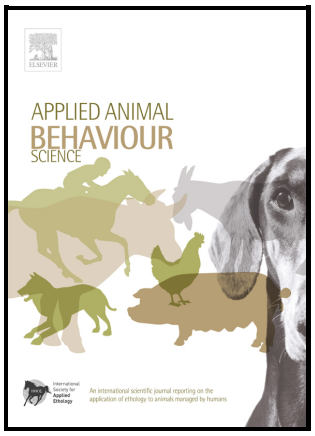
Do you see what I see? Investigating the validity of an equine personality questionnaire.

Aurélie Jolivald, Kelly Yarnell, Carol Hall, Carrie Ijichi



PII: S0168-1591(22)00025-9

DOI: https://doi.org/10.1016/j.applanim.2022.105567

Reference: APPLAN105567

To appear in: *Applied Animal Behaviour Science*

Received date: 29 November 2021
Revised date: 21 January 2022
Accepted date: 23 January 2022

Please cite this article as: Aurélie Jolivald, Kelly Yarnell, Carol Hall and Carrie Ijichi, Do you see what I see? Investigating the validity of an equine personality questionnaire., *Applied Animal Behaviour Science,* (2021) doi:https://doi.org/10.1016/j.applanim.2022.105567

Do you see what I see? Investigating the validity of an equine personality questionnaire.

Aurélie Jolivald[1,*], Kelly Yarnell[1], Carol Hall[1], Carrie Ijichi[1]

[1] School of Animal, Rural and Environmental Science, Nottingham Trent University, Southwell, NG25 0QF.

* Correspondence: aurelie.jolivald@ntu.ac.uk.

Abstract

Subjective equine personality questionnaires have the potential to predict a range of industry-relevant outcomes including fear reactivity, compliance with human cues, pain expression and susceptibility to stereotypies, in a time- and cost-efficient manner. However, to produce meaningful measures of target animals' behavioural tendencies, subjective personality assessment tools must satisfy four criteria: internal consistency, predictive validity, inter-rater reliability, and test-retest reliability. The Equine Personality Test (EPT) has been developed to assess horses on five personality factors based on trait ratings from a familiar observer. While the EPT has been shown to have predictive validity, it has not been assessed for internal consistency, inter-rater reliability or test-retest reliability. To this end, three experienced primary caregivers and three riding instructors assessed 25 familiar horses using the EPT. The internal consistency, inter-rater reliability and test-retest reliability of the five subscales of the EPT were investigated using Cronbach's α and intra-class correlation (ICC) analyses. The Agreeableness, Neuroticism, Extroversion and Gregariousness towards People subscales had high Cronbach α

and inter-rater and test-retest ICC coefficients (α> 0.7; ICC>0.8). By contrast, the Gregariousness towards Horses subscale had low Cronbach α (α=0.39) and inter-rater ICC coefficient (ICC=0.498), and an adequate test-retest ICC coefficient (ICC=0.784). Primary caregivers had higher ICC coefficients than instructors for most subscales and questionnaire items. The EPT therefore provides internally consistent and highly reliable measures of Agreeableness, Neuroticism, Extroversion, and Gregariousness towards People in equines, although measures of Gregariousness towards Horses should be interpreted with caution. The reliability of EPT scores can be further improved by targeting primary caregivers as raters. Taken together with previous findings demonstrating predictive validity for the questionnaire, these results contribute to making the EPT the only subjective equine personality questionnaire to have been checked against all four criteria of a valid and reliable personality assessment tool. This positions the EPT as a highly relevant equine personality assessment tool that may be used to predict behavioural tendencies in industry or research settings alike.

Keywords: subjective questionnaire; horse; internal consistency; reliability; inter-rater; test-retest.

1. Introduction

Horses show individual differences in behaviour that are stable across time and situations, referred to as personality (Gosling, 2001; König von Borstel, 2013). These differences in behaviour are linked with a range of outcomes such as health (McClure, Glickman and Glickman, 1999), learning style (Valenchon *et al.*, 2013; Lansade *et al.*, 2017) and suitability for a particular type of work (Pierard, McGreevy and Geers, 2017). For instance, Neuroticism predicts the magnitude of flight responses in the horse, which has important safety implications (Ijichi *et al.*, 2013), and may explain individual differences in pain and stereotypy vulnerability (Ijichi, Collins and Elwood, 2013, 2014). These industry-relevant outcomes may be predicted through accurate personality assessment. Equine personality may be assessed subjectively through trait ratings provided by a familiar observer (Momozawa *et al.*, 2003; Lloyd *et al.*, 2007; Ijichi *et al.*, 2013). This allows quick assessments without the logistical difficulties associated with standardised behavioural testing (Gosling, 2001). Subjective trait ratings also rely on a broader overview of the target individual's patterns of behaviour (Gosling, 2001) and may therefore provide a more reliable insight into personality than objective behavioural coding (Vazire *et al.*, 2007).

Subjective questionnaires must satisfy four criteria of validity and reliability to meaningfully assess personality (Gosling and Vazire, 2002; Simms and Watson, 2007). First, personality factors are unidimensional and internally coherent constructs (Simms and Watson, 2007). Therefore, all items in the questionnaire should measure the same underlying construct (homogeneity) and produce sufficiently consistent scores (internal consistency). In addition, scores produced by

the questionnaire should be valid, i.e. reflect the expression of conceptually related behaviour (concurrent validity) as well as wider outcomes (predictive validity) (Gosling and Vazire, 2002). Finally, the personality assessment tool should satisfy several aspects of reliability. Scores should reflect the animal's inherent behavioural tendencies rather than the rater's biases or implicit theories of personality (Gosling and Vazire, 2002). Therefore, independent raters should agree in their ratings of a given target individual (inter-rater reliability) (Gosling, 2001). In addition to the quality of the psychometric instrument itself, reliability of scoring may be impacted by the degree of familiarity of the rater with the target animal and the variety of contexts in which the target animal could be observed (Gosling, 2001). This has been tentatively identified in the horse (Lloyd *et al.*, 2007). Finally, personality represents "temporally stable patterns of affect, cognition, and behaviour" (Gosling, 2008). Therefore, repeated testing of the same adult animal by the same rater should yield consistent scores (test-retest reliability). To date, no published equine personality questionnaire has been checked against all four criteria: predicting real-world outcomes, internal consistency, inter-rater reliability and test-retest reliability.

The Equine Personality Test (EPT) provides a trait-based assessment of horses on 5 personality factors: Agreeableness, Neuroticism, Extroversion, Gregariousness towards People and Gregariousness towards Horse (Ijichi *et al.*, 2013). While other questionnaire-based personality tests are available for equines (Le Scolan, Hausberger and Wolff, 1997; Morris, Gale and Howe, 2002; Seaman, Davidson and Waran, 2002; Momozawa *et al.*, 2003; Lloyd *et al.*, 2007), the EPT offers several advantages over these. Most of these questionnaires were not designed using formal scale construction methods recognised by psychometric research (Le Scolan,

Hausberger and Wolff, 1997; Seaman, Davidson and Waran, 2002; Momozawa *et al.*, 2003) or are not specific to the target species (Morris, Gale and Howe, 2002). By contrast, the EPT was developed specifically for horsesd using formal scale construction methods based on psychometric research (Ijichi *et al.*, 2013). In addition, traits are organised in a factor-based model of equine personality and the resulting factors are framed using the common language of personality assessment in multiple species, thus facilitating within and cross-species comparisons (Gosling, 2001).

The EPT has been shown to have good concurrent and predictive validity. Scores on the Neuroticism and Extraversion scales of the EPT predict conceptually related behaviour in standardised behavioural tests (Ijichi *et al.*, 2013), as well as the expression of horses' pain responses in a veterinary context (Ijichi, Collins and Elwood, 2014). However, the internal consistency, inter-rater reliability and test-retest reliability of the EPT have not yet been investigated. This study therefore aimed to investigate how well the EPT meets the criteria for internal consistency and reliability for a personality assessment tool. To this end, personality data of 25 horses was collected from 6 raters using the EPT and used to compute indices of internal consistency, inter-rater reliability and test-retest reliability for the EPT.

2. Materials and Methods

Ethics

This study was approved by the Nottingham Trent University ethical review process. Raters used for this study were over the age of 18 and no personal or sensitive

information was collected. Data was stored anonymously according to GDPR legislation. Animal subjects were not tested or manipulated in any way for the purpose of this study.

2.1. Horses

Personality data was collected for 25 adult riding school horses housed at Nottingham Trent University's Brackenhurst campus (9 mares, 16 geldings; mean age: 14.0±4.1 years). Ten breeds were represented, including Irish Sports Horse (n=5), Connemara (n=4), Cob (n=3), British Warmblood (n=2) and Thoroughbred (n=2). All horses were used to teach equitation and horse management during the academic year (October-May) and for non-invasive research all year round. All horses lived on the same premises and were kept under a similar management regime tailored to individual requirements. During the academic year, horses were kept indoors in individual stables (n=21) or in pairs in outdoor paddocks with field shelters (n=4) during the day and turned out at night when the weather allowed. Horses received a minimum of 1 hour exercise or turnout daily, and workload did not exceed 3 hours of ridden work per day, 5 days a week. They were fed forage (hay or haylage) according to body weight as well as a balancer in the form of hard feed and had ad lib access to water. During the COVID-19 lockdown and the university summer break horses were turned out to pasture in pre-established groups. They had ad lib access to forage (grass, supplemented with hay where necessary) and water. They were not exercised from March to August, then were exercised following a gradual program aiming to build fitness back up from August to September.

2.2. Raters

Six raters were recruited via email from a pool of Nottingham Trent University staff meeting two inclusion criteria aimed at maximising the accuracy of the personality assessment. First, raters had to be familiar with the horses in the sample. All had known the horses for a minimum of two years and interacted with them on a daily to weekly basis. In addition, raters recruited had to possess a strong knowledge of the species as a whole (Gosling, 2001). All raters had 10+ years of professional experience in the equine industry. Three raters were members of the technical team responsible for the day-to-day care of the horses (referred to thereafter as “primary caregivers”). One of those raters later indicated that she had completed some of the questionnaires with input from a fourth primary caregiver, who also met the inclusion criteria. Those jointly-assigned scores were subsequently excluded from analysis (see sections 2.3 and 2.4.1 for further detail). Three additional raters were academics who regularly used the horses in the sample to teach management and riding lessons (referred to thereafter as “instructors”).

2.3. Subjective ratings collection

Upon inclusion in the study all raters were sent a link to digital questionnaire forms for each horse. Instructions for filling in the questionnaire was enclosed within each form. Further explanations and technical guidance regarding the online files could be obtained from the experimenter upon request. Raters were instructed to fill in the questionnaire only for horses they felt confident they were familiar with. One primary caregiver completed the questionnaires in August 2019, while the remaining five raters all completed them in April 2020. All raters were asked not to discuss their

assessments of the horses amongst themselves and to the best of our knowledge completed the questionnaire independently from each other.

In October 2020, two of the primary caregivers were invited to fill in the questionnaires a second time after a delay of six months to provide data for a test-retest reliability analysis. The primary caregiver who had filled in the first batch of questionnaires in consultation with a third party was not included in this analysis as the extent of the input from the fourth caregiver was unknown and could not be reproduced to ensure the test and retest responses were comparable. In addition, instructors were not asked to take part in the test-retest analysis as they had not had regular contact with the horses in the intervening six months. This was due to the impact of the COVID-19 lockdown (April-June) followed by the University summer break (June-October). Online access to the first batch of questionnaires was restricted ahead of raters' recruitment for the second batch to ensure test and retest responses were independent from each other. While available throughout to answer technical questions, the experimenter did not provide any input in the personality assessment.

The full text of the Equine Personality Test is available as Supplementary material in Ijichi *et al.* (2013) and was used without modifications. Briefly, horses were rated on 22 personality traits, divided into two sections. In the first section, raters described the target individual by placing a mark along a visual analogue scale between 12 pairs of opposite adjectives (e.g. Spirited/Steady). Scores between 1 and 5 were obtained for traits in this section as described in Ijichi et al. (2013). In the second section, 5-points Likert scales were used to rate horses on a further 7 traits. Traits were presented in randomised order and the polarity of pairs of adjectives was

randomly reversed to reduce the risk of superficial scoring (Ijichi *et al.*, 2013). Scores for the personality factors were obtained by averaging the scores of all questionnaire items pertaining to each factor, rather than adding them as recommended in Ijichi et al. (2013). This was because raters had overlooked one or more questions in some questionnaires (n=10 questionnaires). Because the questionnaire has yet not been shown to produce consistent responses over time (test-retest reliability), it was deemed inappropriate to collect this missing data in a second sitting. Discarding the questionnaires altogether would have led to an important reduction in sample size for inter-rater and test-retest reliability studies (n=11 horses with a full set of questionnaires completed by all raters) as the statistic used does not tolerate missing data. Therefore, the decision was made that an average of all scores available for each factor would be used, on the condition that factors showed good homogeneity. Continuous scores between 1 and 5 with decimals to 2 places were therefore obtained for all five personality factors.

2.4. Statistical analysis

Statistical analysis was carried out using SPSS statistical package version 26 (SPSS Inc, Chicago, IL) and R version 3.6.1 (2019-07-05) (R Core Team, 2019).

2.4.1. Internal consistency

Internal consistency was assessed using Cronbach's alpha (α). If a questionnaire has subscales, Cronbach's α must be applied to each subscale rather than the questionnaire as a whole (Simms and Watson, 2007). Therefore, for each rater a

value of Cronbach's α was calculated for each of the 5 subscales of the EPT. For each subscale, the mean ± standard deviation of the Cronbach's α's for the 6 raters were also calculated. Cronbach's α is sensitive to the direction of coding used for Likert-like data (Field, 2009). This analysis was therefore run using the coded data, with the relevant questions reversed, rather than the raw data from the questionnaires. Resulting Cronbach's αs were compared to published thresholds for acceptable internal consistency: a coefficient α higher than 0.7 is generally regarded as indicating acceptable internal consistency in a scale (Field, 2009). The homogeneity of each subscale was also evaluated by calculating the mean and distribution of inter-item correlation coefficients (Spearman correlations) (Simms and Watson, 2007). Values obtained for the subscales of the EPT were compared to published standards for scale homogeneity: a mean inter-item correlation between 0.15 and 0.5, with a distribution of coefficients closely clustered around the mean, indicates a homogenous scale (Simms and Watson, 2007). Conversely, significant variability in the correlation coefficients could indicate multidimensionality in the scale (Simms and Watson, 2007).

2.4.1. Inter-rater reliability

Inter-rater reliability was evaluated using intra-class correlation analysis (ICC). ICC is the recommended method to investigate inter-rater agreement in trait rating-based animal personality studies (Vazire *et al.*, 2007). The overall inter-rater reliability of our set of raters was first evaluated. The primary caregiver who filled in questionnaires in consultation with a third party was excluded as their scores could not be attributed with certainty to a single rater. A total of 5 raters were therefore

entered into this initial analysis (k_O=5). In addition, the inter-rater reliability of primary caregivers and instructors was also compared. To this end, separate ICC analyses were carried out using the scores given by the three primary caregivers on the one hand (k_{PC}=3), and those given by the three instructors on the other (k_I=3). Mean-rating (k_O=5; k_{PC}=3; k_I=3), absolute agreement, two-way random effect models were used throughout. Model selection was based on decision trees in (Hallgren, 2012; Koo and Li, 2016) after (McGraw and Wong, 1996). A two-way model was selected because all horses had been assessed by the same raters. Random effects were chosen because the raters recruited to this study were a random set of raters selected from a wider population. Finally, the model definition was set to reflect absolute agreement, rather than correlations, between scores given by the different raters.

Inter-rater reliability was assessed for each of the 5 subscales measuring the 5 personality dimensions, as these were the outcomes likely to be used in subsequent analysis. In addition, inter-rater reliability was also assessed for each questionnaire item separately, in order to identify if some items yielded particularly high levels of disagreement between raters. For each personality factor and questionnaire item ICC estimates and their 95% confidence interval were calculated. Interpretation in terms of inter-rater reliability for the subscale or item was carried out using the thresholds for poor (ICC<0.5), moderate (0.5<ICC<0.75), good (0.75<ICC<0.9) and excellent (ICC>0.9) agreement proposed by Koo and Li (2016).

2.4.3. Test-retest reliability

Test-retest reliability was assessed using ICC (McGraw and Wong, 1996; Koo and Li, 2016) for each of the 5 subscales measuring the 5 personality factors. It was also assessed for each questionnaire item separately to identify whether some items showed higher inconsistency over time. Two sets of scores, for test and retest, were obtained by averaging the scores given by the two primary caregivers who took part in the test-retest study on the questionnaires they completed in April and October, respectively. The scores for test and retest were then compared using a single-ratings, absolute agreement, two-way mixed effects model (McGraw and Wong, 1996). Model selection was guided by Koo and Li (2016), after Shrout and Fleiss (1979). For intra-rater reliability studies a two-way model is selected because all subjects are rated by the same raters, with mixed effects as rater selection is not random. In addition, absolute agreement rather than consistency should be evaluated when investigating intra-rater reliability. Here, single ratings rather than mean ratings were used, to account for the fact that in subsequent studies the personality scores used will likely only result from a single administration of the EPT rather than be averaged across multiple retests.

3. Results

All three instructors elected not to carry out the personality assessment for some of the horses as they did not feel they were familiar enough with them (n=4 horses with at least one assessment missing). By contrast, the three primary caregivers felt confident rating all 25 horses. In addition, the three instructors, who were familiar with the horses mostly in a ridden context, all expressed that they had found it challenging to score the horses on questionnaire items relating to their behaviour towards other horses. This concern was not shared by the primary caregivers.

3.2. Internal consistency and homogeneity of each personality factor subscale

Cronbach's α were consistent across raters for all personality factors, although they were more variable for Gregariousness towards Horses (Table 1). Cronbach's αs were higher than the threshold of 0.7 indicating good internal consistency (Field, 2009) for the subscales measuring Agreeableness, Neuroticism, Extroversion and Gregariousness towards People (Table 1). However, Cronbach's αs were low and well below the threshold of 0.7 for the Gregariousness towards Horses subscale. The Cronbach α procedure in SPSS automatically identifies items if their removal from the scale improves internal consistency. This procedure revealed that removing the item Q7: "Generally how dependable would you say this horse is?" resulted in an increase of Cronbach's α above the threshold for acceptable internal consistency for all 6 raters. Mean α across the 6 raters with Q7 removed was 0.77 ± 0.38, up from 0.39 ±0.15 when this item was included.

Mean inter-item correlation coefficient (± SD) was 0.64 ± 0.14 for Agreeableness, 0.61 ± 0.26 for Neuroticism, 0.56 ± 0.24 for Extroversion, 0.87 ± 0.06 for Gregariousness towards People, and 0.18 ± 0.39 for Gregariousness towards Horses. Inter-item correlation coefficients clustered relatively closely around the mean for Agreeableness, Neuroticism, Extroversion and Gregariousness towards People. However, there was much more variability for Gregariousness towards Horses.

3.3. Inter-rater agreement

3.3.1. Inter-rater agreement across the whole sample of 6 raters

ICC estimates for Agreeableness, Neuroticism, Extroversion and Gregariousness towards People were all higher than 0.75 (Table 2), indicating good inter-rater reliability for those four factors (Koo and Li, 2016). Based on the 95% confidence intervals for the ICC estimates for those 4 factors, the true level of reliability is moderate to excellent. However, the ICC estimate for Gregariousness towards Horses is lower than 0.5 (Table 2), and the 95% confidence interval indicates that inter-rater reliability for this factor is poor to moderate at best.

The average ICC across all questionnaire items was 0.66 ± 0.22, ranging from 0 to 0.869 (Table 2). Inter-rater agreement was good for 13 items (0.754 ≤ ICC ≤ 0.869) and moderate for another 5 (0.613 ≤ ICC ≤ 0.737). However, it was poor for 4 items, three of which related to the horse's behaviour towards other horses.

3.3.2. Caregivers vs. instructors comparison

Comparisons between the two groups revealed that overall primary caregivers showed better inter-rater reliability than instructors (Table 3). At the subscale level, primary caregivers had good inter-rater agreement for three of the five personality factors (Agreeableness, Neuroticism and Gregariousness towards People), and moderate agreement for the remaining two (Extroversion and Gregariousness towards Horses). By contrast, instructors only had good agreement for two factors (Neuroticism and Extroversion), while agreement was moderate for another two (Agreeableness and Gregariousness towards People), and poor for a third (Gregariousness towards Horses). ICC coefficients were higher for primary caregivers than they were for instructors for all factors expect Extroversion, indicating higher levels of inter-rater agreement within that group. The most obvious difference between groups was for Gregariousness towards Horses. For this factor the ICC coefficient was 0.562 for primary caregivers (moderate agreement), while it was only 0.391 for instructors (poor agreement). Similarly, at the item level, good levels of inter-rater agreement were observed more often for primary caregivers than for the instructors (Table 3). For primary caregivers, reliability was good for 8 questionnaire items, moderate for 12 and poor for only 2. By contrast, for instructors reliability was good for only 2 items, while it was moderate for 15 and poor for 5. For all but 5 items, ICC coefficients were higher for primary caregivers than instructors, indicating better levels of inter-rater agreement.

3.4. Test-retest reliability

The subscales measuring Neuroticism, Extroversion and Gregariousness towards People all had ICC estimates greater than 0.9 (Table 4), with 95% confidence intervals indicating that test-retest reliability for these subscales was good to excellent. However, the subscales measuring Agreeableness and Gregariousness towards Horses had ICC estimates greater than 0.75 but lower than 0.9. Their 95% confidence intervals indicated moderate to excellent test-retest reliability for Agreeableness but only moderate to good reliability for Gregariousness towards Horses (Table 4). At the items level, 14 items showed good or excellent test-retest reliability (ICC>0.75), while 8 performed more poorly, with 7 showing moderate reliability (0.5<ICC<0.75) and one showing poor reliability (ICC_{Q5}<0.5). The Agreeableness, Neuroticism and Extroversion subscales all had a majority of highly reliable items. However, all items on the Gregariousness towards Horses had poorer test-retest reliability (Table 4).

4. Discussion

The Equine Personality Test (EPT) has previously been shown to have good concurrent and predictive validity (Ijichi *et al.*, 2013). However, further checks on its internal consistency, inter-rater reliability, and test-retest reliability had not yet been carried out. The aim of this study was to evaluate the EPT's performance on these three criteria for the sample of horses and raters used in this study. To this end, six raters were asked to use the EPT to assess 25 horses, with two of these raters carrying out the assessment twice over a period of 6 months. Cronbach's α and

intra-class correlations analyses were used to analyse scale internal consistency and inter-rater and test-retest reliability, respectively. While the Agreeableness, Neuroticism, Extroversion and Gregariousness towards People subscales performed well on all three criteria, the Gregariousness towards Horses subscale proved more problematic.

The Agreeableness, Neuroticism, Extroversion and Gregariousness towards People subscales had Cronbach's α's greater than 0.7, indicating good internal consistency. These values of coefficient α are comparable to those obtained by Momozawa et al. (2005) for their equine personality questionnaire. In addition, the mean and distribution of their inter-item correlation coefficients indicated good homogeneity for those subscales. Taken together, these results suggest that these subscales are likely to measure a single underlying construct (Field, 2009). In contrast, the Gregariousness towards Horses subscale had a Cronbach's α well below the threshold for acceptable internal consistency. This may be due to the fact that this subscale is only comprised of three items, as Cronbach's α is negatively affected by the number of scale items (Cortina, 1993). However, the low mean and wide distribution of inter-item correlation coefficients also indicate potential multidimensionality in the scale (Simms and Watson, 2007). It therefore appears likely that there is heterogeneity in the underlying constructs measured by the scale. Indeed, for all 6 raters removing the item Q7: "Generally how dependable would you say this horse is?" resulted in an increase of Cronbach's α above the threshold for acceptable internal consistency. This might point to an issue with item selection for this scale. Therefore, the Agreeableness, Neuroticism, Extroversion and Gregariousness towards People subscales show good internal consistency and

homogeneity, reflecting the fact that all items on the subscales reflect the intended underlying personality construct. However, the Gregariousness towards Horses subscale may not be unidimensional and some items on that subscale may not accurately reflect this personality factor.

The inter-rater reliability analysis resulted in high ICC coefficients (ICC > 0.8) for the Agreeableness, Neuroticism, Extroversion and Gregariousness towards People subscales. This demonstrates good levels of agreement between raters compared to published thresholds for inter-rater reliability (Koo and Li, 2016). By comparison, ICC coefficients ranging from 0.28 (4 raters, Agreeableness) to 0.53 (2 raters, Neuroticism) are reported for the human NEO Personality Inventory (Mccrae and Costa, 1987). While good or acceptable inter-rater reliability has been reported for other equine personality questionnaires (Anderson *et al.*, 1999; Morris, Gale and Duffy, 2002; Lloyd *et al.*, 2007), direct comparisons with the EPT are challenging because these studies did not use ICC coefficients. However, average ICC coefficients of 0.62 and 0.79 have been reported for canine personality assessments (Gosling, Kwan and John, 2003; Ley, McGreevy and Bennett, 2009). Therefore, the first 4 subscales of the EPT show good inter-rater reliability compared to the published standards in human and domestic animal personality assessment. For three of those four subscales, and for most of the individual scale items making them up, primary caregivers achieved better inter-rater reliability than instructors. This was expected, as differential exposure to the target individual is known to affect inter-rater reliability: consistently being exposed to an animal in a particular context may limit the range of behaviour a judge has the opportunity to observe and can therefore influence their perception of the subject's personality (Funder, Kolar and

Blackman, 1995; Gosling, 2001). Instructors were most familiar with the horses while they were being ridden, a relatively narrow context in which behavioural expression is reduced and largely placed under the control of the rider (Hall *et al.*, 2008). However, differences in reliability between the two groups were relatively minimal. In addition, the ICC coefficients obtained by instructors remained well above published thresholds for acceptable agreement (Koo and Li, 2016), especially for the subscale level. Therefore, it appears that the restricted context in which they knew the horses, as well as the behavioural restrictions placed on ridden horses, did not significantly impede riding instructors' ability to reliably judge Agreeableness, Neuroticism and Extroversion. Overall, the Equine Personality Test therefore provides a highly reliable assessment of Agreeableness, Neuroticism, Extroversion, and Gregariousness towards People. Ratings are reliable even when provided by riding instructors who are familiar with the horses in a relatively narrow context. However, reliability is further improved when the ratings are provided by primary caregivers.

The ICC coefficient was low (ICC=0.498) for the Gregariousness towards Horses subscale, indicating poor inter-rater reliability (Koo and Li, 2016). This result may reflect difficulty on the part of the raters to assess Gregariousness towards Horses reliably. Indeed, at the trait level, items related to social behaviour towards other horses (Q3-5) also showed poor reliability. This might be because instructors, who made up the majority of the set of raters (k=3 out of 5), only knew the horses in a relatively narrow context where social behaviour is difficult to observe (Funder, Kolar and Blackman, 1995; Gosling, 2001). Indeed, in their informal feedback, instructors self-reported difficulty in scoring items relating to behaviour towards other horses. This was not the case for primary caregivers, who observe the horses in a much

wider set of circumstances, including when turned out in groups. However, rater familiarity was likely not the only factor driving the poor inter-rater reliability of the Gregariousness towards Horses subscale. While primary caregivers showed better agreement than instructors on this subscale, their ICC coefficients remained relatively low (ICC=0.562) and indicative of only moderate inter-rater reliability. This implies that even raters who had the opportunity to observe horses perform the relevant behaviour had difficulty in scoring those traits accurately. Funder (1995) suggests that some traits are inherently less observable, and therefore more difficult to rate than others. This could be the case here. It might also be that individual behavioural patterns on those traits show only limited stability across time and situations. Dominance rank has been shown to be non-linear in stable social groups (Houpt and Keiper, 1982), while affiliative relationships are developed with a network of preferred partners (Briard, Dorn and Petit, 2015). Therefore, horses' tendencies to initiate aversive or affiliative social contacts might depend on the identity of the social partner present, making it difficult even for familiar raters to generalise their behaviour across situations. Gregariousness towards Horses was therefore assessed with only limited reliability by the panel of raters in this study, due to limited familiarity with the target horses but also to apparent difficulty in rating those traits. For studies concerned specifically with Gregariousness towards Horses, it may be preferable to use only primary caregivers as raters; however, even in this case scores must be interpreted with caution.

In the test-retest reliability study, ICC coefficients were generally high for both subscales and individual items. This indicates that the scores given by raters using the EPT were consistent over time. Test-retest reliability was excellent for

personality factors Neuroticism, Extroversion and Gregariousness towards People, and good for Agreeableness. While a few items on those scales had more limited test-retest reliability, the majority of items were rated consistently across time, suggesting that rater's perception of individual horse's traits remained constant across time. This is consistent with the idea that in adult animals, personality should reflect "temporally stable patterns of affect, cognition, and behaviour" (Gosling, 2008). However, while the ICC coefficient for the Gregariousness towards Horses subscale was also relatively high and showed acceptable consistency across time, all three items on the scale only had moderate test-retest reliability when taken individually. This suggests that, unlike the previous four, ratings on this subscale might show acceptable but limited temporal stability. As discussed above, this might be due to rater's difficulty to rate even familiar horses on those traits. Therefore, our results suggest that personality ratings on Agreeableness, Neuroticism, Extroversion and Gregariousness towards People collected using the EPT can be generalised beyond the time of collection. However, ratings on Gregariousness towards Horses only showed acceptable test-retest reliability and should be interpreted with caution when generalised over time. Despite the importance of demonstrating that the individual characteristics measured as part of personality assessments are stable in time (Dingemanse and Wright, 2020), to the best of our knowledge, to date no other equine personality questionnaire had been assessed for test-retest reliability. This result therefore provides a benchmark for other questionnaires to be evaluated against.

5. Conclusions

This study shows that Agreeableness, Neuroticism, Extraversion and Gregariousness towards People are evaluated with satisfactory internal consistency, inter-rater reliability and test-retest reliability in the horse using the EPT. The Gregariousness towards Horses subscale proved to be problematic both in terms of internal consistency and reliability. Only primary caregivers showed acceptable if modest levels of agreement in their assessments of horses on this factor, and their assessment showed only limited consistency in time. Assessments on the Gregariousness towards Horses subscale should therefore be considered with caution. However, the questionnaire offers valid and reliable measures of the personality factors Agreeableness, Neuroticism, Extroversion, and Gregariousness towards People in equines. Taken together with previous findings demonstrating the predictive validity for two of the subscales, these results contribute to making the EPT the only subjective equine personality questionnaire to have been checked against all four criteria of a valid and reliable personality assessment tool. This positions the EPT as a relevant personality assessment tool in the horse, that may be used to predict industry-relevant outcomes such as fear reactivity and susceptibility to stereotypies both for applied and research contexts.

Author Contributions: Conceptualization, AJ., C.I., K.Y., C.H.; methodology, A.J.; validation, A.J.; formal analysis, A.J.; investigation, A.J.; resources, A.J., C.I., K.Y.; data curation, A.J.; writing—original draft preparation, A.J; writing—review and editing, C.I., K.Y., C.H.; visualization, A.J.; supervision, K.Y., C.I., C.H.; project administration, A.J.; funding acquisition, K.Y and C.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Nottingham Trent University.

Acknowledgments: We would like to thank our raters C. Rhoades, A. Hazelhurst, T. Canton, S. Hallam, L. Taylor, C. Hake and J. Bromley-Fowles for taking the time to provide us with personality ratings for the horses in our sample.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

Anderson, M.K. *et al.* (1999) "Behavioral assessment of horses in therapeutic riding programs," *Applied Animal Behaviour Science*, 63(1), pp. 11–24.

Briard, L., Dorn, C. and Petit, O. (2015) "Personality and Affinities Play a Key Role in the Organisation of Collective Movements in a Group of Domestic Horses," *Ethology*, 121(9), pp. 888–902. doi:10.1111/eth.12402.

Cortina, J.M. (1993) "What Is Coefficient Alpha? An Examination of Theory and Applications," *Journal of Applied Psychology*, 78(1), pp. 98–104.

Dingemanse, N.J. and Wright, J. (2020) "Criteria for acceptable studies of animal personality and behavioural syndromes," *Ethology*, 126(9), pp. 865–869. doi:10.1111/eth.13082.

Field, A. (2009) *Discovering statistics using SPSS*. 3rd editio, *Discovering statistics using SPSS*. 3rd editio. London: Sage.

Funder, D.C. (1995) *On the Accuracy of Personality Judgment: A Realistic Approach*, *Psychological Review*.

Funder, D.C., Kolar, D.C. and Blackman, M.C. (1995) “Agreement Among Judges of Personality: Interpersonal Relations, Similarity, and Acquaintanceship,” *Journal of Personality and Social Psychology*, 69(4), pp. 656–672. doi:10.1037/0022-3514.69.4.656.

Gosling, S.D. (2001) “From mice to men - What can we learn about personality from animal research,” *Psychological Bulletin*, 127(1), pp. 45–86.

Gosling, S.D. (2008) “Personality in Non-human Animals,” *Social and Personality Psychology Compass*, 22(10), pp. 985–1001. doi:10.1111/j.1751-9004.2008.00087.x.

Gosling, S.D., Kwan, V.S.Y. and John, O.P. (2003) “A Dog’s Got Personality: A Cross-Species Comparative Approach to Personality Judgments in Dogs and Humans,” *Journal of Personality and Social Psychology*, 85(6), pp. 1161–1169. doi:10.1037/0022-3514.85.6.1161.

Gosling, S.D. and Vazire, S. (2002) “Are we barking up the right tree? Evaluating a comparative approach to personality,” *Journal of Research in Personality*, 36(6), pp. 607–614. doi:10.1016/S0092-6566(02)00511-1.

Hall, C. *et al.* (2008) “Is There Evidence of Learned Helplessness in Horses?,” *JOURNAL OF APPLIED ANIMAL WELFARE SCIENCE*, 11(3), pp. 249–266. doi:10.1080/10888700802101130.

Hallgren, K.A. (2012) “Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial,” *Tutorials in Quantitative Methods for Psychology*, 8(1), pp. 23–34. doi:10.20982/tqmp.08.1.p023.

Houpt, K.A. and Keiper, R. (1982) “The Position of the Stallion in the Equine Dominance Hierarchy of Feral and Domestic Ponies,” *Journal of Animal Science*, 54(5), pp. 945–950. doi:10.2527/JAS1982.545945X.

Ijichi, C. *et al.* (2013) “Harnessing the power of personality assessment: subjective assessment predicts behaviour in horses,” *Behavioural Processes*, 96, pp. 47–52. doi:10.1016/j.beproc.2013.02.017.

Ijichi, C., Collins, L.M. and Elwood, R.W. (2014) “Pain expression is linked to personality in horses,” *Applied Animal Behaviour Science*, 152, pp. 38–43. doi:10.1016/j.applanim.2013.12.007.

Ijichi, C.L., Collins, L.M. and Elwood, R.W. (2013) “Evidence for the role of personality in stereotypy predisposition,” *Animal Behaviour*, 85, pp. 1145–1151. doi:10.1016/j.anbehav.2013.03.033.

König von Borstel, U. (2013) “Assessing and influencing personality for improvement of animal welfare: a review of equine studies.,” *CAB Reviews: Perspectives in Agriculture, Veterinary Science, Nutrition and Natural Resources*, 8(6), pp. 7–21. doi:10.1079/PAVSNNR20138006.

Koo, T.K. and Li, M.Y. (2016) “A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research,” *Journal of Chiropractic Medicine*, 15(2), pp. 155–163. doi:10.1016/j.jcm.2016.02.012.

Lansade, L. *et al.* (2017) “Personality and predisposition to form habit behaviours during instrumental conditioning in horses (Equus caballus),” *PLoS ONE*. Edited by N. Ravel, 12(2), p. e0171010. doi:10.1371/journal.pone.0171010.

Ley, J.M., McGreevy, P. and Bennett, P.C. (2009) “Inter-rater and test-retest reliability of the Monash Canine Personality Questionnaire-Revised (MCPQ-R),” *Applied Animal Behaviour Science*, 119(1–2), pp. 85–90. doi:10.1016/j.applanim.2009.02.027.

Lloyd, A.S. *et al.* (2007) “Evaluation of a novel method of horse personality assessment: Rater-agreement and links to behaviour,” *Applied Animal Behaviour Science*, 105(1–3), pp. 205–222. doi:10.1016/j.applanim.2006.05.017.

McClure, S.R., Glickman, L.T. and Glickman, N.W. (1999) “Prevalence of gastric ulcers in show horses.,” *Journal of the American Veterinary Medical Association*, 215(8), pp. 1130–3.

Mccrae, R.R. and Costa, P.T. (1987) “Validation of the Five-Factor Model of Personality Across Instruments and Observers,” *Journal of Personality and Social Psychology*, 52(1), pp. 81–90.

McGraw, K.O. and Wong, S.P. (1996) “Forming inferences about some intraclass correlation coefficients.,” *Psychological Methods*, 1(1), pp. 30–46. doi:10.1037/1082-989X.1.1.30.

Momozawa, Y. *et al.* (2003) “Assessment of equine temperament by a questionnaire survey to caretakers and evaluation of its reliability by simultaneous behavior test,” *Applied Animal Behaviour Science*, 84(84), pp. 127–138. doi:10.1016/j.applanim.2003.08.001.

Momozawa, Y. *et al.* (2005) “Assessment of equine temperament questionnaire by comparing factor structure between two separate surveys,” *Applied Animal Behaviour Science*, 92(1–2), pp. 77–84. doi:10.1016/j.applanim.2004.11.006.

Morris, P.H., Gale, A. and Duffy, K. (2002) “Can judges agree on the personality of horses?,” *Personality and Individual Differences*, 33(1), pp. 67–81.

Morris, P.H., Gale, A. and Howe, S. (2002) “The factor structure of horse personality,” *Anthrozoos*, 15(4), pp. 300–322. doi:10.2752/089279302786992414.

Pierard, M., McGreevy, P. and Geers, R. (2017) “Developing behavioral tests to support selection of police horses,” *Journal of Veterinary Behavior: Clinical Applications and Research*, 19, pp. 7–13. doi:10.1016/j.jveb.2017.01.005.

R Core Team. (2019). *R: A Language and Environment for Statistical Computing* (3.6.1 "Action of the Toes"). https://www.r-project.org/

Le Scolan, N., Hausberger, M. and Wolff, A. (1997) “Stability over situations in temperamental traits of horses as revealed by experimental and scoring approaches,” *Behavioural Processes*, 41(3), pp. 257–266. doi:10.1016/S0376-6357(97)00052-1.

Seaman, S.C., Davidson, H.P.B. and Waran, N.K. (2002) “How reliable is temperament assessment in the domestic horse (Equus caballus)?,” *Applied Animal Behaviour Science*, 78(2–4), pp. 175–191.

Shrout, P.E. and Fleiss, J.L. (1979) “Intraclass correlations: Uses in assessing rater reliability.,” *Psychological Bulletin*, 86(2), pp. 420–428. doi:10.1037/0033-2909.86.2.420.

Simms, L.J. and Watson, D. (2007) “The Construct Validation Approach to Personality Scale Construction,” in Robins, R.W., Fraley, R.C., and Krueger, R.F. (eds) *Handbook of research methods in personality psychology*. New York: The Guilford Press, pp. 240–258.

Valenchon, M. *et al.* (2013) “Characterization of long-term memory, resistance to extinction, and influence of temperament during two instrumental tasks in horses,” *Animal Cognition*, 16, pp. 1001–1006. doi:10.1007/s10071-013-0648-5.

Vazire, S. *et al.* (2007) “Measuring Personality in Nonhuman Animals,” in Robins, R.W., Fraley, R.C., and Krueger, R.F. (eds) *Handbook of research methods in personality psychology*. New York: The Guilford Press, pp. 190–206.

Table 1 – Cronbach’s α for each personality factor and each rater (PC1-3: 3 primary caregivers to the horses; I1-3: 3 riding instructors familiar with the horses), obtained by assessing personality in n=25 horses using the Equine Personality Test. For each factor, the lowest α between the 6 raters is presented in italics and the highest in bold; the mean and standard deviation of Cronbach’s α for the 6 raters are also presented. Cronbach’s αs higher than 0.7 indicate good internal consistency (Field, 2009).

	PC1	**PC2**	**PC3**	**I1**	**I2**	**I3**	**Mean**	**St Dev**
Agreeableness	*0.76*	0.80	0.93	0.78	0.86	**0.96**	0.85	0.08
Neuroticism	0.85	**0.89**	0.83	0.79	*0.78*	0.81	0.83	0.04
Extroversion	*0.81*	0.83	0.85	**0.92**	0.90	0.83	0.86	0.04
Greg. Towards People	*0.84*	0.92	0.89	0.92	**0.96**	0.79	0.89	0.06
Greg. Towards Horses	0.42	0.39	0.49	**0.54**	*0.12*	0.41	0.39	0.15

Table 2 - Results of intra-class correlation analyses for all subscales and individual items in the Equine Personality Test administered to n=25 horses by 5 familiar raters. A mean-rating (*k*=5), absolute-agreement, two-way random-effects model was used. For each subscale or item the ICC estimate, 95% confidence interval, and interpretation in terms of inter-rater reliability (ICC<0.5: poor; 0.5<ICC<0.75: moderate; 0.75<ICC<0.9: good; ICC>0.9: excellent) are presented. The subscript [?] highlights subscales or items for which the ICC estimate indicates poor inter-rater reliability.

			95% confidence interval		
	n	**ICC**	Lower	Upper	**Reliability**
Agreeableness	21	0.848	0.715	0.930	Moderate to excellent
Easy-going/Intolerant	21	0.869	0.754	0.940	Good to excellent
Argumentative/Well-mannered	21	0.774	0.578	0.869	Moderate to good
Obedient/Wayward	21	0.788	0.607	0.902	Moderate to excellent
Willing/Stubborn	21	0.764	0.559	0.892	Moderate to good
Gentle/Rough	21	0.613	0.290	0.820	Poor to good
Neuroticism	2	0.84	0.715	0.930	Moderate to

	1	8			excellent
Anxious/Confident	2 1	0.83 1	0.679	0.923	Moderate to excellent
Nervous/Calm	2 1	0.78 2	0.593	0.900	Moderate to good
Relaxed/Tense	2 0	0.76 5	0.553	0.895	Moderate to good
Quiet/Restless	1 9	0.77 1	0.552	0.900	Moderate to good
How fearful is this horse around other horses?	2 0	-0.0 11[?]	-0.470	0.443	Poor
Extroversion	2 1	0.80 6	0.640	0.911	Moderate to excellent
Sluggish/Forward-going	2 1	0.66 3	0.388	0.842	Poor to good
Placid/Active	2 0	0.79 8	0.618	0.909	Moderate to excellent
Adventurous/Habitual	2 1	0.48 7[?]	0.036	0.736	Poor to moderate
Excitable/Laid-back	1 7	0.70 5	0.415	0.878	Poor to good
Spirited/Steady	2 0	0.75 7	0.544	0.890	Moderate to good
How energetic would you say this horse is?	2 0	0.69 6	0.423	0.863	Poor to good
Gregariousness towards people	2	0.82	0.681	0.921	Moderate to

	1	9			excellent
Friendly/Standoffish	2 0	0.79 7	0.611	0.909	Moderate to excellent
How often does this horse initiate interaction with you	2 1	0.75 4	0.539	0.887	Moderate to good
How often does this horse initiate interaction with other people	2 1	0.77 0	0.562	0.895	Moderate to good
Gregariousness towards horses	2 1	0.49 8?	0.140	0.752	Poor to moderate
How often does this horse initiate interaction with other horses?	2 1	0.34 6?	0.015	0.643	Poor to moderate
Does this horse ever show affection towards other horses?	2 1	0.18 2?	-0.095	0.499	Poor
How dependable would you say this horse is?	2 0	0.73 7	0.501	0.884	Moderate to good

Table 3 – Comparison of the inter-rater reliability of primary caregivers vs. instructors for all subscales and individual items in the Equine Personality Test administered to n=25 horses by 3 primary caregivers and 3 familiar riding instructors. Mean-rating (*k*=3), absolute-agreement, two-way random-effects models were used to carry out separate intra-class correlation analyses for the two groups. For each subscale or item the sample size, ICC estimate, and interpretation in terms of inter-rater reliability (ICC<0.5: poor; 0.5<ICC<0.75: moderate; 0.75<ICC<0.9: good; ICC>0.9: excellent) are presented. Italics highlight subscales or items for which instructors had better inter-rater reliability than primary caregivers.

	Primary caregivers			Instructors		
	n	ICC	Reliability	n	ICC	Reliability
Agreeableness	25	0.849	Good	21	0.724	Moderate
Easy-going/Intolerant	*25*	*0.87*	*Good*	*21*	*0.95*	*Good*
Argumentative/Well-mannered	25	0.807	Good	21	0.584	Moderate
Obedient/Wayward	25	0.824	Good	21	0.594	Moderate
Willing/Stubborn	*25*	*0.520*	*Moderate*	*21*	*0.571*	*Moderate*
Gentle/Rough	25	0.733	Moderate	21	0.341	Poor
Neuroticism	25	0.792	Good	21	0.777	Good
Anxious/Confident	25	0.791	Good	21	0.694	Moderate
Nervous/Calm	25	0.736	Moderate	21	0.618	Moderate
Relaxed/Tense	25	0.786	Good	20	0.665	Moderate
Quiet/Restless	2	0.7	Good	1	0.7	Moder

	4	62		9	20	ate
How fearful is this horse around other horses?	2 5	0.2 65	Poor	2 0	-0.0 32	Poor
Extroversion	*2* *5*	*0.7* *20*	*Moder* *ate*	*2* *1*	*0.7* *66*	*Good*
Sluggish/Forward-going	2 4	0.6 14	Moder ate	2 1	0.5 76	Moder ate
Placid/Active	*2* *5*	*0.5* *91*	*Moder* *ate*	*2* *0*	*0.7* *78*	*Good*
Adventurous/Habitual	2 5	0.5 41	Moder ate	2 1	0.4 35	Poor
Excitable/Laid-back	2 1	0.7 31	Moder ate	2 1	0.6 30	Moder ate
Spirited/Steady	2 4	0.7 18	Moder ate	2 1	0.6 38	Moder ate
How energetic would you say this horse is?	*2* *5*	*0.5* *18*	*Moder* *ate*	*2* *0*	*0.6* *53*	*Moder* *ate*
Gregariousness towards people	2 5	0.8 27	Good	2 1	0.7 08	Moder ate
Friendly/Standoffish	2 5	0.8 31	Good	2 0	0.6 04	Moder ate
How often does this horse initiate interaction with you	2 5	0.6 87	Moder ate	2 1	0.6 75	Moder ate
How often does this horse initiate interaction with other people	2 5	0.7 54	Good	2 1	0.6 28	Moder ate

Gregariousness towards horses	25	0.562	Moderate	21	0.391	Poor
How often does this horse initiate interaction with other horses?	25	0.511	Moderate	21	0.293	Poor
Does this horse ever show affection towards other horses?	25	0.494	Poor	21	0.147	Poor
How dependable would you say this horse is?	*24*	*0.571*	*Moderate*	*20*	*0.665*	*Moderate*

Table 4 - Results of test-retest reliability analyses for all subscales and individual items in the Equine Personality Test administered twice to n=25 horses by 2 familiar raters. A single-rating, absolute agreement, two-way mixed effects model was used to carry out intra-class correlations analyses. For each subscale or item the ICC estimate, 95% confidence interval, and interpretation in terms of test-retest reliability (ICC<0.5: poor; 0.5<ICC<0.75: moderate; 0.75<ICC<0.9: good; ICC>0.9: excellent) are presented. The subscript ? highlights subscales or items for which the ICC estimate indicates poor test-retest reliability.

			95% confidence interval		
	n	ICC	Lower	Upper	Reliability
Agreeableness	25	0.868	0.699	0.942	Good
Easy-going/Intolerant	2	0.82	0.639	0.917	Good

	5	1			
Argumentative/Well-mannered	25	0.717	0.283	0.884	Moderate
Obedient/Wayward	25	0.777	0.525	0.899	Good
Willing/Stubborn	25	0.700	0.430	0.855	Moderate
Gentle/Rough	25	0.759	0.526	0.886	Good
Neuroticism	25	0.903	0.792	0.956	Excellent
Anxious/Confident	25	0.915	0.816	0.962	Excellent
Nervous/Calm	25	0.846	0.682	0.929	Good
Relaxed/Tense	25	0.786	0.574	0.900	Good
Quiet/Restless	25	0.647	0.342	0.828	Moderate
How fearful is this horse around other horses?	25	0.352[?]	-0.047	0.653	Poor
Extroversion	25	0.910	0.808	0.959	Excellent
Sluggish/Forward-going	25	0.870	0.731	0.940	Good

Placid/Active	25	0.815	0.626	0.914	Good
Adventurous/Habitual	25	0.858	0.704	0.935	Good
Excitable/Laid-back	25	0.844	0.677	0.928	Good
Spirited/Steady	25	0.840	0.659	0.927	Good
How energetic would you say this horse is?	25	0.724	0.469	0.868	Moderate
Gregariousness towards people	25	0.922	0.818	0.966	Excellent
Friendly/Standoffish	25	0.890	0.766	0.950	Good
How often does this horse initiate interaction with you	25	0.777	0.549	0.896	Good
How often does this horse initiate interaction with other people	25	0.820	0.575	0.923	Good
Gregariousness towards horses	25	0.784	0.572	0.898	Good
How often does this horse initiate interaction with other horses?	25	0.592	0.217	0.806	Moderate
Does this horse ever show affection towards other horses?	25	0.703	0.440	0.856	Moderate
How dependable would you say this horse is?	2	0.72	0.469	0.866	Moder

5	2	ate

CRediT authorship contribution statement

Aurelie Jolivald, Carrie Ijichi, Kelly Yarnell, Carol Hall, Methodology: Aurelie Jolivald, Formal analysis: Aurelie Jolivald, Data curation: Aurelie Jolivald, Writing—original draft preparation: Aurelie Jolivald, Writing—review and editing: Carrie Ijichi, Kelly Yarnell, Carol Hall , Supervision: Kelly Yarnell, Carrie Ijichi, Carol Hall, Project administration: Aurelie Jolivald, Funding acquisition: Kelly Yarnell, Carol Hall All authors have read and agreed to the published version of the manuscript.

Declaration of Competing Interest

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Highlights

- We investigated a subjective equine personality questionnaire (EPT)
- The EPT should be assessed against robust criteria of validity and reliability
- The EPT met those criteria to similar standards as human personality questionnaires
- The EPT may be used to reliably assess equine personality in research or industry