Evaluating results of the welfare quality multi-criteria evaluation model for classification of dairy cattle welfare at the herd level

Marion de Vries, E.A.M Bokkers, G. van Schaik, Raphaëlle Botreau, Bas Engel, T. Dijkstra, I.J. de Boer

J. Dairy Sci. 96:6264–6273
http://dx.doi.org/10.3168/jds.2012-6129

Evaluating results of the Welfare Quality multi-criteria evaluation model for classification of dairy cattle welfare at the herd level

M. de Vries,*[1] E. A. M. Bokkers,* G. van Schaik,† R. Botreau,‡ B. Engel,§ T. Dijkstra,† and I. J. M. de Boer*
*Animal Production Systems Group, Wageningen University, 6700 AH Wageningen, the Netherlands
†GD Animal Health Service, 7400 AA Deventer, the Netherlands
‡UMR1213 Herbivores, INRA, Saint-Genès-Champanelle 63122, France
§Biometris, Wageningen University, 6700 AH Wageningen, the Netherlands

ABSTRACT

The Welfare Quality multi-criteria evaluation (WQ-ME) model aggregates scores of single welfare measures into an overall assessment for the level of animal welfare in dairy herds. It assigns herds to 4 welfare classes: unacceptable, acceptable, enhanced, or excellent. The aim of this study was to demonstrate the relative importance of single welfare measures for WQ-ME classification of a selected sample of Dutch dairy herds. Seven trained observers quantified 63 welfare measures of the Welfare Quality protocol in 183 loose housed- and 13 tethered Dutch dairy herds (herd size: 10 to 211 cows). First, values of welfare measures were compared among the 4 welfare classes, using Kruskal-Wallis and Chi-squared tests. Second, observed values of single welfare measures were replaced with a fictitious value, which was the median value of herds classified in the next highest class, to see if improvement of a single measure would enable a herd to reach a higher class. Sixteen herds were classified as unacceptable, 85 as acceptable, 78 as enhanced, and none as excellent. Classification could not be calculated for 17 herds because data were missing (15 herds) or data were deemed invalid because the stockperson disturbed behavioral observations (2 herds). Herds classified as unacceptable showed significantly more very lean cows, more severely lame cows, and more often an insufficient number of drinkers than herds classified as acceptable. Herds classified as acceptable showed significantly more cows with high somatic cell count, with lesions, that could not be approached closer than 1 m, colliding with components of the stall while lying down, and lying outside the lying area, and showed fewer cows with diarrhea, more often had an insufficient number of drinkers, and scored lower for the descriptors "relaxed" and "happy" than herds classified as enhanced. Increasing the number of drinkers and reducing the percentage of cows colliding with components of the stall while lying down were the changes most effective in allowing herds classified as unacceptable and acceptable, respectively, to reach a higher class. The WQ-ME model was not very sensitive to improving single measures of good health. We concluded that a limited number of welfare measures had a strong influence on classification of dairy herds. Classification of herds based on the WQ-ME model in its current form might lead to a focus on improving these specific measures and divert attention from improving other welfare measures. The role of expert opinion and the type of algorithmic operator used in this model should be reconsidered.

Key words: dairy cattle, Welfare Quality, classification, multi-criteria evaluation

INTRODUCTION

The need for methods to assess the overall level of animal welfare on farms has been stressed frequently (e.g., European Commission, 2002; Blokhuis et al., 2003). An overall level of farm animal welfare can facilitate product labeling, encourage producers to improve animal welfare, and, in the future, might become part of export legislation (Blokhuis et al., 2010). Various measures are used to assess animal welfare; for example, animal behavior, heart rate, or cortisol levels in blood (Broom and Fraser, 2007). Measures need to be combined, however, to determine an overall level of animal welfare on farms. Although it has been argued that science should not attempt to perform overall welfare assessment because value judgments are inherently involved (e.g., Fraser, 1995), others state that overall welfare assessment is not arbitrary and a high level of accuracy can be achieved (Bracke et al., 1999). In spite of different viewpoints, various models have been developed to assess overall level of animal welfare; for example, the Animal Needs Index in Austria and Germany (Bartussek et al., 2000), and a decision support

Received September 5, 2012.
Accepted June 19, 2013.
[1]Corresponding author: marion.devries@wur.nl

system for overall welfare assessment of sows in the Netherlands (Bracke et al., 2002).

More recently, Welfare Quality multi-criteria evaluation (**WQ-ME**) models have been developed for different livestock species in the Welfare Quality project (Botreau et al., 2009). Inputs for the WQ-ME model for dairy cattle are on-farm welfare measures described in the Welfare Quality assessment protocol (Welfare Quality, 2009). Compared with other models that combine welfare measures in an overall score, a large proportion of welfare measures in this WQ-ME model are animal based. Animal-based measures for assessing welfare are increasingly preferred over resource-based measures among animal welfare scientists, because they are more closely linked to the welfare of animals and can measure the actual state of animals, regardless of how they are housed or managed (Bartussek, 1999; Whay et al., 2003; Webster, 2009; Rushen et al., 2011). The WQ-ME model uses different algorithmic operators (e.g., a decision tree or a weighted sum) to aggregate measures into an overall score (Botreau et al., 2008b). These operators were parameterized based on value judgments of animal and social scientists and partners and members of the Welfare Quality project on the relative importance of the different welfare measures in the Welfare Quality protocol (Botreau et al., 2008a,b, 2009). The WQ-ME model assigns dairy herds to 1 of 4 welfare classes: unacceptable, acceptable, enhanced, or excellent. These welfare classes should reflect the multi-dimensional nature of welfare and the relative importance of various welfare measures (Botreau et al., 2007a,b).

The WQ-ME model was tested on 69 commercial European dairy herds visited during the Welfare Quality project and partly adjusted according to these results. Although classification of some of these herds was compared with the general impression of observers who audited the farms (Botreau et al., 2009), it has not been demonstrated to what extent classification reflects the relative importance of welfare measures and the multi-dimensional nature of welfare. Such a validation is essential, however, to determine if the model is suitable for its intended purpose. Moreover, besides validity of the model for the 69 herds of the source population (i.e., internal validity), the validity of the model should be tested in other herds (i.e., external validity; Dohoo et al., 2009). Valid welfare classes are essential because they will guide improvements that should positively affect the welfare of farm animals. The aim of this study, therefore, was to demonstrate the relative importance of single welfare measures for WQ-ME classification of a selected sample of Dutch dairy herds.

MATERIALS AND METHODS

Herd Selection

To properly demonstrate the relative importance of single welfare measures for WQ-ME classification, we aimed for data from herds that spanned a wide range of levels of animal welfare. Therefore, herds were selected based on a composite health score (**CHS**). For 5,000 Dutch dairy herds participating in the health scheme of a Dutch dairy cooperative, we calculated a CHS between 0 (worst) and 50 (best). The CHS, for which we used readily available data in herd databases from January 2008 through June 2009, consisted of 5 variables that have been shown to correlate with animal welfare (de Vries et al., 2011): cow mortality, young stock mortality, bulk tank milk SCC, new udder infections, and fluctuations in standardized milk production. Herds were assigned zero points per variable when it was among the 10% worst values and 10 points when it was among the 90% best values of all dairy herds in 2004. Subsequently, 500 herds were approached to participate in the study: 250 herds were randomly selected from the 5% lowest CHS (i.e., CHS $\leq$40) and 250 herds from the 95% highest CHS (i.e., CHS $>$40). Of the 500 herds, 163 farmers responded positively, 75 responded negatively, and 262 failed to respond. In these 3 respective groups, 45, 49, and 64% were from the 5% lowest CHS (i.e., CHS $\leq$40). Nonresponders were contacted by phone. In total, 196 farmers agreed to participate: 90 from the 5% lowest CHS and 106 from the 95% highest CHS.

Farm Visits

Seven observers, each with previous experience in dairy production and handling, were trained to use the Welfare Quality assessment protocol for dairy cattle (Welfare Quality, 2009). Herds were randomly distributed among these observers, who were blinded to the herds' CHS. Each observer visited 14 to 48 herds once from November 2009 through March 2010, when cows had been denied access to pasture for at least 2 wk. Observations were made on a predefined number of lactating and dry cows (for sample sizes, see Welfare Quality, 2009). Data were collected on the cow and herd level, depending on the type of measurement. After data collection, data were expressed as welfare measures at the herd level. These welfare measures could be either continuous or categorical and were expressed on different scales depending on the measure (e.g., percentage of severely lame cows or mean time to lie down).

Aggregation of Welfare Measures into a WQ-ME Classification

The Welfare Quality assessment protocol for dairy cattle consists of 63 welfare measures, which were aggregated following a 3-step aggregation process (Welfare Quality, 2009; Figure 1): 63 welfare measures were aggregated into 12 criteria, these 12 criteria were aggregated into 4 principles, and these 4 principles were aggregated into 1 classification. Different types of algorithmic operators were used in this aggregation process: decision tree, weighted sum, linear combination, conversion to ordinal score, least squares spline curve fitting, and Choquet integral (Figure 1).

In the first step of the aggregation process, decision trees were used to aggregate categorical measures into 3 criteria. A decision tree leads to several possible outcomes, each of which was attributed a criterion score (based on expert opinion). For other criteria, welfare measures were first combined into a weighted sum or converted to an ordinal score representing, for example, no problem, a moderate problem, or a severe problem. The numbers of moderate and severe problems were then combined into a weighted sum, a so-called index value, on a scale from 0 (worst) to 100 (best). Finally, these index values and remaining welfare measures were converted to a criterion score (expressed on the same 0–100 scale), using spline functions (Ramsay, 1988) that were fitted by least-square methods. A detailed description and the rationale behind the use of algorithmic operators in the construction of criteria can be found in Botreau et al. (2007b, 2008a,b) and Veissier et al. (2011).

In the second step, a Choquet integral (Choquet, 1953; Grabisch et al., 2008) was used to aggregate the 12 criteria into 4 principles (Figure 1). This integral uses weights to combine the different criterion scores into 1 principle score (expressed on the 0–100 scale), while limiting the possibility that a poor score of one criterion is compensated for by excellent scores of others (Botreau et al., 2007b; Veissier et al., 2011). These weights, therefore, depend on the values of the criterion scores, whereas the sum of these weights equals 1. For example, when the criterion score for "absence of prolonged hunger" was lower than the criterion score for "absence of prolonged thirst," the weights attributed to "absence of prolonged hunger" and "absence of prolonged thirst" were 0.73 and 0.27. When the criterion score for "absence of prolonged hunger" was higher than the score for "absence of prolonged thirst," however, the weights attributed to "absence of prolonged hunger" and "absence of prolonged thirst" were 0.12 and 0.88. Values for weights were based on expert opinion (Botreau et al., 2008b).

Finally, herds were assigned to 1 of 4 welfare classes: unacceptable, acceptable, enhanced, or excellent, based on reference profiles for the 4 principles (Botreau et al., 2009): to be classified as excellent, a herd must score >55 for each principle and >80 for 2 principles; to be classified as enhanced, each principle must be >20 and at least 2 principles must be >55; to be classified as acceptable, each principle must be >10 and at least 3 principles must be >20. Herds that did not comply with the minimum scores were classified as unacceptable, which means that at least 1 principle was ≤ 10 or at least 2 principles were ≤ 20.

To parameterize the algorithmic operators used for aggregation of welfare measures and criteria, virtual and empirical data sets were presented to expert panels of 13 animal scientists (measures) and 14 animal and social scientists (criteria), who individually ranked farms and gave an absolute score on the 0–100 scale for each farm presented in each of the data sets (Botreau et al., 2008a,b). Partners of the Welfare Quality project, a task force, and members of the Management Committee and Advisory Committee (i.e., stakeholder representatives) were consulted to agree upon parameters for the aggregation of principles into an overall classification (Botreau et al., 2009).

The WQ-ME model was programmed in GenStat for Windows (release 14; VSN International Ltd., Hemel Hempstead, UK) following the Welfare Quality report for the construction of criteria (Botreau et al., 2008a) and the Welfare Quality assessment protocol for dairy cattle (Welfare Quality, 2009) for the construction of principles and classification.

Data Analyses

To evaluate if herd selection based on CHS resulted in a wider range of animal welfare levels and in a larger proportion of herds in lower WQ-ME classes, we compared welfare measures and classification of herds selected from the 5% lowest CHS with those in herds selected from the 95% highest CHS. In addition, we evaluated whether herds in the 2 CHS groups (5% lowest versus 95% highest) were distributed equally across observers. Mann-Whitney U and Chi-squared tests were used, because the assumption of normality was often not appropriate.

To demonstrate the relative importance of single welfare measures for WQ-ME classification, classification of herds was evaluated in 2 ways: by comparing welfare measures of herds in the 4 WQ-ME classes, to determine whether groups of herds in these classes differed; and by evaluating of the effect of replacing observed values for welfare measures with improved, fictitious values on herd classification (sensitivity analyses), to

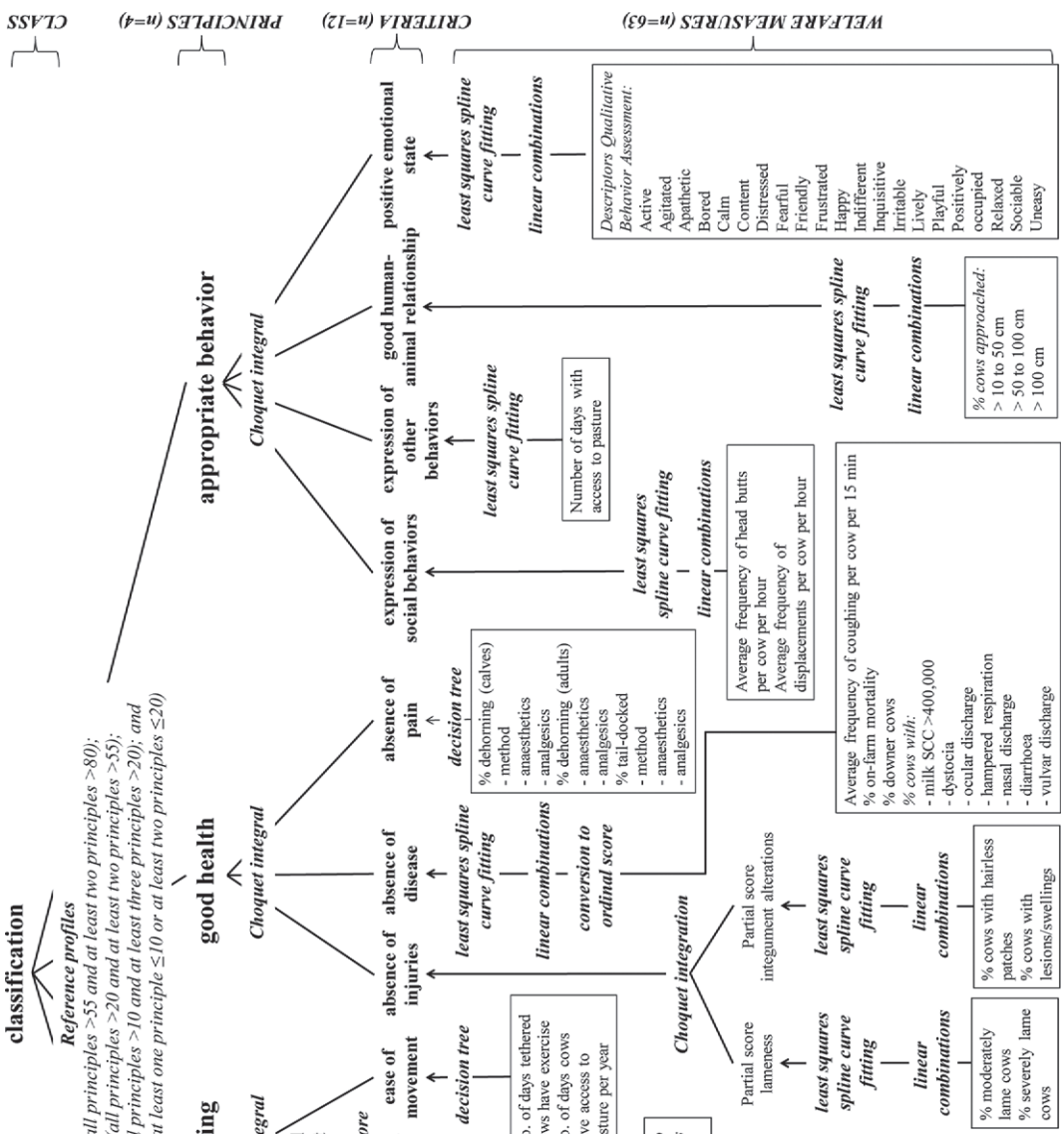

Figure 1. Welfare measures, criteria, and principles of the Welfare Quality multi-criteria evaluation model (adapted from Welfare Quality, 2009).

determine which improvements were most effective in allowing herds to reach a higher classification.

Comparison of WQ-ME Classes. We compared welfare measures for herds in the different WQ-ME classes using the Kruskal-Wallis and Chi-squared tests, because the assumption of normality was often not appropriate. Post hoc pairwise comparisons were made using Mann-Whitney U and Chi-squared tests. Analyses were performed in SPSS 17.0 (IBM SPSS Inc., Chicago, IL). Welfare measures were not considered for analyses when the standard deviation was zero or the prevalence was <5%.

Sensitivity Analyses. For welfare measures that differed between adjacent classes, herds were assigned an improved, fictitious value to see whether improving this single measure enabled a herd to reach a higher WQ-ME classification. The improved value, which replaced the observed value, was the median value of herds classified in the next highest class. This median value was considered to be a realistic and feasible value that farmers might aspire to when aiming to improve their classification. For categorical measures, such as sufficiency of the number of drinkers, the improved value was the mode of herds in the next highest class. After assigning an improved value to a herd, a new classification was computed. For each single measure, the effect of improvement was evaluated by counting the number of herds that reached a higher classification.

RESULTS

Of the selected sample of 196 Dutch dairy herds, the WQ-ME model classified 16 herds as unacceptable, 85 as acceptable, 78 as enhanced, and none as excellent. Classification could not be calculated for 17 herds, because data of one or more welfare measures were missing (15 herds) or data were deemed invalid because the stockperson disturbed behavioral observations (2 herds). Eight welfare measures, related to drinking, tethering, dehorning, and tail-docking, were excluded from the statistical analysis because of no variability (SD = 0) or a prevalence <5%.

Median size of the 179 herds included was 67 lactating cows (ranging from 10 to 211 cows), with a milk production of 25.4 kg/cow per day (ranging from 10.0 to 35.2 kg/cow per day). Cows were in loose housing in 169 herds and tethered in 10 herds. In summer, cows had access to pasture for at least 6 h/d in 132 herds. Herd size, milk production, type of housing, access to pasture, and observer did not differ among WQ-ME classes.

Median (range) of welfare measures for herds selected from the 5% lowest and 95% highest CHS are in Table 1. Herds selected from the 5% lowest CHS showed more cows housed in tiestalls, more with dirty hindquarters, more with SCC >400,000 cells/mL, fewer with diarrhea, higher on-farm mortality, fewer calves disbudded, and with lower scores for 8 descriptors of the Qualitative Behavior Assessment (Rousing and Wemelsfelder, 2006; Wemelsfelder, 2007) than herds selected from the 95% highest CHS ($P < 0.05$). Herds in the 2 CHS groups (5% lowest vs. 95% highest) did not differ in WQ-ME class or in observer.

Comparison of WQ-ME Classes

Median (range) of welfare measures for herds classified as unacceptable, acceptable, and enhanced are given in Table 2. Because no herds were classified as excellent, this class could not be compared with other WQ-ME classes.

Unacceptable Compared with Acceptable and Enhanced. Herds classified as unacceptable showed 5.9 and 7.5% more very lean cows, 4.0 and 5.8% more severely lame cows, and 1.7 and 2.2 times more often an insufficient number of drinkers than herds classified as acceptable and enhanced (Table 2). In addition, herds classified as unacceptable showed 18.1% more cows colliding with components of the stall while lying down compared with herds classified as enhanced. No differences were found for the other 59 welfare measures.

Acceptable Compared with Enhanced. More, but generally smaller, differences in welfare measures were found between herds classified as acceptable and enhanced than between herds classified as unacceptable and other classes. Herds classified as acceptable showed 20.6% more cows colliding with components of the stall while lying down, 1.2% more lying outside the lying area, 13.5% more with lesions or swellings, 2.3% more with an SCC >400,000 cells/mL, 2.2% fewer with diarrhea, 6.6% more that could not be approached closer than 1 m, and 1.3 times more often an insufficient number of drinkers. In addition, herds classified as acceptable scored 18 and 19 points less for the descriptors "relaxed" and "happy" for the Qualitative Behavior Assessment than herds classified as enhanced. Because herds classified as enhanced showed more cows with diarrhea, this measure was not included in the sensitivity analysis.

Sensitivity Analysis

The numbers of herds that changed to a higher classification when observed values of single welfare measures were replaced with an improved value are shown in Table 3.

Table 1. Median (range) of welfare measures[1] for herds selected from the 5% lowest, and 95% highest composite health scores (CHS)[2]

Welfare measure	Herds selected from: 5% lowest CHS (n = 90)	Herds selected from: 95% highest CHS (n = 89)	*P*-value
Percentage of cows			
Very lean	3.1 (0–28.6)	2.0 (0–20.0)	0.086
Dirty udder	15.1 (0–93.9)	11.4 (0–64.7)	0.074
Dirty hindquarters	45.7 (0–100)	28.0 (0–100)	0.015
Lame	26.6 (0–52.5)	21.3 (3.3–58.7)	0.090
Severely lame	6.2 (0–46.9)	3.8 (0–65.9)	0.087
Milk SCC >400,000 cells/mL	13.8 (2.6–36.3)	8.4 (0–24.9)	0.000
Diarrhea	0 (0–46.5)	2.1 (0–34.2)	0.016
On-farm mortality	0.8 (0–30.0)	0.4 (0–3.1)	0.000
Average coughing per cow per 15 min (no.)	0.07 (0–0.4)	0.06 (0–0.2)	0.077
Tethered (no.)	No (81) Yes (9)	No (88) Yes (1)	0.010
Dehorning calves (no.)	No (10) Yes (80)	No (1) Yes (88)	0.005
QBA descriptors[3]			<0.10

[1]Measures with $P > 0.10$ not shown.

[2]CHS based on cow and young stock mortality, bulk tank milk SCC, new udder infections, and fluctuations in standardized milk production.

[3]Median and range of descriptors (relaxed, agitated, calm, content, fearful, happy, irritable, lively, positively occupied) for the Qualitative Behavior Assessment (QBA; Rousing and Wemelsfelder, 2006; Wemelsfelder, 2007) not shown.

Herds Originally Classified as Unacceptable. Replacing observed values of single measures with improved values resulted in a higher class for 14 of the 16 herds originally classified as unacceptable. When the observed percentage of very lean cows was replaced with an improved percentage of 3.3% (i.e., the median score of herds classified as acceptable), 11 of the 16 herds originally classified as unacceptable changed to acceptable. When the number of drinkers was changed to sufficient, 13 herds changed class from unacceptable: 9 to acceptable and 4 to enhanced. When the percentage of severely lame cows was lowered to 5.3%, 1 herd changed to acceptable.

Herds Originally Classified as Acceptable. Replacing observed values of single measures with improved values resulted in an enhanced class for 38 of the 85 herds originally classified as acceptable. Most of these herds changed to enhanced when the percentage of cows colliding with the stall, lying outside the lying area, that could not be approached closer than 1 m was lowered, and when the number of drinkers was changed to sufficient. Replacing the percentage of

Table 2. Median (range) of welfare measures that differed between herds classified as unacceptable, acceptable, and enhanced

Welfare measure	Class: Unacceptable (n = 16)	Class: Acceptable (n = 85)	Class: Enhanced (n = 78)	Overall *P*-value
Percentage of cows				
Very lean	9.2^{a} (0–20.0)	3.3^{b} (0–23.7)	1.7^{b} (0–28.6)	0.001
Colliding with components of the stall while lying down	37.5^{a} (10.0–66.7)	40.0^{a} (0–100)	19.4^{b} (0–88.2)	0.004
Lying outside lying area	0.7^{ab} (0–12.8)	1.5^{a} (0–15.4)	0.3^{b} (0–8.6)	0.001
Severely lame	9.3^{a} (0–65.9)	5.3^{b} (0–46.9)	3.5^{b} (0–25.4)	0.020
Lesions or swellings	40.4^{ab} (3.3–94.7)	42.9^{a} (0–97.6)	29.4^{b} (3.3–95.1)	0.005
Milk SCC >400,000 cells/mL	10.8^{ab} (5.4–20.9)	12.5^{a} (0–26.9)	10.2^{b} (1.1–36.3)	0.045
Diarrhea	0^{ab} (0–36.4)	0^{b} (0–30.3)	2.2^{a} (0–46.5)	0.011
Not approached <1 m	25.3^{ab} (11.9–47.3)	24.4^{a} (0–74.4)	17.8^{b} (0–66.0)	0.049
Sufficient number of drinkers (no.)	No^{a} (14) Yes (2)	No^{b} (44) Yes (41)	No^{c} (31) Yes (47)	0.008
Happy	42^{ab} (1–90)	40^{b} (1–123)	59^{a} (1–115)	0.003
Relaxed	66^{ab} (1–117)	51^{b} (1–118)	69^{a} (1–117)	0.014

$^{a–c}$Medians within a row with different superscripts differ between classes ($P < 0.05$).

Table 3. Number of herds changing to a higher classification when welfare measures, of which the median value differed between adjacent welfare classes, were replaced with an improved value

Original class	Welfare measure	Original value (median)	Improved value[1]	Herds changed class to:[2]		
				Acceptable	Enhanced	Excellent
Unacceptable (n = 16)	Very lean cows (%)	9.2	3.3	11	0	0
	Sufficient number of drinkers (no.)	No (14); Yes (2)	No (0); Yes (16)	9	4	0
	Severely lame cows (%)	9.3	5.3	1	0	0
Acceptable (n = 85)	Sufficient number of drinkers (no.)	No (44); Yes (41)	No (0); Yes (85)		9	0
	Cows colliding with components of the stall while lying down (%)	40.0	19.4		21	0
	Cows lying outside lying area (%)	1.5	0.3		9	0
	Cows with lesions or swellings (%)	42.9	29.4		1	0
	Cows with SCC >400,000 cells/mL (%)	12.5	10.2		0	0
	Cows not approached <1 m (%)	24.4	17.8		7	0
	Happy	40	59		0	0
	Relaxed	51	69		0	0

[1]The improved score was the median score of herds in the next highest classification.
[2]The same herds can appear in different rows.

cows with lesions or swellings, with milk SCC >400,000 cells/mL, and descriptors for the Qualitative Behavior Assessment "happy" and "relaxed" with an improved value rarely resulted in an enhanced class.

Herds Originally Classified as Enhanced. A median value of the next highest class was not available for herds originally classified as enhanced, because no herds were classified as excellent. When we replaced values for welfare measures of herds originally classified as enhanced by an improved value that was equal to the maximum value of all herds, no herds changed to excellent.

DISCUSSION

The WQ-ME model classified 16 herds as unacceptable, 85 as acceptable, 78 as enhanced, and none as excellent. The distribution of herds among classes was not representative of the Dutch dairy sector, because herds in this study were selected based on CHS.

The CHS was useful in selecting for variation in a large number of welfare measures. Although we expected that herd selection based on CHS would increase the proportion of herds in lower WQ-ME classes, no differences among herds with varying CHS were found in the final classification. Selection based on CHS apparently concerned welfare measures other than those that were important for classification. Associations between variables that formed the CHS and welfare measures mainly responsible for classification of herds (e.g., number of drinkers) are also absent in literature (de Vries et al., 2011).

Relative Importance of Welfare Measures for WQ-ME Classification

The most important welfare measures for classifying herds as unacceptable in our study were percentage of very lean cows and sufficiency of drinkers. Herds classified as unacceptable showed a higher percentage of severely lame cows than herds classified as acceptable, but this measure appeared to have little influence on classification when a sensitivity analyses was performed. Although no gold standard exists for the overall level of animal welfare against which results of the WQ-ME model can be validated, results can be compared with expert opinion on the relative importance of welfare measures in other studies. In the study of Lievaart and Noordhuizen (2011), animal welfare experts ranked competition for feed and water as the second most important measure of dairy cattle welfare, which could be considered consistent with the percentage of very lean cows and sufficient number of drinkers being the most important welfare measures for classifying herds as un-

acceptable in our study. However, number of drinkers is a resource-based measure that is less closely linked than an animal-based measure to animal welfare (Webster et al., 2004; Blokhuis, 2008). Water intake is associated with the number and size of drinkers in herds (Pinheiro Machado Filho et al., 2004; Teixeira et al., 2006) but can be influenced by various other factors, such as diet or climate conditions (Dahlborn et al., 1998; Meyer et al., 2004). The value of such a resource-based measure being responsible for classification as unacceptable is, therefore, questionable.

In 2 studies, animal welfare experts ranked lameness as the most important measure of dairy cattle welfare (Whay et al., 2003; Lievaart and Noordhuizen, 2011). In our study, except for one herd, a high prevalence of (severely) lame cows did not result in herds being classified as unacceptable. Percentage of severely lame cows reached 47% in herds classified as acceptable and reached 25% in herds classified as enhanced. Mastitis, which was represented by cows with SCC >400,000 cells/mL in our study, was among the most important measures of dairy cattle welfare in the study of Whay et al. (2003). Although the percentage of cows with SCC >400,000 cells/mL reached 36% in our study, a high prevalence of cows with SCC >400,000 cells/mL did not result in herds being classified as unacceptable. In contrast, a herd with 36% cows having SCC >400,000 cells/mL was classified as enhanced.

Compared with herds classified as unacceptable and acceptable, more differences were found between herds classified as acceptable and enhanced, which was evident in welfare measures of each of the 4 principles of the WQ-ME model. This finding achieved the aim of the WQ-ME model in reflecting the multi-dimensional concept of animal welfare (Botreau et al., 2007c). Improving measures of principles "good feeding" and "good housing" allowed a large number of herds originally classified as acceptable to reach a higher class, whereas improving measures of "good health" was effective in almost none of these herds. This lack of effect was because little difference existed between median measure scores of herds classified as acceptable and enhanced. This showed that, in spite of substantial variation in measure scores among our study herds, relative importance of measures of "good health" for classification was low. This contradicts with results of Whay et al. (2003), in which health records were ranked as the second most important measure of dairy cattle welfare. It should be emphasized, however, that analyses in the current study were limited to single welfare measures. Effects of improving combinations of welfare measures should be further investigated.

None of the herds in our study was classified as excellent. A similar result was found by Botreau et al. (2009), who classified a sample of 69 dairy herds in Austria, Germany, and Italy. The reason that no herds were classified as excellent in our study was a lack of simultaneous excellent scores for a large number of welfare measures. High scores were lacking, especially for welfare measures of the principles "good health" and "appropriate behavior." Improvement of welfare measures in herds originally classified as enhanced did not lead to a changed class of excellent. Apparently, improvement of more than one welfare measure is needed to reach a classification of excellent.

Reasons for a Lack of Influence of Lameness and SCC on WQ-ME Classification

The lack of effect of lameness on herd classification was caused mainly by compensating mechanisms in the first 2 steps of the aggregation process in the WQ-ME model: the construction of the criterion "absence of injuries" and the principle "good health." A herd with 48% moderately lame cows, 29% severely lame cows, 57% cows with lesions and swellings, and 7% cows with hairless patches, for example, obtained a score of 14 for the criterion "absence of injuries." In the construction of the principle "good health," this criterion score was compensated by a score of 65 for the criterion "absence of disease" and a score of 52 for the criterion "absence of pain," leading to a principle score of 26. Given the reference profiles for classification, a herd is classified as unacceptable only when principle scores are <20. Therefore, this principle score did not lead to an unacceptable class.

High percentages of cows with SCC >400,000 cells/mL did not result in herds classified as unacceptable because this measure was converted to an ordinal score (no, moderate, or severe problem) to calculate a score for the criterion "absence of disease." Because this percentage represented a severe problem whenever it was >4.5%, the WQ-ME model did not distinguish between, for example, herds with 27% cows and herds with 5% cows with SCC >400,000 cells/mL. Moreover, a severe problem for the percentage of cows with SCC >400,000 cells/mL was compensated for by other welfare measures that represented no problem, because they were linearly combined for the criterion "absence of disease." Similar to lameness and SCC, other welfare measures of the principle "good health" rarely influenced classification. This is illustrated by the principle "good health," which, despite a large variation in welfare measures, ranged from 21 to 58 (95% range), compared with the principle "good feeding," which ranged from 7.5 to 100 (95% range). As a consequence of the lack of effect on herd classification, farmers might not be motivated to improve welfare measures of "good health."

In summary, 2 major factors explain why severe welfare problems did not result in herds being classified as unacceptable. First, although it was emphasized in the development of the WQ-ME model that welfare scores should not compensate each other (Veissier et al., 2011), compensation occurred for welfare measures that were aggregated using linear combinations and the Choquet integral in the first 2 aggregation steps of the WQ-ME model. The extent of compensation depended on the weight given to welfare measures and criteria, which was derived from expert opinion (Botreau et al., 2008a). The role of expert opinion in the WQ-ME model requires further investigation. Grouping a large number of welfare measures in a principle may have increased compensation. In contrast to the principle "good feeding," for example, which considers 4 welfare measures, the principle "good health" considers 20 welfare measures simultaneously. Second, conversion of welfare measures to an ordinal score makes it impossible for the WQ-ME model to distinguish between herds that slightly or largely exceeded thresholds for severe problems. Consequently, severe welfare problems, such as SCC >400,000 cells/mL in more than 35% of the cows, did not result in a classification of unacceptable. In addition to evaluating the role of expert opinion in the WQ-ME model, reconsidering the choice of algorithmic operator might help ensure that herds with severe welfare problems are classified more appropriately.

CONCLUSIONS

The aim of this study was to demonstrate the relative importance of single welfare measures for WQ-ME classification of a selected sample of Dutch dairy herds. A limited number of welfare measures had a strong influence on classification of dairy herds in this study, especially for herds classified as unacceptable. Classification of herds based on the WQ-ME model in its current form might, on the one hand, lead to improving these specific measures but, on the other hand, divert attention from improving other measures. The role of expert opinion and the type of algorithmic operator used to aggregate welfare measures in the WQ-ME model need to be reconsidered, to assign herds to the most appropriate of the 4 welfare classes.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the farmers for participating in this study, and Kees van Reenen (Livestock Research, Wageningen UR, Wageningen, the Netherlands), Wim Swart (GD Animal Health Service, Deventer, the Netherlands), Jac Thissen (Biometris, Wageningen University, Wageningen, the Netherlands), Ingrid den Uijl (GD Animal Health Service), and Pieter Vereijken (Biometris) for their stimulating discussions and ideas.

REFERENCES

Bartussek, H. 1999. A review of the animal needs index (ANI) for the assessment of animals' well-being in the housing systems for Austrian proprietary products and legislation. Livest. Prod. Sci. 61:179–192.

Bartussek, H., C. H. M. Leeb, and S. Held. 2000. Animal Needs Index for Cattle: ANI35L/2000 cattle. Federal Research Institute for Agriculture in Alpine Regions, BAL Gumpenstein, Irdning, Austria.

Blokhuis, H. J. 2008. International cooperation in animal welfare: The Welfare Quality project. Acta Vet. Scand. 50(Suppl. 1):S10.

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 farm animal well-being. Acta Agric. Scand. A Anim. Sci. 60:129–140.

Botreau, R., M. Bonde, A. Butterworth, P. Perny, M. B. M. Bracke, J. Capdeville, and I. Veissier. 2007a. Aggregation of measures to produce an overall assessment of animal welfare. Part 1: A review of existing methods. Animal 1:1179–1187.

Botreau, R., M. B. M. Bracke, P. Perny, A. Butterworth, J. Capdeville, C. G. Van Reenen, and I. Veissier. 2007b. Aggregation of measures to produce an overall assessment of animal welfare. Part 2: Analysis of constraints. Animal 1:1188–1197.

Botreau, R., J. Capdeville, B. Engel, P. Perny, and I. Veissier. 2008a. Reports on the construction of welfare criteria for different livestock species. Part 2: Subcriteria construction for dairy cows on farm. Deliverable 2.8b, subtask 2.3.1.2. Welfare Quality® (EU Food-CT-2004-506508), Lelystad, the Netherlands.

Botreau, R., J. Capdeville, P. Perny, and I. Veissier. 2008b. Multi-criteria evaluation of animal welfare at farm level: An application of MCDA methodologies. Found. Comput. Decision Sci. 33:1–18.

Botreau, R., I. Veissier, A. Butterworth, M. B. M. Bracke, and L. J. Keeling. 2007c. 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.

Bracke, M. B. M., B. M. Spruijt, and J. H. M. Metz. 1999. Overall animal welfare assessment reviewed. Part 1: Is it possible? Neth. J. Agric. Sci. 47:279–291.

Bracke, M. B. M., B. M. Spruijt, J. H. M. Metz, and W. G. P. Schouten. 2002. Decision support system for overall welfare assessment in pregnant sows. A: Model structure and weighting procedure. J. Anim. Sci. 80:1819–1834.

Broom, D. M., and A. F. Fraser. 2007. Domestic Animal Behavior and Welfare. 4th ed. CABI, Cambridge, MA.

Choquet, G. 1953. Theory of capacities. Annales de l'Institut Fourier 5:132–295.

Dahlborn, K., M. Åkerlind, and G. Gustafson. 1998. Water intake by dairy cows selected for high or low milk-fat percentage when fed two forage to concentrate ratios with hay or silage. Swed. J. Agric. Res. 28:167–176.

de Vries, M., E. A. M. Bokkers, T. Dijkstra, G. van Schaik, and I. J. M. de Boer. 2011. Invited review: Associations between variables of routine herd data and dairy cattle welfare indicators. J. Dairy Sci. 94:3213–3228.

Dohoo, I. R., S. W. Martin, and H. Stryhn. 2009. Veterinary Epidemiologic Research. 2nd ed. VER Inc., Charlottetown, PEI, Canada.

European Commission. 2002. Communication from the European Commission to the Council and the European Parliament on animal welfare legislation on farmed animals in third countries and the implications for the EU. European Union, Brussels, Belgium.

Fraser, D. 1995. Science, values and animal welfare: Exploring the "inextricable connection". Anim. Welf. 4:103–117.

Grabisch, M., I. Kojadinovic, and P. Meyer. 2008. A review of methods for capacity identification in Choquet integral based multi-attribute utility theory applications of the Kappalab R package. Eur. J. Oper. Res. 186:766–785.

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.

Meyer, U., M. Everinghoff, D. Gadeken, and G. Flachowsky. 2004. Investigations on the water intake of lactating dairy cows. Livest. Prod. Sci. 90:117–121.

Pinheiro Machado Filho, L. C., D. L. Teixeira, D. M. Weary, M. A. G. von Keyserlingk, and M. J. Hötzel. 2004. Designing better water troughs: Dairy cows prefer and drink more from larger troughs. Appl. Anim. Behav. Sci. 89:185–193.

Ramsay, J. O. 1988. Monotone regression splines in action. Stat. Sci. 3:425–441.

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

Rushen, J., A. Butterworth, and J. C. Swanson. 2011. Farm animal welfare assurance: Science and application. J. Anim. Sci. 89:1219–1228.

Teixeira, D. L., M. J. Hötzel, and L. C. Pinheiro Machado Filho. 2006. Designing better water troughs: 2. Surface area and height, but not depth, influence dairy cows' preference. Appl. Anim. Behav. Sci. 96:169–175.

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

Webster, A. J. F. 2009. The virtuous bicycle: A delivery vehicle for improved farm animal welfare. Anim. Welf. 18:141–147.

Webster, A. J. F., D. C. J. Main, and H. R. Whay. 2004. Welfare assessment: Indices from clinical observation. Anim. Welf. 13(Suppl.):S93–S98.

Welfare Quality. 2009. Welfare Quality® Assessment Protocol for Cattle. Welfare Quality® Consortium, Lelystad, the Netherlands.

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

Whay, H. R., D. C. J. Main, L. E. Green, and A. J. F. Webster. 2003. 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.