Validation of Qualitative Behaviour Assessment for Dairy Cows at Pasture

Lydiane Aubé, Marie-Madeleine Mialon, Isabelle Veissier, Alice de Boyer Des Roches

Validation of qualitative behaviour assessment for dairy cows at pasture

L. Aubé[ab,1], M.M. Mialon[a], I. Veissier[a], A. de Boyer des Roches[ab]

[a]Université Clermont Auvergne, INRAE, VetAgro Sup, UMR Herbivores, 63122 Saint-Genès-Champanelle, France

[b]Chaire Bien-Être Animal, VetAgro Sup, 1 avenue Bourgelat, 69280 Marcy-l'Etoile, France

Corresponding author: isabelle.veissier@inrae.fr

Abstract

Qualitative Behaviour Assessment (**QBA**) has been validated as a measure of emotional state in dairy cows kept indoors. We investigated the validity of QBA as a measure of emotional state of cows at pasture. We focused on intra- and inter-observer reliability and the selectivity of QBA. First, 5 observers watched 20 videos of dairy cows at pasture 4 times (resulting in 400 observations). The observers performed QBA using the fixed list of 20 terms proposed in Welfare Quality. The first axis of the principal component analysis performed on these observations represented emotional valence (**PC1**, from irritable to content) and a second axis represented arousal (**PC2**, from calm to active). Kendall's concordance coefficients (**KW**) and interclass correlation coefficients (**ICC**) within and between observers were calculated for PC1 and PC2. Intra-observer reliability was mainly high (KW > 0.75 and ICC > 0.75), except for 3 observers during the first session for PC1, for whom reliability was moderate. Inter-observer reliability was low to moderate for PC1 and moderate to high for PC2. Second, two observers performed live QBA on a herd of cows at pasture during 3 contexts designed to induce different emotional valences and arousal levels: **AM**, in the morning after milking (when cows are active); **PM**, in the afternoon before milking (when cows are mainly resting); and during handling to collect cows at pasture for afternoon milking (when cows may be stressed). The effect of context on PC1 and PC2 was investigated using linear mixed effects models. AM and PM contexts had higher scores on PC1 than handling context, indicating a more positive emotional state during AM and PM than during handling. PM context had lower scores on PC2 than AM and handling contexts. Thus, QBA at pasture was able to discriminate between contexts that should differ in emotional valence and arousal. Thirdly, an observer assessed the behavioural responses of cows to handling by direct observations followed by QBA. Pearson correlation coefficients were calculated between the number of behavioural responses and the coordinates on PC1 and PC2. The more cows were observed trotting, the more the cow's mood was perceived as negative (r=-0.71) and the more cows were observed galloping and turning, the more the cow was perceived as excited (r=0.77 and 0.60). In conclusion, the QBA appears to be a valid measure of the emotional and arousal state of dairy cows at pasture, but inter-observer reliability could be improved.

Keywords: emotional state; welfare assessment, reliability, grazing cattle, handling

1 Introduction

Qualitative Behaviour Assessment (**QBA**) has been developed as a measure of emotional state (Forkman and Keeling, 2009; Wemelsfelder et al., 2001). QBA is an integrative assessment of the 'whole animal', where the focus is not on what behaviour or how many behaviours the animal performs, but on how the animal behaves and interacts with its environment (Wemelsfelder, 1997; Wemelsfelder et al., 2001). QBA consists of observing an individual animal or a group of animals over a period of time and assessing the animal's behaviour based on the animal's dynamic body language, using either adjectives spontaneously generated by the observers (free choice profiling) or a fixed list of adjectives (or descriptors) (Fleming et al., 2016). For example, the Welfare Quality protocol includes a QBA with a fixed list of adjectives such as 'excited', 'content', 'anxious' (Forkman and Keeling, 2009). The QBA, which is carried out by observing the animals in their daily environment without disturbance, is intended to measure the general emotional state of the animals. QBA can also be conducted during specific situations/contexts to assess how the situation is perceived by the animals. For example, QBA has been carried out on veal calves during handling to assess the impact of the handling style of the keeper (Ellingsen et al., 2014).

The validation of QBA as a welfare indicator has mainly focused on observer reliability and selectivity. Inter-observer reliability has been confirmed in many studies (Ceballos et al., 2021; Phythian et al., 2013; Rousing and Wemelsfelder, 2006). However, some authors found low inter-observer reliability (Bokkers et al., 2012; Czycholl et al., 2016; Gutmann et al., 2015). Therefore, inter-observer reliability must be assessed before using QBA. The selectivity of a measure corresponds to the ability of the measure to quantify what it is supposed to quantify (Knierim et al., 2021). The selectivity of QBA has been validated through convergent validity by determining the correlation of QBA results with other welfare indicators (e.g. proportion of lame animals in sheep (Phythian et al., 2016), clinical indicators of disease (e.g. in dairy cows, de Boyer des Roches et al, 2018), physiological measures (e.g. in cattle, Stockman et al., 2011) or behavioural measures (e.g. in polar bears, Skovlund et al., 2023; in salmon, Wiese et al., 2023) known to reflect animal welfare and emotional state. The selectivity of QBA has also been validated through construct validity, by investigating whether QBA is able to discriminate between different contexts/situations that are thought to induce different emotional and arousal states in animals (e.g. in salmon, Wiese et al., 2023). QBA has been validated in a wide range of species, including farm animals (e.g. in Welfare Quality protocols for cattle, pigs and poultry), zoo animals (e.g. polar bears, Skovlund et al., 2023) and companion animals (e.g. dogs, Arena et al., 2019), and is now widely used as a welfare indicator (in Welfare Quality and AWIN protocols for cattle, pigs, poultry, sheep, goats and donkeys). In the Welfare Quality protocols (cattle, pigs and poultry), QBA is the only measure used to assess positive emotional states.

QBA has been developed and validated as a measure of emotional state for indoor cows in either tie stalls, cubicles or straw-bedded farms (Forkman and Keeling, 2009). More recently, the Welfare Quality protocol has been applied to cattle at pasture, including QBA (e.g. in dairy cows, Armbrecht et al., 2019; Wagner et al., 2018; in dual-purpose cattle, Hernandez et al., 2017). However, to our knowledge, it has not yet been determined whether QBA developed for indoor cows is relevant and valid for grazing cows. The validation of QBA at pasture seems necessary to have a valid measure of emotional state, as even at pasture cows can be exposed to risks that may affect their welfare (Aubé et al., 2022).

In order to determine whether QBA could be included in a protocol to assess cow welfare at pasture, we propose to investigate the validity (both reliability and selectivity) of this measure applied to dairy cows at pasture. The first objective of this study was to determine if QBA of cows at pasture could be interpreted in a similar way to that of cows indoors. The second objective was to assess observer reliability for QBA at pasture. The third objective was to assess the selectivity of the measure through construct validity by comparing three contexts that are thought to differ in the emotional and arousal state of cows, and through convergent validity by investigating the relationship between QBA and quantitative behavioural reactivity during handling.

2 Materials and methods

The study took place at the INRAE farm 'Le Pin-au-Haras' (France, 48.448N, 0.098E, DOI: 10.15454/1.5483257052131956E12). This study was part of an experiment on the feeding management of dairy cows' in a rotational grazing system. The experimental protocol was assessed by the Rennes Institutional Animal Care and Use Committee of Rennes and approved by the French Ministry of Higher Education, Research and Innovation (APAFIS agreement #20846-2019052810487566 v2).

2.1 Animals, housing, and management

The housing and management of the cows are described in Aubé et al. (2023). Briefly, a herd of 144 cows (48 Holstein, 48 Normande, 48 Jersey) was housed indoors during winter (from December to March) and turned out to pasture in spring after calving. The present study took place when the cows were at pasture (from April to November 2021). The cows were at pasture all day long, except when they were brought indoors for milking, between 07h00 and 08h30 and between 16h00 and 17h30. The cows grazed on plots of 10 ha. They changed plot after 7 to 12 days of grazing.

2.2 Qualitative behaviour assessment (QBA)

2.2.1 Observation conditions

2.2.1.1 Qualitative behaviour assessments from videos.

To assess intra- and inter-observer reliability, twenty videos (1 min each) were taken of cows grazing in different contexts: cows mainly resting or ruminating (n = 5 videos), cows mainly grazing (n = 10), part of the herd grazing and part resting or ruminating (n = 2), cows released to pasture for the first time after wintering indoors (n = 1), cows being bothered by flies in the sun (n = 1), and cows reacting to a truck passing near the pasture (n = 1). Videos included from 4 to approximately 100 cows. To assess intra- and inter-observer reliability, each video was watched twice (at least 24 h apart) during two sessions (4 months apart, one in March, one in July) by 5 observers (A, B, C, D and E), resulting in a total of 400 observations (20 videos x 4 times x 5 observers). Prior to the experiment, observers A and C were trained by two scientists who followed the training programme organised by Welfare Quality and regularly train people in the use of the dairy cow protocol. Subsequently, Observer A trained Observers B, D and E. The QBA training of Observers B, D and E was part of a larger training on 15 animal-based indicators and 3 resource-based indicators (not described in this paper), which took about 15 spread over 5 half days. Of these 15 h of training, about 30 min were spent in the classroom explaining the theoretical background of QBA and about 1 hour was spent in the field (observing a herd of cows at pasture in the morning, in the afternoon and during handling). The observers were female (A, C) or male (B, D and E). Observer A was an ethologist, Observer C was a veterinarian and Observers B, D and E were technicians from the experimental farm who had worked daily with cattle for several years.

2.2.1.2 Live QBA

Live QBA was performed by two observers (Observer A was always present) on the whole herd at pasture on 33 occasions and in different contexts: after returning to pasture from morning milking

(AM context, n = 8), before afternoon milking (PM context, n = 9) and during the collection of cows at pasture for afternoon milking (handling context, n = 16) (see Table 1 for dates, contexts and observers involved).

For the AM context, the two observers remained in the pasture plot as the cows arrived from morning milking until the cows dispersed to graze (about 5 min), after which they performed the observation. For the PM context, as the cows were already present in the pasture plot when the observers arrived, the observer entered the pasture and stood still for about 5 min to allow the cows to get used to the presence of the observers. Prior to the AM and PM observations, the two observers discussed and virtually divided the pasture plot into 4 observation areas, which they defined together in order to observe all animals. The boundaries of the observation areas (using environmental cues such as trees or water troughs) were chosen by the two observers so that each area contained approximately one quarter of the total herd (i.e. between 30 and 40 cows). The time allocated to each observation area was determined according to the Welfare Quality protocol (i.e. 5 min/area for 4 areas for a total of 20 min of observation). The two observers spent 5 min in each observation area, visually scanning the designated area to observe all cows present in the area.

For the handling context, the two observers stood behind the herd in the pasture plot (in relation to the exit of the pasture plot) in order to see the herders leading the cows to the exit. Observation began when the first stockperson entered the pasture and ended when the last cow of the herd left the pasture. As the herd moved forward, the two observers walked quietly behind to follow and observe the herd. Throughout the handling process, the two observers visually scan the entire herd. The duration of the handling context observation depended only on the duration of the handling process itself (i.e. from 3 to 17 min).

2.2.2 *Scoring procedure*

QBA was scored according to the Welfare Quality (2009) protocol. Briefly, each observer observed the group of animals (either 1 min on video or 3 to 20 min live), then assigned a value to 20 adjectives on a 125 mm visual analogue scale, based on how well the adjective describes the group. The adjectives were: Active, Agitated, Apathetic, Bored, Calm, Content, Distressed, Fearful, Friendly, Frustrated, Happy, Indifferent, Inquisitive, Irritable, Lively, Playful, Positively occupied, Relaxed, Sociable, and Uneasy.

2.3 *Quantitative behaviour measures during cow collection*

Cow reactivity to handling at herd level was assessed on 16 occasions (by observer A concurrently with the QBA in handling context) during cow collection in the pasture for afternoon milking (see Table 1 for detailed dates). During cow collection, observer A measured the duration of the entire handling process using a chronometer and recorded the occurrence (all occurrence sampling, Altmann, 1974) of the following responses to the stockperson's actions: cow freezing, startled, kicking, trotting, galloping, slipping, falling, moving backwards, turning around, vocalising (ethogram in Table 2).

The intra-observer reliability of observer A for cow behavioural responses to handling (galloping and trotting behaviours only because they were the most frequent responses) was tested on 30 videos (of 15 s each) recorded during cow collection and viewed twice by the observer. This observer was reliable for these two behaviours with interclass correlation coefficient > 0.95 and percentage of agreement (number of observations where observer A agreed with the total number of observations and multiplied by 100) > 95% for both galloping and trotting behaviour.

2.4 *Statistical analyses*

Statistical analyses were performed using R software version 4.2.2 (R Core Team, 2022). For all analyses, the statistical unit is a QBA observation made by an observer.

2.4.1 Principal component Analysis on QBA and adjectives weights to produce a QBA Score

In order to summarise the information from the QBA of each video observation (n = 400), as in Welfare Quality, a Principal Component Analysis (**PCA**, based on the correlation matrix and without rotation) was performed on the values assigned to the 20 adjectives (from 0 to 125). The PCA was performed using the "ade4" package (Chessel et al., 2004). The coordinates of the live observations on the PCA first axis (**PC1**) and second axis (**PC2**) were further used as the output of the QBA to test intra- and inter-observer reliability. A linear model was applied with the coordinate of observations on PCA first axis (PC1) as explanatory variable and each adjective value as predictors (Welfare Quality, 2009; Brscic et al. (2019)). The effect (i.e. estimate) of an adjective in the linear model is used as the weight for that adjective to calculate the QBA score (see formula below). We checked whether the weights of the adjectives found in our study were similar to the weights of the adjectives used in Welfare Quality for Dairy Cattle. First, we checked that the weights had the same sign. Second, we calculated the Pearson correlation between the weights. Thirdly, the absolute value of the weight of each adjective was used to rank the 20 adjectives to assess the relative importance of the adjectives (from 1, adjective with the highest absolute weight, to 20, adjective with the lowest absolute weight). We then compared the ranking of the 20 adjective weights between our study and Welfare Quality (2009). We considered the rankings of the absolute value of the weight to be close if they were within a range of 2 (e.g. rank 5 vs. rank 3).

The QBA score is a weighted sum where the adjective scores are multiplied by their weights and then summed (Brscic et al., 2019).

$$QBA\ Score = constant + \sum_{k=1}^{20} \omega_k N_k$$

where Nk is the value (in mm) on the scale of a given adjective k for one observation, ωk is the weight given to the adjective and 'constant' is a fixed value corresponding to the intercept of the linear model described above.

We calculated the QBA score of the video observations, using the adjectives' weights and the constant from the present study. To compare our results with those of Welfare Quality (2009), we calculated a QBA score using the weights and the constant defined in Welfare Quality (which differ from ours because the PCA and the linear model were applied to a different set of observations). We will refer to the pasture score (calculated with the weights and constant obtained in the present study) vs. Welfare Quality score (calculated according to the weights and constant defined in Welfare Quality). Pearson correlation coefficient was calculated to assess the relation between the two scores.

2.4.2 Intra- and inter-observer reliability of QBA

Intra- and inter-observer reliability was assessed using the coordinates on PC1 or PC2 of each observation made by each observer from the PCA previously performed on the video observations (see 0).

Kendall's concordance coefficients (**W**) and interclass correlation coefficients (**ICC**) were calculated, using the 'irr' package (Gamer et al., 2019). W and ICC comparing PC1 or PC2 obtained by the same observer between two days of video observations indicate the intra-observer reliability (calculated for each observer for each session). W and ICC calculated on PC1 or PC2 obtained by the different observers on the same day of video observation indicate inter-observer reliability. For W, values less than 0.4 are considered to indicate poor reliability, values between

0.41 and 0.70 - moderate reliability, values between 0.71 and 0.90 – high reliability, and values greater than 0.91 - excellent reliability (Martin and Bateson, 1993). For ICC, values less than 0.5 are considered indicative of poor reliability, values between 0.5 and 0.75 - moderate reliability, values between 0.75 and 0.9 - good reliability, and values greater than 0.90 - excellent reliability (Koo and Li, 2016).

2.4.3 Construct validity

Live QBA observations (AM, PM, and handling contexts) from both observers were added as supplementary observations to the PCA previously conducted with the video observations. The coordinates of the live observations on both PC1 and PC2 were compared between the three contexts (AM, PM, and handling context) with a linear mixed effects model (using the R package lme4) with context as a fixed effect and observer as a random effect. We hypothesised that emotional valence during handling would be less positive and arousal would be lower in PM cows than in the other two contexts. We based our hypothesis on the fact that motivation for cows to be milked is low (Prescott et al., 1998), whereas cows are highly motivated for access to pasture (Von Keyserlingk et al., 2017). We also assumed that emotional valence would be quite similar between the two pasture contexts (AM and PM), whereas arousal levels would likely differ, with cows being more active during AM observations (cows mostly grazing) and less active during PM observations (cows mostly resting and ruminating, personal observation). Variations in PC1 or PC2 with context would therefore indicate that the QBA is valid for discriminating emotional and arousal states of animals (construct validity).

2.4.4 Convergent validity

The Pearson correlation coefficients calculated between the frequency of each behaviour and the coordinates of PC1 and PC2 during cow collection are used to assess convergent validity.

2.4.5 *Stability over the grazing season*

The stability of the QBA results (PC1 and PC2), obtained from the live observations (AM, PM and handling contexts) of observer A only, was examined visually by examining the different observations made on the herd throughout the season. In addition, the variability (amplitude: maximum value minus minimum value obtained) of QBA per context over the season was compared with the total variability of QBA per visit, all contexts together.

3 Results

3.1 *Principal component analysis of QBA from video and adjectives' weights*

The PCA was performed on 399 video observations (instead of 400 initially due to one missing observation). The first axis explained 31% of the variation (eigenvalue, 6.1) and a second one explained 27% of the variation (eigenvalue, 5.4). Adjectives with absolute loadings > 0.6 on PC1 were: content, happy, positively occupied (positive loadings) vs. frustrated irritable, uneasy, and distressed (negative loadings) (Fig. 1 and Table 3). PC1 seems thus to be related to emotional valence: the higher the value on PC1 the more positive the emotional valence. Adjectives with absolute loadings > 0.6 on PC2 were: active, agitated, friendly, playful, lively, inquisitive and sociable (positive loadings) vs. calm (negative loading). PC2 thus seems to relate to arousal: the higher the value on PC2 the more active the animals.

For 16 out of the 20 adjectives, the sign of the weight in the present study was the same as in Welfare Quality (2009). For the adjectives 'indifferent', 'sociable', 'playful' and 'friendly', the sign of the weight differed from that in Welfare Quality (Table 3). In addition, the adjective weights from our video observations and the Welfare Quality weights were highly correlated (r = 0.86). The rankings of the absolute value of the weight were close (within a range of 2 for 12 adjectives), except for 'restless', 'bored', 'calm' and 'lively' (for which the weights at pasture had a higher rank than in Welfare Quality) and 'indifferent', 'friendly', 'apathetic' and 'excited' (for which the weights from Welfare Quality had a higher rank than at pasture) (Table 3).

Pasture scores ranged from 0 to 81.9 and Welfare Quality scores ranged from 0 to 86.4. Mean Pasture score was 53.1 (SD, 18.1) while mean Welfare Quality score was 55.6 (SD, 19.0). Pearson correlation coefficient between the Pasture score and Welfare Quality score was 0.95.

3.2 Intra and Inter-observer reliability of QBA

Intra-observer reliability for PC1 and PC2 was high to excellent for observers A and B. Intra-observer reliability of observers C, D and E for PC1 was moderate in Session 1 and high to excellent in Session 2 except for Observer D (moderate reliability regarding ICC on Session 2). Intra-observer reliability of observers C, D and E for PC2 was high to excellent but moderate for Observer D on Session 1 regarding ICC (Table 4

Inter-observer reliability as assessed from W (Kendall's concordance coefficients) was moderate for PC1 in both sessions and high for PC2 in both sessions. Inter-observer reliability as assessed from ICC was moderate in both sessions for PC1 and PC2 except for PC2 on Session 2 Day 1 for which it was considered high (ICC = 0.76) and for PC1 on Session 1 Day 1 for which it was considered low (ICC = 0.47) (Table 4).

3.3 Construct validity

The observations at pasture during AM and PM had higher coordinates on PC1 (i.e. reflecting a better emotional state) than the observations during handling (Table 5).

The observations at pasture during the PM context had lower coordinates on PC2, i.e. reflecting lower level of arousal, than observations during the AM and handling contexts (Table 5).

3.4 Convergent validity

During the handling observations (n = 16), no kicking, no falling, no moving backwards and no freezing were observed. There was one case of startle and two cases of slipping. These responses were therefore not included in the statistical analysis. Only, turning around, galloping, trotting and vocalizing were retained for analysis. Turning frequency per observation ranged from a minimum of 0 to a maximum of 14 (median = 2.5). Trotting frequency ranged from 0 to 37 (median = 21). Galloping frequency ranged from 0 to 36 (median = 10). Vocalizing frequency ranged from 0 to 8 (median = 0). The duration of the handling process ranged from 3.2 to 17.4 min (median = 11.3 min).

The frequency of galloping and turning around observed during handling were positively correlated with PC2 (arousal) ($r = 0.77$ and 0.60, $P = 0.001$ and 0.015). The duration of handling and the number of trotting behaviours were negatively correlated with PC1 (valence) ($r = -0.70$ and -0.71, $P = 0.012$ and 0.004). Vocalization frequency was not significantly correlated to either PC1 or PC2 ($r < 0.4$, $P > 0.05$).

3.5 Stability over the season

The handling context had always negative coordinates on PC1 (except for one observation at visit 3), whereas the AM and PM contexts always had positive coordinates on PC1 (Fig. 1). Throughout the season, PC1 values (i.e. coordinates of the observations on PC1) ranged from 2.2 to 3.2 for AM, from 0.5 to 2.9 for PM and from -4.0 to 0.7 for handling contexts. The amplitudes of PC1 values were thus, 1.0 for AM, 2.4 for PM and 4.7 for handling. The amplitude of PC1 values all contexts considered, were 6.2, 4.8, 4.3, 4.1, 5.9, 6.4 and 5.0 for Visits 1 to 7 respectively.

The AM and PM contexts always had negative coordinates on PC2, AM and the handling context had either on positive or negative coordinates on PC2 depending on the visit (Fig. 1). Over the whole season, PC2 values (i.e. coordinates of the observations on PC2) ranged from -1.0 to 1.0 for

AM, from -3.0 to 2.9 for PM and from -0.8 to 2.5 for handling context. The amplitudes of PC2 values were thus, 2.0 for AM, 2.9 for PM and 3.2 for handling. The amplitude of PC2 values all contexts considered, were 2.5, 3.8, 1.3, 1.0, 2.5, 3.2 and 5.4 for Visits 1 to 7 respectively.

4 Discussion

Our study shows that QBA has the potential to capture the arousal and emotional state of cows at pasture. However, inter-observer reliability should be improved.

The use of the 20 QBA adjectives as defined in Welfare Quality for indoor cows makes sense for cows at pasture.

Our results showed that PCA axes from QBA of cows at pasture (from video observations) are similar to PCA axes previously obtained from QBA performed on cows indoors (Forkman and Keeling, 2009; Gutmann et al., 2015; Russell et al., 2023; Winckler, 2014) with PC1 being the 'mood' axis and PC2 being the 'arousal' axis. In addition, our PCA axes from QBA of cows at pasture are similar to those found from observations of beef cattle either at pasture in summer or indoors throughout the year (Cooke et al., 2023; except that the axes were reversed in their study (PC1 : 'arousal' and PC2 : 'mood'). Similar results were also found for goats including goats at pasture (observations in eight farms where goats were housed indoors and eight farms where goats were at pasture) (Grosso et al., 2016). Furthermore, the adjectives' weights that we obtained for cows at pasture were similar to those proposed in the Welfare Quality protocol for cows indoor. As a result, the pasture scores calculated from adjectives' weights from our observations were highly correlated with the scores calculated as in the Welfare Quality protocol, and the distributions of the two scores were similar. QBA can therefore be used for dairy cows at pasture as it is described in the Welfare Quality protocol (same adjectives, same formula to calculate the score).

Our construct and convergent validity results support the idea that QBA has the potential to capture emotional state and arousal of cows at pasture.

QBA at pasture is selective in that it is able to discriminate between different contexts that we would expect to be different in terms of both emotional valence and arousal. Indeed, in the present study, the handling situation, which we expected to be less positive in terms of emotional state for the animals compared to situations in which the animals were not disturbed (morning or afternoon observations), obtained lower scores on PC1. In addition, QBA was able to discriminate between contexts in terms of arousal, with afternoon observations (cows are mainly resting and ruminating at this time) scoring lower on PC2 than handling observations and morning observations (cows are all grazing at this time). As QBA at pasture can discriminate between emotional and arousal contexts, we assume that it can also discriminate between farms with different emotional states of the animals.

Gutmann et al. (2015) also found differences in PC1 and PC2 for QBA performed at three different times of the day (early morning, late morning, early afternoon) for cows indoors. This raises the question of the moment of day for QBA, as different moments may give different results. QBA should either be done at different times of the day to get a global view or always at the same time to compare situations (same herd at different times or different herds). If QBA is to be carried out only once during the day, it is advisable to carry out the QBA of grazing cows after morning milking (as in the Welfare Quality protocol for indoor cows), when the cows are coming back to pasture from milking. At this time, all cows are generally active and grazing, making it more likely to observe the cow's interactions with her environment and conspecifics (personal observations). In addition, according to Gutman et al. (2015), performing QBA at the beginning of a welfare assessment day could minimise potential observer bias arising from the assessment of other welfare measures. This is already done in the Welfare Quality protocol for dairy cows; indeed, QBA is carried out at the beginning of the assessment day, immediately after avoidance distance tests and before individual measures carried out in the home pen.

In our study, duration of handling and trotting frequency correlated negatively with PC1. Thus, the longer the handling and the more cows were observed to trot in response to the handlers, the more negative the mood of the cow was perceived by the observer. Galloping and trotting frequencies were positively correlated with PC2, i.e. the more cows were observed galloping away from the handler and turning in response to the handlers, the more the cows were perceived as excited by the observer. These results confirm that QBA is selective in assessing the arousal and emotional state of cows, at least during handling. This is in line with a previous study aimed at identifying appropriate measures of reactivity of dairy cows towards humans (Ebinghaus et al., 2016). Values obtained on PC1 (mood axis) from QBA during two handling tests on individual cows (tolerance to standardised tactile interaction (**TTI**) and release behaviour after restraint (**RB**)) were correlated with reactivity scores during these tests (Ebinghaus et al., 2016). The more reactive cows were to TTI and RB, the more negative emotional state was obtained from QBA. Ebinghaus et al. (2016) concluded that QBA can be used to assess cow reactivity during handling, which is consistent with our results.

As dairy cows at pasture are usually handled at least twice a day for milking, the quality of handling may be an issue for their welfare. We therefore suggest that QBA during handling can be included as a new measure in a future protocol to assess handling quality. As the duration of the handling process in our study ranged from 3 to 17 min, it seems very feasible to include such observations in a protocol. However, the duration and conditions of handling may vary between farms, so QBA during handling may not always be feasible.

Intra-observer reliability was acceptable but inter-observer reliability should be improved.

Intra-observer reliability for PC1 and PC2 was mainly good (except moderate for three observers at the first session), which is in line with previous studies investigating intra-observer reliability of QBA. For example, intra-observer reliability for both PC1 and PC2, as assessed by Spearman rank correlations, was found to be high for QBA used to assess social behaviour in cows (Rousing and Wemelsfelder, 2006). Similarly, the intra-observer reliability of eight observers was considered high for PC1 ($W = 0.71$ to 0.89) and PC2 ($W = 0.74$ to 0.90) for QBA used to assess maternal protective behaviour in dairy cows (Ceballos et al., 2021).

Inter-observer reliability was low to moderate for PC1 and moderate to high for PC2, so we consider that inter-observer reliability was close to acceptable but not fully validated, especially for PC1. This result is not surprising as other studies in cattle have found W coefficients indicating low ($W \leq 0.40$) or moderate ($W \leq 0.70$) inter-observer reliability (Andreasen et al., 2013; Forkman and Keeling, 2009; Gutmann et al., 2015). For example, in indoor dairy cows, inter-observer reliability was low for PC1 ($W = 0.38$) and moderate for PC2 ($W = 0.46$) (Forkman and Keeling, 2009).

In our study and in previous studies (Battini et al., 2021; Forkman and Keeling, 2009; Gutmann et al., 2015), inter-observer reliability seemed to be better for the axis representing arousal than for the axis representing mood. This may be because it is easier to assess the activity, which is external, than an animal's mood, which is an internal state and seems more open to interpretation.

However, acceptable inter-observer reliability of QBA has been found in others studies, probably due to extensive training of the observers. For example, in sheep a high inter-observer reliability was obtained after 3 h of training dedicated to QBA alone (Phythian et al., 2013). In contrast, in our study, the training on QBA was part of a larger training on 15 animal-based indicators, together with 3 resource-based indicators (not described in this paper), which was conducted in about 15 h spread over 5 half days, with about 1.5 h dedicated to QBA. Increasing the training time with the observer on QBA could probably improve inter-observer reliability.

In our study, inter-observer reliability seemed to be better for the second session than for the first. In fact, the W coefficients were all slightly better (or at least equal) during the second session of video observations for both axes than during the first. This suggests that the observers became more reliable with practice. With more time for training before session 1, we would probably have achieved good inter-observer reliability. Providing brief definitions of each adjective, as has been done for several species (donkeys, Minero et al., 2016; goats, Grosso et al., 2016; horses, Minero et al., 2018; sows, Ibach et al., 2024; and shelter dogs, Stubsjøen et al., 2020), to avoid misunderstanding of adjectives by observers, could also improve inter-observer agreement. Definition of adjectives and extensive training of observers should be considered to improve the quality of QBA.

PC1 (“mood”) values appeared to be stable across the grazing season for the AM and PM contexts but not for the handling context, while PC2 (“arousal”) scores did not appear to be stable for any context.

The variation (amplitude) in PC1 scores for the AM and PM contexts was lower than the variation between visits, but the variation in PC1 scores for the handling context seemed to be quite similar to the variation between visits. This suggests that PC1 scores were quite stable over the season, except for the handling context. The variation for the handling context may be due to the pair of handlers not being the same between visits. The variation in PC2 scores (amplitude) within the AM, PM and handling contexts seemed to be close to the variations in PC2 scores between visits (all contexts considered). This suggests that PC2 scores were not stable over the season whatever the context. As the QBA score calculated in the Welfare Quality protocol is based on PC1 only, this measure has the potential to be a stable measure for cows at pasture under undisturbed contexts. We only observed one herd of cows live, the results should be confirmed by observing different herds and contexts.

5 Conclusion

In conclusion, the QBA appears to be a valid measure of the emotional and arousal state of dairy cows at pasture. The time and context of observation during the day must be chosen according to the aim of the study, because QBA results vary during the day (afternoon vs. morning) and in different contexts (undisturbed vs. handling situation). Before using QBA, inter-observer reliability must be checked and training must be intensified if observers' reliability is not good.

Declaration of interest

None.

Acknowledgements

The authors thank Estelle Mollaret (Vetagro Sup), Yves Carbonnier (INRAE), Loïc Leloup (INRAE) and Laurent Chanteloup (INRAE) for observing the cows, Luc Delaby and Frédéric Launay for affording us access to the experimental unit, and the staff from INRAE Le Pin (https://doi.org/10.15454/1.5483257052131956E12) for taking care of the animals. We also thank Marianne Berthelot (Anses) for the discussions about statistical analysis. The experiment was co-financed by the Banque des Territoires and Conseil Régional de Normandie, as part of the France 2030 Territoires d'Innovation LIT OUESTEREL project.

References

Altmann, J., 1974. Observational study of behavior: sampling methods. Behaviour 49, 227–266.

Andreasen, S.N., Wemelsfelder, F., Sandøe, P., Forkman, B., 2013. The correlation of Qualitative Behavior Assessments with Welfare Quality® protocol outcomes in on-farm welfare assessment of dairy cattle. Applied animal behaviour science 143, 9–17.

Arena, L., Wemelsfelder, F., Messori, S., Ferri, N., Barnard, S., 2019. Development of a fixed list of terms for the Qualitative Behavioural Assessment of shelter dogs. PLOS ONE 14, e0212652. https://doi.org/10.1371/journal.pone.0212652

Armbrecht, L., Lambertz, C., Albers, D., Gauly, M., 2019. Assessment of welfare indicators in dairy farms offering pasture at differing levels. Animal 13, 2336–2347. https://doi.org/10.1017/S1751731119000570

Aubé, L., Mialon, M.M., Mollaret, E., Mounier, L., Veissier, I., de Boyer des Roches, A., 2022. Review: Assessment of dairy cow welfare at pasture: measures available, gaps to address, and pathways to development of ad-hoc protocols. animal 16, 100597. https://doi.org/10.1016/j.animal.2022.100597

Aubé, L., Mollaret, E., Mialon, M.M., Mounier, L., Veissier, I., de Boyer des Roches, A., 2023. Measuring the human–animal relationship in cows by avoidance distance at pasture. Applied Animal Behaviour Science 265, 105999. https://doi.org/10.1016/j.applanim.2023.105999

Battini, M., Renna, M., Giammarino, M., Battaglini, L., Mattiello, S., 2021. Feasibility and Reliability of the AWIN Welfare Assessment Protocol for Dairy Goats in Semi-extensive Farming Conditions. Frontiers in Veterinary Science 8, 1210. https://doi.org/10.3389/fvets.2021.731927

Bokkers, E.A.M., Vries, M., Antonissen, I., Boer, I.J.M., 2012. Inter- and intra-observer reliability of experienced and inexperienced observers for the Qualitative Behaviour Assessment in dairy cattle. Animal Welfare 21, 307–318. https://doi.org/10.7120/09627286.21.3.307

Brscic, M., Dam Otten, N., Contiero, B., Kirchner, M.K., 2019. Investigation of a standardized qualitative behaviour assessment and exploration of potential influencing factors on the emotional state of dairy calves. Animals 9, 757.

Ceballos, M.C., Góis, K.C.R., Sant'Anna, A.C., Wemelsfelder, F., da Costa, M.P., 2021. Reliability of qualitative behavior assessment (QBA) versus methods with predefined behavioral categories to evaluate maternal protective behavior in dairy cows. Applied Animal Behaviour Science 236, 105263.

Chessel, D., Dufour, A.B., Thioulouse, J., 2004. The ade4 package-I-One-table methods. R news 4, 5–10.

Consortium, W.Q., 2009. Welfare Quality® assessment protocol for cattle. Welfare Quality Consortium, Lelystad, the Netherlands.

Cooke, A.S., Mullan, S., Morten, C., Hockenhull, J., Le-Grice, P., Cocq, K.L., Lee, M.R.F., Cardenas, L.M., Rivero, M.J., 2023. Comparison of the welfare of beef cattle in housed and grazing systems: hormones, health and behaviour. The Journal of Agricultural Science 161, 450–463. https://doi.org/10.1017/S0021859623000357

Czycholl, I., Kniese, C., Büttner, K., Beilage, E. grosse, Schrader, L., Krieter, J., 2016. Interobserver reliability of the 'Welfare Quality® Animal Welfare Assessment Protocol for Growing Pigs.' SpringerPlus 5, 1114. https://doi.org/10.1186/s40064-016-2785-1

de Boyer des Roches, A., Lussert, A., Faure, Marion., Herry, Vincent., Rainard, Pascal., Durand, Denys., Wemelsfelder, F., Foucras, G., 2018. Dairy cows under experimentally-induced

Escherichia coli mastitis show negative emotional states assessed through Qualitative Behaviour Assessment. Applied Animal Behaviour Science 206, 1–11. https://doi.org/10.1016/j.applanim.2018.06.004

Ebinghaus, A., Ivemeyer, S., Rupp, J., Knierim, U., 2016. Identification and development of measures suitable as potential breeding traits regarding dairy cows' reactivity towards humans. Applied Animal Behaviour Science 185, 30–38. https://doi.org/10.1016/j.applanim.2016.09.010

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. Applied Animal Behaviour Science 153, 10–17. https://doi.org/10.1016/j.applanim.2014.01.011

Fleming, P.A., Clarke, T., Wickham, S.L., Stockman, C.A., Barnes, A.L., Collins, T., Miller, D.W., 2016. The contribution of qualitative behavioural assessment to appraisal of livestock welfare. Anim. Prod. Sci. 56, 1569–1578. https://doi.org/10.1071/AN15101

Forkman, B., Keeling, L.J., 2009. Assessment of animal welfare measures for dairy cattle, beef bulls and veal calves (Welfare Quality Report No. 11). School of City and Regional Planning, Cardiff University, Cardiff, UK.

Gamer, M., Lemon, J., Singh, I.F.P., 2019. irr: Various Coefficients of Interrater Reliability and Agreement. R package version 0.84.1.

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. Applied Animal Behaviour Science 180, 51–57.

Gutmann, A.K., Schwed, B., Tremetsberger, L., Winckler, C., 2015. Intra-day variation of Qualitative Behaviour Assessment outcomes in dairy cattle. Animal Welfare 24, 319–326. https://doi.org/10.7120/09627286.24.3.319

Hernandez, A., Berg, C., Eriksson, S., Edstam, L., Orihuela, A., Leon, H., Galina, C., 2017. The Welfare Quality® assessment protocol: how can it be adapted to family farming dual purpose cattle raised under extensive systems in tropical conditions? Animal Welfare 26, 177–184. https://doi.org/10.7120/09627286.26.2.177

Ibach, S., Chou, J.-Y., Battini, M., Parsons, T.D., 2024. A systematic approach to defining and verifying descriptors used in the Qualitative Behavioural Assessment of sows. Animal Welfare 33, e8. https://doi.org/10.1017/awf.2024.6

Knierim, U., Winckler, C., Mounier, L., Veissier, I., 2021. Developing effective welfare measures for cattle. pp. 81–102. https://doi.org/10.19103/AS.2020.0084.05

Minero, M., Dalla Costa, E., Dai, F., Canali, E., Barbieri, S., Zanella, A., Pascuzzo, R., Wemelsfelder, F., 2018. Using qualitative behaviour assessment (QBA) to explore the emotional state of horses and its association with human-animal relationship. Applied Animal Behaviour Science 204, 53–59. https://doi.org/10.1016/j.applanim.2018.04.008

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. Applied Animal Behaviour Science 174, 147–153. https://doi.org/10.1016/j.applanim.2015.10.010

Phythian, C., Michalopoulou, E., Duncan, J., Wemelsfelder, F., 2013. Inter-observer reliability of Qualitative Behavioural Assessments of sheep. Applied Animal Behaviour Science 144, 73–79. https://doi.org/10.1016/j.applanim.2012.11.011

Phythian, C.J., Michalopoulou, Eleni., Cripps, P.J., Duncan, J.S., Wemelsfelder, Françoise., 2016. On-farm qualitative behaviour assessment in sheep: Repeated measurements across time, and association with physical indicators of flock health and welfare. Applied Animal Behaviour Science, Behavioural indicators of health 175, 23–31. https://doi.org/10.1016/j.applanim.2015.11.013

Prescott, N.B., Mottram, T.T., Webster, A.J.F., 1998. Relative motivations of dairy cows to be milked or fed in a Y-maze and an automatic milking system. Applied Animal Behaviour Science 57, 23–33. https://doi.org/10.1016/S0168-1591(97)00112-3

R Core Team, 2022. R: A language and environment for statistical computing.

Rousing, T., Wemelsfelder, F., 2006. Qualitative assessment of social behaviour of dairy cows housed in loose housing systems. Applied Animal Behaviour Science 101, 40–53. https://doi.org/10.1016/j.applanim.2005.12.009

Russell, A., Randall, L., Kaler, J., Eyre, N., Green, M., 2023. Use of qualitative behavioural assessment to investigate affective states of housed dairy cows under different environmental conditions. Frontiers in veterinary science 10, 1099170. https://doi.org/10.3389/fvets.2023.1099170

Skovlund, C.R., Kirchner, M.K., Contiero, B., Ellegaard, S., Manteca, X., Stelvig, M., Tallo-Parra, O., Forkman, B., 2023. Qualitative Behaviour Assessment for zoo-housed polar bears (*Ursus maritimus*): Intra- and inter-day consistency and association to other indicators of welfare. Applied Animal Behaviour Science 263, 105942. https://doi.org/10.1016/j.applanim.2023.105942

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 naïve and habituated to road transport. Animal Production Science 51, 240–249.

Stubsjøen, S.M., Moe, R.O., Bruland, K., Lien, T., Muri, K., 2020. Reliability of observer ratings: Qualitative behaviour assessments of shelter dogs using a fixed list of descriptors. Veterinary and Animal Science 10, 100145.

Von Keyserlingk, M.A., Cestari, A.A., Franks, B., Fregonesi, J.A., Weary, D.M., 2017. Dairy cows value access to pasture as highly as fresh feed. Scientific reports 7, 44953.

Wagner, K., Brinkmann, J., March, S., Hinterstoißer, P., Warnecke, S., Schüler, M., Paulsen, H.M., 2018. Impact of daily grazing time on dairy cow welfare—Results of the Welfare Quality® protocol. Animals 8, 1.

Wemelsfelder, F., 1997. The scientific validity of subjective concepts in models of animal welfare. Applied Animal Behaviour Science, Basic and Applied Aspects of Motivation and Cognition 53, 75–88. https://doi.org/10.1016/S0168-1591(96)01152-5

Wemelsfelder, F., Hunter, T.E., Mendl, M.T., Lawrence, A.B., 2001. Assessing the 'whole animal': a free choice profiling approach. Animal Behaviour 62, 209–220.

Wiese, T.R., Rey Planellas, S., Betancor, M., Haskell, M., Jarvis, S., Davie, A., Wemelsfelder, F., Turnbull, J.F., 2023. Qualitative Behavioural Assessment as a welfare indicator for farmed Atlantic salmon (Salmo salar) in response to a stressful challenge. Front. Vet. Sci. 10. https://doi.org/10.3389/fvets.2023.1260090

Winckler, C., 2014. Inter-observer agreement for qualitative behaviour assessment in dairy cattle in three different countries, in: Proceedings of the 6th International Conference on the Assessment of Animal Welfare at the Farm and Group Level. Wageningen Academic, pp. 181–181. https://doi.org/10.3920/9789086867981_149

Table 1: Weather conditions, pasture conditions, observers, hour of observation of qualitative behaviour assessment (QBA) performed on a herd grazing herd (n = 144 dairy cows) and handlers who collect the cows on the pasture plot during the handling context for each observation.

		Weather conditions on the test day							Pasture conditions		Context of observation						
											AM [4]		PM [5]		Handling [6]		
visit	Date	Average temperature (°C)	Minimum temperature (°C)	Maximum temperature (°C)	Rainfall (mm)	Solar radiation (Joule/cm²)	Average wind (m/s)	Average humidity (%)	Days in the pasture plot	Pasture access	Observers	Hour	Observers	Hour	Observers	Hour	Handlers
1	15-Apr.	4.3	-1.5	10	0	1937	2.2	70	4	Day only [1]	A and B	10:35	A and B	15:17	A and B	15:54	
2	18-May.	11.7	7.7	16.5	5.5	2200	1.9	75	2	Day and night			A and B	15:30	A and B	16:10	F and G
2	20-May.	10.8	2.5	17.1	0	1859	2.4	77	4	Day and night	A and B	11:08					
3	21-Jun.	16.5	13.3	22.2	29	1910	1.1	86	4	Day and night			A and E	15:33	A and E	15:45	F and H
3	24-Jun.	14.3	8.9	20.1	0.5	2937	1.3	77	1	After morning milking²			A and E	15:34	A and E	15:42	F and I
3	25-Jun.	14.6	9.6	17.6	10.5	994	1.3	91	2	Day and night	A and E	10:16					
4	28-Jul.	16.5	12.5	21.5	0.5	1823	2	77	5	Day and night	A and D	10:57	A and D	15:29	A and D	15:47	F and J
5	30-Aug.	15.8	11.3	19.5	0	858	1.9	87	3	Day and night					A and D	15:48	G and J
5	31-Aug.	16.6	13.7	21.3	0	1859	2.2	76	4	Day and night	A and E	11:11	A and E	15:30	A and E	15:55	F and K
5	2-Sep.	16.8	10.6	23.6	0	2275	2.1	73	6	Day and night	A and E	10:59	A and E	15:24	A and E	15:48	F and L
5	3-Sep.	18.1	10.6	25.4	0	1929	0.8	77	7	Day and night					A and D	15:50	H, G and J
6	4-Oct.	11.9	7	17.6	8.5	1248	1.8	89	5	After morning milking²					A and D	15:50	M and J
6	5-Oct.	12.1	10.8	15.6	3	1349	2.7	85	6	Day and night					A and D	15:58	H and N
6	6-Oct.	11.1	8.4	16.7	0	1100	1.5	84	7	Day and night	A and B	10:50	A and B	15:32	A and B	15:54	M and J
6	7-Oct.	11.4	5.4	16.6	1	773	0.4	90	8	Day and night					A and B	16:19	H and N
7	2-Nov.	7.2	4.7	10.1	0	720	1	93	2	Day and night					A and D	16:00	M and L
7	3-Nov.	6.7	2.6	11.6	5.5	875	1	92	3	Day and night					A and D	15:54	F and O
7	4-Nov.	6.8	3.8	9.9	9	590	1.4	96	4	Day and night	A and E	10:40	A and E	15:33	A and E	15:48	I and M

[1] Pasture access between morning and evening milking only
[2] Cows spent night indoor and were released on pasture after morning milking
[3] At pasture, right after coming back to pasture from morning milking (AM context, n = 8)
[4] At pasture, before the afternoon milking (PM context, n = 9)
[5] During the gathering of cows at pasture for afternoon milking (handling context, n = 16)

Table 2: Ethogram used to describe cows’ responses to handling (quantitative behaviour assessment)

Response	Description	Reference
Slipping	Lowering of the animal's body by slipping or folding of the leg/legs, resulting in loss of balance, with no part of the body other than the hooves and/or legs in contact with the floor.	Welfare Quality (2009) - Fattening cattle at slaughterhouse
Falling	Toss of balance in which parts of the body other than feet and legs are in contact with floor surface.	Welfare Quality ((2009) - Fattening cattle at slaughterhouse
Moving backwards	The animal moves backwards as a reaction to handling. Moves at least two steps backwards. When an animal takes a few steps backwards to achieve balance or changes position in relation to other animals when crowding it is not considered as moving backwards	Welfare Quality (2009) - Fattening cattle at slaughterhouse; Hultgren et al. (2014)
Turning around	The animal turns around, by itself or as a reaction to the handling regime. When/if the animal turns back again to the former direction, the behaviour should not be recorded again.	Welfare Quality ((2009) - Fattening cattle at slaughterhouse
Vocalizing	The cow makes vocal sound	Grandin (1998) ; Grandin (2001) ; Hultgren et al. (2014)
Galloping	The cow is running with both fore and hindlegs placed forward alternately	Bokkers et al 2008
Trotting	The cow is running with the left foreleg and right hindleg are placed forward simultaneously, followed by the right foreleg and left hindleg etc.	Bokkers et al 2009
Kicking	Kicks with hind leg towards stockperson	Hultgren et al. (2014)
Freezing	The route is free in front or behind the animal but the animal refuses to move forwards or backwards within 4 seconds from being touched/coerced by the handler. If the animal takes more than one step and stops again, or moves backwards, a ‘freeze’ is recorded again when a new driving attempt is made. An animal that stops but continues to walk when the handler drives it forwards is not frozen.	Welfare Quality (2009) - Fattening cattle at slaughterhouse; Hultgren et al. (2014)
Startle	The animal flinches, jumps or bucks in response to stimulus.	Gibbons et al (2009)

Table 3: Measure on the scale in mm for each adjective for QBA performed on video of dairy cows at pasture, loading on PC1 and PC2, weights from the Welfare Quality protocol for dairy cows and weight from the present study, rank of each adjectives in Welfare Quality protocol and from the present study at pasture (according to their absolute weights) (n = 399).

	Pasture context								**Welfare Quality**	
	Measure on the scale (mm)									
variable	**Mean**	**SD**	**min**	**max**	**Loadings on PC1**	**Loadings on PC2**	**Weights (Estimate)**	**Rank***	**Weights (Estimate)**	**Rank**
active	71	29.8	1	123	0.1	**0.84**	0.0014	18	0.0077	16
relaxed	66	34.2	1	125	*0.52*	*-0.52*	0.0061	12	0.01	14
fearful	7	8.9	0	95	*-0.35*	*0.42*	-0.0161	5	-0.0129	7
agitated	35	33	0	123	-0.49	**0.74**	-0.0061	13	-0.0162	3
calm	78	34.1	0	122	0.59	**-0.63**	0.007	11	0.0088	15
content	86	26.2	1	124	**0.82**	0.2	0.0127	6	0.0121	8
indifferent	36	25.4	0	116	*0.24*	*-0.46*	0.0038	14	-0.0112	11
frustrated	12	16.7	0	115	**-0.83**	-0.11	-0.0201	2	-0.0161	4
friendly	31	23.4	1	101	-0.14	**0.63**	-0.0024	16	0.0117	10
bored	13	18.2	0	120	*-0.48*	*-0.4*	-0.0108	8	-0.0109	12
playful	19	24.3	0	116	-0.05	**0.75**	-0.0008	20	0.0011	18
positively occupied	82	30.6	1	125	**0.74**	0.28	0.0099	10	0.0118	9
lively	61	35.7	1	123	0.28	**0.81**	0.0032	15	0.0003	20
inquisitive	41	34.4	0	124	0.08	**0.78**	0.0009	19	0.0005	19
irritable	14	18.4	0	120	**-0.87**	-0.04	-0.0191	3	-0.0218	1
uneasy	12	18.6	0	120	**-0.83**	-0.03	-0.0181	4	-0.0103	13
sociable	30	24.2	0	106	-0.13	**0.67**	-0.0022	17	0.0053	17
apathetic	9	10.2	0	98	*-0.26*	*-0.3*	-0.0103	9	-0.0156	5
happy	87	26.5	0	122	**0.8**	0.21	0.0122	7	0.0147	6
distressed	6	12.7	0	117	**-0.79**	-0.17	-0.0253	1	-0.0203	2

*rank determined from absolute value of weight (1 being the adjective with the maximal absolute weight)
In bold, absolute loadings > 0.6
In italics, absolute loadings never > 0.6 on both axe 1 and axe 2
In grey the ranks that differ of more than 2 ranks between WELFARE QUALITY and pasture ranking

Table 4: Intra- and inter-observer reliability of coordinates on PC1 and PC2 obtained from video observations (W, Kendall's concordance coefficient; ICC, intraclass correlation coefficient).

	PCA axis 1 (PC1)				**PCA axis 2 (PC2)**			
	Session 1		**Session 2**		**Session 1**		**Session 2**	
	ICC	**Kendall**	**ICC**	**Kendall**	**ICC**	**Kendall**	**ICC**	**Kendall**
intra-observer reliability								
Observer								
A	0.81	0.84	0.82	0.86	0.96	0.96	0.90	0.96
B	0.88	0.90	0.91	0.90	0.89	0.93	0.92	0.92
C	0.73	0.68	0.93	0.91	0.86	0.90	0.88	0.94
D	0.61	0.66	0.67	0.76	0.74	0.86	0.86	0.90
E	0.75	0.69	0.95	0.89	0.90	0.94	0.89	0.94
Inter-observer reliability								
Day								
1	0.47	0.56	0.62	0.62	0.72	0.89	0.76	0.89
2	0.52	0.52	0.59	0.67	0.68	0.88	0.72	0.88

597 **Table 5:** Results of the linear mixed models (with estimated coefficients and associated statistics) testing
598 the effect of the context of direct observations on the coordinates on axis 1 (PC1) and axis 2 (PC2) of the
599 Principal Component Analysis. Observers were included as a random effect.

Model for PC1*			**Confidence interval**		
Fixed effects: coefficients of the model	Estimate	SE	2.50%	97.50%	t value
Handling context (Intercept)	-2.1	0.43	-3.0	-1.1	-4.8
AM pasture context	4.1	0.30	3.6	4.8	13.6
PM pasture context	3.5	0.29	3.0	4.1	12.0
Random effects: standard deviations	Variance	SD			
Observer	0.61	0.78	0.24	1.75	
Residuals	0.95	0.97	0.81	1.16	
Model for PC2**			Confidence interval		
Fixed effects: coefficients of the model	Estimate	SE	2.50%	97.50%	t value
PM pasture context (Intercept)	-1.7	0.35	-2.4	-1.0	-4.8
Handling context	2.0	0.35	1.3	2.7	5.8
AM pasture context	2.1	0.40	1.3	2.8	5.2
Random effects: standard deviations	Variance	SD			
Observer	0.16	0.41	0.00	1.05	
Residuals	1.35	1.16	0.97	1.39	

*R Formula: lmer(PC1 ~ context+ (1|observer))

**R Formula: lmer(PC2 ~ context+ (1|observer))

SE, standard error; SD, standard deviation

600

FIGURE CAPTIONS

Fig. 1. Two-dimensional plot of the Principal Component Analysis (PCA) showing the loading of all 20 adjectives on PC1 (Valence) and PC2 (Arousal) from Qualitative Behaviour Assessment of dairy cows at pasture observed on video (n = 399 observations in total, 20 videos watched by n = 5 observers in 4 sessions).

Fig. 1. Coordinates on PC1 and PC2 of the QBA from live observations performed by observer A throughout the grazing season. Dots shape represents the context of observation (□: AM observation at pasture; ×: PM observations at pasture; ▽: observation during herd handling).