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).