© 2021 Universities Federation for Animal Welfare
The Old School, Brewhouse Hill, Wheathampstead,
Hertfordshire AL4 8AN, UK
www.ufaw.org.uk

Animal Welfare 2021, 30: 81-90
ISSN 0962-7286
doi: 10.7120/09627286.30.1.081

Evaluation of consistency over time of the use of the Animal Welfare Indicators protocol for horses

I Czycholl*†, K Büttner†‡, P Klingbeil† and J Krieter†

† Institute of Animal Breeding and Husbandry, Christian-Albrechts-University Kiel, Olshausenstraße 40, 24118 Kiel, Germany
‡ Unit for Biomathematics and Data Processing, Facullty of Veterinary Medicine, Justus Liebig University, Frankfurter Str 95, 35392 Giessen, Germany
* Contact for correspondence: iczycholl@tierzucht.uni-kiel.de

Abstract

*Consistency over time is a basic requirement for welfare assessment schemes since consistency must not depend, for example, on the day it is carried out. This study analysed the consistency of the indicators of the Animal Welfare Indicators (AWIN) protocol for horses (*Equus caballus*) over time. Given the multi-dimensionality of animal welfare, the AWIN protocol includes a variety of indicators evaluating, eg the health status or the behaviour of the animals. Fourteen establishments keeping horses in Germany were visited four times each (day 0, day 3, day 42, day 90). For the evaluation of reliability and agreement between the different visits, ie across time, the reference visit on day 0 was compared to the other visits via calculation of Spearman's rank correlation (RS), intra-class correlation (ICC), smallest detectable change (SDC) and limits of agreement (LoA). The indicator, Qualitative Behaviour Assessment (QBA) was analysed by Principal Component Analysis (PCA). Most of the indicators demonstrated sufficient consistency over time. Indicators that were inconsistent included parts of the Horse Grimace Scale, outcomes of behavioural tests, the presence of swollen joints as well as the indicators hoof neglect, alopecia on the legs and water cleanliness. The QBA was consistent for the period of 42 days, but not for 90 days. Overall, those indicators with insufficient consistency over time require to be revised or replaced in future welfare assessment schemes.*

Keywords: *animal-based indicators, animal welfare, AWIN, consistency, horse, reliability*

Introduction

Welfare is generally acknowledged to be a multi-dimensional concept made up of good health and biological functioning, natural behaviour and emotional state. Hence, for the accurate assessment of welfare, multiple indicators are required since a single indicator is insufficient to measure all these different aspects (Fraser 2008; Blokhuis *et al* 2013; Czycholl *et al* 2015). In general, animal-based indicators, which are also outcome-based indicators, are considered the most useful in terms of assessing the true welfare state during on-farm welfare assessments (Blokhuis *et al* 2013). In contrast, management- and resource-based indicators constitute a risk assessment of the surroundings and the possible attainment of a good welfare state for the animals. However, the measurement of animal-based indicators poses the greatest challenge as regards feasibility, reliability and validity. For example, behavioural observations to assess animals' lying behaviour are more time-consuming and require better trained assessors than simply assessing the size of the resting area (Velarde & Geers 2007). Hence, a major issue in animal welfare science continues to be the suitability of animal-based indicators for the purposes of a welfare assessment.

Although generally important for all species in animal welfare assessment, animal-based indicators are probably especially important when it comes to an assessment of equine welfare since the use of this species, not to mention the husbandry conditions, show great variation, eg ranging from leisure horses (*Equus caballus*) to working equids and food production (Dalla Costa *et al* 2017). Animal-based indicators allow comparison between different husbandry conditions and make it possible for objective comparison to take place (Dalla Costa *et al* 2014).

In 2015, a welfare assessment protocol for horses was developed within the framework of the Animal Welfare Indicators (AWIN 2015) project, amongst others. Welfare was defined as a multi-dimensional concept to be measured by a variety of predominantly animal-based indicators. Resource- (eg water provision) or management-based (eg exercise) indicators were only included if no feasible, reliable and valid animal-based indicator was revealed for a certain aspect of animal welfare. However, research concerning the feasibility, reliability and validity of the included indicators remains an ongoing process. Some single indicators were assessed for feasibility, reliability and validity before inclusion in the respective protocols. For example, Dalla Costa *et al* (2014, 2016b) validated the Horse Grimace Scale as a tool to assess pain by comparing horses in pain to control groups without pain or receiving analgesia. Dai *et al* (2015) validated the fear tests by

https://doi.org/10.7120/09627286.30.1.081 Published online by Cambridge University Press

comparing the outcomes to those of infra-red thermography and Dalla Costa *et al* (2015) tested different Human Animal Relationship Tests with regard to their feasible, reliable and valid use in horses. Concerning the AWIN protocol in general, Dalla Costa *et al* (2016a) provides an overall description of the application of the AWIN protocol and a general estimate of feasibility and validity based on expert and stakeholder opinion. Also, the completed and published protocols were tested for feasibility in on-farm studies (Dalla Costa *et al* 2017; Czycholl *et al* 2018). Overall, these studies suggest that feasibility and validity (at least for most of the indicators) are given. Czycholl *et al* (2019a) further addressed the inter-observer reliability of the AWIN protocol demonstrating it to be sufficient for most of the indicators. However, until now, the consistency of the entire protocol over time or of most of the single indicators has not been thoroughly tested (Dalla Costa *et al* 2015).

Nevertheless, consistency over time is a basic requirement of welfare assessment systems. Only with a degree of consistency over time can comparability between farms be assured (Knierim & Winckler 2009). Therefore, in the present study, we analysed the consistency of all included AWIN protocol indicators for horses over time, in different time-periods up to 90 days. The aim was to supplement pre-existing knowledge on the AWIN protocol for horses and explore the potential for different indicators to reliably assess certain aspects of welfare.

Materials and methods

Ethical statement

Animal disturbance was kept to a minimum, in accordance with the nature of the AWIN protocol, with the study designed to be readily incorporated into each establishments' normal routine. The authors declare that the 'German Animal Welfare Act' (German designation: TierSchG 2006) and the 'German Order for the Protection of Animals used for Experimental Purposes and other Scientific Purposes' (German designation: TierSchVersV 2013) were applied. No pain, suffering or injury were inflicted on the animals during the time-frame of the experiment.

Study animals

Data collection took place between December 2016 and May 2017 on 14 equine establishments in Germany. These all took part on a voluntary basis after being invited by their breeding association or the University of Kiel. All the study horses were kept in accordance with the national guidelines for equine husbandry (BMELV 1995), although conditions varied greatly between establishments. Sizes ranged from 14 up to 120 horses with four establishments classifying themselves as mainly breeding stables, three (mainly) as sport stables (dressage and showjumping) and seven as pension stables, ie containing mostly leisure horses. A variety of breeds were kept throughout and German Warmbloods (such as Trakehner, Holsteiner, Hannoverian) made up approximately two-thirds of the total sample, which is typical for Germany.

Welfare assessment

The same well-trained observer carried out all assessments with training having taken place prior to the start of the study via a three-day course organised by the developers of the AWIN protocol for horses. Training was based around video footage material and pictures and took place in various equine establishments until a robust scoring consensus for each indicator and assessment was attained amongst participants and protocol developers. Protocol assessments were carried out in strict accordance with the instructions presented in the AWIN protocol for horses (AWIN 2015), where a detailed description of the entire procedure can be found. An abridged version of the indicators and scores may be found in Table S1 (see supplementary material to papers published in *Animal Welfare*: https://www.ufaw.org.uk/the-ufaw-journal/supplementary-material).

The AWIN protocol for horses is divided into two levels: a first-level welfare assessment which is only performed on a sample of the animals and functions as a quick overview and a second-level assessment which is carried out on every single animal. This includes additional indicators; some of which are assessed in greater detail (eg not only visually but also via palpation). In the present study, both levels were performed at the same establishment: the first on the first day and, irrespective of the outcome, the second the following day. Each of the 14 establishments was visited four times by the same observer, with visit 1 on day 0, visit 2 on day 3, visit 3 after 42 days and visit 4 after 90 days. This produced a total of 56 first-level protocol assessments and 56 second-level protocol assessments, in which a total of 1,055 and 1,582 horses were observed, respectively. The results of visit 1 were classed as the reference visit and compared to those of visits 2, 3 and 4 for each of the (two) levels, separately.

Data analysis

Since the aim of welfare assessment tools tends to be to aid the detection of welfare status, on-farm (Blokhuis *et al* 2013), the results here are expressed at farm level, thus as continuous data, ie showing the percentages of affected animals sorted into the respective scores for each indicator. An approach wholly in accordance with the AWIN protocol (AWIN 2015) that saw each score of each indicator treated as a separate variable.

For the assessment of consistency over time, visits 2, 3 and 4 were compared to the reference visit (visit 1). For comparison, a combination of different reliability, ie Spearman's rank correlation coefficient (RS) and intra-class correlation coefficient (ICC) and agreement parameters, ie smallest detectable change (SDC) and limits of agreement (LoA) were used and interpreted together. For interpretation, it was determined that all parameters had to reach the respective pre-defined thresholds for acceptability (de Vet *et al* 2006).

The RS is a non-parametric measure of rank correlation, for which values can range between –1 and 1. In accordance with the suggestions of Martin and Bateson (2007), an RS equal to or greater than 0.4 was interpreted as an acceptable correlation and RS equal to or greater than 0.7 as a good correlation.

© 2021 Universities Federation for Animal Welfare

https://doi.org/10.7120/09627286.30.1.081 Published online by Cambridge University Press

For the ICC, the following one-way model was calculated in accordance with Shrout and Fleiss (1979):

$x_{jk} = \mu + \alpha_j + \varepsilon_{jk}$

with x_{jk} being the measured value, μ the general average value for each assessed indicator score, α_j the random effect of the difference between the study objects (farm visits) and ε_{jk} the general error term. Then, the variance of the same object (visits) was put in proportion to the total variance using the following formula in accordance with de Vet *et al* (2006):

$$ICC = \frac{\sigma^2(objects)}{\sigma^2(objects) + \sigma^2(residual)}$$

with σ^2 representing the variance of the study objects (visits) and the residual variance, respectively. As proposed by McGraw and Wong (1996), an ICC equal to or greater than 0.4 was interpreted as acceptable reliability and an ICC equal to or greater than 0.7 as good reliability.

The SDC is an expression of the measurement error, which is derived from the previous formulae (de Vet *et al* 2006) and depicts the smallest change that can be detected despite the measurement error:

$$SDC = 1.96 \times \sqrt{2} \times (\sigma^2[visits] + \sigma^2[residual])$$

The values of the SDC are the same as the measurement unit of the assessed indicators, ie percentages in this study. A deviation of up to 10% was interpreted as acceptable agreement and values smaller than or equal to 5% as good agreement.

The LoA were also calculated according to de Vet *et al* (2006) by the formula:

$$LoA = mean \pm 1.96 \times (\sqrt{2} \times \sigma^2[residual])$$

It calculates the range of the difference between two sets of measurement values (first- and second-level protocols). Interpretation was again based on the simple agreement coefficient of de Vet *et al* (2006) and therefore an interval smaller than or equal to –10 to 10% was interpreted as acceptable and –5 to 5% as good agreement.

A Principal Component Analysis (PCA) was performed for the analysis of the Qualitative Behaviour Assessment (QBA), which is one of the additional indicators of the second-level protocol. The QBA consists of 13 descriptors: aggressive, alarmed, annoyed, apathetic, at ease, curious, friendly, fearful, happy, looking for contact, relaxed, pushy and uneasy. A value (in mm) was attained for each horse, for each descriptor. These were aggregated to farm level. For the PCA, a correlation matrix was used, and no rotation applied. Separate PCAs were conducted for the adjectives for the four different visits. For the evaluation of consistency over time, the factor loadings of Principal Components 1 and 2 (PC1 and PC2) of the different visits were compared by the calculation of Spearman's rank correlation coefficients. This procedure is advised for analysis of the QBA by the developers of this methodology (Wemelsfelder & Lawrence 2001; Wemelsfelder & Millard 2009).

All statistical analyses were carried out using SAS® 9.4 (SAS Institute 2008).

Results

Only those indicator scores demonstrating presence of that specific welfare issue are presented. A number of the indicator scores were only observed very rarely (eg strained nostrils and flattening of the profile, score 2 of the Horse Grimace Scale, score 1 of the Body Condition Score (BCS), genital discharge, water points not functioning properly, as well as some of the integument alterations, ie alopecia, skin lesion, deep wound or swelling in different regions of the body). Hence, these scores were excluded from further analysis.

First-level protocol assessment

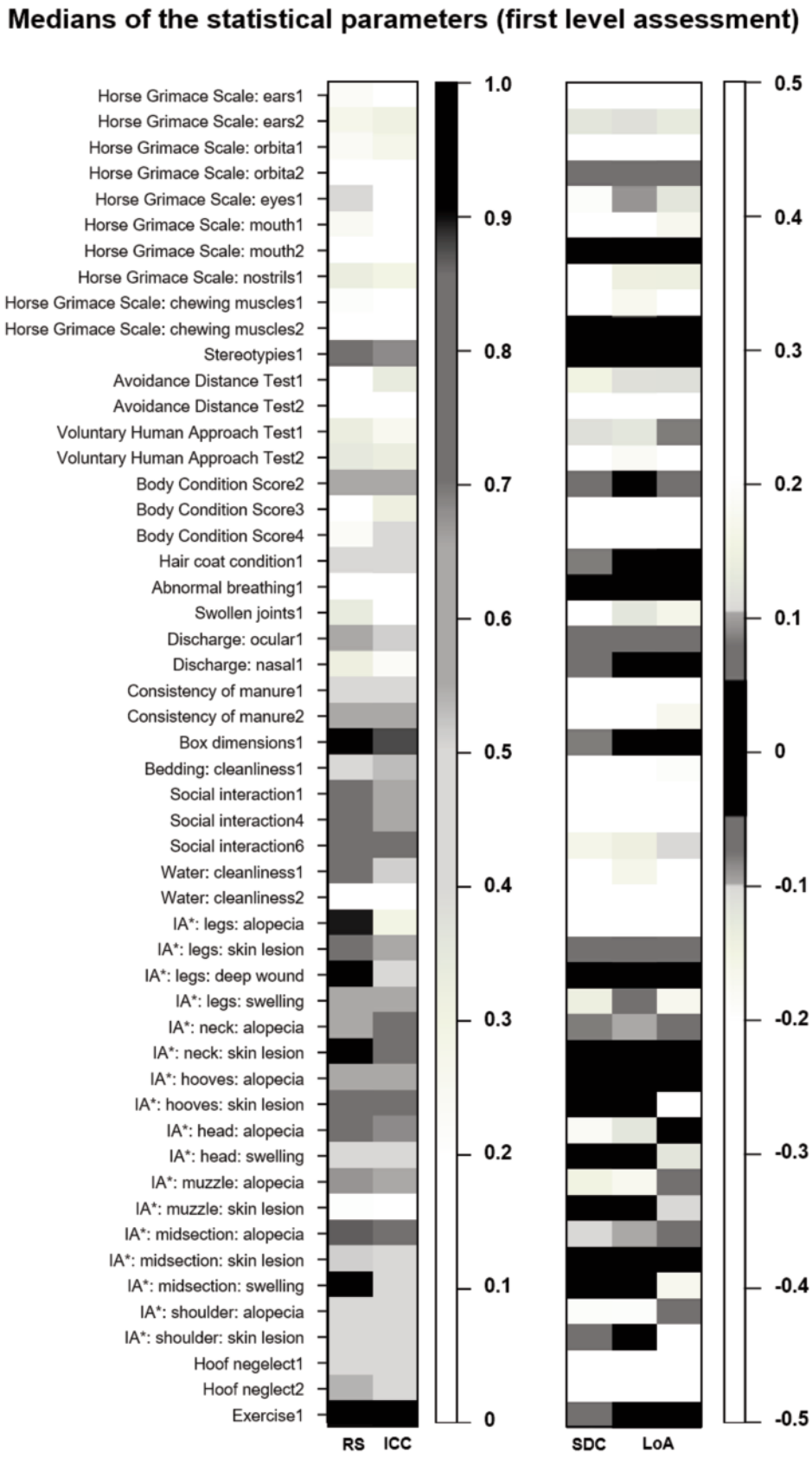
The median values in percent of affected animals for each of the indicator scores (considered as separate variables) for each of the four visits are shown in Table 1 (see supplementary material to papers published in *Animal Welfare*: https://www.ufaw.org.uk/the-ufaw-journal/supplementary-material). The comparison of the different visits for the evaluation of consistency over time is shown in Table 2 (see supplementary material to papers published in *Animal Welfare*: https://www.ufaw.org.uk/the-ufaw-journal/supplementary-material). To facilitate interpretation, the medians of the reliability and agreement parameters of the different comparisons are included in Figure 1. The indicators, stereotypies, ocular discharge, box dimension, alopecia and deep wounds on the legs, alopecia and skin lesion on the neck, alopecia and skin lesion on the hooves, swelling on the head, skin lesion and swelling on the midsection, skin lesion on the shoulder and exercise demonstrated good consistency over time within 90 days. The Horse Grimace Scale (especially those indicators demonstrating only mild pain), the Avoidance Distance Test, the Voluntary Human Approach Test and swollen joints were of insufficient consistency over time in all time-periods under evaluation. The other indicators demonstrated inconsistent results in the sense that either the agreement parameters were of acceptable values while the reliability parameters were not (consistency of manure, bedding cleanliness, social interaction, water cleanliness) or *vice versa* (abnormal breathing, nasal discharge, water functioning, skin lesion on the muzzle).

Second-level protocol assessment

The median percent values of the affected animals for each of the indicators in the assessment of the second-level protocol are shown in Table 3 (see supplementary material to papers published in *Animal Welfare*: https://www.ufaw.org.uk/the-ufaw-journal/supplementary-material). The comparison of the different visits for the evaluation of the consistency in the different indicators over time is shown in Table 4 (see supplementary material to papers published in *Animal Welfare*: https://www.ufaw.org.uk/the-ufaw-journal/supplementary-material). The medians of the statistical parameters are shown in Figure 2. The results for the respective indicators are in accordance with those of the first-level protocol. Exceptions are the different BCS, for which the consistency over time in all the evaluated time-frames improved. A closer look at the additional indicators used in the second-level protocol shows

https://doi.org/10.7120/09627286.30.1.081 Published online by Cambridge University Press

Figure 1

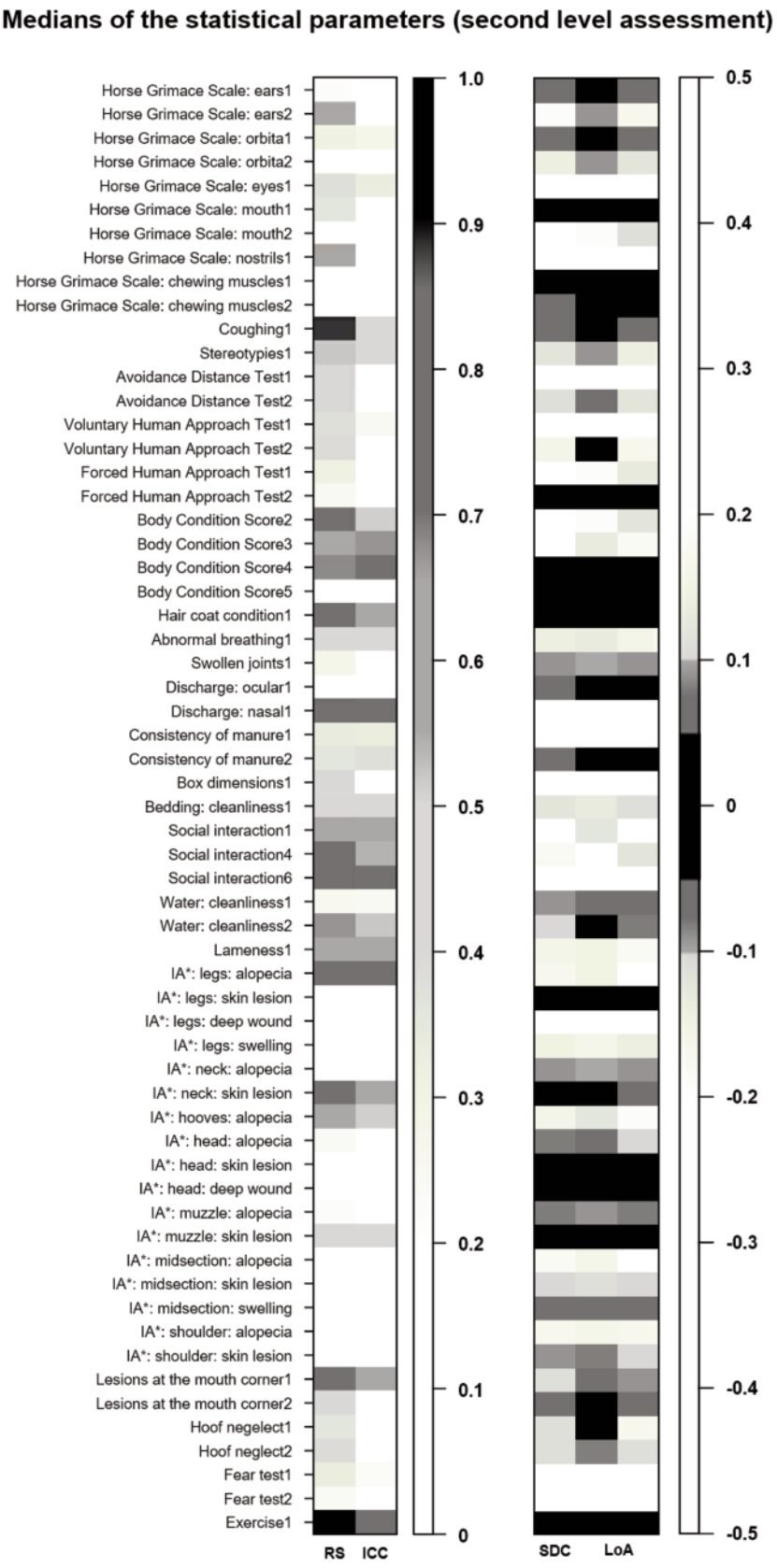


Medians of the different statistical parameters: Spearman's rank correlation coefficient (RS), intra-class correlation coefficient (ICC), smallest detectable change (SDC), limits of agreement (LoA) for the different comparisons of farm visits for each indicator of the AWIN protocol for horses (first-level assessment). LoA is expressed with lower and upper limits. Lighter colours indicate insufficient reliability.

© *2021 Universities Federation for Animal Welfare*

https://doi.org/10.7120/09627286.30.1.081 Published online by Cambridge University Press

Figure 2



Medians of the different statistical parameters: Spearman's rank correlation coefficient (RS), intra-class correlation coefficient (ICC), smallest detectable change (SDC), limits of agreement (LoA) for the different comparisons of farm visits for each indicator of the AWIN protocol for horses (second-level assessment). LoA is expressed with lower and upper limits. Lighter colours indicate insufficient reliability.

https://doi.org/10.7120/09627286.30.1.081 Published online by Cambridge University Press

that for coughing, lameness and lesions at the mouth corner, the statistical parameters met the thresholds for acceptability, while this was not the case for the results of the Forced Human Approach Test.

The results of the QBA are presented in Figure 3. It shows the scores of the different descriptors for PC1 and PC2 in the different comparisons. Correlation between the farm visits was high for the comparison of visit 1 with visit 2 (PC1: RS = 0.96, PC2: RS = 0.74, see Figure 3[a]) and decreased in the longer intervals of comparison (visit 1 to visit 3: PC1: RS = 0.88, PC2: RS = 0.50, see Figure 3[b]); visit 1 to visit 4: RS = 0.87, PC2: RS = 0.13, see Figure 3[c]).

Discussion

First-level protocol assessment

The general prevalence of the single indicators is comparable to those of other studies using the AWIN protocol for horses (Dalla Costa *et al* 2017; Czycholl *et al* 2018).

Those indicators meeting the pre-defined threshold limits of acceptability in all four statistical parameters (stereotypies, ocular discharge, box dimension, alopecia and deep wounds on the legs, alopecia and skin lesion on the neck, alopecia and skin lesion on the hooves, swelling on the head, skin lesion and swelling on the midsection, skin lesion on the shoulder and exercise) can be interpreted as sufficiently constant over time within the 90 days and are able to be considered acceptable for their inclusion as welfare assessment tools. This is reinforced by the inter-observer reliability of these indicators, which proved sufficient in previous studies (Czycholl *et al* 2019a).

For some indicators, all four statistical parameters were below the pre-defined threshold values, which can be interpreted as follows: all Horse Grimace scale scores indicating mild presence of pain in the respective facial region (score 1) were insufficiently consistent over time for all the time-periods under evaluation. Further, the indicator, ears pointing stiffly backwards of score 2 of the Horse Grimace Scale was not reliable. This is supported by the poor inter-observer reliability of these tests, which was insufficient to infer their use (Czycholl *et al* 2019a) plus the fact that the Horse Grimace Scale is only validated for severe acute pain for < 48 h after a painful incident (Thatcher *et al* 2012; Dalla Costa *et al* 2014). Hence, the indicators demonstrating light affection of pain should be removed from the protocol since they cannot be assessed reliably. The indicator, ears pointing stiffly backwards is probably prone to error due to multiple causes of this behaviour (Chamove *et al* 2002). The Avoidance Distance Test was also insufficiently consistent over time in all evaluated time-periods. Although slightly more consistent, the same is basically true for the Voluntary Human Approach Test. Hence, these behavioural tests are unsuitable for welfare assessment in horses. This is in accordance with findings regarding the inter-observer reliability of these tests in the study of Czycholl *et al* (2019a), which also deemed them insufficient. Similarly, within behavioural tests in horses in general, Visser *et al* (2001) also failed to detect long-term consistency in a Novel Object and Handling Test. Jezierski *et al* (1999) found inconsistency in the manageability of young horses within six months and McCann *et al* (1988) detected variances in heart-rate responses in repeated Novel Object Tests. In other species, Czycholl *et al* (2019b) detected that in pigs (*Sus scrofa*), five different behavioural tests were of insufficient consistency over time as did Miller *et al* (2006) for different behavioural tests in quails (*Coturnix coturnix*). This can probably be explained by age effects, learning effects, habituation effects and changes in the motivational state of the animals (Jezierski *et al* 1999; Visser *et al* 2001; Mieloch *et al* 2020).

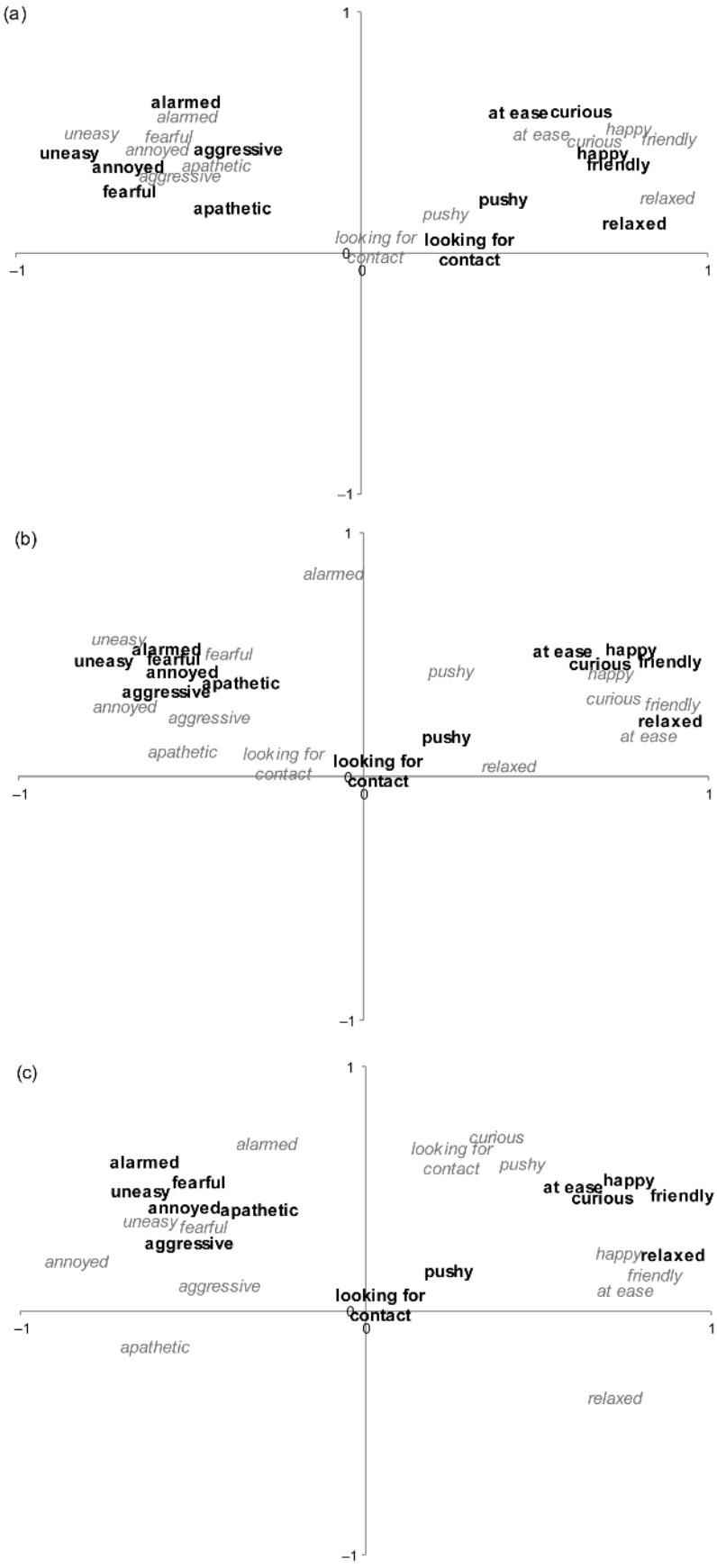
BCS 3 and 4 were unreliable. This is almost certainly because most horses were between both categories and an exact scoring simply using visual inspection was not possible (Czycholl *et al* 2018). Further, swollen joints were of insufficient consistency over time in all evaluated time-periods. However, as with BCS, this is most likely a general problem with the scoring of this indicator given previous reliability studies of the AWIN protocol (Czycholl *et al* 2018, 2019a).

Certain indicators demonstrated inconclusive results with RS and ICC values exceeding threshold and SDC and LoA not doing so (abnormal breathing, nasal discharge, water functioning, skin lesion on the muzzle). This can be explained by specific advantages and disadvantages of the respective statistical parameters, ie according to de Vet *et al* (2006), the parameters, RS and ICC are dependent on the variance among the study objects for reliability (in this case, visits 1–4). Hence, if the prevalence of the visits is very similar, reliability might be underestimated, which occurred for these indicators in the present study. This can be interpreted as merely a statistical flaw, with these indicators still considered to present sufficient consistency over time in all respective time-periods being evaluated. Of course, in the long term, verification is needed through further studies, carried out preferably internationally to enhance the variability of the prevalence of these indicators. The results for the indicators, consistency of manure, cleanliness of the bedding and possibilities for social interaction were also inconclusive but, in so far as the reliability parameters, met the threshold values, ie a moderate reliability was detected without any exact agreement being detected by SDC and LoA. Again, this can be explained by the nature of the statistical parameters. Reliability implies a consistency of results although despite no exact agreement. This means that although the farms are not rated absolutely equally, the ranking of the farms still remains the same (de Vet *et al* 2006). Of course, although this is not perfect in terms of consistency over time, the most important issue as regards welfare assessment would be that the general ranking of specific establishments stays the same (Courboulay *et al* 2009), which would be the case for assessments with the AWIN protocol. So, consistency over time within 90 days can still be interpreted as sufficient.

© 2021 Universities Federation for Animal Welfare

https://doi.org/10.7120/09627286.30.1.081 Published online by Cambridge University Press

Figure 3



Qualitative Behaviour Assessment of the three different comparisons of visits for (a) visits 1 and 2, (b) visits 1 and 3 and (c) visits 1 and 4. The adjectives' position in the chart shows the factor loadings of the adjectives on the first two Principal Components (PC1 and PC2).

https://doi.org/10.7120/09627286.30.1.081 Published online by Cambridge University Press

Second-level protocol assessment

The results for the respective indicators are in agreement with those of the first-level protocol. This was to be expected given that a study comparing the outcomes of the two levels revealed generally good agreement to be present for most indicators (Czycholl *et al* 2018). For BCS 3 and 4, reliability could be improved in comparison with the results of the first-level assessment, supporting the hypothesis that scoring by palpation is especially helpful for this indicator and should be carried out in the first-level assessment (Czycholl *et al* 2018). The new indicators, coughing, lameness and lesions at the mouth corner were sufficiently consistent over time within the 90 days and, given that they were also proven to be of sufficient inter-observer reliability (Czycholl *et al* 2019a), can therefore be recommended for welfare assessment purposes. Moreover, these findings suggest that if these indicators are present, they refer to chronic problems. Coughing, as a clinical sign in acute diseases (bacterial or viral infection), tends to be accompanied by fever and general lethargy (Couetil *et al* 2007), which were not observed in any horses in this study. More often, coughing is associated with chronic lower airway problems, such as recurrent airway obstruction. The prevalence of these chronic inflammations of the lower airways, in which coughing can persist for months to years (Couetil *et al* 2007), has been described for 14 to 35% of horses in different studies (eg Wheeler *et al* 2002; Robinson *et al* 2010). Hence, it is a fairly common problem although these horses are not generally seen as sick by the owners. Lameness was more of a chronic problem and seen commonly in horses aged 15 years or above due to a reduced range of motion in joints and arthritic processes (Ireland *et al* 2011). According to the AWIN protocol (AWIN 2015), lesions at the mouth corner tend to be a sign of incorrect riding or inadequate equipment. The owners are most likely unaware of this problem, ensuring it persists over time. In contrast, the Forced Human Approach Test was of insufficient consistency over time in all time-periods under evaluation. Again, this is probably a general problem of behavioural tests (Miller *et al* 2006; Czycholl *et al* 2019b) making their use in welfare assessments questionable (Czycholl *et al* 2019b). For the QBA, the PCA revealed very good consistency over time in the comparison of visits 1 and 2 with a three-day interval in-between. Consistency over time, however, decreased first to an initially still-acceptable level after the time interval of 42 days. After the largest time interval of 90 days, the correlation was high between PC1 but very low for PC2. Regarding the factor loadings of the adjectives loading highly positively on PC1, it appears that adjectives describing a positive, relaxed emotional state (at ease, curious, friendly, relaxed) were highly positively loaded (using the definition of O'Rourke & Hatcher 2013), whereas adjectives describing a negative, stressed emotional state (aggressive, alarmed, annoyed, apathetic, fearful, uneasy) were highly negatively loaded. The loadings on PC2 were not straightforward: in visits 1 and 2, both negative descriptors (alarmed, annoyed, fearful, uneasy) as well as positive descriptors (at ease, curious, happy) were highly positively loaded. In visits 3 and 4, more negative descriptors were highly positively loaded (fearful, pushy, uneasy). It should be borne in mind that PC1 is usually a more general descriptor since it explains the majority of the variance (O'Rourke & Hatcher 2013). Moreover, in the use of QBA, PCA often reveals the valence of the emotion on PC1 and the level of arousal on PC2 (Wemelsfelder & Millard 2009; Temple *et al* 2013). In the present study, in accordance with this, PC1 describes the valence of the emotion. However, PC2 is less straightforward. Hence, detecting the level of arousal is problematic as regards consistency over time. A possible explanation for these findings is that the emotional state might have changed over the course of time, due, for example, to more time being spent on pastures in spring and summer than in winter. Another explanation might be that the QBA is not robust over long time intervals, not capturing general moods, but more faster changing emotions. Looking at the use of QBA in horses, the finding that reliability was given for PC1 but not for PC2 was also present in the inter-observer reliability study of Czycholl *et al* (2019a). As the present study was carried out with one of the observers involved in that study, it might be that the observer was not consistent in the use of the adjectives that were particularly important for PC2. Interestingly, in the study of Minero *et al* (2018) for which other observers were used, similar factor loadings of the adjectives of PC1 were achieved compared to those on PC1 here, while those loading highly on PC2 differed. In Temple *et al*'s (2013) study on growing pigs, the authors found PC1 to be moderately correlated while PC2 was not of sufficient test-retest reliability. Nevertheless, consistency over time is present over a time-period of 42 days. Further studies are needed to identify the exact reasons and potential influencing factors.

Overall contemplation of the AWIN protocol

Altogether, use of most of the indicators of the AWIN protocol for horses was of sufficient consistency over time within the 90 days, proving its potential as a reliable welfare assessment tool. In considering other studies as regards the feasibility, validity and reliability of the AWIN protocol for horses, this study supports the hypothesis that, overall, the AWIN protocol is a useful tool for a harmonised welfare assessment, on-farm (Dalla Costa *et al* 2016a, 2017; Czycholl *et al* 2018, 2019a). Such a tool is of interest not only scientifically but also to the various stakeholders within the horse industry (Dalla Costa *et al* 2016a). Moreover, veterinary authorities and administrative organisations are also in need of such tools to help highlight welfare issues (Dalla Costa *et al* 2016a). To date, controls of farms by veterinary authorities often rely on resource- or management-based indicators (European Food Safety Authority [EFSA] 2012) and are not standardised in any way. This leads to a certain arbitrariness and impenetrability. This study, along with existing literature concerning the AWIN protocol for horses (eg Dai *et al* 2015; Dalla Costa *et al* 2015, 2016a, 2017; Czycholl *et al* 2017b, 2018, 2019a), reveals that this standardised approach might be

© 2021 Universities Federation for Animal Welfare

https://doi.org/10.7120/09627286.30.1.081 Published online by Cambridge University Press

possible with the AWIN protocol for horses. As in other species, a focus on the improvement of one standardised assessment method might be more purposeful and efficient than the simultaneous development of multiple assessment tools. In order to be feasible, Knierim and Winckler (2009) stated that for the practical application of welfare assessment tools, results should remain consistent within a period of six months. This study proved most indicators at both levels of the AWIN protocol to be consistent within the 90 days. Hence, while working on the further improvement of the robustness, sensitivity and specificity of certain indicators, this is the time-frame that is advisable for practical application to date.

Animal welfare implications

Consistency over time is one of the basic requirements in welfare assessment tools if they are to be feasible, eg for objective and scientifically sound labelling purposes. This paper evaluates the consistency of the use of the indicators of the AWIN protocol for horses over time and thus helps towards its revision and refinement by proving consistency over time for many indicators but also by detecting indicators (eg Horse Grimace Scale, behavioural tests) that need revision, refinement or replacement in the future. It thereby contributes towards the further development and practical implementation of a feasible, reliable and objective welfare assessment for horses.

Conclusion

Although for most indicators sufficient consistency over time was detected after the 90-day period, some problematic indicators were also revealed, ie the Horse Grimace Scale and the behavioural tests (Avoidance Distance Test, Voluntary Human Approach Test, Forced Human Approach Test) cannot be recommended for use as on-farm welfare assessment tools. Further, the indicators, swollen joints and BCS 3 and 4 were of insufficient consistency over time in all time-periods under evaluation. Palpation should be carried out for these indicators, in addition to visual inspection. Moreover, the indicators, hoof neglect, alopecia on the legs and water cleanliness were of insufficient consistency over time in all evaluated time-periods. However, this was most likely due to genuine change in these indicators. Nevertheless, revision of the scoring criteria for these indicators might help to avoid oversensitivity in practical use. The QBA was only consistent for the period of 42 days, but not for the 90 days. In general, this study demonstrates which AWIN protocol indicators for horses are sufficiently consistent over time within 90 days for a feasible welfare assessment under practical conditions. For the problematic indicators, revision or else replacement with more suitable alternatives is required in the future to ensure that all aspects of animal welfare are adequately assessed.

References

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

Blokhuis H, Jones B, Veissier I and Miele M 2013 Improving farm animal welfare. In: Blokhuis H, Miele M, Veissier I and Jones B (eds) *Improving Farm Animal Welfare* pp 1-13. Wageningen Academic Publishers: Wageningen: The Netherlands. https://doi.org/10.3920/978-90-8686-770-7_1

BMELV 1995 Leitlinien zur Beurteilung von Pferdehaltungen unter Tierschutzgesichtspunkten. *Bundesministerium für Ernährung, Landwirtschaft und Verbraucherschutz (Sachverständigengruppe artgerechte Pferdehaltung, BMELV).* BMELV: Bonn, Germany. [Title translation: Guideline for the assessment of horse husbandry with regard to animal welfare]

Chamove AS, Crawley-Hartrick OJE and Stafford KJ 2002 Horse reactions to human attitudes and behavior. *Anthrozoös 15*: 323-331. https://doi.org/10.2752/089279302786992423

Couetil LL, Hoffmann AM, Hodgson J, Buechner-Maxwell V, Viel L, Wood JLN and Lavoie JP 2007 Inflammatory airway disease of horses. *Journal of Veterinary Internal Medicine 21*: 356-361. https://doi.org/10.1111/j.1939-1676.2007.tb02975.x

Courboulay V, Eugene A and Delarue E 2009 Welfare assessment in 82 pig farms. Effect of animal age and floor type on behaviour and injuries in fattening pigs. *Animal Welfare 18*: 515-521

Czycholl I, Büttner K, Grosse Beilage E and Krieter J 2015 Review of the assessment of animal welfare with special emphasis on the Welfare Quality® animal welfare assessment protocol for growing pigs. *Archives Animal Breeding 58*: 237-249. https://doi.org/10.5194/aab-58-237-2015

Czycholl I, Büttner K, Klingbeil P and Krieter J 2018 An indication of reliability of the two-level approach of the AWIN Welfare assessment protocol for horses. *Animals 8*: 7. https://doi.org/10.3390/ani8010007

Czycholl I, Grosse Beilage E, Henning C and Krieter J 2017a Reliability of the qualitative behavior assessment as included in the Welfare Quality® Assessment protocol for growing pigs. *Journal of Animal Science 95*: 3445-3454. https://doi.org/10.2527/jas.2017.1525

Czycholl I, Klingbeil P and Krieter J 2017b Suitability of animal-based indicators for the detection of animal welfare in horses. *Tieraerztliche Umschau 6*: 209-217

Czycholl I, Klingbeil P and Krieter J 2019a Interobserver reliability of the AWIN welfare assessment protocol for horses. *Journal of Equine Veterinary Science 75*: 112-121. https://doi.org/10.1016/j.jevs.2019.02.005

Czycholl I, Menke S, Straßburg C and Krieter J 2019b Reliability of different behavioural tests for growing pigs on-farm. *Applied Animal Behaviour Science 213*: 65-73. https://doi.org/10.1016/j.applanim.2019.02.004

Dai F, Cogi NH, Heinzl EUL, Dalla Costa E, Canali E and Minero M 2015 Validation of a fear test in sport horses using infrared thermography. *Journal of Veterinary Behavior: Clinical Applications and Research 10*: 128-136. https://doi.org/10.1016/j.jveb.2014.12.001

https://doi.org/10.7120/09627286.30.1.081 Published online by Cambridge University Press

Dalla Costa E, Dai F, Lebelt D, Scholz P, Barbieri S and Canali E 2016a Welfare assessment of horses: the AWIN approach. *Animal Welfare 25*: 481-488. https://doi.org/10.7120/ 09627286.25.4.481

Dalla Costa E, Dai F, Lebelt D, Scholz P, Barbieri S, Canali E and Minero M 2017 Initial outcomes of a harmonized approach to collect welfare data in sport and leisure horses. *Animal 11*: 254-260. https://doi.org/10.1017/S1751731116001452

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, Minero M, Lebelt D, Stucke D, Canali E and Leach MC 2014 Development of the Horse Grimace Scale (HGS) as a pain assessment tool in horses undergoing routine castration. *PLoS One 9*: e92281. https://doi.org/10.1371 /journal.pone.0092281

Dalla Costa E, Stucke D, Dai F, Minero M, Leach MC and Lebelt D 2016b Using the horse grimace scale (HGS) to assess pain associated with acute laminitis in horses (*Equus caballus*). *Animals 6*: 47. https://doi.org/10.3390/ani6080047

de Vet HCW, Terwee CB, Knol DL and Bouter LM 2006 When to use agreement versus reliability measures. *Journal of Clinical Epidemiology 59*: 1033-1039. https://doi.org/ 10.1016/j.jclinepi.2005.10.015

European Food Safety Authority (EFSA) 2012 EFSA Panel on Animal Health and Welfare Statement on the use of animal-based measures to assess the welfare of animals. *EFSA Journal 10*: 1-29. https://doi.org/10.2903/j.efsa.2012.2512

Fraser D 2008 Understanding animal welfare. *Acta Veterinaria Scandinavica 50*: S1. https://doi.org/10.1186/1751-0147-50-S1-S1

Ireland JL, Clegg PD, McGowan CM, McKane SA, Chandle KJ and Pinchbeck GL 2011 Disease prevalence in geriatric horses in the United Kingdom: Veterinary clinical assessment of 200 cases. *Equine Veterinary Journal 44(1)*: 101-106. https://doi.org/10.1111/j.2042-3306.2010.00361.x

Jezierski T, Jaworski Z and Górecka A 1999 Effects of handling on behaviour and heart rate in Konik horses: comparison of stable and forest reared youngstock. *Applied Animal Behaviour Science 62*: 1-11. https://doi.org/10.1016/S0168-1591(98)00209-3

Knierim U and Winckler C 2009 On-farm welfare assessment in cattle. Validity, reliability and feasibility issues and future perspectives with special regard to the Welfare Quality® approach. *Animal Welfare 18*: 451-458

Martin P and Bateson P 2007 *Measuring Behaviour. An Introductory Guide*. Cambridge University Press: Cambridge, UK. https://doi.org/10.1017/CBO9780511810893

McCann JS, Heird JC, Bell RW and Lutherer LO 1988 Normal and more highly reactive horses. Part II: the effect of handling and reserpine on the cardiac response to stimuli. *Applied Animal Behaviour Science 19*: 215-226. https://doi.org/ 10.1016/0168-1591(88)90002-0

McGraw KO and Wong SP 1996 Forming inferences about some intraclass correlation coefficients. *Psychological Methods 1(1)*: 30-46. https://doi.org/10.1037/1082-989X.1.1.30

Mieloch FJ, Nietfeld S, Straßburg C, Krieter J, grosse Beilage E and Czycholl I 2020 Factors of potential influence on different behavioural tests in fattening pigs. *Applied Animal Behaviour Science 222*: 104900. https://doi.org/10.1016/j.applanim.2019.104900

Miller KA, Garner JP and Mench JA 2006 Is fearfulness a trait that can be measured with behavioural tests? A validation of four fear tests for Japanese quail. *Animal Behaviour 71*: 1323-1334. https://doi.org/10.1016/j.anbehav.2005.08.018

Minero M, Dalla Costa E, Dai F, Canali E, Barbieri S and Zanella A 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

O'Rourke N and Hatcher L 2013 *Factor Analysis and Structural Equation Modeling*. SAS: Cary, NC, USA

Robinson NE, Karmaus W, Holcombe SJ, Carr EA and Derksen FJ 2010 Airway inflammation in Michigan pleasure horses: prevalence and risk factors. *Equine Veterinary Journal 38(4)*: 293-299. https://doi.org/10.2746/042516406777749281

SAS Institute 2008 *SAS/STAT 9.2. User's Guide*. SAS Institute: Cary, NC, USA

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

Thatcher CD, Pleasant RS, Geor RJ and Elvinger F 2012 Prevalence of over-conditioning in mature horses in Southwest Virginia during the summer. *Journal of Veterinary Internal Medicine 26*: 1413-1418. https://doi.org/10.1111/j.1939-1676.2012.00995.x

TierSchG 2006 *Tierschutzgesetz in der Fassung der Bekanntmachung vom 18 Mai 2006 (BGBl. I S. 1206, 1313), das zuletzt durch Artikel 280 der Verordnung vom 19 Juni 2020 (BGBl I S 1328) geändert worden ist.* [Title translation: German animal protection directive]

TierSchVersV 2013 *Tierschutz-Versuchstierverordnung vom 1 August 2013 (BGBl I S 3125, 3126), die zuletzt durch Artikel 394 der Verordnung vom 31 August 2015 (BGBl I S 1474) geändert worden ist.* [Title translation: German experimental animals directive]

Velarde A and Geers R 2007 *On Farm Monitoring of Pig Welfare*. AE Wageningen: Gelderland, The Netherlands. https://doi.org/10.3920/978-90-8686-591-8

Visser EK, Van Reenen CG, Hopster H, Schilder MBH, Knaap JH, Barneveld A and Blokhuis HJ 2001 Quantifying aspects of young horses' temperament: consistency of behavioural variables. *Applied Animal Behaviour Science 74*: 241-258. https://doi.org/10.1016/S0168-1591(01)00177-0

Wemelsfelder F and Lawrence AB 2001 Qualitative assessment of animal behaviour as an on-farm welfare-monitoring tool. *Acta Agriculturae Scandinavice Section A-Animal Science 51*: 21-25. https://doi.org/10.1080/090647001316923018

Wemelsfelder F and Millard F 2009 Qualitative Behaviour Assessment. In: Forkmann B and Keeling L (eds) *Welfare Quality® Reports*. SLU Service/Reproenheten: Uppsala, Sweden

Wheeler RG, Christley RM and McGowan 2002 Prevalence of owner reported respiratory disease in Pony Club horses. *The Veterinary Record 150*: 79-80. https://doi.org /10.1136/vr.150.3.79

© *2021 Universities Federation for Animal Welfare*

https://doi.org/10.7120/09627286.30.1.081 Published online by Cambridge University Press