The Old School, Brewhouse Hill, Wheathampstead,
Hertfordshire AL4 8AN, UK
www.ufaw.org.uk

Animal Welfare 2022, 31: 455-466
ISSN 0962-7286
doi: 10.7120/09627286.31.3.008

Horse welfare: A joint assessment of four categories of behavioural indicators using the AWIN protocol, scan sampling and surveys

A Ruet*[†], C Arnould[†], J Lemarchand[†], C Parias[†], N Mach[‡], M-P Moisan[§], A Foury[§], C Briant[†] and L Lansade[†]

[†] INRAe, UMR 85 PRC, CNRS, UMR 7247, IFCE, University of Tours, 37380, Nouzilly, France
[‡] INRAe, UMR 1313 GABI, AgroParisTech, University of Paris-Saclay, 78352, Jouy-en-Josas, France
[§] University of Bordeaux, INRAe, Bordeaux INP, NutriNeuro, UMR 1286, 33076 Bordeaux, France
* Contact for correspondence: alice.ruet@ifce.fr

Abstract

*Domesticated horses (*Equus caballus*) can be exposed to a compromised welfare state and detecting a deterioration in welfare is essential to modify the animals' living conditions appropriately. This study focused on four categories of behavioural indicators, as markers of poor welfare: stereotypies, aggressiveness towards humans, unresponsiveness to the environment and hypervigilance. In the scientific literature, at least three assessment methods can be used to evaluate them: the Animal Welfare Indicators (AWIN) protocol, behavioural observations using scans and surveys. The question remains as to whether all these three methods allow an effective assessment of the four categories of behavioural indicators. To address this issue, the repeatability at a three-month interval and convergent validity of each measure (correlations between methods) were investigated on 202 horses housed in loose boxes. Overall, the repeatability and convergent validity were limited, highlighting the difficulty in assessing these indicators in horses. However, stereotypies and aggressiveness measures showed higher repeatability and convergent validity than those of unresponsiveness to the environment and hypervigilance. Behavioural observations using scans enabled the four categories of behavioural indicators to be detected more effectively. Suggestions of improvements are proposed for one-off measures such as those performed with the AWIN protocol. Regardless of the assessment method, very limited correlations were observed between the four categories of behavioural indicators, suggesting that they should all be included in a set of indicators used to assess the welfare state of horses, in conjunction with physiological and health measures.*

Keywords: *aggressiveness, animal-based measure, animal welfare, equine welfare assessment, individual housing, stereotypies*

Introduction

Domesticated horses (*Equus caballus*) can undergo exposure to strong external constraints that may compromise their welfare state. Indeed, the most common housing system remains the stable with loose boxes (eg Hockenhull & Creighton 2015), even although this system can prevent animals from expressing natural behaviours (eg social interactions with conspecifics: Søndergaard *et al* 2011; free movement: Houpt *et al* 2001; grazing: Harris 1999). In addition, interactions with humans, if regularly triggering fear or pain during daily care management and generalised horse-related activities, also constitute risk factors for welfare (Ödberg 1987; Hausberger *et al* 2008, Baragli *et al* 2015). Welfare can be defined as "a state of complete mental and physical health, where the animal is in harmony with its environment" (Hughes 1976), and is dependent in part on the individual's subjective experience (Dawkins 1990; Veissier & Boissy 2007). The multidimensional concept of welfare requires assessment criteria that are primarily animal-based measures (Botreau *et al* 2007). Among them, behavioural indicators are particularly relevant because they can be used to observe how animals deal with their environment and to infer the individual subjective experience (Dellmeier 1989; Dawkins 2003). In recent years, a wide range of behavioural indicators reflecting a compromised welfare state in the living environment have been reported in numerous studies (Mellor 2016). The present study focuses on behavioural indicators identified in the equine literature as especially relevant to assess the welfare state of this species, according to the negative internal conditions that are likely to be inferred from their expression. They have been grouped into four categories: stereotypies, aggressiveness towards humans, unresponsiveness to the environment and hypervigilance.

The first category concerns stereotypies, defined as "repetitive unvarying behaviours without apparent function" and

expressed by individuals in many species living under suboptimal conditions of captivity (Mason 1991; Mason & Latham 2004). In horses, they appear to be linked to physiological indicators of stress (Omidi *et al* 2018), to afflictions of the immune system (Alberghina *et al* 2015) and sometimes to health impairments (eg colic; Curtis *et al* 2019). They can result from a chronic experience of negative affective states caused by the inability to engage in natural behaviours (Sarrafchi & Blokhuis 2013). Numerous risk factors leading to the development and expression of stereotypies have been identified such as social isolation (eg Mills & Davenport 2002), confinement (eg Bachmann *et al* 2003), high starch cereal diets (eg Waters *et al* 2002) and riding (eg Christie *et al* 2006). As in many other species, several behaviours can be observed in stereotypic horses such as repetitive oral (Wickens & Heleski 2010) or motor activity (Ninomiya *et al* 2007).

The second category refers to aggressiveness towards humans which, in horses, includes a series of behaviours ranging from simple threats (looking with ears pinned backward) to more extreme actions (approaching with ears backward and mouth open or turning hindquarters and sometimes raising a leg) or even physical attacks (McGreevy 2004). Evidence shows that these behavioural indicators could emerge due to physical pain (injuries: Popescu & Diugan 2013; chronic back pain: Fureix *et al* 2010) or can reflect a long-lasting negative affective state, inferred from the observation of a more pessimistic judgement bias in the individual horse (Henry *et al* 2017). Aggressiveness towards humans tends to increase with social isolation (eg Normando *et al* 2011) and dietary deprivations (eg Hockenhull & Creighton 2014; Ribeiro *et al* 2019) and can represent a real threat to human safety (Thomas *et al* 2006; Yim *et al* 2007).

The third category concerns unresponsiveness to the environment, displayed in horses as an immobile posture when standing alone in their loose box, with the horse's neck and back at similar heights (a nape/withers back angle of approximately 180°), a fixed gaze and static ear and head positions, as described in detail by Fureix *et al* (2012). This atypical posture, called 'withdrawn posture', is easily recognisable from one horse to another and presents a major difference with the standing resting posture in the opening and fixity of the eyes, with eyelids that rarely blink and which do not become droopy. In addition, the horse does not react physically by turning the head or raising the neck in response to the usual sensory stimuli in the environment. The expression of the 'withdrawn posture' is related to a drop in plasma cortisol (Fureix *et al* 2012), the expression of anhedonia (Fureix *et al* 2015) and impaired selective attention (Rochais *et al* 2016). This behavioural and physiological profile shows strong similarities with some aspects of depressive states in animal models and humans (Post & Warden 2018; Hao *et al* 2019). Other authors looked at so-called 'depressive' states in horses and described the animals as apathetic (eg Pritchard *et al* 2005).

Finally, the fourth category refers to hypervigilance, defined by Richards *et al* (2014) as increased alertness for threats by excessive scanning of the environment, which can be observed by recurrent vigilant behaviours (Wermes *et al* 2018). Hypervigilance could be motivated by negative affective states such as anxiety (Ohl *et al* 2008; Sylvers *et al* 2011; Harro 2018). Although vigilant behaviours have an adaptive value by allowing individuals to detect potential threats in the environment, hypervigilance can sometimes become pathological and compromise the welfare state of the animals (eg Ohl *et al* 2008; Salomons *et al* 2009). Vigilant behaviours in horses are expressed through the alert/alarm posture (ie head held high, rigid body, ears and eyes fixed in the direction of a stimulus for a variable duration: Young *et al* 2012; Statton *et al* 2014; Wathan & McComb 2014). When these vigilant behaviours persist over time and situations, horses are considered hypervigilant and commonly described as alarmed or anxious.

Several methods could be used to assess the four aforementioned categories of behavioural indicators which indicate a compromised welfare state of horses in their living environment. Three of them appear particularly relevant. First, the Animal Welfare Indicators (AWIN) protocol was developed for horses in 2015 (AWIN 2015), based on the concept of the Welfare Quality® protocol, which is a functional implementation of the original Five Freedoms paradigm of animal welfare (Webster 2005). This method allows a rapid assessment of a large number of animals under field conditions and contains animal-based measures of both behavioural and health criteria (Dalla Costa *et al* 2014). To ensure feasibility (Dalla Costa *et al* 2015; Czycholl *et al* 2017), behavioural measures are based on short observations, tests and subjective assessments using the Qualitative Behaviour Assessment (QBA; Wemelsfelder 2007). The latter allows an observer to quantify the main dimensions of the animal's affective state over a given period of time, using a list of 13 subjective descriptors such as 'friendly' or 'relaxed' (Minero *et al* 2016). Second, short but repeated behavioural observations using scans (Altmann 1974) can be used to detect and quantify the expression of the four categories of behavioural indicators studied herein over a specific period of time (stereotypies: Fureix *et al* 2011, 2015; Houpt *et al* 2001; Kwiatkowska-Stenzel *et al* 2016; Pessoa *et al* 2016; aggressiveness towards humans: Ruet *et al* 2019; unresponsiveness to the environment through the 'withdrawn posture' using scans lasting several seconds instead of one: Fureix *et al* 2015; Rochais *et al* 2016; hypervigilance with quantification of the repetition of alert postures: Heleski *et al* 2002; Pessoa *et al* 2016). The scan-sampling method has the advantage of multiplying the observations throughout the day and over time while remaining feasible on a large number of horses. It maximises the chance of observing each behavioural indicator and allows assessment of how often each one is expressed. Third, collecting the opinion of the usual caretakers of the animals through surveys is sometimes used to

https://doi.org/10.7120/09627286.31.3.008 Published online by Cambridge University Press

assess stereotypies (Bachmann & Stauffacher 2002; Normando *et al* 2002; Tadich *et al* 2012), aggressiveness towards humans (Normando *et al* 2002) and hypervigilance in working horses (Pessoa *et al* 2016). It provides an integrative animal point of view, although subjective (see comparisons between caretakers' surveys and ethological observations of stereotypies: Lesimple & Hausberger 2014).

These three methods are suitable for two different contexts of application: either a one-off welfare inspection by an external assessor, for example, for certification purposes, in the case of the AWIN protocol and surveys, or regular self-monitoring, for example, by the horses' handlers who are with the animals over a longer period of time. These fundamental differences in the type of measures and the contexts of application raise the question of whether all three methods lead to an effective assessment of the four categories of behavioural indicators. To address this issue, it is interesting to investigate the reliability and validity of each measure. Reliability includes, in part, the repeatability (or test-retest reliability), which corresponds to the chance that a result will be identical if the measure is repeated (Meagher 2009; Temple *et al* 2013). This central criterion addresses the major issue of variability of measures over time and is particularly relevant in the context of a one-off welfare inspection. Poor reliability would indicate an ineffective ability of the measure to properly detect a compromised welfare state in animals, because it would not systematically identify the category of behavioural indicators of interest. Validity characterises the ability of a measure to evaluate what it is supposed to assess (Czycholl *et al* 2016). Several kinds of validity criteria exist and convergent validity is one such example, which is defined as the degree of correlation between several measures intended to assess the same category of behavioural indicator (Meagher 2009). Low validity would question the relevance of the measures to assess the category of behavioural indicators of interest and could indicate the need for improvement. Finally, looking at the relationships among the four categories of behavioural indicators could help to refine future welfare assessments.

In this study, the three methods mentioned above were tested on 202 horses housed in loose boxes at the same riding school. The aims of this study were to investigate: (i) the repeatability of the measures obtained with the AWIN protocol and scans, at an interval of three months; (ii) the convergent validity of measures evaluating a same category of behavioural indicators (for example, are the unresponsiveness measures correlated among the AWIN protocol, scans and survey assessment); and (iii) the relationships among the four categories of behavioural indicators (for example, is aggressiveness correlated with stereotypies or unresponsiveness to the environment).

Materials and methods

Study animals and management conditions

This study involved 202 sport horses with a mean (± SD) age of 10.5 (± 3.48) years and stabled at the same riding school for at least six months prior to the beginning of the study (seven stallions, 140 geldings, 55 mares). All horses were housed in single loose boxes approximately 9 m² in size and were kept on straw (n = 143), wood-shavings (n = 32) or wood-pellets (n = 27). They all had visual contact with other horses and some also had reduced tactile contact through a grilled window in the wall between two consecutive loose boxes (n = 99). Horses were fed concentrated feed three or four times per day and received approximately 9 kg of hay divided into two meals per day. Water was available *ad libitum* via automatic drinkers. All horses were trained for 90 (± 25) min per day on average, at least six days a week, for sports purposes. Among the 202 horses, only 13 were regularly released into paddocks for free exercise for 65 (± 47) min per day (nine alone and four in pairs).

Assessment of the four categories of behavioural indicators using three methods

All measures were taken by a single observer experienced in equine ethology (PhD in ethology) and unfamiliar to the horses. The three methods were performed over a period of three months (period one). To assess the repeatability of the measures, the AWIN protocol and scans were carried out a second time over the subsequent three-month period (period two). Due to a lack of availability of animal caretakers, only one survey could be completed per horse and the repeatability of this method could not be investigated. For each horse, the measures (AWIN protocol and scans) in both periods were performed at an interval of three months. The definitions and types of each measure assumed to assess the four categories of behavioural indicators are summarised in Table S1.

AWIN protocol

The complete AWIN protocol (AWIN 2015) was conducted, but only the measures assumed to assess the four categories of behavioural indicators studied were analysed. Due to the potential influence of feeding time on the expression of behaviours such as stereotypies and aggressive behaviours towards humans (Hockenhull & Creighton 2014), it was decided to standardise the implementation by carrying out the protocol at least 1 h before or after the horses received concentrated feed or forage.

First, a 1-min observation of the horse from outside the box was carried out to assess stereotypies. Horses were binary recorded as 'not expressed' or 'expressed' if one or more of these behaviours were observed.

Then, a human-animal relationship test consisting of three stages was carried out to assess aggressiveness towards humans: the observer started 2.5 m from the loose-box door

and walked calmly towards the horse with the right arm held forward at a 45° angle from the chest. The observer then opened the door, entered the loose box and tried with their left hand to touch the horse on its neck and along its back. For each horse, aggressiveness was binary coded as 'expressed' or 'not expressed' based on whether aggressive behaviours were observed.

Finally, a Qualitative Behaviour Assessment (QBA) was carried out to assess aggressiveness towards humans, unresponsiveness to the environment and hypervigilance. As defined by the AWIN protocol, the horse was observed from outside the loose box for 30 s and then the observer approached slowly and performed a manual imitation of allogrooming behaviour at the withers for another 30 s. At the end of this sequence, the descriptors were scored using visual linear scales measuring 125 mm: each behaviour was marked between the minimum (score 0) if 'the behaviour was completely absent' and the maximum if 'the behaviour was mainly present' (score 125). The exact score was measured in mm on the scale and reported on a total of 100. The descriptors used for this study were 'aggressive' defined as "hostile, attacking, wants to fight/attack, dominance, defensive, aggression", 'apathetic' defined as "having or showing little or no emotion, disinterested, indifferent, isolated, depressed, unresponsive, motionless" and 'alarmed' defined as "worried/tense, apprehensive, jumpy, nervous, watchful, on guard against a possible threat/danger" (Minero *et al* 2018).

Scans

Repeated behavioural observations were performed by scan sampling to detect and quantify the expression of the four categories of behavioural indicators over time and situations. The observer walked regularly in front of the loose boxes at a distance of at least 1.5 m from the door, making as little noise as possible. Each horse was observed for 5 s, and then the observer recorded whether the animal expressed one of the behavioural indicators in the four categories (Table S1). Per period, each horse was observed on 25 non-consecutive days, with five scans per day during a 90-min observation session, ensuring that the sessions were equally distributed across the time of day (0900 to 1030h, 1030 to 1200h, 1200 to 1330h, 1330 to 1500h and 1500 to 1630h). The average number of total scans analysed per subject was 101 (± 11) for the first period and 97 (± 18) for the second (variations in the number of scans resulted from the absence of the horse or the presence of the caretaker in the loose box at the time of the observation). The frequencies of each behavioural indicator were calculated from the total number of observations per horse.

Survey

To obtain an integrative view of each animal, caretakers were asked about their horse's behaviour using a single close-ended questionnaire. The survey was completed by a total of 38 caretakers (24 men and 14 women), each of whom took daily care of a specific group of horses for at least three months (5.3 [± 2.7] horses per caretaker). The terms used in the survey were chosen to be readily understandable by the caretakers while remaining scientific (see Table S1 in Supplementary material). Stereotypies were binary recorded (expressed/not expressed). Aggressiveness towards humans was assessed using two scales of scores from 0 to 2 (0: never aggressive; 1: sometimes aggressive; 2: always aggressive); the first during grooming and the second while tacking up (putting the saddle and bridle on). The mean of the two scores was calculated for each horse. Unresponsiveness to the environment and hypervigilance were scored using scales from 0 (never expressed) to 10 (always expressed).

Statistical analysis

To investigate repeatability of the measures between the two periods (periods one and two; aim [i]), Spearman's rho correlations were calculated for continuous but non-normally distributed data and unweighted Cohen's kappa coefficients were used for binary data. To investigate convergent validity among the three methods (aim [ii]), as well as the relationships among the four categories of behavioural indicators (aim [iii]), Spearman's rho correlations between continuous but non-normally distributed data, point-biserial correlations between one continuous and one binary data and Fisher's exact tests between two binary data were performed. The variables measured twice during this study (AWIN protocol and scans) were analysed as the sums of the two periods to reduce the number of analyses presented. Correlation coefficients could be interpreted as follows: < 0.30 as negligible, 0.30–0.50 as low, 0.50–0.70 as moderate and > 0.70 as high (Hinkle *et al* 2003). However, in this study, 0.50 was considered as the cut-off point for interpreting coefficients. Cohen's kappa could be interpreted as follows: < 0.40 as minimal, 0.40–0.60 as moderate, 0.60–0.80 as substantial and > 0.80 as strong (McHugh 2012). Again, in this study, the cut-off point for interpreting coefficients were set at 0.50.

All statistical analyses were performed with R software (version 3.3.2, R Development Core Team 2020) and the package stats v3.5.2, except for Cohen's kappa calculations, which were carried out with irr v0.84.1. Significance levels were set at $P \leq 0.05$.

Ethical statement

This study was conducted in compliance with the ethical policy of the International Society for Applied Ethology and approved by the ethics committee of Val de Loire (2019012211274697.V4-18939).

Results

Repeatability

AWIN protocol

Each category of behavioural indicators assessed with the AWIN protocol presented significant coefficients (Spearman's rho or Cohen's kappa) between the two measurements periods ($P < 0.01$ in all cases), but all the values were below 0.50. The measure of stereotypies showed a kappa coefficient of $K^{AWIN\ 1–AWIN\ 2} = 0.12$. The values of the

https://doi.org/10.7120/09627286.31.3.008 Published online by Cambridge University Press

kappa coefficients were higher for the two measures of aggressiveness ($K^{AWIN\ 1–AWIN\ 2} = 0.35$; $K^{AWIN\_QBA\ 1–AWIN\_QBA\ 2} = 0.44$). The term 'apathetic' in the AWIN^{QBA} (category of unresponsiveness to the environment) presented a kappa coefficient of $K^{AWIN\_QBA\ 1–AWIN\_QBA\ 2} = 0.20$. The term 'alarmed' in the AWIN^{QBA} (category of hypervigilance) presented a Spearman's rho of $r_s^{AWIN\_QBA} = 0.37$.

Scans

Each behavioural indicator assessed by scans was significantly correlated between the two periods of measurements ($P < 0.01$ in all cases). Aggressiveness was the most correlated (Spearman's rho: $r_s^{scans} = 0.51$), followed by stereotypies ($r_s^{scans} = 0.45$), unresponsiveness to the environment ($r_s^{scans} = 0.29$) and hypervigilance ($r_s^{scans} = 0.17$). Only aggressiveness towards humans presented a correlation value greater than 0.50.

Convergent validity

Stereotypies

The assessment of stereotypies using the scans and survey showed the highest significant correlation, followed by the AWIN protocol and scans measures. The AWIN protocol and survey measures were not significantly related according to Fisher's exact test. Only the correlation coefficient between the assessment of stereotypies using the scans and survey was greater than 0.50 (point-biserial correlation: $r_{pb} = 0.63$; Table 2).

Aggressiveness towards humans

Aggressiveness was the only category of behavioural indicators for which all measures were significantly correlated with each other, but to varying degrees. The two measures of aggressiveness in the AWIN protocol (human-animal relationship test and the QBA) appeared strongly related according to Fisher's exact test. The assessment using AWIN^{QBA} and scans presented the highest correlation value, followed by AWIN^{QBA} and the survey, AWIN and scans, and AWIN and the survey at relatively similar values. Finally, scans and survey measures were the least correlated. None of the correlation coefficients were greater than 0.50 (Table 3).

Unresponsiveness to the environment

The assessment of unresponsiveness to the environment using the scans and survey showed a significant correlation, but other measures were not significantly correlated. The significant correlation value was below 0.50 (Table 4).

Hypervigilance

The three measures of hypervigilance were not significantly correlated with each other (Table 5).

Relationships among the four categories of behavioural indicators within each method

AWIN protocol

As expected, the two measures of aggressiveness were strongly related (AWIN and AWIN^{QBA}; Fisher's exact test: $\chi^2 = 28.2$; $P < 0.001$). A significant correlation was found between aggressiveness during the human-animal relationship test and the 'alarmed' descriptor (category of hypervigilance) assessed in the AWIN^{QBA} (point-biserial correlation: $r_{pb} = 0.27$; $P < 0.001$), although the value was below 0.50. Aggressive behaviours during the human-animal relationship test (AWIN) appeared to be related to stereotypies (Fisher's exact test: $\chi^2 = 3.59$; $P < 0.05$).

Table 2 Relationships between the three measures of stereotypies among the three methods (Animal Welfare Indicators [AWIN] protocol, scans and survey).

	Stereotypies (AWIN)	Stereotypies (scans)	Stereotypies (survey)
Stereotypies (AWIN)	1	0.16*[1]	ns[2]
Stereotypies (scans)		1	0.63***[1]
Stereotypies (survey)			1

[1] Point-biserial correlations (r_{pb});
[2] Fisher's exact test;
* $P \leq 0.05$; *** $P \leq 0.001$.

Scans

No significant positive correlations were observed between the four scans' measures. A significant negative correlation was found between unresponsiveness to the environment and aggressiveness towards humans (Spearman's rho: $r_s = –0.15$; $P < 0.05$), although the correlation value was below –0.50.

Survey

No significant positive correlations were observed between the survey measures, but a significant negative correlation was found between unresponsiveness to the environment and hypervigilance scores (Spearman's rho: $r_s = –0.40$; $P < 0.001$), although the correlation value was below –0.50.

Discussion

Overall, the results show that the repeatability at an interval of three months and convergent validity of the measures were limited. Measures of stereotypies and aggressiveness showed higher repeatability and convergent validity than those of unresponsiveness to the environment and hypervigilance. Suggestions for the choice of the method to use according to the context of assessment are proposed for each category of behavioural indicators, as well as improvements to existing measures, mainly for those suitable for a one-off welfare inspection. No significant relationship appeared between the four categories of behavioural indicators, which underlines that they measure different mental states.

Repeatability and convergent validity of the measures

Stereotypies

Among the three measures, the most effective but still relatedly poor assessment of stereotypies appears to be using scans, as the repeatability of this measure was estimated with a correlation value slightly below 0.50. The AWIN measure did not show repeatability between the two periods and, in addition, was not related to the measures obtained by the

Table 3 Relationships between the four measures of aggressiveness towards humans among the three methods (Animal Welfare Indicators [AWIN] protocol including a Qualitative Behaviour Assessment [AWIN^{QBA}], scans and survey).

	Aggressiveness (AWIN)	Aggressiveness (AWIN^{QBA})	Aggressiveness (scans)	Aggressiveness (survey)
Aggressiveness (AWIN)	1	***[1]	0.38***[2]	0.33***[2]
Aggressiveness (AWIN^{QBA})		1	0.46***[2]	0.39***[2]
Aggressiveness (scans)			1	0.16*[3]
Aggressiveness (survey)				1

[1] Fisher's exact test;
[2] Point-biserial correlations (r_{pb});
[3] Spearman's rho correlation (r_s)
* $P \leq 0.05$; *** $P \leq 0.001$.

Table 4 Relationships between the three measures of unresponsiveness to the environment among the three methods (Animal Welfare Indicators [AWIN] protocol including a Qualitative Behaviour Assessment [AWIN^{QBA}], scans and survey).

	Unresponsiveness to the environment (AWIN^{QBA})	Unresponsiveness to the environment (scans)	Unresponsiveness to the environment (survey)
Unresponsiveness to the environment (AWIN^{QBA})	1	ns[1]	ns[1]
Unresponsiveness to the environment (scans)		1	0.15*[2]
Unresponsiveness to the environment (survey)			1

[1] Point-biserial correlations (r_{pb});
[2] Spearman's rho correlation (r_s)
* $P \leq 0.05$.

Table 5 Relationships between the three measures of hypervigilance among the three methods (Animal Welfare Indicators [AWIN] protocol including a Qualitative Behaviour Assessment [AWIN^{QBA}], scans and survey).

	Hypervigilance (AWIN^{QBA})	Hypervigilance (scans)	Hypervigilance (survey)
Hypervigilance (AWIN^{QBA})	1	ns[1]	ns[1]
Hypervigilance (scans)		1	ns[1]
Hypervigilance (survey)			1

[1] Spearman's rho correlation (r_s).

scans and survey, although a recent study conducted on a larger number of stables (14) identified an improved consistency over time of this measure (Czycholl *et al* 2021). It should also be noted that the number of stereotypic horses detected was much lower using the AWIN protocol than when using scans (AWIN protocol: 14 stereotypic horses detected, scans: 58 horses, see Table S6). All these results are probably explained by the very short duration of the AWIN measure (1 min), which is carried out only once to maintain feasibility of the overall protocol.

In comparison, the multiplication of observations by scans over time allowed more horses to be detected. Under-detection was also observed in the survey compared to the scans (survey: 15 stereotypic horses detected, see Table S6), which confirms previous results (Lesimple & Hausberger 2014). The under-detection of stereotypic horses probably indicates a lack of knowledge of caretakers regarding the different stereotypic behaviours that exist in horses. Indeed, discreet and less well-known stereotypies (eg lip or tongue movements) were rarely reported. It can also result from an over-exposure effect to animals expressing a compromised welfare state. In that case, the abnormal behaviour becomes the standard for the caretaker or the owner because it is expressed by many horses in the stable (Lesimple & Hausberger 2014).

In the context of one-off welfare inspections, it seems necessary to improve the AWIN measure by extending the duration of the current observation to maximise the chances

https://doi.org/10.7120/09627286.31.3.008 Published online by Cambridge University Press

of detecting stereotypic horses. However, to determine the optimal duration would require some complementary experiments. An alternative would be to replace this measure with a session of scans, following the same protocol as in the current study, with all the horses in the stable-yard being assessed at the same time to maintain feasibility. Here, further research is also needed to determine the optimal duration of this session. For both suggestions, standardising the implementation context, for example, by conducting behavioural observations around feeding could also maximise the chances of detecting stereotypic horses (Hockenhull & Creighton 2014), although the result may not be representative of the expression of stereotypies throughout the day. Furthermore, it may be useful to question the stable owner or the caretaker before the AWIN protocol assessment, as they could indicate the animals that express numerous stereotypies (r_{pb} = 0.63 between scans and survey). However, to increase detection, it appears necessary to inform the person about the different types of stereotypic behaviours that exist in horses.

In the context of regular self-monitoring of welfare by a stable owner or a caretaker, repeated observations at different times of the day using scans probably remains the method that maximises the chances of detecting stereotypic horses and could be easily implemented, for example, each time the person enters the stable (a few seconds per horse but regularly widespread over time). This would also involve training regarding the different stereotypies and how to perform the observation (eg at a distance from the horse, without disturbance) to ensure the reliability of the assessment.

Aggressiveness towards humans

Given the results of repeatability, the most effective assessment of aggressiveness towards humans also seems to be using scans, compared to the AWIN protocol whose measures presented values lower than 0.50. The three methods were significantly correlated with each other, but with correlation values below 0.50. These results could be explained by the variability observed in the detection rates of aggressive horses (AWIN test: 31 horses, AWIN^{QBA}: 19 horses; scans: 86 horses, survey: 67 horses; see Table S6). This variability is probably related to the fact that contexts in which aggressiveness towards humans is assessed are different from one measure to another: when approaching a human from outside the stall (scans), when approaching and touching a human inside the stall (AWIN protocol), and specifically when grooming and tacking up (survey).

In the context of one-off welfare inspections, the two current measures in the AWIN protocol (especially the QBA measure) were significantly related between the two periods of measurement but detected fewer horses expressing aggressiveness towards humans than scans and surveys. To improve the assessment, it could be interesting to standardise the context of implementation of the two measures to prior to meal times, a situation particularly prone to induce the expression of aggressive behaviours towards humans in horses with a compromised welfare state (Hockenhull & Creighton 2014). Performing scans at the same time as the session proposed for the assessment of stereotypies could be complementary.

Within the context of regular self-monitoring of welfare by a stable owner or a caretaker, the scan method seems superior to the others in terms of repeatability and the number of aggressive horses detected.

Unresponsiveness to the environment

Given our results, it seems difficult to determine the most effective method of assessing unresponsiveness to the environment in horses. Indeed, the QBA measure in the AWIN protocol and the scan measure were significantly related between the two periods but with coefficients far below 0.50. In addition, the measures from the three methods were not correlated with each other. Overall, these results suggest that the measures may reflect different mental states and could not be used indiscriminately to assess unresponsiveness to the environment as defined in this study. The score for the 'apathetic' affective state on the AWIN^{QBA} is not easy to interpret. It could either reflect unresponsiveness to the general environment or unresponsiveness to human beings in particular, because of the presence of the experimenter in the loose box (Minero *et al* 2018). A lack of response in the presence of a human could also reflect different mental states such as indifference or fear (Lansade *et al* 2008). However, horses could also have learned to remain immobile when a person (usually the caretaker) is present in their loose box. These different interpretations could explain the lack of correlation with the 'withdrawn posture' assessed by scans, which corresponds to a precise definition and is considered to reflect a 'depressive-like state' (Fureix *et al* 2012). The measure of unresponsiveness to the environment from the survey reports a subjective judgement not corresponding to defined criteria. Although the category of unresponsiveness to the environment seems to be well understood by experienced people (they use the terms 'withdrawn', 'unresponsive' and 'apathy' in association with the mental state 'boredom'; Hötzel *et al* 2019), it cannot be excluded that each caretaker referred to his/her personal interpretation. Whatever the context of the assessment, only the 'withdrawn posture' appears specific for now to evaluate unresponsiveness to the environment, although its expression could vary over time, perhaps due to sensitivity to minor changes in the environment of the stable. In addition, it is important to note that almost all horses were seen expressing the posture at least once through scans (see Table S6). In view of the restrictive living conditions of the animals, it is possible that they may all express a compromised welfare state demonstrated by this behavioural indicator. However, it cannot be excluded that only a high prevalence of the 'withdrawn posture' could be an effective warning sign of welfare deterioration. At this stage, the identification of a threshold at which welfare is compromised when observing this posture would require additional measures, such as the expression of anhedonia (Fureix *et al* 2015), impaired selective attention (Rochais *et al* 2016) or changes of tactile sensory sensitivity (Fureix *et al* 2012; Lansade *et al* 2014).

In the context of one-off welfare inspections, it could be useful to record the 'withdrawn posture' during the scan session proposed to assess stereotypies and aggressiveness towards humans. Further research is required to determine the optimal duration of the session and its timing of implementation to obtain a representative long-term assessment.

In the context of regular self-monitoring of welfare by a stable owner or a caretaker, repeated observations of the 'withdrawn posture' through scans are recommended, for example, each time the person enters the stable, to quantify its expression over time.

Hypervigilance

As for unresponsiveness to the environment, it seems difficult to identify the most effective method of assessing hypervigilance. Indeed, both the QBA measure in the AWIN protocol and scan measures were significantly correlated between the two periods, but with correlation values far below 0.50. In addition, the three measures from the QBA in the AWIN protocol, scans and surveys were not significantly correlated with each other, even though they detected an almost identical number of horses (see Table S6 in Supplementary material). The lack of correlations between the three measures again probably indicates the assessment of different mental states. The measure of the 'alarmed' state on the AWIN^{QBA} may actually reflect a specific temporary response to the experimenter and would not be representative of hypervigilance assessed using scans and monitored by recurrent vigilant behaviours over time and situations (ie the highest frequencies of alert postures). The measure of hypervigilance from the survey again relied on the subjective judgement of the caretakers using personal criteria. These results show that the three measures cannot be used indiscriminately to assess hypervigilance in horses. Further research is required to validate one or several behavioural measures for this category of indicators.

Within the context of one-off welfare inspections, the QBA measure in the AWIN protocol is probably an important measure of the mental state of horses, particularly in relation to humans, but is unlikely to assess hypervigilance as defined in this study. In addition, repeatability of the measure should be improved. It could be useful to add an assessment of alert postures during a scan session as this measure seems to provide a result that most closely approaches the definition of hypervigilance (ie increased alertness for threat by excessive scanning of the environment). However, we observed that the expression of alert postures strongly varies over time, probably due to small changes in the environment of the stable. In addition, alert postures constitute a part of the natural behavioural repertoire of the horse (Austin & Rogers 2014). Thus, only persistent expression of this posture would reflect hypervigilance and a compromised welfare state. There is a need to add specific measures, such as physiological criteria, to identify the threshold at which welfare is effectively compromised.

In the context of regular self-monitoring of welfare by a stable owner or a caretaker, the use of scan measures would probably best assess hypervigilance but this measure remains to be validated.

Overall, the role of the human being is preponderant in the assessment of the four behavioural indicators, whether it be the stable owner, the caretaker or an unknown assessor coming to carry out one-off inspections. Although the measures proposed in this study are based on well-specified behaviours, it is not possible to exclude the influence of human-related factors on the quality of the assessment. Indeed, Šárová *et al* (2011) showed that cattle farmers may underestimate a welfare indicator (ie the prevalence of lameness) in their herd as recognising it would lead to moral conflict and warrant further investigation. In addition, pet studies show that welfare issues are often denied, minimised or considered normal for a given species because owners do not have sufficient psychological distance to recognise them, due to the degree of familiarity with the animal (Serpell 2019). Both results may be applied to stable owners and horses' caretakers and would deserve to be extensively studied. Moreover, the professional affiliation of the unknown assessor for one-off inspections could also have an influence, as it has been shown that consideration for the emotional state of the animals in welfare assessment differed between production advisors, practicing veterinarians and animal welfare control officers in the cattle and pig sectors (Otten *et al* 2017). A low degree of consideration for the emotional state may lead to an underestimation of the four behavioural indicators in a stable.

Limited relationships between the four categories of behavioural indicators

No positive significant correlations were found among the measures of the four categories of behavioural indicators using scans and survey, and one correlation was observed among the AWIN protocol measures (aggressiveness and hypervigilance), but far below 0.50. A significant relationship between the aggressiveness and stereotypies measures in the AWIN protocol was also observed, but the lack of repeatability of the measure of stereotypies questions the validity of the link between the two categories of behavioural indicators. A significant negative correlation (–0.40) was found between the measures of unresponsiveness to the environment and hypervigilance assessed by the survey, although below 0.50. This result probably indicates that the caretakers tended to consider the two categories of behavioural indicators as antagonistic. Overall, these results support the importance of considering at least the four categories among the set of behavioural indicators used to assess the welfare state of horses in their living environment, in conjunction with physiological and health indicators. In the field, stereotypies appear to be the most common welfare indicators and are the subject of the majority of studies (eg McGreevy *et al* 1995; Bachmann *et al* 2003; Christie *et al* 2006; Tadich *et al* 2012). However, a better consideration of the other categories of

https://doi.org/10.7120/09627286.31.3.008 Published online by Cambridge University Press

behavioural indicators would allow a more accurate assessment of deterioration in the welfare state of horses. This is especially true because it remains unclear whether the stereotypic behaviours observed reflect the situation at the time of assessment or previous sub-optimal conditions (Sarrafchi & Blokhuis 2013). The absence of high correlation among the four categories of behavioural indicators was probably because distinct mental states were being assessed involving different underlying physiological mechanisms. For instance, stereotypies in horses could be triggered by changes in central nervous system dopamine physiology (McBride & Hemmings 2009), and the development of pathological aggression is related to glucocorticoids and abnormally low or high hypothalamic-pituitary-adrenal axis functionality in rodents and humans (Walker *et al* 2018). Further studies of the activity of these systems in relation to the four categories of behavioural indicators are needed in horses.

Animal welfare implications and conclusion

To date, very few studies have explored the links between different measures assessing the same behavioural indicator reflecting a compromised welfare state in horses in their living environment. Overall, the measures of stereotypies and aggressiveness showed better repeatability at a three-month interval and correlations between the AWIN protocol, scans and survey, compared to those of unresponsiveness to the environment and hypervigilance. For these two categories, it would be useful to link the measures with physiological indicators to validate them. These results highlight the difficulty in assessing the mental state of animals. Of the three methods of measures used in the present study, repeated measures, such as scans, seem to allow a more effective detection of each category of behavioural indicators of interest and their use should be prioritised whenever possible. However, they can only be implemented into welfare assessment protocols over a short observation period to maintain the feasibility of the overall protocol. Further research is required to determine the optimal duration and timing of this session. Finally, the four categories of behavioural indicators should be included among the set of indicators used to assess the welfare state of horses, in conjunction with physiological and health measures.

Declaration of interest

This project was funded by the IFCE and the 'Fonds Eperon.' This funding source had no role in the study design, data collection or analysis, or in the preparation or submission of the manuscript.

Acknowledgements

The authors are very grateful to the staff of the IFCE (French Horse and Riding Institute) and INRA (Nouzilly, France), particularly Patrick Galloux, Xavier Goupil, Isabelle Burgaud, Sophie Biau, Colonel Patrick Teisserenc, Jean-Marie Yvon, Miléna Trosh and all the animal keepers for their collaboration during the experiment. Thanks also to Sue Edrich from the translation agency Interconnect and the Springer Nature Author Services team for correcting the English manuscript.

References

Alberghina D, De Pasquale A, Piccione G, Vitale F and Panzera M 2015 Gene expression profile of cytokines in leukocytes from stereotypic horses. *Journal of Veterinary Behavior: Clinical Applications and Research 10*: 556-560. https://doi.org/10.1016/j.jveb.2015.08.007

Altmann J 1974 Observational study of behavior: Sampling methods. *Behaviour 49*: 227-267

Austin NP and Rogers LJ 2014 Lateralization of agonistic and vigilance responses in Przewalski horses (*Equus przewalskii*). *Applied Animal Behaviour Science 151*: 43-50. https://doi.org/10.1016/j.applanim.2013.11.011

AWIN 2015 *AWIN welfare assessment protocol for horses*. https://doi.org/10.13130/AWIN_HORSES_2015

Bachmann I, Audigé L and Stauffacher M 2003 Risk factors associated with behavioural disorders of crib-biting, weaving and box-walking in Swiss horses. *Equine Veterinary Journal 35*: 158-163. https://doi.org/10.2746/0425164037761142l6

Bachmann I and Stauffacher M 2002 Prävalenz von verhaltensstörungen in der Schweizer pferdepopulation. *Schweizer Archiv fur Tierheilkunde 144*: 356-368. https://doi.org/10.1024/0036-7281.144.7.356. [Title translation: Prevalence of behavioural disorders in the Swiss horse population]

Baragli P, Barbara P and Telatin A 2015 The role of associative and non-associative learning in the training of horses and implications for the welfare. *Ann Ist Super Sanità 51*: 40-51. https://doi.org/10.4415/ANN

Botreau R, Veissier I, Butterworth A, Bracke MBM and Keeling LJ 2007 Definition of criteria for overall assessment of animal welfare. *Animal Welfare 16*: 225-228

Christie JL, Hewson CJ, Riley CB, McNiven MA, Dohoo IR and Bate LA 2006 Management factors affecting stereotypies and body condition score in nonracing horses in Prince Edward Island. *Canadian Veterinary Journal 47*: 136-143

Curtis L, Burford JH, England GCW and Freeman SL 2019 Risk factors for acute abdominal pain (colic) in the adult horse: A scoping review of risk factors, and a systematic review of the effect of management-related changes. *PLoS ONE 14*: 1-32. https://doi.org/10.1371/journal.pone.0219307

Czycholl I, Büttner K, Klingbeil P and Krieter J 2021 Evaluation of consistency over time of the use of the Animal Welfare Indicators protocol for horses. *Animal Welfare 30*: 81-90. https://doi.org/10.7120/09627286.30.1.081

Czycholl I, Klingbeil P and Krieter J 2017 Suitability of animal-based indicators for the detection of animal welfare in horses. *TIERAERZTLICHE UMSCHAU 72*: 209-217

Czycholl I, Kniese C, Büttner K, Grosse Beilage E, Schrader L and Krieter J 2016 Test-retest reliability of the Welfare Quality® animal welfare assessment protocol for growing pigs. *Animal Welfare 25*: 447-459. https://doi.org/10.7120/09627286.25.4.447

Dalla Costa E, Dai F, Murray LAM, Guazzetti S, Canali E and Minero M 2015 A study on validity and reliability of on-farm tests to measure human-animal relationship in horses and donkeys. *Applied Animal Behaviour Science 163*: 110-121. https://doi.org/10.1016/j.applanim.2014.12.007

Dalla Costa E, Murray L, Dai F, Canali E and Minero M 2014 Equine on-farm welfare assessment: A review of animal-based indicators. *Animal Welfare 23*: 323-341. https://doi.org/10.7120/09627286.23.3.323

Dawkins MS 1990 From an animal's point of view: Motivation, fitness, and animal welfare. *Behavioral and Brain Sciences 13*: 1-9. https://doi.org/10.1017/S0140525X00077104

Dawkins MS 2003 Behaviour as a tool in the assessment of animal welfare. *Zoology 106*: 383-387. https://doi.org/10.1078/0944-2006-00122

Dellmeier GR 1989 Motivation in relation to the welfare of enclosed livestock. *Applied Animal Behaviour Science 22*: 129-138. https://doi.org/10.1016/0168-1591(89)90049-X

Fureix C, Beaulieu C, Argaud S, Rochais C, Quinton M, Henry S, Hausberger M and Mason G 2015 Investigating anhedonia in a non-conventional species: Do some riding horses Equus caballus display symptoms of depression? *Applied Animal Behaviour Science 162*: 26-36. https://doi.org/10.1016/j.applanim.2014.11.007

Fureix C, Gorecka-Bruzda A, Gautier E and Hausberger M 2011 Cooccurrence of Yawning and stereotypic behaviour in horses (*Equus caballus*). *ISRN Zoology 2011*: 1-10. https://doi.org/10.5402/2011/271209

Fureix C, Jego P, Henry S, Lansade L and Hausberger M 2012 Towards an ethological animal model of depression? A study on horses. *PLoS ONE 7*: e39280. https://doi.org/10.1371/journal.pone.0039280

Fureix C, Menguy H and Hausberger M 2010 Partners with bad temper: Reject or cure? A study of chronic pain and aggression in horses. *PLoS ONE 5*: e12434. https://doi.org/10.1371/journal.pone.0012434

Hao Y, Ge H, Sun M and Gao Y 2019 Selecting an appropriate animal model of depression. *International Journal of Molecular Sciences 20*: 1-16. https://doi.org/10.3390/ijms20194827

Harris PA 1999 Review of equine feeding and stable management practices in the UK concentrating on the last decade of the 20th century. *Equine Veterinary Journal. Supplement*: 46-54. https://doi.org/10.1111/j.2042-3306.1999.tb05156.x

Harro J 2018 Animals, anxiety, and anxiety disorders: How to measure anxiety in rodents and why. *Behavioural Brain Research 352*: 81-93. https://doi.org/10.1016/j.bbr.2017.10.016

Hausberger M, Roche H, Henry S and Visser EK 2008 A review of the human-horse relationship. *Applied Animal Behaviour Science 109*: 1-24. https://doi.org/10.1016/j.applanim.2007.04.015

Heleski CR, Shelle AC, Nielsen BD and Zanella AJ 2002 Influence of housing on weanling horse behavior and subsequent welfare. *Applied Animal Behaviour Science 78*: 291-302. https://doi.org/10.1016/S0168-1591(02)00108-9

Henry S, Fureix C, Rowberry R, Bateson M and Hausberger M 2017 Do horses with poor welfare show 'pessimistic' cognitive biases? *The Science of Nature 104*: 8. https://doi.org/10.1007/s00114-016-1429-1

Hinkle DE, Wiersma W and Jurs SG 2003 *Applied Statistics for the Behavioral Sciences, Fifth Edition* pp 756. Houghton Mifflin: Boston, MA, USA

Hockenhull J and Creighton E 2014 Pre-feeding behaviour in UK leisure horses and associated feeding routine risk factors. *Animal Welfare 23*: 297-308. https://doi.org/10.7120/09627286.23.3.297

Hockenhull J and Creighton E 2015 The day-to-day management of UK leisure horses and the prevalence of owner-reported stable-related and handling behaviour problems. *Animal Welfare 24*: 29-36. https://doi.org/10.7120/09627286.24.1.029.

Hötzel MJ, Vieira MC and Leme DP 2019 Exploring horse owners' and caretakers' perceptions of emotions and associated behaviors in horses. *Journal of Veterinary Behavior 29*: 18-24. https://doi.org/10.1016/j.jveb.2018.10.002

Houpt K, Houpt TR, Johnson JL, Erb HN and Yeon SC 2001 The effect of exercise deprivation on the behaviour and physiology of straight stall confined pregnant mares. *Animal Welfare 10*: 257-267

Hughes BO 1976 Behaviour as index of welfare. *5th European Poultry Conference* pp 1005-1018. 5-11 September 1976, Malta

Kwiatkowska-Stenzel A, Sowińska J and Witkowska D 2016 The effect of different bedding materials used in stable on horses behavior. *Journal of Equine Veterinary Science 42*: 57-66. https://doi.org/10.1016/j.jevs.2016.03.007

Lansade L, Bouissou MF and Erhard HW 2008 Fearfulness in horses: A temperament trait stable across time and situations. *Applied Animal Behaviour Science 115*: 182-200. https://doi.org/10.1016/j.applanim.2008.06.011

Lansade L, Valenchon M, Foury A, Neveux C, Cole SW, Layé S, Cardinaud B, Lévy F and Moisan M-P 2014 Behavioral and transcriptomic fingerprints of an enriched environment in horses (*Equus caballus*). *PLoS ONE 9*: e114384. https://doi.org/10.1371/journal.pone.0114384

Lesimple C and Hausberger M 2014 How accurate are we at assessing others' well-being? The example of welfare assessment in horses. *Frontiers in Psychology 5*: 1-6. https://doi.org/10.3389/fpsyg.2014.00021

Mason GJ 1991 Stereotypies: a critical review. *Animal Behaviour 41*: 1015-1037. https://doi.org/10.1016/S0003-3472(05)80640-2

Mason GJ and Latham NR 2004 Can't stop, won't stop: Is stereotypy a reliable animal welfare indicator? *Animal Welfare 13*: 57-69. https://doi.org/10.2307/4493573

McBride S and Hemmings A 2009 A neurologic perspective of equine stereotypy. *Journal of Equine Veterinary Science 29*: 10-16. https://doi.org/10.1016/j.jevs.2008.11.008

McGreevy PD 2004 *Equine Behavior: A Guide for Veterinarians and Equine Scientists* pp 366. Saunders: USA

McGreevy PD, Cripps PJ, French NP, Green LE and Nicol CJ 1995 Management factors associated with stereotypic and redirected behaviour in the Thoroughbred horse. *Equine Veterinary Journal 27*: 86-91. https://doi.org/10.1111/j.2042-3306.1995.tb03041.x

McHugh ML 2012 Interrater reliability: the kappa statistic. *Biochemia Medica 22*: 276-282

Meagher RK 2009 Observer ratings: Validity and value as a tool for animal welfare research. *Applied Animal Behaviour Science 119*: 1-14. https://doi.org/10.1016/j.applanim.2009.02.026

Mellor DJ 2016 Updating animal welfare thinking: Moving beyond the 'Five Freedoms' towards 'A life worth living.' *Animals 6*. https://doi.org/10.3390/ani6030021

Mills DS and Davenport K 2002 The effect of a neighbouring conspecific versus the use of a mirror for the control of stereotypic weaving behaviour in the stabled horse. *Animal Science 74*: 95-101

https://doi.org/10.7120/09627286.31.3.008 Published online by Cambridge University Press

Minero M, Dalla Costa E, Dai F, Canali E, Barbieri S, Zanella A, Pascuzzo R and Wemelsfelder F 2018 Using qualitative behaviour assessment (QBA) to explore the emotional state of horses and its association with human-animal relationship. *Applied Animal Behaviour Science 204*: 53-59. https://doi.org/10.1016/j.applanim.2018.04.008

Minero M, Dalla Costa E, Dai F, Murray LAM, Canali E and Wemelsfelder F 2016 Use of Qualitative Behaviour Assessment as an indicator of welfare in donkeys. *Applied Animal Behaviour Science 174*: 147-153. https://doi.org/10.1016/j.applanim.2015.10.010

Ninomiya S, Sato S and Sugawara K 2007 Weaving in stabled horses and its relationship to other behavioural traits. *Applied Animal Behaviour Science 106*: 134-143. https://doi.org/10.1016/j.applanim.2006.06.014

Normando S, Canali E, Ferrante V and Verga M 2002 Behavioral problems in Italian saddle horses. *Journal of Equine Veterinary Science 22*: 117-120. https://doi.org/10.1016/S0737-0806(02)70123-8

Normando S, Meers L, Samuels WE, Faustini M and Ödberg FO 2011 Variables affecting the prevalence of behavioural problems in horses. Can riding style and other management factors be significant? *Applied Animal Behaviour Science 133*: 186-198. https://doi.org/10.1016/j.applanim.2011.06.012

Ödberg FO 1987 Chronic stress in riding horses. *Equine Veterinary Journal 19*: 268-269. https://doi.org/10.1111/j.2042-3306.1987.tb01402.x

Ohl F, Arndt SS and van der Staay FJ 2008 Pathological anxiety in animals. *Veterinary Journal 175*: 18-26. https://doi.org/10.1016/j.tvjl.2006.12.013

Omidi A, Jafari R, Nazifi S and Parker MO 2018 Potential role for selenium in the pathophysiology of crib-biting behavior in horses. *Journal of Veterinary Behavior: Clinical Applications and Research 23*: 10-14. https://doi.org/10.1016/j.jveb.2017.10.003

Otten ND, Rousing T and Forkman B 2017 Influence of professional affiliation on expert's view on welfare measures. *Animals 7*: 1-10. https://doi.org/10.3390/ani7110085

Pessoa GO, Trigo P, Mesquita Neto FD, Lacreta Junior ACC, Sousa TM, Muniz JA and Moura RS 2016 Comparative well-being of horses kept under total or partial confinement prior to employment for mounted patrols. *Applied Animal Behaviour Science 184*: 51-58. https://doi.org/10.1016/j.applanim.2016.08.014

Popescu S and Diugan EA 2013 The relationship between behavioral and other welfare indicators of working horses. *Journal of Equine Veterinary Science 33*: 1-12. https://doi.org/10.1016/j.jevs.2012.04.001

Post RJ and Warden MR 2018 Melancholy, anhedonia, apathy: the search for separable behaviors and neural circuits in depression. *Current Opinion in Neurobiology 49*: 1-9. https://doi.org/10.1016/j.conb.2018.02.018

Pritchard JC, Lindberg AC, Main DCJ and Whay HR 2005 Assessment of the welfare of working horses, mules and donkeys, using health and behaviour parameters. *Preventive Veterinary Medicine 69*: 265-283. https://doi.org/10.1016/j.prevetmed.2005.02.002

R Core Team 2020 *R: A language and environment for statistical computing*. R Foundation for Statistical Computing: Vienna, Austria. https://www.R-project.org/

Ribeiro LB, Matzkeit TV, Nicolau JT de S, Castilha LD, de Oliveira FCL and Bankuti FI 2019 Determinants of undesirable behaviors in American quarter horses housed in box stalls. *Journal of Equine Veterinary Science 80*: 69-75. https://doi.org/10.1016/j.jevs.2019.07.005

Richards HJ, Benson V, Donnelly N and Hadwin JA 2014 Exploring the function of selective attention and hypervigilance for threat in anxiety. *Clinical Psychology Review 34*: 1-13. https://doi.org/10.1016/j.cpr.2013.10.006

Rochais C, Henry S, Fureix C and Hausberger M 2016 Investigating attentional processes in depressive-like domestic horses (*Equus caballus*). *Behavioural Processes 124*: 93-96. https://doi.org/10.1016/j.beproc.2015.12.010

Ruet A, Lemarchand J, Parias C, Mach N, Moisan M, Foury A, Briant C and Lansade L 2019 Housing horses in individual boxes is a challenge with regard to welfare. *Animals 9*: 621. https://doi.org/10.3390/ani9090621

Salomons AR, Arndt SS and Ohl F 2009 Anxiety in relation to animal environment and welfare. *Scandinavian Journal of Laboratory Animal Science 36*: 37-45

Šárová R, Stěhulová I, Kratinová P, Firla P and Špinka M 2011 Farm managers underestimate lameness prevalence in Czech dairy herds. *Animal Welfare 20*: 201-204

Sarrafchi A and Blokhuis HJ 2013 Equine stereotypic behaviors: Causation, occurrence, and prevention. *Journal of Veterinary Behavior 8*: 386-394 https://doi.org/10.1016/j.jveb.2013.04.068

Serpell JA 2019 How happy is your pet? The problem of subjectivity in the assessment of companion animal welfare. *Animal Welfare 28*: 57-66. https://doi.org/10.7120/09627286.28.1.057

Søndergaard E, Jensen MB and Nicol CJ 2011 Motivation for social contact in horses measured by operant conditioning. *Applied Animal Behaviour Science 132*: 131-137. https://doi.org/10.1016/j.applanim.2011.04.007

Statton R, Cogger N, Beausoleil N, Waran N, Stafford K and Stewart M 2014 Indicators of good welfare in horses. Final report. *Ministry for primary industries technical paper No: 2014/44.* https://www.mpi.govt.nz/resources-and-forms/publications/

Sylvers P, Lilienfeld SO and LaPrairie JL 2011 Differences between trait fear and trait anxiety: Implications for psychopathology. *Clinical Psychology Review 31*: 122-137. https://doi.org/10.1016/j.cpr.2010.08.004

Tadich T, Smulders JP, Araya O and Nicol CJ 2012 Husbandry practices associated with the presentation of abnormal behaviours in Chilean Creole horses. *Archivos de Medicina Veterinaria 44*: 279-284. https://doi.org/10.4067/S0301-732X2012000300011

Temple D, Manteca X, Dalmau A and Velarde A 2013 Assessment of test-retest reliability of animal-based measures on growing pig farms. *Livestock Science 151*: 35-45. https://doi.org/10.1016/j.livsci.2012.10.012

Thomas KE, Annest JL, Gilchrist J and Bixby-Hammett DM 2006 Non-fatal horse related injuries treated in emergency departments in the United States. *British Journal of Sports Medicine 40*: 619-626. https://doi.org/10.1136/bjsm.2006.025858

Veissier I and Boissy A 2007 Stress and welfare: Two complementary concepts that are intrinsically related to the animal's point of view. *Physiology and Behavior 92*: 429-433. https://doi.org/10.1016/j.physbeh.2006.11.008

Walker SE, Papilloud A, Huzard D and Sandi C 2018 The link between aberrant hypothalamic-pituitary-adrenal axis activity during development and the emergence of aggression: Animal studies. *Neuroscience & Biobehavioral Reviews 91*: 138-152. https://doi.org/10.1016/j.neubiorev.2016.10.008

Waters AJ, Nicol CJ and French NP 2002 Factors influencing the development of stereotypic and redirected behaviours in young horses: findings of a four year prospective epidemiological study. *Equine Veterinary Journal 34*: 572-579. https://doi.org/10.2746/042516402776180241

Wathan J and McComb K 2014 The eyes and ears are visual indicators of attention in domestic horses. *Current Biology 24*: R677-R679. https://doi.org/10.1016/j.cub.2014.06.023

Webster J 2005 *Animal Welfare: Limping Towards Eden, First Edition*. John Wiley & Sons Ltd: London, UK

Wemelsfelder F 2007 How animals communicate quality of life: The Qualitative Assessment of Behaviour. *Animal Welfare 16*: 25-31

Wermes R, Lincoln TM and Helbig-Lang S 2018 Anxious and alert? Hypervigilance in social anxiety disorder. *Psychiatry Research 269*: 740-745. https://doi.org/10.1016/j.psychres.2018.08.086

Wickens CL and Heleski CR 2010 Crib-biting behavior in horses: A review. *Applied Animal Behaviour Science 128*: 1-9. https://doi.org/10.1016/j.applanim.2010.07.002

Yim VWT, Yeung JHH, Mak PSK, Graham CA, Lai PBS and Rainer TH 2007 Five year analysis of Jockey Club horse-related injuries presenting to a trauma centre in Hong Kong. *Injury 38*: 98-103. https://doi.org/10.1016/j.injury.2006.08.026

Young T, Creighton E, Smith T and Hosie C 2012 A novel scale of behavioural indicators of stress for use with domestic horses. *Applied Animal Behaviour Science 140*: 33-43. https://doi.org/10.1016/j.applanim.2012.05.008

https://doi.org/10.7120/09627286.31.3.008 Published online by Cambridge University Press