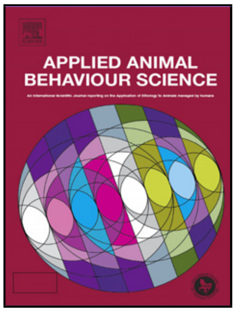


Title: Qualitative Behaviour Assessment of horses exposed to short-term emotional treatments

Authors: Sara Hintze, Eimear Murphy, Iris Bachmann, Francoise Wemelsfelder, Hanno W*ü*rbel

PII: S0168-1591(17)30189-2
DOI: http://dx.doi.org/doi:10.1016/j.applanim.2017.06.012
Reference: APPLAN 4483

To appear in: *APPLAN*

Received date: 25-12-2016
Revised date: 16-6-2017
Accepted date: 18-6-2017

Qualitative Behaviour Assessment of horses exposed to short-term emotional treatments

Sara Hintze [a, b *], Eimear Murphy [a], Iris Bachmann [b], Francoise Wemelsfelder [c], Hanno Würbel [a]

[a] Division of Animal Welfare, University of Bern, Länggassstrasse 120, 3012 Bern, Switzerland

[b] Agroscope, Swiss National Stud Farm, Les Longs-Prés, 1580 Avenches, Switzerland

[c] Animal & Veterinary Sciences Group, SRUC, Roslin Institute Building, Bush Estate, EH25 9RG, United Kingdom

*Corresponding author:

Sara Hintze

Present contact details and address:

Email: sara.hintze@boku.ac.at

Address: Division of Livestock Sciences, Department of Sustainable Agricultural Systems, University of Natural Resources and Life Sciences Vienna, Gregor-Mendel-Strasse 33, 1180 Vienna, Austria

Telephone: 0043 1 47654 93228

Highlights

- QBA uses observers' assessment of behavioural expression to assess animal emotions
- Horses were exposed to two positively and two negatively valenced treatments
- Three main dimensions of behavioural expression were identified
- Each of these dimensions distinguished significantly between treatments

- QBA may complement welfare assessments in situations of varying emotional valence

Abstract

Assessing emotion in animals is fundamental to the study of animal welfare with methodologies for reliable and valid assessments being highly desirable. Qualitative Behaviour Assessment (QBA) is based on the assumption that human observers are capable of integrating details of animals' behavioural expressions using descriptors (e.g. calm, playful) that reflect the animals' putative emotional experiences. Our study aimed at assessing how treatments assumed to induce different short-term emotional states of both positive and negative valence would affect the observers' judgements of horses' behavioural expressions. To this end, 16 horses were each exposed to two positive (grooming, food anticipation) and two negative treatments (food competition, waving a plastic bag) while being video-recorded. Using a Free Choice Profiling methodology, fifteen observers who were blind to treatment were asked to describe and score the horses' behavioural expressions based on 45 second long video clips. General Procrustes Analysis revealed consensus between the observers' judgements. Three main dimensions of behavioural expression were identified, explaining 84.7 % of the variation between horses. Dimension 1 (D1) was positively associated with the terms 'calm/relaxed/content' and negatively with the terms 'nervous/ stressed', whereas dimension 2 (D2) was described as ranging from 'irritated/impatient/angry' to 'frightened/insecure', and dimension 3 (D3) was labelled as ranging from 'curious/interested' to 'aggressive/irritated'. Linear mixed-effect models revealed an effect of treatment on the horse scores on all three dimensions (D1: $F_{4,60} = 86.90$, $p < 0.0001$; D2: $F_{4,60} = 69.57$, $p < 0.0001$, D3: $F_{4,60} = 11.05$, $p < 0.0001$). In line with our hypotheses, horses were judged as 'calm/relaxed/content' (D1) during grooming, whereas they were assessed as 'stressed/nervous' (D1) and 'insecure/frightened' (D2) when exposed to the plastic bag. In the two food treatments (food anticipation, food competition), horses were judged as 'irritated/impatient/angry' (D2). However, on dimension D3, horses during food anticipation were also assessed as more 'curious/interested' than in any other treatment. Our study demonstrates that observers showed consensus in their assessment of horses that were exposed to positive and negative short-term treatments and that they could differentiate between these treatments consistent with our hypotheses. Our results indicate that QBA is a promising tool to complement animal welfare assessments in situations of multiple emotional dimensions of both positive and negative valence.

Keywords: Emotional state, Qualitative Behaviour Assessment, Free Choice Profiling, behavioural expression, horse, animal welfare

1. Introduction

Assessing emotion in animals is challenging but fundamental to the study of animal welfare. Various methodologies have been developed which are aimed at assessing animals' emotional states, including behavioural, physiological and cognitive approaches (overview e.g. in Paul et al., 2005). Qualitative Behaviour Assessment (QBA; Wemelsfelder et al., 2001, 2000) is a 'whole animal' approach focusing on assessing and quantifying the expressive quality of an animal's dynamic interaction with the environment. It is based on the premise that human observers are able to integrate details not only of *what* the animal is doing, but also of *how* it is doing it, describing the animal's behavioural expression or body language in terms (e.g. content, nervous, curious) that reflect its putative emotional experience (Wemelsfelder, 2007). Since the first experimental paper on QBA was published in 2000 (Wemelsfelder et al., 2000), the methodology has been applied to assess emotional states as judged by observers in different species, including pigs (Wemelsfelder et al., 2001, 2000), cattle (e.g. Rousing and Wemelsfelder, 2006), sheep (e.g. Wickham et al., 2015), dogs (e.g. Walker et al., 2016) and horses (e.g. Napolitano et al., 2008), and in a variety of contexts, for example during interactions with humans (Minero et al., 2009), during transport (e.g. Stockman et al., 2011), or as an additional outcome measure during standard behavioural tests (Rutherford et al., 2012). In horses, QBA has been applied to assess the animals' responsiveness to a human handler (Minero et al., 2009) and in an Open Field test (Napolitano et al., 2008), both in association with quantitative measures. Furthermore, it has been used to evaluate horses' behavioural expressions at different time points during endurance rides (Dorman et al., 2010; Fleming et al., 2013).

Besides its application to investigate animals' emotional states in research, QBA is also part of on-farm welfare assessment protocols like Welfare Quality® and AWIN (Animal Welfare Indicators), where it has been applied to assess positive emotional states in several species, including pigs (Welfare Quality®, 2009a), cattle (Welfare Quality®, 2009b), sheep (AWIN, 2015a), goats (AWIN, 2015b), donkeys (AWIN, 2015c), and horses (AWIN, 2015d).

QBA has been shown to be reliable both within and between observers (e.g. Wemelsfelder et al., 2012). In terms of validity, a number of studies have found significant meaningful correlational relationships between QBA and quantitative measures of behaviour (e.g. Minero et al., 2009; Rousing and Wemelsfelder, 2006) as well as between QBA and physiological stress parameters (e.g. Stockman et al., 2011; Wickham et al., 2012), thereby supporting the concurrent validity of QBA (Winckler, 2015). In horses, for example, yearlings exposed to a one-month period of daily handling were characterised as more 'explorative/social' and 'calm/apathetic' and engaged in more physical contact with a human handler compared to before the handling treatment (Minero et al., 2009).

Besides these correlational approaches, QBA has also been validated using pharmacological treatments. To this end, Rutherford and colleagues (2012) specifically manipulated pigs' emotional states using azaperone, which is suggested to have an anxiolytic effect in pigs (Donald et al., 2011). Observers, blind to the treatment, scored the azaperone treated pigs as more confident and curious compared to the untreated pigs. However, in their study only one emotional dimension ('anxious' versus 'confident/ curious') was found to be manipulated by the drug. So far, no study has investigated the ability of QBA to discern the emotional responses of animals exposed to specific test situations aimed at inducing emotional states that potentially vary along multiple emotional dimensions.

In the present study, we exposed 16 horses to a variety of treatments designed to induce both positive and negative emotional states. Responses of the horses to these manipulations were assessed by human observers following the QBA methodology (Wemelsfelder et al., 2001, 2000). Two positive treatments, food anticipation for a high value reward (FA) and grooming (G), were hypothesised to induce states of positive anticipation or relaxation, respectively (e.g. Boissy et al., 2007; Feh and de Mazières, 1993; McBride et al., 2004; Van Den Bos et al., 2003). In contrast, two negative treatments, food competition (FC) and waving of a plastic bag (PB), were hypothesised to result in frustration and fear, respectively (Christensen et al., 2008; König von Borstel et al., 2010; Ninomiya et al., 2004). We have recently demonstrated that these treatments are associated with consistent variation in a measure (angle) of eye wrinkle expression in horses (Hintze et al., 2016). In this paper we focus on the behavioural expression, or body language, of the horse. Our aim was to investigate how the four treatments would affect the observers' judgements of horses' behavioural expression, to determine whether 1) observers would reach consensus in their judgements, and if so, whether 2) observers' judgements would differentiate between the treatments according to our hypotheses.

2. Materials and methods

2.1 Animals and housing

Sixteen horses (15 stallions, 1 mare) were randomly selected from the horse population of the Swiss National Stud Farm of Agroscope (Avenches, Switzerland), excluding horses that did not respond to one of the treatments (food competition) during pilot observations. Subjects ranged in age from 3 to 20 years (10.4 ± 4.7) and represented three breeds (Franches-Montages, warmblood, trotter). All horses were housed in standard individual boxes (3.00 x 3.50 m). The stable was divided into six discrete areas, and each area contained eight boxes in two rows which were separated by an aisle (3 m). The 16 horses were distributed across each of the six areas, with no more than four horses in the same area. Water was provided ad libitum by automatic drinkers and horses were fed hay and concentrates three times a day. On a daily basis, horses were allowed free movement on a sand paddock, they were walked in a horse walker, or otherwise exercised (riding, carriage riding). Details on the individual horses are described in Hintze et al. (2016).

2.2 Experimental design

2.2.1 Overview

Over a series of ten experimental sessions, each horse was exposed to two positive, two negative, and one control treatment. Horses were never confronted with more than one treatment per session. Due to the number of animals, not all horses were tested in each session, but both the order of horses within one session and the order of treatments across sessions were balanced across horses. Responses of the horses were filmed in the test periods of each treatment which lasted 60 seconds. All treatments occurred in the home box of each horse, ensuring that the background of each video clip remained constant across treatments, and to avoid potential influences of novelty or movement on the horses' emotional states.

Prior to the first positive treatment, food anticipation (FA), each horse was trained to associate a specific bucket with a highly-valued food reward. During testing, this bucket was visible to the horse for 60 seconds (test period) before it was given access to the food reward. In the second positive treatment, grooming (G), the experimenter moved the fingertips of both hands along the horse's withers, shoulder and neck for 60 seconds. During G, the horse was positioned between the experimenter and the camera to ensure that the experimenter's hand movements were not visible. However, we could not avoid that parts of the experimenter's body were visible in the video clips. To control for the presence of the experimenter during G, we included a 60 second control phase which preceded the G treatment and in which the experimenter performed the same movements but without actually touching the horse (grooming control, G_{con}). In the first negative treatment, food competition (FC), all horses of one area of the stable were fed one hour later than usual and with the focal horse being fed last. Here, the 60 seconds during which the other horses were fed, were assigned as test period. In the second negative treatment, plastic bag (PB), one experimenter who was positioned outside the box in the aisle, waved a plastic bag and an umbrella for 60 seconds to induce a fear response. All treatments were chosen based on literature demonstrating an either positive or negative effect on horses' emotional states as assessed by behavioural and/ or physiological indicators (e.g. Boissy et al., 2007; Feh and de Mazières, 1993; McBride et al., 2004; Van Den Bos et al., 2003; Christensen et al., 2008; König von Borstel et al., 2010; Ninomiya et al., 2004). For additional details on the treatments and the experimental design see Hintze et al. (2016). The focal horse was filmed by three cameras (GoPro Hero 3, Sony camcorder HDR-CX1115E, Sony camcorder FDR-AX33) mounted at different locations inside and outside the box. Since the behaviour of individual horses in the different treatments was not predictable (e.g. jumping to the back of the box, looking out of the window), having the three cameras allowed us to select the clips that provided the best view of the horses' behavioural expressions. After the last session, one video clip per horse per treatment was chosen based on the quality of the clips and the visibility of the horse's face and body.

2.3 Video clip processing

A total of 80 clips (16 horses x 4 treatments plus 1 control) were cropped to a length of 45 seconds each to ensure the exact same length of all clips. The decision of whether to include the first or the last 45 seconds of the 60 second video clips was based on our assumptions of when the horses' peak emotional response was likely to occur during a particular treatment; this differed between treatments but was consistent within treatments. For FA we selected the first 45 seconds of each clip to minimise the risk of picking up frustration which might occur the longer the horses is waiting for the food. We assumed that horses would also react most strongly early in the PB treatment. In contrast, frustration of horses during FC was deemed the stronger the longer the focal horse had to wait to be fed. Moreover, horses began to show facial expressions typically associated with social grooming later during G, as seen during pilot observations. These expressions included, besides others, a stretched or curved neck, sideways pointing ears and movement of the lips with a prolonged upper and a loose lower lip (so-called ‘grooming face’, Neugebauer and Neugebauer, 2011). Consequently, the first 45 seconds were selected for FA and PB, whereas the last 45 seconds were selected for FC and G (as well as for G_{con} to have an appropriate control).

2.4 Qualitative Behaviour Assessment (QBA)

2.4.1 Observers

Observers were 15 female veterinary students from the Universities of Bern and Zürich, Switzerland. While not specifically trained in observing animal behaviour, they all were experienced in handling and riding horses with their experience ranging from eight to 20 years.

2.4.2 Free Choice Profiling (FCP) procedure

All observers participated in three sessions distributed over two weeks. In the first session terms to describe the horses’ behavioural expressions, e.g. relaxed, curious or agitated, were generated for use in subsequent sessions (‘term generating session’). To this end, observers were told that the aim of the study was to learn more about the body language and the behavioural expression of horses. No information on the different treatments was given. Observers were then asked to watch 24 of the 80 clips which were selected to cover all treatments and a wide range of different behavioural expressions. No clip included sound to ensure that observers were blind to all treatments. After each clip observers were given two minutes to note all terms that they thought would adequately describe the horse’s behavioural expression. Communication language was German but since some observers were French native speakers, they were allowed to note terms in either German or French. All terms were then translated to English by the experimenter and double-checked by a native speaker.

Following this first session, individual scoring sheets were created by collating all terms that a particular observer had generated, and then listing these in random order next to Visual Analogue Scales (VAS). Thus observers only used their own terms for scoring. . Each VAS was 121 mm long and ranged from the ‘minimum’ on the left to the ‘maximum’ on the right with an uncategorised continuous line between these two points.

In the following two sessions observers scored all 80 clips on their individual list of previously generated terms (‘quantification sessions’). In each session, observers were presented with 40 clips which were as far as possible balanced for horse and treatment. Clips were presented in a pseudorandom order with no more than two clips of the same horse or the same treatment shown in a row. After each clip, observers were given two minutes to score the behavioural expression of the horse according to the terms they had previously generated. Observers were instructed to make a vertical mark crossing each VAS, at a point where the distance from the zero-point to the mark was taken to reflect the perceived intensity of the expression indicated by a particular term for a particular clip. Scores were then determined by measuring the distance in millimetres from the ‘minimum’ on the left to the point where the line was crossed by the observer’s mark. An individual spreadsheet per observer was then created which was defined by the number of clips (n = 80) and the number of terms used by that observer.

2.5 Statistical analyses

2.5.1 Calculation and interpretation of consensus profiles

Data were analysed with General Procrustes Analysis (GPA) which has been described in detail elsewhere (Wemelsfelder et al., 2001, 2000). GPA is a multivariate technique that captures underlying patterns in the data without the need for common fixed variables. To this end, GPA transforms the scoring patterns of each observer into multidimensional configurations and determines the ‘best fit’, the so-called consensus profile, by a complex pattern matching process (rotations, transformations). This process is entirely independent of the meaning of the different terms used by the observers. The achieved level of consensus (i.e. the percentage of variation between observer scoring configurations explained by the consensus) is given in the Procrustes Statistic (PS). In order to test whether the consensus profile is a feature of the data set or an artefact of the GPA procedure, each observer’s scores are rearranged randomly, generating new randomised data matrices. A randomised consensus profile is calculated by applying GPA to the randomised data matrices, iterating this process 100 times. Then a one-tailed Student's t-test is run to determine whether the Procrustes Statistic from the 'true' consensus profile significantly differs from the mean of the 100 randomised Procrustes Statistics, with a p-value < 0.001 taken to indicate that the 'true' consensus is a pattern identified by observers, and not merely an artefact of statistical procedures. The number of dimensions of the consensus profile corresponds to the maximum number of terms used by any of the 15 observers (here 39 dimensions). In order to facilitate interpretation, the number of dimensions is reduced through Principal Component Analysis (PCA) to identify those dimensions that explain the majority of the variation between the observed horses. These main dimensions are then semantically interpreted by correlating the animal scores for each of an individual observer's terms with the animal scores for each of the main PCA dimensions, resulting in two-dimensional 'word charts' for each observer in which an observer's terms are projected on to various combinations of the consensus dimensions (e.g. 1st vs 2nd, or 1st vs 3rd dimension). The stronger the correlation between a term and a dimension, the more weight it has as a descriptor for that dimension. Both positive and negative correlations are important to give meaning to either ends of the main dimensions. In this study terms that best described the anchor points of a dimension were chosen according to the following rules: First, only terms with a certain correlation coefficient were included (see below). Second, from these, the two or three most frequently used terms were chosen. Cut-off points were $r > 0.8$ for positive and $r < -0.8$ for negative correlations for dimension 1, $r > 0.5$ and $r < - 0.5$ for dimension 2, and $r > 0.4$ and $r < -0.4$ for dimension 3. Data from all treatments were analysed together in a single GPA (joined analysis), and also for each of the five treatments separately (separate analyses).

2.5.2 The effect of treatment on QBA scores

To investigate the impact of the different treatments on the horses' behavioural expressions as judged by the observers and represented in the horses' scores on the main consensus dimensions, we used linear mixed-effects models which adequately reflect dependencies within the experimental design (nested structure, repeated measures). Data were analysed in R (version 3.2.1) with the lme function of the package nlme (Pinheiro et al., 2016). In the GPA each animal receives one score per treatment on each of the main consensus dimensions. These horse scores were used as primary outcome variables in our analyses. Models were run with the data of the first three dimensions. Treatment was included as fixed effect, and treatment nested in horse was included as random effect. G and G_{con} were treated as independent treatments even though dependency between these two treatments was greater than between the other treatments since horses were exposed to G and G_{con} on the same day. However, with only one control phase (G_{con}), we could not include phase as additional random effect and could thus not correct for it statistically.

In cases of significant differences between treatments, pairwise post-hoc analyses were conducted (function: glht, package: multcomp; Hothorn et al., 2008), and p-values were adjusted for multiple testing by a single-step method incorporating the correlations between the test statistics (Bretz et al., 2011). Model assumptions of the linear models (normal distribution, homogeneity of variance) were verified by graphical inspections of the residuals. No transformation of the data was necessary.

2.5.3 Variation within treatments

To assess consistency of the treatment effects across horses, we also measured within-treatment variation in behavioural expression by calculating Interquartile Ranges (IQRs) of the scaling values for all three dimensions and all four treatments (Mosteller and Tukey, 1977).

2.6 Ethical considerations

This experiment was conducted in compliance with the ethical policy of the International Society for Applied Ethology and approved by the Veterinary Office of the Canton of Vaud according to the Swiss Animal Welfare legislation (Vaud, Switzerland, Approval Number VD 2804).

3. Results

3.1 Joined analysis

3.1.1 Consensus profile and dimensions

The 15 observers generated a total of 137 different descriptive terms (mean ± SD: 31.60 ± 5.70 terms per observer; range: 20 – 39 terms). The consensus profile of the GPA explained 75.41 % (= Procrustes Statistics) of the variation in observer scoring patterns, which differed significantly from the mean Procrustes Statistic of the 100 randomised profiles (mean ± SD Procrustes Statistic: 26.94 % ± 0.22; t_{99} = 224.00, $p < 0.001$, Table 1), indicating that the consensus profile reflected an underlying feature of the data set which was not generated by chance.

Three main dimensions of behavioural expression were identified explaining in total 84.7 % of the variation between horses. 56.5 % of the variation was captured by dimension 1 (D1), 17.4 % by dimension 2 (D2), and 9.8 % by dimension 3 (D3). Even though observers used their own terms to describe the horses' behavioural expressions, terms for the different dimensions were semantically coherent. Dimension 1 was characterised by terms ranging from 'calm/relaxed/content' to 'nervous/stressed', whereas dimension 2 was described by 'irritated/impatient/angry' to 'frightened/insecure', and dimension 3 was labelled ranging from 'curious/interested' to 'aggressive/irritated'. An overview of the terms that were positively and negatively correlated with each dimension is given in Table 2, and two example 'word charts' are presented in Fig. 1.

3.1.2 Treatment effects

Treatment had a highly significant effect on all three main dimensions (D1: $F_{4,60}$ = 86.90, $p < 0.0001$; D2: $F_{4,60}$ = 69.57, $p < 0.0001$, D3: $F_{4,60}$ = 11.05, $p < 0.0001$). For dimensions 1 and 2, post-hoc tests revealed differences between all treatments, except for G and Gcon which did not differ from each other (Table 3). On dimension 1, horses were judged as most 'calm/relaxed/content' during G and Gcon whereas they were assessed as most 'stressed/nervous' during PB. In the two food treatments (FA and FC), horses did not score high on either end of dimension 1 but still differed from each other with horses being judged as more ‘stressed/nervous’ during FC than during FA (Fig. 2A). On dimension 2, observers judged horses during FC as most ‘irritated/impatient/angry’, followed by horses during FA. Horses during PB were again assessed as ‘insecure/frightened’, whereas horses during G and Gcon were neither judged as particularly ‘irritated/impatient/angry’ nor as ‘insecure/frightened’ (Fig. 2B). On dimension 3, horses were judged as significantly more ‘curious/interested’ during FA than in any other treatment (Fig. 2C). Results of all post-hoc tests are given in Table 3.

3.1.3 Variation within treatments

For each of the four treatments, within-treatment variation between horses was investigated based on the IQRs of the scaling values (Table 4, see also Fig. 2 for graphical display of the IQRs). Looking at all three dimensions, variation in the scores was greatest in FC (mean ± SD = 0.063 ± 0.015) and smallest in G (mean ± SD = 0.013 ± 0.006), with intermediate variation in the other two treatments (Gcon: mean ± SD = 0.027 ± 0.006; PB: mean ± SD = 0.030 ± 0.010; FA: mean ± SD = 0.037 ± 0.021).

3.2 Separate Analyses

In addition to the analysis incorporating all treatments (joined analysis, see above), a GPA was also run for each treatment separately (separate analyses). As with the joined analysis, the Procrustes Statistics of the consensus profiles differed significantly from the mean Procrustes Statistic for the 100 randomised profiles for all five separate analyses (Table 1). Moreover, interpretation of at least the two main dimensions of each analysis revealed semantic coherence (Table 2).

Discussion

This study aimed to investigate whether human observers using the QBA methodology would reach consensus in their judgements of horses that were exposed to multiple positive and negative short-term emotional treatments, and if so, whether they would differentiate between these treatments based on the horses’ behavioural expressions in line with our hypotheses. Significant consensus between observers was found for both the joined analysis of all treatments and the separate analyses per treatment. In the joined analysis, three main consensus dimensions of behavioural expression (D1, D2, D3) were identified which together explained 84.7 % of the variation between horses, and each showed a significant effect of treatment. During G and G_{con}, horses were judged as most ‘calm/relaxed/content’ (D1), whereas they were assessed as most ‘stressed/nervous’ (D1) and ‘insecure/frightened’ (D2) in the PB treatment. In the two food treatments (FA, FC), horses were judged as most ‘irritated/impatient/angry’ relative to other treatments (D2). However, on a third dimension (D3), horses during FA were also assessed as more ‘curious/interested’ than in any other treatment.

3.3 Consensus profile and dimensions

Various studies across a range of species have investigated the relationship between QBA outcomes and both behavioural (e.g. Minero et al., 2009; Rousing and Wemelsfelder, 2006) and physiological measures of welfare (e.g. Stockman et al., 2011; Wickham et al., 2012), thereby supporting the concurrent validity of QBA (Winckler 2015). To our knowledge the current study is only the second one to apply QBA in a situation in which animals’ emotional states were experimentally manipulated, allowing us to formulate *a priori* hypotheses on how horses would be affected emotionally by the different treatments. In an earlier study by Rutherford and colleagues (2012), pigs were treated with an anxiolytic drug with a well-studied pharmacological mechanism of action (Posner and Burns, 2009), whereas treatments in our study were chosen based on known effects on behaviour and physiology in horses (e.g. FA: Boissy et al., 2007; Van Den Bos et al., 2003, G: Feh and de Mazières, 1993; McBride et al., 2004; FC: Ninomiya et al., 2004; PB: Christensen et al., 2008; König von Borstel et al., 2010). The advantage of our approach is that we included a number of manipulations aimed at eliciting changes in emotional states across different contexts and of both positive and negative valence. We found that observers reached consensus in their judgements of horses in the different treatments, both when these treatments were analysed jointly, and when analysed separately (research question 1). In addition to the statistical consensus between observer scoring patterns, we also found semantic coherence in the meaning of terms selected to label the three main consensus dimensions. D1 was described by terms ranging from 'calm/relaxed/content' to 'nervous/stressed', whereas D2 was labelled ranging from 'irritated/impatient/angry' to 'frightened/insecure', and D3 was characterised by 'curious/interested' to 'aggressive/irritated'. These dimensions reflect different aspects of emotional valence, with only D1 also reflecting some degree of arousal ('calm', 'relaxed' to 'stressed', 'nervous'). Whereas many studies using QBA report two main dimensions reflecting valence and arousal (e.g. Camerlink et al., 2016; Ellingsen et al., 2014; Minero et al., 2016), others, like our study, describe three main dimensions characterising different aspects of valence (e.g. Grosso et al., 2016; Walker et al., 2010, 2016). Both anchor points of D2 reflect different aspects of negative valence ('irritated/impatient/angry' versus 'insecure/frightened'), thereby underlining the sensitivity of QBA to differentiate not only between positive and negative valence but also between different aspects of the same valence.

3.4 Treatment effects

Qualitative judgements of horses' behavioural expressions made by observers who were blind to treatment differed significantly between all four treatments and were consistent with our hypotheses, with a few exceptions (research question 2). Horses were assessed as most 'calm/relaxed/content' (D1) during G and G_{con}, as most 'stressed/nervous' (D1) and 'insecure/frightened' (D2) during PB, as most 'irritated/impatient/angry' (D2) during FC, and as most 'curious/interested' (D3) during FA. However, horses during FA were not only judged as significantly more 'curious/interested' (D3) than horses in all other treatments, but were also assessed as more 'irritated/impatient/angry' compared to horses during G, G_{con} and PB (D2), even though significantly less so than when exposed to FC. This might be due to individual variation in how horses perceived and consequently responded to FA, with some horses being judged as 'curious/interested' (D3) and others as more 'irritated/impatient/angry' (D2). However, the relatively small variation between horses during FA does not support this argument. Alternatively, emotional states might have varied within individuals rather than between individuals, with some horses showing a transition from positive anticipation (‘curious/interested’, D3) to frustration (‘irritated/impatient/angry’, D2) over time. Such transitions might emerge if the animal lacks control over the situation (Anderson, 2016) or if it takes too long for the anticipated positive event to occur, i.e. in our study until the horse was given access to the food in the bucket (Oppermann Moe et al., 2009). It would be valuable to investigate precisely when frustration starts to develop in such situations, in order to determine if a threshold exists at which frustration starts to override anticipation, or whether anticipation and frustration fluctuate throughout the treatment. In order to detect such changes over time, continuous recording, for example by identifying the most dominant QBA term at any point in time as proposed by Napolitano and colleagues (2015), might be a promising approach in future studies.

When exposed to G and G_{con}, horses were judged as ‘calm/relaxed/content’ (D1) but scores did not differ between G and G_{con}. We would have expected a difference between these two situations, at least if observers had focused predominantly on the face, since most horses showed the typical ‘grooming face’ (Neugebauer and Neugebauer, 2011) during G but not during G_{con}. Several potential explanations for this lack of difference can be considered. We cannot rule out that observers were biased by the presence of the experimenter in the video clips of these two treatments, and consequently assumed horses’ behavioural expressions to be similar, which would mean that our control condition did not work properly. Wemelsfelder and colleagues (2009) previously found that environmental background affected observers’ judgements of pig behavioural expression, but this effect only reflected a slight shift and did not seriously distort the characterisation of individual pigs. Similarly, it could be that, due to the presence of the experimenter, observers failed to note and record (subtle) existing differences between G and G_{con}, yet this would not necessarily invalidate the characterisation of both treatments as ‘calm/relaxed/content’. Whether and how much in the present study the presence of the experimenter affected the observers’ assessments is difficult to answer. Alternatively, the lack of differences between horses’ behavioural expressions in G and G_{con} might be real. Dimension 1 (‘calm/relaxed/content’ to ‘stressed/nervous’) did not just reflect emotional valence but also some degree of arousal and, given that horses showed low levels of arousal during both G and G_{con}, they might have appeared very similar in these two situations compared to all other treatments. Furthermore, even though G_{con} was a proper control condition for G, it might not have been a ‘neutral’ situation for the animal. The presence of the experimenter inside the box might have had a calming effect on the horses independent of any effect of the grooming treatment (e.g. König von Borstel et al., 2011). Finally, it should be noted that treatment effects were established statistically based on the variation induced by the treatments. Since our treatments were designed to exert strong effects, the resulting scaling values might have masked any potential statistical significance of the more subtle differences between G and G_{con}. In this case differences between G and G_{con} might have been perceived and recorded by observers but did not register statistically.

3.5 Variation within treatments

As with the treatment effects, variation between horses within treatments does not have absolute meaning but can only be interpreted in relation to the treatment effects. Variation between horses was smallest during G and greatest during FC across all three dimensions, with intermediate variation in the other treatments (G_{con}, FA, PB). The small variation during G and G_{con} might at least in part be explained by the presence of the experimenter in this treatment, since horses are used to stand still in such a situation. Furthermore, the dependency between these two situations was greater than between all other treatments since the two situations occurred on the same day, with G_{con} directly preceding G.

Variation between horses was greatest in the FC treatment, which might be due to the fact that during FC the horses’ responses were influenced by the neighbouring horses, who were all to some extent undergoing some level of food competition as they were not all fed simultaneously (Creighton and Hockenhull, 2010). Dependent on the behaviour of the neighbouring horses, the strength of the reaction of the focal horse could have differed, leading to greater variation between horses.

4. Conclusion

In the present study we experimentally elicited changes in horses' emotional states across different contexts. Using a Free Choice Profiling methodology commonly used in QBA, we demonstrated that 15 observers showed consensus, both statistically and semantically, in their treatment-blind judgements of the behavioural expressions of horses that were exposed to two positive and two negative short-term emotional treatments, and that these judgements differentiated between the four treatments, thereby supporting the construct validity of QBA. The fact that observers were able to judge the horses' behavioural expressions based on very short video clips consistently and in line with our hypotheses, indicates that our results might be relevant for horse-human interactions in which an immediate assessment of a horse's emotional state is important. Our results demonstrate that QBA is a promising tool to complement animal welfare assessments in situations of multiple emotional dimensions of both positive and negative valence.

Acknowledgements

We are grateful to the vet students from the Vetsuisse who participated as observers in this study. We also would like to thank Astrid Hintze and Ana Stojilkovic for patiently helping to measure all the lines of the VAS. Thanks to Samantha Smith for processing the video clips and to the staff from the stud farm for their logistical support. This work was funded by Agroscope.

References

Anderson, C., 2016. Investigating anticipatory behaviours in lambs (Doctoral thesis). Retrieved from the Swedish University of Agricultural Sciences Library. http://urn.kb.se/resolve?urn=urn:nbn:se:slu:epsilon-e-3738

AWIN, 2015a. AWIN welfare assessment protocol for sheep. doi:10.13130/AWIN_SHEEP_2015

AWIN, 2015b. AWIN welfare assessment protocol for goats. doi:10.13130/AWIN_GOATS_2015 March

AWIN, 2015c. AWIN welfare assessment protocol for donkeys. doi:DOI: 10.13130/AWIN_DONKEYS_2015 March

AWIN, 2015d. AWIN welfare assessment protocol for horses. doi:10.13130/AWIN_HORSES_2015

Boissy, A., Manteuffel, G., Jensen, M.B., Moe, R.O., Spruijt, B., Keeling, L.J., Winckler, C., Forkman, B., Dimitrov, I., Langbein, J., Bakken, M., Veissier, I., Aubert, A., 2007. Assessment of positive emotions in animals to improve their welfare. Physiol. Behav. 92, 375–397. doi:10.1016/j.physbeh.2007.02.003

Bretz, F., Hothorn, T., Westfall, P., 2011. Multiple comparisons using R. Taylor & Francis, Boca Raton. doi:10.2307/1266041

Camerlink, I., Peijnenburg, M., Wemelsfelder, F., Turner, S.P., 2016. Emotions after victory or defeat assessed through qualitative behavioural assessment, skin lesions and blood parameters in pigs. Appl. Anim. Behav. Sci. 183, 28–34. doi:10.1016/j.applanim.2016.07.007

Christensen, J.W., Malmkvist, J., Nielsen, B.L., Keeling, L.J., 2008. Effects of a calm companion on fear reactions in naive test horses. Equine Vet. J. 40, 46–50. doi:10.2746/042516408X245171

Creighton, E., Hockenhull, J., 2010. Feeding routine risk factors associated with pre-feeding behavior problems in UK leisure horses. J. Vet. Behav. Clin. Appl. Res. 5, 48. doi:10.1016/j.jveb.2009.09.012

Donald, R.D., Healy, S.D., Lawrence, A.B., Rutherford, K.M.D., 2011. Emotionality in growing pigs: Is the open field a valid test? Physiol. Behav. 104, 906–913. doi:10.1016/j.physbeh.2011.05.031

Dorman, C., Barnes, A., Fleming, P.A., 2010. Qualitative behavioral assessment (QBA) of horses competing in a 160 km endurance ride. J. Vet. Behav. Clin. Appl. Res. 5, 212. doi:10.1016/j.jveb.2009.11.009

Ellingsen, K., Coleman, G.J., Lund, V., Mejdell, C.M., 2014. Using qualitative behaviour assessment to explore the link between stockperson behaviour and dairy calf behaviour. Appl. Anim. Behav. Sci. 153, 10–17. doi:10.1016/j.applanim.2014.01.011

Feh, C., de Mazières, J., 1993. Grooming at a preferred site reduces heart rate in horses. Anim. Behav. 46, 1191–1194. doi:10.1006/anbe.1993.1309

Fleming, P.A., Paisley, C.L., Barnes, A.L., Wemelsfelder, F., 2013. Application of Qualitative Behavioural Assessment to horses during an endurance ride. Appl. Anim. Behav. Sci. 144, 80–88. doi:10.1016/j.applanim.2012.12.001

Grosso, L., Battini, M., Wemelsfelder, F., Barbieri, S., Minero, M., Dalla Costa, E., Mattiello, S., 2016. On-farm Qualitative Behaviour Assessment of dairy goats in different housing conditions. Appl. Anim. Behav. Sci. 180, 51–57. doi:10.1016/j.applanim.2016.04.013

Hintze, S., Smith, S., Patt, A., Bachmann, I., Würbel, H., 2016. Are eyes a mirror of the soul? What eye wrinkles reveal about a horse’s emotional state. PLoS Biol. 11, e0164017. doi:10.1371/journal.pone.0164017

Hothorn, T., Bretz, F., Westfall, P., Heiberger, R.M., 2008. Simultaneous inference in general parametric models. Biometrical J. 50, 346–363.

König von Borstel, U., Duncan, I.J.H., Lundin, M.C., Keeling, L.J., 2010. Fear reactions in trained and untrained horses from dressage and show-jumping breeding lines. Appl. Anim. Behav. Sci. 125, 124–131. doi:10.1016/j.applanim.2010.04.015

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. Physiol. Behav. 104, 454–463. doi:10.1016/j.physbeh.2011.05.010

McBride, S.D., Hemmings, A., Robinson, K., 2004. A preliminary study on the effect of massage to reduce stress in the horse. J. Equine Vet. Sci. 24, 76–81.

Minero, M., Dalla Costa, E., Dai, F., Murray, L.A.M., Canali, E., Wemelsfelder, F., 2016. Use of Qualitative Behaviour Assessment as an indicator of welfare in donkeys. Appl. Anim. Behav. Sci. 174, 147–153. doi:10.1016/j.applanim.2015.10.010

Minero, M., Tosi, M.V., Canali, E., Wemelsfelder, F., 2009. Quantitative and qualitative assessment of the response of foals to the presence of an unfamiliar human. Appl. Anim. Behav. Sci. 116, 74–81. doi:10.1016/j.applanim.2008.07.001

Mosteller, F., Tukey, J.W., 1977. Data anaylsis and regression - a second course in statistics. Addison-Wesley, Reading.

Napolitano, F., De Rosa, G., Braghieri, A., Grasso, F., Bordi, A., Wemelsfelder, F., 2008. The qualitative assessment of responsiveness to environmental challenge in horses and ponies. Appl. Anim. Behav. Sci. 109, 342–354. doi:10.1016/j.applanim.2007.03.009

Napolitano, F., De Rosa, G., Serrapica, M., Braghieri, A., 2015. A continuous recording approach to qualitative behaviour assessment in dairy buffaloes (Bubalus bubalis). Appl. Anim. Behav. Sci. 166, 35–43. doi:10.1016/j.applanim.2015.01.017

Neugebauer, G.M., Neugebauer, J.K., 2011. Lexikon der Pferdesprache. Eugen Ulmer KG, Stuttgart.

Ninomiya, S., Kusunose, R., Sato, S., Terada, M., Sugawara, K., 2004. Effects of feeding methods on eating frustration in stabled horses. Anim. Sci. J. 75, 465–469. doi:10.1111/j.1740-0929.2004.00214.x

Oppermann Moe, R., Nordgreen, J., Janczak, A.M., Spruijt, B.M., Zanella, A.J., Bakken, M., 2009. Trace classical conditioning as an approach to the study of reward-related behaviour in laying hens: A methodological study. Appl. Anim. Behav. Sci. 121, 171–178. doi:10.1016/j.applanim.2009.10.002

Paul, E.S., Harding, E.J., Mendl, M., 2005. Measuring emotional processes in animals: the utility of a cognitive approach. Neurosci. Biobehav. Rev. 29, 469–491. doi:10.1016/j.neubiorev.2005.01.002

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., 2016. nlme: Linear and nonlinear mixed effects models [WWW Document]. URL http://cran.r-project.org/package=nlme

Posner, L.P., Burns, P., 2009. Veterinary Pharmacology & Therapeutics, 9th editio. ed. Wiley-Blackwell, Ames.

Rousing, T., Wemelsfelder, F., 2006. Qualitative assessment of social behaviour of dairy cows housed in loose housing systems. Appl. Anim. Behav. Sci. 101, 40–53. doi:10.1016/j.applanim.2005.12.009

Rutherford, K.M.D., Donald, R.D., Lawrence, A.B., Wemelsfelder, F., 2012. Qualitative Behavioural Assessment of emotionality in pigs. Appl. Anim. Behav. Sci. 139, 218–224. doi:10.1016/j.applanim.2012.04.004

Stockman, C.A., Collins, T., Barnes, A.L., Miller, D., Wickham, S.L., Beatty, D.T., Blache, D., Wemelsfelder, F., Fleming, P.A., 2011. Qualitative behavioural assessment and quantitative physiological measurement of cattle naive and habituated to road transport. Anim. Prod. Sci. 51, 240–249. doi:10.1071/AN10122

Van Den Bos, R., Meijer, M.K., Van Renselaar, J.P., Van der Harst, J.E., Spruijt, B.M., 2003. Anticipation is differently expressed in rats (Rattus norvegicus) and domestic cats (Felis silvestris catus) in the same Pavlovian conditioning paradigm. Behav. Brain Res. 141, 83–89. doi:10.1016/S0166-4328(02)00318-2

Walker, J., Dale, A., Waran, N., Clarke, N., Farnworth, M., Wemelsfelder, F., 2010. The assessment of emotional expression in dogs using a Free Choice Profiling methodology. Anim. Welf. 19, 75–84.

Walker, J.K., Dale, A.R., BD’Eath, R., Wemelsfelder, F., 2016. Qualitative Behaviour Assessment of dogs in the shelter and home environment and relationship with quantitative behaviour assessment and physiological responses. Appl. Anim. Behav. Sci. 184, 97–108. doi:10.1016/j.applanim.2016.08.012

Welfare Quality®, 2009a. Welfare Quality® assessment protocol for pigs (sows and piglets, growing and finishing pigs). Lelystad.

Welfare Quality®, 2009b. Welfare Quality® assessment protocol for cattle. Lelystad.

Wemelsfelder, F., 2007. How animals communicate quality of life: the qualitative assessment of behaviour. Anim. Welf. 16, 25–31. doi:Article

Wemelsfelder, F., Hunter, T.E.A., Mendl, M.T., Lawrence, A.B., 2001. Assessing the "whole animal": a free choice profiling approach. Anim. Behav. 62, 209–220. doi:10.1006/anbe.2001.1741

Wemelsfelder, F., Hunter, T.E.A., Mendl, M.T., Lawrence, A.B., 2000. The spontaneous qualitative assessment of behavioural expressions in pigs: first explorations of a novel methodology for integrative animal welfare measurement. Appl. Anim. Behav. Sci. 67, 193–215.

Wemelsfelder, F., Hunter, T.E.A., Paul, E.S., Lawrence, A.B., 2012. Assessing pig body language: Agreement and consistency between pig farmers, veterinarians, and animal activists. J. Anim. Sci. 90, 3652–3665. doi:10.2527/jas.2011-4691

Wemelsfelder, F., Nevison, I., Lawrence, A.B., 2009. The effect of perceived environmental background on qualitative assessments of pig behaviour. Anim. Behav. 78, 477–484. doi:10.1016/j.anbehav.2009.06.005

Wickham, S.L., Collins, T., Barnes, A.L., Miller, D.W., Beatty, D.T., Stockman, C.A., Blache, D., Wemelsfelder, F., Fleming, P.A., 2015. Validating the use of qualitative behavioral assessment as a measure of the welfare of sheep during transport. J. Appl. Anim. Welf. Sci. 18, 269–286. doi:10.1080/10888705.2015.1005302

Wickham, S.L., Collins, T., Barnes, A.L., Miller, D.W., Beatty, D.T., Stockman, C.A., Blache, D., Wemelsfelder, F., Fleming, P.A., 2012. Qualitative behavioral assessment of transport-naive and transport-habituated sheep. J. Anim. Sci. 90, 4523–4535. doi:10.2527/jas.2010-3451

Winckler, C., 2015. Qualitative behaviour assessment in animal welfare science. Aktuelle Arbeiten zur artgemäßen Tierhaltung 2015, KTBL-Schrift, 13-25.

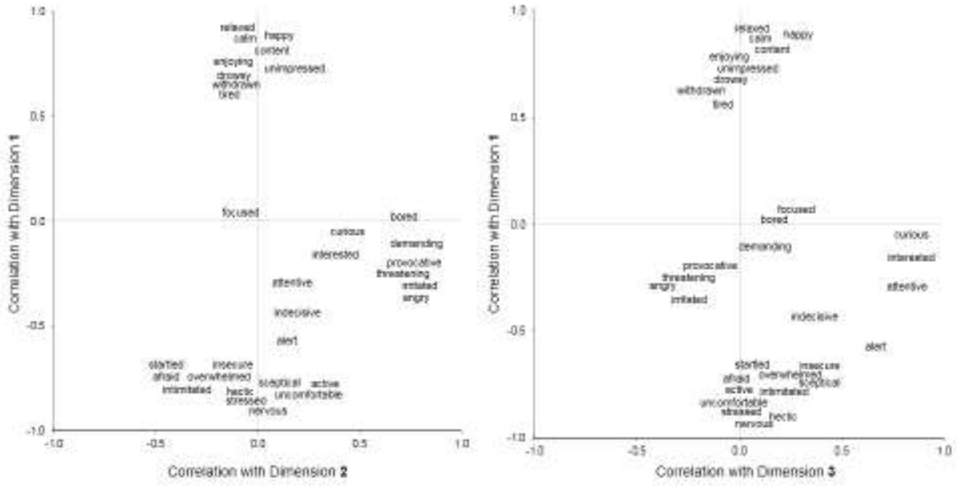
Figure captions

Fig. 1. Example ‘word charts’ of one observer.

The observer’s terms are plotted according to their correlation with Dimension 1 (y-axes) and Dimensions 2 or D3 (x-axes).

Fig. 2. Effect of the five treatments (FA, G, G_{con}, FC, and PB) on the three main QBA dimensions, A: Dimension 1, B: Dimension 2, C: Dimension 3.

The y-axis reflects scaling values for the difference between treatments. Boxplots with medians (black line in box), interquartile range (box) and 1.5 x interquartile range or minimum/maximum value (whiskers) are shown for each treatment.



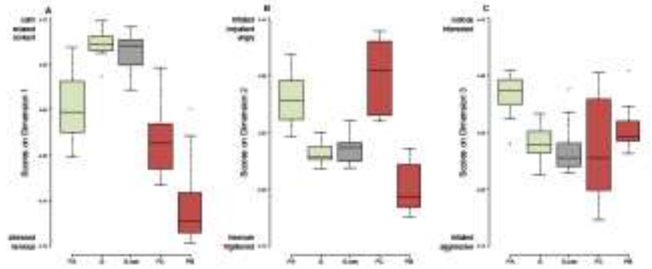

A
calm
relaxed
content
Scores on Dimension 1
B
Scores on Dimension 2
C
Scores on Dimension 3

Table 1
Procrustes Statistics for the joined and separate analyses of treatments

Treatment	**Consensus PS**	**Mean randomised PS ± SD** [1]	**Student's t** (d.f. = 99)
Joined	75.41	26.94 ± 0.22	224.00 *
FA	76.48	61.61 ± 0.33	44.17 *
G	61.72	56.43 ± 0.48	11.58 *
G_{con}	69.55	58.52 ± 0.48	22.99 *
FC	81.63	60.70 ± 0.43	48.20 *
PB	79.91	56.34 ± 0.49	48.18 *

[1] Mean of 100 Procrustes Statistic values obtained through 100 GPAs of randomised data matrices

* P ≤ 0.001

Trt	Dimension (*)	r **	Positive correlation	Negative correlation
Joined	1 (56.5)	$r > 0.8$	**calm** (14), **relaxed** (9), **content** (9)	**nervous** (12), **stressed** (10)
	2 (17.4)	$r > 0.5$	**irritated** (9), **impatient** (7), **angry** (7), aggressive (5)	**frightened** (12), **insecure** (8), pa startled (5)
	3 (9.8)	$r > 0.4$	**curious** (13), **interested** (9), attentive (7)	**aggressive** (10), **irritated** (6), bel angry (5)
FA	1 (43.3)	$r > 0.7$	relaxed (12), content (11), calm (9), drowsy (5), enjoying (4)	nervous (6), stressed (5), impatie
	2 (10.9)	$r > 0.5$	friendly (5), attentive (5) curious (4)	aggressive (6), belligerent (4), irr
	3 (8.5)	$r > 0.4$	no clear terms	focused (4), curious (4)
G	1 (27.0)	$r > 0.7$	relaxed (4), drowsy (4)	curious (4)
	2 (16.4)	$r > 0.4$	enjoying (5), content (4)	irritated (5), nervous (4), indecis
	3 (8.8)	$r > 0.4$	no clear terms	no clear terms
G_{con}	1 (34.4)	$r > 0.6$	relaxed (5)	attentive (8), curious (7), interes
	2 (14.8)	$r > 0.6$	content (5), relaxed (3)	drowsy (4), bored (3)
	3 (11.7)	$r > 0.4$	no clear terms	drowsy (8), enjoying (5), curious
FC	1 (49.4)	$r > 0.8$	friendly (5), relaxed (4), content (4)	irritated (12), angry (5), aggressi (4)
	2 (10.7)	$r > 0.5$	nervous (7), expectant (4)	calm (7)
	3 (8.5)	$r > 0.4$	no clear terms	insecure (7), sceptical (4)
PB	1 (55.2)	$r > 0.9$	relaxed (5), calm (4)	frightened (5), nervous (4)
	2 (8.2)	$r > 0.4$	aggressive (8), irritated (6), belligerent (5)	no clear terms
	3 (6.9)	$r > 0.4$	no clear terms	aggressive (4), irritated (4)

Table 2

Terms used by observers to describe the horses' behavioural expressions across all five treatments (joined analysis) and for each treatment separately (separate analyses)

The presented terms are the ones that were most strongly correlated with each of the three dimensions. Term order is determined by the number of observers using each term (number in brackets). Only terms that were used at a minimum five (joined analysis) or four times (FA, G, G_{con}, FC, and PB) are shown with the terms in bold being used for description of the dimensions of the joined analysis in the text and figures.

* Explained variation in %. **Only the positive correlation coefficients are given here.

Table 3
Post-hoc tests of all contrasts for the three main dimensions

Dimension	Contrast	Test statistic z	P-value [1]
1	FA - G	6.03	< 0.001 ***
	FA - G_{con}	5.34	< 0.001 ***
	FA - FC	-3.64	0.003 **
	FA - PB	-9.83	< 0.001 ***
	G - G_{con}	0.69	0.958
	G - FC	9.67	< 0.001 ***
	G - PB	-15.86	< 0.001 ***
	G_{con} - FC	8.98	< 0.001 ***
	G_{con} - PB	-15.17	< 0.001 ***
	FC - PB	-6.19	< 0.001 ***
2	FA - G	-7.20	< 0.001 ***
	FA - G_{con}	-6.64	< 0.001 ***
	FA - FC	2.99	0.023 *
	FA - PB	-11.64	< 0.001 ***
	G - G_{con}	-0.56	0.981
	G - FC	-10.19	< 0.001 ***
	G - PB	-4.44	< 0.001 ***
	G_{con} - FC	-9.63	< 0.001 ***
	G_{con} - PB	-5.00	< 0.001 ***
	FC - PB	-14.63	< 0.001 ***
3	FA - G	-5.08	< 0.001 ***
	FA - G_{con}	-5.85	< 0.001 ***
	FA - FC	-5.17	< 0.001 ***
	FA - PB	-3.60	0.003 **
	G - G_{con}	0.77	0.940
	G - FC	0.10	1.000
	G - PB	1.48	0.573
	G_{con} - FC	-0.67	0.962
	PB - Gcon	2.25	0.161
	FC - PB	1.58	0.511

[1] Adjusted p-values from single-step method incorporating the correlations between the test statistics (Bretz et al., 2011). * < 0.05, ** < 0.01, *** < 0.001

Table 4
Interquartile ranges (IQRs) per treatment and dimension

Dimension	**FA**	**G**	**G_{con}**	**FC**	**PB**
1	0.06	0.01	0.03	0.05	0.04
2	0.03	0.01	0.02	0.06	0.03
3	0.02	0.02	0.02	0.08	0.02

Numbers show the IQRs of scaling values reflecting differences between treatments.