Applied Animal Behaviour Science 141 (2012) 91–100

Contents lists available at SciVerse ScienceDirect

Applied Animal Behaviour Science

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

Qualitative behaviour assessment of dairy buffaloes (*Bubalus bubalis*)

Fabio Napolitano[a,*], Giuseppe De Rosa[b], Fernando Grasso[b], Françoise Wemelsfelder[c]

[a] *Dipartimento di Scienze delle Produzioni Animali, Università degli Studi della Basilicata, Via dell'Ateneo Lucano 10, 85100 Potenza, Italy*
[b] *Dipartimento di Scienze del Suolo, della Pianta, dell'Ambiente e delle Produzioni animali, Università degli Studi di Napoli "Federico II", Via Università 133, 80055 Portici (NA), Italy*
[c] *Sustainable Livestock Systems, Scottish Agricultural College, Bush Estate, Midlothian, UK*

ARTICLE INFO

Article history:
Accepted 9 August 2012
Available online 6 September 2012

Keywords:
Buffalo
Free choice profiling
Qualitative behaviour assessment
Reliability

ABSTRACT

This study applies qualitative behaviour assessment (QBA) for the first time to dairy buffaloes, using three groups of observers with different cultural backgrounds and different levels of experience in animal behaviour observation and buffalo farming. Eight buffalo heifers aged 16–18 months were subjected to two isolation tests, one performed in the indoor part of their home environment, and one in a novel outdoor paddock. Animals were filmed individually for 2.5 min, and the resulting 16 video clips were shown to three observer panels, consisting of 11 applied animal behaviour scientists from 6 European countries, 11 Italian animal scientists with a background in buffalo farming but no experience in behavioural observation, and 14 Italian undergraduate animal science students with no particular experience. A free choice profiling method was used to instruct observers in QBA, and data for the three panels were analysed separately using Generalised Procrustes Analysis. All three panels showed significant inter-observer agreement ($p < 0.001$) and generated two main consensus dimensions characterised as 'calm-agitated' and 'curious-shy'. There were significant correlations between buffalo scores provided by each of the three observer panels on both these dimensions (*dim1*: Kendall $W = 0.96$, $n = 3$, $\chi^2 = 43.28$, $p < 0.001$; *dim2*: $W = 0.68$, $n = 3$, $\chi^2 = 30.73$, $p < 0.01$). Buffaloes viewed in the familiar indoor pen were assessed by all three panels as more calm and less agitated (dimension 1) than animals viewed in the novel outdoor pen (Wilcoxon $z = -2.52$, $p < 0.01$, $z = -2.52$, $p < 0.01$, $z = -2.38$, $p < 0.01$ for Panels 1, 2, and 3, respectively). Scores on dimension 1 for the same animals viewed in either indoor or outdoor pen were correlated at $r = 0.60$ ($p < 0.10$), 0.74 ($p < 0.05$) and 0.71 ($p < 0.05$) for Panels 1, 2, and 3, respectively. Quantitatively, buffalo in the outdoor pen displayed longer bouts of running and higher frequencies of sniffing (both $p < 0.05$) than those in the indoor pen. Principal component analysis showed meaningful associations between qualitative and quantitative assessments, allowing qualitative dimensions to play a valuable role in interpreting the animals' state. The main outcomes of this study are that QBA can be usefully applied to scientific studies of dairy buffalo, and that substantial differences in observer background do not appear to diminish the reliability of QBA.

1. Introduction

The qualitative assessment of animal behaviour (QBA) is an integrative, 'whole-animal' methodology based on the qualitative interpretation of the dynamic style in which animals interact with their environment. In other words, it describes not 'what' the animals do, but 'how' they do what they do (Stevenson-Hinde, 1983). This method relies on the ability of human observers to integrate perceived details of behaviour and their context into judgements of animal 'body language', using descriptors such as 'calm', 'tense', anxious' or 'content'. Such terms have an

* Corresponding author. Tel.: +39 0971 205078; fax: +39 0971 205099.
E-mail address: fabio.napolitano@unibas.it (F. Napolitano).

0168-1591/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.applanim.2012.08.002

expressive, emotional connotation, and provide information that appears relevant to animal welfare and could be a useful addition to information obtained from quantitative indicators (Wemelsfelder, 1997; Wemelsfelder et al., 2001; Rutherford et al., 2012). QBA has so far been applied to farm animal species such as pigs (Wemelsfelder et al., 2001, 2009a) and cattle (Rousing and Wemelsfelder, 2006), and companion animals such as horses (Napolitano et al., 2008) and kennelled dogs (Walker et al., 2010). This study reports the first application of QBA to Mediterranean buffaloes, a dairy animal recently moved from traditional farming techniques based on the extensive use of humid environments to intensive systems similar to those applied to dairy cattle (De Rosa et al., 2009). Intensification of farming techniques has subjected these animals to environmental challenges so far unknown to this species and potentially impairing their welfare, so there is a need to develop welfare assessment tools suited to address these problems in buffalo.

Qualitative methodologies have in the past been criticised for being based on subjective and unreliable evaluations, however recently their validity has gained renewed interest and discussion (Meagher, 2009; Whitham and Wielebnowski, 2009). Previous QBA studies have shown good intra- and inter-observer reliability (e.g. Wemelsfelder et al., 2001, 2009a,b; Rousing and Wemelsfelder, 2006; Walker et al., 2010), and have supported the validity of QBA in terms of its correlation with ethogram-based behaviour assessments (Napolitano et al., 2008; Minero et al., 2009) and indicators of physiological stress (Stockman et al., 2011). Most recently Rutherford et al. (2012) demonstrated QBA to be highly sensitive (in a blind trial) to whether growing pigs observed in different test situations had been treated either with anti-anxiety drug azaperone or with neutral saline solution. Generally these studies support that the assessment of animal demeanour through QBA can add a valuable layer of expressive information to animal studies, identifying differences in emotional valence that can be difficult to capture quantitatively. Questions that are still to be investigated, however, are for example whether and how different cultural backgrounds in observers, and different levels of experience with animal behaviour observation and animal farming, affect the reliability of QBA. No information is available as yet on these aspects in buffalo or other animal species.

Thus the aim of this study was to compare the qualitative behaviour assessments of buffalo provided by three groups of observers with different nationalities and different levels of experience in animal behaviour observation and buffalo farming. As in most previous QBA studies, a free-choice profiling (FCP) methodology was used to this end. FCP was originally developed in food science (Arnold and Williams, 1985; Oreskovich et al., 1991), and adjusted for use in animal science by Wemelsfelder et al. (2001). It is characteristic of FCP that it asks observers to develop their own descriptors based on direct observation of animals, a feature we consider essential for being able to test whether or not observers perceive animal body language expressions in similar ways. If pre-determined lists of descriptors were given to observers, the actual process of qualitatively interpreting the animals' expressions would remain untested, and any found agreement would only concern the quantitative use of terms. Thus we consider FCP to be a more powerful tool for testing the reliability of QBA. The association of QBA data with a range of quantitative behaviour variables was also investigated using principal component analysis.

2. Materials and methods

2.1. Experimental procedures

Animal subjects were 8 Mediterranean buffalo heifers aged 16–18 months. They were group housed in a 5.0 m × 4.6 m indoor slatted floor pen with free access to an outdoor earth floor 5 m × 4.6 m paddock. These animals had previously been kept in intensive farming conditions, where they were subjected to artificial rearing, early weaning and close confinement. These conditions made the animals accustomed to farming practices involving human presence and handling.

In order to test observer agreement in judging buffalo expressions, it was necessary to show observers a variety of different buffalo expressions. To achieve this, buffaloes were subjected to two different isolation tests. One consisted in isolating individual animals from the rest of the group in the indoor part of the home pen, and one in leading animals individually through a single-file chute to an outdoor paddock (5.0 m × 4.6 m, with earth floor and open metal fencing), which was novel to the animals and located approx. 20 m away from the home pen. Four animals were tested first in the home indoor pen and subsequently in the novel outdoor paddock; the other four were tested in the opposite order. During tests subjects were isolated from tactile and visual contact with conspecifics, but could receive auditory and olfactory stimuli from other animals; they could not receive any stimuli from humans. All testing procedures were performed by a stockperson well-known to the animals.

Each animal was confined individually for 2.5 min in each testing condition, and her behaviour during this time was video-recorded using a DVL-157 JVC video camera equipped with a wide-angle lens, located at a corner of the test area at a distance from the fence of 6 m and operated by remote control. From this material a video tape was created containing 16 clips (8 animals in two conditions) of 2.5 min duration each, showing indoor and outdoor tested buffaloes in random order. To give observers time for recording their assessments, each video clip was followed by a blank frame lasting 1.5 min, which was then followed by the next video clip. Thus the total duration of the video recording presented to observers was 64 min.

2.2. Behaviour assessment

2.2.1. Quantitative assessment

The behaviour shown by buffalo heifers in the 16 video clips was analysed quantitatively by means of continuous recording technique (accurate to 1 s). Observations were performed by one trained observer. Training consisted in the observation of 3 outdoor clips and 3 indoor clips with the aim to instruct the observer in recognizing the

Table 1
Description of behavioural categories recorded during the isolation tests.

Behavioural category	Description
Latency to first movement (duration)	Time lapse between entrance of the animal in the pen and first movement of the animal.
Run (duration)	Rapid forward movement including gallop and trotting.
Flight attempts (frequency)	Fast run towards the fence abruptly interrupted either before or after crashing into it. Final posture with head protruding beyond the fence.
Vocalisation (frequency)	Emission of acoustic signals.
Sniffing (frequency)	Sniffing the ground or the fence.

behavioural categories identified in a previous study (Napolitano et al., 2004). The behaviours recorded are described briefly in Table 1.

2.2.2. Qualitative assessment

For the benefit of this study three groups of observers with different nationalities and cultural and experiential backgrounds were recruited. These were: (1) eleven applied animal behaviour scientists (8 female and 3 male) from 6 different European countries with substantial experience in the assessment of farm animal behaviour (Panel 1), who observed the buffalo video in Edinburgh (Scotland); (2) eleven Italian animal scientists (5 female and 6 male) with a solid background in buffalo farming but no education or experience in assessing animal behaviour (Panel 2), who observed the video in Portici (Naples, Italy); and (3) fourteen Italian undergraduate students in animal science (7 female and 7 male) with no experience in either the observation of animal behaviour or in buffalo farming (Panel 3), who observed the video in Potenza (Italy). These observers all volunteered to participate in the study, and none had previously taken part in any QBA assessment. Observations were conducted in blind: observers received no information on either the animals (i.e. they did not know each animal was tested twice) or on the experimental design (i.e. they did not know the indoor paddock was familiar and the outdoor paddock unknown to the animals), although they could see from the video that two different testing locations were used.

Each group of observers was instructed in free choice profiling procedures by a different instructor (who were all scientists in the field of applied animal behaviour), following Wemelsfelder et al. (2001). These procedures consisted of two phases. In phase 1 observers generated their own descriptive vocabularies by watching the 16 buffalo video clips (of 2.5 min each) and by writing down, during a 1.5 min blank period after each clip, the terms that in their view adequately summed up the animal's style of responding to the test situation. No limit was imposed on observers with regard to the number of terms to be generated, but this number never exceeded forty. In phase 2 observers were instructed to use their personal vocabularies to quantitatively score the responsiveness styles of the same animals. They watched the same videos again in the same order, and during the 1.5 min period after each clip scored each animal on each term of their vocabulary, using visual analogue scales of 125 mm length (0 mm: attribute absent, 125 mm: attribute could not be stronger). Scores for each attribute were measured as the distance in millimetres from the 0-point. For further details of this method see Wemelsfelder et al. (2001).

2.3. Statistical analyses

2.3.1. Analysis of quantitative data

The durations and, in case of incidental behaviours, the frequencies of the recorded quantitative behavioural variables were calculated. A non-parametric Wilcoxon test was used to investigate whether significant differences existed between buffalo behaviour assessed indoors and outdoors.

2.3.2. Generalised Procrustes Analysis

As each panel was instructed by different instructors in different languages, times and locations, their data were analysed separately. For each panel the FCP procedures produced one data matrix for each observer, providing scores for the animals in the 16 clips (8 indoors and 8 outdoors) on the basis of the observers' personal vocabularies. For each panel the concordance between these matrices was analysed using Generalised Procrustes Analysis (GPA), a multivariate statistical technique which does not require fixed variables for its calculation of a consensus. This calculation is essentially a process of complex pattern recognition; the observer matrices are represented in virtual space as multi-dimensional configurations, with the number of dimensions for each configuration determined by the number of terms generated by a particular observer. The 16 clips observed by all observers are placed at different positions in the different observer configurations. The concordance between these configurations is calculated through a complex process of rotation, transformation, aimed at finding a 'best-fit' consensus profile. GPA provides a statistic (called the Procrustes Statistic) which indicates the level of consensus (i.e. the percentage of variation explained between observers) that has been achieved. Whether this consensus is a significant feature of the data set, or, alternatively, an artefact of the Procrustean calculation procedures, is determined through a permutation or randomisation test (Dijksterhuis and Heiser, 1995). This procedure rearranges at random each observer's scores and produces new permutated data matrices. By applying GPA to these permutated matrices, a 'randomised' profile is calculated. This procedure is repeated 100 times, providing a distribution of Procrustes Statistics indicating how likely it is to find an observer consensus based on chance alone. Subsequently a one-way *t*-test is used to determine whether the actual observer consensus profile falls significantly outside the distribution of randomised profiles.

The calculation of the consensus profile takes place independently of the semantic information provided by the terminologies chosen by the observers. Semantic interpretation of this consensus profile takes place after its calculation. Through a principal component analysis (PCA), the number of dimensions of the consensus profile is reduced to one or more main dimensions explaining the majority of variation between the observed animals. These dimensions are subsequently interpreted by correlating them to the original individual observer data matrices. This

step of the analysis produces two-dimensional individual observer interpretative word-charts. In each chart, all terms of a particular observer are correlated with the principal axes of the consensus profile. These observer word charts can be used for the interpretation of the main dimensions, in that the higher a term correlates with an axis, the more weight it has as a descriptor for that axis.

Each clip receives a score on each of these main consensus dimensions. For each panel a non-parametric Wilcoxon test was used to investigate whether the scores obtained for indoor and outdoor animals within the same dataset differed significantly (assessment of differences within animals between the two different environments), whereas the relationship between indoor and outdoor animal scores was determined using the Spearman correlation coefficient (r_s) in order to assess the predisposition of the animals to respond similarly in different environments.

When GPA is performed on groups of animals characterised by disparities in terms of breed, treatment or experience, the analysis tends to show such diversities, and more subtle differences between individuals may be more difficult to discern or disappear altogether (Napolitano et al., 2008). Therefore, in order to take a closer look at differences among individual animals independently from the treatment (location and novelty of test), assessments of indoor and outdoor clips were also analysed separately. For sake of brevity only data provided by Panel 1 (applied animal behaviour scientists) were included in this analysis.

For a more detailed description of GPA calculation and interpretation procedures see Wemelsfelder et al. (2000). All GPA analyses in this study were performed using the programme Senstools.NET v1.x.

2.3.3. The relationship between quantitative and qualitative data

A principal component analysis (PCA) was used to investigate the relationship between quantitative and qualitative data, and to indicate which variables were most closely associated in explaining the variation between clips (Cooper and Weekes, 1983). The PCA was based on the scores of the animals on the first two dimensions of the GPA consensus profile of Panel 1, and on quantitative behaviour data recorded during the two isolation tests. A correlation matrix was used and no rotation was applied. The analysis was performed using the software Unscrambler X v.10.1.

2.3.4. Inter-panel reliability

The degree of agreement between data gathered from the three panels (i.e. the scores of individual animals on the first two components of each panel's consensus profile) was calculated using Kendall's coefficient of concordance (W). In addition, Spearman rank correlation coefficients (r_s) were used to determine the correlation between pairs of panels.

3. Results

3.1. Quantitative behaviour assessment

Table 2 shows the mean durations and frequencies of buffalo heifer behaviour expressed indoors and outdoors.

Table 2
The mean durations in seconds (s) and, for incidental behaviours, the frequencies (f), of behaviours recorded indoors and outdoors.

Behavioural category	Indoors	Outdoors	*P*-Value
Latency to first movement (s)	15.5 ± 3.3	7.0 ± 3.3	0.09
Run (s)	0.0 ± 1.3	3.7 ± 1.3	0.05
Flight attempts (f)	1.2 ± 1.1	3.9 ± 1.1	0.09
Vocalisation (f)	6.1 ± 2.0	7.7 ± 2.0	0.58
Sniffing (f)	4.1 ± 1.0	7.9 ± 1.0	0.02

In the novel outdoor pen the animals exhibited longer bouts of running and greater frequency of sniffing (Wilcoxon signed ranks tests: $Z = -2.84$ and -1.96, respectively, $N = 8$, $p < 0.05$), and tended to show a shorter latency time to first movement and a higher frequency of flight attempts (Wilcoxon signed ranks tests: $Z = 1.68$ and -1.85, respectively, $p < 0.10$) than in the familiar indoor pen.

3.2. Qualitative behaviour assessment and panel reliability

The consensus profiles of the three observer panels explained a high percentage of variation among the observers, and differed significantly from the mean randomised profile (Procrustes Statistic: 77.3%, 77.1% and 77.9% for Panels 1, 2 and 3, respectively; $p < 0.001$).

Two main dimensions of the consensus profiles were identified, explaining 47.0% and 11.5%, 55.2% and 7.6% and 60.3% and 7.7% of the variation between animals for Panels 1, 2 and 3, respectively. Within and between panels the observer word charts interpreting these dimensions were semantically consistent, as they all converged towards similar meanings, albeit using a range of different terms. As it was impossible to show the word charts of all observers, that of observer 1 from Panel 1 was chosen as representative (Fig. 1). The chart characterised the first dimension of the consensus profile with terms ranging from calm to restless/unsettled, whereas the second dimension was described as ranging from curious/confident to shy. These terms displayed correlation coefficients higher than 0.3 with the axes of the consensus profile. To provide an overview of highly correlated terms for all observers, Table 3 lists for each observer the two terms with the highest positive and negative correlation with dimensions 1 and 2. This table shows that the most frequently used descriptors for the positive ends of axes 1 and 2 were 'calm' (used by 2–7 observers per panel) and 'investigatory' (used by at least 1 observer per panel) along with 'explorative' (used by 3 observers from Panel 1) and 'curious' (used by 4 observers from Panel 3), respectively. At least three observers per panel used the term 'agitated' to describe the negative end of axis 1, whereas 'scared', 'shy' (both used by at least 1 observer per panel), and 'frightened' (used by 4 observers from Panel 3) were the most used negative descriptors of axis 2. Thus, axis 1 was labelled as 'calm-agitated' and axis 2 was named as 'curious-shy'.

Fig. 2 shows individual animals as positioned by Panels 1, 2 and 3 on these two consensus dimensions. The animal plots were characterised by a reasonably homogeneous distribution of the individual subjects, thus indicating that

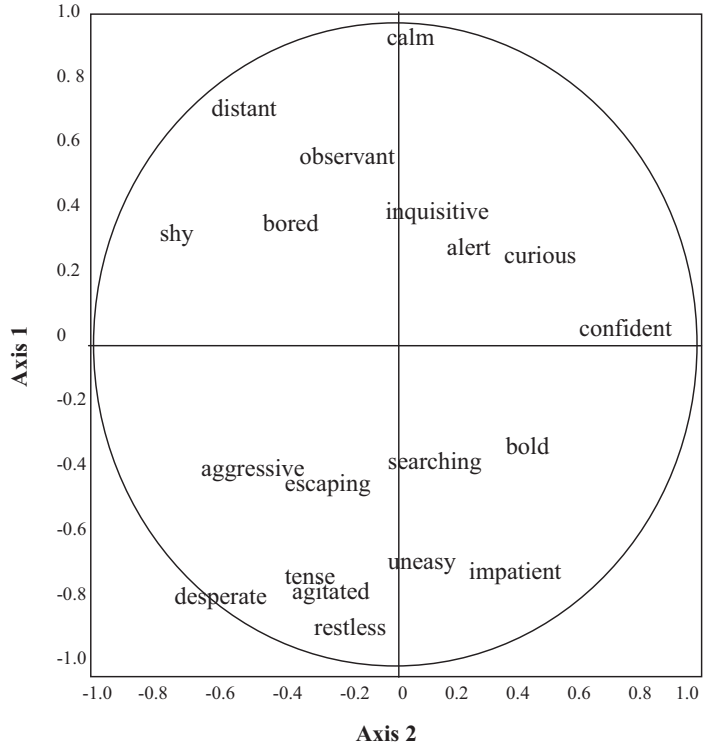

Fig. 1. Word chart of observer 1 from Panel 1. Axes reflect a term's strength of correlation with consensus dimensions 1 and 2.

the two dimensions provided a good resolution of animal behaviour. The scores of individual animals on the first two components of GPA were used to assess inter-panel reliability. The correlation among the three panels was significant for the first and the second dimension (Kendall coefficient of concordance: $W = 0.96$, $\chi^2 = 43.28$, d.f. = 15, $p < 0.001$ and $W = 0.68$, $\chi^2 = 30.73$, d.f. = 15, $p < 0.01$, respectively). When calculated separately, correlation coefficients between Panel 1 and Panel 2 (Spearman $r_s = 0.97$, $p < 0.001$, and $r_s = 0.68$, $p < 0.01$, $N = 16$, for dimensions 1 and 2, respectively) were higher than those between Panel 3 and Panel 1 ($r_s = 0.94$, $p < 0.001$, and $r_s = 0.63$, $p < 0.01$, $N = 16$, for dimensions 1 and 2, respectively,) and between Panel 3 and Panel 2 ($r_s = 0.94$, $p < 0.001$, and $r_s = 0.43$, $p < 0.10$, $N = 16$, for dimensions 1 and 2, respectively).

A marked effect of test location/novelty was observed on the scores of animals on the first dimension (Wilcoxon signed ranks tests: $Z = -2.52$, $N = 8$, $p < 0.01$, $Z = -2.52$, $N = 8$, $p < 0.01$, $Z = -2.38$, $N = 8$, $p < 0.01$ for Panels 1, 2 and 3, respectively), but no effects were observed on the second dimension. Buffalo heifers received significantly higher scores, and were thus assessed as more calm and less agitated, when tested indoors.

For the first dimension high correlation coefficients were found between the scores of the same animals when tested either indoors or outdoors (Spearman rank correlation: $r_s = 0.60$, $N = 8$, $p < 0.10$, $r_s = 0.74$, $N = 8$, $p < 0.05$ and $r_s = 0.71$, $N = 8$, $p < 0.05$ for Panels 1, 2 and 3, respectively), whereas no significant correlations were observed for the second dimension.

Table 3
Terms (one for each observer) showing the highest positive and negative correlation with dimensions 1 and 2 of the consensus profiles for each of three panels.

Dimension	Positive correlation	Negative correlation
Panel 1		
1	Calm (7), docile (1), slow-moving (1), passive (1), tranquil (1)	Active (4), agitated (3), restless (2), fidgety (1), unsettled (1)
2	Explorative (3), confident (3), investigatory (1), searching (1), careful (1), interested (1), self confident (1)	Shy (2), timid (2), scared (1), desperate (1), jumpy (1), rushing (1), flighty (1), challenging (1), assessing (1)
Panel 2		
1	Calm (5), timorous (1), indifferent (1), waiting (1), perplexed (1), tranquil (1)	Agitated (3), restless (2), nervous (2), preoccupied (1), content (1), bold (1)
2	Investigatory (2), self-confident (2), bold (2), nervous (2), active (1), sociable (1)	Shy (2), scared (2), suspicious (2), docile (2), escaping (1), exhausted (1)
Panel 3		
1	Tranquil (5), relaxed (4), calm (2), quiet (1), resigned (1)	Agitated (5), nervous (2), fidgety (2), escaping (1), restless (1), annoyed (1), tense (1)
2	Curious (4), investigatory (3), sociable (1), interested (1), nervous (1), sad (1), tense (1)	Frightened (4), scared (2), lonely (2), escaping (2), flighty (1), shy (1), assessing (1)

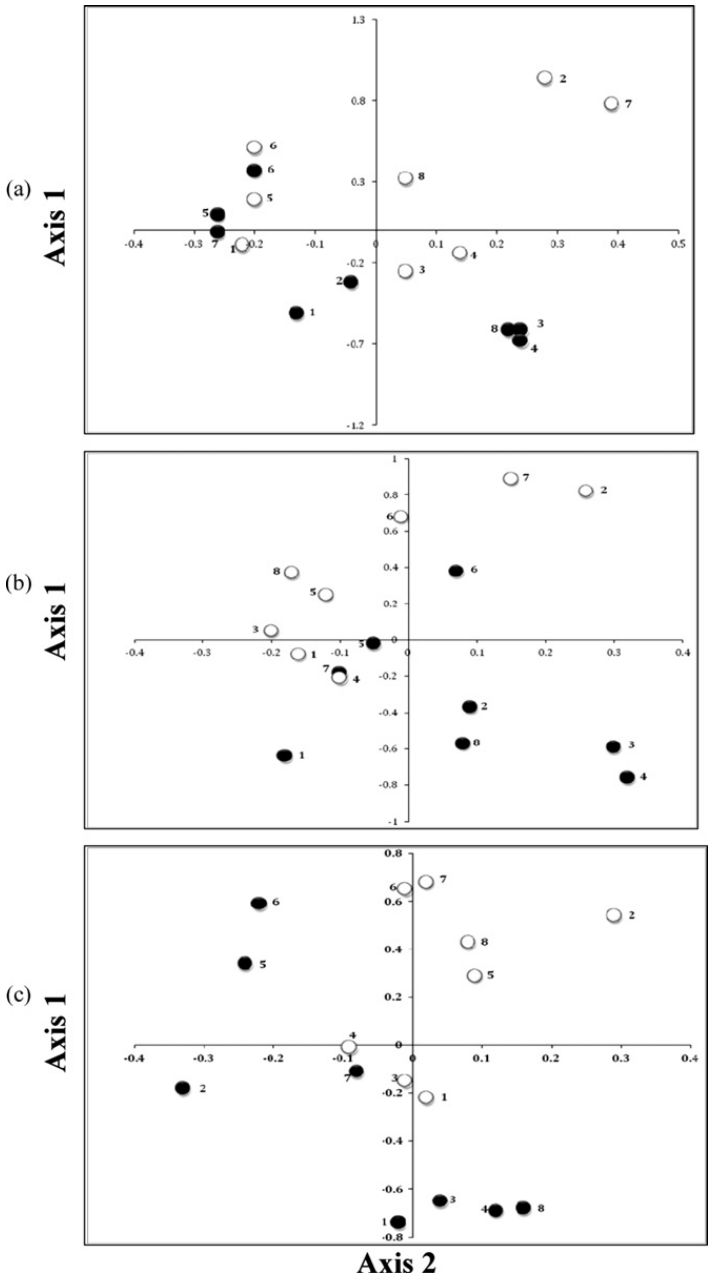

Fig. 2. Position of individual buffalo on the two main consensus dimensions, as scored by Panels 1 (a), 2 (b) and 3 (c) for familiar indoor (◯) and novel outdoor (●) test situations. Axes reflect GPA scaling values for relative sample distance.

The GPA performed on indoor and outdoor animals separately also yielded good consensus among observers (Panel 1 only), and differed significantly from the mean randomised profiles (Procrustes Statistic: 77.7% and 78.9% for indoor and outdoor animals, respectively; $p < 0.001$). The assessors attributed to the first two dimensions of the consensus profile descriptors that were similar to those previously described for all animals, and dimensions 1 and 2 were therefore again labelled as 'calm-agitated' and 'curious-shy', respectively, for both the indoor and outdoor analyses. For the indoor analysis the first two dimensions explained 53.9 and 13.8% of the total variation, respectively, whereas for the outdoor analysis the total variation explained by dimensions 1 and 2 was 49.2% and 15.4%, respectively. Fig. 3 shows the positions of animals on the main consensus dimensions of the separate indoor and outdoor analyses. The Spearman rank correlation between animals indoors and outdoors was 0.64 ($N = 8$, $p < 0.10$) on dimension 1; there was no significant correlation for dimension 2.

3.3. The relationship between qualitative and quantitative assessments

Principal component analysis of qualitative assessments (Panel 1 only) and quantitative data showed two main components explaining 41% and 26% of the variation. Fig. 4 shows the loadings of the different variables on these two components. QBA dimension 1 (calm-agitated) and 'latency to first movement' showed the highest positive

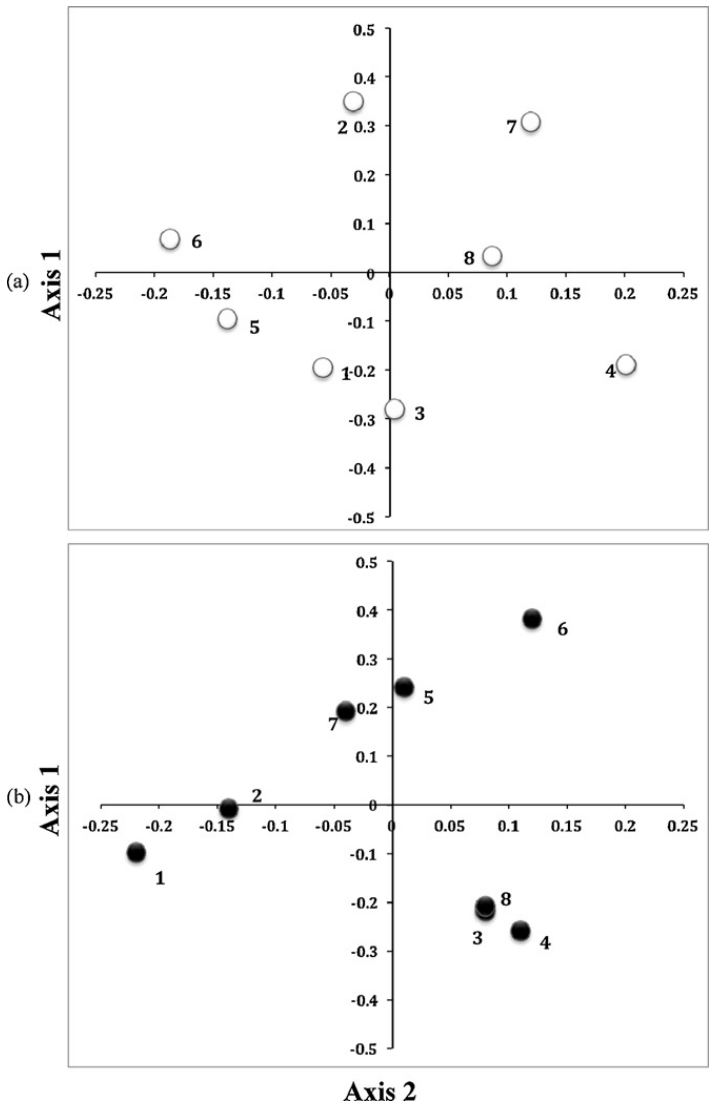

Fig. 3. Position of individual buffalo on the two main consensus dimensions of the 'home indoor test' analysis (a) and the 'novel-outdoor test' analysis (b), as scored by Panel 1. Axes reflect GPA scaling values for relative sample distance.

loadings on the first PCA component (0.54), whereas flight attempts (−0.65) and duration of running (−0.59) displayed the highest negative loadings on this component. The highest positive loadings on the second PCA component were those of QBA dimension 2 (0.65) and vocalisation (0.44), whereas the highest negative loadings were again shown by running (−0.49) and flight attempts (−0.30).

4. Discussion

A first main result of the present study is that, using FCP methodology, high levels of inter observer agreement in qualitative assessments of dairy buffalo behaviour were found within three different observer groups varying in cultural and experiential backgrounds. This is in line with results found for other species in previous QBA studies (e.g. Rousing and Wemelsfelder, 2006; Walker et al., 2010), and generally supports the reliability of QBA. A second main result is that high agreement was also found *between* the different observer groups, in that the rankings of individual buffaloes on the two main consensus dimensions of separate panel analyses were found to be highly correlated. Thus, notwithstanding their differences in nationality, cultural background, and experience with animal behaviour observation and buffalo farming, observers developed similar vocabularies to describe buffalo demeanour, and used these for scoring buffalo in similar ways. This suggests that QBA reliability may be robust even when observers' backgrounds differ substantially in various ways. However more investigation of this question is needed, particularly with observers assessing animals in more complex and variable environments which would require more complex judgments.

That the correlation between the two experienced panels with backgrounds in animal behavioural observation and buffalo farming was highest, suggests that experience and training may be important determinants of QBA reliability (Wemelsfelder et al., 2009a). An interesting

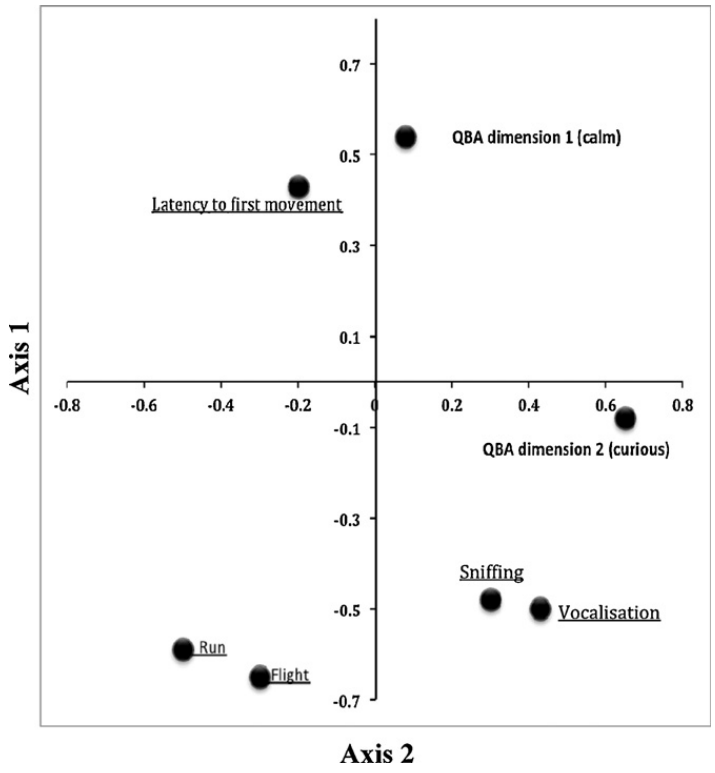

Fig. 4. Loadings of qualitative behaviour dimensions (bold) and quantitative behavioural variables (underlined) on the two main components of a PCA based on data provided by Panel 1 for buffalo in two test environments.

question is how the qualitative QBA assessment process might be embodied physically in observers' perceptive systems. Recent research in humans and some non-human animal species such as birds and monkeys, has indicated that activation of 'mirror neurons' plays a role in the recognition of emotion between individuals (e.g. Keller and Hahnloser, 2009; Rizzolatti and Sinigaglia, 2010; Keysers et al., 2010). It is conceivable that such activation may also be involved in the recognition of emotion between human and animal individuals, however this hypothesis has not yet been investigated but would be interesting to pursue.

Because observers in this study all used different descriptors, calculation of agreement relied on comparison of scoring patterns generated by observers' entire vocabularies, not comparison of scores generated by specific terms. It is possible that if terms were imposed on observers through pre-determined scoring lists, agreement would not be as high as found in this study, or would be high for some terms but not others. Fixed-term lists are more suitable than FCP for welfare monitoring purposes, when a standardised way of assessment is needed for feasibility reasons (Wemelsfelder et al., 2009b). It is important in this case to give observers the opportunity to consider and/or discuss the meaning of each term, and regular tests of observers' agreement on individual terms would allow elimination of terms failing to reach good agreement. FCP also provides information on how observers use particular terms, in that comparison of the semantic structure and tone of word charts gives an idea of how well observers' use of terms converges, but this information is not as quantitatively precise as that provided by fixed lists. Clearly FCP and the use of fixed lists each have advantages and disadvantages in how they support application of QBA, and should be used in situations to which they are best suited.

Given the high levels of observer agreement, it is of interest that the three panels attributed different mean scores to buffaloes tested in a familiar indoor environment or an unfamiliar outdoor pen. Animals tested indoors were on average assessed as more calm and docile, and less agitated and restless, than animals tested in the outdoor pen. A previous study with pigs indicated that the background against which animals are viewed can have a small effect on observers' scores of these animals, but is not likely to significantly alter their overall characterisation (Wemelsfelder et al., 2009a). Thus, although the different backgrounds against which buffalos were viewed may have affected observer assessments, it is certainly feasible that the perceived differences in demeanour shown by the buffaloes in the two test environments were genuinely present. In addition to noting differences in the buffaloes' average response levels, observers also noted a consistency in how buffaloes responded individually to the two test environments, as indicated by the significant correlation between indoor and outdoor scores for individual buffaloes on dimension one. Thus it appears that QBA may be sufficiently sensitive to detect temperamental predispositions in individual animals, over and above mean differences in these animals' responsiveness levels.

The behavioural response of young ruminants to open field testing is generally viewed as driven by a mix

ofmotivations, and it can be difficult to discern exactly which are operative at any given time (Rushen, 2000; Forkman et al., 2007). In the present study multivariate analysis through principal component analysis showed a meaningful association between qualitative and quantitative assessments, suggesting that QBA could potentially add valuable information to assist the interpretation of quantitative data gathered in open fields. For instance, the association between 'calm' and 'latency to first movement' on Component 1 suggests that the latter measure does not in this case reflect fearful freezing, but rather a more tranquil, relaxed response to the test situation compared to animals running around trying to escape. Similarly, the association between 'curious' and 'vocalisation' on Component 2 suggests vocalising in this study is associated more with exploration than fear. This coherence of QBA with quantitative measures is in line with results reported by Rutherford et al. (2012) for pigs, which showed QBA to play an important role in judging the emotional valence of the animals' physical activity in open field and elevated plus maze tests.

Finally, we note that the separate analyses of in- and outdoor buffalo scores for Panel 1 data reproduced the characterisation of the two main consensus dimensions as 'calm-agitated' and 'curious-shy', indicating the robustness of these dimensions in describing the animals' response to the two open field tests. Given this robustness it is not surprising that the position of individual buffaloes on dimension 1 of both separate analyses is virtually identical to their position on dimension 1 in the combined analysis. Thus in the present study it seems that the significant treatment effect of test location/novelty did not, as hypothesised, mask more subtle within-test individual differences on this dimension. However there are some changes in the positions of individual buffalo on dimension 2 when in- and outdoor scores are analysed separately, even though there was no significant effect of test on this dimension in the combined analysis. Whether or not such changes occur may depend on the strength of a dimension, the particular characteristic this dimension captures, or any other trait. Careful thought should thus be given to size and composition of animal samples in QBA testing, in relation to the questions that are asked.

5. Conclusions

In conclusion, the positive results reported here in applying QBA to dairy buffalo heifers in familiar and novel testing environments add to the growing literature supporting the reliability of QBA. A novel contribution of the present study is that reliability was achieved even though observers differed in cultural background, and in their experience with buffalo farming and animal behavioural observation. Given this consistency in assessments, QBA was able to provide expressive information helpful to the interpretation of quantitative behaviour variables. It described the effects of two open field tests on the animals' state, and also uncovered a consistency in how individual animals responded to these tests. Together these results support the guiding hypothesis that QBA, rather than consisting of unfounded projections of human emotion, is empirically grounded in the observation of expressive behavioural criteria.

Acknowledgments

The authors would like to thank A. M. Riviezzi and G. Migliori for expert technical assistance. Thanks are also due to the members of the three panels participating in this study.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.applanim.2012.08.002.

References

Arnold, G.M., Williams, A.A., 1985. The use of generalized procrustes techniques in sensory analysis. In: Piggott, J.R. (Ed.), Statistical Procedures in Food Research. Elsevier Applied Science, London, pp. 233–253.
Cooper, R.A., Weekes, A.J., 1983. Data, Models and Statistical Analysis. Barnes & Noble Books, Totowa.
De Rosa, G., Grasso, F., Braghieri, A., Bilancione, A., Di Francia, A., Napolitano, F., 2009. Behavior and milk production of buffalo cows as affected by housing system. J. Dairy Sci. 92, 907–912.
Dijksterhuis, G.B., Heiser, W.J., 1995. The role of permutation test in exploratory multivariate data analysis. Food Qual. Pref. 6, 263–270.
Forkman, B., Boissy, A., Meunier-Salaün, M.-C., Canali, E., Jones, R.B., 2007. A critical review of fear tests used on cattle, pigs, sheep, poultry and horses. Physiol. Behav. 92, 340–374.
Keller, G.B., Hahnloser, R.H., 2009. Neural processing of auditory feedback during vocal practice in a songbird. Nature 457, 187–190.
Keysers, C., Kaas, J.H., Gazzola, V., 2010. Somatosensation in social perception. Nat. Rev. Neurosci. 11, 417–428.
Meagher, R.K., 2009. Validity and values as a tool for animal welfare research. Appl. Anim. Behav. Sci. 119, 1–14.
Minero, M., Tosi, M.A., Canali, E., Wemelsfelder, F., 2009. Quantitative and qualitative assessment of the response of foals to the presence o fan unfamiliar human. Appl. Anim. Behav. Sci. 116, 74–81.
Napolitano, F., De Rosa, G., Grasso, F., Pacelli, C., Bordi, A., 2004. Influence of space allowance on the welfare of weaned buffalo (*Bubalus bubalis*). Liv. Prod. Sci. 86, 117–124.
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.
Oreskovich, D.C., Klein, B.P., Sutherland, J.W., 1991. Procrustes analysis and its applications to free-choice and other sensory profiling. In: Lawless, H.T., Klein, B.P. (Eds.), Sensory Science: Theory and Applications in Foods. Marcel Dekker, New York, pp. 353–393.
Rizzolatti, G., Sinigaglia, C., 2010. The functional role of the parieto-frontal mirror circuit: interpretations and misinterpretations. Nat. Rev. Neurosci. 11, 264–274.
Rousing, T., Wemelsfelder, F., 2006. Social behaviour as a welfare indicator in loose housing system for dairy cows: a qualitative approach. Appl. Anim. Behav. Sci. 101, 40–53.
Rushen, J., 2000. Some issue in the interpretation of behavioural responses to stress. In: Moberg, G.P., Mench, J.A. (Eds.), The Biology of Animal Stress—Basic Principles and Implications for Animal Welfare. CABI Publishing, Wallingford, pp. 23–42.
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.
Stevenson-Hinde, J., 1983. Individual characteristics: a statement of the problem. In: Hinde, R.A. (Ed.), Primate Social Relationships: An Integrated Approach. Blackwell Scientific Publications, Oxford, pp. 28–34.
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.

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.

Wemelsfelder, F., 1997. The scientific validity of subjective concepts in models of animal welfare. Appl. Anim. Behav. Sci. 5, 75–88.

Wemelsfelder, F., Hunter, E.A., Mendl, M.T., Lawrence, A.B., 2000. The spontaneous qualitative assessment of behavioural expression 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., Mendl, M.T., Lawrence, A.B., 2001. Assessing the whole animal: a free choice profiling approach. Anim. Behav. 62, 209–220.

Wemelsfelder, F., Nevison, I., Lawrence, A.B., 2009a. The effect of perceived environmental background on qualitative assessment of pig behaviour. Anim. Behav. 78, 477–484.

Wemelsfelder, F., Millard, F., De Rosa, G., Napolitano, F., 2009b. Qualitative behaviour assessment. In: Forkman, B., Keeling, L. (Eds.), Welfare Quality® Report No. 11—Assessment of Animal Welfare Measures for Dairy Cattle, Beef Bulls and Veal Calves. Cardiff University, Cardiff, pp. 215–224.

Whitham, J.C., Wielebnowski, N., 2009. Animal-based welfare monitoring: using keeper ratings as an assessment tool. Zoo. Biol. 28, 545–560.