1-15-2013

The Correlation of Qualitative Behavior Assessments with Welfare Quality® Protocol Outcomes in On-Farm Welfare Assessment of Dairy Cattle

S. N. Andreasen
University of Copenhagen

F. Wemelsfelder
Scottish Agricultural College

P. Sandøe
University of Copenhagen

B. Forkman
University of Copenhagen

Recommended Citation

Andreasen, S. N., Wemelsfelder, F., Sandøe, P., & Forkman, B. (2013). The correlation of Qualitative Behavior Assessments with Welfare Quality® protocol outcomes in on-farm welfare assessment of dairy cattle. Applied animal behaviour science, 143(1), 9-17.

The Correlation of Qualitative Behavior Assessments with Welfare Quality® Protocol Outcomes in On-Farm Welfare Assessment of Dairy Cattle

S.N. Andreasen[1], F. Wemelsfelder[2], P. Sandøe[1], and B. Forkman[1]

[1] *University of Copenhagen*
[2] *Scottish Agricultural College*

KEYWORDS

Welfare Quality® protocol, Qualitative Behavior Assessment, welfare assessment, dairy cow

ABSTRACT

The effort to develop methods for assessing animal welfare at farm level has grown dramatically since the end of the 1990s, culminating in the protocols developed by the European-wide project Welfare Quality® (WQ). However, these protocols are time consuming and lack transparency in how scores are aggregated into welfare outcomes. The current study investigates the potential of Qualitative Behavior Assessment (QBA), a much less time-consuming approach, to be used as a stand-alone integrative screening tool for identifying farms with compromised welfare before applying the full WQ protocol. QBA is a 'whole-animal' approach asking human observers to summarize animals' expressive demeanor and its context into descriptors such as relaxed, anxious, content or frustrated –terms which given their emotional connotation appear to have direct relevance to animal welfare. Two trained QBA-assessors, and one trained Welfare Quality® assessor visited 43 Danish dairy cattle farms at different times, the former focusing on QBA and the latter making a full WQ protocol assessment. The QBA scores were analyzed using Principal Component Analysis (correlation matrix, no rotation), and WQ protocol data were analyzed and integrated according to the WQ protocol. The resulting QBA and WQ protocol outcomes were correlated using non-parametric methods (Spearman Rank and Kendall W). Highly significant inter-observer agreement was found between the two QBA-assessors ($P < 0.0001$). QBA scores showed some weak correlations to WQ measures but no meaningful pattern of relationship between these measures emerged. The present study does not support the application of QBA as a stand-alone welfare assessment tool capable of predicting the outcome of the larger WQ protocol.

1. Introduction

The effort to develop methods for assessing animal welfare at farm level has grown dramatically since the end of the 1990s. A culmination of these efforts can be witnessed in the protocols developed by the EU-project Welfare Quality® (WQ; Keeling, 2009). The WQ protocols' basic premise is that welfare is related

to an animal's mental states, and that welfare indicators should therefore be primarily animal-based and be aggregated to reflect an animal's perception of its situation (Botreau et al., 2007). The protocols cover different aspects of animal welfare using 12 different criteria (consisting of one or more welfare assessment measures), which are added up to four principles, and finally to one overall score for animal welfare on-farm. These protocols consist of measures that have been tested for feasibility, inter-observer reliability and validity, though not all of these tests were done on farm. Thus the WQ protocol can be assumed to have relatively high validity as an on-farm welfare assessment tool. Assessors wanting to use the protocols have to attend courses learning how to perform the assessments (Forkman & Keeling, 2009; Anonymous, 2011a).

However, many welfare assessment protocols have two main limitations from a practical point of view. Firstly, the methods used are becoming more and more time consuming: an evaluation of an ordinary dairy farm with 200 dairy cows using the Welfare Quality® protocol takes around seven to eight hours by specially trained staff (Welfare Quality, 2009). This makes it unlikely that the method will be widely used and be of practical importance to ordinary farmers. Secondly, when talking about the Welfare Quality® protocols, the way the measures relating to the criteria and principles are added together to a single figure is not very transparent (Veissier et al., 2011), and this is likely to generate discussions about the fairness and relevance of the welfare assessments based on this tool.

It would be useful in this context to find methods that could summarize the Welfare Quality® protocol or key aspects of it, and thus serve as a screening tool for identifying farms with compromised welfare before applying the full WQ protocol. One method which may be suitable for this purpose is Qualitative Behavior Assessment (QBA), a method developed by Wemelsfelder and co-workers which relies on a human observer's ability to integrate perceived details of an animal's demeanor and its context, using 'whole animal' descriptors such as calm, playful, content, indifferent or frustrated (Wemelsfelder et al., 2000, 2001, 2009c). Such terms specify not so much what an animal does physically, as how it expresses itself dynamically (Feaver et al., 1986), and as such have an emotional connotation that has been suggested to be directly relevant to its welfare (Wemelsfelder, 1997, 2007).

Based on these premises, Wemelsfelder et al. (2000, 2001) have argued that QBA could potentially function as an integrative welfare assessment tool, complementing existing tools that focus on specific aspects of behavior and welfare such as abnormal behavior or lameness (QBA is also included as one measurement in the Welfare Quality® protocol). The QBA approach has two obvious prima facie advantages compared to more traditional protocols such as the WQ protocol: Firstly, using QBA is not very time-consuming – as it stands on-farm assessment takes at most 30 min; secondly, like WQ it strongly focuses on the animal, and given its reliance on 'whole-animal' observation, avoids the theoretically contentious issue of how to merge scores on different aspects of welfare into a single measure. However, qualitative characterisations of animal expression have long been dismissed by scientists as anthropomorphic and unscientific (Wemelsfelder, 2007), creating the need to address the general reliability and validity of this approach.

QBA was first explored in 2000 using growing pigs and 18 observers. The observations were done on living pigs entering a trial pen (Wemelsfelder et al., 2000). Most of the subsequent studies have focused on video clips of single animals but some later studies have also used live animals as well as groups of animals. The studies include video-clips of a range of production animals (pigs, cattle and poultry) and of dogs and horses, shown in varying situations such as experimental or on-farm home pen, human–animal interaction, and live transport. Observers in these studies ranged from having none to extensive experience with the animals being observed. The outcomes of these previous studies show QBA to generally have good inter-observer reliability, and to correlate significantly with measures of individual animals' physical behavior and/or physiological stress responses (Wemelsfelder et al., 2001; Rousing &

Wemelsfelder, 2006; Napolitano et al., 2008; Minero et al., 2009; Wemelsfelder and Millard, 2009; Wemelsfelder et al., 2009a,b,c; Walker et al., 2010; Stockman et al., 2011; Rutherford et al., 2012). One on-farm study of veal calves in addition found a positive relationship between a fear/agitation QBA component and the frequency of visits by unfamiliar visitors such as veterinarians, and between an activity/play component and frequency of visits by the familiar farmer, suggesting that QBA may also be sensitive to the quality of human contact (Brscic et al., 2009).

The objective of the present study was to investigate whether and how outcomes provided by the Qualitative Behavior Assessment of dairy cattle on 43 Danish farms would correlate with the overall score, the principles and/or the individual criteria of the Welfare Quality® protocol developed for dairy cattle. If significant and meaningful patterns of correlation between these measures are found then this would support QBA's use as a stand-alone screening tool for welfare assessment predicting the WQ protocol's outcome.

2. Materials and methods

This study is build up as a cross-sectional observational study where 44 Danish dairy farms are assessed using two welfare assessment protocols on each farm. The two protocols used are Qualitative Behavior Assessment and the Welfare Quality® protocol. Only adult dairy cattle were assessed, so calves and young stock were left out. All farms were assessed during the autumn and winter of 2010/2011 (October–March) when all animals were housed indoors.

2.1. Qualitative Behavior Assessment

QBA was in the first instance developed using a Free Choice Profiling (FCP) methodology that relies on larger groups of observers developing their own personal descriptors (Wemelsfelder et al., 2001). Because this is impractical in applied settings and difficult to standardize, a fixed-list approach to QBA was subsequently developed for the Welfare Quality® project. For this a set of 20 qualitative descriptors were selected from previous FCP research on dairy cattle, and subsequently tested for inter-observer reliability on farms in various European countries (Wemelsfelder et al., 2009b). At the request of the QBA-assessors in the current study one further term, 'aggressive', was added to this list. This fixed list was then translated into Danish (Table 1).

All the terms were connected to a Visual Analog Scale (VAS-scale) of 125 mm where minimum and maximum represented the ends of the scale. The two assessors were instructed in how to use the scale. The assessment of the animals took place from the feeding table. The animals were observed for 20 min during which the QBA-assessors were allowed to freely move around on the feeding table and choose as many observation points as they felt necessary, as long as they reached both ends of the feeding table and as long as the observation time and points were equally distributed across the feeding table. If the stable was divided so that two feeding tables were present the QBA-assessors were instructed to spend 10 min on each feeding table. After observing the animals for 20 min the two QBA-assessors went to a place where the dairy cows were not visible, where they scored the animals using the VAS-scale for the 21 qualitative terms. All QBA-assessments were done between 09:40 and 16:45 with noon being the mean, and the assessors were always dressed in the same type and color of clothes at all farms.

2.1.1. Qualitative Behavior Assessment assessors

Two assessors conducted the Qualitative Behavior Assessment.

Assessor one (QBA-obs1): Female, educated as a farmer and later as an agricultural technologist. This assessor had been working with dairy cows for 12 years and had no prior experience with welfare assessment protocols.

Assessor two (QBA-obs2): Female, educated as a farmer and later as a farm manager and agricultural technician. This assessor had been working with animals for 35 years, primarily dairy cows, but also pigs, and had experience with welfare assessment protocols.

Table 1. Terms used by the two QBA-assessors in the Qualitative Behavior Assessment.

Terms used in the Qualitative Behavior Assessment (Danish word in Brackets)	
Active (Aktiv)	Positively occupied (Positivt beskæftiget)
Relaxed (Afslappet)	Lively (Livlig)
Fearful (Frygtsom)	Inquisitive (Nysgerrig)
Agitated (Ophidset)	Irritable (Irritabel)
Calm (Rolig)	Uneasy (Urolig)
Content (Tilfreds)	Sociable (Social)
Indifferent (Ligeglad)	Apathetic (Apatisk)
Frustrated (Frustreret)	Happy (Glad)
Friendly (Venlig)	Distressed (Trist)
Bored (Keder sig)	Aggressive (Aggressiv)
Playful (Legesyg)	

2.1.2. Training of the QBA-assessors

The two QBA-assessors had a one day introduction course where they were introduced to the concept behind Qualitative Behavior Assessment. They were made familiar with the terms used in the method and were given time to discuss the meaning of the terms so that a consensus about the meanings was reached. They were introduced to the VAS-scale and how to use it. The assessors tried out the method using 16 video clips showing dairy cattle. Many of these clips were the same as those used during the training of assessors for the Welfare Quality® protocol. After clip one, two and three the assessors were allowed to talk about their assessment and compare their results. Possible disagreements and reasons for disagreements were discussed. During the last 13 clips they were not allowed to discuss until after the final clip. After introduction and training they were told about the setup of the study.

2.2. The Welfare Quality® protocol

A full welfare assessment was made on each farm using the Welfare Quality® protocol for dairy cattle (Welfare Quality, 2009). The Welfare Quality® protocol for dairy cattle consists of 29 validated measures which are carried out on-farm (Forkman & Keeling, 2009). The 29 measures are used to calculate 12 criteria (one of which, C12, is QBA), which are aggregated into four principles, and these principles are then translated into a final result that classifies farms as either “Not classified”, “Acceptable”, “Enhanced” or “Excellent”. Criteria and principles are also individually calculated to reach one of the four scores. Table 2 summarizes the WQ protocol.

On-farm assessment was done by the WQ-assessor and started approximately 15 min after morning feeding. For the farms participating in this study assessment started between 4:15 am to 9:00 am. Before starting the assessment a sketch of the stable outline where the dairy cows were held was made, this was to be used in defining stable segments for use later in the protocol. Hereafter the assessment

started. At first the avoidance distance at the feeding table was assessed. The avoidance distance test was followed by the QBA protocol developed for Welfare Quality®. After this the rest of the WQ protocol was employed. For a full description of the Welfare Quality® protocol and assessment see Welfare Quality®, 2009.

2.2.1. Welfare Quality® assessor

One assessor, WQ-obs, conducted the Welfare Quality® assessment. This assessor was female, educated as a veterinarian with work experience from cattle practice and experience as a control officer at the authorities.

2.2.2. Training of the WQ-assessor

The WQ-assessor took part in a four day course in Lyon, France where she together with other new assessors was trained thoroughly in the methods of dairy cattle welfare assessment using the Welfare Quality® protocol. This course is mandatory for using the Welfare Quality® protocol and consists of both theory and practical assessment.

2.3. Farms

Forty-four Danish dairy farms with loose-housed Danish Holstein-Friesian cows participated in this study. Only farms with more than 50 cows were included. In 2009 approximately 70% of all Danish dairy cows were Danish Holstein-Friesian cows (Anonymous, 2009), therefore only farms with this breed was included. A list containing 63 farms was generated on the basis of the above mentioned criteria. All 63 farms were contacted by telephone to see if they were interested in participating. If they were interested a letter describing the project was sent to them. All 63 farmers agreed to receive more information about the study. A week after receiving the letter the farmers were once again contacted by telephone to see whether or not they would participate in the study. Forty-four farms wished to participate; this gave a positive response-rate of 70%. Forty of the farms were situated in Jutland, four on Funen. The total number of cows was 8106 cows with a mean of 184 cows on each farm (min 101, max 452). Two of the farms were organic. Four of the farms had deep bedding the rest used cubicles, here 12 used rubber mats as bedding material, 17 used mattresses, eight used sand, two used latex and one used straw. Twenty-seven of the farms milked in parlor, 17 used automatic milking i.e. by means of a robot.

Before the visit all farmers were forwarded a questionnaire either by e-mail or ordinary mail. Here they were asked questions concerning their management of the farm.

2.3.1. Farm visits

Each of the 44 farms received two visits, one from the two QBA-assessors and one from the WQ-assessor. The visits could either be on different days or on the same day. If the farm was visited by both types of assessors on the same day it was made sure that the WQ-assessor had finished her assessment well in advance of the two QBA-assessors arriving. The longest period of time between the two visits was five days, mean time between Welfare Quality® and QBA visits were 1.5 days (min 0, max 5).

2.4. Statistical analysis

For statistical analysis farm 24 was left out because QBA-obs1 did not do the Qualitative Behavior Assessment as the protocol prescribed. This left a total of 43 farms. For statistical analysis the software SAS JMP© and SAS 9.2© (SAS Institute Inc., SAS Campus Drive, Cary, North Carolina) was used. The

Welfare Quality® calculations were done according to the outline in the protocol (Welfare Quality, 2009). For calculation the data was send off to INRA (Institut National de la Recherche Agronomique), France.

Table 2. Structure of the Welfare Quality® protocol.

Criteria 1 (C1) – Absence of prolonged hunger	Principle 1 (P1) – Good feeding	Overall assessment
Criteria 2 (C2) – Absence of prolonged thirst		
Criteria 3 (C3) – Comfort around resting	Principle 2 (P2) – Good housing	
Criteria 4 (C4) – Thermal comfort (no measure)		
Criteria 5 (C5) – Ease of movement	Principle 3 (P3) – Good health	
Criteria 6 (C6) – Absence of injuries		
Criteria 7 (C7) – Absence of disease		
Criteria 8 (C8) – Absence of pain induced by management procedures		
Criteria 9 (C9) – Expression of social behaviors	Principle 4 (P4) – Appropriate behavior	
Criteria behavior 10 (C10) – Expression of other behaviors		
Criteria 11 (C11) – Good human–animal relationship		
Criteria 12 (C12) – Positive emotional state		

2.4.1. Qualitative Behavior Assessment

For each farm three Qualitative Behavior Assessments were made, two from the independent QBA-assessors and one from the WQ-assessor. QBA-obs1 and QBA-obs2 did the Qualitative Behavior Assessment at exactly the same time at each farm, but were instructed to do so strictly independently from each other. WQ-obs did the Qualitative Behavior Assessment as a part of the Welfare Quality® protocol.

For the purpose of data analysis, for each of the words the distance from minimum to where the assessor ticked the VAS-scale was measured in mm. The data was handled using Principal Component Analysis (PCA), using a correlation matrix with no rotation. PCA was firstly done on individual assessor data for the three assessors separately, to see whether or not the QBA-assessors independently of each other would describe the farms using the same dimensions. The word aggressive was left out to enable comparison between assessors 1 and 2 on the one hand and 3 on the other hand. However, to be able to correlate each assessor's outcomes, their data were also analyzed as part of one PCA, to ensure that the components generated by individual assessors were aligned. This was done both for the two QBA assessors together, and for the two QBA assessors and the WQ assessor together.

2.4.2. Welfare Quality®

The data collected on farm was typed into an excel sheet for each farm. The data from each farm was randomly checked for entry errors. For each farm the data for the Welfare Quality® protocol was summed up using an excel sheet provided by INRA. All the Welfare Quality® data was sent off to INRA and the analyzed results were sent back.

2.4.3. Comparing Qualitative Behavior Assessment and Welfare Quality®

Inter-observer agreement between the assessors' QBA scores was calculated using a Spearman Rank Correlation Coefficient if two assessors were compared or Kendall's Coefficient of Concordance W for three assessors (i.e. the two QBA-assessors and the WQ-assessor).

To assess whether or not QBA can be used as a screening tool QBA scores on the two main PCA components for the 43 farms, both for the separate assessor analyses and for the aligned analyses, were correlated to the farm scores obtained in the Welfare Quality® protocol. This was done for the 12 criteria, the four principles and the overall assessment, using Spearman Rank Correlation Coefficient.

3. Results

3.1. Qualitative Behavior Assessment

3.1.1. Observers individually

Using the cumulative percentage and the scree-plot two principle components (PC1 and PC2) were found to explain the majority of variation between farms, both for QBA-obs1 and QBA-obs2 (cumulative percentage = 52.3% and 61.8% respectively; QBA-obs1: PC1 36.3%, PC2 16.0%, QBA-obs2: PC1 46.2%, PC2 15.6%). The terms with the highest loadings on these dimensions for each observer can be found in Table 3. As can be seen in the table there is a good agreement between the two observers in how they characterize the two principle components.

Table 3. Terms and loadings describing Principal Components 1 and 2 for QBA-obs1 and QBA-obs2. Terms which load more than ± 0.24 are reported in the table. (±0.24 is used as cutoff since no rotation was used).

	QBA-Obs1	**QBA-Obs2**
Principal component 1	Calm (−0.32)	Calm (−0.27)
	Relaxed (−0.32)	Relaxed (−0.27)
	Friendly (−0.25)	Content (−0.26) Friendly (−0.24)
	Uneasy (0.28)	Positively occupied (−0.24)
	Frustrated (0.28)	
	Agitated (0.26)	Uneasy (0.29)
	Distressed (0.25)	Frustrated (0.28)
	Active (0.25)	Agitated (0.27)
	Lively (0.25)	Irritable (0.26)
	Irritable (0.25)	Active (0.24)
	Fearful (0.24)	
Principal component 2	Bored (−0.39)	Indifferent (−0.36)
	Indifferent (−0.38)	Distressed (−0.31)
	Positively occupied (0.36)	Inquisitive (0.31)
	Content (0.32)	Playful (0.30)
	Lively (0.30)	Happy (0.27)
	Active (0.29)	Lively (0.27)
	Playful (0.24)	Active (0.24)

3.1.2. QBA-obs1 and QBA-obs2 analyzed together

For the PCA encompassing both QBA-obs1 and QBA-obs2 scores, two main dimensions were found using cumulative percentage and the scree-plot. These dimensions explained 56.3% of the variation between farms (PC1: 44.1%, PC2: 12.2%). Table 4 shows terms which load more than ±0.24 on each

Principal Component, indicating that PC1 can be described as ranging from calm/relaxed to uneasy/agitated, and PC2 from indifferent/distressed to lively/playful.

It is assumed that a good farm is described on Principal Component 1 by a negative score and on Principal Component 2 by a positive score. The Spearman Rank correlations between QBA-obs1 – and QBA-obs2 – farm scores on Principal Component 1 and Principal Component 2 were 0.72 ($P < 0.0001$) and 0.56 ($P < 0.0001$) respectively.

3.2. Welfare Quality

As mentioned previously the results for the 43 farms are given on three levels (criteria, principle and overall) and the farms could be assigned one of four possible assessments ("Not Classified", "Acceptable", "Enhanced" or "Excellent") depending on the score reached. For results see Table 5.

When looking at criterion five there is no spread between the farms, that is all the farms are scored to be "Excellent". Criterion five is describing "Ease of movement" and is composed of the measures "Presence of tethering" and "Access to outdoor loafing area or pasture". None of the farms were tie stalls, so the score for all farms regardless of access to outdoor loaf or pasture are 95 according to the guidelines on how to calculate the scores (Welfare Quality, 2009). Also in criteria eight and nine the farms are given the same score. Criterion eight is describing "Absence of pain induced by management procedures" and is composed of the measures "Disbudding/dehorning" and "Tail docking". In Denmark tail docking is prohibited by law and therefore all the farms automatically score 100. Also there are distinct rules concerning dehorning/disbudding. The rules concern both age of the calves and provision of anesthesia and analgesia, therefore the farms all receive the same score. Criterion nine is describing "Expression of social behaviours", this criterion is composed of the measure "Agonistic behaviour". All the farms score "Excellent" because not much agonistic behavior was present on any of the farms. When looking at criterion 10 it was found that 29 of the 43 farms scored "Not classified", one scored "Acceptable" and 13 scored "Enhanced". This criterion describes "Expression of other behaviours" and consists of the measure "Access to pasture". On the 29 farms scoring "Not classified" the cows did not in any season of the year have access to pasture, on the farm scoring "Acceptable" the cows had access to pasture, but for a more limited amount of time than the cows from the farms scoring "Enhanced".

3.3. Qualitative Behavior Assessment and Welfare Quality®

Table 6 shows the Spearman rank correlations of QBA-obs1 and QBA-obs2 in both the separate and combined analyses, and with the different Welfare Quality criteria. Because there is no measure for criterion four and because criteria 5, 8 and 9 showed no variability, a comparison with these criteria was not possible. For a comparison between the two QBA assessors' scores and Welfare Quality criterion 12, which is also based on QBA, see below in paragraph 3.4.

The results show that neither QBA-obs1 nor QBA-obs2 reach a significant correlation with the Welfare Quality Overall Assessment (OA) score, although QBA-obs1 PC2 scores shows a tendency toward correlating significantly with OA in both the individual and aligned analysis. No significant correlation occurs with the four principles or the twelve criteria when these are tested separately against the individual Principal Components PC1 and PC2, although there is a tendency for QBA-obs1 PC1 to correlate with C3 ("Comfort around resting"). When the results of the aligned Principal Components were tested against the results of the Welfare Quality® three significant correlations were found; for QBA-obs1 PC2 with C2 ("Absence of prolonged thirst") and for both observers' PC2 with P1 ("Good feeding").

Table 4. Terms and loadings describing Principal Component 1 and Principal Component 2 for the aligned analyses of QBA-obs1 and QBA-obs2. Terms which load more than ±0.24 are reported in the table.. (±0.24 is used as cutoff since no rotation was used).

	Principal Component 1	Principal Component 2
Negative loading	Calm (−0.28)	Indifferent (−0.41)
	Relaxed (−0.27)	Distressed (−0.31)
	Friendly (−0.25)	Bored (−0.25)
	Content (−0.24)	
Positive loading	Uneasy (0.28)	Lively (0.29)
	Agitated (0.27)	Playful (0.28)
	Irritable (0.26)	Inquisitive (0.27)
	Frustrated (0.24)	Active (0.27)

Table 5. Distribution of the Welfare Quality® scores and assessment achieved by the 43 farms, for the twelve criteria, the four principles and the Overall Assessment.

Criteria, principles and overall assessment	Score min	Score max	Std. deviation	Welfare assessments
Criteria 1 (C1)	30	100	22.53	Acceptable to Excellent
Criteria 2 (C2)	3	100	38.21	Not classified to Excellent
Criteria 3 (C3)	16	54	12.00	Not classified to Acceptable
Criteria 4 (C4)	No measure			
Criteria 5 (C5)	95	95	0	Excellent
Criteria 6 (C6)	12	71	15.10	Not classified to Enhanced
Criteria 7 (C7)	37	86	13.49	Acceptable to Excellent
Criteria 8 (C8)	58	58	0	Enhanced
Criteria 9 (C9)	100	100	0	Excellent
Criteria 10 (C10)	0	73	32.46	Not classified to Enhanced
Criteria 11 (C11)	22	78	14.08	Acceptable to Enhanced
Criteria12 (C12)	3	89	23.03	Not classified to Excellent
Principle 1 (P1)	7	100	24.84	Not classified to Excellent
Principle 2 (P2)	46	69	7.33	Acceptable to Enhanced
Principle 3 (P3)	23	63	10.78	Acceptable to Enhanced
Principle 4 (P4)	14	71	14.37	Not classified to Enhanced
Overall assessment (OA)	1	3	0.55	Not classified to Enhanced

3.4. Qualitative Behavior Assessment done by WQ-obs

In the Welfare Quality® protocol the measure used in criterion 12 is Qualitative Behavior Assessment. This measure is as mentioned previously done in the beginning of the data collection on-farm, just after the avoidance distance test. For the sake of this study, these scores were also used separately from the WQ protocol to investigate whether and how they agreed with the scores of QBA-obs1, QBA-obs2 and with the Welfare Quality® outcomes. The main loadings of this PCA are shown in Table 7, indicating PC1 to range from bored/indifferent to happy/content, and PC2 from calm/relaxed to active/lively.

Table 6. The Spearman Rank correlations (ρ) between the Welfare Quality® protocol and the Qualitative Behavior Assessment for the two QBA assessors individually and for their aligned QBA assessment. 1: Describing the parts of the Welfare Quality® protocol which were used in the analysis, that is the criteria: C1 ("Absence of prolonged hunger"), C2 ("Absence of prolonged thirst"), C3 ("Comfort around resting"), C6 ("Absence of injuries"), C7 ("Absence of disease"), C10 ("Expression of other behaviours") and C11 ("Good human–animal relationship"), and the principles: P1 ("Good feeding"), P2 ("Good housing"), P3 ("Good health") and P4 ("Appropriate behaviour") and the Overall Assessment: OA. 2: PC1 – Principal Component 1. 3: PC2 – Principal Component 2. *: Significant findings (P<0.05).

WQ[1]	Observers individually				Observers aligned			
	QBA-obs1		QBA-obs2		QBA-obs1		QBA-obs2	
	PC1[2]	PC2[3]	PC1	PC2	PC1	PC2	PC1	PC2
C1	ρ=–0.1580 *P*=0.3115	ρ=–0.0011 *P*=0.9946	ρ=0.1426 *P*=0.3617	ρ=0.1321 *P*=0.3983	ρ=0.1834 *P*=0.2393	ρ=0.0734 *P*=0.6399	ρ=0.1380 *P*=0.3776	ρ=0.1316 *P*=0.4002
C2	ρ=–0.1223 *P*=0.4345	ρ=0.2828 *P*=0.0661	ρ=–0.1701 *P*=0.237	ρ=0.1415 *P*=0.3654	ρ=–0.1563 *P*=0.3169	ρ=*0.3033* *P*=0.0480*	ρ=–0.1913 *P*=0.2191	ρ=0.1536 *P*=0.3253
C3	ρ=–0.2991 *P*=0.0513	ρ=0.1403 *P*=0.3696	ρ=–0.2562 *P*=0.0973	ρ=–0.0550 *P*=0.7260	ρ=–0.2809 *P*=0.0680	ρ=0.1526 *P*=0.3285	ρ=–0.2523 *P*=0.1026	ρ=–0.0444 *P*=0.7775
C6	ρ=–0.1016 *P*=0.5168	ρ=0.1978 *P*=0.2035	ρ=–0.0465 *P*=0.7670	ρ=0.1542 *P*=0.3234	ρ=–0.0555 *P*=0.7236	ρ=0.2182 *P*=0.1597	ρ=–0.0437 *P*=0.7807	ρ=0.1350 *P*=0.3882
C7	ρ=–0.0929 *P*=0.5536	ρ=–0.0927 *P*=0.5543	ρ=–0.1674 *P*=0.2834	ρ=0.0013 *P*=0.9933	ρ=–0.0618 *P*=0.6938	ρ=0.0503 *P*=0.7487	ρ=–0.1732 *P*=0.2666	ρ=–0.0641 *P*=0.6829
C10	ρ=–0.1086 *P*=0.4883	ρ=0.0644 *P*=0.6815	ρ=0.0649 *P*=0.6793	ρ=0.1408 *P*=0.3679	ρ=–0.0925 *P*=0.5550	ρ=0.0422 *P*=0.7881	ρ=0.0602 *P*=0.7012	ρ=0.0929 *P*=0.5535
C11	ρ=–0.0354 *P*=0.8219	ρ=0.0160 *P*=0.9188	ρ=0.0799 *P*=0.6104	ρ=0.2634 *P*=0.0879	ρ=–0.0054 *P*=0.9724	ρ=0.0279 *P*=0.8592	ρ=0.0070 *P*=0.6237	ρ=0.2372 *P*=0.1257
P1	ρ=–0.0113 *P*=0.9428	ρ=0.261	ρ=–0.0625 *P*=0.6906	ρ=0.2873 *P*=0.0618	ρ=–0.0338 *P*=0.8296	ρ=0.3179 *P*=0.0378*	ρ=–0.1017 *P*=0.5162	ρ=0.3028 *P*=0.0484*
P2	ρ=–0.2991 *P*=0.0513	ρ=0.1403 *P*=0.396	ρ=–0.2562 *P*=0.0973	ρ=–0.0550 *P*=0.7260	ρ=–0.2809 *P*=0.0680	ρ=0.1526 *P*=0.3285	ρ=–0.2523 *P*=0.1026	ρ=–0.0444 *P*=0.7775
P3	ρ=–0.0845 *P*=0.5901	ρ=0.1607 *P*=0.3032	ρ=–0.0707 *P*=0.6522	ρ=0.0977 *P*=0.5330	ρ=–0.0361 *P*=0.8183	ρ=–0.2237 *P*=0.1492	ρ=–0.0744 *P*=0.6353	ρ=0.0674 *P*=0.6676
P4	ρ=–0.1286 *P*=0.4112	ρ=0.1646 *P*=0.2914	ρ=0.0314 *P*=0.8413	ρ=0.2331 *P*=0.1326	ρ=–0.1017 *P*=0.5165	ρ=0.1586 *P*=0.3097	ρ=0.0084 *P*=0.9574	ρ=0.1973 *P*=0.2047
OA	ρ=–0.1649 *P*=0.2907	ρ=0.2917P=0.0846 *P*=0.0577	ρ=–0.2319 *P*=0.1345	ρ=0.1558 *P*=0.3184	ρ=–0.1504 *P*=0.3358	ρ=0.2663 *P*=0.0843	ρ=–0.2500 *P*=0.1059	ρ=0.1703 *P*=0.2749

To align the PCA dimensions of WQ-obs with QBA-obs1 and QBA-obs2, WQ QBA scores were also analyzed with those of the other two observers in one PCA, with PC1 ranging from calm/content to agitated/irritable, and PC2 from indifferent/distressed to playful/lively. Correlating the scores of the 3 assessors in this PCA showed a statistically significant agreement of Kendall's $W = 0.55$ ($\chi^2 = 68.81$, $P < 0.01$) for PC1 and $W = 0.58$ ($\chi^2 = 72.57$, $P < 0.005$) for PC2. Looking in more detail at pairs of observers it emerged that only QBA-obs1 and QBA-obs2 agreed significantly with each other on both dimensions (Dimension one: QBA-obs1:QBA-obs2 – $\rho = 0.7147$, $P < 0.0001$; QBA-obs1:WQ-obs – $\rho = 0.1892$, $P = 0.2243$; QBA-obs2:WQ-obs – $\rho = 0.0464$, $P = 0.7678$. Dimension two: QBA-obs1:QBA-obs2 – $\rho = 0.6179$, $P < 0.0001$; QBAobs1:WQ-obs – $\rho = 0.2756$, $P = 0.0737$; QBA-obs2:WQ-obs – $\rho = 0.2910$, $P = 0.0583$).

Table 8 shows that the QBA results on PC1 for the WQ-assessor (analyzed separately as in Table 7) were significantly correlated with criteria 1, 7 and 11, and with Principle 4. Only PC1 scores were used

here because, when QBA scores are analyzed as part of the Welfare Quality® database then only PC1 scores are used.

Table 7. Terms and loadings describing Principal Component 1 and Principal Component 2 for the WQ-obs. Terms which load more than ±0.24 are reported in the table. (±0.24 is used as cutoff since no rotation was used).

	WQ-Obs
Principal Component 1 50.4%	Bored (−0.27) Indifferent (−0.25) Agitated (−0.24) Uneasy (−0.24) Happy (0.29) Content (0.26) Sociable (0.26) Friendly (0.24)
Principal Component 2 13.3%	Calm (−0.28) Relaxed (−0.20) Active (0.45) Lively (0.37) Playful (0.35) Irritable (0.30) Frustrated (0.25)

Table 8. Correlations that reached significance between QBA scores of WQ-obs and WQ protocol outcomes.

WQ	**WQ-obs Correlation between QBA and WQ Principal Component 1**
C1	ρ=0.3141; P=0.0402[a]
C7	ρ=0.3337; P=0.0288[a]
C11	ρ=0.5149; P=0.0004[a]
P4	ρ=0.4453; P=0.0028[a]

[a] Significant at 0.05 level.

4. Discussion

The aim of the current study was to investigate whether Qualitative Behavior Assessment (QBA) could potentially be used as a screening tool for summarizing the large, time-consuming Welfare Quality® protocol, to identify farms with compromised welfare requiring further inspection. This was done by correlating the scores provided by two QBA assessors, who visited 43 dairy farms at the same time but scored them independently, to the outcomes of the Welfare Quality® assessment protocol applied by one Welfare Quality observer to the same farms. Highly significant inter-observer agreement was found between the two QBA-assessors, but no significant correlations were found between their assessments and the final overall score of the Welfare Quality® protocol. Comparison to the principles and criteria of the WQ protocol separately showed only two significant positive correlations: Principle 1 (Good feeding) with PCA dimension two (indifferent/distressed – lively/playful) for both QBA-obs1 and QBA-obs2, and criterion 2 (Absence of prolonged thirst) with PCA dimension two (indifferent/distressed – lively/playful) for QBA-obs1 only. However none of these correlations were very strong. The QBA done by the WQ assessor as part of the Welfare Quality® protocol did show several significant positive correlations of PC1 (bored/indifferent – happy/content) with the rest of the Welfare Quality® protocol, including for example

criterion 11 (Human–animal relationship) and Principle 4 (Positive emotional state). In this case the QBA and Welfare Quality® assessments were made by the same assessor (WQ-assessor), and even though the QBA assessments were made prior to the other WQ assessments (with the exception of the human–animal relationship), this observer had been highly trained in using the Welfare Quality® protocol, which makes it difficult to assume independence between the QBA and WQ measurements.

Thus, the current study did not find evidence for a meaningful pattern of correlation between QBA and WQ assessments, and therefore does not support that QBA as a stand-alone tool could potentially summarize welfare outcomes provided by the WQ protocol. The absence of such correlations raises the issue of how QBA's welfare content compares to that of the WQ protocol –QBA may either reflect on-farm animal welfare differently, or to a lesser degree.

There may be a variety of reasons of a theoretical or practical nature for why so few meaningful correlations were found. Firstly, one could question the extent to which the Welfare Quality® protocol itself can be trusted to reflect the animals' overall welfare state. The method for aggregating measures into an overall score in the Welfare Quality® protocol is based on expert assessments and therefore possibly biased. However, this would not explain why there were so few significant correlations between QBA and the individual criteria and principles of the WQ protocol. The measures constituting WQ criteria have all been through a thorough revision, both as concerns their validity as well as their reliability (Forkman & Keeling, 2009). Some measures were not validated on farm, and it may be that under variable on-farm conditions these measures did not perform as expected. Generally it seems reasonable to suppose that the criteria used in the WQ protocol are meaningful indicators of animal welfare that can be used to evaluate other measures of animal welfare.

It could be argued that since QBA addresses the 'whole animal' (Wemelsfelder et al., 2001), it is a question of finding the corresponding weightings that sum up WQ measures into criteria, principles and overall assessment. However, the almost total absence of strong significant correlations makes it improbable that there is a simple way of combining criterion values to improve the present results, and so the question remains whether the two methods might differ quite fundamentally in their integration/evaluation of measured/perceived details of behavior and demeanor.

One could suppose in this light that QBA addresses animal welfare from a different angle than the Welfare Quality® protocol. QBA, using expressive whole-animal descriptors such as anxious or content purports to give an account of the actual experience of the animals. It may succeed in doing this and it may even give a good indication of the presence of these states viewed over longer stretches of time. However, the relationship between experience and welfare is not straightforward (Fraser, 2003; Appleby and Sandoe, 2002). Good animal welfare seems on most accounts not only to imply that the animals as a matter of fact are content and not anxious. It also seems to imply that the animals are provided with resources which in a robust way will secure this. The final version of the Welfare Quality® protocol, for this reason also includes some resource-based elements, and therefore strictly speaking cannot be said to be purely animal-based; this may be part of the explanation of why QBA and WQ outcomes do not seem to align.

Secondly, there may be various practical factors playing a role in the lack of correlation between the two approaches. Most of the farms included in the study achieved a good overall assessment (20 "Acceptable" and 22 "Enhanced"), with only one farm scoring "Not classified" (similar distributions have been reported for Belgium, France and Sweden – Anonymous, 2011b), and the welfare status of herds was therefore unlikely to vary widely. This has two consequences. The first is that it Qualitative assessments are anchored by perceived differences between animals, and depend on contrasting expressions to anchor quantification of intermediate welfare values. Previous QBA studies have generally

included contrasting expressive qualities, but this may be more difficult to accomplish in on-farm studies at herd level. It is therefore very possible; even likely, that the spread between the farms in this study was too small to robustly anchor an effective qualitative welfare scale to correspond with WQ outcomes. That the only significant correlations for the two QBA-assessors were those for the principle and criterion with the largest standard deviation (SD) tentatively supports this explanation; however as the standard deviation for several of the principles and criteria was fairly large, this is not an obvious explanation. However the reverse could of course also possibly be true – it may have been that QBA was able to pick up small welfare differences between herds not captured by WQ. Certainly, the effect of farm sample size and range on the efficacy of welfare assessment tools merits further study, and projects facilitating this are currently underway.

The two non-WQ QBA assessors were trained by one of the authors (SNA) and reached very good agreement, supporting results from previous studies (Wemelsfelder et al., 2000, 2001, 2009b,c). There was on the other hand, no significant correlation between these assessors' QBA scores and the QBA scores provided by the WQ-assessor for the Welfare Quality® protocol. The reason for this is unclear but several things differed among the QBA-assessors and the WQ-assessor, including their educational background and level of experience with welfare evaluation schemes. Thus more knowledge about how previous experience affects on-farm QBA at herd level is needed. The Welfare Quality® assessment was done by one person, and we do not know therefore how representative this assessment is. Yet, if the WQ assessor had been biased in recording some of the protocol's indicators, this would not have explained the general absence of strong correlations to QBA; for this, a more systemic bias would have to be presumed. The assessor had previous experience with cattle as well as the required training on the use of the protocol, and should therefore in theory be able to make a representative WQ assessment.

Another practical reason for the disparity between assessments may be that most of the QBA assessments were not done on the same day as the Welfare Quality® assessment (the mean interval was 1.5 days). But when asked, none of the farmers reported that any extraordinary events had happened between the QBA and the Welfare Quality® assessment (e.g. power cuts, renovation works, many animals ill). While several studies have tested the inter-observer reliability of QBA from video tapes with good results (Wemelsfelder et al., 2001; Rousing & Wemelsfelder, 2006; Wemelsfelder et al., 2009c), there are no published studies looking at repeatability over time in a commercial setting using the same observers, and it is therefore not clear what to expect from QBA in this respect. Repeatability studies of measures used in the Welfare Quality® protocol on the other hand have mostly shown good repeatability over both a short and longer term (weeks to months; Forkman & Keeling, 2009), and we should therefore not expect a slight time difference to significantly affect the results.

Finally, to date most published studies applying Qualitative Behavior Assessment were done with animals that appear lively and responsive to a human observer's eye, such as pigs, horses and dogs (although studies on poultry have also been successful: Wemelsfelder et al., 2009a). It could be speculated that dairy cattle, given their apparently less dynamic nature, may be less easy to assess, particularly when observed in large fairly stationary groups. Rousing & Wemelsfelder (2006) assessed the social demeanor of cattle at the drinker, and found both good inter-observer reliability and meaningful correlation with various social behaviors. This study used short video clips showing a small number of cows showing selected social behaviors in a particular situation and this may have focused the observers' attention in ways harder to achieve with a large number of dairy cattle showing a wider range of behaviors. Dairy farmers and stock-people however are bound to disagree that cattle are hard to read, arguing that for anyone with their level of daily hands-on experience, subtle disturbances of expression in either individuals or the entire herd are easily detected (Kielland et al., 2010). In the current study the two QBA-

assessors both had extensive experience with dairy cattle (12 and 35 years respectively) and it is therefore not probable that this is the cause of the relative lack of correlations in the current study.

In conclusion, the current study does not support the idea that Qualitative Behavior Assessment can be used as a stand-alone welfare assessment tool in on-farm conditions, capable of predicting the outcomes of the larger Welfare Quality protocol. This outcome raises questions about the background and training of assessors, the farm sample size and range used in comparative studies, and finally the welfare content of different welfare assessment methods.

Acknowledgements

The authors wish to thank the farmers who participated in the study and the two assessors who performed the Qualitative Behavior Assessment. Also we wish to thank “R. Nørtoft Thomsens legat” and “H.P. Olsen og Hustrus Mindefond” for financially supporting the project.

References

Anonymous, 2009. Kvægbruget i Tal. Dansk Kvæg, Kødbranchens fællesråd, Dansk Landbrugsrådgivning and Landbrug & Fødevarer. http://www.lf.dk/Tal og Analyser/Aarstatistikker/Kvaegbruget i tal/~/media/lf/Tal%20og%20analyser/Aarsstatistikker/Kvaegbruget%20i%20tal/2009/kvaegbruget tal2009%20dk.ashx

Anonymous, 2011a. http://www.welfarequality.net/everyone

Anonymous, 2011b. http://www1.clermont.inra.fr/wq/index.php?id=farms#overall

Appleby, M.C., Sandoe, P., 2002. Philosophical debate on the nature of wellbeing: implications for animal welfare. Anim. Welf. 11, 283–294.

Botreau, R., Veissier, I., Butterworth, A., Bracke, M.B.M., Keeling, L.J., 2007. Definition of criteria for overall assessment of animal welfare. Anim. Welf. 16, 225–228.

Brscic, M., Wemelsfelder, F., Tessitore, E., Gottardo, F., Cozzi, G., Van Reenen, C.G., 2009. Welfare assessment: correlations and integration between a qualitative behavioural assessment and a clinical/health protocol applied in veal calves farms. Ital. J. Anim. Sci. 8, 601–603.

Feaver, J., Mendl, M., Bateson, P., 1986. A method for rating the individual distinctiveness of domestic cats. Anim. Behav. 34, 1016–1025.

Forkman, B., Keeling, L. (Eds.), 2009. Assessment of Animal Welfare Measures for Dairy Cattle, Beef Bulls and Veal Calves. Welfare Quality Reports No. 11. Cardiff University, United Kingdom.

Fraser, D., 2003. Assessing animal welfare at the farm and group level: the interplay of science and values. Anim. Welf. 12, 433–443.

Keeling, L. (Ed.), 2009. An Overview of the Development of the Welfare Quality Project© Assessment Systems. Welfare Quality Reports no. 11. Cardiff University, United Kingdom.

Kielland, C., Skjerve, E., Osterås, O., Zanella, A.J., 2010. Dairy farmer attitudes and empathy toward animals are associated with animal welfare indicators. J. Dairy Sci. 93, 2998–3006.

Minero, M., Tois, M.V., Canali, E., Wemelsfelder, F., 2009. Quantitative and qualitative assessment of the response of foals to the presence of an unfamiliar human. Appl. Anim. Behav. Sci. 116, 74–81.

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

Rousing, T., Wemelsfelder, F., 2006. Qualitative assessment of social behaviour of dairy cows housed in loose housing systems. Appl. Anim. Behav. Sci. 101, 40–53.

Rutherford, K.M.D., Donald, R.D., Lawrence, A.B., Wemelsfelder, F., 2012. Qualitative Behavioural Assessment of emotionality in pigs. Appl. Anim. Behav. Sci. 139 (3–4), 218–224.

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

Veissier, I., Jensen, K.K., Botreau, R., Sandøe, P., 2011. Highlighting ethical decisions underlying the scoring of animal welfare in the Welfare Quality® scheme. Anim. Welf. 20, 89–101.

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

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

Wemelsfelder, 1997. The scientific validity of subjective concepts in models of animal welfare. Appl. Anim. Behav. Sci. 53, 75–88.

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

Wemelsfelder, F., Millard, F., 2009. Qualitative Behaviour Assessment. In: Forkman, B., Keeling, L. (Eds.), Assessment of Animal Welfare Measures for Sows, Piglets and Fattening Pigs. Welfare Quality reports No. 10, Sixth Framework Programme. University of Cardiff, Cardiff, pp. 213–219.

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

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

Wemelsfelder, F., Knierim, U., Schulze Westerath, H., Lentfer, T., Staack, M., Sandilands, V., 2009a. Qualitative Behaviour Assessment. In: Forkman, B., Keeling, L. (Eds.), Assessment of Animal Welfare Measures for Layers and Broilers. Welfare Quality reports No. 9, Sixth Framework Programme. University of Cardiff, Cardiff, pp. 113–119.

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

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