Quantification of pain from facial expression characteristics assessed from photos of horses and donkeys

W.W.A.J. (Whitney) Zwitserloot
Student Number : 4265645
Master Veterinary Medicine
Research thesis 2019
Supervisor: dr. J.P.A.M. (Thijs) van Loon
Department of Equine sciences, Faculty of Veterinary Medicine,
Utrecht University, The Netherlands

Table of Contents

1 Abstract

In the past decade, there has been more attention paid to the evaluation of pain in animals. Assessment of pain has become more important to horse and donkey welfare. This study describes the assessment of objective scoring on photos of horses and donkeys measured with two newly developed pain scales: the Horse Pain Face (HPF) and the Donkey Pain Face (DPF). In this research, 1654 horse photos and 534 donkey photos were used. The photos were scored by two trained observers, with variable levels of equine experience, to determine the inter- and intra-observer reliability by intra-class correlation analysis (ICC), determination of Cronbach's alpha and Bland Altman analysis. For the intra-observer reliability 20% of the photos from both the horse and donkey photo sets were rescored (n=331 ; n=107 respectively). The assessment of photos were performed simultaneously and the observers did not discuss their findings. This research also analyzed the HPF and DPF inter-observer reliability of three observers. The third observer scored 20% photos that were used to asses intra-observer reliability. The HPF scored a very good inter-observer reliability (Cronbach's alpha = 0.92 ; P<0.001) and the DPF scored an acceptable inter-observer reliability (Cronbach's alpha = 0.79 ; P<0.001), while good intra-observer reliabilities were found for both the HPF and the DPF by the first two observers. (Observer 1: Cronbach's alpha = 0.81 for the HPF ; Cronbach's alpha = 0.84 for the DPF ; P<0.001) (Observer 2 : Cronbach's alpha = 0.88 for the HPF ; Cronbach's alpha = 0.97 for the DPF ; P<0.001). Both the HPF and the DPF scored an excellent inter-observer reliability by the three observers. (Cronbach's alpha = 0.91 for the HPF ; P<0.001 (n=331) and Cronbach's alpha = 0.92 for the DPF; P<0.001 (n=107)). Both pain scales scores proved valid for objective pain and repeatable assessment of pain in horse and donkey photos. This study showed the importance of training to obtain valid and reproducible results with the two pain scales. The results of this study will be used in a future project using neural networks and deep learning techniques for automated pain scoring from facial expressions.

2 Introduction

2.1 Horse and donkey

In recent decades, increasing attention has been paid to the issue of pain evaluation in animals, in line with growing concerns around the ethics of animal welfare (van Loon, and van Dierendonck, 2019).
Pain is defined by the International Association for the Study of Pain (IASP) as an '*unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage*' (Merskey and Bogduk, 1994). The physiology of pain ranges from the concept of a simple reflex arc to a complex of interactions at all levels of the peripheral and central nervous system (McFadzean and Love, 2019). Pain expression depends on the origin and type of pain, as well as the differences between species. Acute, chronic, somatic and visceral pain all manifest differently (Robertson, 2006). Acute pain can result from a specific injury or disease, whereas chronic pain is considered a disease in itself (Lamont, et al., 2000; McFadzean and Love, 2019).

The horse and the donkey were domesticated thousands of years ago and have since lived alongside humans all around the world. Humans have often denigrated the donkey as a lowly beast of burden (Burden and Thiemann, 2015). Considering it the 'poor relation' of its often more respected 'cousin', the horse. The domestic donkey is an undervalued species. However, there are many similarities between horses and donkeys. Both are used for the provision of milk, dairy products and meat. They also have roles as pets and companions (Burden and Thiemann, 2015; Librado et al., 2016; Morrow et al., 2011).
There are also differences between the horse and the donkey. Horses, unlike other ungulates, were not only used as a source of dairy products and meat, as their stamina and speed were also used to revolutionize warfare (Librado et al., 2016). The donkey differs from the horse in its physical traits, behaviour and propensity to disease. Variations in physiology also lead to differences in its metabolism and distribution of drugs. Donkeys have more subtle behaviour traits than the horse. They show different behavioural repertoires when frightened, in pain, or under pressure. The anatomical differences are shown especially in the hoof, upper airway, and their conformation (Burden and Thiemann, 2015). The variation in chromosome numbers, the horse having 64 chromosomes and the donkey 62, makes the offspring of these two species infertile (Morrow et al., 2011). The donkey is also less of a flight animal than the horse and can even better be considered as a fight animal (Burden and Thiemann, 2015). The horse, having evolved as a flight animal, cannot express pain too openly, as it seeks to avoid predation (Taylor et al., 2002).

2.2 Pain scoring systems

Animals are not capable of expressing themselves using words. Therefore, adequate diagnosis and treatment depends on accurate recognition of their experiences of pain. This recognition depends on a variety of factors, including environmental characteristics, breed, individual variations, and drugs (Bussieres et al., 2008). There is no 'gold standard' method available to veterinarians for the assessment of pain (Dalla Costa et al., 2014). In contrast to humans, where the gold standard is that a person gives a score to its own pain and expresses it in words. A pain scoring system that was originally used for humans is the Visual Analogue Scale (VAS). A human patient marks its pain on a horizontal 10 cm line, representing pain intensity from none

at the beginning (left) of the line to the worst imaginable pain (right) (de Grauw and van Loon, 2016; DeLoach et al., 1998).

In practice, it is important that pain scoring systems are quick, reliable, and easy-to-use. Additionally, the system should be repeatable, with different observers producing consistent results (Wagner, 2010). The purpose of the pain assessment systems is to detect mild, moderate, and severe pain (Ashley et al., 2005). Several studies have found behavioural, endocrinal and physiological parameters for pain in horses using the Visual Analogue Scale (VAS) and the Composite Pain Scales (CPS) (de Grauw and van Loon, 2016; Wagner, 2010). However, VAS scores in horses, provided by humans, are very subjective and have a low inter-observer reliability (Lindegaard et al., 2010).

Facial expression-based pain scales have been widely used for humans, as people are able to use facial expressions to communicate with one another (Haxby et al., 2000). A study by Prkachin and Mercer on adult human facial expressions asked participants to conceal their pain and noted that information was nevertheless 'leaked' through their facial expression (Prkachin and Mercer, 1989). With young children who are not capable of complex communication, facial expressions are an important means of registering pain (Poole and Craig, 1992). Facial expressions have been structurally described by facial action coding systems. The method describes a contraction of a single muscle or a group of muscles that lead to a specific facial expression, referred to as 'Facial Action Units'. This method is considered accurate and reliable for humans (Friesen and Ekman, 1978).

A new facial action coding system for horses was developed in 2015, named 'EquiFACS'. This system provides an extensive list of multiple facial expressions, identified by underlying facial musculature and muscle movement (Wathan et al., 2015). The system defines 17 distinct Action Units. With EquiFACS, the horse community now possesses a standardized language with which it can share information about all possible horse facial expressions. At the same time, researchers published the first study on facial expressions of horses with acute pain, the Horse Grimace Scale (HGS). HGS is a pain coding system using the facial expressions detectable in horses with acute laminitis (Dalla Costa et al., 2016) or in horses that have undergone surgical routine castration (Dalla Costa et al., 2014). The HGS is a valid and effective method of assessing pain, with good inter-observer reliability. The method is painless for the horse and can be easily repeated several times. Researchers at Utrecht University constructed the Equine University Utrecht Scale for Facial Assessment of Pain (EQUUS-FAP). This facial pain score scale has been successfully used for the facial assessment of pain in horses with acute colic (van Loon and van Dierendonck, 2015) and in those with acute or postoperative pain originating in the head, including dental pain, ocular pain and trauma to the skull (van Loon and van Dierendonck, 2017). Its inter-observer reliability and validity are high, and it scores well for sensitivity and specificity. Most recently, EQUUS-FAP was also used in horses with orthopaedic trauma and after orthopaedic surgery (van Loon and van Dierendonck, 2019). The EQUUS-FAP has proven successful in repeatable scoring of pain. At the same time, another study was conducted using induced experimental pain, specifically the application of two noxious stimuli, topical capsaicin, an irritating ointment with pepper extracts and a tourniquet to the antebrachium. The objective of this study was to investigate whether an equine pain face existed and whether it could be described in detail. This is known as the Equine Pain Face. This latter study also showed the possibilities of facial expressions in the characterisation of pain in horses (Gleerup et al., 2015).

With over 44 million donkeys worldwide, it is essential that veterinarians and donkey owners are able to understand and treat the animal effectively (Burden and Thiemann, 2015). Donkeys are often wrongly considered to be incapable of feeling pain, being seen as stubborn and stupid (Regan et al., 2014).

There have been numerous studies on the welfare of donkeys. Working horses, mules and donkeys are an essential transportation resource in developing countries all over the world. They are often owned by poor people and live and work under very harsh conditions. For these owners, animal welfare should be of great concern. Pritchard et al. developed a protocol for the assessment of the welfare of these working animals in urban and peri-urban areas such as Afghanistan, Egypt and Pakistan. This research directly observed various health and behaviour parameters and was performed on 4,903 animals in the period of December 2002 and April 2003 (Pritchard et al., 2005). Another study by Regan et al. (2014) investigated whether gender, time of day or day of the week influenced the behaviour of working donkeys in order to develop an evidence-based working donkey ethogram (Regan et al., 2014). Regan et al., 2016) also investigated pain-related behaviour in working donkeys, intending to help owners and veterinarians to recognize such. The nonsteroidal anti-inflammatory drug meloxicam (Metacam) was hereby used (Regan et al., 2016). Until recently, there have been no studies describing a facial pain scale for donkeys. However, the University of Utrecht has carried out a study with donkeys and has developed a new facial pain scale: 'the Facial Expression Pain Scale for Donkeys' (van Dierendonck et al., submitted data).

2.3 Aim of this study

The aim of the current study was to assess inter- and intra- observer reliability in the assessment of facial characteristics that can be seen in horses and donkeys with acute pain.

Two newly developed pain-scoring systems were used in this research: The Horse Pain Face and The Donkey Pain Face. Both scales are based on the currently available pain scoring systems from facial characteristics. The objective was to determine whether these scales were reliable instruments for objective scoring using photos of horses and donkeys both with and without pain.

The photos and the results will be used in the future for a follow-up study on training computer systems for objective pain recognition from facial characteristics of photos. Using artificial intelligence techniques (deep learning and neural networks), computers will be trained to recognize these patterns.

The hypothesis is as follows: *'The horse and donkey pain scores are reliable for objective assessment of facial characteristics indicative of acute pain, with both good inter- and intra-observer reliability.'*

3 Materials and method

3.1 Horse and donkey photos collecting

Horse and donkey owners were asked to upload photos of the faces of their horses and donkeys. Dr T. van Loon advertised in different journals, on the Equine Pijn en Welzijns App (EPWA) website, and on social media in order to request equid owners to upload photos of their horses. Owners that uploaded their photos automatically gave permission to use their photos for this research. The owners only uploaded a photo and were not asked to provide any background information on their animal. The national resting place for horses, De Paardenkamp, in Soest was also asked to send horse photos. In March and April 2019, two students from the University of Utrecht went to the Donkey Sanctuary in Sidmouth, UK, to carry out research. The students were asked to send back photos of donkeys, that they collected during their time at the sanctuary. Dr T. van Loon also collected photos of horses and of donkeys during other research at Utrecht University. The number of photos collected of horses and donkeys is reported in **Table 1**.

Table 1.
Collected horse and donkey photos.

Horse	
EPWA Website	1320
Other research UU	354
De Paardenkamp	210
Total	**1884**
Donkey	
EWPA Website	13
Other research UU	23
The Donkey Sanctuary	535
Total	**571**

3.2 Photo selection

To increase validity, the 2455 photos were first checked for their usefulness. The photos had to meet conditions concerning quality, visibility of facial elements, and camera position. Selected images were numbered, cropped, and rotated using the program 'Photos'. Blurry photos, images in the wrong position, and duplicates were excluded.

On many of the photos, one or more facial elements were not visible. This was for example because the horse was wearing a bridle with a bit, preventing the corners of the mouth from being assessed, or the horse had its forelock in its eyes. Despite all Facial Action Units (FAUs) not being visible, it was decided to use these photos anyway as other facial elements could still be assessed.

The total study population consisted of a total of 1654 horses- and 534 donkey photos. The number of used photos is reported in **Table 2**.

Table 2
Useful horse and donkey photos.

Horse	
All FAUs visible	666
Not all FAUs visible	805
Horses with bit	183
Duplicate	-71
Not useful	-159
Useful total	**1654**
Donkey	
All FAUs visible	324
Not all FAUs visible	210
Duplicate	-25
Not useful	-12
Useful total	**534**

3.3 The Horse Pain Face and the Donkey Pain Face

The Horse Pain Face is based on the Horse Grimace Scale (**Figure 1**), EQUUS-FAP (**Table 3**) and Equine Pain Face (**Table 4** and **Figure 2a-c**). The Donkey Pain Face is based on the Facial Expression Pain Scale for Donkeys (**Table 5**), which is not yet published.

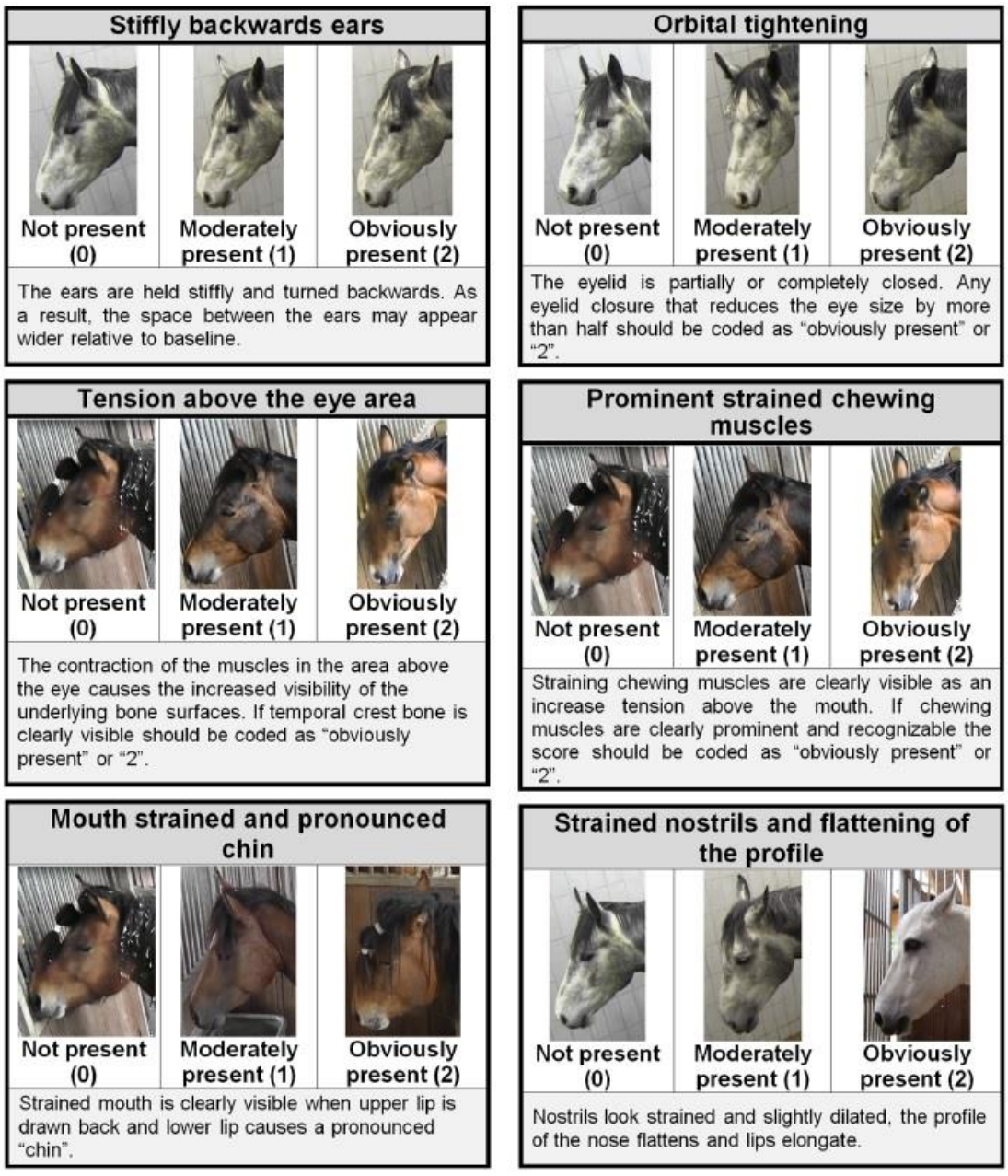


Figure 1. The Horse Grimace Scale (HGS).
HGS is based on six FAUs and is used in horses that have undergone surgical routine castration (Dalla Costa et al., 2014) and in horses with acute laminitis (Dalla Costa et al., 2016). The scores of these FAUs can range between 0 to 2. Whereas in this 3-point scale score 0 means no pain present, score 1 means pain is moderately present and score of 2 means pain is obviously present. Therefore, the maximum score varies between 0 and 12, in which 0 means no pain and 12 means maximal pain (Dalla Costa et al., 2014).

Table 3
The Equine University Utrecht Scale for Facial Assessment of Pain (EQUUS-FAP).
The EQUUS-FAP is used in horses with acute colic pain (van Loon and van Dierendonck, 2015), horses with head-related pain (van Loon and van Dierendonck, 2017) and in horses with acute orthopaedic pain (van Loon and van Dierendonck, 2019). The EQUUS-FAP is based on nine different facial expression characteristics. The nine facial expressions need to be observed and independently scored between 0 and 2. Therefore, the total pain score can range from 0 to 18. A score of 0 means there are no signals of pain. Whereas a score of 18 is the maximum (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
Descriptions of the Equine Pain Face.
The Equine Pain Face consists out of 6 different features and is used in horses with induced experimental pain (Gleerup et al., 2015).

Pain face feature	Detailed description
Asymmetrical/low ears	Both ears are moving in different directions or are placed in asymmetrical positions with neither of the ears facing directly forward or back. There may be lowering of both ears (increased distance between them) with the opening of the ears facing the sides or slightly back. The ears may be both asymmetrical and low.
Angled eye	There is tension of the *m. levator anguli oculi medialis* (Fig. 7).
Withdrawn and tense stare	The quality of the glance changes to become withdrawn and tense.
Nostrils – square-like	The nostrils are dilated mediolaterally; especially the medial wing of the nostril may be tense. This is most obvious during inspiration.
Tension of the muzzle	There is increased tonus of the lips and tension of the chin resulting in an edged shape of the muzzle.
Tension of the mimic muscles	There is tension of the muscles visible on the lateral aspect of the head, especially *m. zygomaticus* and *m. caninus*, but *m. masseter* may also be tense.

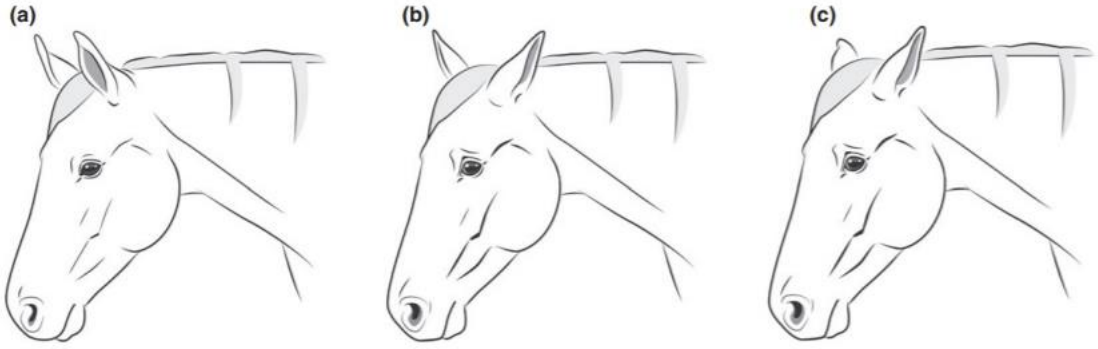


Figure 2a-c. The Equine Pain Face.
(a) Horse with a pain free and relaxed facial expression. (b) Horse in pain, showing all comprising features of a pain face including asymmetrical ears. (c) Horse in pain, showing all comprising features of a pain face including low ears (Gleerup et al., 2015).

Table 5
The Facial Expression Pain Scale for Donkeys.
The Pain Scale for Donkeys contains 12 FAUs and is used in donkeys with acute colic, acute orthopaedic pain and acute head-related pain and postoperative pain. The 12 FAUs of this pain scale can only be scored as either a 0 or 2, or can be scored between 0,1 or 2. The total pain score that contains all different elements could be situated between 0 and 24 in which 0 means no signs of pain and 24 means a maximal pain score (van Dierendonck et al., submitted data).

Head	Normal movement Less/no or more/ exaggerated movement	0 2
Eyelids	Opened More opened eyes or tightening of eyelids Obviously more opened eyes or obvious tightening of eyelids	0 1 2
Focus	Focused on environment Less focused on environment Not focused on environment	0 1 2
Nostrils	Relaxed A bit more opened, nostrils lifted, wrinkles seen Obviously more opened, nostril flaring, possibly audible breathing	0 1 2
Corners mouth/lips	Relaxed Lifted	0 2
Muscle tone head	No fasciculation's Mild fasciculation's Obvious fasciculation's	0 1 2
Flehming/yawning/smacking	Not seen Seen	0 2
Teeth grinding and/or moaning	Not been heard Heard	0 2
Ear response	Clear response with both ears or ear closest to source Delayed/reduced response to sounds No response to sounds	0 1 2
Ear position	Normal position Abnormal position (hang down/backwards)	0 2
Startle/headshaking	No startle/headshaking At least one startle (a sudden abrupt movement with the head as if suddenly aware of danger)/period of head shaking	0 2
Sweating behind the ears	No signs of sweating Signs of sweating	0 2
Total		/24

Although there are many other pain score scales for horses, those listed above are regularly used for live observations or video recording. To use these aforementioned scales, the assessors observed how frequently FAUs appeared during a set time period. These FAUs consisted of head movements, a focus on environment and other behaviour such as yawning and teeth grinding. Therefore, these dynamic FAUs could not be assessed from a photo. In fact, it was for this reason that a completely new pain scale was developed to score photos. The new pain scale was created by pieces, integrating FAUs that were thought to be more specific for the assessment of photos from the pain scales mentioned previously.

The newly created pain scales for horses and donkeys are similar and are based on six Facial Action Units: ears, orbital tightening, angulated upper eyelid, visibility of the sclera, corners of the mouth/lips, and nostrils (**Table 6** and **Table 7**).

Each of the six FAUs can receive a score between 0 to 2:

- A score of 0 indicates the absence of the FAU.
- A score of 1 indicates the moderate appearance of the FAU.
- A score of 2 indicates the obvious appearance of the FAU.

The total pain score can range from 0 (no signs of pain) to 10 (maximum pain). Normally, the maximum pain score is 12 (6 FAUs x 2). However, in this new pain scale there were FAUs related to one another. Thus, if you have a score of 2 for one FAU, another FAU must receive a score of 0. For example, if an animal shows a lot of sclera, its eyes must be wide open, resulting in an orbital tightening of 0. On the other hand, if an animal has orbital tightening score of 2, eyes closed, it is impossible to see the sclera, so the sclera would receive a score of 0.

Table 6.
The Horse Pain Face (HPF).

Ears	Both ears turned forwards At least one ear lateral position or further to backwards Both ears turned backwards	0 1 2
Orbital Tightening	Relaxed A bit tightening of the eyelids Obviously tightening of eyelids / eye closed	0 1 2
Angulated upper eyelid	Relaxed A bit more visible Obviously more visible	0 1 2
Visibility of the sclera	Sclera is not visible An edge of the sclera is visible Obviously more visible	0 1 2
Corners mouth/lips	Relaxed Lifted a bit Obviously lifted / strained	0 1 2
Nostrils	Relaxed A bit more opened Obviously more opened (dilated mediolaterally)	0 1 2
Total		/10

Table 7
The Donkey Pain Face (DPF).

Ears	Both ears turned forwards At least one ear lateral position or further to backwards Both ears turned backwards	0 1 2
Orbital Tightening	Relaxed A bit tightening of the eyelids Obviously tightening of eyelids / eye closed	0 1 2
Angulated upper eyelid	Relaxed A bit more visible Obviously more visible	0 1 2
Visibility of the sclera	Sclera is not visible An edge of the sclera is visible Obviously more visible	0 1 2
Corners mouth/lips	Relaxed Lifted a bit Obviously lifted / strained	0 1 2
Nostrils	Relaxed A bit more opened Obviously more opened (dilated mediolaterally)	0 1 2
Total		/10

3.4 Experimental design

The assessors were three trained observers. Two of them were both veterinary students with variable levels of equine experience. One first year veterinary student, with little equine experience and one first year master student, who had several years of experience with horses. They performed their evaluations simultaneously, but did not discuss their findings. The third observer was an experienced pain-behaviour researcher, who had helped to develop this research.

Before beginning this study, the two student observers had the opportunity to familiarize themselves with the Horse Pain Face and the Donkey Pain Face. They underwent a training session of approximately six days. The first 5 days of training were used to familiarize themselves with photos of horses. The purpose of the second training period of one day was to become familiar with photos of donkeys. Two instruction manuals (**Attachment 1** and **2**) were developed for the training programmes by a researcher in this current study. The manuals, one for the horse and one for the donkey, each consists of 54 photos that describe the different FAUs and their scores.

A total of 2626 photos were assigned to each assessor for evaluation in a specific order. Each photo was assessed separately by the two student observers, using a Microsoft Excel table with the new pain scales to document the presence or absence of an FAU as well as the colour of the animal. The tables used to record the pain scale scores per photo is found in **Attachment 3** and **4**. As an example, a scored horse photo and a donkey photo are shown in **Attachment 5** and **6**.

After assessment of all of the horse photos (1654) and all of the donkey photos (534) 20% of each set were randomly chosen and re-scored to assess intra-observer reliability. The random set of photos was chosen by selecting every first and tenth photo from the original photo set, resulting in an intra-observer set of 331 horse photos and 107 donkey photos. Each assessment took on average one minute per photo. The photographs were available in electronic copies (which could be enlarged). Pain scoring was performed for five weeks, between 9.00 am and 4:00 pm.

Notice, the third observer scored 331 horse photos and 107 donkey photos, that were scored twice by the two student observers. However, the third observer did not have the opportunity to train and score together with the other two observers.

3.5 Data processing and statistical analysis

Inter-observer reliability and the intra-observer reliability were examined on two pain scoring systems: the Horse Pain Face (HPF) and the Donkey Pain Face (DPF) scores between the two participating observers. Inter-observer reliability and the intra-observer reliability were assessed using Intra-class Correlation Coefficient analysis with Cronbach's Alpha. The scatter plots were used to visually analyze the large amount of data presented to determine if there was a positive relationship between two observations and the Bland-Altman plots were used to visually evaluate correlations and determine bias and limits of agreement. The total FAU scores were used in these analyses as well as the individual FAU scores from the two participating observers.

Further analysis was made of the two pain scoring scales when examining scores between three observers scoring 331 horse and 107 donkey photos, that were used to assess intra-observer reliably. These scores underwent the same analysis and processing as the scores described above.

The program IBM SPSS Statistics 26 was used for all of the statistical analysis. The statistical significance was accepted at $P < 0,05$. Microsoft Excel was used to create all graphs, scatter plots and Bland-Altman plots.

4 Results

4.1 Inter-observer reliability for photos of horses

Table 8 and **Fig. 3** show the results of correlation analyses of the Horse Pain Face between two independent observers, observer 1 and 2. There was an excellent and significant correlation (Cronbach's Alpha = 0.92, P < 0.001). The Bland Altman analysis yielded a bias of -0.03 and the limits of agreement of -1.33 and +1.28 (n=1654).

Table 8
Inter-observer reliability analysis of the Horse Pain Face.
The Cronbach's Alpha, average scores of Intra-class Correlation Coefficient (ICC), with the 95%-confidence interval reported between brackets and the P-value, are presented for each FAU and the total FAUs score (n=1654).

Facial Action Unit	Cronbach's Alpha	ICC (95% confidence interval)	P-value
Ears	0.96	0.96 (0.95-0.96)	< 0.001
Orbital Tightening	0.92	0.92 (0.91-0.93)	< 0.001
Angulated upper eyelid	0.80	0.79 (0.77-0.81)	< 0.001
Visibility of the sclera	0.94	0.94 (0.93-0.94)	< 0.001
Corners mouth/lips	0.53	0.52 (0.45-0.59)	< 0.001
Nostrils	0.80	0.80 (0.78-0.82)	< 0.001
Total FAUs scores	0.92	0.92 (0.92-0.93)	< 0.001

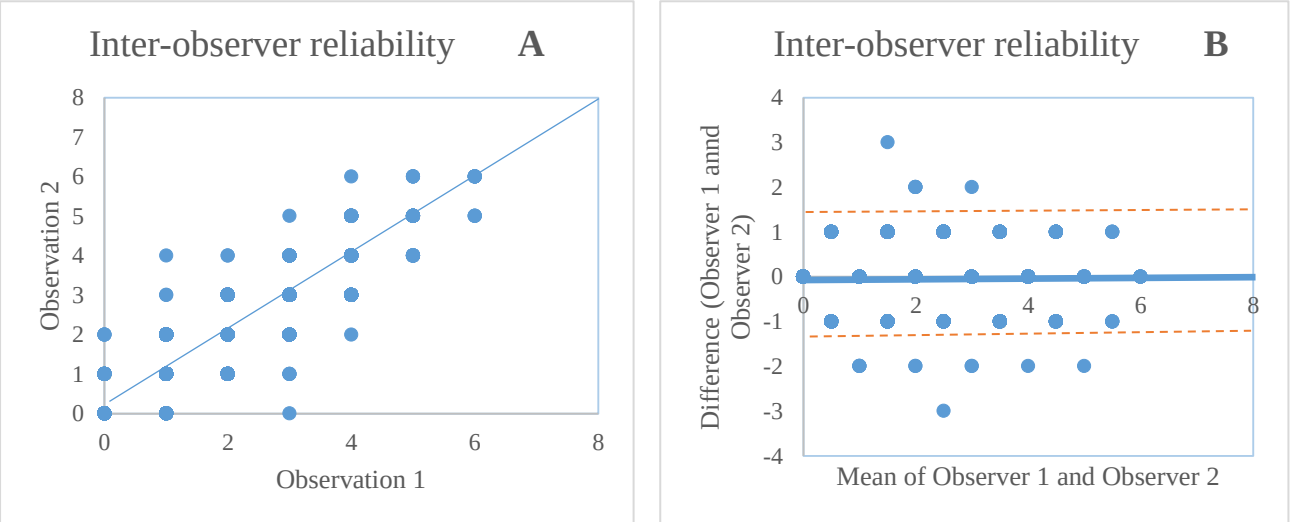


Fig. 3. Inter-observer reliability.
(A) Scatter plot of the Horse Pain Face scores simultaneously assessed by two independent observers (Cronbach's Alpha = 0.92 and ICC= 0.92 (P<0.001)). (B) Bland-Altman plots of the Horse Pain Face presenting bias = -0.03 (Solid line) and limits of agreement between -1.33 and +1.28 (Dashed lines) (n=1654; NB many points overlap).

4.2 Intra-observer reliability for photos of horses

Table 9 and **Fig. 4** show the results of correlation analyses of the Horse Pain Face of observer 1 scores. There was a good and significant correlation (Cronbach's Alpha = 0.81, P < 0.001). The Bland Altman analysis yielded a bias of -0.02 and the limits of agreement of -2.01 and +1.98 (n=331). **Table 10** and **Fig. 5** show the results of correlation analyses of the Horse Pain Face of observer 2 scores. There was also a good and significant correlation (Cronbach's Alpha = 0.88, P < 0.001). The Bland Altman analysis yielded a bias of -0.01 and the limits of agreement of -1.68 and +1.67 (n=331).

Table 9
Intra-observer reliability analysis of the Horse Pain Face by observer 1.
The Cronbach's Alpha, average scores of Intra-class Correlation Coefficient (ICC), with the 95%-confidence interval reported between brackets and the P-value, are presented for each FAU and the total FAUs score (n=331).

Facial Action Unit	Cronbach's Alpha	ICC (95% confidence interval)	P-value
Ears	0.88	0.88 (0.88-0.90)	< 0.001
Orbital Tightening	0.85	0.85 (0.82-0.88)	< 0.001
Angulated upper eyelid	0.81	0.80 (0.75-0.85)	< 0.001
Visibility of the sclera	0.90	0.90 (0.88-0.92)	< 0.001
Corners mouth/lips	0.49	0.46 (0.26-0.61)	< 0.001
Nostrils	0.79	0.79 (0.74-0.83)	< 0.001
Total FAUs scores	0.81	0.81 (0.77-0.85)	< 0.001

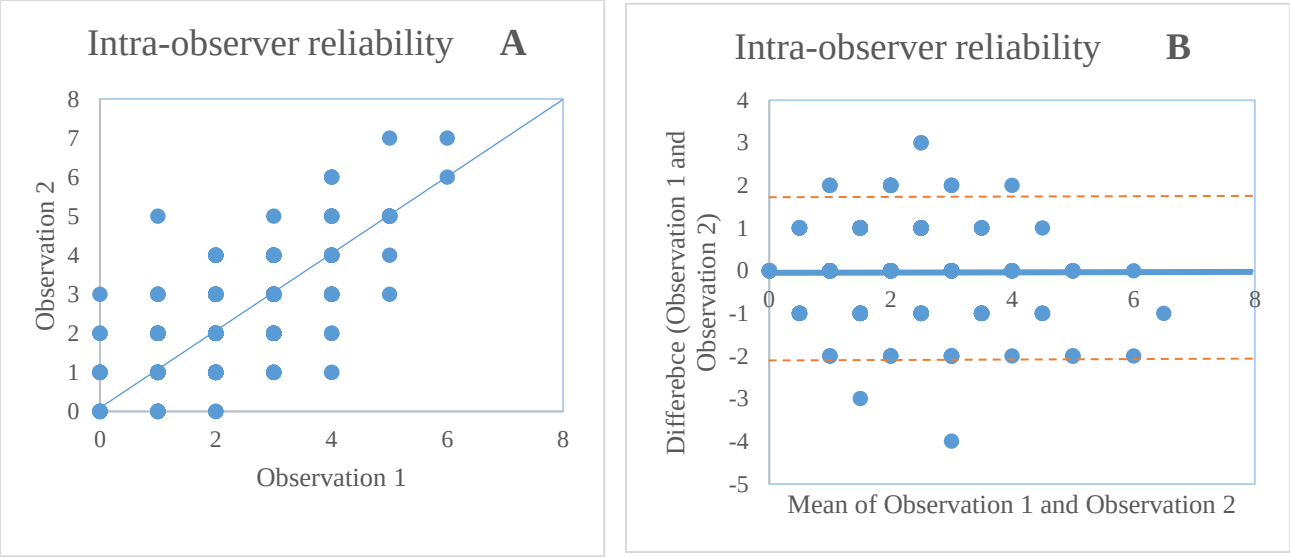


Fig. 4 Intra-observer reliability.
(A) Scatter plot of the Horse Pain Face scores assessed by observer 1 (Cronbach's Alpha = 0.81 and ICC= 0.81 (P<0.001)). (B) Bland-Altman plots of the Horse Pain Face presenting bias = -0.02 (Solid line) and limits of agreement between -2.01 and +1.98 (Dashed lines) (n=331; NB many points overlap).

Table 10
Intra-observer reliability analysis of the Horse Pain Face by observer 2.The Cronbach's Alpha, average scores of Intra-class Correlation Coefficient (ICC), with the 95%-confidence interval reported between brackets and the P-value, are presented for each FAU and the total FAUs score (n=331).

Facial Action Unit	Cronbach's Alpha	ICC (95% confidence interval)	P-value
Ears	0.97	0.97 (0.96-0.97)	< 0.001
Orbital Tightening	0.92	0.92 (0.90-0.93)	< 0.001
Angulated Upper Eyelid	0.81	0.81 (0.75-0.85)	< 0.001
Visibility of the Sclera	0.90	0.89 (0.87-0.92)	< 0.001
Corners mouth/lips	0.71	0.71 (0.60-0.79)	< 0.001
Nostrils	0.80	0.80 (0.75-0.84)	< 0.001
Total FAUs scores	0.88	0.85 (0.85-0.90)	< 0.001

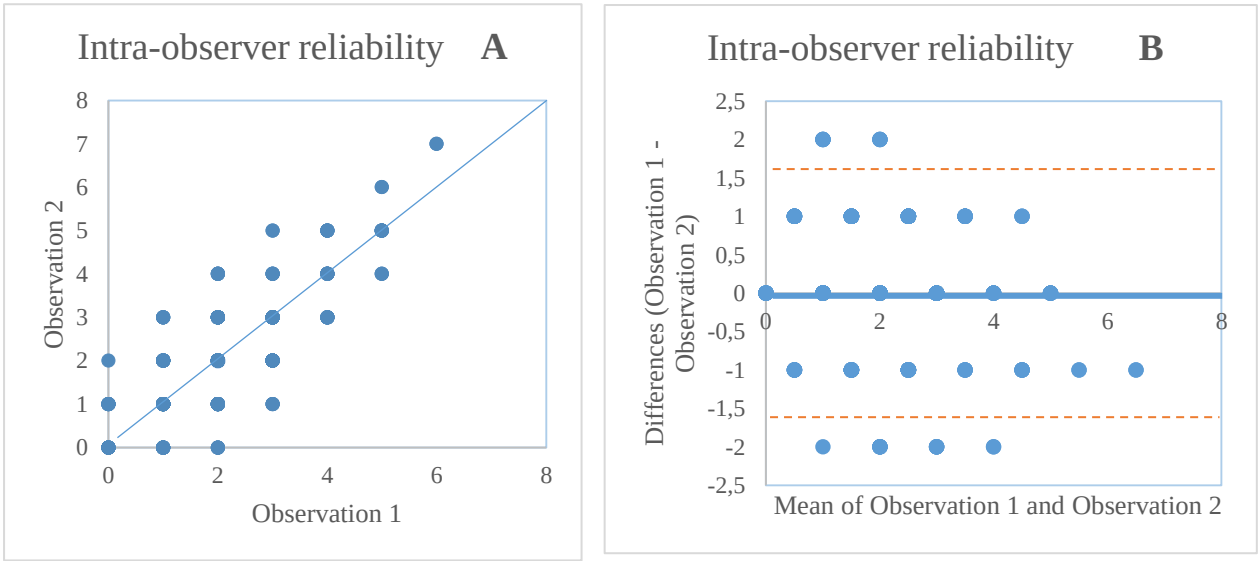


Fig. 5. Intra-observer reliability.
(A) Scatter plot of the Horse Pain Face scores assessed by observer 2 (Cronbach's Alpha = 0.88 and ICC= 0.85 (P<0.001)). (B) Bland-Altman plots of the Horse Pain Face presenting bias = -0.01 (Solid line) and limits of agreement between -1.68 and +1.67 (Dashed lines) (n=331; NB many points overlap).

4.3 Inter-observer reliability for photos of horses between all three observers

Table 11 show the results of correlation analyses of the Horse Pain Face between three independent observers. There was an excellent and significant correlation (Cronbach's Alpha = 0.91, P < 0.001).

Table 11
Inter-observer reliability for three observer's analysis of the Horse Pain Face.
The Cronbach's Alpha, average scores of Intra-class Correlation Coefficient (ICC), with the 95%-confidence interval reported between brackets and the P-value, are presented for each FAU and the total FAUs score (n=331).

Facial Action Unit	Cronbach's Alpha	Observers 1 and 3	Observers 1 and 2	Observers 2 and 3	ICC (95% confidence interval	P-value
Ears	0.97	0.88	0.94	0.93	0.97 (0.96-0.98)	< 0.001
Orbital Tightening	0.77	0.48	0.87	0.51	0.74 (0.65-0.80)	< 0.001
Angulated Upper eyelid	0.86	0.63	0.68	0.71	0.86 (0.82-0.88)	< 0.001
Visibility of the sclera	0.94	0.76	0.96	0.76	0.94 (0.92-0.95)	< 0.001
Corners mouth/lips	0.67	0.37	0.43	0.43	0.66 (0.55-0.74)	< 0.001
Nostrils	0.86	0.61	0.73	0.67	0.85 (0.80-0.88)	< 0.001
Total FAUs score	0.91	0.71	0.87	0.70	0.89 (0.86-0.92)	< 0.001

4.4 Inter-observer reliability for photos of donkeys

Table 12 and **Fig. 6** show the results of correlation analyses of the Donkey Pain Face between two independent observers, observer 1 and 2. There was an acceptable and significant correlation (Cronbach's Alpha = 0.79, P < 0.001). The Bland Altman analysis yielded a bias of -0.09 and the limits of agreement of -3.63 and +1.93 (n=534).

Table 12
Inter-observer reliability analysis of the Donkey Pain Face.
The Cronbach's Alpha, average scores of Intra-class Correlation Coefficient (ICC), with the 95%-confidence interval reported between brackets and the P-value, are presented for each FAU and the total FAUs score (n=534).

Facial Action Unit	Cronbach's Alpha	ICC (95% confidence interval)	P-value
Ears	0.95	0.95 (0.94-0.96)	< 0.001
Orbital Tightening	0.86	0.86 (0.84-0.89)	< 0.001
Angulated upper eyelid	0.70	0.68 (0.56-0.76)	< 0.001
Visibility of the sclera	0.91	0.91 (0.89-0.99)	< 0.001
Corners mouth/lips	0.68	0.68 (0.60-0.74)	< 0.001
Nostrils	0.57	0.52 (0.35-0.64)	< 0.001
Total FAUs score	0.79	0.74 (0.51-0.84)	< 0.001

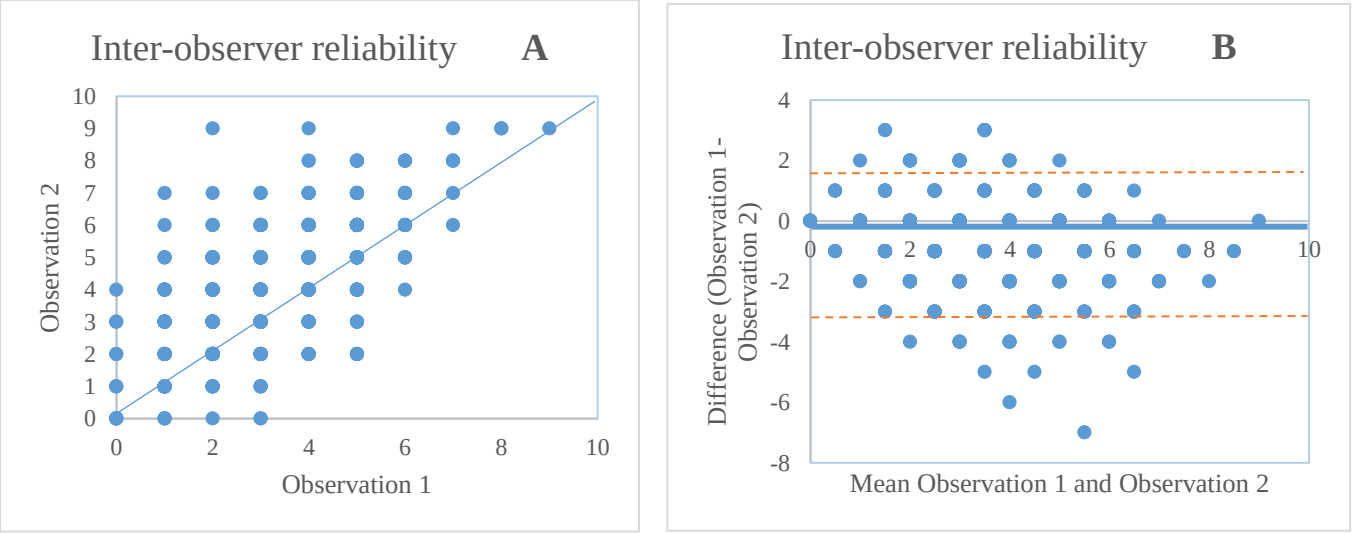


Fig. 6. Inter-observer reliability.
(A) Scatter plot of the Donkey Pain Face scores simultaneously assessed by two independent observers (Cronbach's Alpha = 0.79 and ICC= 0.74 (P<0.001)). (B) Bland-Altman plots of the Donkey Pain Face presenting bias = -0.09 (Solid line) and limits of agreement between -3.63 and +1.93 (Dashed lines) (n=534; NB many points overlap).

4.5 Intra-observer reliability for photos of donkeys

Table 13 and **Fig. 7** show the results of correlation analyses of the Donkey Pain Face for observer 1 scores. There was a good and significant correlation (Cronbach's Alpha = 0.84, P < 0.001). The Bland Altman analysis yielded a bias of 0.77 and the limits of agreement of -1.92 and +3.45 (n=107). **Table 14** and **Fig. 8** show the results of correlation analyses of the Donkey Pain Face for observer 2 scores. There was an excellent and significant correlation (Cronbach's Alpha = 0.97, P < 0.001). The Bland Altman analysis yielded a bias of -0.01 and the limits of agreement of -1.10 and +1.08 (n=107).

Table 13
Intra-observer reliability analysis of the Donkey Pain Face for observer 1's scores.
The Cronbach's Alpha, average scores of Intra-class Correlation Coefficient (ICC), with the 95%-confidence interval reported between brackets and the P-value, are presented for each FAU and the total FAUs score (n=107).

Facial Action Unit	Cronbach's Alpha	ICC (95% confidence interval)	P-value
Ears	0.97	0.97 (0.96-0.98)	< 0.001
Orbital Tightening	0.75	0.71 (0.47-0.82)	< 0.001
Angulated upper eyelid	0.72	0.72 (0.57-0.82)	< 0.001
Visibility of the sclera	0.91	0.91 (0.86-0.94)	< 0.001
Corners mouth/lips	0.76	0.75 (0.62-0.84)	< 0.001
Nostrils	0.55	0.53 (0.32-0.68)	< 0.001
Total FAUs score	0.84	0.78 (0.60-0.89)	< 0.001

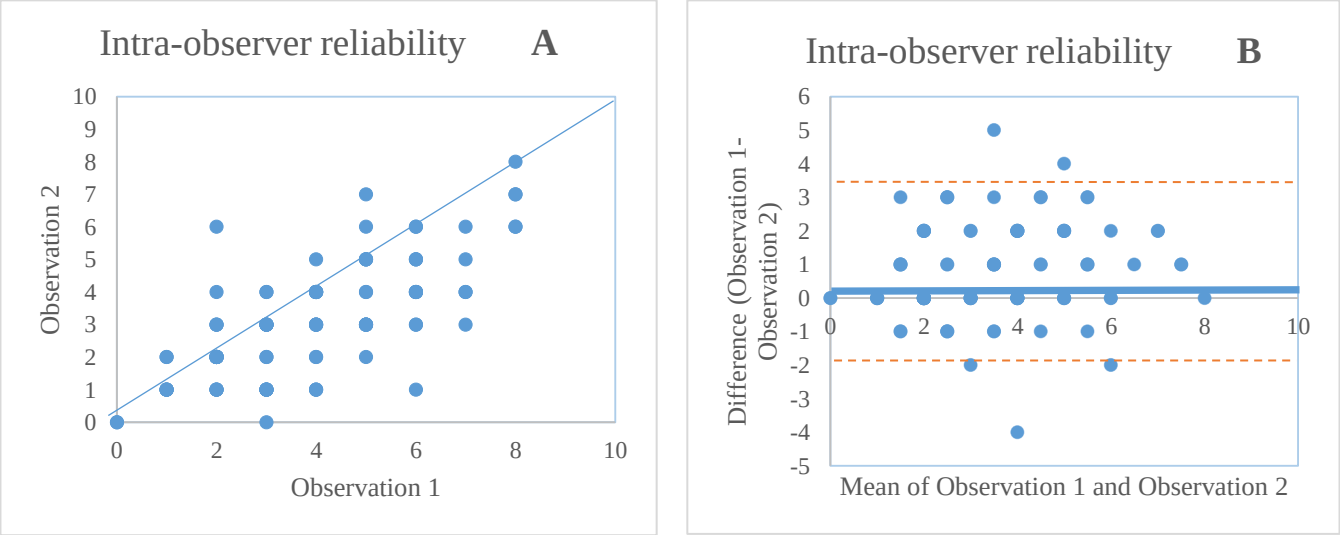


Fig. 7. Intra-observer reliability.
(A) Scatter plot of the Donkey Pain Face scores assessed by observer 1 (Cronbach's Alpha = 0.84 and ICC= 0.78 (P<0.001)). (B) Bland-Altman plots of the Donkey Pain Face presenting bias = 0.77 (Solid line) and limits of agreement between -1.92 and +3.45 (Dashed lines) (n=107; NB many points overlap).

Table 14
Intra-observer reliability analysis of the Donkey Pain Face for observer 2's scores.
The Cronbach's Alpha, average scores of Intra-class Correlation Coefficient (ICC), with the 95%-confidence interval reported between brackets and the P-value, are presented for each FAU and the total FAUs score (n=107).

Facial Action Unit	Cronbach's Alpha	ICC (95% confidence interval)	P-value
Ears	0.99	0.99 (0.99-1.00)	< 0.001
Orbital Tightening	0.83	0.82 (0.71-0.88)	< 0.001
Angulated upper eyelid	0.83	0.83 (0.70-0.90)	< 0.001
Visibility of the sclera	0.98	0.98 (0.97-0.99)	< 0.001
Corners mouth/lips	0.87	0.88 (0.80-0.92)	< 0.001
Nostrils	0.82	0.82 (0.74-0.88)	< 0.001
Total FAUs score	0.97	0.97 (0.95-0.98)	< 0.001

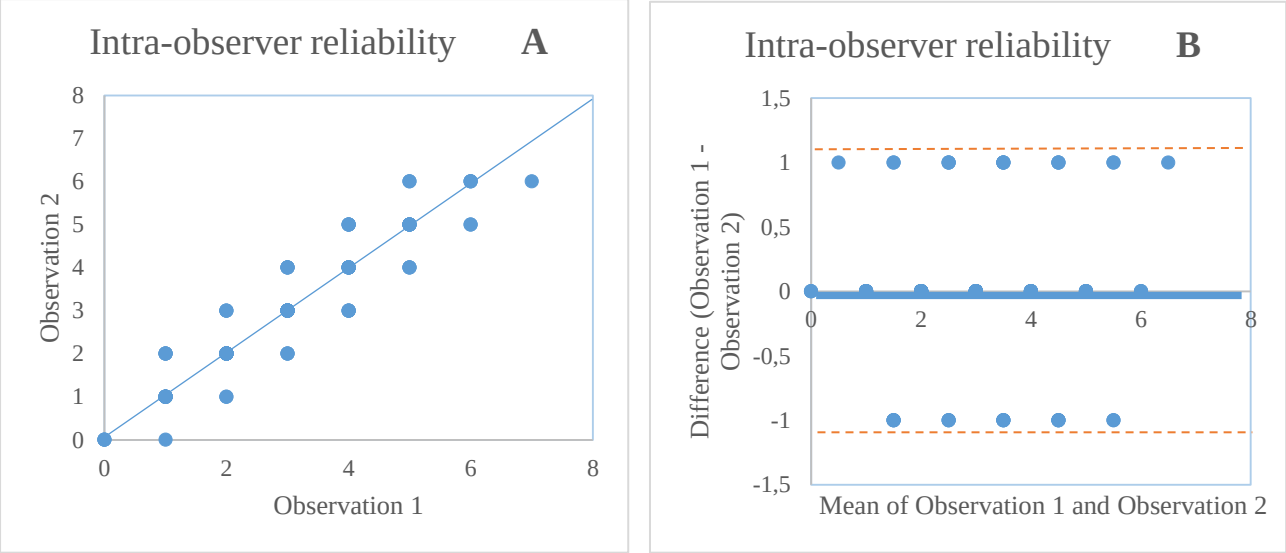


Fig. 8. Intra-observer reliability.
(A) Scatter plot of the Donkey Pain Face scores assessed by observer 2 (Cronbach's Alpha = 0.97 and ICC= 0.97 (P<0.001)). (B) Bland-Altman plots of the Donkey Pain Face presenting bias = -0.01 (Solid line) and limits of agreement between -1.10 and +1.08 (Dashed lines) (n=107; NB many points overlap).

4.6 Inter-observer reliability for photos of donkeys between all three observers

Table 15 show the results of correlation analyses of the Donkey Pain Face between three independent observers. There was an excellent and significant correlation (Cronbach's Alpha = 0.93, P < 0.001).

Table 15
Inter-observer reliability for three observers analysis of the Donkey Pain Face.
The Cronbach's Alpha, average scores of Intra-class Correlation Coefficient (ICC), with the 95%-confidence interval reported between brackets and the P-value, are presented for each FAU and the total FAUs score (n=107).

Facial Action Unit	Cronbach's Alpha	Observers 1 and 3	Observers 1 and 2	Observers 2 and 3	ICC (95% confidence interval)	P-value
Ears	0.97	0.90	0.92	0.91	0.97 (0.95-0.98)	< 0.001
Orbital Tightening	0.85	0.64	0.69	0.62	0.84 (0.77-0.89)	< 0.001
Angulated Upper eyelid	0.76	0.49	0.56	0.54	0.71 (0.54-0.83)	< 0.001
Visibility of the sclera	0.91	0.67	0.92	0.75	0.91 (0.87-0.94)	< 0.001
Corners mouth/lips	0.68	0.37	0.49	0.41	0.66 (0.49-0.77)	< 0.001
Nostrils	0.79	0.49	0.60	0.57	0.79 (0.70-0.85)	< 0.001
Total FAUs score	0.93	0.74	0.84	0.77	0.91 (0.88-0.94)	< 0.001

5 Discussion

5.1 Horse Pain Face and Donkey Pain Face

This study demonstrated that the horse and donkey pain scores are reliable for objective assessment of facial characteristics indicative of acute pain. Both scales, the Horse Pain Face (HPF) and the Donkey Pain Face (DPF) were used to assess pain from photos of horses and donkeys. The photos were uploaded without giving any background information. It should be noted, that while no information was given on specific animals, all donkey photos came from the Donkey Sanctuary in the UK. Therefore, observers knew that the donkeys had medical conditions, but were unaware of the health status of the individual donkeys. The Horse Pain Face was created using multiple published works, the Donkey Pain Face was based on one scale, Equine Utrecht University Scale for Donkey Facial Assessment of Pain (EQUUS-Donkey-FAP), which is not yet published. Both the HPF and DPF were useful for the purpose of photo assessment because both scales yielded reliable and reproducible results, by observers that were trained to use these scales.

The findings of the results will be used for two applications. One for the development of a training tool to train observers by providing feedback on scored photos in a smart phone application. And another for the development of an automated pain recognition tool, based on analysis of facial characteristics of photos of horses and donkeys, based on deep learning techniques and neural networks. This will make the application easier to interpret and more user friendly. To make this goal a reality, it is necessary to determine if prior training would have an influence on scoring by teaching individuals how to consistently and correctly score a pain face. Notice that, there will be necessary a larger amount of photos to train computers to recognize the facial expressions. An Automatic Pain Facial Expression Detection System for Sheep has already been developed (McLennan and Mahmoud, 2019). This system automatically detects pain in sheep. It is used by producers to increase the chances of controlling spread and prevention of the diseases. (Lu et al., 2017; McLennan and Mahmoud, 2019).

5.2 Training for horse and donkey photos

Theoretically, training would make it possible for individuals with different level of experience to give the same pain score, making the scoring system more observer independent. In order to test the value of training individuals for scoring, two observers simultaneously underwent training and afterward scored 1654 horse photos and 571 donkey photos. The observers had five training days dedicated to the Horse Pain Face and one training day for the Donkey Pain Face. Additionally, while both observers were veterinary students, one observer had several years of experience with horses while the other student had very little equine experience. Therefore, the resulting data showed that the training of two individual observers with vastly different experience can produce pain scores with good inter- and intra-observer reliability. In a comparison study it would be interesting to change the sequence of the research. Thus, the observers start with training and scoring of donkey photos and then start with the horse. This is because evaluation of about 2,000 horse photos gave the observers a lot of experience scoring an equine face before they even trained for scoring donkey photos. This experience could have altered the results for the donkey photos, mostly because the HPF and DPF are very similar

pain scoring systems. In order to ensure that the donkey training is just as successful, it would be valuable to reverse the order of the study.

5.3 Inter-observer reliability for photos of horses

When analyzing the inter-observer and intra-observer reliability, it was interesting to analyze the total Facial Action Unit (FAU) scores as well as the individual FAU scores because the total FAU score is not always representative for all the individual FAU scores. For example, if a horse has a total pain score of 4 by both observers, the score could be coming from different FAUs. It is therefore necessary to also analyze the individual FAU in order to not only see patterns in the pain score, but also identify where observers had difficulty scoring their equine subjects. In addition, looking at the data from individual FAUs can provide feedback for future pain scale training. If a particular FAU has a low consistency then perhaps the training of that particular FAU as well as the definition of the scoring categories of the FAU should be reassessed in order to determine both where the confusion is coming from and if there is a better way to define the parameters for future projects.

A good inter-observer reliability was found in the literature for facial pain scales for horses with live observation. In a study by van Loon and van Dierendonck, (2015) an excellent inter-observer reliability for the Equine Utrecht University Scale for Facial Assessment of Pain (EQUUS-FAP) in horses with acute colic was found (Cronbach's Alpha = 0.84) with limits of agreements of -2.2 to +2.8. Similarly, in another study by van Loon and van Dierendonck, (2017) an excellent inter-observer reliability for the EQUUS-FAP in horses with acute and postoperative head-related pain, including dental pain, ocular pain and trauma to the skull was found (ICC = 0.92). Most recently, van Loon and van Dierendonck, (2019) also found an excellent inter-observer reliability for the EQUUS-FAP in horses with orthopaedic trauma and after orthopaedic surgery (Cronbach's Alpha= 0.93), with low bias (-0.08) and a narrow range of limits of agreements (between -1.9 and +1.9). Another study, (Dalla Costa et al., 2016), used videos and photos for assessment of pain with the Horse Grimace Scale (HGS) in horses with acute laminitis. The inter-observer reliability was good for both the observations (ICC =0.95 for video and ICC = 0.85 for photos). In a different study, the Horse Grimace Scale for equines undergoing castration demonstrated high inter-observer reliability with an overall ICC value of 0.92 (Dalla Costa et al., 2014). The pain scales used in these studies were the basis for the one created for the purpose of this research. As these scales all have high Cronbach's Alpha and ICC values, it can be said that these scales have indeed been valid for reproducible scoring of pain in equines. Notice that Gleerup et al., (2015) investigated the existence of an equine pain face, however the inter- and intra-observer reliability was not described. In the current study, the inter-observer reliability for the total scores of the horse photos was also excellent (Cronbach's Alpha = 0.92) with a low bias (-0.03) and a narrow range of limits of agreements (between -1.33 and +1.28).

The FAU 'Ears' showed the highest inter-observer reliability (Cronbach Alpha = 0.96). Similarly, in previous studies by Dalla Costa et al., (2014) and Dalla Costa et al., (2016) 'Ears' was found to have the highest inter-observer reliability (ICC = 0.95 ; ICC = 0.97 respectively). This shows that the 'Ears' are a relatively easy FAU to score, leading to high accuracy and precision. A reason for the high inter-observer reliability of the 'Ears' FAU is most likely

because of the vast difference in scoring categories. It is difficult to incorrectly score the 'Ears' because they are either forward, lateral, or backwards. Alternatively, in this study 'Corners mouth/lips' was found to have the lowest inter-observer reliability (Cronbach Alpha = 0.53). This FAU is more open to interpretation, leading to a lower inter-observer reliability. The difficulty of scoring this FAU was noted by observers after the scoring process and observers hypothesized that this FAU would be the most inconsistent for the horses due to the slight changes in the angle of the mouth. The mouth is already difficult to score because of the (vague) criteria for scoring, coupled with other hindrances such as bad camera angles. Notice, the FAU 'Corners mouth/lip' was not scored on a photo if the horse was wearing a bridle with bit.

5.4 Intra-observer reliability for photos of horses

The two observers re-scored 331 photos of horses to assess intra-observer reliability. Both observers yielded a good and significant intra-observer correlation (Observer 1: Cronbach's Alpha = 0.81; $P < 0.001$) (Observer 2: Cronbach's Alpha = 0.88; $P < 0.001$). These values indicate that the observers were able to duplicate their scores with high accuracy in relation to their previous scores. The highest intra-observer reliability for observer 1 was the 'Visibility of the sclera' (Cronbach's Alpha = 0.90) and the lowest intra-observer reliability for observer 1 was the 'Corners mouth/lips' (Cronbach's Alpha = 0.49). Meanwhile, the highest intra-observer reliability for observer 2 was the 'Ears' (Cronbach's Alpha = 0.97) and the lowest intra-observer reliability was also the 'Corners mouth/lips' (Cronbach's Alpha = 0.71).
Intra-observer reliability was higher for observer 2 compared to observer 1. This could be explained by the relative horse experience of each observer. As stated previously, one observer had a lot of experience with horses and thus could easily detect subtle nuances, while the other observer had relatively little experience and therefore could not as easily categorize the FAUs into the different scoring categories. In a follow up study, it would be interesting to have more example photos in the training manual and a gold standard for scoring the 'Corners mouth/ lips' FAU. Having these tools may lead to less error and variability when scoring.

5.5 Influence of different levels of experience between all three observers assessing photos of horses

The dataset of photos (n=331), that were used to analyze intra-observer reliability, was also scored by a third observer. This third observer was an experienced veterinarian, already familiar with the pain scoring systems. This gave us the opportunity to assess inter-observer reliability among three observers with all three different levels of expertise as well.
It was expected that the 'Corners mouth/lips' would yield a low consistency because of the ambiguity in scoring this FAU. However, it was unexpected that the 'Orbital tightening' would be so inconsistent because 'Orbital tightening' is considered to be a relatively clear FAU. The consistency between observers 1 and 2 is high (Cronbach's Alpha = 0.87). In sharp contrast, the consistency between observer 3 and the other two observers is low (Cronbach's Alpha = 0.48 and 0.51 respectively). This unexpected outcome could be due to the difference in training between observer 3 and observers 1 and 2. If this study were to be replicated, it would be

interesting to see the results when all observers, with the same equine experience, train and score the entire photo set together.

5.6 Inter-observer reliability for photos of donkeys

At this moment there is no published literature available to compare the inter- and intra-observer reliability with other studies. Unlike horses, there is no previous facial pain scale published for donkeys yet. Therefore, this study will only compare the results with one unpublished study by the University of Utrecht.
The unpublished study presents the construction and testing of two pain scales, the Equine Utrecht University Scale for Donkey Composite Pain Assessment and for Donkey Facial Assessment of Pain (EQUUS-DONKEY-COMPASS and EQUUS-DONKEY-FAP), in live observations of donkeys with acute pain. The observers were two trained master veterinary students from the Utrecht University. Both pain scales showed excellent and significant correlation (Cronbach's alpha = 0.97, $P < 0.001$ for EQUUS-DONKEY-COMPASS, Cronbach's alpha = 0.94, $P< 0.001$ for EQUUS-DONKEY-FAP) (van Dierendonck et al., submitted data). In comparison, this study's inter-observer reliability for the assessment of donkey photos is acceptable and significant (Cronbach's Alpha = 0.79, $P < 0.001$ for Donkey Pain Face). Both the current study and study on acute pain in donkeys have yielded high values for Cronbach's Alpha and show significant correlations. Individual FAUs in this study cannot be compared to the unpublished study because the unpublished study did not show results for each FAU, only the total of all the FAUs. Despite not being able to compare and contrast these individual FAUs, the discussion for this research will be focused on the highest and lowest Cronbach's Alpha for the individual FAUs of the DPF. 'Ears' was found to have the highest inter-observer reliability (Cronbach Alpha = 0.95). This can also be seen in the analysis of scored horse photos and is most likely due to reasons described previously. 'Nostrils' was found to have the lowest inter-observer reliability (Cronbach Alpha = 0.57). There are multiple factors that could have played a role in the low inter-observer reliability of this facial action unit. Certain breeds of donkeys have larger nostrils than others, so a slightly flared nostril on one breed of donkey can look like a completely closed nostril on another breed, making it difficult to distinguish between scoring categories. Additionally, donkeys with darker coat colour tend to have black hair around their noses. The hair is so dark that it is nearly impossible to tell where the opening of the nose ends, and the black muzzle begins. Furthermore, donkeys tend to be shorter than the average human. Therefore, when the photo is taken, the angle of the camera is generally pointing down. All of these factors contribute to the nostrils being difficult to assess.

Currently, there is not a lot of information on facial expressions in donkeys in general. In a follow-up study, it would be interesting for researchers to look at the Donkey Pain Face when a donkey is experiencing either acute pain or chronic pain. Similar research has already been completed with horses; however, no such published research exists for donkeys.

5.7 Intra-observer reliability for photos of donkeys

To assess intra-observer reliability, the two observers re-scored 107 donkey photos. The intra-observer reliability of observer 1 was good and significant (Cronbach's Alpha = 0.84; P<0.001). Meanwhile, the intra-observer reliability of observer 2 was excellent and significant (Cronbach's Alpha = 0.97, P < 0.001). The highest intra-observer reliability for both observers was the 'Ears' (Cronbach's Alpha = 0.97 and 0.99 respectively). The lowest intra-observer reliability for both observers was the 'Nostrils' (Cronbach's Alpha = 0.55 and 0.85 respectively). These values indicate that the observers were able to re-score the photos with scores that were nearly identical to their previous scores. The reasons for these values have been previously described. However, the difference between both observers was most likely due to the observer's familiarity with equines.

5.8 Influence of different level of experience between all three observers assessing photos of donkeys

As stated previously, observer 3 did not undergo the same training or score the photos in the same timespan that observers 1 and 2 did. Besides this, observer 2 developed the training manuals and therefore had more time to familiarize with the different FAUs. Consequently, observer 2 has more training than observer 1. This difference in procedure and different level in experience with donkeys may have impacted the results. The highest Cronbach's Alpha between the three observers was found to be the 'Ears' FAU (Cronbach's Alpha = 0.97; ICC = 0.97). Surprisingly, the 'Corners mouth/lips' FAU had the lowest score (Cronbach's Alpha = 0.68; ICC = 0.66). Before analyzing the data, observers hypothesized that the 'Corners mouth/lips' FAU in the donkey would have a higher score than that of the horse (Cronbach's Alpha = 0.67) because of the more distinct definitions for each donkey 'Corners mouth/lips' scoring category. However, the results showed that the 'Corner mouth/lips' for donkeys is much lower than anticipated. A reason for this may be that observers had one training day for learning how to properly score donkey photos, whereas they previously had five days to learn how to score a horse pain face. One training day may not have been enough time to fully learn how to see all of the nuances of the donkey 'Corners mouth/lips' FAU. The 'Nostrils' FAU for the inter-observer reliability between the three observers was also significantly lower than the other FAUs (Cronbach's Alpha = 0.79). Observers expected this outcome for reasons described previously. It is noticeable that despite intense training there are still slight differences between the three observers. The consistency between observers 1 and 2 is higher (Cronbach's Alpha = 0.84) than the consistency between observer 3 and the other two observers (Cronbach's Alpha = 0.74 and 0.77 respectively). This outcome could indicate the importance of training, because it is imaginable how precise or not precise the agreement would be, if a veterinarian or a donkey owner without training starts working with the Donkey Pain Face. In a follow up study, it would be interesting to have all three observers score the entire photo set. A larger data set could lead to different inter- and intra-observer reliability and provide a more accurate score, better showing whether or not the training for the donkeys is successful.

5.9 Comparison of horse and donkey

For both species, the scores of the inter- and intra-observer reliability had some slight discrepancies. A factor that might have played a role is the difference in the visibility of FAUs and definition of the FAUs scoring categories. For example, observers expected the 'Nostrils' FAU to be easy to score for the donkey photos because of prior experience with scoring horse photos. However, the 'Nostrils' Facial Action Unit proved to be harder to score because of reasons stated previously. Alternatively, another factor that may have influenced this score is the inability to see a FAU. Donkeys have longer coats than the average horse. This shaggy coat can make it difficult to see orbital tightening, the angle of upper eyelid or sclera. If a FAU was not clearly visible, observers left it blank. These extra blank categories for the donkeys may have impacted the intra- and inter-observer reliability.

Looking at the results of all three observers, both the horse and the donkey share many similar scores. The highest scoring FAU for both species is the 'Ears' FAU. On the other hand, there is a trend with the lowest scores being either the 'Corners mouth/lips' or the 'Nostrils' FAU. As the reasoning behind these trends have been discussed previously, they will not be discussed again.

Observers noticed that the average pain score for the donkey pain face was higher than the horse pain face, with average of the total FAU of the donkey being above 3 and the average of the total FAU of the horse being close to 2. The difference in the pain scores may be because the donkey photos were mainly acquired from a sanctuary in the UK. This sanctuary houses very old donkeys with chronic health problems. Alternatively, a lot of horse photos are of healthy horses that owners uploaded onto the EPWA website. In a follow up study, it would be interesting to compare the average of the total scores of the FAUs of both donkey and horse photos when there are the same number of photos for each animal and the photos are all acquired from owners, so that there will be no bias in the pain scoring.

5.10 Limitations of this research

There are many factors that can affect the pain score and thus have an effect on the consistency of scores as well as the accuracy. One of these factors is the fact that different coats make it difficult to assess certain parameters. The assessment of useful photos of brown coated horses and donkeys presented no difficulties. However, the assessment of very dark or very light coated horses and donkeys was more complicated. This is especially true when scoring the angulated upper eyelid, corners of the mouth/lips, and the nostrils. The length of an animal's coat also interfered with scoring. Donkeys tend to have long coats and horses can sometimes have long forelocks. These cosmetic features can inhibit FAU scoring of orbital tightening, angulated upper eyelid and the presence of the sclera of eye. Other studies also had difficulty in scoring animals. In a previous study by Dalla Costa et al., (2014), researchers noted that photos of dark brown or black coloured horses taken in a profile view were more difficult to score than horses that had lighter coats. This was especially true for the orbital tightening and prominent strained chewing muscles categories (12% and 16% respectively).

Furthermore, the way photos were taken also played a role in the observer's ability to assess FAUs. The lighting or the angle of the photo influenced the possibility to score certain parameters. Bad lighting can exaggerate a feature, making it appear more prominent. Alternatively, sometimes it is difficult to see any features on the horse or donkey because the photo is so dark. The angle of the photo greatly determines which parameters can be scored. Sometimes the angle of the eye, mouth or nostrils were not visible. Action photos, such as when the horse is working, or eating can also affect the parameters of the nostril and the mouth. Due to inability to clearly distinguish the FAUs, a number of photos have been rejected and have not been used. In addition, there were also a lot of photos on which not all of the FAUs were visible. In the future, when owners are asked to upload photos for research, more sample photos could be placed on the website, showing how the lighting and angle of the photo should be taken so that all the FAUs are visible.

The environment is another factor that could be a limitation of this research. The environment has an influence on the behaviour of the animals. Under different circumstances not all FAUs are equally reliable, for example if we look at the positioning of the ears. The rotation of the ears both with the horse and donkey can rapidly change under different circumstances. A person passing by with food or a photoshoot with the perfect setting, or noise in the surroundings can change the positioning of the ears. In addition, it was noticed that when a series of photos of the same horse were taken, all the FAUs were the same except for the 'Ears'. Therefore, the photo would have been more representative if the horse had been in its natural state. If the owner had taken a photo from a distance the photo would be more representative of the animal's health. The importance of a representative photo has not been of any influence on the reliability of this research. However, it becomes essential to collect representative photos when the research is held to determine whether a horse or donkey is in pain.

The consistency of the pain scores can be influenced by all the previously mentioned factors that make distinguishing specific features a difficult task. Furthermore, it can also be affected by the experience of the observers. After scoring the photos, it was noticed that previous photos had some influence on the score given to the current photo. For example, if a horse with very wide nostrils was given a score of 2 and the following photo was a horse with flared nostrils, the observers may score the horse with flared nostrils as a 1. However, if the preceding photo was a horse with very small, relaxed nostrils, that received a score of 0, the flared nostrils may appear much larger and be given a score of 2. Therefore, the scoring of the exact same photo by the same observer may differ based on the preceding photo, which lead to some disturbances in the intra-observer reliability. However, as can be seen from the data, while this may have had a slight impact, the impact was not enough to be considered significant.

The last limitation in this research concerns the number of assessors. Only two observers were assessing the entire set of photos of horses and donkeys. If there had been more observers scoring all photos, the reproducibility of the assessment of photos could have been higher. In the future it will be interesting to use a larger amount of observers. Furthermore, the effect of the observers training will also be investigated.

6 Conclusion

Both pain scales are reproducible instruments for objective pain evaluation in horse and donkey photos. This study shows that the inter- and intra- observer reliability for total Facial Action Units (FAUs) were acceptable and significant, indicating that the new scoring systems, Horse Pain Face (HPF) and Donkey Pain Face (DPF), as well as the training for each scoring system was effective. The results support that the training protocol is useful for producing reliable scoring of equine facial characteristics from objective observers, proving that anyone can use the scales properly if they are trained. Although, the HPF scores show a weak inter-observer reliability for the individual FAU 'Corners mouth/lips' and similarly, the DPF scores show a weak inter-observer reliability for the FAU 'Nostrils'. While this study was a success, further research is required to ensure that the definitions of individual parameters of the HPF and DPF become more explicit and easy to distinguish between different scoring categories. Both the HPF and DPF show great promise to become valuable tools in scoring equine pain. In a follow-up study the results of this research will be used in a project using artificial intelligence techniques (deep learning and neural networks) for automated pain scoring from facial expressions of horses and donkeys.

7 Acknowledgements

I would like to thank my supervisor Dr. J.P.A.M. (Thijs) van Loon for assisting in the preparation and performance of this research and for giving me feedback during the writing process. I would like to thank Amber Irick for helping me to collect data.

8 References

Ashley, F. H., Waterman-Pearson, A. E., Whay, H. R., (2005). Behavioural assessment of pain in horses and donkeys: Application to clinical practice and future studies. *Equine Veterinary Journal, 37*(6), 565-575.

Burden, F., Thiemann, A., (2015). Donkeys are different. *Journal of Equine Veterinary Science, 35*(5), 376. doi:10.1016/j.jevs.2015.03.005

Bussieres, G., Jacques, C., Lainay, O., Beauchamp, G., Leblond, A., Cadore, J. L., Desmaiziéres, L.M., Cuvelliez, S.G., Troncy, E., (2008). Development of a composite orthopaedic pain scale in horses. *Research in Veterinary Science, 85*(2), 294-306. doi:S0034-5288(07)00246-9 [pii]

Dalla Costa, E., Minero, M., Lebelt, D., Stucke, D., Canali, E., & Leach, M. C., (2014). Development of the horse grimace scale (HGS) as a pain assessment tool in horses undergoing routine castration. *PloS One, 9*(3), e92281. doi:10.1371/journal.pone.0092281 [doi]

Dalla Costa, E., Stucke, D., Dai, F., Minero, M., Leach, M. C., Lebelt, D., (2016). Using the horse grimace scale (HGS) to assess pain associated with acute laminitis in horses (equus caballus). *Animals : An Open Access Journal from MDPI, 6*(8), 10.3390/ani6080047. doi:10.3390/ani6080047 [doi]

de Grauw, J. C., van Loon, J. P., (2016). Systematic pain assessment in horses. *Veterinary Journal, 209*, 14-22. doi:10.1016/j.tvjl.2015.07.030 [doi]

DeLoach, L. J., Higgins, M. S., Caplan, A. B., Stiff, J. L., (1998). The visual analog scale in the immediate postoperative period: Intrasubject variability and correlation with a numeric scale. *Anesthesia and Analgesia, 86*(1), 102-106. doi:10.1097/00000539-199801000-00020 [doi]

Friesen, E., Ekman, P., (1978). Facial action coding system: A technique for the measurement of facial movement. *Palo Alto, 3*

Gleerup, K. B., Forkman, B., Lindegaard, C., Andersen, P. H., (2015). An equine pain face. *Veterinary Anaesthesia and Analgesia, 42*(1), 103-114. doi:10.1111/vaa.12212 [doi]

Haxby, J. V., Hoffman, E. A., Gobbini, M. I., (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences, 4*(6), 223-233. doi:S1364-6613(00)01482-0 [pii]

Lamont, L. A., Tranquilli, W. J., Grimm, K. A., (2000). Physiology of pain. *The Veterinary Clinics of North America.Small Animal Practice, 30*(4), 703-28, v.

Librado, P., Fages, A., Gaunitz, C., Leonardi, M., Wagner, S., Khan, N., Hanghoj, K., Alquraishi, S.A., Alfarhan, A.H., Al-Rasheid, K.A., der Sarkissian, C., Schubert, M., Orlando, L., (2016). The evolutionary origin and genetic makeup of domestic horses. *Genetics, 204*(2), 423-434. doi:10.1534/genetics.116.194860

Lindegaard, C., Thomsen, M. H., Larsen, S., Andersen, P. H., (2010). Analgesic efficacy of intra-articular morphine in experimentally induced radiocarpal synovitis in horses. *Veterinary Anaesthesia and Analgesia, 37*(2), 171-185. doi:10.1111/j.1467-2995.2009.00521.x [doi]

Lu, Y., Mahmoud, M., Robinson, P., (2017) *Estimating sheep pain level using facial action unit detection.* doi:10.1109/FG.2017.56

McFadzean, W. J. M., Love, E.J., (2019) *Perioperative pain management in horses.*

McLennan, K., Mahmoud, M., (2019). Development of an automated pain facial expression detection system for sheep (ovis aries). *Animals : An Open Access Journal from MDPI, 9*(4), 10.3390/ani9040196. doi:E196 [pii]

Merskey, H., Bogduk, N., (1994). Classification of chronic pain, IASP task force on taxonomy. *Seattle, WA: International Association for the Study of Pain Press (also Available Online at Www.Iasp-Painorg),*

Morrow, L. D., Smith, K.C., Piercy, R.J., du Toit, N., Burden, F.A., Olmos, G., Gregory, N.G., Verheyen, K.L.P., (2011). Retrospective analysis of post-mortem findings in 1,444 aged donkeys. *Journal of Comparative Pathology, 144*(2-3), 145. doi:10.1016/j.jcpa.2010.08.005

Poole, G. D., Craig, K. D., (1992). Judgments of genuine, suppressed, and faked facial expressions of pain. *Journal of Personality and Social Psychology, 63*(5), 797-805.

Pritchard, J. C., Lindberg, A.C., Main, D.C.J., Whay, H.R., (2005). Assessment of the welfare of working horses, mules and donkeys, using health and behaviour parameters. *Preventive Veterinary Medicine, 69*(3-4), 265. doi:10.1016/j.prevetmed.2005.02.002

Prkachin, K. M., Mercer, S. R., (1989). Pain expression in patients with shoulder pathology: Validity, properties and relationship to sickness impact. *Pain, 39*(3), 257-265.

Regan, F. H., Hockenhull, J., Pritchard, J.C., Waterman-Pearson, A.E., Whay, H.R., (2014). Behavioural repertoire of working donkeys and consistency of behaviour over time, as a preliminary step towards identifying pain-related behaviours. *PLoS ONE, 9*(7), e101877. doi:10.1371/journal.pone.0101877

Regan, F. H., Hockenhull, J., Pritchard, J.C., Waterman-Pearson, A.E., Whay, H.R., (2016). Identifying behavioural differences in working donkeys in response to analgesic administration. *Equine Veterinary Journal, 48*(1), 33. doi:10.1111/evj.12356

Robertson, S., (2006). The importance of assessing pain in horses and donkeys. *Equine Veterinary Journal, 38*(1), 5. doi:10.2746/042516406775374379

Taylor, P., Pascoe, P., Mama, K., (2002). Diagnosing and treating pain in the horse. where are we today? *The Veterinary Clinics of North America.Equine Practice, 18*(1), 1-19, v.

van Dierendonck, M. C., Burden, F., Rickards, K., van Loon, J.P.A.M., (Submitted data) Monitoring acute pain in donkeys with the Equine Utrecht University Scale for Donkeys Composite Pain Assessment (EQUUS-DONKEY-COMPASS) and the Equine Utrecht University Scale for Donkey Facial Assessment of Pain (EQUUS-DONKEY-FAP).

van Loon, J. P.A.M., van Dierendonck, M. C., (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. *Veterinary Journal, 206*(3), 356-364. doi:10.1016/j.tvjl.2015.08.023 [doi]

van Loon, J. P.A.M., van Dierendonck, M. C., (2017). Monitoring equine head-related pain with the equine utrecht university scale for facial assessment of pain (EQUUS-FAP). *Veterinary Journal, 220*, 88-90. doi:S1090-0233(17)30011-4 [pii]

van Loon, J. P. A. M., van Dierendonck, M. C., (2019). Pain assessment in horses after orthopaedic surgery and with orthopaedic trauma. *Veterinary Journal, 246*, 85-91. doi:S1090-0233(19)30010-3 [pii]

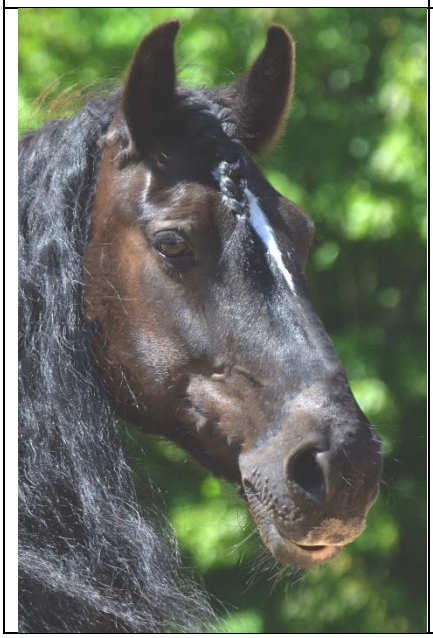
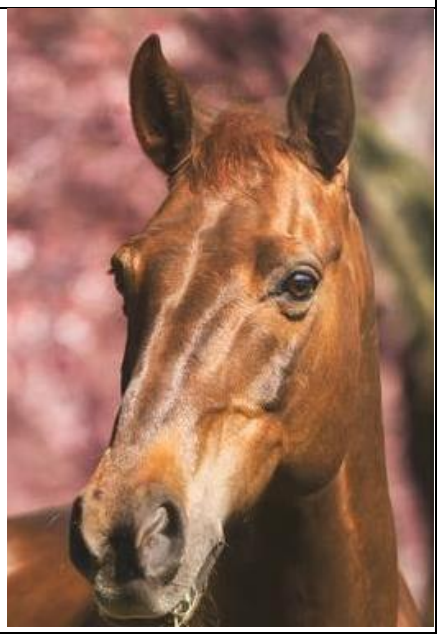
Wagner, A. E., (2010). Effects of stress on pain in horses and incorporating pain scales for equine practice. *The Veterinary Clinics of North America.Equine Practice, 26*(3), 481-492. doi:10.1016/j.cveq.2010.07.001 [doi]

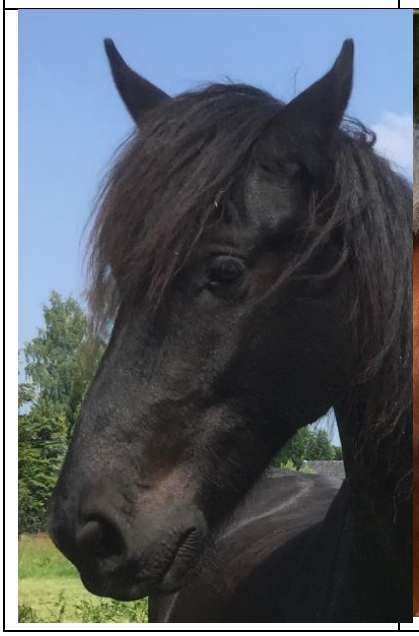
Wathan, J., Burrows, A. M., Waller, B. M., McComb, K., (2015). EquiFACS: The equine facial action coding system. *PloS One, 10*(8), e0131738. doi:10.1371/journal.pone.0131738 [doi]

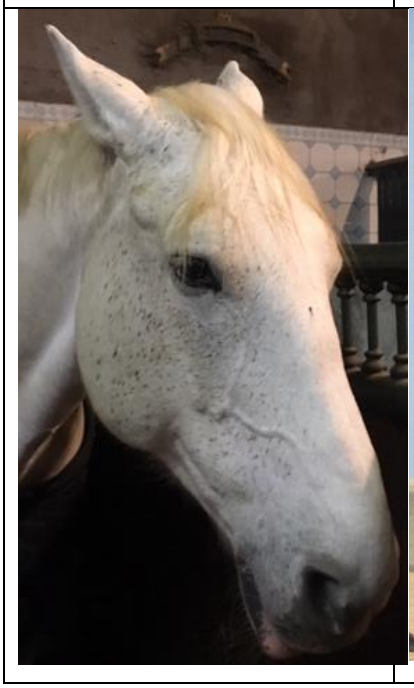
9 Attachments

9.1 Instruction Manual Horse

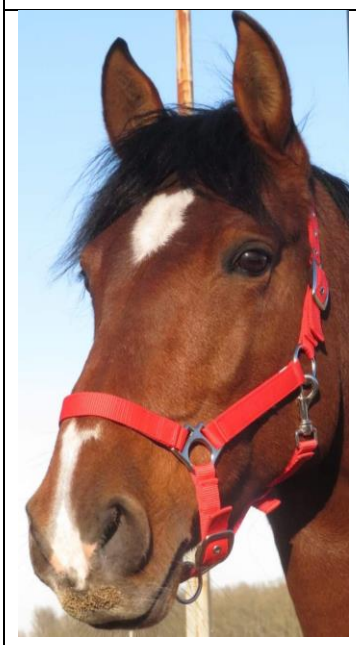
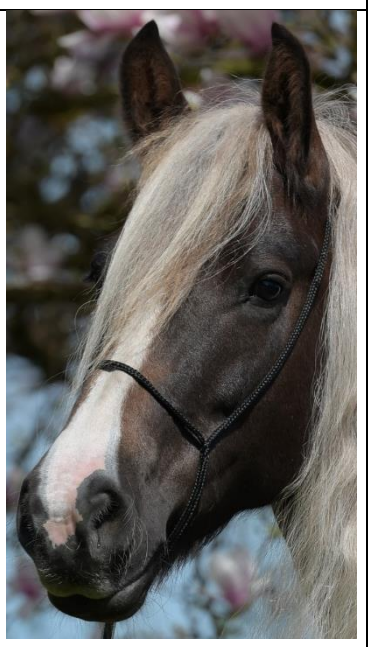
Ears

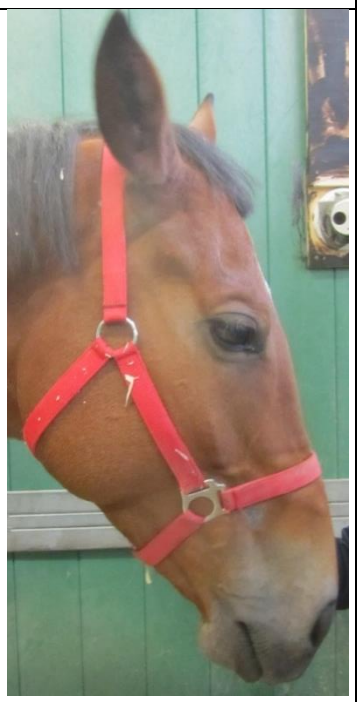
0	0	0
		

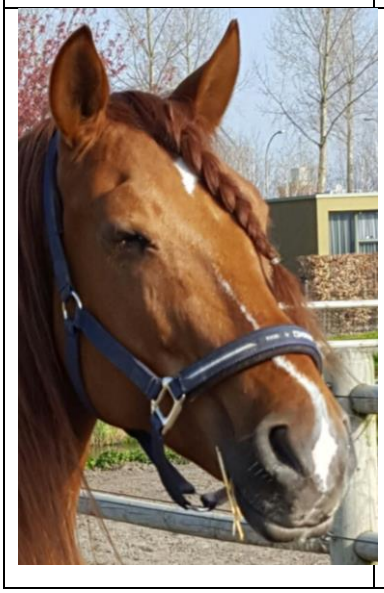
1	1	1
		

2	2	2
		

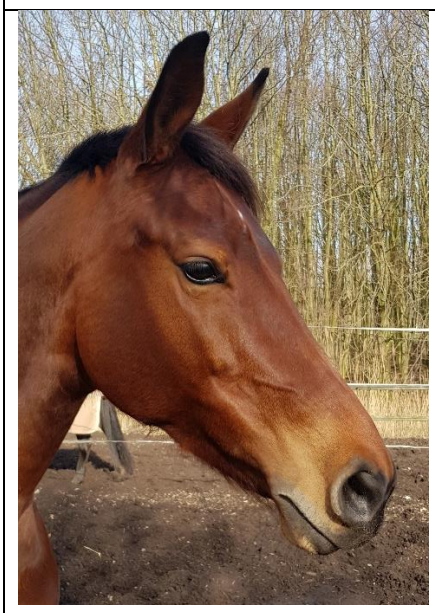
Orbital Tightening

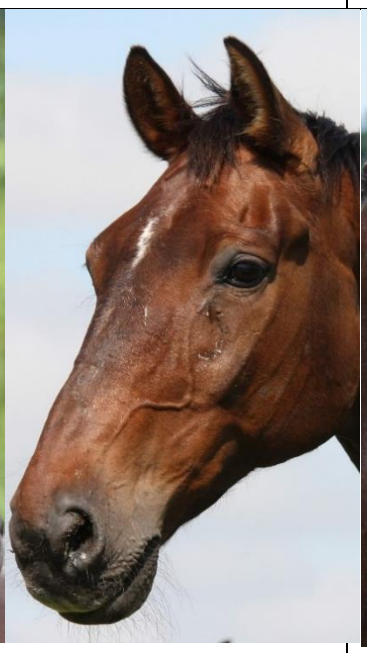
0	0	0
		

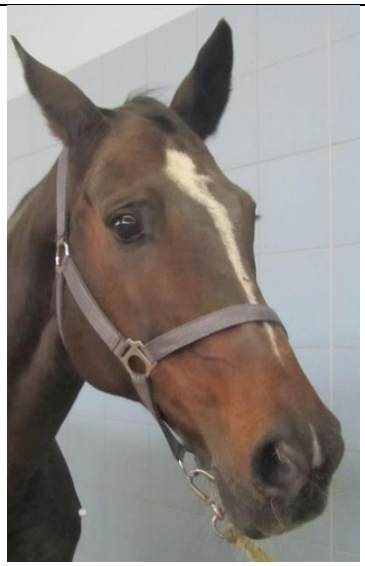
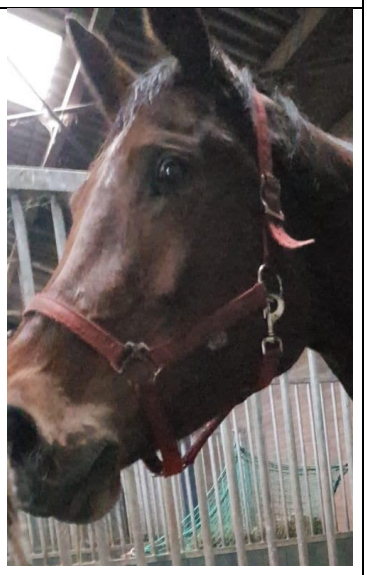
1	1	1
		

2	2	2
		

Angulated upper eyelid

0	0	0
		

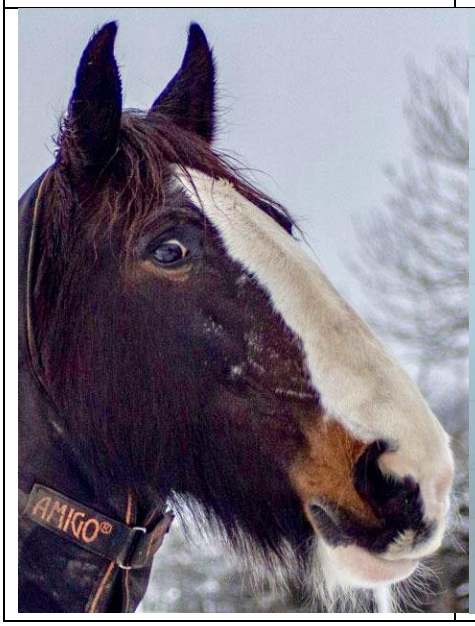
1	1	1
		

2	2	2
		

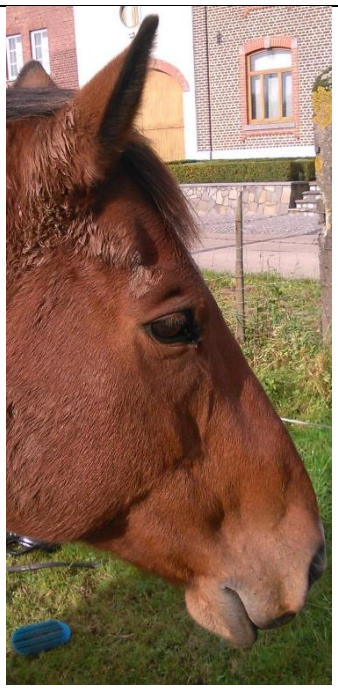
Visibility of the sclera

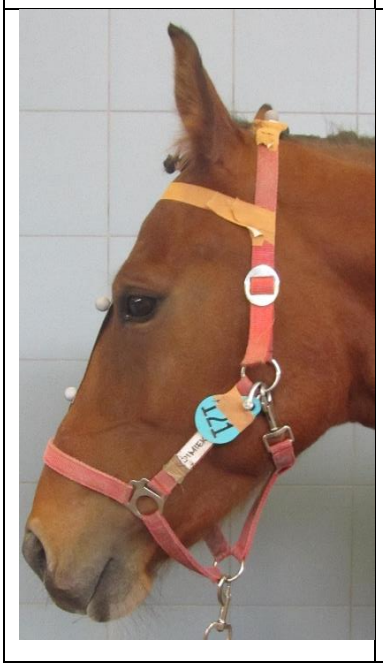
0	0	0
		

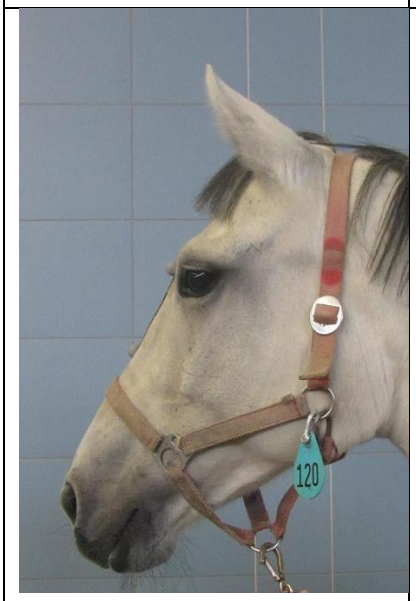
1	1	1
		

2	**2**	**2**
		

Corners mouth/lips

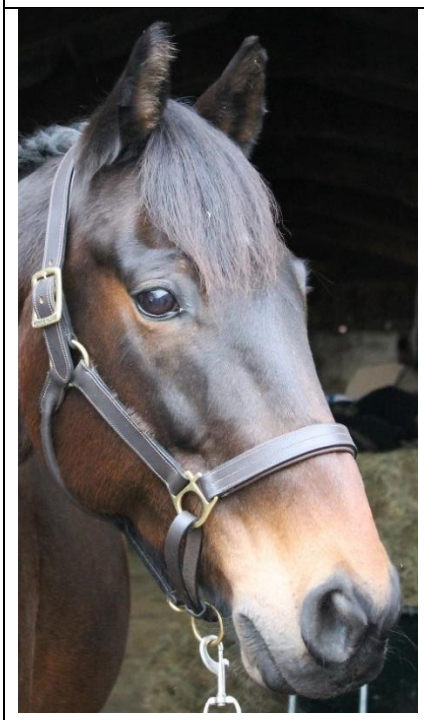
0	0	0
		

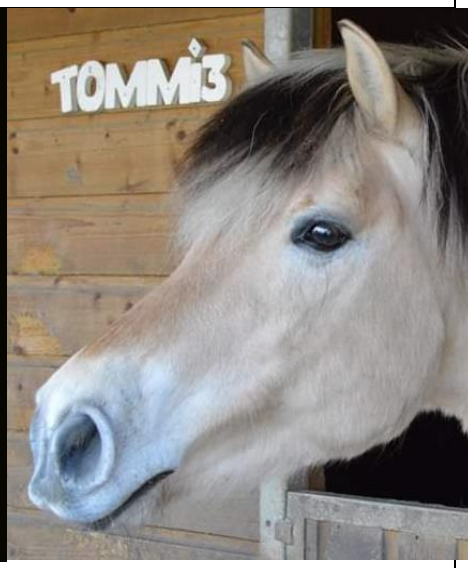
1	1	1
		

2	**2**	**2**
		

Nostrils

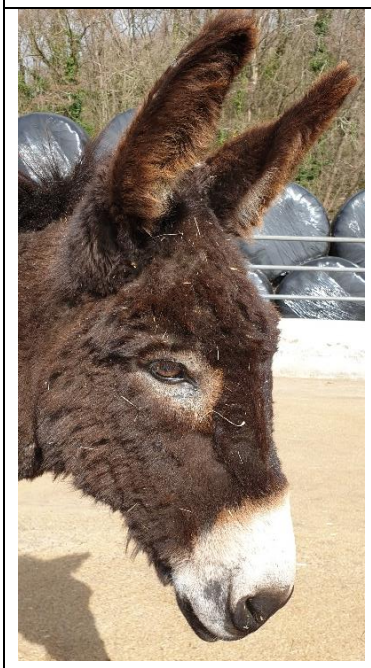
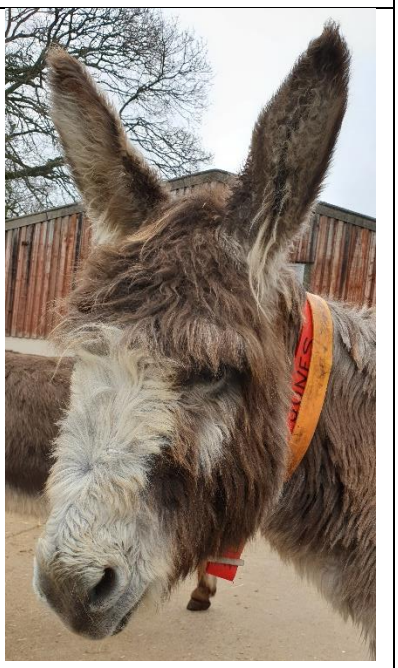
0	0	0
		

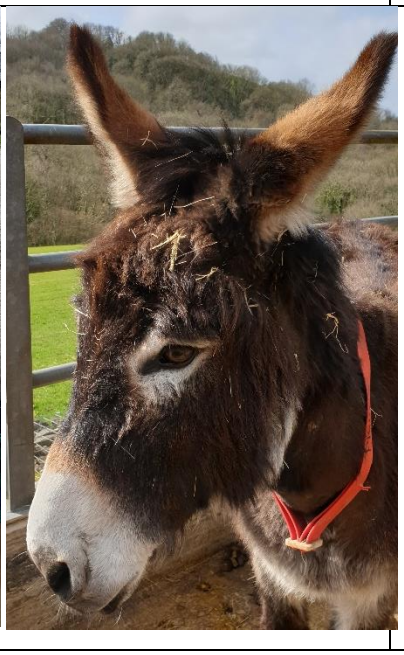
1	1	1
		

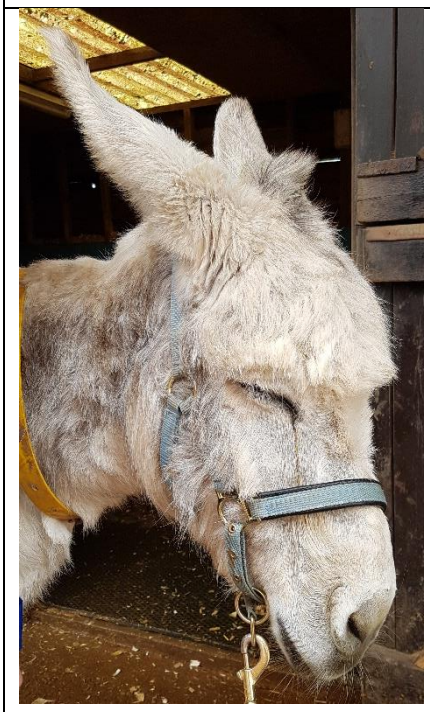
2	2	2
		

9.2 Instruction Manual donkey

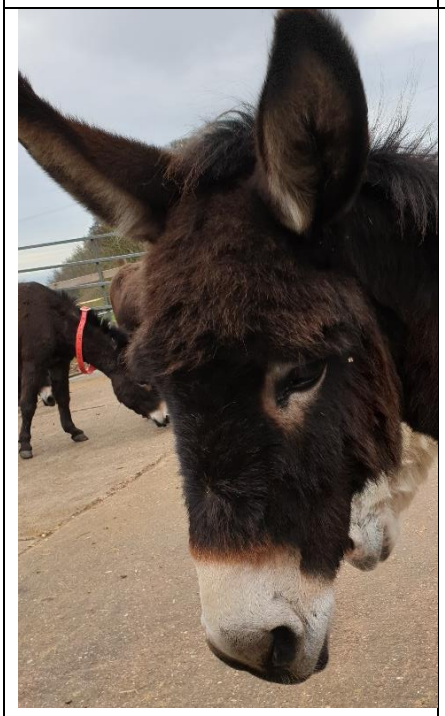
Ears

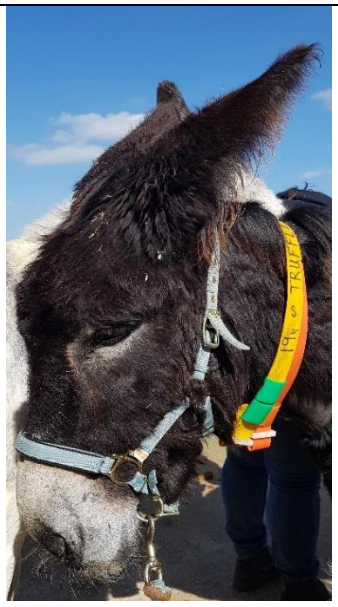
0	0	0
		

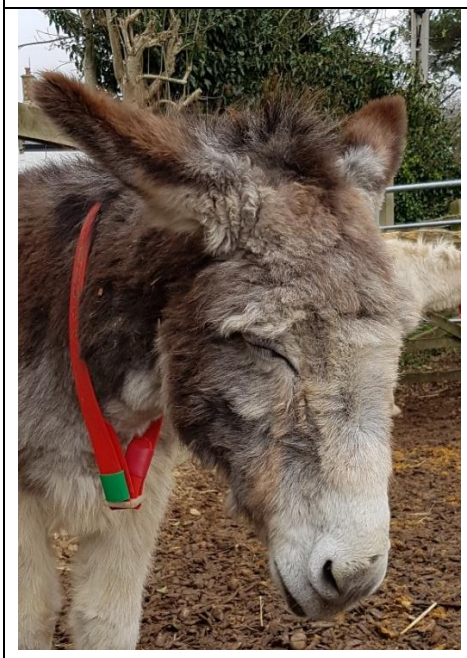
1	1	1
		

2	**2**	**2**
		

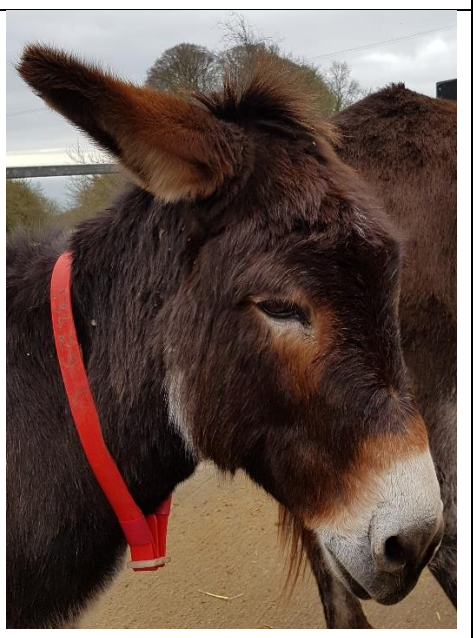
Orbital Tightening

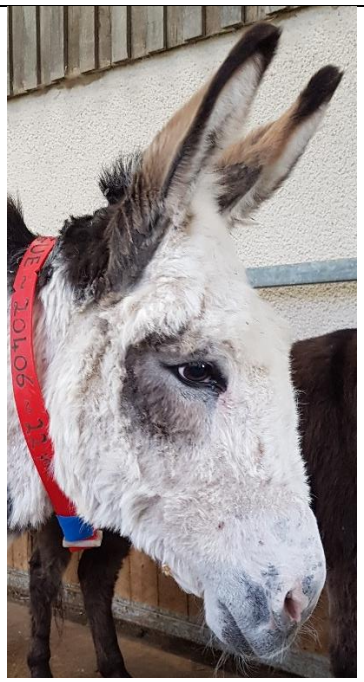
0	0	0
		

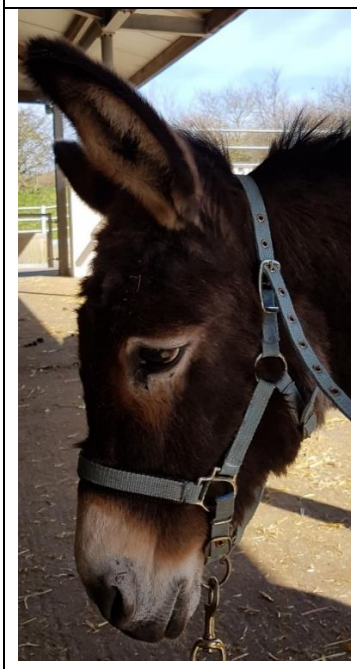
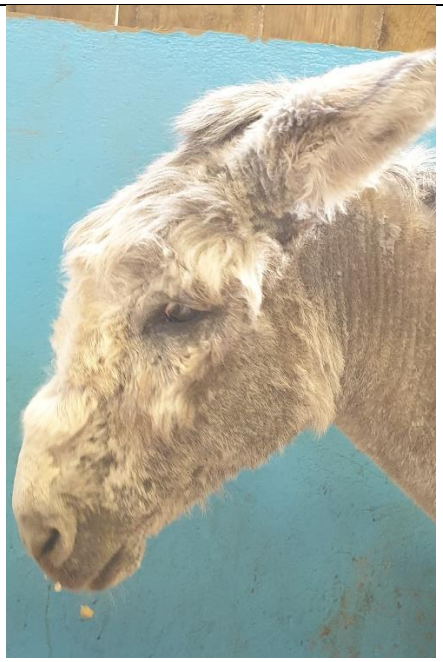
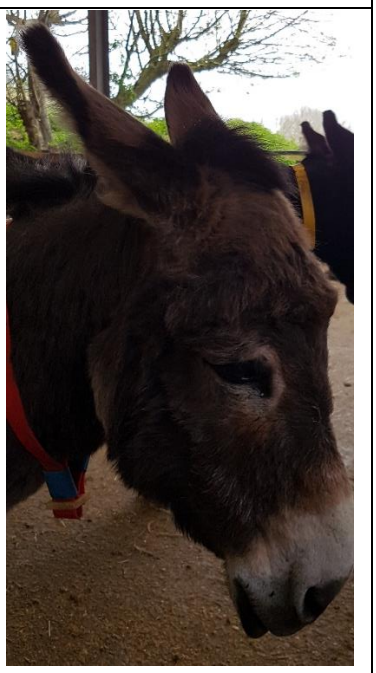
1	1	1
		

2	**2**	**2**
		

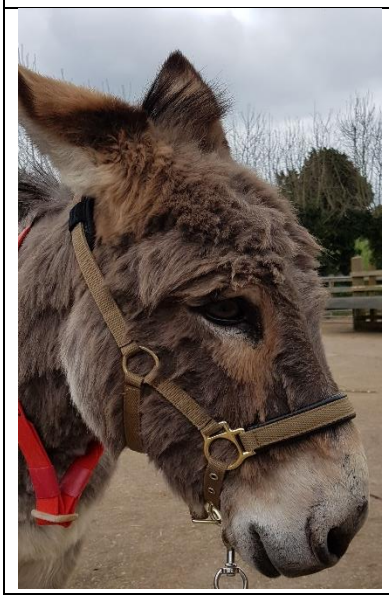
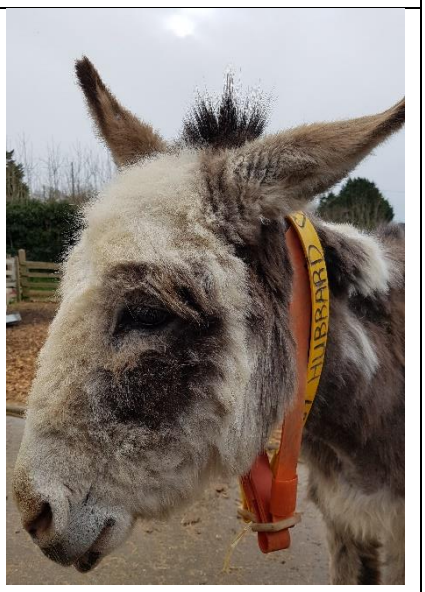
Angulated upper eyelid

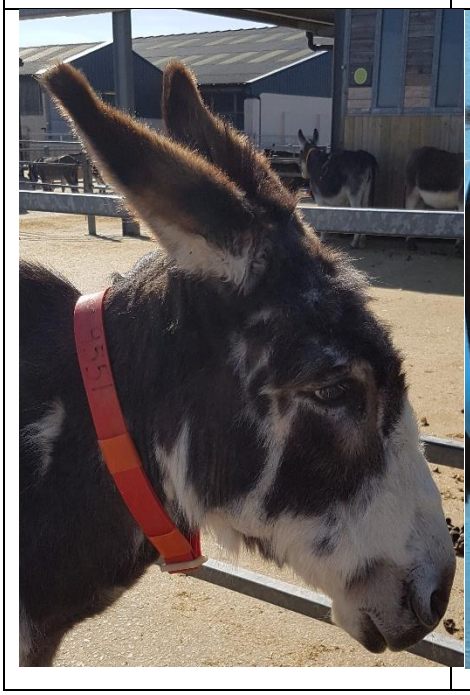
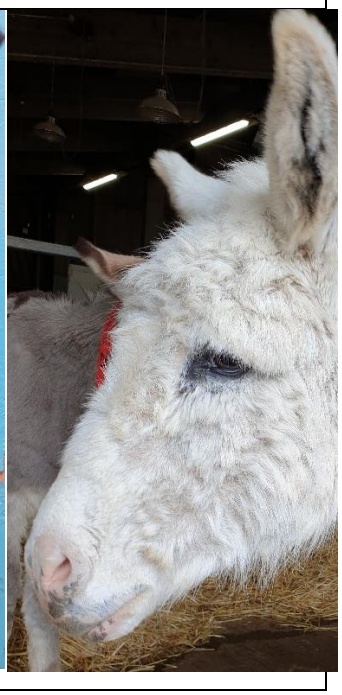
0	0	0
		

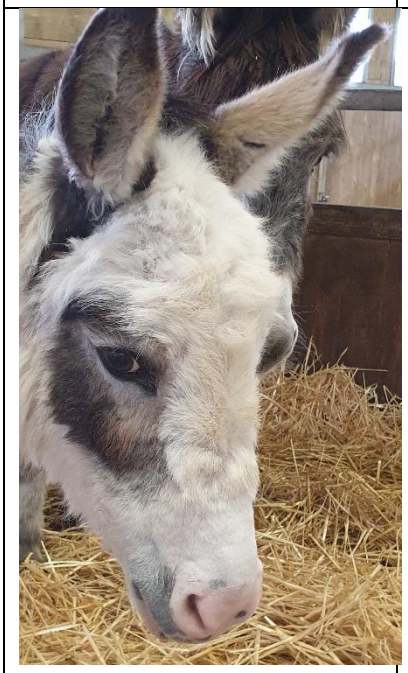
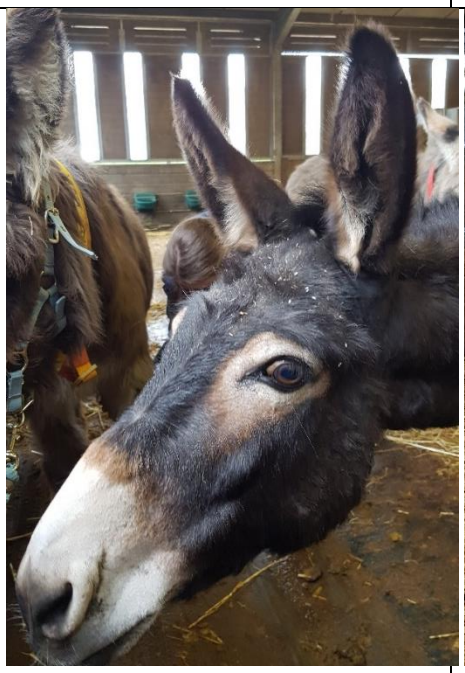
1	1	1
		

2	2	2
		

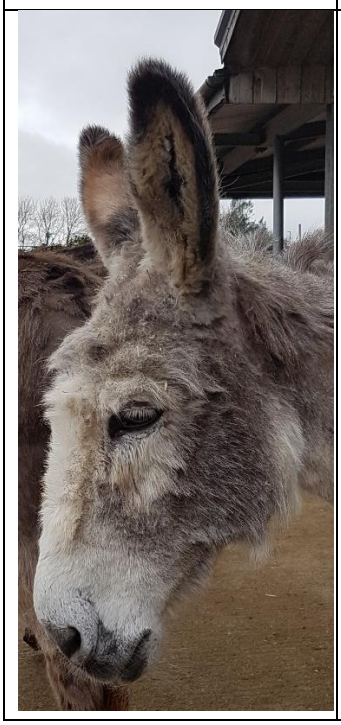
Visibility of the sclera

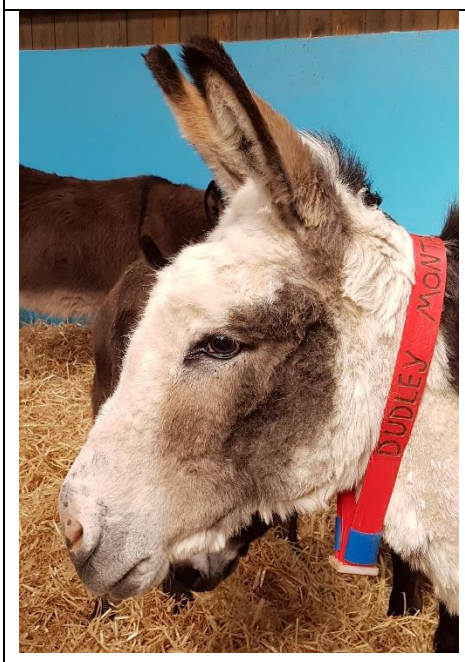
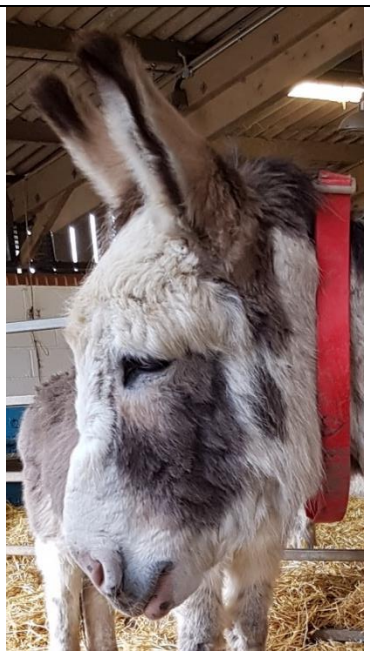
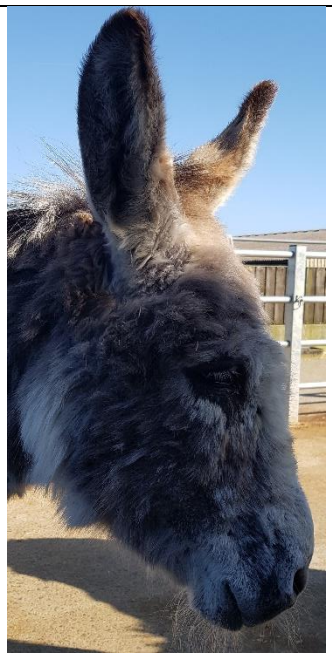
0	0	0
		

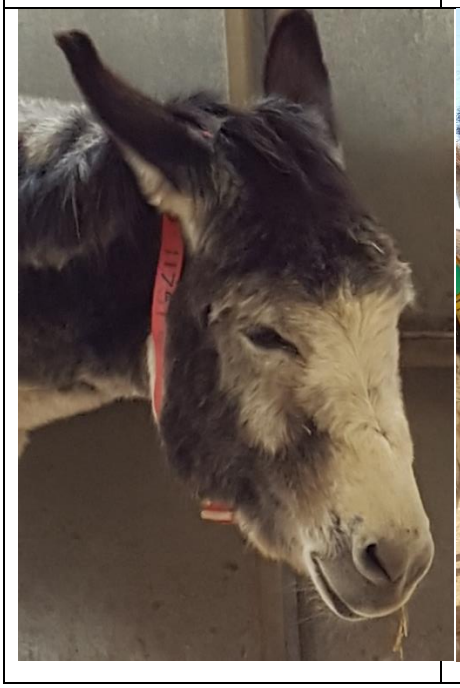
1	1	1
		

2	2	2
		

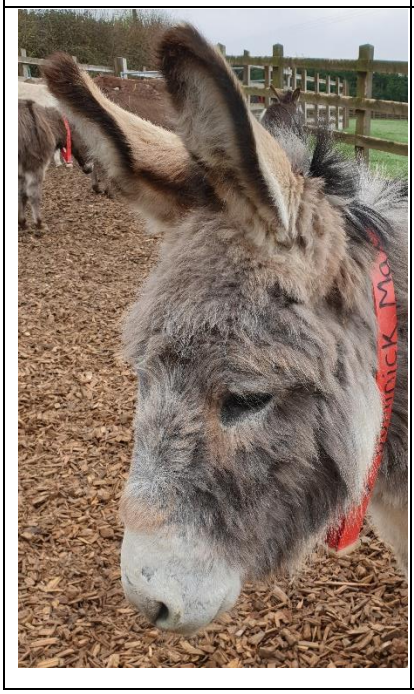
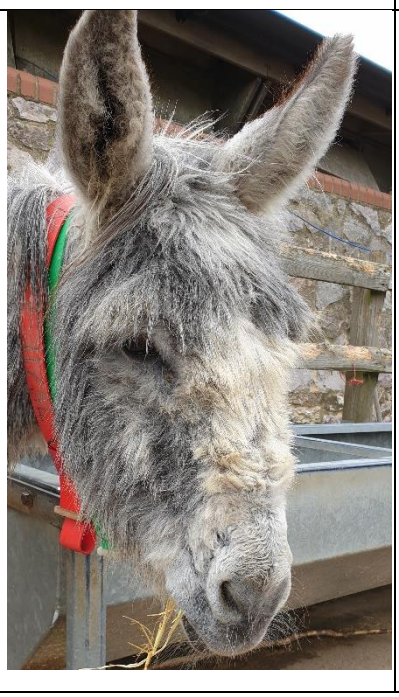
Corners mouth/lips

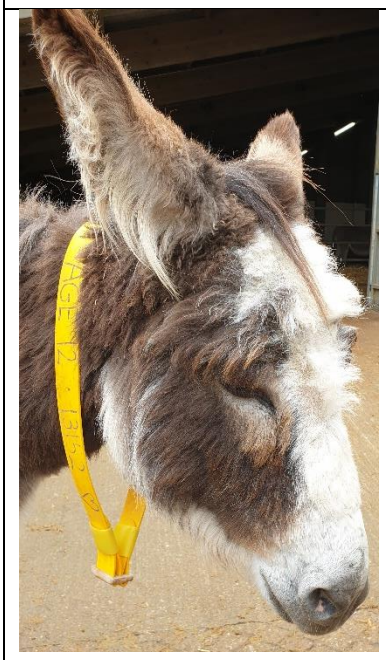
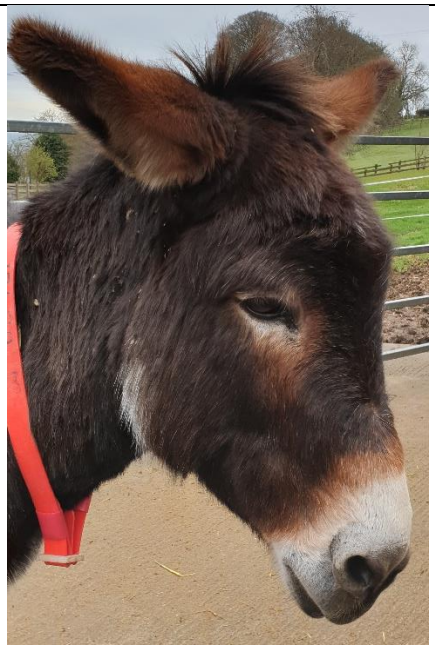
0	0	0
		

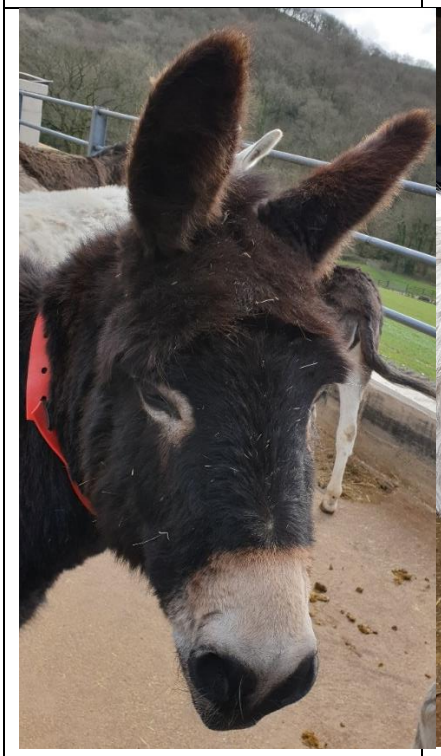
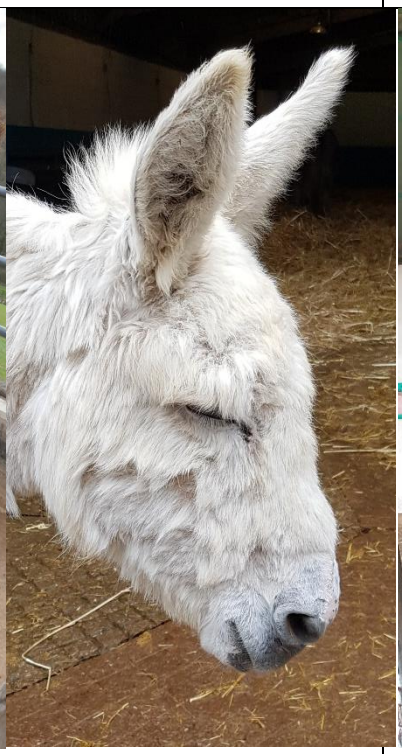
1	1	1
		

2	2	2
		

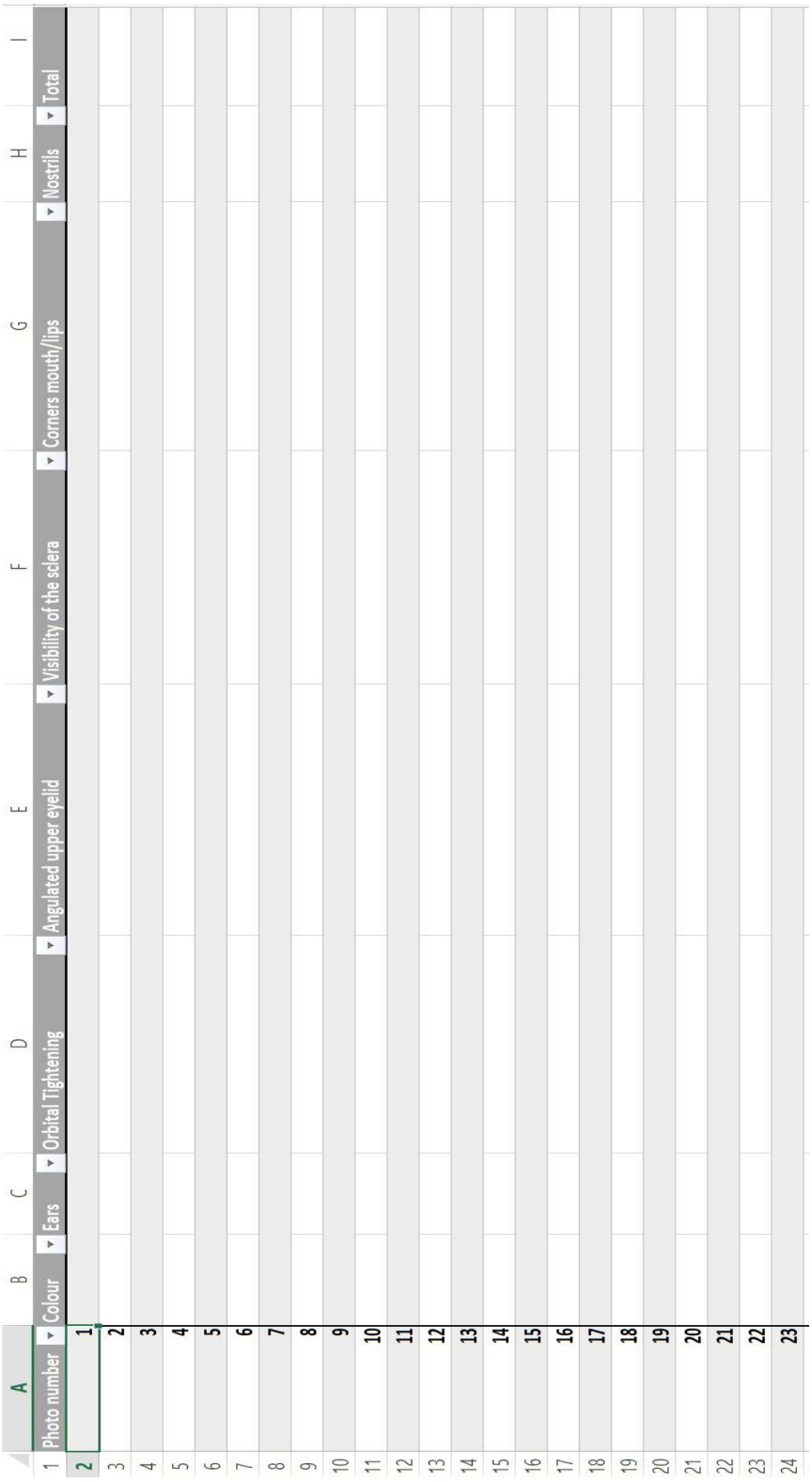
Nostrils

0	0	0
		

1	1	1
		

2	2	2
		

9.3 Scoring table for the assessment of horse photos



	A	B	C	D	E	F	G	H	I
1	Photo number	Colour	Ears	Orbital Tightening	Angulated upper eyelid	Visibility of the sclera	Corners mouth/lips	Nostrils	Total
2	1								
3	2								
4	3								
5	4								
6	5								
7	6								
8	7								
9	8								
10	9								
11	10								
12	11								
13	12								
14	13								
15	14								
16	15								
17	16								
18	17								
19	18								
20	19								
21	20								
22	21								
23	22								
24	23								

9.4 Scoring table for the assessment of donkey photos

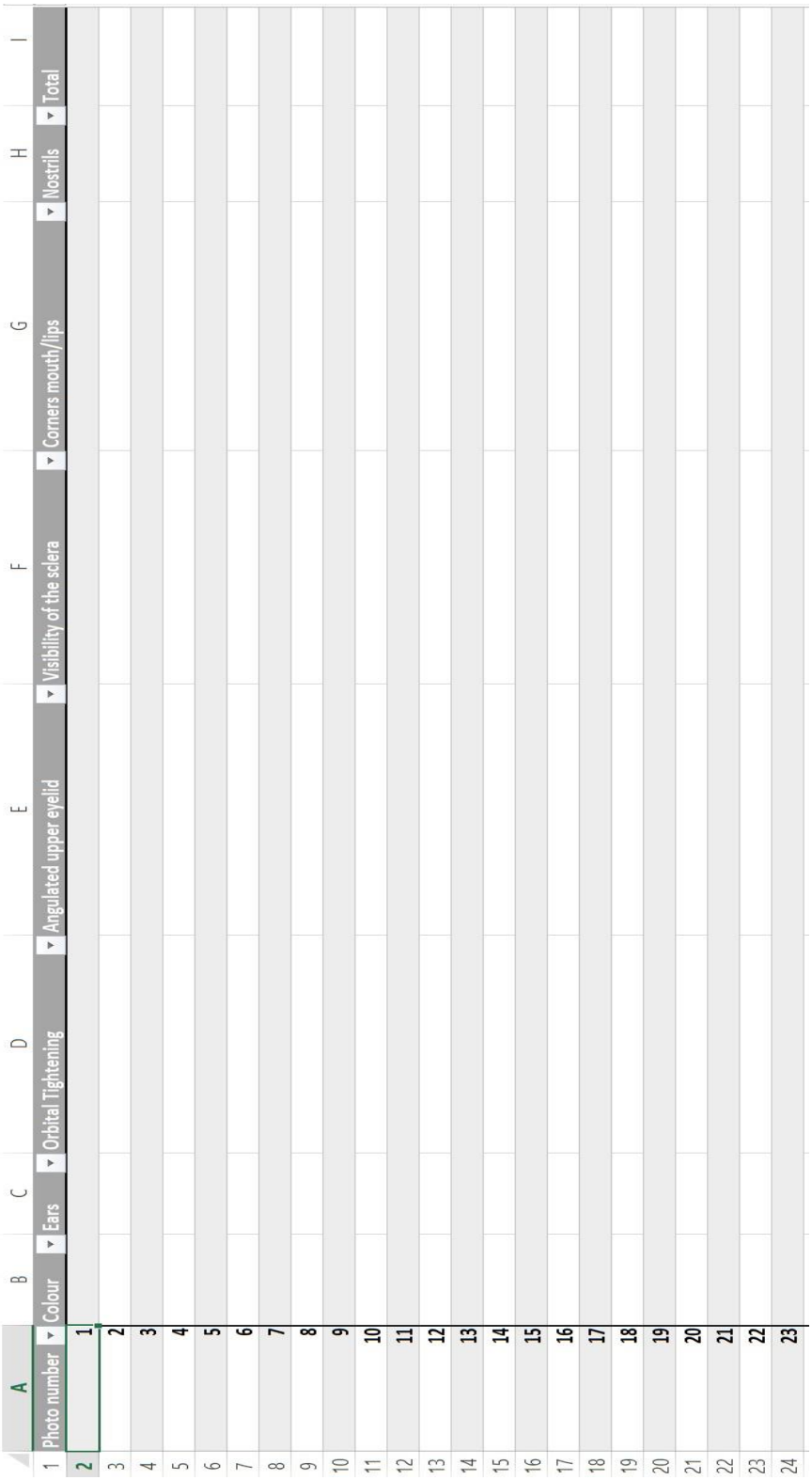


Photo number	Colour	Ears	Orbital Tightening	Angulated upper eyelid	Visibility of the sclera	Corners mouth/lips	Nostrils	Total
1								
2								
3								
4								
5								
6								
7								
8								
9								
10								
11								
12								
13								
14								
15								
16								
17								
18								
19								
20								
21								
22								
23								

9.5 Example of a scored horse photo

Ears	Both ears turned forwards At least one ear lateral position or further to backwards Both ears turned backwards	**0** 1 2
Orbital Tightening	Relaxed A bit tightening of the eyelids Obviously tightening of eyelids / eye closed	**0** 1 2
Angulated upper eyelid	Relaxed A bit more visible Obviously more visible	**0** 1 2
Visibility of the sclera	Sclera is not visible An edge of the sclera is visible Obviously more visible	**0** 1 2
Corners mouth/lips	Relaxed Lifted a bit Obviously lifted / strained	**0** 1 2
Nostrils	Relaxed A bit more opened Obviously more opened (dilated mediolaterally)	0 **1** 2
Total		1 /10

9.6 Example of a scored donkey photo

Ears	Both ears turned forwards At least one ear lateral position or further to backwards Both ears turned backwards	0 1 **2**
Orbital Tightening	Relaxed A bit tightening of the eyelids Obviously tightening of eyelids / eye closed	0 1 **2**
Angulated upper eyelid	Relaxed A bit more visible Obviously more visible	**0** 1 2
Visibility of the sclera	Sclera is not visible An edge of the sclera is visible Obviously more visible	**0** 1 2
Corners mouth/lips	Relaxed Lifted a bit Obviously lifted / strained	0 1 **2**
Nostrils	Relaxed A bit more opened Obviously more opened (dilated mediolaterally)	**0** 1 2
Total		6 /10