A study on validity and reliability of on-farm tests to measure human–animal relationship in horses and donkeys

a r t i c l e i n f o

Keywords:
Horses
Donkeys
Human–animal relationship
On-farm welfare assessment

a b s t r a c t

The development and maintenance of a positive human–horse/donkey relationship is essential in order to decrease accidents and reduce negative states of equine welfare. In many animal species the reaction of animals to humans during specific behavioural tests is influenced by their past interaction and is linked to the level of fear felt in the presence of a human. The present research aims to assess whether a set of on-farm behavioural tests allow differentiation between horse facilities with excellent or sub-optimal human–animal relationship. Furthermore, we evaluated mid-term repeatability (3-month intervals), inter-observer reliability and on-farm feasibility of these behavioural tests in single stabled horses and in group housed donkeys. Eleven horse and eight donkey facilities ($N = 313$ adult horses; $N = 47$ adult donkeys) were visited twice at 3-month intervals. Horse facilities were selected on the basis of reports of inspections on animal welfare conducted by competent local authorities; they were classified as "excellent" ($N = 5$) and "sub-optimal" ($N = 6$). Four observers, with no experience in assessing equine welfare and not aware of the a priori category of the facility, were trained to perform and score standardised human–equine behavioural tests: avoidance distance test (AD), voluntary animal approach test (VAA), forced human approach test (FHA), walking down side and tail tuck.

All the behavioural tests carried out proved to be feasible in an on-farm environment. In spite of the fact that the reactions of horses were largely positive, those kept in facilities with "sub-optimal" relationship showed avoidance and aggressive behaviours more often when approached (GLMM $P < 0.05$). As for donkeys, less than 30% of the animals exhibited negative behaviour towards the assessor.

Observer's agreement of AD, VAA, FHA, WDS and tail tuck scoring was consistent for both species (percentage agreement ranged between 67.7% and 93.3%). Repeatability of tests was good for all the tests and no significant differences were found between two repetitions at 3-month intervals. Our results support the findings described for working donkeys and show that, also on-farm, the assessment of donkeys' reactions to an unknown human during standardised tests could prove useful in evaluating the quality of their relationship with humans.

Further research is needed to verify if our findings can be generalised for different husbandry conditions.

1. Introduction

The human–animal relationship is a continually changing process that can be defined as the mutual perception that develops and expresses itself in the mutual behaviour (Estep and Hetts, 1992). This relationship is based on repeated interaction, defines each subject's expectation during the encounters that follow (Fureix et al., 2009; Hausberger and Muller, 2002; Ligout et al., 2008; Waiblinger et al., 2006) and is, in addition, linked to the level of fear felt in the presence of a human (Hemsworth et al., 1993; Rushen et al., 1999). The quantity and quality of interaction influences the emotional, cognitive and productive behaviour of the animal (Hemsworth, 2003; Mendl et al., 2010). In horses several factors can affect the relationship, such as early experience and training (Henry et al., 2006; Sankey et al., 2010), breed and temperament (Hausberger and Muller, 2002; Lesimple et al., 2011), and even chronic discomfort (Fureix et al., 2010). The relationship will range from confidence to fear, implying different emotion involved, in accordance with the perceived importance of the interactions (positive/negative) (Hemsworth et al., 1993; Lansade et al., 2008; Søndergaard and Halekoh, 2003). Different studies were carried out to evaluate the human–animal relationship where the animal-based measures used to assess this relationship are based on how they react to humans (for a review on horses see Hausberger et al., 2008). Broadly speaking, tests designed to test the reaction of equines to people take into consideration the measurement of: reaction to a standing human (Table 1), reaction to a moving person (Table 2) or the reaction to a particular handling (Table 3) (for review, see Waiblinger et al., 2006).

Normally the tests are very simple but when drafting them it is vital to assess their validity and reliability. To gauge the predictive validity of a human–horse/donkey test – a measure of the efficiency of a test to predict results (Acock, 2008; Cronbach and Meehl, 1955) – the selection of the facilities to be assessed is crucial and must be carefully performed in order to ensure coherence with the pre-determined level of quality in the human–animal relationship. One option is to take into consideration previous assessments performed by competent local authorities operating in the area of control of animal welfare. A peculiar aspect of horse management is that, within the same facility, each horse can have a different owner; therefore, the quality of interactions with both owner and groom is reflected in how the animal reacts to unknown humans. As already demonstrated in past research (Chamove et al., 2002; Hemsworth and Coleman, 1998; Waiblinger et al., 2006; Windschnureremail et al., 2009), attitude and behaviour of the stockman/groom are themselves an indication of the quality of the relationship with the animal concerned. In practice, attitude is difficult to measure directly and it is usually pinpointed through a questionnaire. However, an important disadvantage of questionnaires aimed at evaluating personal traits is the tendency for people to present a favourable image of themselves. This bias, called socially desirable responding, confounds research results by creating false relationships between variables (Van de Mortel, 2008).

The observer is considered "the measurer" of behaviour and as is the case with any measuring instrument, his or her measurement can be distorted or imprecise. An observer's reliability is defined by the repeatability of their results. Inter-observer reliability measures the agreement between different assessors and the agreement between observations on the same individual on at least two different occasions (test–retest reliability) is used to verify whether the measure remains the same (Acock, 2008). If correctly carried out, observer training is the best weapon for guaranteeing coherence of measurements and bias control.

Other factors that can affect the reliability of the results are tied to the nature of the behaviour itself and the technique applied for its measurement. For example, pain can make a horse more unwilling to approach a human or even aggressive (Fureix et al., 2010). Furthermore, a motivation, such as hunger, makes it difficult to assess how willing the animal is to approach a human because there is a built in positive aspect perceived or because the human is associated with satisfying the specific motivation (Hausberger and Muller, 2002). If the test is carried out in a place that is unfamiliar to the animal or that is not the animal's usual environment, another confounding factor can be behavioural inhibition brought on by the physical and social novelty of the surroundings (Søndergaard and Halekoh, 2003).

The concepts related to validity and reliability of the above mentioned behavioural tests are the subject of this research paper aimed at investigating the relationship between humans and horses and donkeys in an on-farm environment. At present, there is no research available which evaluates the validity of these on-farm tests on horses, whereas research on working donkeys can be found in specific literature (Burn et al., 2009). The objectives of the present research were to assess whether the tests of avoidance distance, voluntary animal approach and forced human approach are a suitable means of differentiating between horse facilities with excellent or sub-optimal human–animal relationship. Furthermore, we evaluated mid-term repeatability (3-month intervals), inter-observer reliability and on-farm feasibility of the above mentioned behavioural tests and donkey behavioural tests in single stabled horses and in group housed donkeys.

2. Materials and methods

The present study was in compliance with the European Communities Council Directive of 24 November 1986 (86/609/EEC) and followed the requirements of ISAE Ethical Guidelines.

2.1. Horses

2.1.1. Subjects

Eleven horse facilities, housing a total of 313 horses, were visited. The facilities took part in the study voluntarily. Riding schools, sport and leisure horse stables were selected on the basis of official reports of inspections on animal welfare conducted by veterinarians of the competent local authorities (Atto C18, art. 3 e 4 Dir. 1998/58/CE D.Lgs. n. 146 del 26 marzo 2001). Reports were reviewed by

Table 1
Test for reactions to stationary human.

Procedures and other factors[a]	Variables	Species[b]	References
P walks to centre of paddock and stands still	Latency to touch human	H	Fureix et al. (2009) and Søndergaard and Halekoh (2003)
P appears suddenly at door (closed) of box and notes horse's first reaction	First reaction score A–E (friendly–indifferent–very aggressive)	H	Hausberger and Muller (2002)
A is left alone in arena (phase 1, 3 min), P enters and stands still next to wall (phase 2, 3 min), A is left alone (phase 3, 3 min), A is caught	Restlessness, exploration, vocalising, standing alert. Latencies to first contact, of contacts. Time taken to capture; heart rate: mean, deviation from baseline	H	Søndergaard and Halekoh (2003)
P enters the pen, stands stationary opposite the door	Time spent in certain squares, of immobilisation. Latencies to first neigh, to sniffing P. Mean duration sniffing. Number sniffs, glances at P, neighs, defecations, squares entered	H	Lansade et al. (2004)
A left alone in test box for 3 min, P stands in front of box for 3 min, then enters box and holds horse for 3 min	Latencies to first pawing. Frequencies of restless behaviour (pawing, rearing, striking, head shaking). Locomotion; heart rate: mean, variability	H	Visser et al. (2001, 2002)
A released in paddock and left alone for 3 min	Behaviour scored in four situations (catching, led away, hooves picked up, approached). Ease of manipulation score 1–5 (1 = not executed, 5 = executed very easily), sum of scores = total behavioural score; mean HR	H	Jezierski et al. (1999)
P stood in the centre of the 10 m circle, with shoulders rounded and head down, without looking directly at the horse; P stood with head up and shoulders back in an erect, rigid posture and direct eye contact was maintained with the horse	Times to enter each circle and approach the person (maximum of 10 min). Behavioural responses	H	Seaman et al. (2002)
P stood motionless, quiet, and looking down in the middle of the test arena for 3 min. If A did not approach P voluntarily within 3 min, P called the horse	The latency to approach P	H	Maros et al. (2010)
A P standing next to each other in the circle	The total time the horse spent besides P without walking away from him	H	Maros et al. (2010)

[a] P = person; A = animal.
[b] H = horse; D = donkey.

Table 2
Test for reactions to moving human.

Procedures and other factors[a]	Variables	Species[b]	References
P enters paddock and approaches horse/horses slowly (1 step/s, hands at sides); P attempts to touch horse's neck	Score 1–4 (1 = horse moves away, 4 = person could touch the horse)	H	Fureix et al. (2009) and Søndergaard and Halekoh (2003)
A released in paddock and left alone for 3 min	Behaviour scored in four situations (catching, led away, hooves picked up, approached). Ease of manipulation scored 1–5 (1 = not executed, 5 = executed very easily), sum of scores = total behavioural score (TBS); mean heart rate	H	Jezierski et al. (1999)
P approaches the animal's head from 3 to 5 m away, at angle of approximately 45°	Friendly approach: animal turns head towards observer. Avoidance/aggression: animal does one or more of following: turns head away, moves away, flattens ears, attempts to bite or kick	H, D	Burn et al. (2010), Popescu and Diugan (2013) and Pritchard et al. (2005)
P walks down side of animal's body at distance of 30 cm from its side, turning at tail and walking back to head	Any acknowledgement of observer's presence, e.g. ear turn, head turn, move away, kick Tail tuck (donkeys only)	H, D	Burn et al. (2010), Popescu and Diugan (2013) and Pritchard et al. (2005)
P made the horse follow him along a predetermined route in the arena	Total time of following P	H	Maros et al. (2010)

[a] P = person; A = animal.
[b] H = horse; D = donkey.

Table 3
Reaction to handling.

Procedures and other factors[a]	Variables	Species[b]	References
P enters stall, quietly approaches A and attempts to stroke it for 1.5 min. Horses were equipped with ECG telemetry transmitters	Heart rate	H	McCann et al. (1988)
P lead horse around pre-determined course	Head position, ear movements and position, resistance	H	Chamove et al. (2002)
P tries to lead horse across a bridge (maximum three attempts)	Attempts to cross bridge, reluctance behaviour (pawing, rearing, striking, head shaking, walking sideways, pulling backwards), locomotion; heart rate: mean, variability	H	Visser et al. (2001, 2002)
P stroke horses for 90 s. Horses were equipped with wireless ECG monitor recordings	ECG	H	Hama et al. (1996)
P approaches the foal in test pen, halters, picks up feet, leads A through corridor	Time taken to fit with halter, pick up feet, 'walk ratio', defences	H	Lansade et al. (2004)
A left alone in test box for 3 min, P stands in front of box for 3 min, then enters box and holds horse for 3 min	Latencies to first pawing. Frequencies of restless behaviour (pawing, rearing, striking, head shaking). Locomotion; heart rate: mean, variability	H	Visser et al. (2001, 2002)
A is left alone in arena (phase 1, 3 min), P enters and stands still next to wall (phase 2, 3 min), A is left alone (phase 3, 3 min), A is caught	Restlessness, exploration, vocalising, standing alert. Latencies to first contact, of contacts. Time taken to capture; heart rate: mean, deviation from baseline	H	Søndergaard and Halekoh (2003)
A released in paddock and left alone for 3 min	Behaviour scored in four situations (catching, led away, hooves picked up, approached). Ease of manipulation scored 1–5 (1 = not executed, 5 = executed very easily), sum of scores = total behavioural score (TBS); mean heart rate	H	Jezierski et al. (1999)
P tries to lead horse across a bridge (wooden planks on the ground)	Total time to cross bridge, retreat, jumping. Standing still	H	Wolff et al. (1997)
P tries to touch the chin	Proportion of animals avoiding contact or withdrawing head when hand was placed lightly under the chin	H, D	Burn et al. (2010), Popescu and Diugan (2013) and Pritchard et al. (2005)

[a] P = person; A = animal.
[b] H = horse; D = donkey.

a team composed of three official veterinarians working at competent local authorities, taking into account answers (yes/no) and evidences about questions relevant to the human–animal relationship:

- Horses are cared for by a suitable number of stable grooms?
- Staff possess the appropriate ability, advanced and updated knowledge and professional competence?
- Staff attend to specific training regarding the welfare of horses?
- All animals are inspected several times a day?
- In the case of extraordinary management procedures, which are likely to cause suffering to any of the animals, all the precautions to avoid any pain/distress are adopted?
- Staff/owners have positive interactions (e.g. behaving and talking calmly, stroking) with the horse during routine handling procedures?

Following the review of the reports, the horse facilities were classified as "excellent" ($N = 5$; a total of 139 horses) and "sub-optimal" ($N = 6$; a total of 174 horses). All the facilities classified as "excellent" scored positively in all the questions related to the human–animal relationship, whilst facilities considered "sub-optimal" scored negatively in at least three out of six questions. Sub-optimal facilities differed regarding questions which were scored negatively. Further criteria for horse facilities to be included were: horses were primarily managed by stable grooms; preferably 20 horses or more per facility and all horses being stabled indoors in single boxes for at least half of the day, horses have been stabled in the same facility for, at least, the last 2 years. The average farm size was 53 horses (varying from 33 to 180 per farm). All the horses were warmblood sport and leisure horses, aged between 5 and 35 years (mean = 10.04 ±6.8 years). The gender split was 44% mares, 54% geldings, 2% stallions. All, or at least 70% of the adult horses (>5 years) in the facilities, were tested. In order to minimise the effect of familiarisation and so that it could be reasonably assumed that the horses did not see or hear the experimenter before being tested, the assessors tested horses in areas of the stable separated by some distance and never tested horses in adjoining boxes. To allow inter-observer reliability to be tested, assessors worked in pairs with one performing the tests and the other observing from a distance. They tested 90 horses kept in three horse facilities. To allow repeatability to be tested, one assessor repeated the assessment on all the horses 3 months after their initial assessment. To assess the on-farm feasibility, single box was chosen as at the moment it is assumed to be the most common housing systems for horses in Europe.

2.1.2. Behavioural tests

All the behavioural tests were performed on horses not restrained in their home box. A scoring system for each test was developed. Data was collected during regular working days and only healthy horses were tested between meals, at least 1 h before being put to work, in order to avoid confounding food motivation and any possible distractions. Assessors were requested not to talk or discuss their findings during on-farm assessments. Three behavioural tests were performed sequentially as follows (Fig. 1).

2.1.2.1. Avoidance distance test (AD). This test was developed by Waiblinger et al. (2003) and has already been used on cows (Windschnurer et al., 2008), sheep (Napolitano et al., 2011) and goats (Mattiello et al., 2010). The avoidance distance was performed without entering the box. The assessor waited for the horse to be attentive to his/her presence before approaching the animal. If the horse did not take any notice of the assessor, the assessor attracted its attention by clicking their tongue three times. Horses were approached by the test person in a standardised manner, starting from a distance of 2 m, walking at measured pace of one step per second, looking at the horse's chest without staring at it, and keeping the right arm raised 45° in front of the body, the back of the hand facing upwards (Fig. 1A). The test ended as soon as the horse showed any avoidance behaviour (e.g. moving away, turning its head away). Avoidance distance was estimated at the moment of horse withdrawal as the distance between the observer's hand and the animal's head with a resolution of 10 cm. A distance of 0 cm was assigned when the horse did not show any avoidance behaviour.

2.1.2.2. Voluntary animal approach test (VAA). This test was originally developed by Søndergaard and Halekoh (2003) on horses kept in paddocks and in this study it has been adapted to single housed horses. The voluntary animal approach was performed without entering the box. The assessor stood still outside the box with their hand on the door latch and their body at an angle of 45° from the box door (Fig. 1B). The maximum test time was 20 s. The latency time in seconds until the horse had its nose/mouth within a distance of 2 cm from the assessor's hand was recorded. Furthermore, the horse's reaction to the presence of the assessor was scored from 0 to 2:

- *Score 0*: The horse moved or turned its head away from the assessor, the horse showed signs of aggression (i.e., ears laid backwards, bite attempts).
- *Score 1*: The horse did not show any interest in the presence of the assessor and did not stop whatever it was doing (e.g. standing still in a corner of the box).
- *Score 2*: The horse was positively interested in the presence of the assessor and sniffed their hand.

2.1.2.3. Forced human approach test (FHA). This test, too, was adapted from the study done by Søndergaard and Halekoh (2003) on horses kept in a paddock. The assessor opened the box door and observed the horse's reaction for 5 s, then entered the box and approached the horse slowly at approximately one step per second with their hands by their side. If the horse stood still calmly, the assessor slowly raised their hand, touched the withers and moved their hand along the back of the subject. For safety reasons, one is advised to remain 30 cm away from the horse body (Fig. 1C). The horse's reaction was scored from 0 to 2 on the following scale:

- *Score 0*: The horse showed an aggressive behaviour (e.g. tried to bite or kick).
- *Score 1*: The horse moved away from the person as soon as he/she touched the withers.
- *Score 2*: The horse stood still calmly for the entire duration of the test or showed positive signs of interest (i.e., sniffing or staying in contact with the assessor).

2.2. Donkeys

2.2.1. Subjects

Eight donkey facilities, ranging in size from 5 to 60 animals per farm (mean 20 donkeys per farm), were visited twice at 3-month intervals. A total of 47 donkeys (2 geldings, 8 stallions, 37 jennies), aged between 1 and 18 years old (mean = 7.85±3.9 years) were assessed. The donkeys were of different breed and attitude – companion animals, rescue animals, assisted therapy, tourist trekking, and dairy production. All, or at least 70% of the adult donkeys in the facilities, were tested. Private practitioner veterinarians were consulted as to their knowledge of donkey farmers in the area that might have been interested in taking part in the research. All those involved participated voluntarily. To allow inter-observer reliability to be tested, assessors worked in pairs with one performing the tests and the other observing from a distance. They tested 47 donkeys kept on seven farms. To allow repeatability to be tested, one assessor repeated the assessment on all the donkeys 3 months after their initial assessment. To assess the on-farm feasibility, group housing was chosen as it is the most common housing system for keeping donkeys in Europe.

2.2.2. Behavioural tests

The methodology developed and validated by Burn et al. (2009) on working equines was adapted to reflect farming conditions of donkeys in Western European Countries to assess the reactions of donkeys towards humans. Data was collected during regular working days. Healthy donkeys were tested between meals, at least 1 h before being put to work, in order to avoid confounding food motivation and any possible distractions. The assessors worked in pairs and one carried out and scored the tests, while the other scored the reactions of the donkeys from a distance, without interfering with the test performance. For tests to remain consistent, assessors always wore the same type and colour of clothing during all farm visits, which included appropriate safety clothing (e.g. accident prevention shoes) to reduce the risk of injury. Assessors were requested not to talk or discuss their findings during on-farm assessments. Two behavioural tests were performed consecutively as follows (Fig. 2).

2.2.2.1. Avoidance distance test (AD). Each test donkey was chosen at random from a paddock and brought to a quiet

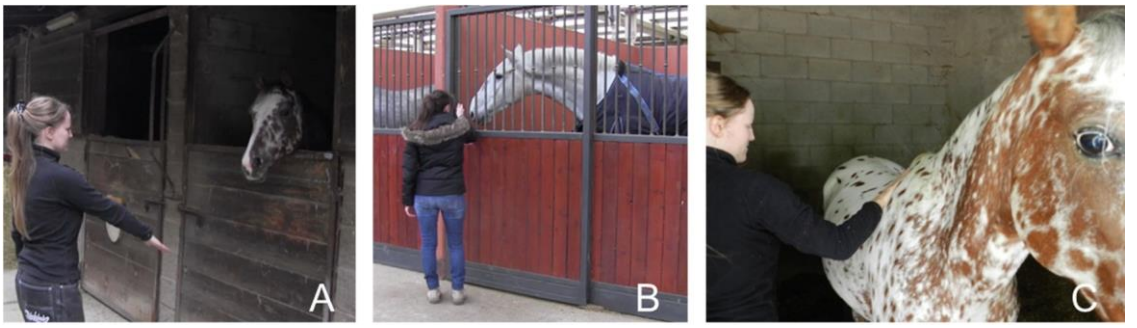

Fig. 1. Example of human–horse relationship tests: (A) avoidance distance; (B) voluntary animal approach; (C) forced human approach.

Fig. 2. Example of human–donkey relationship tests: (A) avoidance distance; (b) walking down side.

area of the farm by the stockperson, who restrained the donkey by holding the lead rope fixed to the head collar, allowing enough movement away from the approaching observer should the donkey want to retreat. Observers ensured that each test donkey was within visual and auditory reach of the other donkeys to prevent/minimise separation-related behaviours. Donkeys were approached by the observer in a standardised way, directly from the front, starting at a distance of 3 m. The test began when the donkey noticed he observer. If the donkey did not pay any attention to the observer, the observer would attract their attention by clicking their tongue three times. Once the donkey's attention was gained, the observer walked calmly at a measured pace (one step per second), with their arm raised 45° from their chest and the back of their hand facing upwards (Fig. 2A). The test ended as soon as the donkey exhibited any avoidance behaviour (e.g. moving away, turning the head), after which, the distance (cm) between the tip of the assessor's fingers and the head of the donkey was recorded (with a resolution of 10 cm). A distance of 0 cm was assigned when the donkey did not show any avoidance behaviour.

2.2.2.2. Walk down side test (WDS) and tail-tuck. To walk down the side of a donkey whilst touching it can be seen as a positive interaction with a human or a negative one, depending of course, on the perception of the donkey. Immediately following the AD test, the observer began by standing on the left side of the donkey, maintaining a distance of approximately 30 cm from its body and gently placed a hand on the donkey's withers. The observer proceeded to walk down the left side of the body towards the rear of donkey, stopping for a few seconds to note whether a tail-tuck is present, then continued to walk along the right side of the donkey towards the head, removing their hand and terminating the test at the withers (Fig. 2B). The observer recorded any signs of the donkey being alert to their presence during WDS. A two-point scale (0/1) was used to evaluate the donkey's reaction to the WDS test, as follows:

- *Score 0*: If the donkey showed any negative reaction to the movement of the observer during WDS test (ears flat back, trying to flee, attempting to kick, defecation).
- *Score 1*: If the donkey showed no interest or if the donkey showed any positive reaction to the movement of the observer during WDS test (remained calm and stationary, ear rotation towards observer, maintaining contact with observer, sniffing).

A two-point scale (0/1) was also used to evaluate tail-tuck:

- *Score 0*: If the donkey tucked in or clamped down its tail and/or tucked in or tensed its hindquarters during the WDS test (negative).
- *Score 1*: If the donkey did not tuck in or clamp down its tail and/or tuck in or tense its hindquarters during the WDS test (positive/neutral).

2.3. Training of assessors

Horse assessments were carried out by four, third-year undergraduate Animal Welfare students with a good knowledge of horse behaviour but no previous experience in assessing equine welfare. None of the assessors were acquainted with the horse facilities classification. Donkey assessments were conducted by four experimenters, all experienced in the field of animal welfare and applied ethology. Prior to training, all assessors familiarised themselves with the human–animal-relationship tests from relevant scientific literature. The assessors were trained to perform and score the tests by a senior veterinarian with over 10 years' experience in assessing horse welfare (silver standard). Assessor training procedures required 1 day per species and were comprised of two phases: theoretical training and practical assessment on-farm. During the theoretical phase, assessors were taught how to perform and score the tests with the use of a written training guide and pictorial and audio-visual presentations containing detailed explanations of each test, scoring systems and test videos. The assessors were also instructed to take the necessary precautions when handling the animals to minimise the risk of injuries. Exhibition of any behaviour that could compromise the animals' or the assessors' safety resulted in immediate termination of the test. On-farm training was then performed in order to achieve the skills necessary to perform and score the tests accurately on farm. In pairs with a senior veterinarian, assessors conducted live assessments of horses/donkeys until each assessor had performed a minimum of five consecutively accurate assessments. The training was considered completed when the assessors achieved ≥80% agreement with the silver standard, on both video and live scoring.

2.4. Data collection and statistical analysis

Data analysis was performed using the statistical software package IBM SPSS Statistics 21 (IBM Corp, 2012). Proportion of each score for all the behavioural tests was calculated, for the avoidance distance test measurements ">0 cm" was combined in the category "avoidance". We assessed the probability that horses showed negative reactions to the human–animal tests using a generalised linear mixed model (Wald-like test). The farm was considered as random-effect to account for multiple horses stabled in the same facility. Fixed effects of farm classification (excellent or sub-optimal) and observer (two levels) were also included in the model. Inter-observer reliability was initially assessed from videos by the observers that scored independently a random sample of 10 clips for each test. The reliability was calculated by means of percentage agreement that was always more than 80%. On-farm, inter-observer reliability was evaluated by comparing individual scores recorded by the two assessors independently and simultaneously. Prevalence indices for all the categories of the test scores were calculated. The prevalence index is the absolute difference between the agreed numbers for the two categories, divided by the total number of animals:

$$\text{Prevalence index} = \frac{|a-d|}{n}$$

where a is the number of agreed-upon animals in one of the categories and d is the number of agreed-upon animals for the other categories; n is the total number of possible agreements, i.e., the number of animals. A prevalence index of 0 indicates a completely balanced population, while an index of 1 would be a homogenous population in which only one of the categories is represented (Burn et al., 2009). Inter-observer reliability was analysed by calculating, according to the type of variable (categorical, scale), percentage agreement (the proportion of ratings where the raters agree), Kappa values, Kendall's coefficient of concordance (W) and interclass correlation coefficient (ICC). McNemar's and Wilcoxon's tests were performed in order to assess test repeatability by comparing results of the first and the second assessment. P values <0.05 were considered statistically significant.

3. Results

3.1. Horses

The behavioural tests used in this study proved to be feasible under field conditions in horses stabled in single boxes. No safety issues were encountered. All the owners showed good acceptability of the procedure adopted to test the animals. Total time required to perform all the tests on each horse varied from 90 to 180 s. Results of the three tests are reported in Table 4. In the avoidance distance test, most of the horses did not show any sign of avoidance (53.7%). In 38% of the subjects this test was not applicable (NA) because the horse did not take any notice of the assessor (e.g. the horse was looking out of the window or nibbling the floor looking for hay). In the VAA and FHA tests, positive reactions were displayed by most of the subjects whereas aggressiveness had the lowest prevalence in both tests. For safety reasons, horses that showed an aggressive reaction in the VAA test were not tested with the FHA (13.4%). voluntary animal approach and forced human approach tests were not applicable in very few cases, 18.8% and 13.4%, respectively.

3.1.1. Validity

There is no indication that the horses' reactions to humans during the tests varied as consequence of a random farm effect (negligible variance) or because of different observers performing the tests (avoidance distance GLMM, $P = 0.662$; voluntary animal approach GLMM, $P = 0.687$; forced human approach GLMM, $P = 0.065$). Only being housed in a facility classified as having a excellent or sub-optimal human–horse relationship significantly affected the reactions of horses to the tests (avoidance

Table 4
Behavioural responses expressed as proportion (%) among the 313 horses (11 horse facilities) and 49 donkeys (8 farms). In the avoidance distance test measurements >0 cm were combined in the category "avoidance".

Species[a]	Behavioural test	Response	Proportion of responses (%)
H	Avoidance distance (AD)	Avoidance (0)	8.3
		No avoidance (1)	53.7
		NA	38.0
H	Voluntary animal approach (VAA)	Aggressive (0)	4.2
		Not interested (1)	9.9
		Positive (2)	67.1
		NA	18.8
H	Forced human approach (FHA)	Aggressive (0)	8.3
		Avoidance (1)	9.2
		Neutral/positive (2)	64.9
		Not tested (aggressive at VAA)	4.2
		NA	13.4
D	Avoidance distance (AD)	Avoidance (0)	25.0
		No avoidance (1)	75.0
D	Walking down side (WDS)	Aggressive/Avoidance (0)	27.3
		Neutral/Positive (1)	72.7
D	Tail tuck	Presence (0)	13.6
		Absence (1)	86.4

[a] H = horse; D = donkey.

distance GLMM, P = 0.005; voluntary animal approach GLMM, P = 0.035; forced human approach GLMM, P = 0.01). In the avoidance distance test, only 1.4% of horses showed avoidance reaction on the farms classified as "excellent", whilst in "sub-optimal" farms 13.8% of the subjects avoided the assessor when approached (see Fig. 3). In the voluntary approach test, horses on "excellent" farms approached the assessor in 3 ± 3 s, whilst they needed 6 ± 7 s on "sub-optimal" farms. Furthermore, on farms classified as "sub-optimal" horses more often showed aggressive behaviour compared to "excellent" farms, with 6.9% and 0.7%, respectively (see Fig. 3). 10.3% of horses from "sub-optimal" farms reacted in an aggressive way when approached, whereas only 5.8% of subjects belonging to "excellent" farms showed the same behaviour (see Fig. 3).

3.1.2. Inter-observer reliability

On-farm, observers agreement of the avoidance distance test scoring was consistent (Cohen's Kappa = 0.89; percentage agreement = 93.3%; prevalence index = 0.52), voluntary animal approach test (latencies: Spearman's Rho = 0.85; behaviour: Cohen's Kappa = 0.40; percentage agreement = 67.7%; prevalence index = 0.57) and forced human approach test (Cohen's Kappa = 0.75; percentage agreement = 85.5%; prevalence index = 0.58).

3.1.3. Test re-test reliability

Repeatability of tests was good for all the tests and no significant differences were found between two repetitions at 3-month intervals (avoidance distance test: McNemar

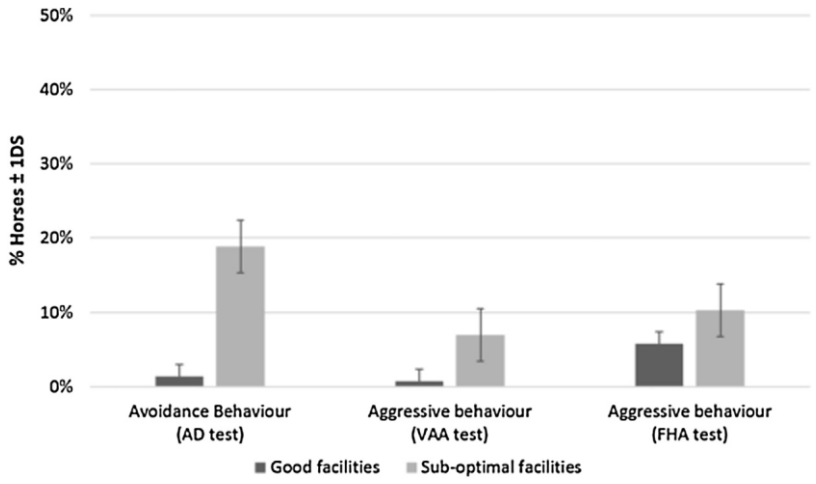

Fig. 3. Percentage of negative behaviours (avoidance and aggressive) showed during the three test (Mean ± 1DS).

$P = 0.61$; voluntary animal approach test and forced human approach test Wilcoxon $P = 0.60$ and $P = 0.56$, respectively).

3.2. Donkeys

The behavioural tests, developed for working equines, proved to be feasible for on-farm use as well. As all the donkeys were restrained, no safety issues were encountered. Most of the owners showed good acceptability of the procedure adopted to test the animals and in addition all owners were very willing to help in collecting and restraining the animals. Total time required to perform all the tests in each donkey varied from 60 to 90 s. Descriptive results of the three tests are presented in Table 4. In the avoidance distance test, most of the donkeys did not show any sign of avoidance (75.0%). In the walking down side test, aggressive/avoidance behaviour as well as tail tuck was displayed rarely, 27.3% and 13.6% respectively. Most of the donkeys showed neutral or positive reaction when approached by the assessor (72.7%). All the tests were always feasible, due to the fact that all the donkeys were restrained. However, test feasibility could be impaired in conditions where the donkeys cannot be restrained.

3.2.1. Inter-observer reliability

On-farm, observers agreement of the avoidance distance test scoring was consistent (Cohen's Kappa = 0.54; percentage agreement = 81%; prevalence index = 0.43), walking down side (Cohen's Kappa = 0.67; percentage agreement = 86%; prevalence index = 0.39), tail tuck (Cohen's Kappa = 0.83; percentage agreement = 95.3%; prevalence index = 0.70).

3.2.2. Test re-test reliability

Repeatability of tests was good for all the tests and no significant differences were found between two repetitions at 3-month intervals (avoidance distance test: McNemar $P = 1.00$; walking down side and tail tuck tests McNemar $P = 0.77$ and $P = 0.12$, respectively).

4. Discussion

One of the most important results of the present research was that all the behavioural tests performed made it possible to differentiate between horse facilities with excellent or sub-optimal human–animal relationship, as previously evaluated by official veterinarians. It is to be noted that the human–horse relationship tests were performed by assessors unaware of the farm classification; thus they were not biased in their evaluation. In spite of the fact that the reactions of horses were largely positive, those kept in facilities with "sub-optimal" relationship showed avoidance and aggressive behaviours more often when approached. They also needed more time to approach the assessor voluntarily. We propose that these measurements at farm level are sensitive and allow even relatively minor differences to be detected between farms. Human–animal relationship has an important impact on equine welfare as horses are handled on a daily basis. The presence of negative responses of equines to humans as an increased avoidance distance can be linked to the lack of confidence and/or fear of humans and suggests a variation in this relationship. Such negative responses can furthermore lead to flight reactions which can be dangerous for both horse and man. Our results are compatible with previous experimental studies showing that horses kept on farms where the management is focused on enhancing the relationship with horse and reducing their level of stress around humans, improve their reaction when they are facing a novel encounter with an unknown person (Fureix et al., 2009; Popescu and Diugan, 2013; Sankey et al., 2010; Søndergaard and Halekoh, 2003). Furthermore, the results of the present study provide evidence as to what extent the day to day behaviour of humans with horses can influence their reactions to simple on-farm behaviour tests. In fact, in facilities where horses are primarily managed by stable grooms, who possess an advanced and updated knowledge of welfare of horses, are inspected several times a day, and where positive interactions during routine handling procedures are enhanced, the horses showed more positive behaviours towards the assessors and they seemed to be more confident when approached by humans. In the present study, the possible confounding factors taken into account for assessing the validity of these behavioural tests were: the presence of pain, the underlying motivation of the horse and the context. Different studies showed that the presence of pain can fundamentally affect the behaviour of horses (Ashley et al., 2005; Fureix et al., 2010; Pritchett and Ulibarri, 2003). Indeed, one of the possible parameters taken into account when scoring pain in horses is how they react to the approach of an unknown human (Bussières et al., 2008; Van Loon et al., 2010). Therefore, in the present study only healthy horses were assessed. The underlying motivation of the horse can also affect how they react to human presence; for example, hungry horses can be more prone to approach a human if they have already identified the person as a source of food. To avoid this possible confounding factor we tested horses between meals, when no food was around. Horses are prey species and Søndergaard and Halekoh (2003) described that an unfamiliar and less spatial environment can affect their reaction to an unknown human during behavioural tests. It is for this reason that for the purposes of our research, all the horses were tested in their home box. It would be possible to argue that the avoidance distance could be influenced simply by habituation. That means, interacting more frequently with horses in their box may make them more used to the presence of humans in the box but not result in a general reduction of responsiveness towards humans in other situations. However, to some extent there might be an integrative effect of habituation and positive interaction as regards the effect of context. We report that in dairy cattle it was shown that the avoidance distance is not context specific, i.e., avoidance behaviour of animals under different test conditions is significantly related (Waiblinger et al., 2003; Windschnurer et al., 2008). Moreover, avoidance distances of dairy cows were shown to be related to milkers' behaviour during milking (Waiblinger et al., 2003). In the home environment, it is desirable to have horses that are easy to approach. Studies to date reported that breed and age of the horse can also play an important role in how they react to humans (Fureix et al., 2009;

Górecka-Bruzda et al., 2011; Søndergaard and Halekoh, 2003). In the present study all the subjects were adult warmblood sport and leisure horses. Further studies, with appropriate experimental designs, should verify whether differences in response to these tests, ascribable to the effect of breed, are narrower than the differences caused by the human–animal relationship. Given the findings of Søndergaard and Ladewig (2004) that young foals deprived of social contact with other horses may be more inclined to seek human contact, it would be interesting/useful to evaluate this effect on the reaction to the above mentioned behavioural tests. Søndergaard and Halekoh (2003) found that age can affect the reaction to humans in VAA and FHA tests, but in their study all the horses were young (less than 2 years). The same authors described that the way their horses responded during these tests from 12 months of age onwards was no different to that of a horse 24 months of age. Therefore, they concluded that "the effect of age in the human and animal approach tests may be an effect of familiarity to humans due to them being fed daily by people but it could also be an effect of the psychological development that horses undergo with age". We tested only adult horses, routinely handled on a daily basis and therefore both fully grown and used to human presence.

As for donkeys, in the present study most of the donkeys exhibited positive behaviour towards the assessor with no signs of avoidance during the AD test, no aggressive/avoidance reaction during the WDS, and no tail tuck display. Our results support the findings described by other authors in working donkeys (Burn et al., 2009; Popescu and Diugan, 2013; Pritchard et al., 2005), moreover, they show that, not only in a working environment, but also on-farm, the assessment of donkeys' reactions to an unknown human during standardised tests could prove useful in evaluating the quality of their relationship.

The accuracy of the assessment of the human–animal relationship is crucial, mostly when different observers in different countries perform this assessment as a decision support tool in animal welfare valuation. In the present study, agreement among different assessors was good (Cohen's Kappa > 0.60) in both species for most of the human–animal relationship tests. The voluntary animal approach test in horses and the avoidance distance in donkeys were the tests with the lowest agreement among observers. This may be due to the position of the observer while the assessor was performing the test. Indeed, position is crucial when observing the reaction of the animal, most especially because the reaction is sometimes rapid and not so obvious. When evaluating inter-observer reliability, scientists should always take into account the prevalence of the different scores in the population assessed. As already pointed out by Burn and colleagues (2009), "the prevalence of certain observations reduces the reliability ratings". In the present study, the prevalence was unbalanced – with a certain score more present compared to others – only for avoidance distance in horses and tail tuck in donkeys. Therefore, reliability for these tests was difficult to establish. This limitation was already described in similar studies carried out on working equines (Burn et al., 2009). For both horses and donkeys, all the tests performed to assess human–animal relationship were proved feasible in an on-farm environment. Little time was required to perform them (maximum of 3 min per animal), the step by step procedure guaranteed the safety of the animals and people involved, and they required minimal handling of the subjects. Among the horses' behavioural tests, the avoidance distance and voluntary animal approach tests can be performed without entering the box, and therefore can be carried out even though the owner is not available to help and without interfering in any way with the daily routine of the horses. For the same reasons, all the behavioural tests were well accepted by horse and donkey owners. Given the subjective nature of the scoring process, training should be considered as a key issue so that the achievement of consistent evaluation by different assessors is obtained. In the present study, training of assessors was both theoretical and practical. Working both through videos, as a class or an on-line exercise, and on-farm paired with a silver standard assessor on a purposely selected population (where the scores of the studied variable vary) proved vital in targeting good inter-observer reliability.

As far as mid-term repeatability is concerned, the results of the present study showed that the behavioural reactions of both horses and donkeys to unknown assessors did not change in a 3-month interval. These findings confirmed that the relationship between horses and humans, based on repeated interactions, reflects each subject's expectations during the encounters that follow. Thus, if the animals are kept in the same management conditions, the behavioural reaction to standardised human–animal relationship tests does not significantly change over time.

5. Conclusion

The avoidance distance, voluntary animal approach and the forced human approach tests proved useful to assess the human–horse relationship at farm level due to their feasibility, reliability, 3-months interval repeatability and ability to identify differences between horse facilities. Our results concerning the ability of these tests to reflect horses' previous experience with humans and their expectations on future interactions are encouraging. Moreover, our findings reveal that horses, kept in facilities where they are cared for by grooms with not only professional competence, but also advanced and updated knowledge on welfare of horses, react in a more positive way when approached by an unknown person. However, further studies are needed to determine to what extent responses to these tests are ascribable to the effect of breed and deprivation of social contact with other horses. We suggest that future work should investigate specifically these factors and assess the reliability of these tests on group housed horses. Given the relatively higher proportion of horses where the avoidance distance test was not applicable and the unbalanced prevalence of certain scores, this test should be considered as the preferred choice only as a first step in the assessment of the human–horse relationship and when the forced human approach test is not feasible.

As for donkeys, our findings show that the avoidance distance, walking down side and tail tuck tests are feasible

and reliable measurements in a typical Western Countries farm environment. The prevalence of tail tuck was unbalanced; therefore, we suggest taking into account this possible limitation during the training of assessors.

In general our results support the findings in other species (Waiblinger et al., 2006, 2003; Windschnureremail et al., 2009) that good human–animal relationship can be identified through looking at the reaction of the animal in a standardised interaction with an unknown assessor and underlines the importance of human behaviour in the interaction. Further research is needed to increase the number of equine farms, under different husbandry conditions and with different breeds.

Conflict of interest

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

Acknowledgements

The authors wish to thank the EU VII Framework programme (FP7-KBBE-2010-4) for financing the Animal Welfare Indicators (AWIN) project. The authors would like to thank Chantal Bonaita, Sara Pedretti, Alessandra Guzzeloni, Elisa Govoni, Alessandra Meazza, for their help in assessing the animals; we also acknowledge Kirk Ford for his extensive and professional revisions of language and structure.

References

Acock, A.C., 2008. A Gentle Introduction to Stata. Stata Press, College Station, USA.

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 Vet. J. 37, 565–575.

Burn, C.C., Dennison, T.L., Whay, H.R., 2010. Relationships between behaviour and health in working horses, donkeys, and mules in developing countries. Appl. Anim. Behav. Sci. 126, 109–118.

Burn, C.C., Pritchard, J.C., Whay, H.R., 2009. Observer reliability for working equine welfare assessment: problems with high prevalences of certain results. Anim. Welf. 18, 177–187.

Bussières, G., Jacques, C., Lainay, O., Beauchamp, G., Leblond, A., Cadoré, J.-L., Desmaizières, L.-M., Cuvelliez, S.G., Troncy, E., Desmaizie, L., 2008. Development of a composite orthopaedic pain scale in horses. Res. Vet. Sci. 85, 294–306.

Chamove, A., Crawley-Hartrick, O., Stafford, K., 2002. Horse reactions to human attitudes and behaviour. Anthrozoos: Multidiscip. J. Interact. People Anim. 15, 323–331.

Cronbach, L.J., Meehl, P.E., 1955. Construct validity for psycological tests. Psycol. Bull. 52, 283–302.

Estep, D., Hetts, S., 1992. Interactions relationships and bonds: the conceptual basis for scientist–animal relations. In: Davis, H., Balfour, D. (Eds.), The Inevitable Bond: Examining Scientist–Animal Interactions. Cambridge University Press, Cambridge, UK, pp. 6–26.

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

Fureix, C., Pagès, M., Bon, R., Lassalle, J.-M., Kuntz, P., Gonzalez, G., 2009. A preliminary study of the effects of handling type on horses' emotional reactivity and the human–horse relationship. Behav. Process. 82, 202–210.

Górecka-Bruzda, A., Jastrzębska, E., Sosnowska, Z., Jaworski, Z., Jezierski, T., Chruszczewski, M.H., 2011. Reactivity to humans and fearfulness tests: field validation in polish cold blood horses. Appl. Anim. Behav. Sci. 133, 207–215.

Hama, H., Yogo, M., Metsyyama, Y., 1996. Effects of stroking horses on both humans' and horses' heart rate responses. Jpn. Psychol. Res. 38, 66–73.

Hausberger, M., Muller, C., 2002. A brief note on some possible factors involved in the reactions of horses to humans. Appl. Anim. Behav. Sci. 76, 339–344.

Hausberger, M., Roche, H., Henry, S., Visser, E.K., 2008. A review of the human–horse relationship. Appl. Anim. Behav. Sci.

Hemsworth, P., Coleman, G., 1998. Human–Livestock Interactions: The Stockperson and the Productivity and Welfare of Intensively Farmed Animals. CAB International, Oxon, UK.

Hemsworth, P., 2003. Human–animal interactions in livestock production. Appl. Anim. Behav. Sci. 81, 185–198.

Hemsworth, P.H., Barnett, J.L., Coleman, G.J., 1993. The human–animal relationship in agriculture and its consequences for the animal. Anim. Welf. 2, 33–51.

Henry, S., Richard-Yris, M.A., Hausberger, M., 2006. Influence of various early human–foal interferences on subsequent human–foal relationship. Dev. Psychobiol. 48, 712–718.

IBM Corp, 2012. IBM SPSS Statistics for Windows.

Jezierski, T.A., Jaworski, Z., Górecka-Bruzda, A., 1999. Effects of handling on behaviour and heart rate in Konik horses: comparison of stable and forest reared youngstock. Appl. Anim. Behav. Sci. 62, 1–11.

Lansade, L., Bertrand, M., Boivin, X., Bouissou, M.F., 2004. Effects of handling at weaning on manageability and reactivity of foals. Appl. Anim. Behav. Sci. 87, 131–149.

Lansade, L., Bouissou, M., Erhard, H., 2008. Fearfulness in horses: a temperament trait stable across time and situations. Appl. Anim. Behav. Sci. 115, 182–200.

Lesimple, C., Fureix, C., LeScolan, N., Richard-Yris, M.A., Hausberger, M., 2011. Housing conditions and breed are associated with emotionality and cognitive abilities in riding school horses. Appl. Anim. Behav. Sci. 129, 92–99.

Ligout, S., Bouissou, M.-F., Boivin, X., 2008. Comparison of the effects of two different handling methods on the subsequent behaviour of Anglo-Arabian foals toward humans and handling. Appl. Anim. Behav. Sci. 113, 175–188.

Maros, K., Boross, B., Kubinyi, E., 2010. Approach and follow behaviour – possible indicators of the human – horse relationship. Interact. Stud. 11, 410–427.

Mattiello, S., Battini, M., Andreoli, E., Minero, M., Barbieri, S., Canali, E., 2010. Avoidance distance test in goats: a comparison with its application in cows. Small Rumin. Res. 91, 215–218.

McCann, J., Heird, J., Bell, R., Lutherer, L., 1988. Normal and more highly reactive horses. Part II: the effect of handling and reserpine on the cardiac response to stimuli. Appl. Anim. Behav. Sci. 19, 215–226.

Mendl, M., Burman, O.H.P., Paul, E.S., 2010. An integrative and functional framework for the study of animal emotion and mood. Proc. R. Soc. B 277, 2895–2904.

Napolitano, F., De Rosa, G., Girolami, A., Scavone, M., Braghieri, A., 2011. Avoidance distance in sheep: test–retest reliability and relationship with stockmen attitude. Small Rumin. Res. 99, 81–86.

Popescu, S., Diugan, E.-A.A., 2013. The relationship between behavioral and other welfare indicators of working horses. J. Equine Vet. Sci. 33, 1–12.

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. Prev. Vet. Med. 69, 265–283.

Pritchett, L., Ulibarri, C., 2003. Identification of potential physiological and behavioral indicators of postoperative pain in horses after exploratory celiotomy for colic. Appl. Anim. Behav. Sci. 80, 31–43.

Rushen, J., Taylor, A.A., Passille A.M., De, 1999. Domestic animals' fear of humans and its effect on their welfare. Appl. Anim. Behav. Sci. 65, 285–303.

Sankey, C., Richard-Yris, M.-A., Leroy, H., Henry, S., Hausberger, M., 2010. Positive interactions lead to lasting positive memories in horses, *Equus caballus*. Anim. Behav. 79, 869–875.

Seaman, S.C., Davidson, H.P.B., Waran, N.K., 2002. How reliable is temperament assessment in the domestic horse (Equus caballus)? Appl. Anim. Behav. Sci. 78, 175–191.

Søndergaard, E., Halekoh, U., 2003. Young horses' reactions to humans in relation to handling and social environment. Appl. Anim. Behav. Sci. 84, 265–280.

Søndergaard, E., Ladewig, J., 2004. Group housing exerts a positive effect on the behaviour of young horses during training. Appl. Anim. Behav. Sci. 87, 105–118.

Van de Mortel, T., 2008. Faking it: social desirability response bias in self-report research. Aust. J. Adv. Nurs. 25, 40–48.

Van Loon, J., Back, W., Hellebrekers, L.J., van Weeren, P.R., Loon, J.P.A.M., Van Weeren, V., 2010. Application of a composite pain scale to objectively monitor horses with somatic and visceral pain under hospital conditions. J. Equine Vet. Sci. 30, 641–649.

Visser, E.K., van Reenen, C.G., van der Werf, J.T.N., Schilder, M.B.H., Knaap, J.H., Barneveld, A., Blokhuis, H.J., 2002. Heart rate and heart rate variability during a novel object test and a handling test in young horses. Physiol. Behav. 76, 289–296.

Visser, E.K.K., van Reenen, C.G., Hopster, H.B.H., Schilder, J.H., Knaap, J.H., Barneveld, A., Blokhuis, H.J., 2001. Quantifying aspects of young horses' temperament: consistency of behavioural variables. Appl. Anim. Behav. Sci. 74, 241–258.

Waiblinger, S., Boivin, X., Pedersen, V., Tosi, M.-V., Janczak, A.M., Visser, E.K., Jones, R.B., 2006. Assessing the human–animal relationship in farmed species: a critical review. Appl. Anim. Behav. Sci. 101, 185–242.

Waiblinger, S., Menke, C., Fölsch, D., 2003. Influences on the avoidance and approach behaviour of dairy cows towards humans on 35 farms. Appl. Anim. Behav. Sci. 84, 23–39.

Windschnurer, I., Schmied, C., Boivin, X., Waiblinger, S., 2008. Reliability and inter-test relationship of tests for on-farm assessment of dairy cows' relationship to humans. Appl. Anim. Behav. Sci. 114, 37–53.

Windschnureremail, I., Boivin, X., Waiblinger, S., 2009. Reliability of an avoidance distance test for the assessment of animals' responsiveness to humans and a preliminary investigation of its association with farmers' attitudes on bull fattening farms. Appl. Anim. Behav. Sci. 117, 117–127.

Wolff, A., Hausberger, M., Le Scolan, N., 1997. Experimental tests to assess emotionality in horses. Behav. Proc. 40, 209–221.