.

Rutherford, K.M.D. (2002). Assessing pain in animals. Animal Welfare, 11:31-53.

Sandem, A., A. Janczac, B. Braastad. (2004). A short note on effects of exposure to a novel stimulus (umbrella) on behavior and percentage of eye-white in cows. Applied Animal Behavior Science 89, 309e314.

Sandem, A.I., B.O. Braastad. (2005). Effects of cow–calf separation on visible eye white and behaviour in dairy cows—A brief report. Applied Animal Behaviour Science, 95(3–4):233-9.

Seibert, L.M., V. Parthasarathy, C.M. Trim, and S.L. Crowell-Davis. (2003). An ethogram of post-anesthetic recovery behaviors in horses: comparison of pre- and post-anesthetic behaviors. Proceedings of the American College of Veterinary Anesthesiologists 27th Annual Meeting, Orlando, Florida, 10–11 October 2002 – abstract. Veterinary Anaesthesia Analgesia, 30:100-120.

Sotocinal, S., R. Sorge, A. Zaloum, A. Tuttle, L. Martin, J. Wieskopf, J. Mapplebeck, P. Wei, S. Zhan, S. Zhang, J. McDougall, O. King, J. Mogil. (2011). The Rat Grimace Scale: a partially automated method for quantifying pain in the laboratory rat via facial expressions. Molecular Pain 7, 55.

Sutton, G.A. and L. Bar. (2016). Refinement and revalidation of the Equine Acute Abdominal Pain Scale (EAAPS). Israel Journal of Veterinary Medicine, 71:15–23.

Sutton, G.A., O. Paltiel, M. Soffer, D. Turner. (2013b). Validation of two behaviour-based pain scales for horses with acute colic. The Veterinary Journal, 197:646–650.

Sutton, G.A., R. Dahan, D. Turner, O. Paitiel. (2012). A behaviour based pain scale for horses with acute colic: scale construction. Veterinary Journal, 196:394-401.

Sutton, G.A., R. Dahan, D. Turner, O. Paltiel. (2013a). A behaviour-based pain scale for horses with acute colic: Scale construction. The Veterinary Journal, 196:394–401.

Sweeting, M., C. Houpt, and K. Houpt. (1985). Social facilitation of feeding and time budgets in stabled ponies. Journal of Animal Science, 60:369-374.

Swerdlow, N.R., L. Wasserman, J. Talledo, R. Casas, P. Bruins, D.L. Braff et al. (2003). Prestimulus modification of the startle reflex: relationship to personality and physiological markers of dopamine function. Biological Psychology 62, 17e26.

Taffarel, M.O., S.P. Luna, F.A. de Oliveira, G.S. Cardoso, J.M. Alonso, J.C. Pantoia, J.T. Brondani, E. Love, P. Taylor, K. White, J.C. Murrell. (2015). Refinement and partial validation of the UNESP-Botucatu multidimensional composite pain scale for assessing postoperative pain in horses. BMC Veterinary Research, 11:83.

Taylor, P.M., P.J. Pascoe and K.R. Mama. (2002). Diagnosing and treating pain in the horse – Where are we today? Veterinary Clinics of North America: Equine Practice, 18:1–19.

Tuyttens, F.A.M., S. de Graaf, J.L.T. Heerkens, L. Jacobs, E. Nalon, S. Ott, L. Stadig, E. van Laer, B. Ampe. (2014). Observer bias in animal behaviour research: can we believe what we score, if we score what we believe? Animal Behaviour 90, 273–280.

Universiteit Utrecht, Faculteit Diergeneeskunde / Stichting De Paardenkamp. (2018). Equine Pain and Welfare App. Retrieved from https://www.epwa.nl/en/.

Van Dierendonck, M.C. and J.P.A.M. van Loon. (2016). Monitoring acute equine visceral pain with the Equine Utrecht University Scale for Composite Pain Assessment (EQUUS-COMPASS) and the Equine Utrecht University Scale for Facial Assessment of Pain (EQUUS-FAP): A validation study. The Veterinary Journal, 216:175-77.

Van Loon, J.P.A.M. and M.C. van Dierendonck. (2015). Monitoring acute equine visceral pain with the Equine Utrecht University Scale for Composite Pain Assessment (EQUUS-COMPASS) and the Equine Utrecht University Scale for Facial Assessment of Pain (EQUUS-FAP): A scale-construction study. The Veterinary Journal, 206:356-64.

Van Loon, J.P.A.M. and M.C. van Dierendonck. (2017). Monitoring equine head-related pain with the Equine Utrecht University scale for facial assessment of pain (EQUUS-FAP). The Veterinary Journal, 220:88-90.

Van Loon, J.P.A.M. and M.C. van Dierendonck. (2018). Objective pain assessment in horses (2014-2018). The Veterinary Journal, 242:1-7.

Van Loon, J.P.A.M. and M.C. van Dierendonck. (2019). Pain assessment in horses after orthopaedic surgery and with orthopaedic trauma. The Veterinary Journal, 246:85-91.

Van Loon, J.P., V.S. Jonckheer-Sheehy, W. Back, P. Rene van Weeren, and L.J. Hellebrekers. (2014). Monitoring equine visceral pain with a composite pain scale score and correlation with survival after emergency gastrointestinal surgery. The Veterinary Journal, 200:109–115.

Van Weeren, P.R., T. Pfau, M. Rhodin, L. Roepstorff, F. Serra Bragança, and M.A. Weishaupt. (2017). Do we have to redefine lameness in the era of quantitative gait analysis? Equine Veterinary Journal, 49:567-569.

Von Borstel, U., I. Duncan, A. Shoveller, K. Merkies, L. Keeling, S. Millman. (2009). Impact of riding in a coercively obtained Rollkur posture on welfare and fear of performance horses. Applied Animal Behavior Science 116, 228e236.

Wagner, A.E. (2010). Effects of stress on pain in horses and incorporating pain scales for equine practice. Veterinary Clinics of North America. Equine Practice, 26:481–492.

Wathan J., K. McComb. (2014). The eyes and ears are visual indicators of attention in domestic horses. Current Biology, 24(15):R677–9.

Wathan, J., A.M. Burrows, B.M. Waller and K. McComb. (2015). EquiFACS: The Equine Facial Action Coding System. PLOS One, 10(8):e0131738.

Whalen P.J., J. Kagan, R.G. Cook, F.C. Davis, H. Kim H, S. Polis, et al. (2004). Human Amygdala Responsivity to Masked Fearful Eye Whites. Science. 306(5704):2061.

Williams, A.C.D.C. (2002). Facial expression of pain: an evolutionary account. Behavioral and Brain Sciences, 25:439-488.

Wyn-Jones, G. (1988). Equine Lameness. Blackwell Scientific, Oxford, UK.