Sensitivity of the integrated Welfare Quality® scores to changing values of individual dairy cattle welfare measures

de Graaf, S; Ampe, B; Buijs, S; Andreasen, S; de Boyer des Roches, A; Haskell, MJ; Kirchner, M; Mounier, L; Radeski, M; Winckler, C; Bijttebier, J; Lauwers, L; Verbeke, W; Tuyttens, F

Published in:
Animal Welfare

DOI:
10.7120/09627286.27.2.157

First published: 01/05/2018

Document Version
Peer reviewed version

Citation for pulished version (APA):
de Graaf, S., Ampe, B., Buijs, S., Andreasen, S., de Boyer des Roches, A., Haskell, MJ., Kirchner, M., Mounier, L., Radeski, M., Winckler, C., Bijttebier, J., Lauwers, L., Verbeke, W., & Tuyttens, F. (2018). Sensitivity of the integrated Welfare Quality® scores to changing values of individual dairy cattle welfare measures. *Animal Welfare*, *27*(2), 157 - 166. https://doi.org/10.7120/09627286.27.2.157

Animal Welfare

Sensitivity of the integrated Welfare Quality® scores to changing values of individual dairy cattle welfare measures

Journal:	*Animal Welfare*
Manuscript ID	F2001.R2
Manuscript Type:	Original Article
Date Submitted by the Author:	09-Nov-2017
Complete List of Authors:	de Graaf, Sophie; ILVO, Farm animal welfare and behaviour Ampe, Bart; Institute for Agricultural and Fisheries Research (ILVO), Animal Sciences Buijs, Stephanie Andreasen, Sine; SEGES P/S, Danish Pig Research Center de Boyer des Roches, Alice; Université de Lyon, VetAgro Sup, UMR 1213 Herbivores; INRA, UMR1213 Herbivores van Eerdenburg, Frank; Faculty of Veterinary Medicine, Farm Animal Health Haskell, Marie; SRUC, Animal and Veterinary Sciences Kirchner, Marlene; University of Copenhagen, Large Animals Mounier, Luc; INRA, UMR1213 Herbivores, ; Université de Lyon, VetAgro Sup, UMR1213 Herbivores, Radeski, Miroslav; Ss. Cyril and Methodius University Winckler, Christoph; University of Natural Resources and Life Sciences, Department of Sustainable Agricultural Systems Bijttebier, Jo; Institute for Agricultural and Fisheries Research (ILVO), Social Sciences Unit Lauwers, Ludwig; Institute for Agricultural and Fisheries Research (ILVO) , Social Sciences Unit Verbeke, Wim; Universiteit Gent, Faculty of Bioscience Engineering, Department of Applied Biosciences Tuyttens, Frank; ILVO, Animal Sciences Unit
Keywords:	animal welfare, animal-based welfare indicators, dairy cattle, integrated welfare index, sensitivity analysis, Welfare Quality

Dairy cattle welfare assessment using Welfare Quality®

Sensitivity of the integrated Welfare Quality® scores to changing values of individual dairy cattle welfare measures

S. de Graaf[1,10], B. Ampe[1], S. Buijs[1], S.N. Andreasen[2], A. De Boyer Des Roches[3,7], F.J.C.M. van Eerdenburg[4], M.J. Haskell[5], M.K. Kirchner[6], L. Mounier[3], M. Radeski[8], C. Winckler[9], J. Bijttebier[1], L. Lauwers[1], W. Verbeke[10], F.A.M. Tuyttens[1]*

[1]Institute for Agricultural and Fisheries Research (ILVO), Burgemeester van Gansberghelaan 92, 9820 Merelbeke, Belgium

[2]Department of Large Animal Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, Groennegaardsvej 8, DK-1870 Frederiksberg, Denmark

[3]Université de Lyon, VetAgro Sup, UMR1213 Herbivores, 69280 Marcy-L'Étoile, France

[4]Department of Herd Animal Health, Utrecht University, 3508 TD Utrecht, The Netherlands

[5] SRUC, West Mains Road, Edinburgh EH9 3JG, Scotland, United Kingdom

[6]University of Copenhagen, Dept. of Veterinary and Animal Sciences, Section of Animal Welfare and Disease Control, Grønnegårdsvej 8, 1870 Frederiksberg Copenhagen, Denmark

[7]Institut National de la Recherche Agronomique, UMR1213 Herbivores, Equipe Comportement Animal, Robustesse et Approche Intégrée du Bien-Etre, 63122 Saint Genes Champanelle, France

[8]Animal Welfare Center, Faculty of Veterinary Medicine, Ss. Cyril and Methodius University in Skopje, Lazar Pop-Trajkov 5-7, 1000 Skopje, Republic of Macedonia

[9]Division of Livestock Sciences, Department of Sustainable Agricultural Systems, University of Natural Resources and Life Sciences, Gregor-Mendel Straße 33, 1180 Vienna, Austria

[10]*Faculty of Bioscience Engineering, Ghent University, Coupure links 653, 9000 Ghent, Belgium*

* Contact for correspondence and requests for reprints: frank.tuyttens@ilvo.vlaanderen.be

Abstract

The Welfare Quality® (WQ) protocol for on-farm dairy cattle welfare assessment describes 33 measures and a step-wise method to integrate the outcomes into 12 criteria scores, grouped into four principle scores and into an overall welfare categorization with four possible levels. The relative contribution of various welfare measures to the integrated scores has been contested. Using a European dataset (491 herds), we investigated 1) variation in sensitivity of integrated outcomes to extremely low and high values of measures, criteria and principles by replacing each actual value with minimum and maximum observed and theoretically possible values and 2) the reasons for this variation in sensitivity. As intended by the WQ consortium, the sensitivity of integrated scores depends on 1) the observed value of the specific measures/criteria, 2) whether the change was positive/negative, and 3) the relative weight attributed to the measures. Additionally, two unintended factors of considerable influence appear to be side-effects of the complexity of the integration method. Namely 1) the number of measures integrated into criteria and principle scores, and 2) the aggregation method of the measures. Therefore, resource-based measures related to drinkers, of which validity to assess absence of prolonged thirst was criticized, have a much larger influence on integrated scores than health-related measures like 'mortality rate' and 'lameness score'. Hence, the integration method of the WQ protocol for dairy cattle should be revised to ensure that the relative contribution of the various welfare measures to the integrated scores more accurately reflect their relevance for dairy cattle welfare.

Keywords: animal welfare, animal-based welfare indicators, dairy cattle, integrated welfare index , sensitivity analysis, Welfare Quality®

Introduction

Accurate welfare assessment is vital for improving animal welfare. In dairy cattle, measures have been developed and validated for a wide variety of both negative and positive aspects of welfare. However, only a few protocols exist that aggregate the scores of multiple welfare measures into one score or index reflecting the overall welfare status of a given herd. Such an overall welfare status score might be used for example in the communication with consumers (food labelling), as an incentive for on-farm welfare improvements and as regulative target (Blokhuis et al 2010). Examples of schemes that calculate an overall welfare status of dairy cattle are a protocol by Whay et al (2003) based on the “Five Freedoms” (Farm Animal Welfare Council 1992) which generates a ranking of herds’ welfare status. The Animal Needs Index (ANI) produces an overall welfare score based on integrating mostly resource-based measures (measures of environmental aspects that affect welfare) (Bartussek et al 2000). Finally, the Welfare Quality® (WQ) protocol categorizes overall welfare status of a herd as ‘excellent’, ‘enhanced’, ‘acceptable’ or ‘not classified’ based on a step-wise integration procedure (Welfare Quality® 2009). The current study focuses on the WQ protocol, as this is the only protocol that predominantly uses animal-based measures to calculate an integrated welfare index. Such measures are generally preferred over resource-based measures as the latter tend to reflect risk factors for welfare impairments instead of directly measuring welfare (Blokhuis et al 2003, 2010).

In the EU project Welfare Quality® (WQ), protocols for the welfare assessment of the main types of farm animals (cattle, pigs and chickens) were proposed. The dairy cattle protocol describes 33 welfare measures performed on-farm by means of behavioural observations, qualitative behaviour assessment, an avoidance distance test, a management questionnaire, a resource checklist and clinical scoring (Table 1). Subsequently, three steps are used to integrate separate measures into one overall welfare category. Measures are first integrated into criteria scores on a scale of 0 – 100 which are in turn collated into four welfare principles (‘good feeding’, ‘good housing’, ‘good health’ and ‘appropriate behaviour’). These principle scores are then used to determine herds’ overall welfare category (Welfare Quality® 2009). Integration methods are intended to limit compensation of poor scores with better scores on other welfare aspects (Veissier et al 2011). Expert opinion of social and animal scientists and stakeholders was used to determine weights for the integration method (Botreau et al 2007). Additionally, the protocols were designed with the intention of modifying and updating assessment methods according to advances in animal welfare science (www.welfarequalitynetwork.net/network/45848/7/0/40).

Discussion has arisen recently about WQ’s measures and integration methods. Some of the measures have been criticised for their poor or undocumented reliability, validity or feasibility (Knierim and Winckler, 2009; de Vries et al., 2013; de Jong et al., 2015; Tuyttens et al., 2015; de Graaf et al., in press). In addition, studies have indicated that a few, resource-based measures have a disproportionately large influence on the overall welfare category (Heath et al 2014; de Vries et al 2013). Both critical findings may harm the credibility and validity of the WQ protocol in assessing herd welfare. To further examine the functioning of the WQ protocol for dairy cattle, the aim of the current study was to examine 1) if there is variation in sensitivity of integrated outcomes (criteria and principle scores and overall welfare category) to extremely low and high values of measures, criteria and principles and 2) the reasons for this variation in sensitivity. More specifically, we aimed to critically evaluate whether differences in sensitivity appear to be deliberate and justifiable rather than unintentional side-effects of the complex integration method. To this end, we performed a sensitivity analysis by replacing individual observed values for a given herd with both the theoretically possible and the actually observed worst and best values. The latter values were based on a large database of WQ data that reflect a wide range of herd types in Europe and thereby ensuring a substantial but realistic spread in observed values.

Materials and methods

WQ protocol

Only a brief description of the integration method of the WQ protocol for on-farm dairy cattle welfare assessment is given here. The full protocol can be found at http://www.welfarequalitynetwork.net/.

Step 1: from measures to criteria scores

Aggregation starts by combining 33 measures into 11 rather than 12 criteria (Table 1), because no data is collected on-farm for the criterion 'thermal comfort'. Because the recording scales of measures differ, various aggregation methods are used. For categorical measures, decision trees are used resulting in a score between 0 – 100 where 100 indicates the best possible score. Other measures are converted to ordinal scores where required (e.g. scores within 'comfort around resting' are converted into three categories: normal, moderate problem or serious problem using thresholds in seconds for time needed to lie down and percentages of cows for the other measures) and then combined into index values using weighted sums. Spline functions are used to re-weight these sums based on their severity according to expert opinion. Finally, when multiple spline functions were used, Choquet integrals are used to combine these functions into criteria scores on a scale of 0 – 100 (Botreau et al 2007). These algorithmic operators calculate the criteria scores in such a way that a poor score cannot be fully compensated for by a better score in another measure (Botreau et al 2007). Consequently, poor scores will have a bigger influence on the integrated scores than good scores. Using Choquet integrals, the weight given to each element (measures or criteria) depends on its value relative to the other elements, where the poorest score always gets the highest weight (Botreau et al., 2008; Welfare Quality 2009).

Step 2: from criterion scores to principle scores

To integrate criterion scores into principle scores, Choquet integrals are used (Welfare Quality 2009). The resulting principle scores range from 0 (worst) to 100 (best). Because no data is collected on-farm for the criterion 'thermal comfort', this criterion score is replaced with the best score among 'comfort around resting' and 'ease of movement'.

Step 3: from principle scores to overall welfare category

The third and final integration step is from principle scores to overall welfare category. Dairy welfare in a herd is considered 'excellent' when it scores >50 for each principle and >75 on two of them. When a herd scores >15 on each principle and >50 on at least two of them, it is classified as 'enhanced'. 'Acceptable' herds score >5 for all principles and >15 for at least three principles. Herds that do not reach the thresholds for the category 'acceptable' are considered 'not classified' (Botreau et al 2009).

<Table 1>

Data collection and collation

To reflect the current range present in Europe across various herding systems, pre-existing research datasets of assessments using the WQ protocol for on-farm dairy cattle welfare were collated from seven European research institutes and included data from 10 countries. The collected samples were selected by the research institutes to be representative for 1) small scale dairy herds in Macedonia (n = 12); 2) non-organic and non-tie stall dairy herds in The Netherlands (n = 60) and France (n = 128); 3) random herds with individual Somatic Cell Count data available (SCC, to be able to calculate WQ scores) in Belgium (n = 140), Scotland (n = 16) and Denmark (n = 42); 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). The total number of herds in the collated database was 491. To ensure a homogenous integration method for all data, 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 programme is available on request. The results were checked for coherence with the INRA WAFA webtool (http://www1.clermont.inra.fr/wq/), in which WQ measure values can be entered (for dairy cows, fattening pigs, growing pigs and broilers), and WQ criteria, principle and classification scores can be calculated.

Sensitivity analysis

In order to investigate the extent to which values for separate measures affected the criteria and principle scores and the overall welfare category, each herd-level observation for each measure and each herd was replaced one by one with both the theoretically possible and the observed (of the entire dataset of 491 herds) worst and best values. This was repeated for individual criteria and principle scores to assess the impact of criteria and principle scores on the overall welfare category. For these calculations, farms that were already in the highest or lowest overall welfare category were excluded. This decision was made because these excluded farms were not able to shift categories, therefore retaining them would give a distorted picture of the results. Subsequently, the median increase and decrease in criteria and principle scores and the percentage of herds that shifted to a lower or higher overall welfare category were quantified for each replacement by the theoretically and observed worst and best values.

For most measures, values that were altered were scored as either percentage of cows (e.g. % of severely lame cows) or 'yes' and 'no' (e.g. for cleanliness of drinkers). However, for some measures (avoidance distance at the feed rack (ADF), lameness and integument alterations) the aggregated measure indexes rather than individual percentages were replaced with worst and best scores. Because these measures together add up to 100% of animals, changing percentages within these could create an impossible situation (i.e. percentages would add up to over 100%). In addition, the theoretical best score for the measures 'length of drinking trough' and 'number of drinking bowls' depends on the average number of cows on the herd. Therefore, we replaced these with scores that would meet the requirements for all herds in the dataset (10,000 cm for drinking trough length and 100 for number of drinking bowls) as best scores. For the measures of dehorning and tail docking, we replaced the actual methods used at each herd with the methods which would generate the best (i.e. no dehorning, no tail docking respectively) and the worst score (i.e. dehorning using surgery with no anaesthetics or analgesics, tail docking using a rubber band without anaesthetics and analgesics, respectively).

Results

None of the 491 herds were originally (i.e. before replacement with worst/best scores) in the 'excellent' category, 174 (35%) were in the 'enhanced' category, 308 (63%) in the 'acceptable' category and nine (2%) in the 'not classified' category. For eight of the nine 'not classified' herds, classification was due to a 'good feeding' principle score below 5 (the threshold for the not-classified category). The median, minimum, and maximum scores are given at the measure (Table 2) and principle and criterion level (Table 4). For several measures, the observed range spanned the entire theoretical range (i.e. 0 – 100 for percentages, 0 – 24 for hours and 0 – 365 for days). However, for several other measures (18 out of 33), criteria (6 out of 12) and principles (3 out of 4), the observed data range was narrower than was theoretically possible (Tables 2 and 3). Only 5% of herds were not dehorned or disbudded, 18% were disbudded using caustic paste, 76% using thermocautery, and 1% was dehorned using surgery. Analgesics and/or anaesthetics were used during these procedures in 24% and 60% of the herds, respectively. Only 5 (ca. 1%) herds were tail-docked (3 by rubber ring and 2 by surgery). Analgesics were never used during tail docking whilst anaesthetics were used in two herds.

Sensitivity analysis using observed values: measurement level

Sensitivity of the overall welfare category

When separate measure values were increased to the observed maximum value (i.e. to the level of the herd that scored best for that specific measure) fewer herds shifted between overall categories than when separate scores were decreased to the observed minimum value (Table 2). Regarding the overall welfare categories between which the shifts occurred, for most measures, the highest percentage of shifts occurred between the 'enhanced' and 'acceptable' category (percentage of shifts ranging from 0 – 99%). However, for increases in some measures ('% of lean cows', 'number of water bowls', 'cleanliness of drinker' and 'loose versus tied housing') highest % of shifts to a higher category were between 'not classified' and 'acceptable' (percentage of shifts ranging from 22 - 100%).

Replacements of measure values only rarely led to negative shifts of more than one category and never to positive shifts of more than one category (Table 2). The effects of replacing a measure often differed greatly, even between measures that belong to the same principle.

'Good health' was the only principle for which changing the values of any of its underlying measures did not result in a substantial (>10%) effects on herd classification. All measures that were the only measure of a certain criterion caused a relatively high percentage of herds to shift category: '% of lean cows', 'loose or tied housing' and the 'QBA index' when replaced with the worst possible score, with the exception of the 'ADF index'. Although seemingly combined with many other measures, most measures of the criterion 'absence of prolonged thirst' had a relatively large influence as well. Most upgrades to a higher overall welfare category were achieved by increasing (to the observed maximum levels) 'number of water bowls', 'trough length', and to a lesser extent '% of cows colliding'. Within the two criteria that contained most measures, either sensitivity was very low for all measures ('absence of disease') or sensitivity was greater for those measures that were attributed the highest weight (i.e. within 'comfort around resting', the measures for resting behaviour are given a higher weight than cleanliness).

<Table 2>

Sensitivity of the principles and criteria scores

The sensitivity analysis of the effect of changes in separate measure values on the principles scores and on the criteria scores (Table 3) showed the same pattern as the sensitivity analysis of the overall welfare category. The decrease caused by changing a measure to the lowest observed value was usually greater than the increase caused by changing the same measure to its highest observed value. Exceptions to this trend often concerned measures of which the observed values were very poor. Furthermore, measures that caused the greatest difference tended to belong to criteria that contain few other measures. Exceptions to this trend once again concerned most measures within 'absence of prolonged thirst' and the measure '% of cows colliding with housing'. There was a difference in the sensitivity of the principles and the criteria in that measure values have a more direct influence on criteria scores, and therefore had a greater influence on criteria scores than on principle scores.

<Table 3>

Sensitivity analysis using observed values: criteria and principle level

Of all principles, alteration of 'good feeding' led to the highest number of negative shifts as well as positive shifts (Table 4). Moreover, replacing the 'good feeding' score to the lowest observed score in the database caused all 'enhanced' herds to be re-categorised as 'non-classified'. Alterations to the other principle scores never caused a change of more than one overall welfare category. Alteration of the 'good housing' principle caused the fewest positive shifts of all principles, as most farms already scored relatively high for this principle (median score of 54).

Of all criteria, replacement with the lowest observed score was most effective in generating negative shifts for 'absence of prolonged hunger' followed by 'absence of prolonged thirst'. Replacement with the highest observed score was most effective in generating a positive shift for 'absence of prolonged thirst'. Both criteria within the principle 'good housing' ('comfort around resting' and 'ease of movement') caused 27% of herds to be downgraded when replaced by the observed minimum. Effects of replacing criteria scores within the 'good health' and 'appropriate behaviour' principles varied considerably between criteria.

<Table 4>

Differences between replacement with observed and theoretically possible scores

For several measures, criteria and principles, the observed range did not span the entire theoretical range. For three measures ('lameness index', 'head butts/cow/15 min' and 'ADF index'), four criteria ('absence of injuries', 'absence of diseases', and 'absence of pain induced by management procedures') and three principles ('good housing', 'good health' and 'appropriate behaviour'), replacement with the theoretically possible scores instead of the observed scores resulted in a higher % of herds shifting between overall welfare categories (Table 5). For four measures ('% lean cows', 'lameness index', 'number of coughs/cow/15 min.', '% cows with hampered respiration' and 'ADF index), this resulted in a higher median increase or decrease of the principle and criteria scores than when worst or best observed scores were used (Table 6).

<Table 5>

<Table 6>

Discussion

This study investigated the sensitivity of the integrated scores of the WQ protocol for on-farm dairy cattle welfare assessment to extreme changes in individual measure, criterion and principle scores. The impact of one by one replacement of observed herd-level measure, criteria and principle scores by extremely low or high values had variable effects on the more highly integrated scores and on the overall welfare category. Investigation into what type of replacements have a large versus negligible impact suggests that a considerable part of this variation appears to be an unwanted side-effect of the complex step-wise integration method rather than being intentional or justifiable.

Sensitivity analysis using observed values: measurement level

Generally, the impact of a replacement with an extremely low score was bigger than replacement with an extremely high score. This reflects the intention of the WQ integration method to limit compensation of poor scores with better scores on other welfare aspects (Veissier et al 2011). The effect of replacing observed measure scores with extreme values on more highly integrated scores (criteria and principles) and on the overall welfare category was very variable and seemed to depend on various aspects. Replacements of the measures '% of lean cows', 'loose/tied housing', the 'QBA index', 'drinker trough length' and 'cleanliness of drinkers', had a bigger impact on overall classification compared to other measures (particularly when substituted by observed worst scores). The common feature shared by the first three measures is that they are the only measure of the criterion they belong to ('absence of prolonged hunger', 'ease of movement' and 'positive emotional state', respectively). One other criterion is also documented by a single measure, namely 'expression of other normal behaviour' measured with the ADF-test. This measure had less impact compared with the aforementioned three measures, presumably because the ADF-index was already poor for most farms to begin with (so the change by replacing the actual score with the worst possible score was often very small) .

The relatively large impact of drinker space and cleanliness of drinkers is in accordance with previous findings for both the dairy cattle protocol (de Vries et al 2013; Heath et al 2014) and the WQ broiler chicken protocol (Buijs et al 2016). This seems to be caused by a combination of factors. First, these measures both belong to the criterion of 'absence of prolonged thirst' which contains few measures that matter for calculating the criterion scores (in the decision tree only number/length of drinkers and cleanliness are taken into account). The other measures are either prerequisites for the required number/length of drinkers and therefore less directly influence criterion scores ('water flow'), or are related to the number of drinkers ('at least 2 drinkers/cow'). Second, the principle 'good feeding' contains only one other criterion apart from 'absence of prolonged thirst', whereas most other principles are composed of more criteria. It could be argued that the large impact of these measures is not necessarily problematic if they are valid indicators of an important welfare problem. However, as resource-based measures, drinker space and cleanliness would appear to be potential risk factors rather than direct measures of thirst (Sprenger et al 2009; Vanderhasselt et al 2014). Moreover, to our knowledge, the validity of these measures of thirst has not yet been tested. Therefore, the finding that these measures have a relatively large influence on integrated scores can be considered problematic. Animal-based indicators of thirst have been developed, such as blood sodium concentrations, plasma osmolality (Reece, 2009; Vanderhasselt et al., 2013) and voluntary water consumption (in broiler chickens; Sprenger et al., 2009; Vanderhasselt et al., 2014). Whereas blood parameters are too invasive to perform in on-farm welfare monitoring, it could be promising to develop voluntary water consumption tests further. Identifying the most reliable, valid and feasible measure of prolonged thirst in dairy cattle should be a priority in future animal welfare assessment research.

Replacements of measures within the principle 'good health' with the best or worst scores had little influence on principle and criterion scores and on overall classification, in accordance with previous results (de Vries et al 2013; Heath et al 2014; Nielsen et al 2015). This is remarkable because it includes measures which indicate important welfare problems in dairy cattle according to many experts, such as mortality, mastitis and lameness (Nielsen et al 2014; Lievaart and Noordhuizen, 2011). In addition, Tuyttens et al (2010) reported that both consumers and farmers rank health aspects as the most important for farm animal welfare.
The very limited effect of extreme changes in measures within the criterion 'absence of diseases' on integrated WQ scores seems to be caused, at least partially, by the aggregation method of this criterion. In this aggregation, prevalence of symptoms of diseases is compared to warning and alarm thresholds (e.g. warning threshold for nasal discharge is 5% of cows and

alarm threshold 10% of cows). Subsequently, a weighted sum is calculated of warnings and alarms, with a weight of 1 for warnings and 3 for alarms, which is computed into the criterion score using a spline function. Because of this method, increasing prevalences that were already above the alarm threshold (or decreasing those that were already below the threshold) will not affect classification at all. Also, when the prevalence of one disease symptom changes, it has only a limited effect on the criterion scores because it is aggregated with many other disease symptoms.

Similarly to measures within ‘absence of diseases’, measures within ‘absence of injuries’ also had a small impact on the integrated scores. However, a different method is used to integrate the measures within ‘absence of injuries’ to one score. Partial scores for lameness and integument alterations are first calculated using weighted sums and i-spline curves, and are then combined using a Choquet integral. The lameness index had most influence, but still caused only 10% of herds to be downgraded when replaced with the theoretically worst possible score (i.e. 100% severely lame cows). This surprisingly low impact seems to be due to the large number of criteria within the principle ‘good health’, and to the observation that herds often score relatively low for these criteria. Therefore, changing another score within this principle to a low score is likely to have a smaller effect than when it is done for a score in another principle with fewer criteria such as ‘good feeding’. Due to the limited impact of good health measures on overall welfare categorisation, in theory a situation could occur where farms categorised as ‘acceptable’ or better have 100% severely lame animals, while this may obviously be considered a major welfare problem.

Regarding positive shifts, the percentage of cows colliding with housing had a relatively large positive impact when replaced with best observed score. This is likely because a large proportion of farms (55%) were classified as having a serious problem for this measure to begin with, so for many farms a vast improvement was possible (compared to 37% for % of cows laying out and 28% which were above the threshold value of 6.3 seconds for mean time needed to lie down).

Sensitivity analysis using observed values: criteria and principle level

There are two, three, or four criteria per principle. This difference in the number of criteria is reflected in the results of the sensitivity analysis: replacement with the worst criteria scores within the principle ('good feeding') containing only two criteria ('absence of prolonged hunger' and 'absence of prolonged thirst') generated most shifts towards a different welfare category. The principle 'good housing' also consists of only two criteria for which measures have been developed (for its third criterion 'thermal comfort' no measure is available). The impact of both criteria are smaller compared to the two criteria of 'good feeding'. However, even though for 'thermal comfort' no data are collected, the missing criterion score is replaced with the best score among 'comfort around resting' and 'ease of movement'. This dilutes the effect of a very low score on either of these two criteria. Although some validated measures for thermal comfort exist for dairy cattle (e.g. respiration rate, Schutz et al., 2010), inclusion of such measures may complicate timing of farm visits, as the outcomes of these measures are highly influenced by ambient temperature and humidity. Therefore, climatic conditions should be similar during farm visits to capture farm-level differences in thermal comfort rather than differences based on ambient weather conditions. Further research on how to deal with these complexities in the WQ protocol is necessary, or removal of 'thermal comfort' as a criterion for dairy cattle welfare should be considered.

In line with the criteria, of all principles, alteration of ‘good feeding’ led to the most negative and positive shifts when replaced with observed worst and best scores. For negative shifts this was because ‘good feeding’ was the only principle for which scores <5 were observed, which automatically categorizes a herd as ‘not classified’. For positive shifts, this was because this principle caused more ‘not classified’ and ‘acceptable’ categorizations than any other principle (as 131 farms originally had a score between 5 and 15 for this principle, as opposed to 9 for housing, 3 for health and 23 for behaviour). Therefore, more positive shifts could occur when ‘good feeding’ was altered than when the other principles were replaced with observed maximum scores.

Differences between replacement with observed and theoretically possible scores

As the sample size in the current study was large and contained a wide variety of herds (given the different sampling aims), we can draw some conclusions about the observed scores in relation to theoretical possible scores. For most measures, observed scores spanned the entire theoretical range. This means that for the dairy cattle protocol, most limits set by WQ seem realistically attainable. For some measures however, observed scores were less extreme than the theoretically possible scores. In most cases, this did not affect criterion scores as these were within the criterion ‘absence of diseases’, where warning and alarm thresholds are used to integrate scores. For lameness index and ADF index however, fewer shifts of the overall welfare category were observed when replaced with the observed scores. This was also reflected in the corresponding criteria and principle scores, of which the worst possible score never occurred. This is one of the reasons that the principles ‘good health’ and ‘appropriate behaviour’ never caused herds to be categorized as ‘not classified’ when replaced by the observed minimum score.

Conclusion

The results of the current study provide insight into the functioning of the integration methods for the dairy cattle WQ protocol. Findings indicate that the sensitivity of integrated scores to replacement of individual scores by extreme scores is dependent on a number of factors which were intended by the WQ protocol: 1) the observed value of the specific measure (or criterion), relative to the values of the other measure in the same criterion (or principle); 2) whether the values were replaced by an extremely low or an extremely high value (more impact of the former); 3) the relative weight WQ attributes to the measures. However, two other factors that were not intended and appear to be unwanted side-effects of the complexity of the step-wise integration method also had considerable influence. These factors were: 1) the number of measures that are integrated into criteria and principle scores; and 2) the aggregation method of the measures (e.g. decision trees or weighted sums). The effect of both integration method and grouping is problematic, as it should be the severity of the welfare problem that affects the overall category. As a result, sensitivity is highest for changes in measures of the 'good feeding' principle, of which a large proportion of the measures are criticized for their validity (i.e. measures of 'absence of prolonged thirst'). On the contrary, measures within the principle 'good health' have the lowest impact while some of these measures are considered to most severely affect dairy cattle welfare. For instance, a farm in the 'acceptable' category or higher could theoretically have 100% severely lame animals. The unwanted side-effects of the current WQ integration methods shown in this study warrant research to develop and evaluate alternative integration methods.

Animal welfare implications

This study indicates that the WQ integration method does not adequately balance the relative importance of all welfare measures that are included in order to adhere to the multidimensional nature of animal welfare. Therefore, using the current integrated WQ scores could lead to a focus on a limited set of (often resource-based) measures which is hard to justify. As this harms the credibility of the assessment protocol, we recommend a revision of the integration method, so that the relative contribution of the various welfare measures to the integrated scores more correctly reflects their relevance for dairy cattle welfare.

References

Bartussek H, Leeb C and Held S 2000 Animal Needs Index for Cattle. ANI 35, L/2000. http://www.bartussek.at/veroeffentlichungen/511134991b0db8204/index.html

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

Blokhuis HJ, Veissier I, Miele M and Jones B 2010 The Welfare Quality® project and beyond: Safeguarding herd animal well-being. *Acta Agriculturae Scand* 60: 129-140

Buijs S, Ampe B and Tuyttens FAM 2016 Sensitivity of the Welfare Quality® Broiler chicken protocol to differences between intensively reared indoor flocks: which factors explain overall classification? *Animal* 15: 1-10

Botreau R, Veissier I, Butterworth A, Bracke MBM and Keeling LJ 2007 Definition of criteria for overall assessment of animal welfare. *Animal Welfare* 16: 225-228

Botreau R, Capdeville J, Perny P, and Veissier I 2008 Multicriteria evaluation of animal welfare at farm level: An application of MCDA methodologies. *Foundation of Computing and Decision Sciences* 33: 1–18

de Graaf S, Ampe B and Tuyttens FAM in press Assessing dairy cow welfare at the beginning and end of the indoor period using the Welfare Quality® protocol. *Animal Welfare*.

de Vries M, Bokkers EAM, van Schaik G, Botreau Rl, Engel B, Dijkstra T and de Boer IJM 2013 Evaluating results of the Welfare Quality multi-criteria evaluation model for classification of dairy cattle welfare at the herd level. *Journal of dairy science* 96: 6264-6273

Herd Animal Welfare Council 1992 FAWC updates the five freedoms. *Veterinary Record* 17: 357

Heath CAE, Browne, WJ, Mullan S, Main DCJ 2014 Navigating the iceberg: reducing the number of parameters within the Welfare Quality®assessment protocol for dairy cows. *Animal* 8: 1978-1986

Lievaart JJ and Noordhuizen JPTM 2011 Ranking experts' preferences regarding measures and methods of assessment of welfare in dairy herds using Adaptive Conjoint Analysis. *Journal of dairy science* 94: 3420-3427.

Nielsen BH, Angelucci A, Scalvenzi A, Forkman B, Fusi F, Tuyttens F, Houe H, Blokhuis H, Sørensen JT, Rothmann J , Matthews L, Mounier L, Bertocchi L, Richard M, Donati M, Nielsen PP, Salini R, de Graaf S , Hild S, Messori S, Nielsen SS, Lorenzi V, Boivin X and Thomsen PT 2014 Use of animal based measures for the assessment of dairy cow welfare-ANIBAM. *EFSA External scientific report.*

Reece WO 2009. Functional anatomy and physiology of domestic animals. John Wiley & Sons. Iowa, USA.

Schütz KE, Rogers AR, Poulouin YA, Cox NR, and Tucker CB 2010 The amount of shade influences the behavior and physiology of dairy cattle. *Journal of dairy science* 93: 125–133.

Sprenger M, Vangestel C and Tuyttens FAM 2009 Measuring thirst in broiler chickens. *Animal Welfare* 18: 553-560

Tuyttens FAM, Vanhonacker F, Van Poucke E and Verbeke W 2010 Quantitative verification of the correspondence between the Welfare Quality® operational definition of herd animal welfare and the opinion of Flemish herders, citizens and vegetarians. *Livestock Science* 131: 108-114

Vanderhasselt RF, Buijs S, Sprenger M, Goethals K, Willemsen H, Duchateau L and Tuyttens FAM 2013 Dehydration indicators for broiler chickens at slaughter. *Poultry Science* 92: 612-619.

Vanderhasselt RF, Goethals K, Buijs S, Federici JF, Sans ECO, Molento CFM, Duchateau L and Tuyttens, FAM 2014 Performance of an animal-based test of thirst in commercial broiler chicken herds. *Poultry science* 93:1327-1336.

Veissier I, Jensen KK, Botrea R and Sandøe P 2011 Highlighting ethical decisions underlying the scoring of animal welfare in the Welfare Quality® scheme. *Animal Welfare* 20:89-101

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

Whay HR, Main DCJ, Webster AJF and Green LE 2003 Assessment of the welfare of dairy cattle using animal based measurements: direct observations and investigation of herd records. *The Veterinary Record* 153: 197-202

Appendix

Percentages of herds[1] (n = 491) that were downgraded and upgraded by 1 or 2 overall welfare categories when individual values at measure level within the criterion 'absence of diseases' were replaced with theoretical worst and best values per measure

Measures	Observed worst score	Observed best score
	% downgraded 1 category	% upgraded 1 category
Number of coughs/cow/minute	2	0
% cows with nasal discharge	2	0
% cows with ocular discharge	2	0
% cows with hampered respiration	1	0
% cows with diarrhoea	2	0
% cows with vulvar discharge	2	0
% cows with SCC >400.000	2	1
% cows mortality	2	1
% calvings with dystocia	1	0
% downer cows	1	1

[1]Percentages were based on the herds that were actually able to shift one or two categories. For downgrades of 1 category n = 482, for downgrades of 2 categories n = 174. For upgrades of 1 category n = 491.

Table 1: All principles, the corresponding criteria and indicators used in the Welfare Quality® assessment protocol for dairy cattle welfare

Principles	Criteria	Measures	Aggregation method measures
Good feeding	Absence of prolonged hunger	Body Condition Score (% very lean animals)	Spline curve fitting
	Absence of prolonged thirst	Availability & cleanliness water	Decision tree
Good housing	Comfort around resting	Lying down duration; collisions during lying down ; on edge/outside of lying area; cleanliness	Converted to ordinal scores, combined in weighted sums and spline curve fitting
	Thermal comfort	No measure for dairy cattle	
	Ease of movement	Free stalls or presence of tethering and exercise	Decision tree
Good health	Absence of injuries	Lameness; integument alterations	Combined in weighted sums, spline curve fitting and Choquet integration
	Absence of disease	Respiration/digestive diseases; mastitis; mortality; dystocia, downer cows	Converted to ordinal scores, combined in weighted sums and spline curve fitting
	Absence of pain induced by management procedures	Mutilations (dehorning; tail docking; use of anaesthetics/analgesics)	Decision tree
Appropriate behaviour	Expression of social behaviour	Incidence agonistic interactions	Combined in weighted sums and spline curve fitting
	Expression of other behaviours	Access to pasture	Spline curve fitting
	Good human-animal relationship	Avoidance distance at feeding place	Combined in weighted sums and spline curve fitting
	Positive emotional state	Qualitative Behavioural Assessment	Combined in weighted sums and spline curve fitting

Table 2: Percentages of herds[1] (n = 491) that were downgraded or upgraded 1 or 2 overall welfare categories when individual values at measure level (continuous and binary) were replaced with observed worst and best values per measure

Principles	Criteria, Continuous measures	Observed median, min - max	Observed worst score		Observed best score
			% downgraded 1 category	% downgraded 2 categories	% upgraded 1 category
Good feeding	Absence of prolonged hunger				
	% of lean cows[2]	4, 0 – 88	53	0	5
Good housing	Comfort around resting				
	Mean time needed to lie down (s)	6, 3 – 20	10	0	6
	% of cows colliding with housing	33, 0 – 100	5	0	12
	% of cows lying outside of lying area	0, 0 – 73	11	0	8
	% of cows with dirty flanks	64, 0 – 100	0	0	7
	% of cows with dirty lower legs	80, 0 – 100	2	0	7
	% cows with a dirty udder	37, 0 – 100	2	0	7
Good health	Absence of injuries				
	Lameness index	88, 37 – 100	6	0	5
	Integument alterations index	53, 0 – 100	2	0	4
	Absence of diseases				
	Range of all disease-measures[2]	-	1-2	0	0-1
Appropriate behaviour	Expression of social behaviour				
	Head butts/cow/15 min.	0.5, 0 – 7	13	0	1
	Displacements/cow/15 min.	0.4, 0 – 5	16	0	4
	Expression of other normal behaviour				
	Number of hours on pasture)	7.5, 0 - 24	9	0	1
	Number of days on pasture	175, 0 - 365	9	0	1
	Human-animal interaction				
	ADF index	67, 23 – 100	13	0	6
	Positive emotional state				
	QBA index	0.3, -11 – 5	24	1	7
	Criteria, Binary measures	**% farms with best score**			
Good feeding	Absence of prolonged thirst				
	Water flow	82	22	3	3
	Trough length	18	26	1	19
	Number of water bowls		11	1	20
	Drinker cleanliness	76	23	0	8
	At least 2 drinkers/cow	84	9	0	1
Good	Ease of movement				

housing	Loose or tied housing	93	38	2	3
Good health	Absence of pain induced by management procedures				
	Dehorning method	5	9	0	3
	Tail docking method	95	8	0	0

[1]Percentages were based on the herds that were actually able to shift one or two categories. For downgrades of 1 category n = 482, for downgrades of 2 categories n = 174. For upgrades of 1 category n = 491.

[2]As absence of disease contains a very high number of measures with a very small range of shifts, we present only the range here. All separate measures can be found in the Appendix.

Table 3: Median (min – max) decrease and increase in principle and criteria scores when measure scores were replaced with worst and best observed measure scores

Principles / Criteria	Measures	Changes in principles scores		Changes in criteria scores	
		Median decrease in worst scenario	Median increase in best scenario	Median decrease in worst scenario	Median increase in best scenario
Good feeding					
Absence of prolonged hunger	% lean cows	24 (0 – 71)	5 (0 – 69)	67 (0 - 98)	30 (0 - 98)
Absence of prolonged thirst	Water flow	11 (0 – 85)	0 (0 – 85)	29 (0 - 97)	0 (0 - 0)
	Trough length	25 (0 – 85)	0 (0 – 85)	29 (0 - 97)	0 (0 - 97)
	Number of water bowls	0 (0 – 85)	10 (0 – 85)	0 (0 - 97)	12 (0 - 97)
	Drinker cleanliness	12 (0 – 60)	0 (0 – 60)	40 (0 - 68)	0 (0 - 68)
	At least 2 drinkers per animal	0 (0 – 35)	0 (0 –35)	20 (0 – 97)	0 (0 - 40)
Good housing					
Comfort around resting	Mean time to lie down	6 (0 – 20)	5 (0 – 20)	10 (0 – 32)	8 (0 - 31)
	% cows colliding with housing	0 (0 – 19)	11 (0 – 17)	0 (0 - 32)	18 (0 - 27)
	% cows lying outside of lying area	10 (0 – 20)	0 (0 – 29)	16 (0 – 32)	0 (0 – 46)
	% cows with dirty flanks	0 (0 – 5)	4 (0 – 14)	0 (0 – 12)	6 (0 – 22)
	% cows with dirty lower legs	0 (0 – 9)	4 (0 – 12)	0 (0 – 15)	6 (0 – 18)
	% cows with a dirty udder	0 (0 - 9)	4 (0 – 8)	0 (0 – 15)	6 (0 – 18)
Ease of movement	Loose or tied housing	24 (0 – 37)	0 (0 – 40)	66 (0 – 66)	0 (0 – 85)
Good health					
Absence of injuries	Lameness index	13 (0 – 37)	5 (0 – 35)	27 (3 – 69)	33 (0 – 57)
	Integument alteration index	4 (0 – 24)	5 (0 – 26)	10 (0 – 44)	26 (0 – 42)
Absence of disease	Number of coughs/cow/minute	0 (0 – 0)	0 (0 – 0)	0 (0 – 0)	0 (0 – 0)
	% cows with nasal discharge	1 (0 – 12)	0 (0 – 10)	8 (0 – 35)	0 (0 – 21)
	% cows with ocular discharge	1 (0 – 12)	0 (0 – 8)	8 (0 – 35)	0 (0 – 35)
	% cows with hampered respiration	1 (0 – 5)	0 (0 – 1)	4 (0 – 14)	0 (0 – 14)

		% cows with diarrhoea	2 (0 – 12)	0 (0 – 10)	9 (0 – 35)	0 (0 – 35)
		% cows with vulvar discharge	3 (0 – 12)	0 (0 – 7)	10 (0 – 35)	0 (0 – 24)
		% cows with SCC >400.000	2 (0 – 12)	1 (0 – 12)	8 (0 – 35)	4 (0 – 35)
		% cows mortality	2 (0 – 11)	0 (0 – 12)	8 (0 – 35)	0 (0 – 35)
		% calvings with dystocia	1 (0 – 12)	0 (0 – 13)	7 (0 – 35)	4 (0 – 35)
		% downer cows	2 (0 – 12)	1 (0 – 13)	0 (0 – 35)	3 (0 – 35)
	Absence of pain induced by management procedures	Dehorning method (none, surgery)	15 (0 – 35)	6 (0 – 40)	50 (0 – 89)	48 (0 – 98)
		Tail docking method (none, ring)	14 (0 – 34)	0 (0 – 6)	6 (0 – 89)	0 (0 – 0)
Appropriate behaviour						
	Expression of social behaviour	Head butts/cow/15 min.	13 (0 – 37)	1 (0 – 16)	69 (0 – 100)	8 (0 – 49)
		Displacements/cow/15 min.	16 (0 – 44)	2 (0 – 30)	69 (0 – 100)	19 (0 – 93)
	Expression of other behaviour	Number of hours on pasture	15 (0 – 38)	0 (0 – 34)	64 (1 – 100)	0 (0 – 85)
		Number of days on pasture	15 (0 – 38)	1 (0 – 24)	64 (1 – 100)	15 (0 – 86)
	Good human-animal relationship	ADF index	10 (0 – 37)	9 (0 – 37)	31 (0 – 87)	56 (0 – 87)
	Positive emotional state	QBA index	20 (0 – 50)	7 (0 – 44)	52 (0 – 93)	40 (0 – 93)

Table 4: Percentages of herds[1] (n = 491) that shifted into a different overall welfare category when individual scores were replaced with observed worst and best criteria or principle scores (observed median, min. and max. score given in column b)

Principles, Criteria	Original observed median, min - max	Observed worst score % farms downgraded 1 category	 % farms downgraded 2 categories	Observed best score % farms upgraded 1 category	 % farms upgraded 2 categories
Good feeding	40, 4 – 100	64	100	36	1
Absence of prolonged hunger	70, 3 – 100	59	0	6	0
Absence of prolonged thirst	60, 3 – 100	35	3	30	1
Good housing	54, 6 – 86	37	0	13	0
Comfort around resting	27, 0 – 80	27	0	13	0
Ease of movement	100, 15 – 100	27	0	0	0
Good health	34, 8 – 86	37	0	23	0
Absence of injuries	35, 4 – 100	21	0	8	0
Absence of diseases	40, 12 – 100	4	0	7	0
Absence of pain induced by management procedures	52, 2 – 100	9	0	3	0
Appropriate behaviour	35, 6 – 86	37	0	25	0
Expression of social behaviour	69, 0 – 100	16	0	5	0
Expression of other normal behaviour	64, 0 -100	9	0	8	0
Good human-animal relationship	44, 13 – 100	14	0	8	0
Positive emotional state	53, 0 – 93	24	1	7	0

[1]Percentages were based on the herds that were actually able to shift one or two categories. For downgrades of 1 category n = 482, for downgrades of 2 categories n = 174. For upgrades 1 category n = 491, for upgrades of 2 categories n = 317.

Table 5: Percentages of herds[1] (n = 491) that shifted into a different overall welfare category when scores at the measure, criterion, and principle level[2] were replaced with theoretically possible[1] worst and best scores

	Worst score		Best score
	% downgraded 1 category	% downgraded 2 categories	% upgraded 1 category
Measures[1]			
Lameness index[3]	10	0	5
Head butts/cow/15 min.[3]	16	0	1
ADF index[3]	20	0	6
Criteria[1]			
Absence of injuries[4]	29	1	8
Absence of diseases[4]	36	1	7
Absence of pain induced by management procedures[4]	12	0	3
Good human-animal relationship[4]	23	0	8
Principles[1]			
Good housing[4]	64	100	13
Good health[4]	64	100	23
Appropriate behaviour[4]	64	100	25

[1]Percentages were based on the herds that were actually able to shift one or two categories. For downgrades of 1 category n = 482, for downgrades of 2 categories n = 174. For upgrades of 1 category n = 491.

[2]Scores shown are of those measures, criteria and principles where replacement with theoretical score generated different results than when replaced with observed score.

[3] theoretical possible worst score was 100, theoretical best score was 0

[4] theoretical possible worst score was 0, theoretical best score was 100

Table 6: Median (min – max) decrease and increase in principle and criterion scores when measures were replaced with worst and best theoretically possible values

Principles, criteria	Measures	Change in principle scores		Change in criteria scores	
		Median decrease in worst scenario	Median increase in best scenario	Median decrease in worst scenario	Median increase in best scenario
Good feeding[1]					
Absence of prolonged hunger	% lean cows[2]	25 (2 – 73)	5 (0 – 69)	69 (2 – 100)	30 (0 – 98)
Good health[1]					
Absence of injuries	Lameness index [3]	15 (2 – 39)	5 (0 – 35)	27 (3 – 69)	33 (0 – 57)
Absence of disease	Number of coughs/cow/15 min.[2]	4 (0 – 12)	0 (0 – 0)	10 (5 – 35)	0 (0 – 0)
	% cows with hampered respiration[2]	4 (1 – 12)	0 (0 – 1)	10 (6 – 35)	0 (0 – 14)
Appropriate behaviour[1]					
Good human-animal relationship	ADF index[2]	46 (11 – 82)	9 (0 – 37)	44 (13 – 100)	55 (0 – 87)

[1] Scores shown are of those where replacement with theoretical score generated different results than when replaced with observed score

[2] theoretical possible worst score was 100, theoretical best score was 0

[3] theoretical possible worst score was 0, theoretical best score was 100