The Old School, Brewhouse Hill, Wheathampstead,
Hertfordshire AL4 8AN, UK
www.ufaw.org.uk

Animal Welfare 2012, 21: 307-318
ISSN 0962-7286
doi: 10.7120/09627286.21.3.307

Inter- and intra-observer reliability of experienced and inexperienced observers for the Qualitative Behaviour Assessment in dairy cattle

EAM Bokkers*, M de Vries, ICMA Antonissen and IJM de Boer

Animal Production Systems Group, Wageningen University, PO Box 338, 6700 AH Wageningen, The Netherlands
* Contact for correspondence and requests for reprints: eddie.bokkers@wur.nl

Abstract

Qualitative Behaviour Assessment (QBA) is part of the Welfare Quality® protocol for dairy cattle, although its inter- and intra-observer reliability have not been reported. This study evaluated inter- and intra-observer reliability of the QBA for dairy cattle in experienced and inexperienced observers using videos. Eight experienced observers performed the QBA (20 descriptors) twice for 16 video clips (60 s per clip; series 1) showing 4–17 animals. They assessed another 11 video clips showing herds (4 shots of 30 s per clip; series 2). Ten inexperienced observers performed the QBA on both video series one time. Inter-observer reliability of experienced observers ranged from slight to moderate (both assessments of series 1), and from low to high (series 2) for descriptors, and from slight to moderate for the QBA score. Inter-observer reliability of inexperienced observers ranged from low to moderate (series 1), and from low to high (series 2) for descriptors, and was moderate (both series) for the QBA score. Intra-observer correlations varied largely per descriptor and observer. They were both negative and positive, and ranged from low to very high. High correlations, however, were not necessarily associated with low paired differences. Values of half of the descriptors and the QBA score differed amongst experienced and inexperienced observers. The QBA appears insufficiently reliable as a tool for welfare assessment in dairy cattle.

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

Introduction

Worldwide, consumers are increasingly concerned about the welfare of farm animals. European consumers, for instance, expect their food to be produced and processed with a significant attention for animal welfare (Blokhuis *et al* 2003). To meet consumer requirements, various on-farm welfare schemes have been developed to assess and improve farm animal welfare (Blokhuis *et al* 2003). Recently, an on-farm welfare assessment protocol for dairy cattle was developed within the European Welfare Quality® project. This protocol contains several measurements of which the outcomes are used in a three-step multi-criteria evaluation model to enable the assignment of dairy herds to one of four welfare classifications (not acceptable, acceptable, enhanced, and excellent) (Welfare Quality® 2009). A specific measurement within this on-farm assessment protocol is the Qualitative Behaviour Assessment (QBA). In the Welfare Quality® Assessment protocol (WQ protocol) for dairy cattle, the QBA is the only measurement that is linked to the criterion 'positive emotional state' (Welfare Quality® 2009). The QBA is a qualitative measurement that records the expressive quality of behaviour (Welfare Quality® 2009). In other words, how animals interact with herd mates and with their environment, which is visible by their 'body language'(Wemelsfelder *et al* 2006). The QBA, as used in the WQ protocol for dairy cattle, consists of 20 descriptors, eg active, relaxed, fearful and happy, that are valued on a rating scale (Welfare Quality® 2009) .

Measurements in an on-farm assessment protocol that will be used for certification of farms or farm products must be valid and reliable (Knierim & Winckler 2009). Measurements are reliable if they are precise and consistent (Martin & Bateson 1993). For observers, a differentiation can be made between two aspects of reliability: inter-observer reliability and intra-observer reliability. Inter-observer reliability measures the extent to which two or more observers achieve the same results when measuring the same observation (Martin & Bateson 1993). Intra-observer reliability measures the extent to which the same observer achieves a corresponding result when measuring the same observation at different moments (Martin & Bateson 1993). Observer reliability was tested for several welfare measurements from the Welfare Quality® project for cattle (Windschnurer *et al* 2008; Bokkers *et al* 2009; Plesch *et al* 2010), but little is known about the inter- and intra-observer reliability of the QBA in dairy cattle. In one study involving four observers experienced in cattle behaviour and handling, none of the descriptors reached

satisfactory inter-observer reliability (Kendall's *W* ranged from 0.19–0.65 for 29 descriptors) with an on-farm assessment with dairy cattle (Wemelsfelder *et al* 2006). According to the WQ protocol for dairy cattle, QBA assessors must have experience with cattle and must be trained. However, the study of Wemelsfelder *et al* (2006). indicates that training and experience do not guarantee high inter-observer reliability. Because the QBA is an anthropomorphic qualification that is based on a human interpretation of the emotional state of an animal, we argue that experience does not contribute to an increase in reliability of the QBA. We expected that experienced and inexperienced observers are equally able to interpret what they see and feel when observing animals, although the level of qualification might be different.

In this study, we had access to the video clips belonging to the official QBA training material of Welfare Quality®. These video clips, however, were considered not to be completely representative of how QBA is performed on-farm. The video clips varied in quality, were not recorded in a standardised way, focused on only few animals on the farm, and showed both dairy and beef cattle. Therefore, additional, good quality video clips were made, recorded in a standardised way and representing whole dairy herds, which suited our aim to study inter-observer reliability comparable to a QBA performed on-farm but in a controlled way. The aim of this study was to evaluate the inter- and intra-observer reliability of experienced and inexperienced observers for the QBA in dairy cattle.

Materials and methods

Eighteen observers participated in this study. Measurements to test inter- and intra-observer reliability of the QBA were done with two series of video clips. The experimental procedure consisted of three observation sessions, one in September 2009 and two in June 2010.

Qualitative behaviour assessment

The WQ protocol explains in detail how to perform the QBA on farms (Welfare Quality® 2009). The QBA is based on visual scanning of cattle. Depending on herd size, an observer selects one to eight observation points in the barn(s) of the farm jointly providing an overview of the whole herd. The total observation time of 20 min is divided equally over the chosen observation points. Interference with animals is to be avoided. After the observation time of 20 min a value is given while the animals are no longer in sight. The expressive quality of behaviour is valued on a QBA rating scale with 20 descriptors in the following order: 'active', 'relaxed', 'fearful', 'agitated', 'calm', 'content', 'indifferent', 'frustrated', 'friendly', 'bored', 'playful', 'positively occupied', 'lively', 'inquisitive', 'irritable', 'uneasy', 'sociable', 'apathetic', 'happy', and 'distressed'. For each descriptor, a line is drawn on a visual analogue scale (VAS) ranging from 0 mm (expressive quality was absent in any of the observed animals) to 125 mm (expressive quality was dominant across all observed animals) (Welfare Quality® 2009). Video clips are used to train observers in performing the QBA before applying the assessment on live animals on farms.

Video series

Two video series with sound were used. Video series 1 consisted of 16 video clips, belonging to the official training material of Welfare Quality®. Thirteen of these clips showed dairy cows and three showed fattening bulls, all loose housed. The number of animals shown in each video clip ranged from four to 17. Each clip lasted about 1 min. Five of the video clips were rather dark, in three of these the image was slightly blurred, whereas in one clip the person who was recording zoomed in on one animal. In two video clips the camera switched from one group of cows to another group of cows in the same barn.

Because the video clips of video series 1 were of varying quality, not recorded in a standardised way, representing just a part of a herd, and did not contain only dairy cattle, new video clips (video series 2) were made by one of the authors (IA). Video series 2 consisted of eleven video clips. Each video clip included four shots of 30 s, recorded from four different observation points in one barn, with a total duration of 2 min per herd. The reason for showing shots of four observation points of one herd within each video clip was to create a situation similar to a QBA assessment on-farm, although observation time was reduced from 20 to 2 min. Observation time was reduced to stay close to the other video series but also because of practical reasons: it is hard to find volunteers to observe a series of video clips of 20 min each.

For the recordings, a Sony camcorder (DCR SX, Sony Corporation, Tokyo, Japan) was installed on a tripod and placed on a platform in the feeding corridor (camera height: 2.5 m). Recordings were made of the living area of dairy herds in a loose-housing system with cubicles on Dutch dairy farms. The number of cows shown in a 2-min clip ranged from 30 to 70 cows. The clips of video series 2 had the same quality in terms of focus, light, clarity, etc.

Observers

The experienced group consisted of eight observers, ie six males and two females, aged 24–53 years (one PhD student, one MSc student, and one technical assistant from Wageningen University [Wageningen, The Netherlands], one technical assistant from the Institute for Agricultural and Fisheries Research [Merelbeke, Belgium], and four advisors from Animal Health Service [Deventer, The Netherlands]). All were familiar with dairy cattle. The inexperienced group consisted of ten students from Wageningen University, ie five males and five females, aged between 18 and 26 years They were not familiar with farm animals and had no experience with observing farm animal behaviour. None of the students followed a study programme with any relation to animals, except for one who studied Biology.

Experimental procedure

Observation session 1 — experienced group

In September 2009, all observers from the experienced group were trained to execute the WQ protocol for dairy cattle in a three-day course. This training was given in The Netherlands, by two delegates from the Welfare Quality®

https://doi.org/10.7120/09627286.21.3.307 Published online by Cambridge University Press

project. As part of the training, observers were taught how to perform the QBA. The QBA training started with a theoretical introduction followed by a discussion about the meaning of the 20 descriptors of the QBA rating scale. When agreement was obtained about the meaning of the descriptors, the observers practised performing the QBA with video clips and with two farm visits.

The QBA training ended with a session together at one location in which each observer performed the QBA for video series 1. Clips of video series 1 were shown on a slide screen. Directly after having watched a video clip, observers valued the 20 descriptors of the QBA rating scale. When all observers had finished, the next video clip was shown. Each observer completed 16 QBA rating scales in total. Observers were not allowed to deliberate during the assessment.

Between September 2009 and April 2010, the experienced observers applied their skills by executing the QBA on commercial dairy farms. They visited between 10 and 48 dairy farms each.

Observation session 2 — experienced group

In June 2010, the experienced group again performed the QBA for video series 1. Conditions were similar to observation session 1. Before watching the video clips, descriptors were not discussed and no further instructions were given because observers were still familiar with the QBA. The same 16 video clips were shown on a slide screen in the same order as in observation session 1. After scoring video series 1, observers had a short break and then continued to assess the eleven video clips from video series 2. The same procedure was followed as for assessing video series 1. Observers were at the same location to assess the video clips, except for one observer who did this assessment at another location but under similar conditions. As in the first session, observers were not allowed to deliberate.

Observation session 3 — inexperienced group

In June 2010, the inexperienced group performed the QBA for video series 1 and 2, together in a quiet room. Observers were given instructions about expectations and the way to fill out the VAS. The meanings of the descriptors were not discussed before watching and scoring the video clips. Observers had to read the descriptors and were allowed to use a dictionary. If there was ambiguity about the meaning of a certain descriptor the instructor explained its meaning for the whole group of observers. Next, the inexperienced group watched the video clips of both video series in the same order as the experienced group. Video clips were shown on a TV screen, which was smaller (37 inches) than the slide screen but of a similar quality. Between scoring the two video series there was a 15-min break. Observers were not allowed to deliberate during the session.

Statistical analysis

In total, 743 QBA rating scales were completed by the experienced and inexperienced observers. VAS values for each descriptor were determined by measuring the distance with a ruler from the left side (minimum) to the point where the observer drew a line on the VAS. Next, VAS values were aggregated into one overall QBA score. We calculated this QBA score according to the procedure described in the WQ protocol (Welfare Quality® 2009). In this procedure, VAS values are first weighted into a single index per clip, observer, and observation session. Weighting factors are derived from a Principle Component Analysis. This index, expressed on a scale from –8 to 8, was then transformed into one QBA score per clip, observer, and observation session, expressed on a 0 to 100 scale. This transformation was based on I-spline functions that reflect expert opinion. We analysed statistically VAS values of the individual descriptors and QBA scores, using SPSS 17.0 (IBM SPSS Inc, Chicago, USA).

To test if experience affected VAS values and QBA scores, independent samples *t*-tests were performed. To test for differences among observers within a group, a one-way analysis of variance was performed on VAS values of all descriptors and QBA scores. When an observer effect was found, a *post hoc* pair-wise comparison (Bonferroni) was conducted for all observers.

To determine inter-observer reliabilities for both video series, Kendall's coefficients of concordance (*W*-coefficient) were calculated for each descriptor and for the QBA score in both groups of observers. This test is often used for expressing inter-rater agreement among independent judges who are rating the same stimuli (see for example Bokkers *et al* 2009) and was used earlier for inter-observer agreement in a QBA study (Wemelsfelder *et al* 2006).

To analyse if the valuing of descriptors were related, descriptors were correlated (Spearman's rank correlation, r_s) over video series and observation sessions for experienced and experienced observers. Only significant correlation coefficients ($r_s > 0.7$) between descriptors are described.

To determine intra-observer reliability among experienced observers, paired samples *t*-tests were conducted to identify differences in VAS values and QBA scores between video series 1 assessed in September 2009 and in June 2010. Additionally, the paired-samples correlation coefficients (Pearson) are given to indicate whether observers valued descriptors and QBA scores for a video clip the same in September 2009 and June 2010.

For coefficients of concordance (*W*) and correlation coefficients (*r*), we followed Martin and Bateson (1993) who distinguished five categories, ie slight correlation: coefficients from 0 to 0.2; low correlation: coefficients from 0.2 to 0.4; moderate correlation: coefficients from 0.4 to 0.7; high correlation: coefficients from 0.7 to 0.9; and very high correlation: coefficients from 0.9 to 1.0. For each descriptor, the range is given from the lowest to the highest correlation found for both groups.

Table 1 Mean (± SD) VAS values, Kendall's coefficients of concordance (*W*) per descriptor, and QBA scores for video series 1 in September 2009 (experienced observers) and June 2010 (experienced and inexperienced observers).

	September		June				
	Experienced		Experienced		Inexperienced		
Descriptor	**Mean (± SD)**	**Kendall's *W***	**Mean (± SD)**	**Kendall's *W***	**Mean (± SD)**	**Kendall's *W***	***P*-value**
Active	59.1 (± 31.3)	0.59	62.4 (± 31.5)	0.58	45.1 (± 30.3)	0.34	0.000
Relaxed	55.6 (± 35.7)	0.37	50.2 (± 33.8)	0.32	62.6 (± 32.6)	0.41	0.002
Fearful	23.4 (± 26.8)	0.64	21.1 (± 25.4)	0.63	15.8 (± 18.7)	0.65	0.050
Agitated	35.0 (± 31.3)	0.38	33.0 (± 32.4)	0.41	30.5 (± 28.2)	0.46	0.485
Calm	65.9 (± 33.4)	0.31	57.7 (± 33.7)	0.33	64.8 (± 34.2)	0.40	0.060
Content	57.9 (± 30.9)	0.23	52.4 (± 31.6)	0.47	59.4 (± 31.7)	0.48	0.078
Indifferent	35.2 (± 28.3)	0.25	31.9 (± 26.4)	0.13	38.6 (± 30.1)	0.40	0.048
Frustrated	35.4 (± 33.0)	0.35	35.6 (± 31.5)	0.39	27.6 (± 28.7)	0.56	0.028
Friendly	46.5 (± 31.4)	0.30	41.3 (± 29.8)	0.51	56.9 (± 33.0)	0.68	0.000
Bored	48.3 (± 33.3)	0.29	38.9 (± 30.9)	0.16	53.9 (±30.3)	0.38	0.000
Playful	32.0 (± 26.6)	0.38	33.3 (± 27.5)	0.59	27.5 (± 26.3)	0.45	0.071
Positively occupied	46.3 (± 30.1)	0.24	54.0 (± 31.4)	0.52	45.7 (± 32.4)	0.60	0.029
Lively	50.0 (± 28.9)	0.48	54.5 (± 27.2)	0.51	47.9 (± 32.8)	0.39	0.067
Inquisitive	45.1 (± 31.5)	0.49	40.5 (± 29.5)	0.42	37.5 (± 30.0)	0.56	0.396
Irritable	35.1 (± 32.3)	0.25	35.3 (± 31.1)	0.31	29.4 (± 29.0)	0.51	0.099
Uneasy	36.4 (± 29.7)	0.35	35.8 (± 31.3)	0.31	32.6 (± 28.5)	0.51	0.358
Sociable	44.8 (± 30.1)	0.38	41.5 (± 28.7)	0.38	45.4 (± 32.6)	0.48	0.292
Apathetic	19.3 (± 20.5)	0.30	14.9 (± 19.1)	0.46	31.3 (± 30.2)	0.42	0.000
Happy	48.1 (± 27.6)	0.37	46.8 (± 26.8)	0.57	44.9 (± 33.3)	0.62	0.591
Distressed	24.5 (± 25.8)	0.25	23.1 (± 26.1)	0.61	64.2 (± 40.4)	0.62	0.000
QBA score	27.4 (± 27.4)	0.14	31.5 (± 26.7)	0.35	23.4 (± 18.9)	0.48	0.004

The *P*-value reflects the difference in VAS values per descriptor between experienced and inexperienced observers given in June 2010.

Results

Effect of experience

Video series 1

Mean VAS values of the experienced group for video series 1 varied from 19.3 ('apathetic') to 65.9 ('calm') in September and from 14.9 ('apathetic') to 62.4 ('active') in June (Table 1). Mean VAS values of the inexperienced group for video series 1 varied from 15.8 ('fearful') to 64.8 ('calm') in June (Table 1). In June, the experienced group valued the descriptors 'active', 'frustrated', and 'positively occupied' higher, and the descriptors 'relaxed', 'indifferent', 'friendly', 'bored', 'apathetic' and 'distressed' lower than the inexperienced group for video series 1 (Table 1). The average QBA score for video series 1 was higher for the experienced than for the inexperienced group in June (Table 1).

Video series 2

Mean VAS values of the experienced group for video series 2 varied from 9.6 ('distressed') to 70.6 ('content') (Table 2). Mean VAS value of the inexperienced group for video series 2 varied from 11.0 ('fearful') to 79.4 ('calm') (Table 2). For video series 2, the experienced group valued the descriptors 'active', 'playful', 'positively occupied', and 'lively' higher, and the descriptors 'calm', 'indifferent', 'friendly', 'bored', 'apathetic' and 'distressed' lower than the inexperienced group (Table 2). The average QBA score for video series 2 was higher for the experienced than for the inexperienced group (Table 2).

In the experienced group, the descriptors 'relaxed', 'calm', 'content', and 'happy' showed high positive correlation coefficients across video series and observation sessions. Moreover, 'positively occupied' was also highly positively correlated with 'content' and 'happy' ($r_s > 0.7$; $P < 0.05$); 'agitated', 'frustrated', and 'irritable' were highly positively correlated ($r_s > 0.7$; $P < 0.05$), and 'fearful' was highly positively correlated with 'agitated', and 'uneasy' with 'frustrated' and 'irritable' ($r_s > 0.7$; $P < 0.05$).

In the inexperienced group, the descriptor 'relaxed' was highly positively correlated with 'calm' across video series, and 'frustrated' was highly positively correlated with 'irritable', 'fearful', and 'agitated' ($r_s > 0.70$; $P < 0.05$).

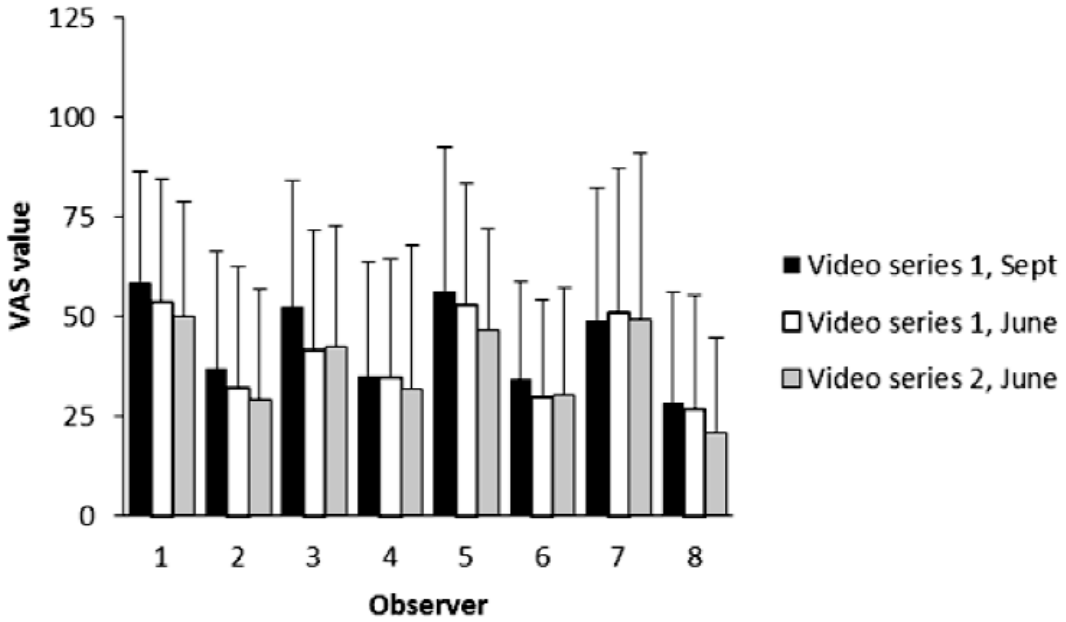
https://doi.org/10.7120/09627286.21.3.307 Published online by Cambridge University Press

Table 2 Mean (± SD), Kendall's coefficients of concordance (*W*) per descriptor, and QBA score for video series 2 in June 2010.

	Experienced		Inexperienced		
Descriptor	**Mean (± SD)**	**Kendall's *W***	**Mean (± SD)**	**Kendall's *W***	***P*-value**
Active	69.5 (± 28.8)	0.72	36.0 (± 21.3)	0.23	0.000
Relaxed	69.7 (± 25.3)	0.42	75.7 (± 27.6)	0.60	0.112
Fearful	12.9 (± 15.2)	0.71	11.0 (± 13.2)	0.84	0.347
Agitated	18.0 (± 20.9)	0.66	21.2 (± 20.0)	0.72	0.267
Calm	69.9 (± 27.9)	0.66	79.4 (± 27.7)	0.67	0.018
Content	70.6 (± 24.9)	0.47	68.4 (± 28.1)	0.74	0.561
Indifferent	21.9 (± 18.7)	0.32	41.5 (± 30.3)	0.57	0.000
Frustrated	18.8 (± 19.2)	0.61	15.8 (± 18.2)	0.66	0.267
Friendly	43.3 (± 27.6)	0.67	59.4 (± 28.6)	0.77	0.000
Bored	24.3 (± 27.8)	0.65	54.3 (± 29.0)	0.48	0.000
Playful	26.3 (± 24.0)	0.83	18.3 (± 17.7)	0.65	0.009
Positively occupied	68.4 (± 27.2)	0.64	55.6 (± 29.4)	0.56	0.002
Lively	56.6 (± 24.7)	0.52	43.5 (± 25.6)	0.48	0.000
Inquisitive	33.3 (± 23.9)	0.74	28.5 (± 20.2)	0.57	0.135
Irritable	15.9 (± 17.9)	0.62	20.0 (± 20.8)	0.58	0.143
Uneasy	19.9 (± 23.6)	0.59	21.2 (± 21.8)	0.53	0.694
Sociable	36.3 (± 24.7)	0.68	42.8 (± 25.7)	0.65	0.071
Apathetic	9.7 (± 12.3)	0.77	36.0 (± 28.7)	0.67	0.000
Happy	56.6 (± 23.8)	0.52	58.1 (± 25.0)	0.60	0.676
Distressed	9.6 (± 12.4)	0.76	76.4 (± 39.9)	0.81	0.000
QBA score	48.8 (± 22.1)	0.62	28.4 (± 14.0)	0.66	0.000

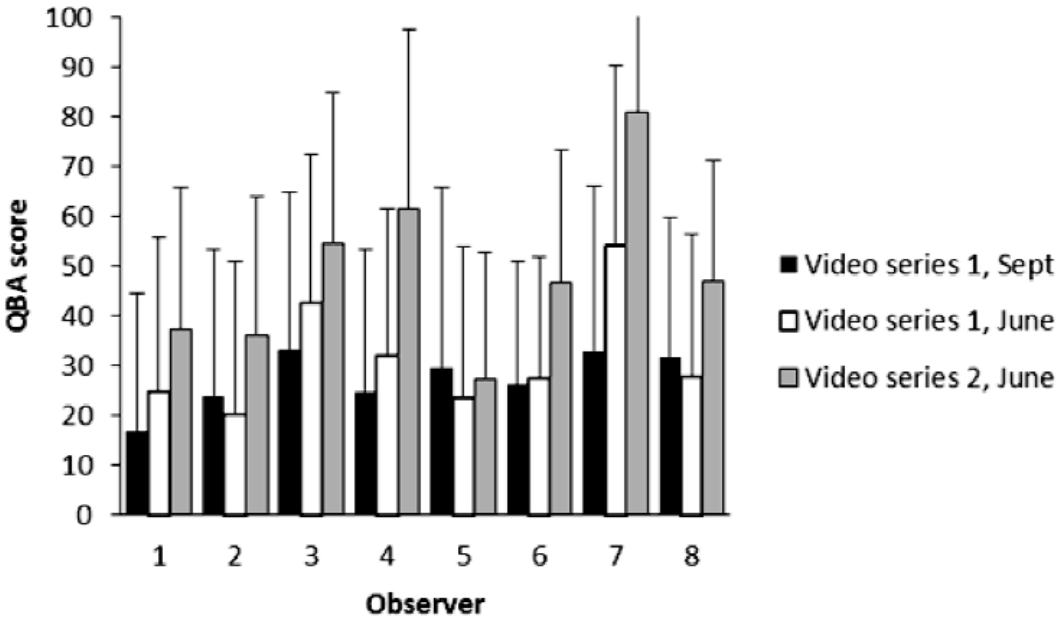
The *P*-value reflects the difference in VAS values per descriptor between experienced and inexperienced observers.

Figure 1



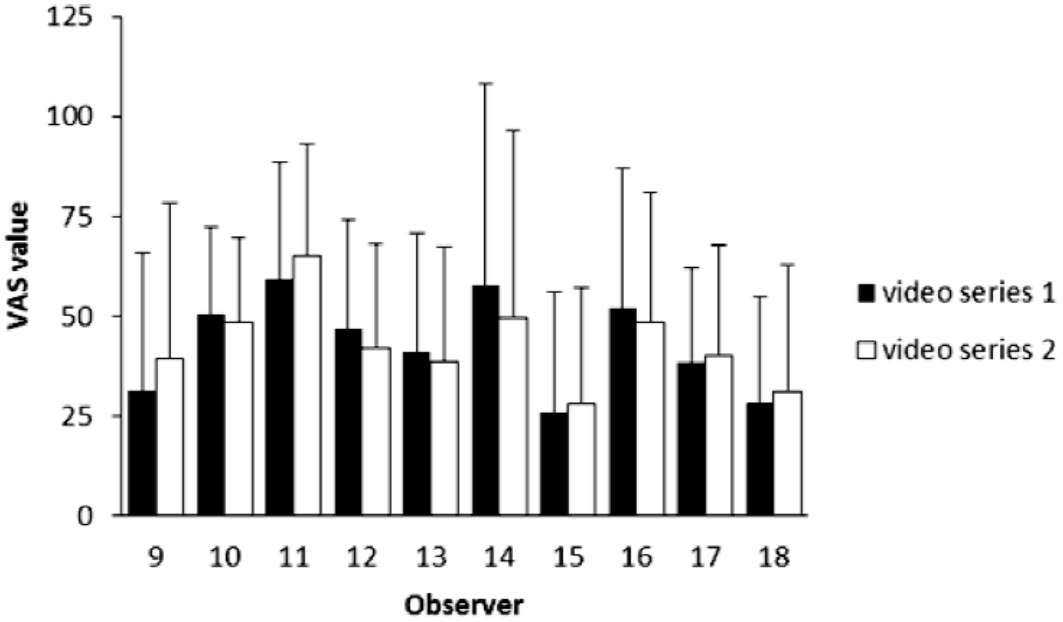
Mean (± SD) VAS values for video series 1 and 2 of the experienced observers in September 2009 and June 2010.

Figure 2



Mean (± SD) QBA scores for video series 1 and 2 of the experienced observers in September 2009 and June 2010.

Figure 3



Mean (± SD) VAS values for video series 1 and 2 of the inexperienced observers in June 2010.

Differences among observers in a group

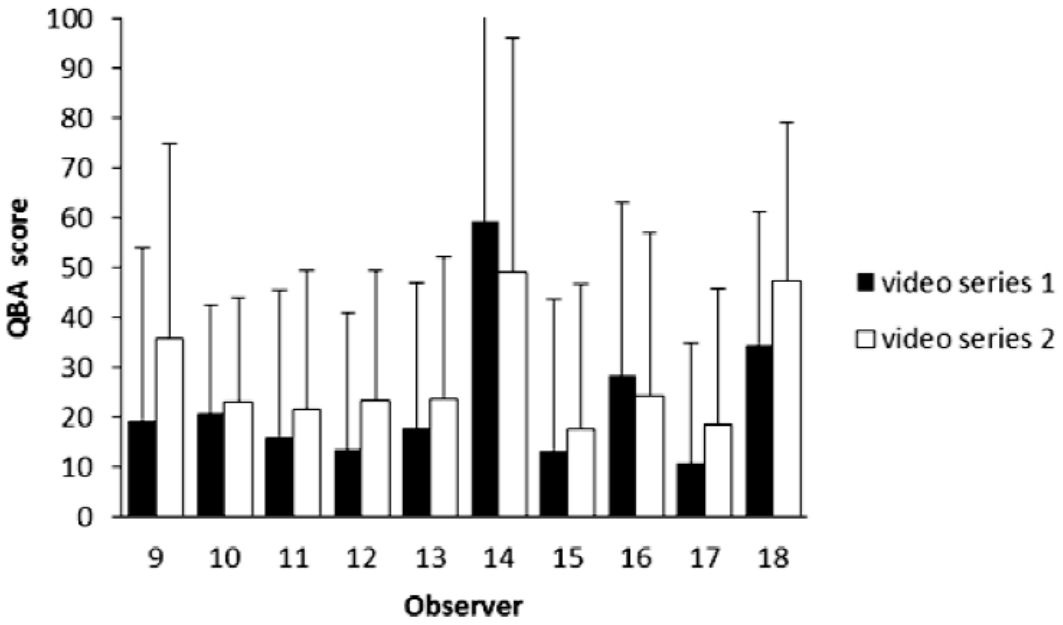
Experienced observers

VAS values of all descriptors differed among experienced observers in both video series ($P < 0.001$; Figure 1). The multiple, pair-wise comparison showed that experienced observers 2, 4, 6, and 8 had systematically lower VAS values compared to observers 1, 3, 5, and 7 in September and June ($P < 0.05$), except for observer 4 who did not have a different VAS value compared with observer 3 in June.

In contrast to the VAS values, overall QBA scores did not differ among experienced observers for video series 1 in September (Figure 2). In June, however, observer 7 had higher QBA scores than observers 1, 2, and 5 ($P < 0.05$) for video series 1. For video series 2, observer 3 had a higher

https://doi.org/10.7120/09627286.21.3.307 Published online by Cambridge University Press

Figure 4



Mean (± SD) QBA scores for video series 1 and 2 of the inexperienced observers in June 2010.

QBA score than observer 5 ($P < 0.01$); observer 4 had a higher QBA score than observer 1, 2, and 5 ($P < 0.05$); and observer 7 had higher QBA scores than all other observers, except for observer 4 ($P < 0.01$; Figure 2).

Inexperienced observers

VAS values and overall QBA scores differed among inexperienced observers for video series 1 and 2 (Figure 3). For all descriptors, with respect to video series 1, the inexperienced observers 9, 15, and 18 had lower VAS values than observers 10, 11, 12, 13, 14, 16 and 17 ($P < 0.05$), except for observer 9 whose values did not differ from those from observer 17 (Figure 3). For video series 2, only observer 15 had lower VAS values than all other inexperienced observers ($P < 0.05$; Figure 3). Moreover, for video series 1, observer 14 had higher QBA scores than all other observers ($P < 0.001$); observer 16 had higher QBA scores than observer 17 ($P < 0.05$); and observer 18 had higher QBA scores than observers 11, 12, 13, 15, and 17 ($P < 0.05$; Figure 4). For video series 2, observer 9 had higher QBA scores than observers 11 and 17 ($P < 0.05$), and observer 14 had higher QBA scores than all other observers except observer 18 ($P < 0.05$; Figure 4).

Inter-observer reliability

Video series 1

In the experienced group, *W*-coefficients for descriptors ranged from 0.23 ('content') to 0.64 ('fearful') for video series 1 in September (Table 1). *W*-coefficients were slight for the QBA score, low for 15 descriptors, and moderate for five descriptors. In June 2010, *W*-coefficients ranged from 0.16 ('indifferent') to 0.63 ('fearful') for video series 1. *W*-coefficients were slight for two descriptors, low for six descriptors and the QBA score, and moderate for twelve descriptors (Table 1). In the inexperienced group, *W*-coefficients ranged from 0.34 ('active') to 0.68 ('friendly') for video series 1. *W*-coefficients were low for three descriptors and moderate for seventeen descriptors and the QBA score (Table 1).

Video series 2

In the experienced group, *W*-coefficients ranged from 0.32 ('indifferent') to 0.83 ('playful') for video series 2 (Table 2). *W*-coefficients were low for one descriptor, moderate for thirteen descriptors and the QBA score, and high for six descriptors (Table 2). In the inexperienced group, *W*-coefficients ranged from 0.23 ('active') to 0.84 ('fearful') for video series 2 (Table 2). *W*-coefficients were low for one descriptor, moderate for 14 descriptors and the QBA score, and high for five descriptors (Table 2).

Intra-observer reliability

Experienced observers 1, 2, 3, and 6 valued descriptors lower in June than in September (Table 3). In total, 44 significant differences were found for VAS values of individual observers between September and June. Of these differences, 34 were lower in June than in September. When looking per descriptor, the number of observers that showed a difference between September and June varied. For one descriptor, five differences were found. For nine descriptors, three differences were found. For five descriptors, two differences were found, and for four descriptors, one difference was found. For one descriptor none of the observers differed between

Table 3 Correlation coefficients and scores of experienced observers over all descriptors for video series 1 in September 2009 and June 2010.

All cases	Paired-samples correlations	September: Mean (± SD)	June: Mean (± SD)	*P*-value
Observer 1	0.61*	58.5 (± 27.7)	53.5 (± 30.8)	0.000
Observer 2	0.76*	36.8 (± 29.5)	31.9 (± 30.5)	0.000
Observer 3	0.69*	52.5 (± 31.6)	41.5 (± 29.8)	0.000
Observer 4	0.65*	35.2 (± 28.5)	34.7 (± 29.5)	0.685
Observer 5	0.51*	56.3 (± 36.1)	52.8 (± 30.4)	0.061
Observer 6	0.65*	34.1 (± 24.7)	29.8 (± 24.4)	0.000
Observer 7	0.65*	49.0 (± 33.0)	50.8 (± 36.2)	0.279
Observer 8	0.58*	28.1 (± 28.1)	26.9 (± 28.5)	0.396

* Correlation was significant when $P < 0.05$.

Table 4 Pair-wise correlation and comparison of the QBA scores of experienced observers for video series 1 in September 2009 and June 2010.

All cases	Paired-samples correlations	September: Mean (± SD)	June: Mean (± SD)	*P*-value
Observer 1	0.86*	16.8 (± 20.0)	24.8 (± 27.8)	0.043
Observer 2	0.85*	23.9 (± 24.2)	20.2 (± 19.9)	0.278
Observer 3	0.78*	33.1 (± 24.8)	42.5 (± 26.0)	0.043
Observer 4	0.81*	24.8 (± 19.7)	31.9 (± 19.6)	0.029
Observer 5	0.71*	29.7 (± 29.6)	23.3 (± 25.9)	0.259
Observer 6	0.74*	26.2 (± 22.6)	27.5 (± 23.8)	0.761
Observer 7	0.88*	32.9 (± 28.5)	53.9 (± 32.9)	0.000
Observer 8	0.74*	31.6 (± 24.3)	27.7 (± 23.0)	0.373

* Correlation was significant when $P < 0.05$.

Table 5 Example for the descriptor 'active'. Pair-wise correlation and comparison of the VAS values of the experienced observers performed for video series 1 in September 2009 and June 2010.

All cases	Paired-samples correlations	September: Mean (± SD)	June: Mean (± SD)	*P*-value
Observer 1	0.52	59.4 (± 30.8)	74.4 (± 29.0)	0.058
Observer 2	0.89*	48.6 (± 22.3)	45.3 (± 30.6)	0.390
Observer 3	0.30	82.9 (± 24.3)	71.9 (± 15.9)	0.097
Observer 4	0.74*	38.6 (± 30.2)	55.9 (± 27.4)	0.005
Observer 5	0.45	92.9 (± 13.4)	87.0 (± 21.8)	0.254
Observer 6	0.84*	41.9 (± 19.1)	33.6 (± 17.3)	0.006
Observer 7	0.40	78.5 (± 21.6)	91.3 (± 15.1)	0.027
Observer 8	0.81*	46.6 (± 30.1)	40.0 (± 32.1)	0.191

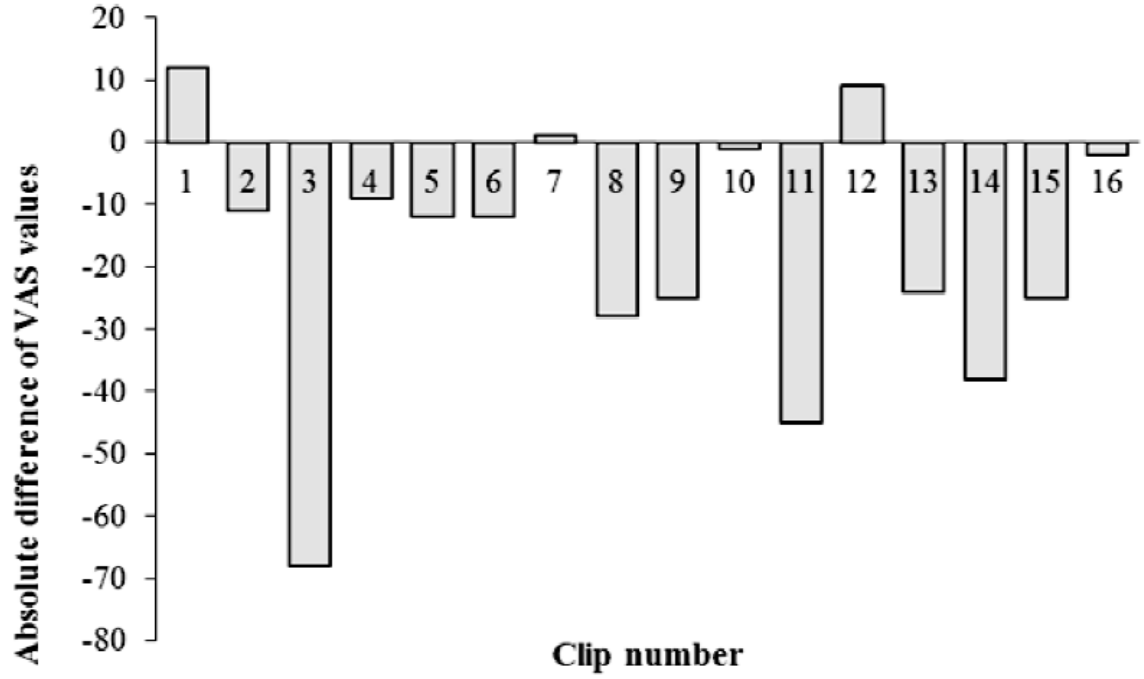
* Correlation was significant when $P < 0.05$.

September and June. Observers 1, 3 and 7 had different VAS values for nine to twelve descriptors. Observers 2, 4, 5 and 6 had different VAS values for three to four descriptors, and observer 8 had a different VAS value for one descriptor only. Although high positive correlations were found for QBA scores between September and June for all experienced observers, scores were lower in September than in June for experienced observers 1, 3, 4, and 7 (Table 4).

Table 5 shows an example of the correlations and comparison of VAS values in September and June performed by the experienced observers for the descriptor 'active'. For descriptor 'active', the VAS values in September and June

https://doi.org/10.7120/09627286.21.3.307 Published online by Cambridge University Press

Figure 5



Absolute differences for video clips of video series 1 between VAS values of September 2009 compared to June 2010 (reference) given by observer 4 for the descriptor 'active'

from observers 4 and 6 were positively correlated, but observer 4 had higher VAS values in June while observer 6 had higher VAS values in September. The VAS values of observers 2 and 8 were also positively correlated, but no difference was found between VAS values of September and June. Furthermore, no significant correlation for descriptor 'active' was found for observer 7, but this observer gave higher VAS values in June than in September. Figure 5 shows the absolute difference for descriptor 'active' for observer 4. Observer 4 valued the descriptor 'active' lower in 13 of the video clips in September than s/he did in June. Positively correlated VAS values with significant mean differences were found in 26 cases (16%) over all observers and descriptors (8 observers × 20 descriptors = 160 cases in total).

Discussion

The aim of this study was testing the inter- and intra-observer reliability of the QBA with experienced and inexperienced observers.

Inter-observer reliability

For video series 1, inter-observer reliability of each descriptor was low to moderate in both observer groups. No correlation coefficients above 0.7 were found, which is referred to as the threshold for an acceptable correlation coefficient for inter- and intra-observer reliability (Martin & Bateson 1993). For video series 2, inter-observer reliability was low to high for the different descriptors, but most correlation coefficients were moderate. These results indicate that observers (trained or not trained) are not able to give a similar VAS value for each descriptor when assessing the same video clips. This is in accordance with the results of Wemelsfelder *et al* (2006) who also found low inter-observer reliabilities for individual descriptors assessed on dairy cattle farms (none of the 29 descriptors reached an inter-observer reliability above 0.7).

For the experienced group, inter-observer reliability for the QBA score was slight (September 2009) to low (June 2010) for video series 1, and moderate for video series 2. For the inexperienced group, this reliability was moderate for both video series. Our outcomes were comparable to the inter-observer reliability based on PCA factors in Wemelsfelder *et al* (2006). They found inter-observer reliabilities of 0.38 for the first and 0.46 for the second factor, based on four observers performing qualitative assessments on farms.

Inter-observer reliability was similar for QBA scores and most of the individual descriptors. To obtain a reliable QBA score, reliability of all underlying descriptors is of importance. If a herd, for example, obtains a low QBA score in an on-farm welfare assessment, one would look into the values of the underlying individual descriptors to identify which of the descriptors are causing this low QBA score. If the low QBA score is caused mainly by a high value for, eg 'frustrated', there would be no use in improving this value if this descriptor cannot be assessed with sufficient inter-, as well as, intra-observer reliability. Moreover, if two observers give different VAS values for a certain descriptor, the differ-

ence in this valuation could even lead to different classifications of a herd using the Welfare Quality® multi-criteria evaluation model (Welfare Quality® 2009).

Although experienced observers were trained, it can be questioned whether observers (experienced and inexperienced) interpreted the descriptors in a similar way. Clear definitions of descriptors using objective criteria is essential to get acceptable inter-observer reliabilities (Meagher 2009). In her review, Meagher (2009) suggests that terms such as 'joyful'or 'sisterly', which have a comparable vagueness as for instance the QBA descriptors 'happy' or 'content', may not be suitable in research because there are no clear criteria to define these descriptors.

Intra-observer reliability

Four observers from the experienced group gave lower average VAS values in June compared to September. Among these four were observers 2 and 6, who visited a large number of farms to execute the WQ protocol for dairy cattle including the QBA. Even more striking is the fact that intra-observer reliability for various descriptors varied from low to high among observers. In other words, the reliability level was not necessarily related to an increase or decrease of the VAS values in September and June. QBA scores were higher for four observers in June compared to September, but these four observers were not all the same observers that valued the VAS differently. The computation of the QBA score is affecting this because descriptors are weighted differently. From these results it can be concluded that there is no consistency in scores of individual descriptors within an observer when assessing the same video clips twice. Although Wemelsfelder and Lawrence (2001) conclude that intra-observer reliability of spontaneous qualitative assessments of pig behaviour are high in experimental conditions, it might be different in on-farm conditions and with a prescribed format of descriptors instead of having a free choice of descriptors. Also, such qualitative assessments might be more suitable for pigs than for dairy cows because observers might find it easier to assess and interpret their behaviour and translate it to descriptors.

Effect of experience

The number of descriptors with low *W*-coefficients decreased from 15 in September to six in June. Twelve descriptors with low *W*-coefficients changed to moderate. *W*-coefficients of the QBA score improved from slight to low. The effect of performing QBA assessments on farms could be the reason for these changes. Practice and experience influences reliability positively (Martin & Bateson 1993). Although in previous research inexperienced observers were able to show acceptable agreements when performing a spontaneous qualitative assessment in pigs (Wemelsfelder *et al* 2000, 2001) and in cattle (Napolitano *et al* 2007), the WQ protocol for dairy cattle prescribes that observers should be fully trained to be capable of performing the QBA. In the WQ protocol for dairy cattle, however, observers have to value predefined descriptors, which is different from the spontaneous qualitative assessment in which observers choose their own descriptors. In this study, we hypothesised that with predefined descriptors experienced and inexperienced observers would show similar inter-observer reliability, because descriptors are assessed qualitatively and are based on a human interpretation of the emotional state of an animal. This is illustrated by descriptors such as 'happy', 'inquisitive', 'bored', etc, which are anthropomorphic qualifications.

Although experienced (either trained or not trained to perform the QBA) and inexperienced people may have a different levels of QBA qualification, it may be expected that inexperienced people are equally capable of interpreting what they have seen and felt when watching a group of animals for a while. Many behaviours and activities animals perform are so close to the general human understanding and perspective that experience is not needed to interpret what is seen. Though, this human emotional interpretation is not necessarily the correct interpretation of how animals feel.

It was not surprising, therefore, that the inexperienced group, without training or any knowledge about dairy cattle, reached similar or even higher reliability levels for descriptors and the QBA score compared with the experienced group. This showed that being trained and experienced had no, or even an adverse, effect on inter-observer reliability of the QBA. Looking at the results, it appears that individual differences within groups are higher than differences between experienced and inexperienced observers, which possibly reflects different emotional states of the observers due to, for instance, different attitudes, characters, personal background, and age.

Another explanation for the large individual differences could be that the experienced observers were not trained well enough. One of the aims of training is to ensure that observers record measures with a consistent rate of accuracy (Kazdin 1977). When the highest level of training is reached, observers are supposed to continue to utilise the same definitions of measures and record accurately (Kazdin 1977; Meagher 2009). After the assessment in September, there was no direct feedback on the results. Observers did not know whether their values were right. On the longer term, observer 'drift' might have happened. Drift means observers change the way they use definitions over time. This is not always visible if inter-observer reliability is determined. Observers could develop similar variations of the original definitions with a high level of inter-observer agreement but accuracy will decrease (Kazdin 1977). The effect of drift might have been reduced if observers were trained not only before the farm visits but also in the period when the farm visits were executed.

Video series

Inter-observer reliability was clearly different between video series 1 and 2. *W*-coefficients of most descriptors in video series 1 were low to moderate, whereas coefficients in video series 2, were mostly moderate and sometimes high. Video series 1 showed 16 video clips, each clip with a duration of approximately 30 s. Recording was done from one observation point in the barn in each clip. Therefore, in each video clip, a limited number of animals in the herd

https://doi.org/10.7120/09627286.21.3.307 Published online by Cambridge University Press

were observed from the beginning to the end of the clip and some clips were focused on special behaviours (eg agonistic interactions). Exceptions were two video clips in which the camera switched from one group of cows to another in the same barn. Video series 2 showed eleven different herds and from every herd four shots, each from another observation point, of 30 s, were viewed. Consequently, the number of observed cows was much higher compared to the number of animals in video series 1. For this reason, a poorer inter-observer reliability was expected for video series 2, but results showed the opposite. One possible explanation for this outcome could be the similarity of the video clips from video series 2. All clips of video series 2 showed dairy cows in a barn and behaviour of cows consisted mainly of lying, feeding and walking, which are the main activities of dairy cows in a barn (Gomez & Cook 2010; Uzal & Ugurlu 2010). Few social interactions were shown in video series 2, contrary to video series 1

Another explanation for the higher inter-observer reliability for video series 2 is that scoring should be done in proportion to all observed cows. This could lead to a more average scoring pattern. The difficulty of scoring video series 2 is being able to remember what happened in the first 30 s of the 2-min video clip and value the descriptors in proportion to all the observed cows. If, for example, in the first 30 s 'fearful' cows were shown and in the next 90 s no 'fearful' cows were shown, observers should attempt to average their value because the descriptor 'fearful' has to be valued in proportion to all observed cows. It can be questioned, however, if an observer is able to do so for all descriptors. This situation seems to be comparable with performing QBA on farms, in which an observer watches the whole herd divided over different observation points in the barn and after 20 min of observation, a value has to be given to all descriptors of the QBA.

Although it is one of the instructions for the QBA to value descriptors independently and value each descriptor in its own right, it may be questioned whether observers are able to do so, because the meanings of some descriptors are overlapping. This was confirmed by the results of the experienced group, in which we found high positive correlations for descriptors with a similar loading such as 'happy', 'content', 'calm', and 'relaxed', and also for 'agitated', 'frustrated', 'irritable', and 'fearful'. No high negative correlations, however, were found between these contrary descriptors. In the inexperienced group, correlations between descriptors were much lower, indicating that training did have an effect on how the observers perceive and value descriptors. This, however, is not evidence that they value them consistently, as illustrated by the reliability results. Because of the number of descriptors with overlapping meaning, it is inevitable observers will link certain descriptors with each other.

Also, observers' declining concentration may have led to a more average scoring pattern (Caro *et al* 1979). First, the observers watched the 16 clips of video series 1, each approximately 1 min in duration, and after each video clip, they needed approximately 1 min to fill in a QBA form for each video clip. Then, after a break of 10 min, the observers watched video series 2 with a total duration of 22 min, and approximately 1 min between every clip to fill in the QBA rating scale. So, total time to complete the whole session was approximately 1 h and 15 min. The level of concentration might be different for each observer which could have negatively affected the inter-observer reliability.

A measurement may or may be not repeatable between and within observers, for a truly reliable measurement it also should be valid, or provide useful information about the underlying state. This study was not designed to assess validity, but if QBA aims to become a valuable measurement for welfare assessment, both repeatability and validity of the QBA must be proven scientifically. With the results of our study it is questionable if the QBA is a useful and reliable tool for welfare assessment. A high inter- and intra-observer reliability is needed if the QBA is to be used in the future for certification purposes (Knierim & Winckler 2009).

Conclusion

The results of this study showed that inter-observer reliability was slight to high for individual QBA descriptors and slight to moderate for QBA scores. Values of half the descriptors and the QBA score differed amongst experienced and inexperienced observers. Intra-observer reliability varied largely per descriptor and observer. The QBA appears insufficiently reliable as a tool for welfare assessment in dairy cattle.

Acknowledgements

We would like to thank all the observers for their time of scoring the videos. Christoph Winckler is acknowledged for delivering video series 1.

References

Blokhuis HJ, Jones RB, Geers R, Miele M and Veissier I 2003 Measuring and monitoring animal welfare: transparency in the food product quality chain. *Animal Welfare 12*: 445-455

Bokkers EAM, Leruste H, Heutinck LFM, Wolthuis-Fillerup, M van der Werf JTN, Lensink BJ and van Reenen CG 2009 Inter-observer and test-retest reliability of on-farm behavioural observations in veal calves. *Animal Welfare 18*: 381-390

Caro TM, Roper R, Young M and Dank GR 1979 Inter-observer reliability. *Behaviour 69*: 303-315. http://dx.doi.org/10.1163/156853979X00520

Gomez A and Cook NB 2010 Time budgets of lactating dairy cattle in commercial free-stall herds. *Journal of Dairy Science 93*: 5772-5781. http://dx.doi.org/10.3168/jds.2010-3436

Kazdin AE 1977 Artifact, bias and complexity of assessment: ABCs of reliability. *Journal of Applied Behavior Analysis 10*: 141-150. http://dx.doi.org/10.1901/jaba.1977.10-141

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 1993 *Measuring Behaviour, An Introductory Guide*. Cambridge University Press: Cambridge, UK

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

Napolitano F, De Rosa G, Caporale G, Carlucci A, Grasso F and Monteleone E 2007 Bridging consumer perception and on-farm assessment of animal welfare. *Animal Welfare 16*: 249-253

Plesch G, Broerkens N, Laister S, Winckler C and Knierim U 2010 Reliability and feasibility of selected measures concerning resting behaviour for the on-farm welfare assessment in dairy cows. *Applied Animal Behaviour Science 126*: 19-26. http://dx.doi.org/10.1016/j.applanim.2010.05.003

Uzal S and Ugurlu N 2010 The dairy cattle behaviors and time budget and barn area usage in freestall housing. *Journal of Animal and Veterinary Advances 9*: 248-254. http://dx.doi.org/10.3923/javaa.2010.248.254

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

Wemelsfelder F and Lawrence AB 2001 Qualitative assessment of animal behaviour as an on-farm welfare-monitoring tool. *Acta Agriculturae Scandinavica, Section A, Animal Science 30*: 21-25

Wemelsfelder F, Hunter EA, 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, Millard F, De Rosa G and Napolitano F 2006 Qualitative indicators for the on-farm monitoring of cattle welfare. *EU deliverable D2.18.10, subtask 2.2.4* 29

Windschnurer I, Schmied C Boivin X and Waiblinger S 2008 Reliability and inter-test relationship of tests for on-farm assessment of dairy cows' relationship to humans. *Applied Animal Behaviour Science 114*: 37-53. http://dx.doi.org/10.1016/j.applanim.2008.01.017

https://doi.org/10.7120/09627286.21.3.307 Published online by Cambridge University Press