© 2015 Universities Federation for Animal Welfare
The Old School, Brewhouse Hill, Wheathampstead,
Hertfordshire AL4 8AN, UK
www.ufaw.org.uk

Animal Welfare 2015, 24: 319-326
ISSN 0962-7286
doi: 10.7120/09627286.24.3.319

Intra-day variation of Qualitative Behaviour Assessment outcomes in dairy cattle

AK Gutmann, B Schwed, L Tremetsberger and C Winckler*

Division of Livestock Sciences, Department of Sustainable Agricultural Systems, University of Natural Resources and Life Sciences, Gregor-Mendel-Str 33, 1180 Vienna, Austria
* Contact for correspondence and requests for reprints: anke.gutmann@boku.ac.at

Abstract

Qualitative Behaviour Assessment of cattle expression using a fixed rating scale of 20 descriptors is one of the measures of the Welfare Quality® (WQ) assessment protocol for dairy cattle. As for other on-farm measures of welfare, reliability is an important issue especially if farms are to be certified. This study investigated the repeatability of QBA results across three different observation times during the day (early morning, late morning, early afternoon). For this purpose, 13 observers assessed a total of 30 video clips from ten commercial dairy farms using visual analogue scales to score the 20 QBA terms. QBA scores for 'emotional state' were computed according to the Welfare Quality® protocol (WQ_QBA) and, additionally, a Principal Component Analysis was carried out. The latter revealed two main dimensions which may be described as 'mood' and 'activity', the former thus corresponding to the 'emotional state' score of the WQ protocol. Both for scores derived from the WQ protocol and from PCA, mixed model analysis for repeated measures revealed a significant effect of observation time depending on the farm. Mixed model analysis for repeated measures revealed a significant effect of observation time for three farms out of ten on both the WQ_QBA score and the PCA 'mood' dimension; a similar effect was found for eight out of ten farms for the PCA 'activity' dimension. These results indicate that observation time potentially affects WQ (and other QBA) outcomes on a proportion of farms. However, given that outcomes for WQ_QBA and PCA 'mood' were consistent for the majority of farms, procedures suggested in the Welfare Quality® protocol may constitute a reasonable compromise between reliability and feasibility. If the QBA assessment should reflect the 'mean mood', multiple assessments throughout the day may be carried out.

Keywords: *animal welfare, dairy cattle, observation time, Qualitative Behaviour Assessment, reliability, Welfare Quality®*

Introduction

Farm animal welfare is becoming an important issue among European consumers (European Commission 2007). In order to accommodate these societal concerns substantial efforts have been undertaken to produce a reliable system to assess animal welfare on-farm that is based on scientific findings (Blokhuis *et al* 2008). The assessment protocols that were developed within the framework of the EU project Welfare Quality® (WQ) provide such a system for the assessment of animal welfare on-farm. The WQ approach is based on four principles (Good feeding, Good housing, Good health, Appropriate behaviour) and, in total, twelve criteria allocated to these principles. Besides a number of quantitative, mostly animal-based measures, Qualitative Behaviour Assessment (QBA) also forms part of the protocol (Wemelsfelder *et al* 2009c). QBA is the only measure that is linked to the WQ criterion 'positive emotional state' (Welfare Quality® 2009a). The qualitative assessment relies on the ability of human observers to integrate perceived behavioural details into descriptions of an animal's 'body language', using descriptors such as 'relaxed', 'fearful' or 'content' (Wemelsfelder *et al* 2009c). A large number of studies covering different species, eg poultry (*Gallus gallus domesticus*): Wemelsfelder *et al* (2009a); horses (*Equus caballus*): Fleming *et al* (2013); pigs (*Sus scrofa*): Wemelsfelder *et al* (2009b); buffalo (*Bubalus bubalis*): Napolitano *et al* (2012); beef cattle (*Bos primigenius taurus*): Stockman *et al* (2012); Wemelsfelder *et al* (2009c) provides strong evidence that observers consistently distinguish expressive behavioural patterns into dimensions from positive to negative mood, and from low to high arousal within these moods. This first component, in particular, thus provides integrated information which is directly relevant to emotional experience and thus to animal welfare. Qualitative terms describing patterns of behaviour and emotional experience have been used before, eg in the field of animal temperament and personality research (for examples of a review, see Uher & Asendorpf 2008; Meagher 2009). However, the assessment

https://doi.org/10.7120/09627286.24.3.319 Published online by Cambridge University Press

of those qualities is based mainly on quantitative measures such as latency, frequency, intensity or duration of certain behaviours. In contrast, QBA aims at taking the expressive emotional component of behaviour into account (eg for cattle social behaviour: Rousing & Wemelsfelder 2006). Developing and implementing QBA, Wemelsfelder *et al* (2000, 2001) demonstrated its reliability regarding various critical factors, such as inter- and intra-observer agreement, contextual bias with regard to environmental background (Wemelsfelder *et al* 2009[d]), and validity in terms of correlation with quantitative behavioural and physiological measures (eg Rousing & Wemelsfelder 2006; Minero *et al* 2009; Stockman *et al* 2011; Stockman *et al* 2012). The above-mentioned studies used a Free Choice Profiling approach, which allows each single observer to create and use his or her own descriptors for scoring. However, analysis requires a data set of several observers. To better meet the requirements of on-farm welfare assessment schemes in terms of work and labour, fixed terms lists have been developed on the basis of FCP results (Wemelsfelder *et al* 2009[d] for cattle). To the knowledge of the authors there is no published study aimed at the direct comparison of QBA outcomes using FCP versus fixed terms lists, however the analysis of own unpublished data of QBA using video clips of dairy cow herds revealed that both the dimensions or components of behavioural expression as well as the actual scores for the herds on these dimensions were highly comparable (correlations of dimensions 0.96–0.67, all $P < 0.01$). Especially if used for certification purposes, welfare assessment systems should be sufficiently reliable with regard to, eg intra- and inter-observer agreement or test-retest reliability (Knierim & Winckler 2009), ie results should be independent of when and by whom the assessment is carried out. Intra-day variation is especially relevant for behavioural measures of welfare, eg social behaviour (Winckler *et al* 2002). Concerning reliability of QBA using fixed terms lists, recent studies yielded both promising and rather critical results. During the developmental phases within the WQ® project, observer agreement was good concerning dimensions of emotional expression as well as scores on these dimensions both using FCP and fixed terms' lists in pigs (Wemelsfelder *et al* 2009b), poultry (Wemelsfelder *et al* 2009a), and beef cattle (Wemelsfelder *et al* 2009c), however, only moderate using fixed terms' lists in dairy cattle (Wemelsfelder *et al* 2009c). Phythian *et al* (2013) found good inter-observer agreement independent of the professional background of observers (veterinary students and surgeons versus farm assurance inspectors) using QBA fixed-terms' lists for a video-based assessment of emotional expression in sheep (*Ovis aries*). Bokkers *et al* (2012) reported only moderate inter-observer agreement for QBA scores as derived from the WQ protocol when assessing videos of dairy cow herds (clips made of four shots of 30 s each), however independent of the level of experience. Intra-observer reliability of QBA WQ-scores in the study of Bokkers *et al* (2012) using repeated video assessment (ten months between assessments) was good in terms of high pair-wise correlations, however a pair-wise *t*-test revealed significant differences in four out of eight observers. In a study on the suitability of QBA as a stand-alone integrative screening tool for animal welfare assessment Andreasen *et al* (2013) reports high inter-observer agreement between two trained QBA assessors, but no satisfying correlations between them and a third observer who carried out QBA at a different time as a part of a full WQ assessment. Finally, Temple *et al* (2013) investigated the test-retest reliability of QBA and other welfare measures over time, on-farm in pigs (one year between visits, following the WQ assessment protocol for pigs; Welfare Quality® 2009b). Correlations over time were moderate, but QBA scores differed significantly between the assessments. Since a number of welfare measures differed after one year it cannot be distinguished whether these differences resulted from intra-observer reliability problems or changes in the welfare state of the pigs.

According to the WQ protocol, QBA is scheduled following the morning milking in order to ensure observers to be least influenced by other quantitative measures taken in the course of the farm visit (Welfare Quality® 2009a). However, cattle behaviour is subject to a circadian rhythm. Cattle on pasture show the highest activity at dawn and dusk (Winckler 2009; p 89) whereas the late morning and the early afternoon are used for resting and ruminating (Houpt 2011; p 80). This typical distribution of active and less active periods can also be observed in indoor housing when *ad libitum* feeding is provided (Winckler 2009; p 89); times of milking and delivery of fresh feed then usually serve as pacemakers (DeVries *et al* 2003). The pattern of the expressive quality of behaviour in the course of the day is however not known. It was therefore the aim of this study to evaluate whether QBA assessments at various observation times during the day lead to different results concerning the qualitative evaluation of dairy cow behaviour. We were also interested if effects differ depending on whether data are computed according to the WQ protocol or independently analysed using a Principal Component Analysis.

Materials and methods

Study animals and housing

Ten private dairy farms located in lower and upper Austria were visited once for on-farm welfare assessment between November 2011 and January 2012. All farms had loose-housing systems with deep-littered cubicles, and herd size ranged from 27 to 38. The most prevalent breed was Austrian Fleckvieh (eight farms), and one farm each kept Holstein-Friesian and Brown Swiss cows. All herds were zero-grazed and the feed ration contained grass silage or grass and maize silage, hay and concentrates. Concentrate dispensers were present in five farms, two farms fed a TMR and three farms manually fed concentrates at the feed bunk. In four farms at least part of the floor in the alleys was slatted and two herds had access to an outdoor run. Cows were milked twice daily in a milking parlour.

https://doi.org/10.7120/09627286.24.3.319 Published online by Cambridge University Press

Video recordings

Behaviour of dairy cows was video-recorded using a Panasonic HDC-SD99 HC Camcorder (Panasonic, Tokyo, Japan) at three different observation times: early morning (following the morning milking; ± 0800h); late morning (± 1100h); and early afternoon (± 1300h). Following the WQ protocol, the pen holding the lactating cows was divided into four to six observation segments to cover the complete area evenly (Welfare Quality® 2009a). At all three observation times, the segments were video-recorded consecutively for a total of 20 min with the duration per segment depending on the number of segments (ie 20 min divided by number of segments). The camera was always installed at the same pre-defined positions on an extensible tripod.

For the QBA assessment sessions, four-minute video clips for each of the ten farms and each of the three observation times were created which comprised 2-min recordings from two out of the four to six barn segments. The final video footage thus comprised 30 4-min video clips. The selection of the segments chosen for these clips was based on three principles:

- Minimum number of five animals visible at the beginning of the video recording;
- All barn areas covered, ie feeding places, alleys, concentrate dispensers and lying areas; and
- No overlap of two consecutive segments.

The behaviour of the animals displayed in the video recordings was not taken into account when selecting the clips. From each of the selected segment recordings, the first 30 s were discarded and the following 2 min were used. The same two segments were used for all three observation times. To check how representative the selected clips were as compared to the full information of the 20-min video recordings, we compared the latter with the selected 4-min clips regarding number of animals present as well as percentage of animals lying, standing and feeding (scan sampling with 1-min scan interval). Pearson's correlations indicated that we had generated a representative sub-sample (Pearson's r for early morning video clips: 0.88, 0.92, 0.87, and 0.67; for late morning clips: 0.94, 0.96, 0.47, and 0.90; for early afternoon clips: 0.95, 0.93, 0.46 and 0.93, respectively).

Observers and assessment sessions

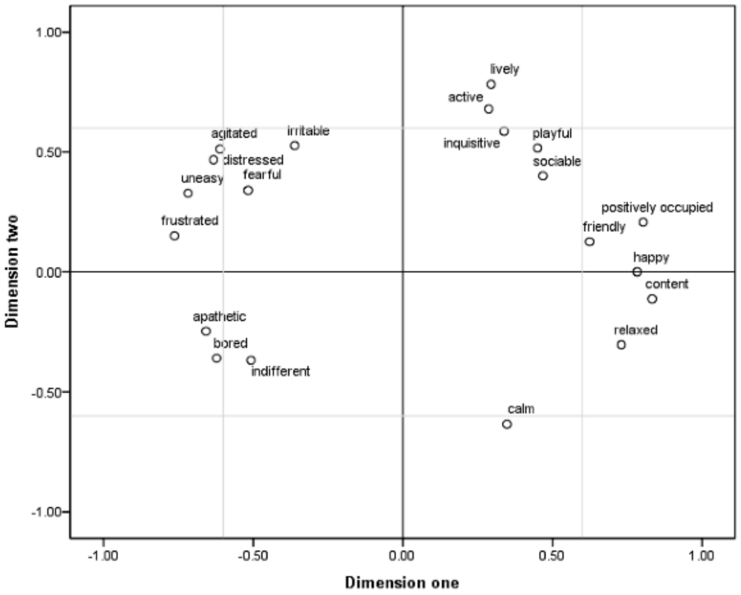
A group of ten female and three male observers (scientists in the field of animal husbandry or MSc students of agricultural/livestock science) assessed the 30 video clips in two sessions. The observers were told that the study aimed at assessing the emotional state of dairy herds under commercial housing conditions and that farms may repeatedly be shown, but they were blind to the actual goal (ie effect of observation time during the day) and design of the study. They were all familiar with the basics of cattle behaviour. The observers underwent a 1-h training session consisting of a general introduction to QBA and an explanation of the WQ QBA list of 20 pre-determined terms (Wemelsfelder *et al* 2009c). Based on video clips for training purposes, all terms (active, relaxed, fearful, agitated, calm, content, indifferent, frustrated, friendly, bored, playful, positively occupied, lively, inquisitive, irritable, uneasy, sociable, apathetic, happy, distressed) were discussed until a common basic understanding of their meaning was achieved. The assessment sheets consisted of visual analogue scales (VAS) of 125-mm length for each for the 20 terms with the far left and right defined as 'minimum' and 'maximum', respectively. 'Minimum' means that the expressive quality indicated by this term is entirely absent in any of the animals observed while 'maximum' indicates it is pervasively dominant across the whole group of animals. The observers scored each clip after watching it completely by ticking the scale for each of the 20 given terms at a point they perceived as appropriate. Clips were shown in colour and with sound and in a randomised order to ensure unbiasedness concerning observation time.

Data processing and statistical analysis

VAS values for each of the 20 fixed terms were determined by measuring the distance in mm from the left side (minimum) to the point where the observer ticked the line. A weighted sum approach to integrate the single QBA ratings into WQ-QBA scores (possible range 0–100) was applied according to the WQ protocol. The weighting coefficients used for this aggregation had been derived from a reference data set using Principal Component Analysis (PCA); they are related to the loadings of the terms on the components or dimensions (Welfare Quality® 2009a). Since these loadings depend on the underlying data set, additionally a PCA on the raw data set was carried out to examine whether similar components, loadings of terms and corresponding scores per farm are obtained from PCA (PCA_QBA scores). PCA was computed once on the complete data set (390 cases) based on correlation matrices. The initial solution (unrotated components) was used and the first two factors with Eigenvalues greater than one were further considered. Terms with loadings ≥ 0.6 were considered to describe the meaning of the dimensions.

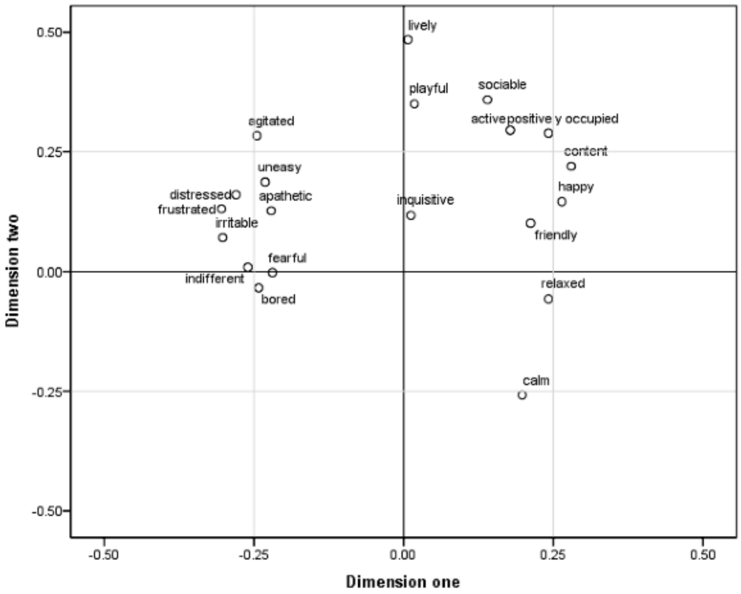
Score data were suitable for parametric statistics (Kolmogorov-Smirnov test for normality on residuals $P > 0.05$ for WQ_QBA and PCA_QBA scores, respectively). Effects of observation time on WQ_QBA scores and PCA_QBA scores were analysed using a mixed model for repeated measures including 'farm ID', 'observation time', and their interaction term, where 'observation time' was specified as repeated measures factor and 'farm ID' as group factor to account for random intercepts and to indicate that data are correlated on the same farm while permitting different covariance structures. To evaluate differences at farm-level, *post hoc* contrast tests between scores of different observation times were carried out for each farm applying Bonferroni adjustment for multiple pair-wise comparisons. Kendall's coefficient of concordance (W-coefficient) was used to determine inter-observer reliability for each observation time. Statistical analyses were done in SPSS (version PASW Statistics 18) and SAS (version 9.2). The alpha level was set to 0.05.

Figure 1



PCA loading diagram for the 20 pre-defined fixed terms of the Welfare Quality® protocol analysed using Principal Component Analysis on the raw data set comprising ratings of 13 observers for all ten farms at three different observation times.

Figure 2



Loading plot of the 20 pre-defined terms analysed in the course of development of the Welfare Quality® protocol (adapted from F Wemelsfelder, personal communication 2013). Weights of the weighted-sum approach according to the Welfare Quality® protocol are based on these loadings (Welfare Quality® 2009).

Results

Dimensions derived from PCA

QBA raw data were suitable for PCA (Kaiser-Mayer-Olkin 0.878). PCA revealed two main dimensions explaining 35 and 18% of variance, respectively. The first dimension was described by the terms 'content', 'positively occupied', 'happy', 'relaxed', 'friendly' (positive loadings) and 'frustrated', 'uneasy', 'apathetic', 'distressed', 'bored' (negative loadings) and denominated as 'mood'. The second dimension 'activity' was defined by the terms 'lively', 'active', 'inquisitive' (positive loadings) and 'calm' (negative loading) (Figure 1). Loading diagrams of WQ_QBA and PCA_QBA dimensions resembled each other noticeably (Figures 1 and 2), and loadings on the single terms were highly correlated (Pearson's $r = 0.96$ and 0.75 for dimension one and two, respectively).

Inter-observer reliability (IOR)

For early morning, late morning and afternoon data, Kendall's W for WQ_QBA scores were 0.37, 0.28, and 0.46, respectively. For PCA_QBA scores on the dimension 'mood' Kendall's W were 0.37, 0.29 and 0.50, and for scores on the dimension 'activity' 0.53, 0.57, and 0.50 for early morning, late morning and afternoon data, respectively.

https://doi.org/10.7120/09627286.24.3.319 Published online by Cambridge University Press

Table 1 Results of mixed model analysis. Effects of 'farm ID', 'observation time' and their interaction term on WQ_QBA and PCA_QBA scores.

Analysis	Model factors	df	*F*-value	*P*-value
WQ_QBA score	Farm ID	9	7.3	< 0.001
	Observation time	2	13.8	< 0.001
	Observation time × Farm ID	18	1.7	0.022
PCA_QBA score, dimension 'mood'	Farm ID	9	8.6	< 0.001
	Observation time	2	9.6	< 0.001
	Observation time × Farm ID	18	1.9	0.020
PCA_QBA score, dimension 'activity'	Farm ID	9	11.3	< 0.001
	Observation time	2	49.2	< 0.001
	Observation time × Farm ID	18	5.4	< 0.001

Table 2 Least square means (± SEM) of WQ_QBA and PCA_QBA scores for the ten farms at three different observation times based on mixed model analyses.

	WQ_QBA scores			PCA_QBA 'mood' scores			PCA_QBA 'activity' scores		
	EM	**LM**	**AF**	**EM**	**LM**	**AF**	**EM**	**LM**	**AF**
Overall	27.9 (± 1.49)	38.6 (± 1.49)	30.8 (± 1.49)	–0.16 (± 0.08)	0.28 (± 0.08)	–0.11 (± 0.08)	0.52 (± 0.07)	–0.39 (± 0.07)	–0.12 (± 0.07)
Farm ID 1	45.7 (± 5.60)	38.1 (± 5.60)	38.3 (± 5.60)	0.79 (± 0.27)	0.23 (± 0.27)	0.22 (± 0.27)	**0.65 (± 0.22)**^{a}	**–0.53 (± 0.22)**^{b}	**–0.74 (± 0.22)**^{b}
Farm ID 2	**17.5 (± 4.61)**^{a}	**42.4 (± 4.61)**^{b}	**25.7 (± 4.61)**^{a}	**–0.75 (± 0.26)**^{a}	**0.33 (± 0.26)**^{b}	**–0.40 (± 0.26)**^{ab}	**0.73 (± 0.22)**^{a}	**–1.13 (± 0.22)**^{b}	**–0.14 (± 0.22)**^{c}
Farm ID 3	34.5 (± 5.18)	40.4 (± 5.18)	42.5 (± 5.18)	0.29 (± 0.24)	0.61 (± 0.24)	0.43 (± 0.24)	**0.92 (± 0.22)**^{a}	**0.80 (± 0.22)**^{a}	**–0.21 (± 0.22)**^{b}
Farm ID 4	19.4 (± 3.39)	25.6 (± 3.39)	21.1 (± 3.39)	–0.79 (± 0.21)	–0.53 (± 0.21)	–0.72 (± 0.21)	–0.66 (± 0.20)	–0.87 (± 0.20)	–0.80 (± 0.20)
Farm ID 5	31.7 (± 5.02)	33.5 (± 5.02)	17.9 (± 5.02)	**0.10 (± 0.26)**^{a}	**0.16 (± 0.26)**^{a}	**–0.90 (± 0.26)**^{b}	**0.82 (± 0.23)**^{a}	**0.52 (± 0.23)**^{a}	**–0.80 (± 0.23)**^{b}
Farm ID 6	**16.6 (± 3.92)**^{a}	**43.4 (± 3.92)**^{b}	**28.9 (± 3.92)**^{c}	**–0.76 (± 0.19)**^{a}	**0.48 (± 0.19)**^{b}	**–0.14 (± 0.19)**^{b}	**0.32 (± 0.19)**^{a}	**–0.46 (± 0.19)**^{b}	**–0.29 (± 0.19)**^{ab}
Farm ID 7	30.5 (± 4.82)	37.0 (± 4.82)	30.2 (± 4.82)	–0.07 (± 0.24)	0.22 (± 0.24)	–0.17 (± 0.24)	–0.01 (± 0.20)	–0.71 (± 0.20)	**–0.27 (± 0.20)**
Farm ID 8	**19.4 (± 4.33)**^{a}	**35.5 (± 4.33)**^{b}	**17.7 (± 4.33)**^{a}	–0.44 (± 0.25)	0.11 (± 0.25)	–0.73 (± 0.25)	**1.42 (± 0.24)**^{a}	**–0.27 (± 0.24)**^{b}	**1.01 (± 1.24)**^{a}
Farm ID 9	27.4 (± 5.13)	39.9 (± 5.13)	34.5 (± 5.13)	–0.17 (± 0.27)	0.28 (± 0.27)	0.33 (± 0.27)	**0.57 (± 0.20)**^{a}	**–0.82 (± 0.20)**^{b}	**0.64 (± 0.20)**^{a}
Farm ID 10	33.9 (± 5.44)	50.5 (± 5.44)	49.2 (± 5.44)	0.19 (± 0.27)	0.88 (± 0.27)	0.94 (± 0.27)	**0.39 (± 0.18)**^{a}	**–0.45 (± 0.18)**^{b}	**0.36 (± 0.18)**^{a}

EM: early morning; LM: late morning; AF: early afternoon. Within-farm significant differences are highlighted in bold, different superscripts within rows and parameters indicate significant differences ($P < 0.05$; Bonferroni adjustment for multiple pairwise comparisons).

Effect of observation time on QBA results

Single WQ_QBA scores per video clip ranged from 0 to 78 in the early morning, from 0 to 90 in the late morning and from 0 to 88 in the early afternoon. PCA_QBA scores per video clip for 'mood' ranged from –2.63 to 2.49 in the early morning, from –2.36 to 2.69 in the late morning, and from –2.99 to 2.83 in the early afternoon, and for the dimension 'activity' from –2.10 to 2.74, from –2.33 to 2.13, and from –2.60 to 3.55, respectively.

'Observation time' and 'farm ID' significantly affected all three QBA scores (Table 1). These effects were modified by a disordinal interaction effect of 'observation time' and 'farm ID' and thus are not independently interpretable at a global level. Least square means for the different farms and observation times are provided in Table 2. In three farms (Farm 2, 6, and 8), WQ_QBA scores were significantly higher in the late morning, ie the emotional state of the animals was perceived to be better than in the early morning and the early afternoon while no significant differences were found for the other farms. Differences in mean scores per farm over observation time ranged from < 1 up to 27 (Table 2, eg farm 1 and 6).

In two of the three farms showing significant effects as regards WQ_QBA scores (Farm 2 and 6), PCA_QBA scores for the dimension 'mood' during late morning were significantly higher, ie 'mood' was perceived to be better as compared to early morning. In a further farm (Farm 5) PCA_QBA scores were higher during the morning assessments as compared to the early afternoon. Numerically, mean scores of six farms in total ranged from positive to negative values, ie from a rather 'positive mood' to a rather 'negative mood' (Table 2).

https://doi.org/10.7120/09627286.24.3.319 Published online by Cambridge University Press

PCA_QBA scores for the dimension 'activity' differed significantly in eight out of ten farms depending on observation time. In six farms, 'activity' scores were significantly lower in the late morning than in the early morning, ie cows were perceived as less active in the late morning. This was followed by an increase in 'activity' in the early afternoon in most farms (significant for four farms). However, 'activity' scores in the early afternoon did reach early morning levels in only four farms (Table 2).

Discussion

The aim of the present study was to test the intra-day effect on QBA outcomes regarding the assessment of the emotional state in dairy cattle. This is the first such examination and adds to previous work on reliability testing of QBA (Rousing & Wemelsfelder 2006; Wemelsfelder *et al* 2009d; Bokkers *et al* 2012; Andreasen *et al* 2013).

We used video clips which cannot truly replace on-farm assessments. Apart from the duration of observations (20 min for on-farm assessments according to WQ [Welfare Quality® 2009a] vs 4-min video clips in the present study), impressions such as smell or the general atmosphere are not covered, and the selected scenes might not be representative. The latter might be particularly important for the present study design, since the once-chosen segments were shown repeatedly over the different observation times without checking again for, eg number of animals present. However, correlations between full 20-min video recordings and the selected 4-min clips were high, indicating a representative sub-sample. Furthermore, given the specific research question, the benefits of video assessment may outweigh disadvantages: it allows the same situation to be assessed by a larger group of observers, and by randomising clip order, a bias due to recognising farm and observation time can be avoided. To achieve a balance between feasibility and validity, longer video clips were used as compared to other QBA studies (eg Rousing & Wemelsfelder 2006; Wemelsfelder *et al* 2009d; Rutherford *et al* 2012), and clips contained recordings of two different segments of the barn simulating the on-farm assessment situation where assessors also change their position (Welfare Quality® 2009a). In the video assessment as well as in the on-farm assessment, observers are challenged by integrating the information gained in different segments into one single rating.

First, we were interested whether the pre-defined term-loadings of the emotional dimension in WQ (as determined by the WQ farm sample) would affect results concerning intra-day variation as compared to PCA on raw data. It was assumed that PCA would lead to different results by reflecting the observers' definition of dimensions more adequately. However, the first dimension of PCA resembled the WQ dimension closely: as in WQ, the PCA dimension explaining most of the variation distinguished between positive and negative mood which bears direct relevance to the assessment of the WQ criterion 'Positive emotional state'. The second dimension differentiated between high and low activity which in itself is not necessarily related to welfare and therefore is considered less meaningful in WQ (Wemelsfelder *et al* 2009c). The WQ definition of the mood-dimension could be reproduced using PCA, indicating that the WQ QBA analysis is suitable to generally reflect observers' perception of the terms.

Inter-observer reliability for the QBA outcomes was low to moderate in the observer group. Correlation coefficients of 0.7, which is referred to as a threshold for an acceptable correlation coefficient for inter-observer reliability (Martin & Bateson 2007; p 51), were not obtained. This is in accordance with the results of Wemelsfelder *et al* (2009c) and Bokkers *et al* (2012) who found low inter-observer reliabilities for individual descriptors in dairy cattle assessment. Although it is important of course to reach an acceptable inter-observer reliability for a truly reliable measurement, it was of secondary importance for our research question since we were interested in the agreement on intra-day differences rather than in agreement on evaluation in absolute terms. Furthermore, detecting consistent effects of observation time despite considerable variation between observers rather underlines the findings.

The underlying data set included ten rather similar farms, with no apparent differences, eg in size or husbandry system. This could have negatively influenced observer agreement due to a rather narrow reference bandwidth — agreement at the extreme ends of any dimension is often easier to achieve. However, scores ranged nearly over the whole WQ_QBA scale indicating that observing the emotional expression of the animals was sufficient to differentiate between farms.

Mixed model analysis revealed similar results concerning main and interaction effects of 'observation time' and 'farm ID' for both WQ_QBA and PCA_QBA scores. For the scores related to 'emotional state' (WQ-QBA) and 'mood' (PCA_QBA), significant differences depending on observation time were found in three farms each, with two of the farms showing differences in both scores. In all four farms for which statistical analysis at least once revealed significant effects, the pattern of scores was similar. Concerning the mean difference between different observations times over all farms, the magnitude of the effect might be tolerable: numerical differences do not necessarily imply different meanings, as long as scores stay within the same broader interpretive range. However, for farms where significant effects were found, the difference is qualitatively important and meaningful since, for example, in the case of WQ_QBA scores changes amounted to up to a quarter of the scale (eg 27 out of 100, farm 6). Similarly, concerning PCA_QBA scores, farms changed from 'better than mean' (ie positive scores) to 'worse than mean' (negative scores) depending on observation time.

Scores for 'activity' differed significantly in eight out of ten farms. Except that in all farms activity-scores were highest in the morning (differing significantly or not), no consistent pattern was found. In contrast to the emotional state, activity in itself does not explicitly refer to the welfare state. However, the diverse patterns of activity scores across farms might reflect the influence of management routines on the

https://doi.org/10.7120/09627286.24.3.319 Published online by Cambridge University Press

diurnal rhythm of the animals (DeVries & von Keyserlingk 2005). While for the first observation time 'early morning' the situation for the animals is comparable across farms (feed delivery directly after morning milking), the situation might differ considerably a few hours later (eg concerning daily management routines, or veterinary treatments), and thus might also affect the emotional state. Since PCA produces orthogonal dimensions, 'mood' and 'activity' were evaluated independently by the observers. Animals could, for example, have been perceived 'active' and in a 'good mood' as well as 'calm' in a 'bad mood'. Still, the differing activity patterns over farms may indicate the occurrence of different events which may in turn influence the mood of the animals. Indeed, there was no consistent pattern regarding 'emotional state'/'mood' scores. However, only one farm achieved the highest scores on 'emotional state'/'mood' in the 'early morning' assessment as compared to the other observation times. Thus, QBA assessment carried out only once and only in the early morning might often underestimate the 'mean-mood' at the farm, but on the other hand probably leads to least variation in the underlying situation.

The aim of the WQ protocol is to provide a standardised instrument for the on-farm assessment of dairy farms (Blokhuis *et al* 2008). Following the prescribed schedule, QBA is carried out as the second measure of the protocol (right after morning milking and following the assessment of 'Avoidance distance at the feeding rack') to minimise observer bias through assessing other WQ animal and resource based measures. At the same time this is meant to ensure least disturbance to the animals. For on-farm QBA assessments, based on the results of the present study, there seems to be a trade-off between observer bias and representativeness of results. For some farms, repeated QBA assessments during the farm visit would most likely better reflect the emotional state on a farm than a single one, but the risk of being biased through recording of other animal-based measures would inevitably increase. A single QBA assessment might misestimate the average mood of animals at single farms, but observers are unbiased and results more comparable due to the more uniform situation after morning milking. The issue of balancing representativeness and feasibility applies to all behavioural measures of welfare in the WQ protocol. To our knowledge it has only been addressed in a small-scale study on social behaviour (Winckler *et al* 2002). Intra- as well as between-day variation was high, and preference was given to a standardised situation (after morning feeding) because it best reflected the mean incidence of social interactions over the day as well as over consecutive days. If a representative evaluation of the dairy cows' emotional state is aspired, on the basis of our results, repeated QBA assessment at various times during the day may be favourable. QBA assessment carried out only in the early morning can rather allow making a statement about the emotional state at the specific time-period when QBA is carried out.

Concerning QBA studies, priority should be given to comparability of the assessed situation to achieve optimal preconditions for acceptable repeatability. For example, observation times varied considerably when Andreasen *et al* (2013) investigated the potential of QBA as a stand-alone integrative screening tool for identifying farms with compromised welfare. QBA assessments were carried out between 0940 and 1645h by two assessors who only performed the qualitative assessments, whereas a full welfare assessment by another assessor using the WQ protocol for dairy cattle started between 0415 and 0900h. Inter-observer reliability between the two QBA-assessors who observed simultaneously was highly significant, but no significant correlation of the QBA scores between the two QBA-assessors and the WQ-assessor was found. According to our results, this lack of agreement may have resulted from different observation times, ie observed results could merely reflect changes in the observed situation. Thus, the present study strongly supports the need to retain constant observation times in QBA reliability studies.

Animal welfare implications

The present study tackled the challenge of improving the measurement of animals' emotional state using the WQ assessment system. The results can help improve the reliability of on-farm assessments for valid statements regarding the welfare state of a certain farm. This is important for the farmers' trust in the assessment and further improvement of on-farm assessment can build upon it.

Conclusion

The results of this study show that for some farms QBA outcomes may vary significantly depending on the time of day the assessment is carried out. For three out of ten farms, this effect was obtained for both the WQ_QBA score for emotional state and scores on a similar dimension 'mood' derived from PCA. However, a standardised observation time, as suggested in the WQ protocol, allows observing the animals in a similar situation (ie after feed delivery) thus ensuring high comparability between farms. The procedure reveals consistent results for the majority of farms and may therefore constitute a reasonable compromise between reliability and feasibility. If the QBA assessment should reflect potential changes in mood throughout the day, multiple assessments may be necessary. The number and timing of such comprehensive assessments requires further investigation.

Acknowledgements

For statistical support we thank the Centre for Experimental Design, Institute of Applied Statistics and Computing (IASC), Department of Landscape, Spatial and Infrastructure Sciences, University of Natural Resources and Life Sciences (BOKU), Vienna. We would like to thank all farms for participating in the study and for the warm tea when we were freezing during the observation. Many thanks also go to the observers who invested their precious time in this study. Last but not least we extend our appreciation to all the cows that were substantially involved in this work.

References

Andreasen SN, Wemelsfelder F, Sandøe P and Forkman B 2013 The correlation of Qualitative Behaviour Assessments with Welfare Quality® protocol outcomes in on-farm welfare assessment of dairy cattle. *Applied Animal Behaviour Science 143*: 9-17. http://dx.doi.org/10.1016/j.applanim.2012.11.013

Blokhuis HJ, Keeling LJ, Gavinelli A and Serratosa J 2008 Animal welfare's impact on the food chain. *Trends in Food Science and Technology 19*: 79-87. http://dx.doi.org/10.1016/ j.tifs.2008.09.007

Bokkers EAM, de Vries M, Antonissen ICMA and de Boer IJM 2012 Inter-and intra-observer reliability of experienced and inexperienced observers for the Qualitative Behaviour Assessment in dairy cattle. *Animal Welfare 21*: 307-318. http://dx.doi.org/10.7120/09627286.21.3.307

DeVries TJ and von Keyserlingk MAG 2005 Time of feed delivery affects the feeding and lying patterns of dairy cows. *Journal of Dairy Science 88*: 625-631. http://dx.doi.org /10.3168/jds.S0022-0302(05)72726-0

DeVries TJ, von Keyserlingk MAG and Beauchemin KA 2003 Short communication: Diurnal feeding pattern of lactating dairy cows. *Journal of Dairy Science 86*: 4079-4082. http://dx.doi.org/10.3168/jds.S0022-0302(03)74020-X

European Commission 2007 Attitudes of EU citizens towards animal welfare. *Special Eurobarometer 270*. http://ec.europa.eu/ public_opinion/archives/ebs/ebs_270_en.pdf

Fleming PA, Paisley CL, Barnes AL and Wemelsfelder F 2013 Application of Qualitative Behavioural Assessment to horses during an endurance ride. *Applied Animal Behaviour Science 144*: 80-88. http://dx.doi.org/10.1016/j.applanim.2012.12.001

Houpt KA 2011 *Domestic Animal Behaviour for Veterinarians & Animal Scientists*. Wiley-Blackwell: Ames, USA

Knierim U and Winckler C 2009 On-farm welfare assessment in cattle: validity, reliability and feasibility issues and future perspectives with special regard to the Welfare Quality® approach. *Animal Welfare 18*: 451-458

Martin P and Bateson P 2007 *Measuring Behaviour. An Introductory Guide*. Cambridge University Press: Cambridge, UK. http://dx.doi.org/10.1017/CBO9780511810893

Meagher RK 2009 Observer ratings: Validity and value as a tool for animal welfare research. *Applied Animal Behaviour Science 119*: 1-14. http://dx.doi.org/10.1016/j.applanim.2009.02.026

Minero M, Tosi MV, Canali E and Wemelsfelder F 2009 Quantitative and qualitative assessment of the response of foals to the presence of an unfamiliar human. *Applied Animal Behaviour Science 116*: 74-81. http://dx.doi.org/10.1016/j.applanim.2008.07.001

Napolitano F, De Rosa G, Grasso F and Wemelsfelder F 2012 Qualitative behaviour assessment of dairy buffaloes (*Bubalus bubalis*). *Applied Animal Behaviour Science 141*: 91-100. http://dx.doi.org/10.1016/j.applanim.2012.08.002

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

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

Rutherford KMD, Donald RD, Lawrence AB and Wemelsfelder F 2012 Qualitative Behaviour Assessment of emotionality in pigs. *Applied Animal Behaviour Science 139*: 218-224. http://dx.doi.org/10.1016/j.applanim.2012.04.004

Stockman CA, Collins T, Barnes AL, Miller DW, Wickham SL, Beatty DT, Blache D, Wemelsfelder F and Fleming PA 2011 Qualitative behavioural assessment and quantitative physiological measurement of cattle naïve and habituated to road transport. *Animal Production Science 51*: 240. http://dx.doi.org/10.1071/AN10122

Stockman CA, McGilchrist P, Collins T, Barnes AL, Miller D, Wickham SL, Greenwood PL, Cafe LM, Blache D, Wemelsfelder F and Fleming PA 2012 Qualitative Behavioural Assessment of Angus steers during pre-slaughter handling and relationship with temperament and physiological responses. *Applied Animal Behaviour Science 142*: 125-133. http://dx.doi.org/10.1016/j.applanim.2012.10.016

Temple D, Manteca X, Dalmau A and Velarde A 2013 Assessment of test–retest reliability of animal-based measures on growing pig farms. *Livestock Science 151*: 35-45. http://dx.doi.org/10.1016/j.livsci.2012.10.012

Uher J and Asendorpf JB 2008 Personality assessment in Great Apes: Comparing ecologically valid behavior measures, behavior ratings, and adjective ratings. *Journal of Research in Personality 42*: 821-838. http://dx.doi.org/10.1016/j.jrp.2007.10.004

Welfare Quality® 2009a *Welfare Quality® Assessment Protocol for Cattle*. Welfare Quality Consortium: Lelystad, The Netherlands

Welfare Quality® 2009b *Welfare Quality® Assessment Protocol for Pigs*. Welfare Quality Consortium: Lelystad, The Netherlands

Wemelsfelder F, Hunter TEA, Mendl MT and Lawrence AB 2000 The spontaneous qualitative assessment of behavioural expressions in pigs: first explorations of a novel methodology for integrative animal welfare measurement. *Applied Animal Behaviour Science 67*: 193-215. http://dx.doi.org/10.1016/S0168-1591(99)00093-3

Wemelsfelder F, Hunter TEA, Mendl MT and Lawrence AB 2001 Assessing the 'whole animal': a free choice profiling approach. *Animal Behaviour 62*: 209-220. http://dx.doi.org/10.1006/anbe.2001.1741

Wemelsfelder F, Knierim U, Schulze Westerath H, Lentfer T, Staack M and Sandilands V 2009a Qualitative behaviour assessment. In: Forkman B and Keeling L (eds) *Assessment of Animal Welfare Measures for Layers and Broilers* pp 113-119. Welfare Quality Reports no 9, Cardiff, UK

Wemelsfelder F and Millard F 2009b Qualitative behaviour assessment. In: Forkman B and Keeling L (eds) *Assessment of Animal Welfare Measures for Sows, Piglets and Fattening Pigs* pp 213-219. Welfare Quality Reports no 10, Cardiff, UK

Wemelsfelder F, Millard F, De Rosa G and Napolitano F 2009c Qualitative Behaviour Assessment. In: Forkman B and Keeling L (eds) *Assessment of Animal Welfare Measures for Dairy Cattle, Beef Bulls and Veal Calves* pp 215-224. Welfare Quality Reports no11, Cardiff, UK

Wemelsfelder F, Nevison I and Lawrence AB 2009d The effect of perceived environmental background on qualitative assessments of pig behaviour. *Animal Behaviour 78*: 477-484. http://dx.doi.org/10.1016/j.anbehav.2009.06.005

Winckler C 2009 Verhalten der Rinder. In: Hoy S (ed) *Nutztierethologie*. Ulmer: Stuttgart, Germany. [Title translation: Cattle Ethology]

Winckler C, Bühnemann A and Seidel K 2002 Social behaviour of commercial dairy herds as a parameter for on-farm welfare assessment. In: Koene P (ed) *Proceedings of the 36th International Congress of the ISAE* p 86. 6-10 August 2002, Egmond aan Zee, The Netherlands

https://doi.org/10.7120/09627286.24.3.319 Published online by Cambridge University Press