Applied Animal Behaviour Science 219 (2019) 104836

Contents lists available at ScienceDirect

Applied Animal Behaviour Science

journal homepage: www.elsevier.com/locate/applanim

Validation of a temperament test in the Norwegian horse breeds

Hanne Fjerdingby Olsen*, Gunnar Klemetsdal

Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, NMBU, P.O. Box 5003, N-1432 Aas, Norway

ARTICLE INFO

Keywords:
Temperament
Temperament test
Factor analysis
Heart rate
Validation
Horse

ABSTRACT

The competitive edge of the National horse breeds in Norway is assumed to be strengthened by utilizing their temperament as a trait in the breeding goal. Previously, the temperament of the Norwegian horse breeds is described through five common factors, and there is a need for developing a test for temperament. In this study, 63 horses were tested at 6 different farms, through a proposed test procedure consisting of seven test moments composed of different obstacles/tasks. The performance of the horses was video recorded and individual heart rate was monitored. Through a developed ethogram, in total 43 traits were scored and related to the five factors from the former study and to the heart rate registrations. The proposed temperament test identified four of the five factors describing the temperament of the Norwegian horse breeds, except agreeableness. The seven test moments could be grouped into static objects, dynamic objects and precision tasks. The static and the dynamic, novel objects in the test triggered especially expressions of anxiousness and openness. Flight behaviour appeared as a suitable trait to record anxiousness, which was supported by increased heart rate, and openness seemed to be well recorded through explorative behaviour. Further, conscientiousness seemed to be best caught through the precision tasks when recording traits like focus on task and task performance. Dominance was not clearly related to any of the traits in the ethogram and there was a certain overlap with conscientiousness, which calls for more knowledge of the correlation between these two traits. The proposed test gave promising results, but the description of the traits should be improved, and a future test should vary test moments within category to avoid habituation.

1. Introduction

The Norwegian horse breeds, the Fjord, the Dole and the Nordland/Lyngen, are vulnerable breeds that have experienced a reduction of the genetic variance the past decades (Olsen et al., 2010). In addition, the number of foals born has decreased rapidly far below the recommended sustainable level of foals born per year (Olsen and Klemetsdal, 2010). The market situation is difficult, as the Norwegian breeds are almost outdistanced by imported breeds like the pony breeds, the Icelandic horse and the Warmblood riding horse, which are all bred for specific sports purposes. To strengthen the Norwegian breeds' competitive edge, their breeding organizations have concluded that the horses' overall temperament and their strong physique are traits for which these breeds have strengths relative to imported sports breeds. These traits make the breeds mentally and physically robust, well-suited for tough challenges in rough terrain, but also well-suited for horse-human relationship and activities such as health work (Norsk Hestesenter, 2012; Hem and Iversen, 2013). The breeding organizations also concluded that these strengths should be reflected when evolving future sports activities, which initiated the start-up of a common sport for the three breeds named Skeid. Skeid combines Norwegian cultural history and the rough, Norwegian nature, with the physical and mental benefits of these breeds and has similarities to the French equestrian sport named TRÉC (Techniques de Randonnée Équestre de Compétition). Skeid allows for different degrees of difficulty, from beginner to elite level, and the horses' competition performances have the potential to contribute with information of the mental and physical traits of interest, which also should be part of the breeding goal.

Since 1995, the three breeds have had separate breeding plans, approved by the Norwegian Equine Centre, and all three have included temperament as a trait in their breeding goal. Data recorded for traits in the breeding plan were digitalized in 2010 to found the basis of the work of Selle (2010) and revealed a lack of precisely described phenotyping and hence several of the traits were calculated with approximately no heritability. Despite this, temperament showed moderate heritabilities, and by far the highest when the veterinarian scored the horse for temperament in the health control. Obviously, a successful inclusion of temperament in the breeding goal relies on clear definitions and descriptions of the temperamental traits. A similar conclusion was reached in Germany, where it was found that evaluation of

* Corresponding author at: Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences (NMBU), P.O. Box 5003, N-1432 Ås, Norway.
E-mail address: hanne.fjerdingby@nmbu.no (H.F. Olsen).

https://doi.org/10.1016/j.applanim.2019.104836
Received 21 December 2018; Received in revised form 6 July 2019; Accepted 4 August 2019
Available online 07 August 2019

temperamental traits lacked objectivity with no clear definitions, and the need for changes in the evaluation system was precarious (König von Borstel et al., 2013).

There exists a discrepancy in the definition of temperament between human and non-human research. In human research, personality and temperament are considered closely related, whereas in non-human research the concern of being accused of anthropomorphism leads to differentiated definitions (Weinstein et al., 2008; Finkemeier et al., 2018). The definition of Clarke and Boinski (1995) has gained some joint acceptance; temperament refers to behavioural styles or tendencies that show continuity over time and that can be identified in early infancy, and which are reflected in the degree and responsivity to novel or stressful stimuli. Yet, it is gradually clear that the continuity also is relative, as the temperamental traits that are seen in early infancy most likely are moderated or strengthened by environmental influences (Weinstein et al., 2008). In horses, Le Scolan et al. (1997) rather points out the stability of the temperamental traits over situations than over time. It is also accepted that the temperament of non-human species is a multi-dimensional trait, following the same pattern across a wide range of species (Gosling and John, 1999).

Olsen and Klemetsdal (2017) described five temperamental factors common to the Norwegian horse breeds; 'anxiousness', 'agreeableness', 'conscientiousness', 'openness' and 'dominance'. The challenge now is how these underlying traits better can contribute to the inclusion of temperament in the breeding program, and to develop practically feasible assessment methods to reliably measure the temperamental traits. Strictly objective methods including physiological measurements, such as heart rate and salivary cortisol response, are reliable and predictable, but unfortunately often complicated and time-consuming to implement in practical testing of the animals. These objective measures are instead often used as validation criteria (e.g. Visser et al., 2002; König von Borstel et al., 2011), to develop a practical test utilizing trait scoring. Graf et al. (2014) suggested a temperament test consisting of five stimuli, aiming at scoring various temperamental traits and concluded that it is possible to implement temperament testing into performance tests.

In 2009, a test course with elements of different well-known temperament tests (e.g. bridge, novel object and unknown handler) was established of practitioners from the breeding organisation of the Fjord horse, and was accomplished for 3-year-old stallions attending the yearly breeding show, which is a part of the official breeding program, until 2015. This pioneering work made a good basis for further development of a test for the temperamental traits. A personality test has the purpose to correctly assess the personality of the horses. Not to state 'good' or 'bad' temperament, but rather to clearly present the advantages and drawbacks of an individual regarding its purpose of use (Lansade et al., 2016). Thus, the main goal of this study was to propose and validate a test of temperament in the Norwegian horse breeds, by calculating expressions for bias, repeatability and reliability, in relation to the five temperamental factors found for the Norwegian horse breeds in a former study.

2. Material and methods

2.1. Animals

The test included 63 horses of the breed Nordland/Lyngen stabled at 6 different farms (in total 26 owners). There were 17 stallions, 19 geldings and 27 mares of ages 1 to 25 years (on average 8.2 years). The horses were accustomed to head collar and girth. The youngest horses were at least accustomed to halter. The horses were familiar with the arena used for the test but naïve to the stimuli in the test. The horses had not been exercised on the test day. All horses were used mainly for hobby purposes, light working tasks or breeding. Stabling of the horses varied from outdoor housing in groups to indoor in boxes.

2.2. The test procedure

The test procedure was accomplished on the 6 different farms, which were spread out on a geographically large area in Troms County in northern Norway. Five test arenas were located outdoor on riding grounds, while one was in an indoor riding arena. The first test was carried out during four days in April 2013, and the repeated test was done four weeks later in May 2013. The repeated test included 26 of the 63 horses, as three of the farms did not manage to participate the second time. The weather conditions were quite even across tests, except less snow on the ground in the repeated test.

The horses wore a head collar and were equipped with a heart rate monitor (Polar ProTrainer 5, equine edition) recording the heart rate every five seconds. Due to measurement errors, the total number of animals with a valid pulse registration was 52.

The horses were led by an unfamiliar handler. The test consisted of seven test moments with different stimuli. The test moments were marked with a start and a stop line. All the horses were video recorded (with audio) through the test by a stationary camera mounted on a tripod. The tripod was manually handled such that the horse's midline was perpendicular to the lens as far as practically feasible. The time used in each test moment was recorded with a stopwatch. If the horse refused to enter the test moment within 1 min, showed extreme fear of the situation or the situation became potentially dangerous the handler was instructed to interrupt.

2.2.1. Test moment 1: log

A log free of branches with a diameter of 20–25 cm and length of 150–200 cm was placed across the moving direction. The horse was led towards the obstacle and was then let to decide how to pass the log.

2.2.2. Test moment 2: serpent slope

The slope was made up of four cones placed linearly in the moving direction, with 2-meter gaps between (Fig. 1). The horse was led in a zigzag pattern through the cones.

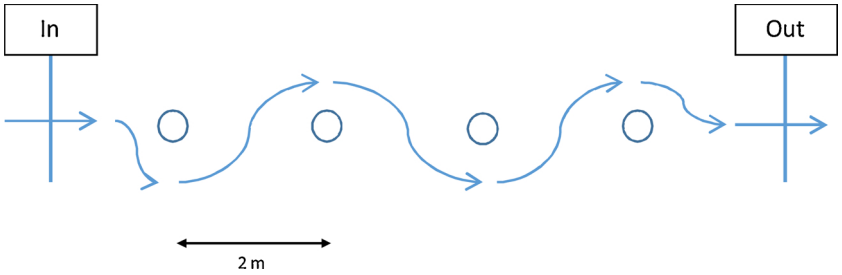

Fig. 1. Test moment 2 – Serpent slope.

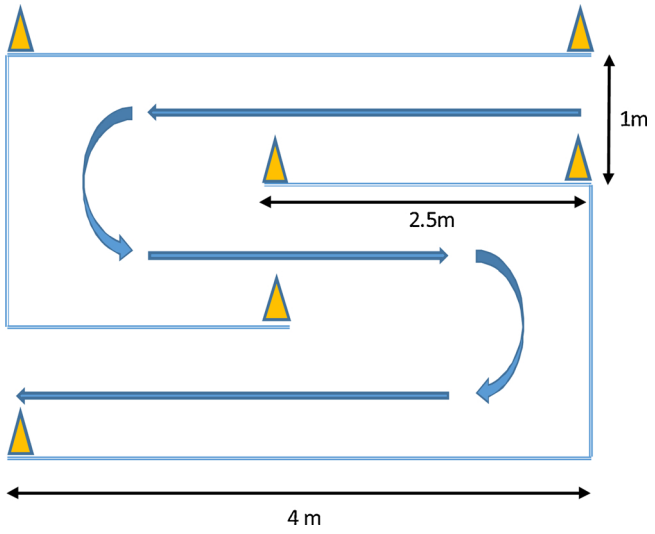

Fig. 2. Test moment 3 – Double-U, with orange cones indicating the borders of the slope.

2.2.3. Test moment 3: double-U

The slope was marked with six cones and formed a double-U of width 1 m and a total length of 4 m (Fig. 2). The horse was led through the slope.

2.2.4. Test moment 4: bridge

The test moment consisted of a "bridge", made up of a pallet covered with an even surface of plywood. The plywood was roughened with sawdust and gravel attached with glue. The horse was led towards the obstacle and was then let to decide how to enter or pass the bridge.

2.2.5. Test moment 5: visual obstacle

As the horse entered this test moment, a person placed 3 m from the moving direction lifted up a large flag and started to wave the flag in a figure of eight with an even movement until the horse had passed through.

2.2.6. Test moment 6: sound obstacle

As the horse entered the test moment, a person placed 3 m from the moving direction started thudding with a stick on an empty plastic can until the horse had passed through.

2.2.7. Test moment 7: tarpaulin

A tarpaulin of size 2 × 1.5 m was placed on the ground across the moving direction. The horse was led towards the unknown material on the ground as if to walk across but was then let to decide how to pass the material.

2.3. Behavioural observations

The horses' behaviour through the test moments were scored by analysing the video recordings of the test, using an ethogram consisting of nine traits (Fig. 3) with an associated scoring scheme (Fig. 4). All traits were scored on a 5-point Likert scale from 0 to 4, where '0' indicated no expression of a trait and '4' indicated full expression of the trait. In addition, the time spent on the task was recorded separately in test moments 1, 4, 5, 6 and 7 (Fig. 4). In scoring, the traits were colour coded to ease navigation in the ethogram.

The video recordings of the two tests were scored by the test leader. In addition, to investigate the inter-rater reliability, the initial test was scored by four observers asked to attend the study. The observers consisted of one professional educated horse judge and three persons with relevant professional skills combined with long experience with horses. The observers were trained in the scoring technique by the test leader. An instruction manual with video examples of how to use the scale for the different traits was made.

2.4. The horse personality questionnaire

In total, 32 of the horses in the present test had prior to this study been scored in a Horse Personality Questionnaire (HPQ) study (Olsen and Klemetsdal, 2017). To be able to use individual factor scores for temperamental traits interpreted in the HPQ study in the validation, the horse owners of the remaining 31 horses were requested to answer the HPQ, resulting in answers for additionally 21 horses.

2.5. Statistical analysis

The video recording of the initial test was analysed by using an ethogram (Fig. 4). Then, the standardised scorings of the 43 trait-moment variables in the ethogram were analysed using factor analysis. A principal component method was used with the largest absolute correlation between the variables as prior to determining the

Flight behaviour 0 = no expression of trait 1 = lower the speed somewhat, signs of hesitation 2 = as in 1, tries to increase the distance to the object with body position or small increase in speed, may show small startle reactions, may stop, leaning backwards 3 = as in 2, but clear startle reaction one or several times, move backwards or increase the speed considerably 4 = tries to run away	**Head posture (in walk)*** 0 = head bowed down to the ground, touches/almost touches the ground 1 = long neck, head hanging 2 = relaxed, natural head position 3 = somewhat raised head position 4 = very raised head position, overextended * = not when performing explorative behaviour	**Task performance, log** 0 = not accomplished, pass on the outside 1 = stumble the obstacle with uneven tempo as consequence 2 = touch the obstacle with one or more legs 3 = pass the obstacle without touching, but somewhat uneven tempo 4 = as in 3, but with even tempo and high precision
Threatening behaviour 0 = not threatening 1 = show small warnings; low snorting, tramping or bowing 2 = clear warnings; tramping, snorting, ears backwards 3 = as in 2, but also small attacks against object; tries to hit, kick or bite 4 = active attack towards object; rear, hit, kick or bite	**Focus on task (task=move forwards)** 0 = not interested in task, stops, focus other things, does not respond to the handler's signals after several reminders 1 = as in 0, but respond to the handler after few reminders 2 = not focused, lack of flow in movements, but respond to the handler's reminders 3 = move forwards quite focused, perhaps small reduction in the speed and less need for reminders from the handler 4 = focused and goal-oriented without reduction in speed or need for reminders from handler	**Task performance, serpent slope and double-U** 0 = bring down/move 3 or more cones 1 = bring down/move 1-2 cones 2 = touch at least one cone, otherwise ok precision 3 = accomplish without touching the cones, ok precision with some variation in speed 4 = accomplish perfect, with even tempo and high precision
Explorative behaviour 0 = no expression of trait 1 = visually focus on object without necessarily lower the speed 2 = as in 1, but lowers the speed, can stop, shows interest without necessarily physical contact with the object 3 = as in 2, but stop and is in physical contact with the object, investigates thorough (max 10 seconds) 4 = as in 3, but use very long time to investigate the object (>10 seconds)	**Pulling of rope** 0 = no pulling 1 = trace of pulling 2 = pull some in rope 3 = obvious pulls in the rope, can move sideways of the moving direction 4 = as in 3, but in obvious opposition to the handler with threatening elements	**Task performance, bridge/tarpaulin** 0 = not accomplished, pass on the outside 1 = pass partly over the obstacle, but interrupt the attempt 2 = pass over the obstacle, but use either long time, not in a straight line or "trips" over 3 = pass over the obstacle with somewhat uneven tempo 4 = as in 3, but with an even, controlled tempo

Fig. 3. Ethogram for scoring the various traits. The colour codes correspond with the colour codes in the scoring scheme in Fig. 4.

communalities. The number of factors was determined by keeping eigenvalues greater than 1. A varimax rotation was chosen in the interpretation of the factor loadings.

The new data, collected through the HPQ for individuals in the test, was included in the original HPQ data set for the Nordland/Lyngen horse, and then standardized with new means and standard deviations. A factor analysis of the scorings from the HPQ was accomplished as described in Olsen and Klemetsdal (2017), producing the same five factors as previous, producing individual factor scores for 53 of the 63 horses in the present study.

To calculate the relation between a horse's performance in the proposed test and its' temperament profile as obtained through the HPQ, Pearson correlation coefficients were calculated between test scores in the initial test and the individual factor scores from the factor analysis of the HPQ. Also, to explore whether there was a relation between the factors obtained from the HPQ and the factors from the proposed test, Pearson correlation coefficients were calculated between individual factor scores from the factor analysis of the test scores and the individual factor scores from the factor analysis of the HPQ.

The heart rate recordings from the initial test was summed per animal and moment, and normalised with an ln-transformation.

To examine the relationship between expressed heart rate and the individual test scores from the initial test, the normalised, summed heart rate of an individual *i* per moment *m* was regressed on the variables scored by the test leader for that moment, as expressed by Model 1:

$$\ln(\sum HR_{m_i}) = y_{m_i} = \beta_0 + \beta_1 \bullet X_1 + ... + \beta_j \bullet X_j + e_{m_i}$$

where β_0 is the intercept, β_1 is the regression coefficient associated with the first of j explanatory variables scored in that moment (X_1) (given in Fig. 4), and e_{m_i} is the random error term.

The relationship between the per moment (*m*) normalised, summed heart rate of an individual *i* and the factor scores for the eight factors (X_i) kept out of the 43 trait-moment variables, were examined by use of the following regression model (Model 2):

$$\ln(\sum HR_{m_i}) = y_{m_i} = \beta_0 + \beta_1 \bullet X_1 + ... + \beta_8 \bullet X_8 + e_{m_i}$$

where the remaining variables are as explained with Model 1.

In both models, multicollinearity among the predictor variables was evaluated by the variance inflation factor (criteria: < 5) as well as the condition number of the eigenvalues (criteria: < 30) (Montgomery et al., 2001), and all variables were kept.

To measure the consistency of individual measurement for the same trait, made in the initial and repeated tests, the Spearman rank correlation coefficients were calculated between individual test scores from video analysis of the initial test and the repeated test, utilizing data on the same 26 horses. In fact, the Spearman rank correlation coefficients are the Pearson correlation coefficient between the rank values in the test and in the repeated test (e.g. Bell et al., 2009).

To examine the degree of agreement among raters in the scoring of a trait within a moment, the inter-rater reliability of the scores done by the test leader and the four observers for the 63 horses in the initial test was estimated by the intraclass correlation (ICC) and its 95% confidence interval. Following Shrout and Fleis (1979), assuming that each of the n subjects was rated by the same k raters and the results address only these k raters, the ICC could be calculated as:

$$ICC = \frac{\sigma_h^2}{\sigma_h^2 + \sigma_e^2}$$

where σ_h^2 and σ_e^2 are variance components from the following model used to analyse a trait in a specific moment m:

$$y_{m_{ijk}} = \mu + rater_i + horse_j + e_{ijk}$$

where $rater_i$ denotes the random effect of the i-th rater, $\sim N(0, \sigma_r^2)$, $horse_j$ is the random effect of the j-th horse, $\sim N(0, \sigma_h^2)$, and e_{ijk} is the random residual, $\sim N(0, \sigma_e^2)$. Utilizing ANOVA, ICC can also be written:

$$ICC = \frac{MS_h - MS_e}{MS_h + (k-1)MS_e}$$

where MS_h is the mean square for horse and MS_e is the mean square for error.

The lower and upper confidence intervals (CI) for the ICC were calculated as follows:

$$CI_{lower} = \frac{F_{lower} - 1}{F_{lower} + (k-1)}, \text{ where } F_{lower} = \frac{F_h}{FINV(\frac{\alpha}{2}, df_h, df_e)}$$

$$CI_{upper} = \frac{F_{upper} - 1}{F_{upper} + (k-1)}, \text{ where } F_{upper} = F_h \bullet FINV(\frac{\alpha}{2}, df_e, df_h)$$

where F_h is the Fischer F-statistics and FINV is the inverse of the right-tailed F-probability distribution for a probability of 0.05.

The ICC and CI were calculated by using Excel, while the remaining calculations were performed using SAS/STAT® software.

3. Results

Fig. 5 visualizes the relation, through correlation coefficients ($P < 0.05$), between the individual trait scores from the video analysis of the initial temperament test and the individual factor scores from the analysis of the HPQ (Olsen and Klemetsdal, 2017). More detailed information can be found in Appendix A. Individual factor scores for the first factor, labelled anxiousness, relates to expressed flight behaviour in four out of the five possible test moments (1, 4, 6 and 7). In addition, anxiousness relates to raised head posture (test moments 5–7), reduced focus on task (test moments 1, 4 and 7), reduced task performance (test

	Test moment 1	Test moment 2	Test moment 3	Test moment 4	Test moment 5	Test moment 6	Test moment 7
Flight behaviour	0 1 2 3 4			0 1 2 3 4	0 1 2 3 4	0 1 2 3 4	0 1 2 3 4
Threatening behaviour					0 1 2 3 4	0 1 2 3 4	
Explorative behaviour	0 1 2 3 4			0 1 2 3 4	0 1 2 3 4	0 1 2 3 4	0 1 2 3 4
Task performance, log	0 1 2 3 4						
Task performance, serpent slope and double-U		0 1 2 3 4	0 1 2 3 4				
Task performance, bridge/tarpaulin				0 1 2 3 4			0 1 2 3 4
Head posture	0 1 2 3 4	0 1 2 3 4	0 1 2 3 4	0 1 2 3 4	0 1 2 3 4	0 1 2 3 4	0 1 2 3 4
Focus on task	0 1 2 3 4	0 1 2 3 4	0 1 2 3 4	0 1 2 3 4	0 1 2 3 4	0 1 2 3 4	0 1 2 3 4
Pulling of rope	0 1 2 3 4	0 1 2 3 4	0 1 2 3 4	0 1 2 3 4	0 1 2 3 4	0 1 2 3 4	0 1 2 3 4
Time spent on task	s			s	s	s	s

Fig. 4. Scheme for scoring the various traits per test moment, from zero to four and in seconds (s). The colour codes correspond with the colour codes in the ethogram in Fig. 3. A trait not scored within a test moment is indicated with a grey cell.

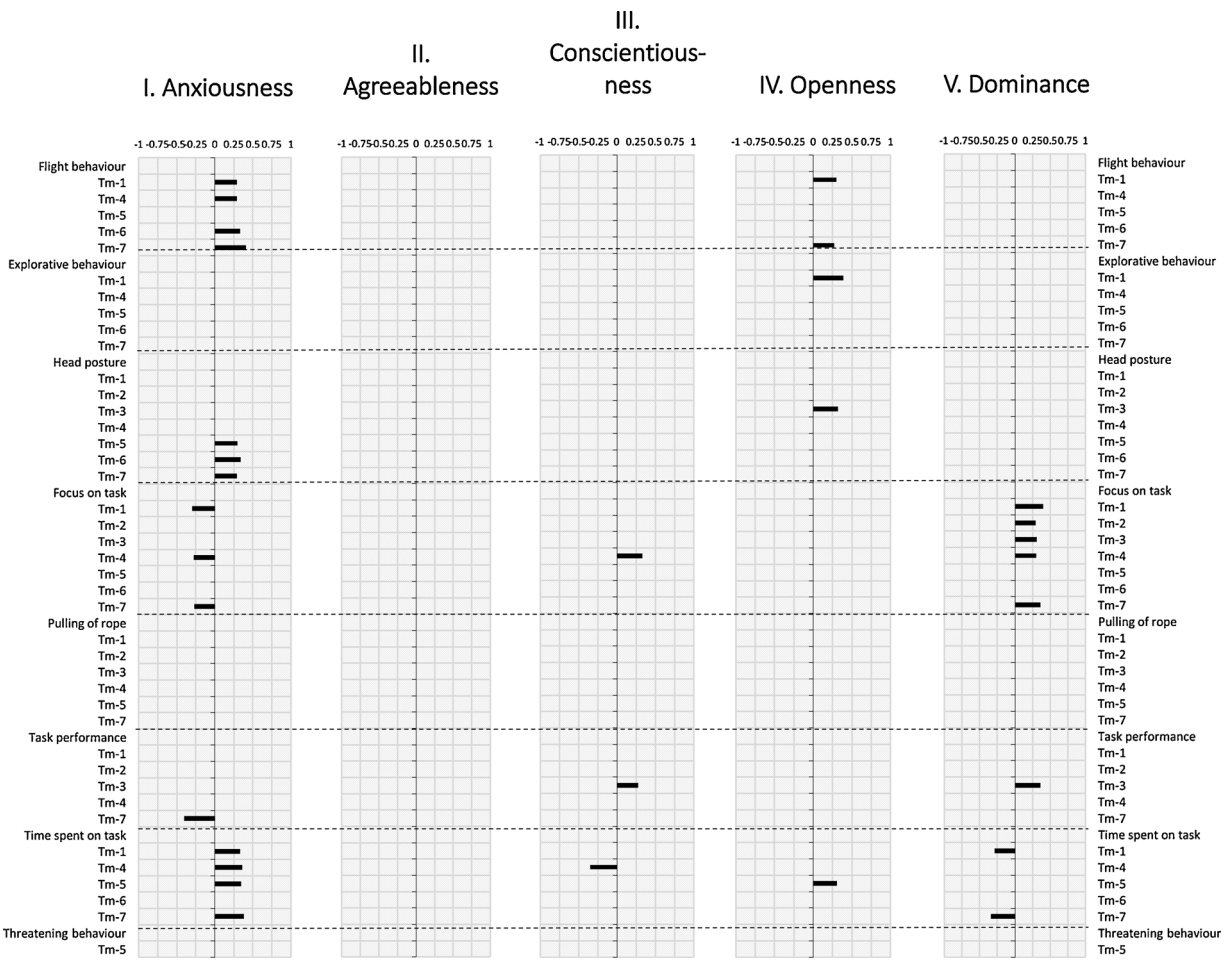

Fig. 5. Significant values ($P < 0.05$) of Pearson correlation coefficients between the individual test scores of the traits, where scored, for the test moments (Tm) 1–7 in the initial test and the individual factor scores of the five factors from the factor analysis of the Horse Personality Questionnaire (Olsen and Klemetsdal, 2017) (N = 53).

moment 7) and increased time spent on task (test moments 1, 4, 5 and 7). Actually, the tarpaulin (test moment 7) was represented with an expression of all the traits related to anxiousness. In addition, there was a similarity in the results for test moment 1 (the **log)**, test moment 4 (the bridge) and test moment 7 (the tarpaulin), where flight behaviour and time spent on task correlated positively with the factor anxiousness, and the correlation was consistently negative for focus on the task. Flight behaviour in test moments 1 and 7 also correlated with the fourth factor, openness, and traits from all three test moments correlated with the fifth factor, dominance. Special for test moment 4 was the positive correlation for focus on the task and the negative correlation for time spent on task relative to the factor conscientiousness. Conscientiousness also showed a relation to better task performance in test moment 3 (the double-U). In addition to flight behaviour, the factor openness correlated positive to explorative behaviour in test moment 1 **(log)**, head posture in test moment 3 (double-U) and time spent on task in test moment 5 (sound). Individual factor scores for the factor dominance from the HPQ had a positive correlation to focus on task in five of the seven test moments (except the sound and the flag). Dominance was also related to high task performance (test moment 3) and reduced time spent on task (test moments 1 and 7). Notice that the second factor labelled agreeableness did not significantly correlate to any of the traits in the temperament test.

Table 1 shows estimates of the regression coefficients from a multiple regression analysis of the ln-transformed sum of heart rates within each test moment on test scores within the same test moment. Flight behaviour was an important explanatory variable increasing heart rate in the test moments 5 and 7 ($P < 0.01$), but also in the test moments 1 and 6 ($P < 0.05$). Other traits that were strongly significant ($P < 0.01$) and enhancing the heart rate were (low) focus on task (test moments 3 and 4), (low) task performance (test moment 4) and explorative behaviour (test moment 7). A weaker relationship was found to pulling of rope (test moments 1 and 5) and time spent on task (test moment 5). The largest portions of the variance of the transformed sum of heart rate were explained in test moments 4 and 5, followed by moment 7, with R-squared values ranging from 0.68 to 0.57 (Table 1).

Fig. 6 shows the rotated factor loadings ($\geq |40|$) in the eight factors from the factor analysis of the standardized, individual trait scores. About one-fifth of the variance (12.75%) was due to the first factor (Appendix B). The seven remaining factors explained quite even shares of the variance, although somewhat declining with increasing factor number (from 8.57% to 4.89%, not shown). The first factor had mainly large loadings for focus on the task, (low) pulling of rope and task performance, of which all three traits appeared in both test moments 2 and 3. In addition, pulling of rope appeared in test moments 4 and 7, and focus on the task appeared in test moment 6. Somewhat smaller loadings were found for (low) head posture in test moments 2 and 4. For the second factor, flight behaviour had loadings above the threshold in test moments 1, 4, 6 and 7, corresponding well with similar findings for (low) focus on the task in test moment 1, and (low) task performance and time spent on task in test moment 7. The third factor loaded heavily on pulling of rope in test moments 5 and 7, time spent on task in test moments 1, 4 and 7 and (low) focus on the task in test moment 1. The fourth and fifth factors had only loadings for different traits

Table 1
Regression coefficient estimates (with level of significance in parenthesis) and the R-squared (R^2) statistics, as obtained when ln- transformed sum of heart rates per test moment ($ln\text{-}\Sigma HR_{moment}$) was regressed on test scores within the test moment in the initial test (N = 52). Numbers in bold for significant values (P < 0.05).

	Flight behaviour	Threatening behaviour	Explorative behaviour	Head posture	Focus on task	Pulling of rope	Task performance	Time spent on task	R^2
$ln\text{-}\Sigma HR_1$	**0.20 (0.012)**	- [1]	0.06 (0.138)	0.02 (0.861)	0.02 (0.817)	**0.43 (0.012)**	0.03 (0.416)	0.01 (0.339)	0.51
$ln\text{-}\Sigma HR_2$	- [1]	- [1]	- [1]	0.29 (0.063)	−0.12 (0.173)	−0.01 (0.977)	−0.004 (0.960)	- [1]	0.20
$ln\text{-}\Sigma HR_3$	- [1]	- [1]	- [1]	−0.13 (0.199)	**−0.26 (< 0.001)**	0.11 (0.447)	0.08 (0.064)	- [1]	0.36
$ln\text{-}\Sigma HR_4$	0.08 (0.285)	- [1]	−0.002 (0.955)	0.08 (0.371)	**−0.24 (0.003)**	−0.02 (0.872)	**−0.13 (< 0.001)**	−0.001 (0.750)	0.68
$ln\text{-}\Sigma HR_5$	**0.16 (0.004)**	0.16 (0.565)	−0.02 (0.721)	0.04 (0.567)	−0.07 (0.181)	**−0.60 (0.036)**	- [1]	**0.01 (0.029)**	0.60
$ln\text{-}\Sigma HR_6$	**0.20 (0.025)**	- [2]	0.15 (0.197)	0.07 (0.581)	0.05 (0.680)	- [2]	- [1]	0.004 (0.665)	0.34
$ln\text{-}\Sigma HR_7$	**0.58 (0.004)**	- [1]	**0.30 (< 0.001)**	−0.08 (0.549)	−0.02 (0.859)	0.11 (0.619)	0.001 (0.991)	0.0003 (0.962)	0.57

[1] Variable not included in test moment.
[2] No expression of the trait.

expressed in the test moments 5 and 6. For the fourth factor, loadings were highest in test moment 5, most expressed for (low) focus on the task, followed by flight behaviour, (high) head posture, (long) time spent on task and explorative behaviour. For factor five, loadings were consistently high for time spent on task and explorative behaviour in test moments 5 and 6. The sixth factor loaded heavily on traits in test moment 7, with (low) focus on the task, explorative behaviour and (long) time spent on task, and also to the traits flight behaviour and (low) task performance in test moment 4. The seventh factor loaded mainly on explorative behaviour and (low) focus on the task in test moment 4, but also task performance in test moment 1 and explorative behaviour in test moment 7. In the last factor, loadings were for head posture in the test moments 1, 2, 6 and 7.

Regarding the relation between the individual factor scores from the test scores and the individual factor scores from the Horse Personality Questionnaire (HPQ) (Olsen and Klemetsdal, 2017), there were two significant correlations (Table 2). A positive correlation between Factor 2 from the test scores and Factor 1 (anxiousness) from the HPQ, and a negative correlation between Factor 6 from the test scores and Factor 5 (dominance) from the HPQ.

Table 3 shows the estimates of the regression coefficients from a multiple regression analysis of the ln-transformed sum of heart rates within each test moment on individual factor scores for the eight factors resulting from factor analysis of the test scores. The largest number of significant (P < 0.05) regression coefficients were obtained for factors 1, 2 and 4. Further, the largest portion of the explained phenotypic variance of the ln-transformed sum of heart rate was in test moment 5, followed by test moments 4 and 7, with R-squared ranging from 0.55 to 0.49 (Table 3).

The consistency of trait measures, as scored in the initial and the repeated tests, are shown in Table 4. The highest Spearman rank correlations were found for time spent on task in test moment 6, followed by flight behaviour in test moments 1 and 6, ranging from 0.71 to 0.64. In fact, flight behaviour correlated significantly (P < 0.05) in four out of five test moments, followed by focus on the task, which was estimated with a significant correlation in five of the seven test moments. The other traits correlated less frequent and especially explorative behaviour, which was not found significant in any of the five test moments. In test moment 7 all traits but one had correlations above 0.5, while in test moment 5 none of the Spearman rank correlations were significant.

The intraclass correlation (ICC) shows the consistency between

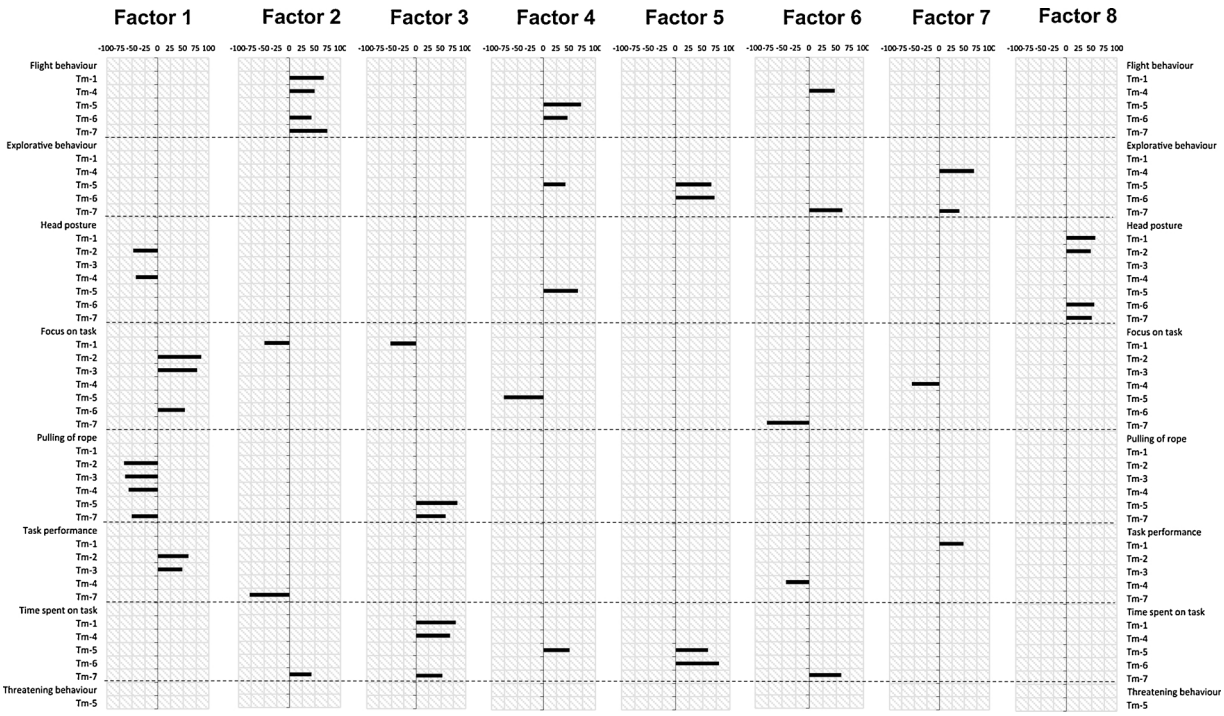

Fig. 6. Rotated factor loadings (x100) ≥|40| of the standardized individual trait scores, where scored, for the test moments (Tm) 1–7, in the initial test (N = 63). Threatening behaviour and pulling of rope in test moment 6 were omitted from the analysis, due to no expression of these traits.

Table 2
Pearson correlation coefficients (with level of significance in parenthesis) between individual factor scores for the eight first factors from the factor analysis of the test scores in the initial test, and individual factor scores of the five factors from the factor analysis of the Horse Personality Questionnaire (Olsen and Klemetsdal, 2017). Numbers in bold for significant values (P < 0.05) (N = 53).

	I. Anxiousness	II. Agreeableness	III. Conscientiousness	IV. Openness	V. Dominance
Factor 1	−0.06 (0.661)	0.08 (0.557)	0.12 (0.391)	−0.04 (0.761)	0.24 (0.082)
Factor 2	**0.34 (0.013)**	−0.001 (0.996)	−0.02 (0.897)	0.26 (0.060)	−0.19 (0.173)
Factor 3	0.20 (0.156)	−0.09 (0.535)	−0.21 (0.126)	−0.01 (0.937)	−0.11 (0.448)
Factor 4	0.23 (0.098)	−0.19 (0.168)	0.01 (0.919)	0.25 (0.072)	0.02 (0.863)
Factor 5	0.18 (0.193)	−0.04 (0.753)	0.13 (0.350)	0.21 (0.123)	−0.21 (0.128)
Factor 6	0.12 (0.386)	−0.002 (0.988)	−0.11 (0.416)	0.04 (0.803)	**−0.31 (0.022)**
Factor 7	0.03 (0.846)	−0.02 (0.866)	−0.19 (0.183)	0.17 (0.229)	−0.15 (0.295)
Factor 8	0.20 (0.159)	−0.001 (0.997)	−0.10 (0.468)	0.05 (0.718)	0.06 (0.659)

raters and is given with confidence intervals in Table 5. When using the guidelines of Cicchetti (1994) for interpretation, where an ICC value between 0.4 - 0.59 is considered fair, ICC between 0.6 - 0.74 is good and ICC between 0.75–1.0 is excellent, a total of 27 of the 38 traits scored were estimated with an ICC from fair to excellent. Four of the traits obtained the favoured ICC values in all test moments, with the most expressed values for task performance, followed by flight behaviour, explorative behaviour and focus on the task. For head posture, ICC was fair only in test moments 2, 5 and 6, while pulling of rope and threatening behaviour was consistently low, with an exception for the former being fair in test moments 4 and 7. In test moment 7, all traits but one (head posture) was above 0.5.

4. Discussion

The Nordland/Lyngen pony constitutes a small population of approximately 3000 animals, with few animals per unit and the units spread over a geographically large area. Therefore, it is difficult to gather many horses for scientific purposes, which in this study was approached by testing horses in six stables with altogether 63 horses. The test animals were assumed to be representing a random sample of future performance tested horses, even though such a small sample most likely does not stretch out the environmental and phenotypic variance for the population, causing less statistical insight. A larger spread on age classes will have the same effect, unless the expression of a temperamental trait is stable over time, as suggested by e.g. Lansade et al. (2008). Although both objections will reduce the power of the experiment, it is still assumed that the results found also will be valid for future performance tested horses. The definition of the linear scale is somewhat affected by the actual expression of the traits in the test sample. Due to our limited number of animals, the defined scale should be adjusted as more samples are observed over time. Also, a higher resolution to the scale would probably be advantageous to classify expressed behaviour with higher precision, although distinguishing between the classes in the scale description and when scoring could become a challenge.

An ethogram requires precise trait definitions, allowing for different coping strategies, as these might be different expressions of the same response, as suggested by Ijichi et al. (2013). With the main objective to reveal the relationship between the proposed tests and the five different temperamental factors identified through a Horse Personality Questionnaire (HPQ) in Olsen and Klemetsdal (2017), the behavioural traits that were chosen for the ethogram were *a priori* assumed to be relevant in covering the expression of the factors. Still, some of the traits were somewhat difficult to identify and to score in the most suitable category. During the video analysis, it was especially hard to separate explorative behaviour from behaviour related to anxiousness. For example, if a horse suddenly stopped ahead of an obstacle, it could both be an expression of a startle reaction (flight behaviour) or an expression for inquisitive interest for the novel object (explorative behaviour), seeking information on distance. This mixed expression was reflected in the results, as for instance in Fig. 6 where it appears a certain co-loading between anxiousness and openness (factor 6), and in Table 2 where the factors 2 and 4 both relate (although not significantly) to anxiousness and to openness (see more detailed explanation below).

The test moments consisted of different tasks or obstacles for the horse to cope with, comparable to everyday challenges in ordinary use. An important question was whether the test moment scores could reveal individual information of the five temperamental factors from the HPQ (Olsen and Klemetsdal, 2017). As there was no distinct pattern within test moment regarding correlation to the individual factor scores from the HPQ (Fig. 5), it was rejected that the test moments tested separate traits. Rather, there seemed to be a common pattern between the type of obstacle, roughly divided into technical challenges (serpent slope and the double-U) and dynamic (moving, varying) or static, novel objects, and expressions of the temperamental traits. The dynamic, novel objects were represented by the flag and the sound obstacles, and the static, novel objects were represented by the log, the bridge and the tarpaulin. The three last mentioned items could also be regarded as variants of obstacles on the ground, which is most important for a flight animal to be aware of.

Anxiousness, being the first factor from the HPQ in Olsen and Klemetsdal (2017), correlated significantly to traits in all test moments with either visual or auditive novel objects (except test moments 2 and 3 involving technical challenge with cones). All traits associated with high individual factor score for anxiousness (Fig. 5) are easily

Table 3
Regression coefficient estimates (with level of significance in parenthesis) and the R-squared statistics (R^2), obtained from regressing ln- transformed sum of heart rates per test moment (ln-ΣHR_m, where m is moment 1 to 7) on individual factor scores, for the eight factors resulting from the factor analysis of the individual test scores in the initial test (N = 52). Numbers in bold for significant values (P < 0.05).

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8	R^2
ln-ΣHR_1	−0.08 (0.102)	**0.21 (< 0.001)**	0.07 (0.055)	0.08 (0.110)	0.04 (0.333)	−0.002 (0.958)	0.07 (0.089)	0.07 (0.117)	0.44
ln-ΣHR_2	**−0.20 (0.001)**	**0.13 (0.016)**	0.03 (0.486)	**0.12 (0.043)**	−0.0002 (0.997)	0.01 (0.822)	−0.004 (0.941)	0.05 (0.296)	0.24
ln-ΣHR_3	**−0.14 (0.003)**	0.05 (0.195)	0.06 (0.090)	**0.15 (0.001)**	−0.02 (0.489)	−0.02 (0.683)	−0.01 (0.758)	0.02 (0.650)	0.27
ln-ΣHR_4	**−0.17 (0.003)**	**0.21 (< 0.001)**	0.04 (0.394)	0.06 (0.267)	−0.05 (0.231)	**0.19 (< 0.001)**	**0.10 (0.044)**	**0.12 (0.017)**	0.53
ln-ΣHR_5	**−0.10 (0.029)**	**0.13 (0.002)**	−0.04 (0.259)	**0.27 (< 0.001)**	**0.11 (0.005)**	−0.001 (0.982)	0.03 (0.487)	−0.002 (0.965)	0.55
ln-ΣHR_6	**−0.16 (0.007)**	**0.16 (0.004)**	0.003 (0.955)	**0.19 (0.002)**	**0.13 (0.007)**	−0.02 (0.765)	−0.0004 (0.994)	0.08 (0.124)	0.41
ln-ΣHR_7	**−0.22 (0.014)**	**0.28 (< 0.001)**	−0.06 (0.407)	0.12 (0.174)	0.08 (0.241)	**0.39 (< 0.001)**	0.11 (0.175)	0.08 (0.287)	0.49

Table 4
Spearman rank order correlation coefficients (with level of significance in parenthesis) between individual test scores from the initial and the repeated tests (N = 26). Numbers in bold for significant values (P < 0.05).

	Test moment 1	Test moment 2	Test moment 3	Test moment 4	Test moment 5	Test moment 6	Test moment 7
Flight behaviour	**0.65 (< 0.001)**	- [1)]	- [1)]	**0.40 (0.044)**	0.27 (0.176)	**0.64 (< 0.001)**	**0.51 (0.008)**
Threatening behaviour	- [1)]	- [1)]	- [1)]	- [1)]	−0.10 (0.612)	- [2)]	- [1)]
Explorative behaviour	0.34 (0.085)	- [1)]	- [1)]	0.14 (0.506)	−0.09 (0.653)	0.30 (0.132)	0.20 (0.325)
Head posture	**0.43 (0.029)**	0.39 (0.051)	0.27 (0.183)	0.37 (0.063)	0.15 (0.467)	0.10 (0.624)	**0.57 (0.003)**
Focus on task	**0.53 (0.006)**	0.33 (0.104)	**0.39 (0.049)**	**0.41 (0.036)**	0.27 (0.185)	**0.60 (0.001)**	**0.55 (0.004)**
Pulling of rope	−0.06 (0.779)	−0.04 (0.846)	**0.40 (0.045)**	0.20 (0.337)	- [2)]	- [2)]	**0.51 (0.008)**
Task performance	0.23 (0.265)	**0.47 (0.015)**	0.16 (0.442)	0.01 (0.944)	- [1)]	- [1)]	**0.61 (0.001)**
Time spent on task	0.30 (0.134)	- [1)]	- [1)]	**0.58 (0.002)**	0.13 (0.524)	**0.71 (< 0.001)**	**0.61 (0.001)**

[1)] Variable not included in test moment.
[2)] No expression of trait in one or both of the tests.

Table 5
The intraclass correlation (ICC) (with 95% confidence interval (C.I.) in brackets) for test scores in the initial test (N = 63), scored by five different judges. Numbers in bold for ICC values above 0.4, considered as fair to excellent (Cicchetti, 1994).

	Test moment 1		Test moment 2		Test moment 3		Test moment 4		Test moment 5		Test moment 6		Test moment 7	
	ICC	C.I.	ICC	C.I.	ICC	C.I.	ICC	C.I.	ICC	C.I.	ICC	C.I.	ICC	C.I.
Flight behaviour	**0.71**	[0.61, 0.79]	- [1)]	- [1)]	- [1)]	- [1)]	**0.71**	[0.62, 0.80]	**0.82**	[0.76, 0.88]	**0.77**	[0.69, 0.84]	**0.68**	[0.58, 0.77]
Threatening behaviour	- [1)]	- [1)]	- [1)]	- [1)]	- [1)]	- [1)]	- [1)]	- [1)]	0.08	[-0.01, 0.19]	0	[-0.07, 0.09]	- [1)]	- [1)]
Explorative behavior	**0.86**	[0.80, 0.90]	- [1)]	- [1)]	- [1)]	- [1)]	**0.86**	[0.81, 0.91]	**0.59**	[0.48, 0.70]	**0.68**	[0.58, 0.77]	**0.90**	[0.86, 0.93]
Head posture	0.30	[0.19, 0.43]	**0.54**	[0.42, 0.65]	0.31	[0.20, 0.45]	0.28	[0.17, 0.41]	**0.49**	[0.37, 0.61]	**0.50**	[0.38, 0.62]	0.23	[0.13, 0.36]
Focus on task	**0.75**	[0.67, 0.82]	**0.71**	[0.62, 0.80]	**0.60**	[0.49, 0.70]	**0.71**	[0.62, 0.79]	**0.62**	[0.51, 0.72]	**0.54**	[0.43, 0.66]	**0.73**	[0.64, 0.81]
Pulling of rope	0.27	[0.16, 0.40]	0.30	[0.19, 0.44]	0.20	[0.10, 0.32]	**0.40**	[0.28, 0.52]	0.11	[0.02, 0.23]	0.07	[-0.01, 0.18]	**0.51**	[0.39, 0.63]
Task performance	**0.90**	[0.86, 0.93]	**0.70**	[0.61, 0.79]	**0.78**	[0.71, 0.85]	**0.93**	[0.90, 0.96]	- [1)]	- [1)]	- [1)]	- [1)]	**0.90**	[0.87, 0.94]

[1)] Variable not included in test moment.

associated with expressions for anxiousness, such as flight behaviour, spending long time on a task, low concentration level and low performance. Further, several of these traits were also highly significant explanatory variables to an increase in heart rate, as shown in Table 1, where flight behaviour was significant in four of the five test moments involving a novel object. Exposing horses to novel stimuli is known to increase the heart rate (e.g. Christensen et al., 2005; McCall et al., 2006) and has been found as a good method for measuring fearfulness in horses (Wolff et al., 1997; McCall et al., 2006; Marsbøll and Christensen, 2015). The traits mentioned above, that was associated with anxiousness, were also found in the factors 2 and 4 from the factor analysis of the individual trait scores, as shown in Fig. 6. Factor 2 mainly loaded on traits in the test moments with static, novel objects (the log, the bridge and the tarpaulin), with flight behaviour present in all. Factor 2 also showed a positive correlation to the factor anxiousness from the HPQ, as shown in Table 2. Flight behaviour from the test moment with the sound (dynamic object) also mapped on factor 2, but the design of the sound obstacle, with a large plastic can being visible to the horse, might cause a certain covariance with the static objects. Factor 4 consisted of traits associated with anxiousness only from the test moments with the dynamic obstacles (flag and sound). Also, the factors 2 and 4 increased the heart rate strongest in main through the static or dynamic novel objects (Table 3). Overall, this suggests a grouping into anxiousness for static, novel objects and anxiousness for dynamic, novel objects, where anxiousness being expressed differently when an individual is exposed to static or dynamic novel objects. In addition, factor 6 loaded on traits associated to anxiousness in test moment 4, but this factor also loaded on traits associated to openness in test moment 7, such as explorative behaviour, and could be an example of the confusion between the expressions of flight behaviour and explorative behaviour as mentioned above. The results indicate that flight behaviour appears as a good indicator for anxiousness and that a variety of both dynamic and static objects should be included in the further development of a temperament test.

Horses scoring high on conscientiousness, which is a factor associated with work and use of the horses in the HPQ (Olsen and Klemetsdal, 2017), showed high focus on task and solved the task rapidly when crossing the bridge (test moment 4), and performed the task well in the technical challenge through the double-U (test moment 3) (Fig. 5). Conscientiousness could be recognized in factor 1 from the factor analysis of the individual trait scores (Fig. 6), primarily through traits such as high focus on task, low pulling of rope and high task performance, mainly from the test moments with the technical challenges (test moments 2 and 3). High focus on task significantly reduced heart rate in the test moments 3 (double-U) and 4 (bridge), as also did high task performance in test moment 4 (Table 1). In addition, the factor interpreted as conscientiousness from Fig. 6 (factor 1) significantly reduced heart rate in all test moments, except the first. These results indicate that individuals scoring high on conscientiousness performed better and had a generally lower stress level through the test. Several studies have reported a negative correlation between fearfulness and performance (e.g. Fiske and Potter, 1979; Lindberg et al., 1999), and there is suggested that nervous horses are more easily distracted, leading to lower performance (Mendl, 1999). In addition, Valenchon et al. (2013) showed that non-fearful horses, measured in heart rate and heart rate variability, performed better than fearful horses under stressful conditions. This supports our findings of a connection between less nervous individuals expressed through lower heart rate levels, and high scorings for focus on task and task performance. Thus, it seems likely that the trait conscientiousness could be recorded by using technical challenges and could be evaluated through traits like focus on task and task performance.

The factor openness from Olsen and Klemetsdal (2017), constituting traits like curiosity, sociability and playfulness, had the highest and most significant correlation to explorative behaviour in test moment 1 (log) (Fig. 5) and correlated close to significant with explorative behaviour in test moment 4 (bridge), but did not correlate significantly with explorative behaviour in any of the other test moments (see Appendix A). On the other hand, the factor openness showed a significant correlation of around 0.3 to flight behaviour expressed in the test

moments with the log and the tarpaulin (test moments 1 and 7). Further, the individual trait score for explorative behaviour did not affect heart rate significantly in the test moment with the log, but instead, it affected increased heart rate highly significant in test moment 7 (tarpaulin) (Table 1). From the factor analysis of the individual trait scores (Fig. 6), the factors 3, 5 and 7 could be interpreted in the direction of openness due to the inclusion of traits like explorative behaviour, long time spent on task and low focus on task. Factor 5 covers the test moments with the dynamic, novel objects (flag and sound) with the traits explorative behaviour and time spent on task, while factor 3 and factor 7 covers the test moments with the static, novel objects (log, bridge and tarpaulin), where time spent on task in general loaded on factor 3, and explorative behaviour in general loaded on factor 7. However, factor 3 also loaded on pulling of rope in the test moments 5 and 7. Pulling of rope was included in the ethogram with the purpose to record dominance (towards the handler), but in addition to this, the trait was observed also in other situations during the test. A horse could pull the rope when being uncomfortable with the situation, or it could pull the rope when being eager to explore. In total, factors 3 and 7 might be interpreted as different expressions of openness, separating out different traits. Neither of the factors 3, 5 or 7 correlated significantly with the factor openness or any other factors from Olsen and Klemetsdal (2017) (Table 2), but factor 5 increased the heart rate significantly in the test moments with the dynamic, novel objects (flag and sound), and factor 7 increased the heart rate significantly in test moment 4 (bridge) (Table 3). Factor 6 from the factor analysis of the individual trait scores (Fig. 6), which was mentioned to be associated with both anxiousness and openness, was also a highly explanatory variable for increased heart rate in the test moments with the bridge and the tarpaulin (static objects). But again, it was through this study hard to separate whether this relation is caused by anxiousness or openness (exploration). Exploration can trigger exaggerative behaviour, such as extreme sniffing, stumping and similar, leading to more motoric activity, which in turn can increase the heart rate. Nevertheless, the connection between openness and flight behaviour might show the difficulty in separation of flight behaviour from explorative behaviour in scoring, as mentioned above, but also might illustrate that all traits are expressions of a mixture of underlying factors (Finkemeier et al., 2018). This was also seen for time spent on task in test moment 5, which correlated both to anxiousness and openness (Fig. 5), while flight behaviour and explorative behaviour did not. However, more time spent on a task might yet be another expression of exploration, indicating that the same pattern, or response, might be recorded by different traits across test moments, i.e. an example of different coping strategies across traits, in analogy with Ijichi et al. (2013). Overall, the test animals showed different levels of explorative behaviour towards dynamic and static objects, suggesting a similar structure as with anxiousness.

Dominance was the last of the five factors derived from the HPQ (Olsen and Klemetsdal, 2017), and correlated strongly positive to the trait focus on task (Fig. 5). Further, dominance showed a positive correlation to the traits high task performance and less time spent on task (Fig. 5). These three traits are also in general associated with conscientiousness, and task performance from the double-U and focus on task from the bridge also correlated positively to conscientiousness. Nevertheless, only focus on task in the test moments 3 (double-U) and 4 (bridge) significantly reduced the heart rate (Table 1), as mentioned before for conscientiousness. To follow this, the factor dominance from the HPQ was positive, although not significant, correlated to factor 1 from the test scores interpreted as conscientiousness and correlated significantly negative to factor 6, which was associated both with openness and anxiousness (Table 2). This means that horses scoring high for dominance in the HPQ also scored high on conscientiousness and low on the split factor explaining both openness and anxiousness. Dominance, as interpreted from the HPQ, consisted of adjectives like dominant to horses and non-subordinate (Olsen and Klemetsdal, 2017). Also, the review of Gosling and John (1999) states that dominance, across studies, was compounded by assertiveness, physical aggression and low fearfulness. It has been shown that baseline stress hormones can be related to rank in horses (Christensen et al., 2012) and that horses with low stress reactions perform better in cognitive tests (Mengoli et al., 2014), making it likely that dominant horses perform better, and which can explain the close relationship to conscientiousness. At the same time, it reveals a challenge to separate conscientiousness and dominance through the suggested test, and that selection for willingness to work potentially can include undesirable levels of dominance. From the factor analysis of the individual trait scores (Fig. 6), none of the factors could easily be interpreted in the direction of dominance. Factor 8, that not yet has been explained by the other traits, consists solely of the trait high head posture in the test moments 1, 2, 6 and 7, and could thus be interpreted in terms of self-confidence and dominance. But high head posture could also likely be an expression of arousal or alertness, as described for the factor anxiousness. Besides, factor 8 is a significant explanatory variable to increased heart rate in test moment 4, the bridge (Table 3), which do not strengthen the connection of factor 8 to dominance, but rather in direction of a general expression of anxiousness.

The proposed test did not at all correlate with the second factor from the former HPQ, agreeableness, and none of the other results could be explained in terms of this trait. Agreeableness is considered important for the horse-human relationship, and thus for making an objective recommendation to prioritize areas of use for individual horses, such as to the health segment. Thus, there is a need to further explore how to record agreeableness, which needs to be prioritized in future research.

The test for repeatability (Table 4) showed that flight behaviour (test moments 1, 6 and 7) and time spent on task (test moments 6 and 7) represented the traits with the highest consistencies over time, ranging from 0.51–0.65 and 0.61–0.71, respectively. This means that horses that expressed flight behaviour and spent a long time on the tasks the first time also, to some extent, expressed the same traits the second time four weeks later. Notice also that focus on task was estimated with a significant rank correlation in five out of seven traits. All other traits, except threatening behaviour and explorative behaviour, showed a rank correlation larger than 0.5 in one or more test moments. A lack of repeatability for explorative behaviour could be caused by habituation, as seen in König von Borstel et al. (2012) and Leiner and Fendt (2011). To prevent habituation to test situations, König von Borstel et al. (2012) suggested a temperament index, consisting of repeated measures and both their absolute scores and the improvement together with behavioural data collected during training. Due to the risk of habituation in this test, the novel objects within a category should vary over time if applied to a test allowing for multiple testing of individuals. Further, the rank stability of a test moment across traits is important to evaluate, being especially high in test moment 7.

Lansade et al. (2016) identified a personality test for horses, suitable for use in the field, which were stable across both situation and time, given sufficient training of the test personnel. Imprecise descriptions in an ethogram allow for subjectivity in scoring and can cause low consistency between observers (Pierard et al., 2015). When evolving the ethogram, a large effort was put into making the traits objective to record, and the observers were thoroughly trained in how to use the scale to minimize the variation between observers. This was supported by the intraclass correlation (ICC) between judges, measuring this consistency, which was at least fair for 27 out of 38 traits (Table 5). In addition, flight behaviour, explorative behaviour, focus on task and task performance, showed significantly high ICC throughout all the test moments (Table 5). Still, the ethogram needs to be evolved to better be able to distinguish the traits, like for instance anxiousness and openness or conscientiousness and dominance.

5. Conclusion

The proposed test moments with its suggested traits were related to four of the five temperamental factors described in a previous study of a Horse Personality Questionnaire; anxiousness, conscientiousness, openness and dominance. Agreeableness was not identified through the test results. The seven test moments could be grouped into static objects, dynamic objects and precision tasks. The static and the dynamic, novel objects in the test triggered especially expressions of anxiousness and openness, and the expression of a trait for a static obstacle did not necessary trigger the same expression for a dynamic one. Flight behaviour appeared as a suitable trait to record anxiousness, which was supported by increased heart rate, and openness seemed to be well recorded through explorative behaviour. Nevertheless, there were signs of confusion between flight behaviour and explorative behaviour when scoring, caused by an overlap in traits, signalizing that the description of these traits in the ethogram must be evaluated. Further, conscientiousness seemed to be best caught through precision tasks when recording traits like focus on task and task performance. Whether the test could reveal dominance in a good way was somewhat unclear, as this trait was not clearly related to any of the traits in the ethogram and there was a certain overlap with conscientiousness, which calls for more knowledge of the correlation between these two traits. In future development and use of such a test for temperament, it is advised to use both static and dynamic novel objects and include technical challenges for the horse. To avoid habituation, it is also advised to develop several test moments representing the different categories rather than literary use the test moments suggested here. In addition, the description of the recorded traits should be improved according to the findings in this study.

Declaration of Competing Interest

None.

Acknowledgements

Thanks to the Norwegian Research Council for financial support through project number 190166, to Birgit D. Nielsen and Stine Samsonstuen for practical assistance, to the expert panel and to all the horse owners placing their horses at disposal for the project.

Appendix A. Pearson correlation coefficients (with level of significance in the parenthesis) between the test scores in the initial test and the individual factor scores of the five factors from factor analysis of the Horse Personality Questionnaire (Olsen and Klemetsdal, 2017) (N = 53). Numbers in bold are for significant values (P < 0.05).

	I.	II. Agreeableness	III. Conscientious-ness	IV.	V.
Test moment 1: Log					
Flight behaviour	**0.29 (0.032)**	0.03 (0.847)	−0.13 (0.366)	**0.30 (0.031)**	−0.22 (0.106)
Explorative behaviour	0.16 (0.265)	−0.04 (0.753)	−0.06 (0.661)	**0.39 (0.004)**	−0.22 (0.107)
Head posture	0.14 (0.331)	0.02 (0.892)	−0.11 (0.413)	−0.07 (0.643)	−0.14 (0.305)
Focus on task	**−0.30 (0.028)**	0.09 (0.500)	0.14 (0.330)	−0.21 (0.140)	**0.40 (0.003)**
Pulling of rope	0.15 (0.296)	0.09 (0.500)	0.03 (0.840)	0.04 (0.780)	−0.17 (0.212)
Task performance	0.01 (0.926)	−0.05 (0.725)	−0.08 (0.560)	0.20 (0.148)	−0.03 (0.812)
Time spent on task	**0.33 (0.016)**	−0.13 (0.343)	−0.22 (0.116)	0.15 (0.273)	**−0.29 (0.037)**
Test moment 2: Serpent					
Head posture	0.07 (0.617)	−0.21 (0.123)	−0.06 (0.688)	0.08 (0.558)	0.02 (0.890)
Focus on task	−0.11 (0.413)	0.11 (0.417)	0.12 (0.399)	0.06 (0.674)	**0.29 (0.034)**
Pulling of rope	0.13 (0.353)	0.04 (0.752)	−0.06 (0.661)	0.19 (0.168)	−0.20 (0.157)
Task performance	−0.18 (0.186)	0.09 (0.515)	0.20 (0.151)	0.02 (0.868)	0.24 (0.084)
Test moment 3: Double-U					
Head posture	0.13 (0.365)	−0.09 (0.543)	−0.17 (0.219)	**0.32 (0.021)**	−0.08 (0.579)
Focus on task	−0.16 (0.259)	0.22 (0.120)	0.14 (0.322)	−0.09 (0.526)	**0.31 (0.022)**
Pulling of rope	0.14 (0.328)	−0.12 (0.387)	−0.15 (0.273)	0.20 (0.161)	−0.15 (0.289)
Task performance	−0.10 (0.469)	−0.10 (0.492)	**0.28 (0.040)**	0.06 (0.675)	**0.36 (0.009)**
Test moment 4: Bridge					
Flight behaviour	**0.29 (0.034)**	0.02 (0.893)	−0.10 (0.474)	0.11 (0.433)	−0.24 (0.087)
Explorative behaviour	−0.00 (0.991)	−0.08 (0.564)	−0.16 (0.259)	0.27 (0.054)	0.09 (0.516)
Head posture	0.26 (0.060)	−0.20 (0.149)	−0.16 (0.258)	0.18 (0.208)	−0.07 (0.617)
Focus on task	**−0.28 (0.046)**	0.14 (0.318)	**0.33 (0.015)**	−0.20 (0.141)	**0.30 (0.029)**
Pulling of rope	0.06 (0.650)	0.07 (0.635)	−0.15 (0.294)	0.16 (0.252)	−0.22 (0.112)
Task performance	−0.02 (0.859)	−0.01 (0.945)	−0.03 (0.846)	−0.06 (0.651)	0.09 (0.537)
Time spent on task	**0.36 (0.008)**	−0.15 (0.282)	**−0.35 (0.011)**	0.17 (0.234)	−0.17 (0.233)
Test moment 5: Visual					
Flight behaviour	0.12 (0.399)	−0.18 (0.209)	−0.00 (0.998)	0.16 (0.255)	0.00 (0.999)
Threatening behaviour	0.10 (0.458)	0.16 (0.243)	0.18 (0.192)	0.15 (0.281)	0.00 (0.978)
Explorative behaviour	0.19 (0.170)	−0.07 (0.606)	0.06 (0.652)	0.21 (0.128)	−0.11 (0.427)
Head posture	**0.30 (0.030)**	−0.16 (0.253)	0.10 (0.456)	0.23 (0.101)	−0.15 (0.286)
Focus on task	−0.21 (0.130)	0.20 (0.161)	−0.02 (0.859)	−0.16 (0.238)	0.08 (0.551)
Pulling of rope	0.16 (0.250)	−0.15 (0.288)	−0.20 (0.150)	0.01 (0.960)	−0.01 (0.946)
Time spent on task	**0.35 (0.009)**	−0.16 (0.257)	0.01 (0.935)	**0.31 (0.027)**	−0.21 (0.136)
Test moment 6: Sound					
Flight behaviour	**0.33 (0.015)**	−0.04 (0.770)	−0.16 (0.239)	0.24 (0.086)	−0.20 (0.147)
Threatening behaviour	- [1)]	- [1)]	- [1)]	- [1)]	- [1)]
Explorative behaviour	0.26 (0.063)	−0.02 (0.871)	0.12 (0.374)	0.19 (0.167)	−0.19 (0.179)
Head posture	**0.34 (0.012)**	−0.22 (0.106)	−0.15 (0.276)	0.21 (0.131)	−0.06 (0.687)
Focus on task	−0.12 (0.389)	−0.01 (0.925)	−0.05 (0.702)	−0.19 (0.181)	−0.02 (0.897)
Pulling of rope	- [1)]	- [1)]	- [1)]	- [1)]	- [1)]
Time spent on task	0.05 (0.747)	−0.08 (0.561)	0.05 (0.698)	0.11 (0.413)	−0.20 (0.144)

Test moment 7: Tarpaulin					
Flight behaviour	**0.41 (0.002)**	0.03 (0.816)	−0.04 (0.764)	**0.27 (0.049)**	−0.14 (0.303)
Explorative behaviour	0.04 (0.797)	0.15 (0.286)	−0.01 (0.967)	−0.03 (0.849)	−0.23 (0.105)
Head posture	**0.29 (0.034)**	0.08 (0.571)	0.04 (0.797)	−0.02 (0.909)	−0.08 (0.562)
Focus on task	**−0.27 (0.048)**	0.11 (0.429)	0.18 (0.197)	−0.14 (0.302)	**0.36 (0.008)**
Pulling of rope	0.24 (0.089)	−0.00 (0.993)	−0.16 (0.257)	0.12 (0.389)	−0.24 (0.078)
Task performance	**−0.40 (0.003)**	0.05 (0.749)	0.10 (0.488)	−0.11 (0.446)	0.24 (0.078)
Time spent on task	**0.38 (0.005)**	−0.08 (0.575)	−0.22 (0.111)	0.09 (0.510)	**−0.34 (0.013)**

[1)] No expression of trait.

Appendix B. Rotated factor loadings (x100) of the standardized individual trait scores in the initial test (N = 63), and the variance explained by each of the eight first factors. Threatening behaviour and pulling of rope in test moment 6 were omitted from the analysis, due to no expression of these traits. Numbers in bold for factor loadings (x100) > |40|.

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8
Test moment 1: Log								
Flight behaviour	−9	**67**	22	4	6	27	−19	−12
Explorative behaviour	−4	31	19	13	34	3	39	−15
Head posture	−39	−19	27	−7	6	19	1	**57**
Focus on task	37	**−49**	**−52**	−4	−21	−14	−16	12
Pulling of rope	−31	27	7	−15	3	5	24	27
Task performance	4	−20	−2	15	12	3	**48**	1
Time spent on task	−22	35	**80**	2	24	11	8	4
Test moment 2: Serpent								
Head posture	**−48**	13	−8	2	15	12	−25	**48**
Focus on task	**85**	−1	−13	6	−7	−12	1	−7
Pulling of rope	**−66**	6	−5	9	5	3	−24	6
Task performance	**60**	−6	−5	3	10	−13	−12	−19
Test moment 3: Double-U								
Head posture	−38	0	−4	5	39	−11	15	28
Focus on task	**77**	−11	−6	−5	−8	−4	−15	−3
Pulling of rope	**−64**	14	1	8	9	15	−6	−6
Task performance	**48**	0	−9	33	−18	−11	−14	−23
Test moment 4: Bridge								
Flight behaviour	−17	**49**	9	10	−11	**47**	−13	18
Explorative behaviour	8	−8	−6	10	5	6	**69**	−7
Head posture	**−43**	16	20	24	2	−7	−15	20
Focus on task	32	−30	−31	−18	8	−30	**−55**	−16
Pulling of rope	**−57**	25	23	−8	15	5	−1	34
Task performance	13	−26	−4	8	4	**−43**	9	−12
Time spent on task	−10	20	**68**	18	−3	33	35	23
Test moment 5: Visual								
Flight behaviour	−12	22	14	**72**	−5	−8	8	3
Threatening behaviour	8	−20	1	38	−3	5	9	−8
Explorative behaviour	10	−9	−8	**42**	**66**	0	−10	4
Head posture	9	−4	1	**66**	25	−4	14	24
Focus on task	23	−1	−4	**−76**	−20	−17	−15	9
Pulling of rope	6	7	**83**	12	−4	−3	−16	19
Time spent on task	−2	−8	−3	**50**	**60**	13	−2	−7
Test moment 6: Sound								
Flight behaviour	−32	**43**	28	**46**	−7	−4	4	25
Explorative behaviour	−30	12	8	1	**72**	−16	10	21
Head posture	−10	27	12	34	1	−4	7	**55**
Focus on task	**53**	−16	−22	−28	−31	5	−9	−4
Time spent on task	−18	−1	12	−14	**80**	5	15	−4
Test moment 7: Tarpaulin								
Flight behaviour	−7	**74**	21	−2	0	6	−8	24
Explorative behaviour	−9	−17	−12	2	10	**61**	**40**	−4
Head posture	−23	26	23	−9	4	7	−20	**50**
Focus on task	15	−25	−29	−10	2	**−78**	−19	3
Pulling of rope	**−51**	29	**59**	−1	0	15	−20	7
Task performance	18	**−78**	−12	4	3	−10	1	−10
Time spent on task	−8	**43**	**53**	3	1	**59**	15	8
% variance	12,75	8,93	8,57	7,12	6,59	5,64	5,12	4,89
Cumulated		21,68	30,25	37,37	43,96	49,60	54,72	59,61

References

Christensen, J.W., Keeling, L.J., Nielsen, B.L., 2005. Responses of horses to novel visual, olfactory and auditory stimuli. Appl. Anim. Behav. Sci. 93, 53–65.

Christensen, J.W., Ahrendt, L.P., Lintrup, R., Gaillard, C., Palme, R., Malmkvist, J., 2012. Does learning performance in horses relate to fearfulness, baseline stress hormone and social rank? Appl. Anim. Behav. Sci. 140 (1-2), 44–52.

Cicchetti, D.V., 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psych. Ass. 6 (4), 284–290.

Clarke, A.S., Boinski, S., 1995. Temperament in nonhuman primates. Am. J. Prim. 37, 103–125.

Finkemeier, M.-A., Langbein, J., Puppe, B., 2018. Personality research in mammalian farm animals: concepts, measures and relationship to welfare. Front. Vet. Sci. 5, 131.

https://doi.org/10.3389/fvets.2018.00131.
Fiske, J.C., Potter, G.D., 1979. Discrimination reversal learning in yearling horses. J. Anim. Sci. 49, 583–588.
Gosling, S.D., John, O.P., 1999. Personality dimensions in nonhuman animals: a cross-species review. Am. Psychol. Soc. 8 (3), 69–75.
Graf, P., König von Borstel, U., Gauly, M., 2014. Practical considerations regarding the implementations of a temperament test into horse performance test: results of a large-scale test run. J. Vet. Behav. 9, 329–340.
Hem, L.E., Iversen, N.M., 2013. Forprosjekt: Markedsføringsstrategier for de nasjonale hesterasene/Marketing strategies for the Norwegian horse breeds (In Norwegian). Project report. Norwegian Equine Centre, Starum.
Ijichi, C., Collins, L.M., Creighton, E., Elwood, R.W., 2013. Harnessing the power of personality assessments: subjective assessment predicts behaviour in horses. Behav. Proc. 96, 47–52.
König von Borstel, U., Euent, S., Graf, P., König, S., Gauly, M., 2011. Equine behaviour and heart rate in temperament tests with or without rider or handler. Phys. Behav. 104, 454–463.
König von Borstel, U., Pirsich, W., Gauly, M., Bruns, E., 2012. Repeatability and reliability of scores from ridden temperament tests conducted during performance tests. Appl. Anim. Behav. Sci. 139, 251–263.
König von Borstel, U., Pasing, S., Gauly, M., Christmann, L., 2013. Status quo of the personality trait evaluation in horse breeding: judges' assessment of the situation and strategies for improvement. J. Vet. Behav. 8, 326–334.
Lansade, L., Bouissou, M.F., Erhard, H.W., 2008. Fearfulness in horses: a temperament trait stable across time and situations. Appl. Anim. Behav. Sci. 115, 182–200.
Lansade, L., Philippon, P., Hervé, L., Vidament, M., 2016. Development of personality tests to use in field, stable over time and across situations, and linked to horses' show jumping performance. Appl. Anim. Behav. Sci. 176, 43–51.
Leiner, L., Fendt, M., 2011. Behavioural fear and heart rate responses of horses after exposure to novel objects: Effects of habituation. Appl. Anim. Behav. Sci. 131, 104–109.
Le Scolan, N., Hausberger, M., Wolff, A., 1997. Stability over situations in temperamental traits of horses as revealed by experimental and scoring approaches. Behav. Proc. 41, 257–266.
Lindberg, A.C., Kelland, A., Nicol, C.J., 1999. Effects of observational learning on acquisition of an operant response in horses. Appl. Anim. Behav. Sci. 61, 187–199.
Marsbøll, A.F., Christensen, J.W., 2015. Effects of handling on fear reactions in young Icelandic horses. Equine Vet. J. 47, 615–619.
Mendl, M., 1999. Performing under pressure: stress and cognitive function. Appl. Anim. Behav. Sci. 65, 221–244.
McCall, C.A., Hall, S., McElhenney, W.H., Cummins, K.A., 2006. Evaluation and comparison of four methods of ranking horses based on reactivity. Appl. Anim. Behav. Sci. 96, 115–127.
Mengoli, M., Pageat, P., Lafont-Lecuelle, C., Monneret, P., Giacalone, A., Sighieri, C., Cozzi, A., 2014. Influence of emotional balance during a learning and recall test in horses (Equus caballus). Behav. Proc. 106, 141–150.
Montgomery, D.C., Peck, E.A., Vining, G.G., 2001. Introduction to Linear Regression Analysis, 3rd ed. John Wiley & Sons, inc ISBN 0-471-31565-6.
Norsk Hestesenter, 2012. Handlingsplan for Nasjonale Hesteraser 2011-2020/Plan of Action for the Norwegian Horse Breeds 2011-2020 (In Norwegian). Norwegian Equine Centre, Starum (approved by the board 24.4.2012).
Olsen, H.F., Klemetsdal, G., Ruane, J., Helfjord, T., 2010. Pedigree structure and genetic variation in the two, endangered Norwegian horse breeds: Døle and Nordland/Lyngen. Acta Agric. Scand. Sect. A 60, 13–22.
Olsen, H.F., Klemetsdal, G., 2010. Management to ensure effective population size in a breeding programme for the small Norwegian horse breeds – a simulation study. Acta Agric. Scand. Sect. A 60, 60–63.
Olsen, H.F., Klemetsdal, G., 2017. Temperament of the Norwegian horse breeds – a questionnaire based study. Appl. Anim. Behav. Sci. 193, 60–66.
Pierard, M., Hall, C., König von Borstel, U., Averis, A., Hawson, L., McLean, A., Nevison, C., Visser, K., McGreevy, P., 2015. Evolving protocols for research in equitation science. J. Vet. Behav. 10, 255–266.
Selle, T., 2010. Genetic Analysis of Show Results for the Norwegian Horse Breeds; the Dole Horse, the Fjord Horse and the Nordland/Lyngen Pony. M. Sc. Thesis. Norwegian University of Life Sciences 66 pp. (In Norwegian).
Shrout, P.E., Fleis, J.L., 1979. Intraclass correlations: Uses in assessing rater reliability. Psychol. Bull. 86, 420–428.
Valenchon, M., Lévy, F., Fortin, M., Leterrier, C., Lansade, L., 2013. Stress and temperament affect working memory performance for disappearing food in horses, Equus caballus. Anim. Behav. 86, 1233–1240.
Visser, E.K., van Reenen, C.G., van der Werf, J.T.N., Schilder, M.B.H., Knaap, J.H., Barneveld, A., Blokhuis, H.J., 2002. Heart rate and heart rate variability during a novel object test and a handling test in young horses. Phys. Behav. 76, 289–296.
Weinstein, T.A.R., Capitanio, J.P., Gosling, S.D., 2008. Personality in animals. In: John, O.P., Robins, R.W., Pervin, L.A. (Eds.), Handbook of Personality – Theory and Research, 3rd edition. The Guilford Press, New York ISBN 978-1-59385-836-0.
Wolff, A., Hausberger, M., Le Scolan, N., 1997. Experimental tests to assess emotionality in horses. Behav. Proc. 40, 209–221.