J. Dairy Sci. 100:6376–6388
https://doi.org/10.3168/jds.2016-12255

Trained-user opinion about Welfare Quality measures and integrated scoring of dairy cattle welfare

S. de Graaf,*† B. Ampe,* C. Winckler,‡ M. Radeski,§ L. Mounier,# M. K. Kirchner,‖ M. J. Haskell,¶ F. J. C. M. van Eerdenburg, A. de Boyer des Roches,# S. N. Andreasen,‖ J. Bijttebier,* L. Lauwers,*† W. Verbeke,† and F. A. M. Tuyttens*[1]**

*Institute for Agricultural and Fisheries Research (ILVO), Burg. van Gansberghelaan 92, 9820 Merelbeke, Belgium
†Department of Agricultural Economics, Ghent University, Coupure Links 653, 9000 Ghent, Belgium
‡Division of Livestock Sciences, Department of Sustainable Agricultural Systems, University of Natural Resources and Life Sciences, Gregor-Mendel Straße 33, 1180 Vienna, Austria
§Animal Welfare Center, Faculty of Veterinary Medicine, Ss. Cyril and Methodius University, Lazar Pop-Trajkov 5-7, 1000 Skopje, Republic of Macedonia
#UMR1213 Herbivores, L'Institut national de la recherche agronomique (INRA), VetAgro Sup, Clermont Université, Université de Lyon, F-63122 Saint-Genès-Champanelle, France
‖University of Copenhagen, Department of Veterinary and Animal Sciences, Section of Animal Welfare and Disease Control, Grønnegårdsvej 8, 1870 Frederiksberg, Copenhagen, Denmark
¶SRUC (Scotland's Rural College), West Mains Road, Edinburgh EH9 3JG, Scotland, United Kingdom
**Department of Herd Animal Health, Utrecht University, 3508 TD Utrecht, the Netherlands

ABSTRACT

The Welfare Quality (WQ) protocol for on-farm dairy cattle welfare assessment describes 27 measures and a stepwise method for integrating values for these measures into 11 criteria scores, grouped further into 4 principle scores and finally into an overall welfare categorization with 4 levels. We conducted an online survey to examine whether trained users' opinions of the WQ protocol for dairy cattle correspond with the integrated scores (criteria, principles, and overall categorization) calculated according to the WQ protocol. First, the trained users' scores (n = 8–15) for reliability and validity and their ranking of the importance of all measures for herd welfare were compared with the degree of actual effect of these measures on the WQ integrated scores. Logistic regression was applied to identify the measures that affected the WQ overall welfare categorization into the "not classified" or "enhanced" categories for a database of 491 European herds. The smallest multivariate model maintaining the highest percentage of both sensitivity and specificity for the "enhanced" category contained 6 measures, whereas the model for "not classified" contained 4 measures. Some of the measures that were ranked as least important by trained users (e.g., measures relating to drinkers) had the highest influence on the WQ overall welfare categorization. Conversely, measures rated as most important by the trained users (e.g., lameness and mortality) had a lower effect on the WQ overall category. In addition, trained users were asked to allocate criterion and overall welfare scores to 7 focal herds selected from the database (n = 491 herds). Data on all WQ measures for these focal herds relative to all other herds in the database were provided. The degree to which expert scores corresponded to each other, the systematic difference, and the correspondence between median trained-user opinion and the WQ criterion scores were then tested. The level of correspondence between expert scoring and WQ scoring for 6 of the 12 criteria and for the overall welfare score was low. The WQ scores of the protocol for dairy cattle thus lacked correspondence with trained users on the importance of several welfare measures.
Key words: animal welfare, welfare assessment, trained-user opinion, Welfare Quality

INTRODUCTION

Assessing animal welfare is a highly complex task. Animal welfare is a multidimensional concept that calls for a multicriteria assessment using a multitude of welfare indicators (Mason and Mendl, 1993; Fraser et al., 1997). To express the overall welfare status of a group of farm animals in a single score or index, indicator data should be integrated, which requires interpretation and balancing. No standardized and commonly agreed-on method for assessing the overall welfare status of a group of farm animals exists (i.e., there is no gold standard), which implies that some degree of subjectivity is inevitable when weighting different measures (Spoolder et al., 2003). To be widely accepted, an overall welfare index ought to correspond

Received November 4, 2016.
Accepted April 2, 2017.
[1]Corresponding author: Frank.Tuyttens@ilvo.vlaanderen.be

with society's concept of animal welfare and with the opinion of experts (i.e., people who are seen by society to have adequate knowledge and expertise about animal welfare). However, opinions on the concept of animal welfare may differ between and even within experts and society. For example, producers tend to highlight basic health and functioning of farm animals, whereas nonproducers tend to emphasize farm animals' need for a natural living environment (reviewed by Sørensen and Fraser, 2010). It can be argued that it is too difficult for people without expertise in dairy cattle welfare and the specific welfare measures involved to adequately balance the importance of different welfare measures. It has been shown that providing detailed information about on-farm collection methods of welfare measures significantly influences the relative weights they are given by experts (Rodenburg et al., 2008). Therefore, the current study elicited experienced animal scientists on only the specific welfare measures involved.

To date, the Welfare Quality (**WQ**) protocols are most likely the most renowned and comprehensive method for overall welfare assessment of different farm animal species (chickens, pigs, and cattle; Welfare Quality Consortium, 2009). Unlike some other welfare assessment protocols, WQ relies predominantly on animal-based measures. Resource-based and management-based measures, in contrast, mostly reflect risk factors for welfare impairments instead of directly measuring welfare (Blokhuis et al., 2003, 2010). The WQ protocols are based on 4 main welfare principles (good feeding, good housing, good health, and appropriate behavior), which are split into 12 independent welfare criteria (Table 1). Various welfare measures (n = 27 for dairy cows) were selected by animal scientist to assess these welfare criteria based on validity, reliability, and feasibility of performing the measure on farm. The WQ protocol describes 3 steps for integrating these welfare measures into an overall final welfare category. Methods of integration aim to be widely acceptable by society and therefore are based on expert opinion of social and animal scientists and stakeholders (Botreau et al., 2007), depending on the integration step. For interpretation of measures into criteria scores, animal scientists (n = 6) who were involved in the choice and development of the WQ measures were consulted (Botreau et al., 2008). They were asked to score several situations per criterion that could occur on farm. For example, for integument alterations within the criterion "absence of injuries," experts were asked to score 11 hypothetical farms with varying prevalence of hairless patches, wounds, and swellings. Calculation of criterion scores is based on expert scoring. Social scientists were also involved for aggregation from criteria to principle scores using a similar approach. For the final step, several scenarios for reference profiles were developed to aggregate principle scores into an overall category. First, these scenarios were tested for 69 European dairy farms (Austrian, German, and Italian) to compare their ability to discriminate between farms. Second, stakeholders were consulted to assess which scenario was most appropriate. Third, the degree to which each scenario matched with the general impression of observers for 44/69 dairy farms was assessed. The 4 overall categories (excellent, enhanced, acceptable, or not classified; Welfare Quality Consortium, 2009) were constructed to reflect both the multidimensional nature of welfare and the relative importance of the various welfare measures using mathematical operators that limit the amount of compensation that may occur between welfare measures (i.e., when a combination of positive scores compensates for 1 negative score; Botreau et al., 2009).

Recent critical evaluations of the WQ integration methods indicate that in the dairy cattle protocol a few resource-based measures appear to have a dispro-

Table 1. Principles, the corresponding criteria, and measures used in the Welfare Quality assessment protocol for dairy cows

Principle	Criterion	Measure
Good feeding	Absence of prolonged hunger	BCS (percentage very lean animals)
	Absence of prolonged thirst	Availability and cleanliness of water
Good housing	Comfort around resting	Lying duration, collisions during lying down, on edge or outside of lying area, cleanliness
	Thermal comfort	No measure for dairy cattle
	Ease of movement	Free stalls or presence of tethering and exercise
Good health	Absence of injuries	Lameness, integument alterations
	Absence of disease	Respiration or digestive diseases, mastitis, mortality, dystocia, downer cows
	Absence of pain induced by management procedures	Mutilations (dehorning, tail docking, use of anesthetics or analgesics)
Appropriate behavior	Expression of social behavior	Incidence of agonistic interactions
	Expression of other behaviors	Access to pasture
	Good human–animal relationship	Avoidance distance at feeding place
	Positive emotional state	Qualitative behavioral assessment

portionately large influence on integrated scores (de Vries et al., 2013; Heath et al., 2014). For example, the measures for the criterion "absence of prolonged thirst" (i.e., number, adequate functioning, and cleanliness of drinkers) have a relatively large influence on integrated scores, although they are criticized for their low or undocumented validity (Knierim and Winckler, 2009; de Vries et al., 2013; Tuyttens et al., 2014; de Jong et al., 2016). In contrast, some of the most pressing welfare problems for dairy cattle as highlighted by epidemiological studies (Main et al., 2003; Whay et al., 2003a; de Boyer des Roches et al., 2014) and assessed by experts (i.e., mortality, lameness, and mastitis; Whay et al., 2003b; Lievaart and Noordhuizen, 2011; Nielsen et al., 2014) had a smaller influence on overall welfare categorization (de Vries et al., 2013; Heath et al., 2014; Buijs et al., 2017). These findings point toward potential discrepancies between the dairy cattle welfare assessment of certain welfare experts and the WQ scores.

The WQ protocols were designed with the intention of modifying and updating assessment methods according to advances in animal welfare science. Currently, a large group of researchers has become familiar with the protocol, and these researchers (further referred to as trained users) have performed many farm visits, allowing for a thorough evaluation of the effect that measures have on overall welfare categorization. Therefore, analyzing the correspondence between WQ integrated scores and the opinion of such trained users has become feasible. Hence, the objective of the current study was to analyze the correspondence between welfare assessment by trained users and the WQ scores (criterion and overall welfare category). We did this by examining whether measures that affect WQ categorization most are also those that are deemed most important by trained users.

MATERIALS AND METHODS

WQ Protocol

A brief description of the WQ protocol for on-farm dairy cattle welfare assessment is presented here; the full protocol can be found at http://www.welfarequalitynetwork.net/. In short, the protocol describes 27 on-farm welfare measures (Table 1) that are subsequently integrated in a 3-step process to arrive at an overall welfare category. First, 27 welfare measures of various scales are combined into scores for 12 welfare criteria on a scale of 0 (worst) to 100 (best; Table 1) using various aggregation methods (for details, see Welfare Quality Consortium, 2009). Second, criteria are integrated into scores for 4 welfare principles using Choquet integrals—algorithmic operators that ensure that a poor score cannot be fully compensated by a better score in another criterion (Botreau et al., 2008). Principle scores can range from 0 (worst) to 100 (best). The third and final integration step is an outranking procedure from principle scores, arriving at an overall welfare category. Dairy welfare in a herd is considered excellent when that herd scores >50 for each principle and >75 on 2 of them. When a herd scores >15 for each principle and >50 for at least 2 of them, it is classified as enhanced. Acceptable herds score >5 for all principles and >15 for at least 3 principles. Herds that do not reach the thresholds for the acceptable category are considered not classified. These reference profiles for overall welfare categorization were based on data from 69 herd assessments in the European Union (Botreau et al., 2009).

Collating WQ Data

Data sets of assessments using the WQ protocol for on-farm dairy cattle welfare were collated from 7 European research institutes. Data from 10 countries (Macedonia, The Netherlands, France, Belgium, Scotland, Denmark, Romania, Northern Ireland, Spain, and Austria) and 491 herds were used. The collected samples were selected to be representative of (1) small-scale dairy herds in Macedonia (n = 12); (2) nonorganic and non-tiestall dairy herds in the Netherlands (n = 60) and France (n = 128); (3) random herds with individual SCC data available (to be able to calculate WQ scores) in Belgium (n = 140), Scotland (n = 16), and Denmark (n = 40); (4) typical herds for the regional low-input herding systems in Romania, Northern Ireland, and Spain (n = 30); and (5) loose-housed dairy herds with at least 20 cows in Austria (n = 65). Integrated WQ scores were calculated from raw data using a custom-made integration procedure programmed in R 3.2.2 (R Foundation for Statistical Computing, Vienna, Austria). The R integration program is available on request. The resulting welfare scores were in agreement with the L'Institut National de la Recherche Agronomique Welfare Assessment of Farm Animals webtool (http://www1.clermont.inra.fr/wq/), in which WQ measure scores can be entered (for dairy cows, fattening cattle, growing pigs, and broilers) and WQ criteria, principle, and categorization scores are provided.

Survey

The survey was sent to 31 trained users; it was partially completed by 14 to 15 users (depending on the question) and totally completed by 8 users. The survey

was sent to animal welfare scientists who the coauthors knew to be experienced in the WQ assessment protocol for dairy cow welfare. These trained users were in turn asked to provide contact details of any additional animal welfare scientists who would be suitable (i.e., trained to use the WQ protocol). No trained users who filled out the survey were involved in creating the survey. All trained users had experience with the WQ protocol for dairy cattle (i.e., were trained to perform the WQ protocol for dairy cattle and had performed on-farm WQ assessment of dairy herds), were animal scientists, and had authored at least 1 peer-reviewed scientific paper about dairy cattle welfare involving the WQ protocol. Trained users were all European, and a total of 8 nationalities were represented (British, Spanish, Macedonian, Dutch, Finnish, Austrian, German, and French). Trained users were surveyed on their judgement of the reliability, validity, and importance of all WQ measures. In questions based on data from the WQ European Union database, they were asked to score the farms for each WQ criteria and to assign an overall welfare score.

Reliability, Validity, and Ranking of All WQ Measures for Dairy Cattle. The trained users were asked to indicate how acceptable they judged the reliability and validity of all measures using a tagged visual analog scale from 0 to 100. Tags were "not acceptable (<25)," "just acceptable (25–50)," "acceptable (50–75)," and "very acceptable (75–100)." Reliability was defined in the survey as "a combination of interobserver, intraobserver, and test–retest reliability." Validity was defined as "the measure measures what it is supposed to." Trained users were then asked to rank all WQ measures according to importance for the overall welfare status of a herd of dairy cattle from 1 (most important) to 27 (least important). It was mentioned that reliability, validity, perceived relevance, and prevalence may be considered for ranking.

Expert Scoring Based on All WQ Measurements. The trained users were then asked to score overall welfare based on all measures from the WQ protocol. They were shown one figure with box plots for all measures (part of the figure for one criterion: Figure 1). These showed the same herds as in Figure 1 using the same colored triangles. Trained users were asked to score the overall welfare of 7 focal herds using a 0-to-100 tagged visual analog scale with the tags "not classified (<20)," "acceptable (20–55)," "enhanced (55–80)," and "excellent (>80)." For this purpose, we randomly selected 5 herds from the acceptable welfare category and 2 herds from the enhanced category out of the entire data set. This reflects the distribution of the data set in which 1.8% of the herds (9 herds) were categorized as not classified, 62.7% (308 herds) were categorized as acceptable, 35.4% (174 herds) were categorized as enhanced, and none were categorized as excellent.

Comparing WQ Criteria Scores Using Trained-User Opinion. To assess the degree to which integrated WQ criteria scores correspond to trained-user opinion, the trained users were shown separate graphs of all measures per criterion showing the distribution of all herds in the database (for an example of one criterion, see Figure 2; data shown in Table 2). The focus herds were highlighted using triangles in different colors, and tables stated the data for each. Trained users were asked to score the herds for all 11 criteria (excluding the criterion "thermal comfort," which was not measured on farm for dairy cattle) on a 0-to-100 tagged visual analog scale using the tags "not classified (<20)," "acceptable (20–55)," "enhanced (55–80)," and "excellent (>80)."

Statistical Analysis

The statistical analysis was performed in R 3.2.2 (R Foundation for Statistical Computing). The analyzed data (except overall welfare categorization) were considered to be sufficiently normally distributed based on the graphical evaluation (histogram and quantile–quantile plot) of the residuals.

Reliability, Validity, and Ranking of All WQ Measures for Dairy Cattle

To examine the influence of median reliability and validity scores and their interaction on median ranking of all measures, we used a linear mixed regression model with reliability and validity scores as independent variables and importance rank as a dependent variable. A random effect for expert was included in the model to account for the repeated measures.

Predicting Overall Welfare Categorization Using WQ Measures

To analyze which measures affected the WQ overall categorization into both the lowest (not classified) and highest (enhanced, as no farms were categorized as excellent) categories, welfare categories of the entire European data set (n = 491) were divided into 2 binary variables (1 = enhanced, 0 = other for variable 1; 1 = not classified, 0 = other for variable 2). Logistic regression was used to identify measures that affected overall categorization both univariate and multivariate. For the latter, a model was built using stepwise forward selection, retaining measures with a P-value <0.05 while maintaining the highest coefficient of determina-

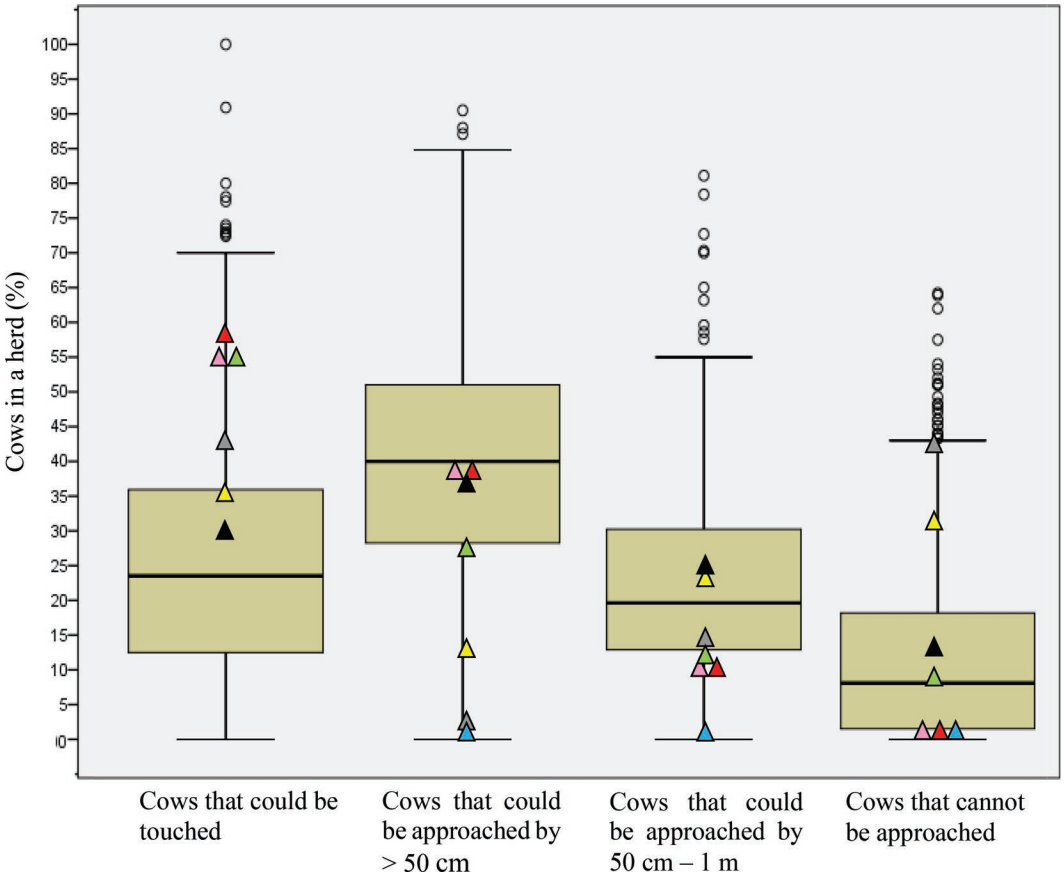

Figure 1. Sample box plot figure from the survey among trained users portraying the distribution of all herds in the database (n = 491) for the measures of the avoidance distance at the feed rack test within the criterion "human–animal relationship." Colored triangles mark the 7 focus herds. Boxes indicate medians and interquartile range; whiskers indicate data within 1.5× the interquartile range.

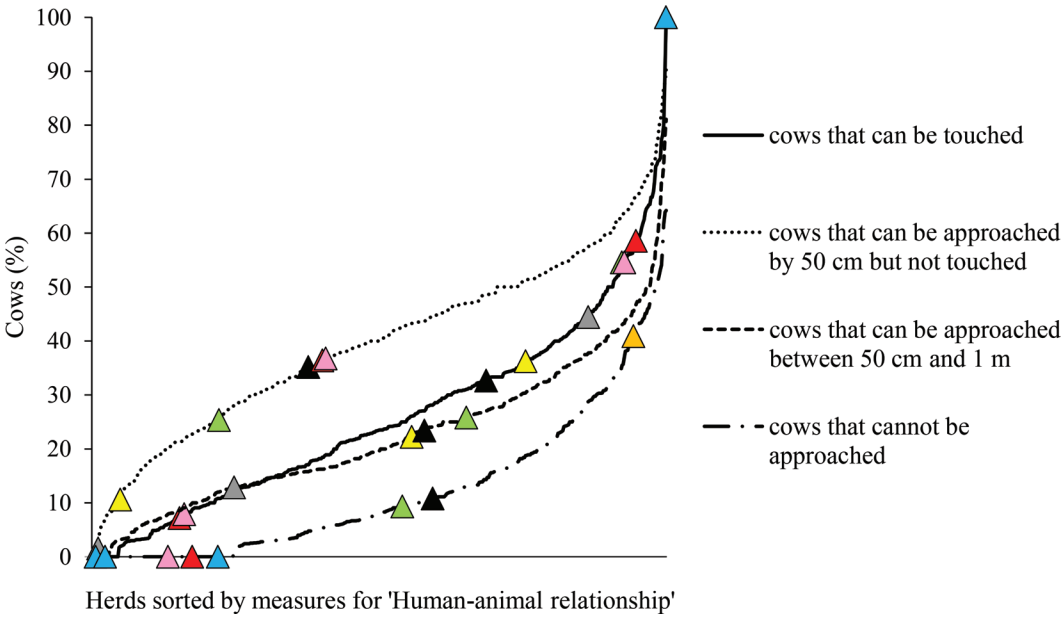

Figure 2. Sample figure from the survey among trained users portraying the distribution of all herds in the database (n = 491) for the measures of the avoidance distance at the feed rack test within the criterion "human–animal relationship." Colored triangles mark the 7 focus herds.

tion. Collinearity was checked for measures used within the models. Model outcome was assessed by calculating specificity and sensitivity using the following formulae:

$$\text{Specificity} = \text{TN}/(\text{TN} + \text{FP})$$

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN}),$$

where TN = true negatives, FP = false positives, TP = true positives, and FN = false negatives. Negatives were those farms categorized as other, and positives were those farms categorized as enhanced for the first binary variable or not classified for the second.

Comparing WQ Criteria Scores with Trained-User Opinion

To assess the systematic difference between the median trained-user opinion score and the WQ criteria scores for each focal herd (n = 7), a paired *t*-test was performed. To model the correspondence of median scores allocated by the trained users and the WQ crite-

Table 2. Measure values of each of the 7 herds presented to trained users in the survey

Criterion and measure	1	2	3	4	5	6	7
Absence of prolonged hunger							
Lean cows (%)	0	3	17	5	11	3	24
Absence of prolonged thirst							
Water bowls/cow (no.)	0.5	0.0	0.0	0.0	0.0	0.6	0.05
Trough length/cow (cm)	0.0	7.9	4.7	28.6	9.0	0.0	0.0
Drinker cleanliness	Yes	Yes	No	Yes	Yes	Yes	Yes
At least 2 drinkers/cow	No	Yes	No	No	Yes	No	Yes
Resting comfort							
Mean time needed to lie down (s)	4.6	4.6	7.5	4.1	6.6	5.4	6.8
Cows colliding with housing equipment (%)	16	15	72	0	37	8	33
Cows lying outside of lying area (%)	50	11	0	0	0	35	0
Cows with dirty flanks (%)	34	55	81	14	67	79	70
Cows with dirty lower legs (%)	57	37	85	38	20	79	100
Cows with a dirty udder (%)	18	21	77	10	42	48	95
Ease of movement							
Housing	Tied	Loose	Loose	Loose	Loose	Tied	Loose
Absence of injuries							
Moderately lame cows (%)	0	13	88	0	23	0	84
Severely lame cows (%)	32	0	12	10	17	27	5
Cows with at least 1 lesion (%)	7	12	72	28	13	20	68
Cows with no lesions but at least 1 hairless patch (%)	98	18	28	38	21	100	32
Absence of disease							
Number of coughs/cow per minute	0.05	0.00	0.13	0.10	0.06	0.17	0.00
Cows with nasal discharge (%)	59	0	0	0	5	18	0
Cows with ocular discharge (%)	0	0	0	0	0	0	0
Cows with hampered respiration (%)	0	0	0	5	0	0	0
Cows with diarrhea (%)	5	0	0	0	0	0	16
Cows with vulvar discharge (%)	0	0	0	0	0	0	3
Cows with SCC >400,000 (%)	8	21	25	0	14	8	12
Cow mortality (%)	5	3	4	0	4	3	4
Calvings with dystocia (%)	0	21	0	0	1	6	3
Downer cows (%)	0	6	0	0	0	6	5
Absence of pain induced by management procedures							
Dehorning method[1]	T	P	P	N	P	P	T
Use of analgesics	No	Yes	Yes	No	Yes	No	No
Use of anesthetics	No	Yes	Yes	No	Yes	No	Yes
Expression of social behavior							
Number of head butts/cow per 15 min	0.8	4.0	0.7	0.0	0.4	0.4	1.0
Number of displacements/cow per 15 min	0	1.2	0.1	0.0	0.2	0.0	0.8
Expression of other normal behavior							
Hours on pasture (no.)	214	180	0	0	0	214	195
Days on pasture (no.)	19	9	0	0	0	8	9
Human–animal relationship							
Could be touched (%)	36	55	59	100	55	44	30
Closer than 50 cm but not touched (%)	11	36	37	0	26	2	35
Between 50 cm and 1 m (%)	23	9	9	0	11	14	24
>1 m (%)	30	0	0	0	9	41	11
Positive emotional state							
Qualitative behavior assessment score	43	40	8	91	77	66	54

[1]T = thermal; P = caustic paste; N = none.

ria scores, a linear model was fitted and the coefficient of determination was calculated. Additionally, the intraclass correlation coefficient (**ICC**) was calculated to assess the degree of coherence between individual trained-user opinions.

RESULTS

Perceived Reliability, Validity, and Ranking of WQ Measures

Median validity and reliability scores for all measures were acceptable to very acceptable (i.e., median scores were >50; Table 3). Nevertheless, there was variation in median scores for the various measures, ranging from 60 to 100 for reliability and from 50 to 90 for validity. The highest median ranking was attached to lameness score (rank 2), BCS (4), mortality rate (7), and integument alterations (7). Lameness score and integument alternations received the highest median validity scores (89 and 90, respectively), along with "lying outside the lying area" (89) and "tail docking method" (88). "Tied versus loose housing" (100), measures of drinker space ["centimeters of trough per cow (minimum 6 cm), number of water bowls per cow (minimum 0.10), and at least 2 drinkers available for each cow" (93)], and "water flow" (90) received the highest median reliability scores. The measure "qualitative behavior assessment" (**QBA**) was given the worst median importance rank (22) and the lowest median reliability score (60) and was among the lowest median validity scores (57). Measures of drinker space were given the lowest median validity score (50). Water flow was among the lowest ranking measures in terms of importance (20) and among the lowest median validity scores (60). The highest variation in reliability scores between trained users (SD) was found for QBA (32), and the lowest variation was found for BCS (10). For validity scores, the highest variation between trained users was found for validity scores of water flow (28), and the lowest variation was found for integument alterations (8). For ranking, scores for the measures "tail docking method," "head butts and displacements," and "avoidance distance test" (9) were most variable, and scores for mortality and integument alterations were least variable (4).

The importance rank of the measure was negatively associated with both the reliability and validity scores, although validity had a somewhat higher estimate (i.e., higher importance as indicated by a lower ranking was associated with higher reliability and validity scores; $P = 0.03$ for both; estimates $= -0.66$ and -0.74, respectively; adjusted $R^2 = 0.20$). A very small but significant interaction was found between reliability and validity scores where they did not strengthen each

Table 3. Median (interquartile range) reliability and validity scores and rankings for each Welfare Quality measure by trained users

Measure	Reliability (n = 15)	Validity (n = 15)	Ranking (n = 13)
BCS	89 (11)	79 (35)	4 (8)
Centimeters of trough/cow (minimum 6 cm), no. of water bowls/cow (minimum 0.10) and at least 2 drinkers/cow	93 (15)	50 (34)	13 (6)
Water cleanliness (judged visually)	80 (28)	70 (36)	19 (9)
Water flow	90 (33)	60 (40)	20 (15)
Time needed to lie down	75 (38)	78 (21)	9 (7)
Cows colliding with housing	70 (39)	82 (28)	16 (10)
Cows lying outside of lying area	85 (33)	89 (28)	16 (10)
Cleanliness of udders, flanks, and lower legs	75 (12)	81 (24)	15 (5)
Tied versus loose housing	100 (6)	84 (28)	11 (13)
Lameness score	69 (36)	89 (11)	2 (2)
Integument alterations	75 (15)	90 (14)	7 (4)
Coughing	69 (44)	75 (35)	19 (13)
Nasal discharge	84 (35)	80 (11)	18 (8)
Ocular discharge	85 (31)	80 (12)	18 (11)
Hampered respiration	88 (36)	86 (12)	21 (12)
Diarrhea	75 (21)	70 (22)	15 (8)
Vulvar discharge	77 (39)	86 (14)	18 (8)
SCC >400,000	83 (19)	81 (11)	13 (14)
Mortality	79 (47)	81 (16)	7 (6)
Dystocia	79 (37)	80 (17)	13 (10)
Downer cows	79 (47)	81 (16)	15 (14)
Dehorning method	90 (26)	86 (16)	11 (10)
Tail docking method	95 (16)	88 (17)	17 (18)
Head butts and displacements	70 (26)	75 (17)	14 (16)
Access to pasture (no. of hours and no. of days on pasture)	90 (18)	75 (33)	19 (8)
Avoidance distance test	66 (24)	76 (28)	17 (15)
Qualitative behavior assessment	60 (37)	57 (20)	22 (11)

Table 4. *P*-values of the univariate logistic regression models examining predictability of single measures for a herd to be categorized as "enhanced" or "not classified" based on the collated European data set (n = 491)

Criterion and measure	Enhanced	Not classified
Absence of prolonged hunger		
Lean cows (%)	<0.001	<0.001
Absence of prolonged thirst		
Water bowls (no.)	0.070	0.863
Water flow	<0.001	0.505
Trough length/cow (cm)	0.001	0.008
At least 2 drinkers/cow	<0.001	0.006
Drinker cleanliness	<0.001	0.068
Resting comfort		
Mean time needed to lie down	<0.001	0.577
Cows colliding with housing (%)	<0.001	0.365
Cows lying outside of lying area (%)	<0.001	0.014
Cows with dirty flanks (%)	0.101	0.172
Cows with dirty lower legs (%)	0.023	0.110
Cows with a dirty udder (%)	0.374	0.258
Ease of movement		
Loose or tied housing	<0.001	0.016
Absence of injuries		
Moderately lame cows (%)	0.002	0.392
Severely lame cows (%)	<0.001	0.096
Cows with at least 1 lesion or swelling (%)	<0.001	0.014
Cows with at least 1 hairless patch (%)	0.141	0.075
Absence of disease		
No. of coughs/cow per min	0.168	0.350
Cows with nasal discharge (%)	0.092	0.165
Cows with ocular discharge (%)	0.044	0.426
Cows with hampered respiration (%)	0.293	0.385
Cows with diarrhea (%)	0.386	0.546
Cows with vulvar discharge (%)	0.588	0.936
Cows with SCC >400,000 (%)	0.130	0.014
Cow mortality (%)	<0.001	0.189
Calvings with dystocia (%)	0.619	0.841
Downer cows (%)	0.742	0.423
Absence of pain induced by management procedures		
Method of dehorning	0.130	0.021
Use of analgesics during and after dehorning	0.618	0.540
Use of anesthesia during dehorning	0.759	0.110
Method of tail docking	0.150	0.974
Use of analgesics during and after tail docking	0.011	0.008
Use of anesthesia during tail docking	0.025	0.010
Expression of social behavior		
Head butts/cow per 15 min	0.033	0.759
Displacements/cow per 15 min	0.615	0.159
Expression of other normal behavior		
Hours on pasture (no.)	0.467	0.153
Days on pasture (no.)	0.810	0.454
Human–animal relationship		
Could be touched (%)	0.711	0.188
Could be approached <50 cm but not touched (%)	0.012	0.379
Could be approached 50 cm to 1 m (%)	0.253	0.924
>1 m (%)	0.011	0.547
Positive emotional state		
Qualitative behavior assessment index score	0.079	<0.001

other's negative effect on ranking ($P = 0.048$, estimate $= -0.009$).

Predicting Overall Welfare Categorization Using WQ Measures

When analyzed univariately, 20 out of 41 measures significantly ($P < 0.05$) affected overall welfare categorization into the enhanced category (Table 4), and 11 measures significantly affected categorization into the not classified category for the entire European data set (n = 491).

The multivariable model that had the fewest variables while maintaining the highest percentage of both sensitivity and specificity (67 and 85%, respectively) for the enhanced category contained the following

Table 5. *P*-values and model estimates of measures in the multivariate logistic regression models predicting a herd to be categorized as "enhanced" or "not classified" based on the collated European data set (n = 491)

	Enhanced model		Not classified model	
Outcome variable	Estimate	*P*-value	Estimate	*P*-value
No. of lean cows	—	—	1.8	<0.001
Water flow	1.1	<0.001	—	—
At least 2 drinkers/cow	2.4	<0.001	−3.7	0.007
Drinker cleanliness	0.6	<0.001	—	—
Mean time needed to lie down	−0.7	<0.001	—	—
Cows lying outside of lying area (%)	−0.9	<0.001	—	—
Cows with at least 1 lesion or swelling (%)	−0.5	<0.001	—	—
No. of displacements/cow per hour	—	—	0.7	0.043
Qualitative behavior assessment index score	—	—	−1.6	0.002

measures (from most to least influence): at least 2 drinkers/cow, water flow, percentage of animals lying outside the lying area, mean time needed to lie down, drinker cleanliness, and percentage of animals with at least 1 lesion/swelling (Table 5). For the not classified category, the measures (from most to least influence) at least 2 drinkers/cow, number of lean cows, QBA index, and number of displacements/cow per hour contributed to the model with fewest variables but the highest sensitivity (44%) and specificity (100%).

Comparing WQ Overall Welfare Category and Criteria Scores with Trained-User Opinion

For 2 of 5 acceptable herds and for 1 of 2 enhanced herds, the majority of trained users (n = 8) scored in accordance with WQ (Figure 3). Regarding scores that were not in accordance with WQ, the vast majority were a lower category than the WQ calculation (25/29 expert scores). Overall, ICC for overall welfare scores by trained users was 0.5.

The following criteria were systematically scored lower by trained users than the WQ score: absence of injuries, absence of pain induced by management procedures, expression of social behavior, and good human–animal relationship (Table 6). The expert and WQ scores were not significantly related for 2 criteria: absence of prolonged thirst and absence of prolonged hunger (Table 6). The correspondence between trained users was insufficient (ICC < 0.6) for 2 criteria: absence of injuries and absence of disease. The number of measures within a criterion tended to be negatively related to ICC ($P = 0.06$, estimate = −0.04).

DISCUSSION

This study gives insight into the relationship between integrated scores of the WQ dairy cattle protocol and trained-user opinion. The specific research design imposes some limitations but also provides challenges for future research. For example, we chose to select only dairy cattle welfare experts who were trained users

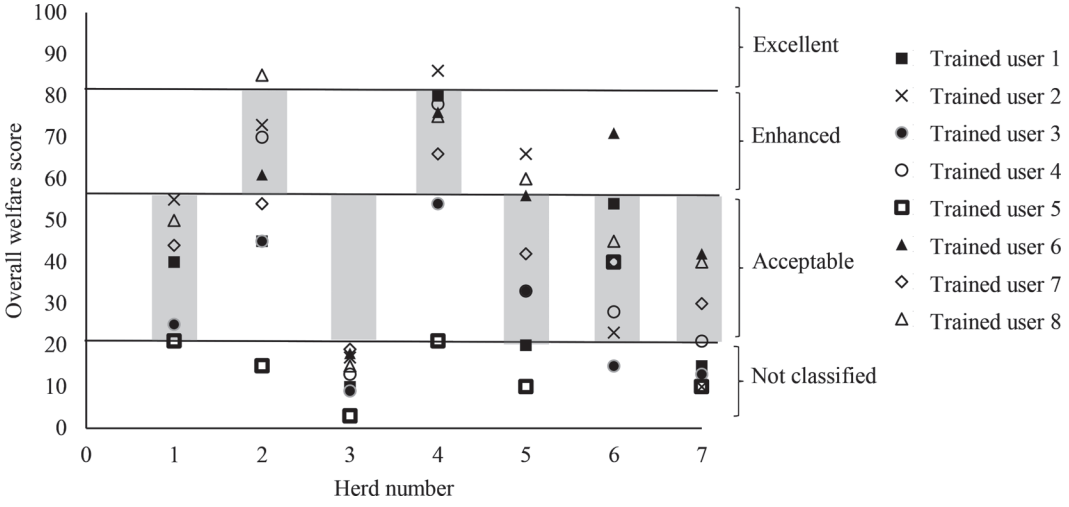

Figure 3. Overall welfare score for all 7 focus herds by 8 trained users. Gray boxes indicate Welfare Quality overall welfare category.

of the WQ dairy cattle protocol. This ensured that trained users had a proper knowledge of the protocol and all measures but limited the number of possible respondents. The results show discrepancies between trained-user opinion and WQ scores.

Trained-User Opinion on Ranking, Reliability, and Validity of Measures

The measures that the trained users ranked highest in terms of perceived importance for the overall welfare status of a herd (namely lameness score, BCS, mortality rate, and integument alterations) are in agreement with earlier studies in which dairy cattle welfare trained users were asked to score the importance of welfare measures (Whay et al., 2003b; Lievaart and Noordhuizen, 2011; Nielsen et al., 2014). Both reliability and validity scores influenced ranking positively (based on the negative relationship between reliability and validity scores and ranking) but did not positively interact. This means that highest ranked measures in the current study did not necessarily receive both the highest validity and the highest reliability scores. In addition, although the set-up of this study was such that trained users had to consider validity and reliability before ranking, other (unknown) factors appeared to influence the trained users' opinion on the importance of the various measures for overall herd welfare as well (further supported by the models' low R^2 of 0.20). This was the case for lameness, for example, which was ranked highest for importance, although its reliability was among the lowest.

Overall, QBA was scored among the lowest by the trained users with regard to reliability and validity (although it was still within the "acceptable" range) and was ranked lowest on importance for dairy cattle welfare status. The QBA is a method that uses descriptors such as "frustrated" or "content" to interpret the behavior and body language of an animal and integrates these details of animal behavior into a qualitative judgment of overall welfare state (Wemelsfelder and Lawrence, 2001; Rousing and Wemelsfelder, 2006; Wemelsfelder, 2007). Interobserver reliability was tested and deemed acceptable for a QBA method using "free" descriptors (i.e., not set but rather determined by observers themselves) and was validated by correlating results to behavioral observations (Rousing and Wemelsfelder, 2006; Napolitano et al., 2012). The fixed-term method and specific set of descriptors used in the WQ protocol were tested for interobserver reliability in a study by Bokkers et al. (2012) and judged as not satisfactory by the authors involved (i.e., Kendall's coefficient of concordance was <0.7), whereas Wemelsfelder et al. (2009) reported satisfactory observer agreement of those descriptors in beef, dairy cattle, and veal calves. In addition, recently published papers demonstrated internal validity by testing the correlation between QBA and other behavioral and physiological measures (Coignard et al., 2014; Phythian et al., 2016; Serrapica et al., 2017).

Although some measures scored highest for reliability, they scored lowest for validity [e.g., measures related to the criterion "absence of prolonged thirst" ("centimeters of trough per cow")] or were ranked lowest on importance for dairy cattle welfare ("water flow"). Criticism expressed in earlier studies for these measures is related to their resource-based nature and the effect these specific measures have on the WQ integrated scores, whereas preference generally is given to animal-based measures (de Vries et al., 2013; Heath et al., 2014; Buijs et al., 2017). Measuring functioning of water points, water provision, and water cleanliness refers to assessing a risk for cows being in a certain welfare state and may therefore in some cases not be the

Table 6. Systematic *t*-test *P*-value, linear regression coefficient of determination, and intraclass correlation coefficient (ICC) of Welfare Quality (WQ) integrated scores and trained-user median scores (n = 14) for the focus herds (n = 7) for each WQ criterion

Criterion	Median (IR[1]) WQ score	Median (IR) expert score	Systematic *t*-test *P*-value	Regression R^2	ICC
Absence of prolonged hunger	67 (39)	50 (75)	0.475	0.237	0.6
Absence of prolonged thirst	20 (97)	50 (71)	0.737	0.007	0.7
Comfort around resting	27 (20)	25 (33)	0.181	0.880**	0.8
Freedom of movement	100 (33)	90 (90)	0.125	1.000***	1.0
Absence of injuries	28 (19)	18 (29)	0.006	0.926***	0.5
Absence of disease	40 (32)	42 (34)	0.296	0.903**	0.4
Absence of pain induced by management procedures	58 (18)	10 (50)	0.023	0.521*	0.8
Expression of social behavior	84 (24)	58 (50)	0.020	0.869**	0.6
Expression of other normal behavior	73 (78)	60 (78)	0.828	0.978***	0.9
Good human–animal relationship	54 (37)	52 (50)	0.023	0.984***	0.7
Positive emotional state	54 (32)	50 (37)	0.901	0.997***	0.8

[1]IR = interquartile range.

*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$.

most valid measure of an actual welfare state in dairy cattle, in this case due to prolonged thirst. Additionally, to our knowledge, no actual validity testing of the WQ drinker measures has occurred. This could explain the relatively low perceived validity score attached to these measures by the trained users. Further testing of reliability and validity on certain measures is needed based on the results of the current study and previous research (Knierim and Winckler, 2009). If from such studies it appears that measures are not sufficiently reliable or valid, then research should be performed to propose improved measures.

The trained users did not always agree on the relative importance of the overall welfare status of dairy herds of different welfare measures given the high variations in ranking and in reliability and validity scores between trained users. This possibly reflects diverging views in what trained users find most important for dairy cattle welfare, as Fraser et al. (1997) showed in his study on animal welfare conceptualization among animal welfare scientists. This indicates that when using trained-user opinion to determine weights for various measures, such variation should be accounted for when selecting the expert panel. Therefore, it is not likely that an overall welfare score will always perfectly reflect an individual trained user's opinion. Methods for achieving more consensus among trained users exist. Examples are deliberative processes using a workshop such as that performed by Rodenburg et al. (2008) or more complex processes, such as a Delphi method with multiple rounds of expert elicitation and feedback (Linstone and Turoff, 1975).

Comparison of the Measures' Effect on Overall Welfare Categorization and Trained-User Opinion

Compared with previous studies (Heath et al., 2014; Buijs et al., 2017), more measures affected both the "enhanced" and the "not classified" categorization in the current study. This is likely attributable to a larger variation in data in the current study, which used a much larger (and diverse, as data were collected in more than 1 country) database compared with both other studies. To specify, the current sample comprised 491 herds as opposed to 92 herds and 22 flocks for Heath et al. (2014) and Buijs et al. (2017), respectively. In accordance with Heath et al. (2014), drinker measures had the biggest influence for both the enhanced and not classified models, whereas in the current study these received some of the lowest ranks or validity scores by the trained users. Additionally, the QBA score, which scored lowest overall, was among the best predictors for the "not classified" categorization. By contrast, although little agreement on the importance of various welfare measures often exists among trained users, some measures that are regarded as highly important to cattle welfare by certain welfare trained users did not have a great influence on the overall welfare status categorization. For example, although lameness score and mortality rate contributed to the "enhanced" categorization in univariate models, they did not when combined into a multivariable model. These results show that the relative influence of measures on WQ integrated scores may not be in accordance with the trained users' opinion of this study. We tested this by comparing expert scoring of WQ criteria and overall welfare with calculated WQ scores.

Comparing WQ Integrated Scores with Trained-User Opinion

Overall Welfare Category. For only 3 out of the 7 herds, the majority of trained users scored in accordance with the WQ overall welfare categorization. The 2 herds that were scored as "not classified" by at least half of the trained users (herds 3 and 7) both scored badly (i.e., relatively high prevalence) on measures that were ranked as highly important by the trained users (namely lesions and swellings and moderately lame cows).

Variation between trained users was shown for the overall welfare scoring given the relatively low ICCs. This was also shown for criteria scores, where ICCs tended to be lower for criteria that contained the most measures. This can indicate that (1) trained users did not agree on their assessment of overall welfare caused by a different view of animal welfare (as mentioned previously) or (2) some trained users may have had difficulties in aggregating many welfare measures into a single overall score. The latter explanation is supported by the fact that 6 of the 14 trained users who completed the questions on criterion scores did not complete the question on overall welfare scores.

Criteria Scores. The following criteria were systematically scored lower by trained users than the WQ integrated scores: absence of injuries, absence of pain induced by management procedures, expression of social behavior, and good human–animal relationship. In the WQ protocol, poor scores have more influence on integrated scores than do good scores (Buijs et al., 2017). Therefore, lower scores on each of these criteria would have a major effect on principle scores and overall welfare category.

The correspondence between the expert and WQ score for the criterion "absence of prolonged thirst" was extremely low. The finding that the trained users considered some of these measures to be of relatively poor validity may partly explain this lack of correspondence.

It is a strong indication that trained users of the present study did not agree with the way that the criterion score for absence of prolonged thirst is calculated in the WQ protocol.

Four complementary explanations can be put forward for the poor correspondence between trained users' scores and WQ integrated scores. First, except for the first step of the integration procedure, WQ consulted a much wider group of stakeholders (including animal scientists, social scientists, producers, and retailers) than we did in the current study. These stakeholders' views on the relative effect of the various measures on dairy cattle welfare may differ substantially from those of the trained users in the current study. We opted to limit the current study to trained users only because it could be argued that they are best qualified to assess overall dairy cattle welfare state and the relative importance of the various WQ measures.

Second, because the protocol was not yet published when stakeholder opinion was elicited during the WQ project, stakeholders could not have gained as much experience in performing the various WQ measures as the trained users in this study. It has previously been shown that detailed information on welfare measures (e.g., practical implications) can significantly influence relative weight attributed to these welfare measures by trained users (Rodenburg et al., 2008).

Third, there was considerable variation between trained users in the present study regarding importance ranking, although no information on the degree of variation between the original WQ trained users is readily available. The variation in prioritizing certain aspects of welfare in the current group of trained users could arise from different concepts of animal welfare, such as what Fraser (2008) described as "basic health and functioning," "natural living," and "affective states."

Fourth, WQ integration methods likely contribute to differences between trained-user opinion and WQ integrated scores. de Graaf et al. (2016) identified 2 factors that influence the effect a measure has on the integrated WQ scores but that seem unintended by the Welfare Quality Consortium: (1) the number of integrated measures per criterion or principle and (2) the various aggregation methods of measures into criteria scores that influence the effect individual measures have on integrated scores. In the present study a low level of correspondence was found between welfare measures that affect WQ categorization most and those that were scored as most important by trained users. Also, poor correspondence between trained-user opinion and some criterion scores indicated that this lack of correspondence starts in the first step of integration.

These findings indicate a lack of correspondence between WQ welfare scores and trained users' assessment of herd welfare. The opinion of these trained users is the only silver standard we have for validating animal welfare integrated scores because these users are arguably best equipped to assess and quantify the welfare of a given herd. Moreover, these trained users may be considered authorities for animal welfare assessment in society, and it is important that scientists who use this method support it. Future research could focus on determining whether the way trained users assess welfare corresponds with the assessment of other stakeholders. Improvements for WQ may be derived from the observed discrepancies between WQ overall welfare assessment and the assessment of the trained users. In some cases, the trained users scored lower than WQ, and in other cases (e.g., water provision) they were less stringent. Because WQ allocates more weight to low scores, this is likely to have a significant effect on the overall assessment. For example, higher criterion scores for absence of thirst (following our trained users' opinion) would reduce the effect of this criterion on the overall assessment. On the contrary, lameness should be given more effect because our trained users ranked this as highly important.

CONCLUSIONS

Trained-user opinion on the most and least important measures for the overall welfare status of a herd did not correspond well with the influence of these measures on the WQ overall welfare categorization. Some of the measures that were ranked as least important for herd welfare by trained users (e.g., measures relating to drinkers) had the highest influence on the WQ overall welfare categorization. On the contrary, measures ranked as most important by the trained users (e.g., lameness and mortality) had a lower effect on the WQ overall category. In addition, results indicate poor correspondence between trained users' scoring and 6 of 11 WQ criteria and the overall welfare category. In both cases, trained users mostly allocated more negative scores, indicating a lower level of welfare. The WQ scores of the protocol for dairy cattle thus lacked correspondence with those of selected trained users on the importance of several welfare measures.

ACKNOWLEDGMENTS

We thank all trained users who filled out the survey and Miriam Levenson (ILVO, Belgium) for editing the language in the article.

REFERENCES

Blokhuis, H. J., R. B. Jones, R. Geers, M. Miele, and I. Veissier. 2003. Measuring and monitoring animal welfare: Transparency in the food product quality chain. Anim. Welf. 12:445–455.

Blokhuis, H. J., I. Veissier, M. Miele, and B. Jones. 2010. The Welfare Quality® project and beyond: Safeguarding herd animal well-being. Acta Agric. Scand. Anim. Sci. 60:129–140.

Bokkers, E. A. M., M. de Vries, I. C. M. A. Antonissen, and I. J. M. de Boer. 2012. Inter-and intra-observer reliability of experienced and inexperienced observers for the Qualitative Behavior Assessment in dairy cattle. Anim. Welf. 21:307–318.

Botreau, R., J. Capdeville, P. Perny, and I. Veissier. 2008. Multicriteria evaluation of animal welfare at farm level: An application of MCDA methodologies. Found. Comp. Decis. Sci. 33:287–317.

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

Botreau, R., I. Veissier, and P. Perny. 2009. Overall assessment of animal welfare: Strategy adopted in Welfare Quality®. Anim. Welf. 18:363–370.

Buijs, S., B. Ampe, and F. A. M. Tuyttens. 2017. Sensitivity of the Welfare Quality® broiler chicken protocol to differences between intensively reared indoor flocks: Which factors explain overall categorization? Animal 11:244–253.

Coignard, M., R. Guatteo, I. Veissier, A. Lehebel, C. Hoogveld, L. Mounier, and N. Bareille. 2014. Does milk yield reflect the level of welfare in dairy herds? Vet. J. 199:184–187.

de Boyer des Roches, A., I. Veissie, M. Coignard, N. Bareille, R. Guatteo, J. Capdeville, E. Gilot-Fromont, and L. Mounier. 2014. The major welfare problems of dairy cows in French commercial farms: An epidemiological approach. Anim. Welf. 23:467–478.

de Graaf, S., B. Ampe, S. Buijs, S. N. Andreasen, A. De Boyer Des Roches, F. J. C. M. van Eerdenburg, M. J. Haskell, M. K. Kircher, L. Mounier, M. Radeski, C. Winckler, J. Bijttebier, L. Lauwers, W. Verbeke, and F. A. M. Tuyttens. 2016. Sensitivity of the integrated Welfare Quality® scores of the dairy cattle protocol to changes in individual measures. Page 12 in Proc. Benelux ISAE Conf 2016, Berlicum, the Netherlands.

de Jong, I. C., V. A. Hindle, A. Butterworth, B. Engel, P. Ferrari, H. Gunnink, T. P. Moya, F. A. M. Tuyttens, and C. G. Van Reenen. 2016. Simplifying the Welfare Quality® assessment protocol for broiler chicken welfare. Animal 10:117–127.

de Vries, M., E. A. M. Bokkers, G. van Schaik, R. I. Botreau, B. Engel, T. Dijkstra, and I. J. M. de Boer. 2013. Evaluating results of the Welfare Quality multi-criteria evaluation model for categorization of dairy cattle welfare at the herd level. J. Dairy Sci. 96:6264–6273.

Fraser, D. 2008. Understanding animal welfare. Acta Vet. Scand. 50(Suppl. 1):S1. 10.1186/1751-0147-50-S1-S1.

Fraser, D., D. M. Weary, E. A. Pajor, and B. N. Milligan. 1997. A scientific conception of animal welfare that reflects ethical concerns. Anim. Welf. 6:187–205.

Heath, C. A. E., W. J. Browne, S. Mullan, and D. C. J. Main. 2014. Navigating the iceberg: Reducing the number of parameters within the Welfare Quality® assessment protocol for dairy cows. Animal 8:1978–1986.

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

Lievaart, J. J., and J. P. T. M. Noordhuizen. 2011. Ranking experts' preferences regarding measures and methods of assessment of welfare in dairy herds using Adaptive Conjoint Analysis. J. Dairy Sci. 94:3420–3427.

Linstone, H. A., and M. Turoff. 1975. The Delphi method: Techniques and applications. Addison-Wesley, London, UK.

Main, D. C. J., H. R. Whay, L. E. Green, and A. J. F. Webster. 2003. Effect of the RSPCA Freedom food scheme on the welfare of dairy cattle. Vet. Rec. 153:227–231.

Mason, G., and M. Mendl. 1993. Why is there no simple way of measuring animal welfare? Anim. Welf. 2:301–319.

Napolitano, F., G. De Rosa, F. Grasso, and F. Wemelsfelder. 2012. Qualitative behavior assessment of dairy buffaloes (*Bubalus bubalis*). Appl. Anim. Behav. Sci. 141:91–100.

Nielsen, B. H., A. Angelucci, A. Scalvenzi, B. Forkman, F. Fusi, F. A. M. Tuyttens, H. Houe, H. Blokhuis, J. T. Sørensen, J. Rothmann, L. Matthews, L. Mounier, L. Bertocchi, M. Richard, M. Donati, P. P. Nielsen, R. Salini, S. de Graaf, S. Hild, S. Messori, S. S. Nielsen, V. Lorenzi, X. Boivin, and P. T. Thomsen. 2014. Use of animal based measures for the assessment of dairy cow welfare—ANIBAM. European Food Safety Authority, Parma, Italy.

Phythian, C. J., E. Michalopoulou, P. J. Cripps, J. S. Duncan, and F. Wemelsfelder. 2016. On-farm qualitative behaviour assessment in sheep: Repeated measurements across time, and association with physical indicators of flock health and welfare. Appl. Anim. Behav. Sci. 175:23–31.

Rodenburg, T. B., F. A. M. Tuyttens, K. De Reu, L. Herman, J. Zoons, and B. Sonck. 2008. Welfare assessment of laying hens in furnished cages and non-cage systems: Assimilating trained user opinion. Anim. Welf. 17:355–361.

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

Serrapica, M., X. Boivin, M. Coulon, A. Braghieri, and F. Napolitano. 2017. Positive perception of human stroking by lambs: Qualitative behaviour assessment confirms previous interpretation of quantitative data. Appl. Anim. Behav. Sci. 187:31–37.

Sørensen, J. T., and D. Fraser. 2010. On-farm welfare assessment for regulatory purposes: Issues and possible solutions. Livest. Sci. 131:1–7.

Spoolder, H., G. De Rosa, B. Horning, S. Waiblinger, and F. Wemelsfelder. 2003. Integrating parameters to assess on-farm welfare. Anim. Welf. 12:529–534.

Tuyttens, F. A. M., S. de Graaf, J. L. Heerkens, L. Jacobs, E. Nalon, S. Ott, L. Stadig, E. Van Laer, and B. Ampe. 2014. Observer bias in animal behavior research: Can we believe what we score, if we score what we believe? Anim. Behav. 90:273–280.

Welfare Quality Consortium. 2009. Welfare Quality assessment protocol for cattle. Accessed Apr. 11, 2016. http://www.welfarequalitynetwork.net/.

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

Wemelsfelder, F., and A. B. Lawrence. 2001. Qualitative assessment of animal behaviour as an on-farm welfare-monitoring tool. Acta Agric. Scand. Anim. Sci. 51:21–25.

Wemelsfelder, F., F. Millard, G. De Rosa, and F. Napolitano. 2009. Qualitative behaviour assessment. Pages 215–224 in Welfare Quality Reports No. 11. B. Forkman and L. J. Keeling, ed. Welfare Quality Consortium, Lelystad, the Netherlands.

Whay, H. R., D. C. J. Main, L. E. Green, and A. J. F. Webster. 2003a. Assessment of the welfare of dairy cattle using animal-based measurements: Direct observations and investigation of farm records. Vet. Rec. 153:197–202.

Whay, H. R., D. C. J. Main, L. E. Greent, and A. J. F. Webster. 2003b. Animal-based measures for the assessment of welfare state of dairy cattle, pigs and laying hens: Consensus of expert opinion. Anim. Welf. 12:205–217.