Animal 17 (2023) 101018

Contents lists available at ScienceDirect

Animal
The international journal of animal biosciences

Improvement of the Welfare Quality scoring model for dairy cows to fit experts' opinion

R. Lardy [a,*], R. Botreau [a], A. de Boyer des Roches [a], F.J.C.M. van Eerdenburg [b], S. de Graaf [c], M.J. Haskell [d], M.K. Kirchner [e], L. Mounier [a], M Kjosevski [f], F.A.M. Tuyttens [c,g], I. Veissier [a]

[a] Université Clermont Auvergne, INRAE, VetAgro Sup, UMR Herbivores, 63122 Saint-Genès-Champanelle, France
[b] Department of Farm Animal Health, Faculty of Veterinary Medicine, Utrecht University, 3584 CL Utrecht, The Netherlands
[c] Flanders Research Institute for Agriculture, Fisheries and Food (ILVO), Burgemeester van Gansberghelaan 92, 9820 Merelbeke, Belgium
[d] SRUC, West Mains Road, Edinburgh EH9 3JG, UK
[e] FOUR PAWS International, Linke Wienzeile 236, 1150 Vienna, Austria
[f] Animal Welfare Center, Faculty of Veterinary Medicine, Ss Cyril and Methodius University in Skopje, Lazar Pop-Trajkov 5-7, 1000 Skopje, Republic of Macedonia
[g] Department of Veterinary and Biosciences, Faculty of Veterinary Medicine, Ghent University, Heidestraat 19, 9820 Merelbeke, Belgium

ARTICLE INFO

Article history:
Received 5 July 2023
Revised 9 October 2023
Accepted 10 October 2023
Available online 17 October 2023

Keywords:
Animal well-being
Cattle
Expert opinion
Multicriteria evaluation
Sensitivity analysis

ABSTRACT

After several years of implementation, the original Welfare Quality scoring model for dairy cows appears to be highly sensitive to the number and cleanliness of drinkers and not enough to the prevalence of diseases, and as a consequence may not fit the opinion of some animal welfare experts. The present paper aims to improve the Welfare Quality calculations for the criteria 'Absence of prolonged thirst' and 'Absence of disease' in dairy cows, so that the results are more sensitive to input data and better fit experts' opinion. First, we modified the calculation of 'Absence of prolonged thirst' by linearising the calculation for drinkers' availability to avoid threshold effects. Second, we modified the calculation of 'Absence of disease' by applying a Choquet integral on the three lowest spline-based scores for each health disorder to limit compensation between health disorders. Third, we performed a global sensitivity analysis of the original and the alternative scoring models. Fourth, we compared the results obtained with the original and the alternative models with eight experts' opinions on two subsets composed of 44 and 60 farms, respectively, inspected using the Welfare Quality protocol and on which experts gave their opinion on the overall level of animal welfare. Results show that the alternative model significantly reduced the 'threshold effects' related to the number of drinkers and the compensation between health disorders. On the first subset, the alternative model fits the experts' opinion slightly better than the original model ($P = 0.061$). On the second subset, the models performed equally. In conclusion, the proposed refinements for calculating scores are validated since they significantly reduced 'threshold effects' and the influence of measures related to drinkers. It also reduced the compensation between health disorders by considering only the three lowest scores and thus increasing the influence of measures related to health disorders, and slightly improve at overall score level the accordance with experts' opinion.

Implications

We propose to modify the Welfare Quality scoring model to increase its sensitivity and its convergence with experts' opinion. We propose a linearised version for the interpretation of 'drinkers' availability. This significantly reduces the threshold effects initially applied on the number of drinkers. We propose to calculate the 'Absence of disease' by combining the three lowest scores obtained across health disorders using a Choquet integral. This largely reduces compensation between health disorders. The proposed modifications will be implemented in an updated version of the Welfare Quality protocol.

Introduction

The welfare of animals is considered to involve multiple dimensions (e.g. Fraser, 2003). Therefore, the assessment of welfare, as a whole, requires a multicriteria approach. Animal welfare assessment is, by nature, value-based (Veissier et al., 2011). It implies combining facts (e.g. measures) and interpretation of these facts, therefore combining objective and subjective information

* Corresponding author.
E-mail address: romain.lardy@inrae.fr (R. Lardy).

https://doi.org/10.1016/j.animal.2023.101018
1751-7311/

(e.g. interpretation and aggregation of measures) to get welfare scores (Spoolder et al., 2003). For dairy cattle, several protocols have been proposed to define welfare measures and aggregate their results into a score reflecting the overall welfare status of a herd (e.g. Bartussek, 1999; Burow et al., 2013; Welfare Quality®, 2009) or simplified versions of this protocol (Tuyttens et al., 2021; Stomp et al., 2023). Welfare Quality proposes a comprehensive scoring model based on measures whose relevance has been checked by concurrent, construct or consensus validity (Knierim et al., 2021) and on a scoring system fine-tuned according to experts' opinion (Welfare Quality®, 2009). The experts were animal scientists – for their knowledge on animals and on the meaning of measures –, social scientists – for their knowledge of how various societal groups value animal welfare – and stakeholders – for their knowledge of what can be done in practice – (Veissier et al., 2011). Since its original publication in 2009, Welfare Quality protocols have been extensively used, including for certifying farms in Spain and Finland. However, the original Welfare Quality scoring model received criticism. First, the results of the model for dairy cows are too sensitive to the number and cleanliness of drinkers (Heath et al., 2014; de Graaf et al., 2017, 2018; van Eerdenburg et al. 2021). Thereby, this measure is resource-based and does not necessarily reflect the level of thirst. Second, it is not sensitive enough to the prevalence of lameness or mastitis due to compensating mechanisms. For example, a farm where 50% of the cows are affected by mastitis, but no other disease is noticed on the farm would still receive a high score − 64.5 – for the 'absence of disease criterion', because the criterion is calculated on the proportion of alarming problems (here only one problem: mastitis) out of the eight potential disease problems. Therefore, the original scoring model does not correspond to the opinion of some dairy cattle welfare experts, nor does it encourage farmers to reduce such disorders (e.g. de Vries et al., 2013; Heath et al., 2014; van Eerdenburg et al., 2018). Doubts were also expressed on the reliability and the validity of the 'Qualitative Behaviour Assessment' (de Graaf et al., 2017). While work continues on refinement of the measures, the General Assembly of Welfare Quality Network decided to put efforts into improving the scoring model, namely the calculation of scores for two criteria 'Absence of prolonged thirst' and 'Absence of disease'.

Amendments have been suggested to improve the scoring of the provision of water or the health status. For criterion 'Absence of prolonged thirst', Van Eerdenburg et al. (2018) proposed to divide the number of drinkers by their average cleanliness to produce a score. For criterion 'Absence of diseases', de Vries et al. (2013) argued for a limited compensation between the results obtained across health disorders (nasal discharge, ocular discharge, hampered respiration, diarrhoea, vulvar discharge, milk somatic cell counts, mortality, dystocia, downer cows). Indeed, in the original Welfare Quality protocol, to calculate the score for Criterion 'Absence of disease', warning and alarm thresholds are defined for each health disorder and a weighted sum of alarms and warnings is calculated, resulting in the impact of each of the nine health disorders being diluted into a whole.

Sensitivity analyses of a model are essential to identify how variations in the inputs (here the results obtained for each welfare measure) influence the outputs (here the variation in the overall assessment of a farm) (see (Iooss and Lemaître, 2015) for a review). Due to a lack of time during the Welfare Quality project, no sensitivity analysis was performed. Two studies (de Vries et al., 2013; de Graaf et al., 2018) looked at how the model performed in scenarios in which the results from a farm were replaced by higher values (de Vries et al., 2013) or by the best or worst possible values (de Graaf et al., 2018). Their approaches had two main methodological biases: first, the increase applied to results varied between measures (e.g. the higher the initial value the smaller the shift to the best possible value) and second, interactions between measures and non-linear effects of measures (e.g. threshold effects) were not addressed. These biases, in turn, may lead to interpretation bias (Saltelli et al., 2006; Saltelli and Annoni, 2010). This paper presents a formal sensitivity analysis of the Welfare Quality scoring model in its original and alternative versions.

The present paper aims to improve the Welfare Quality calculations for the criteria 'Absence of prolonged thirst' and 'Absence of disease' in dairy cows, so that the results are more sensitive to input data and better fit experts' opinion. We propose new calculations for these two criteria. To check the benefit of the new calculations proposed, we perform a global sensitivity analysis, using the Morris method (which avoids the above-mentioned biases (Saltelli and Annoni, 2010)) on the original Welfare Quality scoring model and the alternative model that include the new calculations. We then compare the results of the two models to experts' opinion.

Material and methods

Welfare Quality scoring model for dairy cattle

The Welfare Quality protocol for assessing the welfare of dairy cattle on-farm can be found at https://www.welfarequalitynetwork.net/. It includes 49 measures taken on animals or their environment, grouped into 11 criteria then four principles before an overall assessment is produced. The scoring model that builds from scores on individual measures to an overall assessment comprises three steps briefly described here.

Step 1: From measures to criterion scores

Aggregation starts by combining 49 measures (Supplementary Table S1) into 11 criterion scores expressed on a 0–100 scale, with 100 as the best score. Several aggregation methods are used depending on the measures included in a criterion. For 'Absence of prolonged thirst', the five measures (the total length of water troughs; the number of water bowls; the number of water troughs; the cleanliness of water points and the water flow) are aggregated by the use of a decision tree. At each node of the tree, a decision is taken based on a Yes/No answer to a specific question (e.g. are the drinkers clean (drinkers with fresh feed residuals are not counted as dirty)? Are drinkers in sufficient number?). The decision tree finally defines seven possible situations, all assigned a score.

For 'Absence of disease', the percentage of cows affected by each of ten health problems are converted into three classes: 'below warning threshold', 'above warning threshold and below alarm threshold' or 'above alarm threshold', with warning threshold being half of the alarm threshold (for example, warning and alarm threshold for nasal discharge are 5 and 10%). The number of warnings and alarms is then combined into a weighted sum (with more weight attributed to alarms) which is in turn translated into a score by the use of a spline function.

Step 2: From criterion scores to principle scores

Criterion scores are aggregated into principle scores expressed on the same 0–100 scales as for criteria. For instance, Principle 'good feeding' embraces 'Absence of hunger' and 'Absence of prolonged thirst' and Principle 'Good health' embraces 'Absence of injuries', 'Absence of disease' and ''Absence of pain due to management procedures'. Choquet integrals are used for this aggregation, which allows to limit the possible compensation of poor scores by good ones while considering Criterion 'Absence of disease' is more important than Criterion 'Absence of injuries' that is more important than Criterion 'Absence of pain due to management procedures'.

Step 3: From principle scores to overall welfare category

The final aggregation is from principle scores to the overall welfare category. The welfare is considered 'excellent' when the farm scores ≥50 for each principle and ≥75 on two of them. When the farm scores ≥15 on each principle and ≥50 on at least two of them, it is classified as 'enhanced'. 'Acceptable' farms score ≥5 for all principles and ≥15 for at least three principles. The remaining farms are 'not classified'.

We used the INRAE 'Welfare Assessment of Farm Animals' webtool (https://www1.clermont.inrae.fr/wq/) to calculate scores and to assign farms to welfare categories as defined in the original Welfare Quality protocol. We used a modified version of the webtool to implement new calculations taking into account the proposed improvements in the scoring model.

Proposed improvements of the Welfare Quality scoring model

The Welfare Quality Network (https://www.welfarequality.net/) discussed alternative ways to calculate 'Absence of prolonged thirst' and 'Absence of disease' so that the scores obtained better match with experts' opinion. These alternatives are tested in the present paper.

Absence of prolonged thirst

A linearised version for the interpretation of 'drinkers' availability' weighted by a 'cleanliness' score is proposed to avoid threshold effects. To do so:

1. We calculate the total number of water bowls, and we convert it into trough length (1 bowl = 60 cm of trough). We then calculate the cumulated length of water troughs. Finally, we add the cumulated length of water troughs to the cumulated length of the bowls (previously transformed into trough length). As in the original model, if a drinker is not functioning properly or the water flow is insufficient (i.e. lower than 20 L/min for a trough or lower than 10 L/min for a bowl), then its length is divided by two.
2. We then divide the total drinkers' length by the number of cows.
3. We calculate a 'drinker availability' score based on the length of trough per cow according to one linear equation (if cows have access to only one drinker) and to a two-piecewise linear equation (if cows have access to at least two drinkers) (Fig. 1). The following equations were used:

$$Score = \begin{cases} \min\left(10 * cm_{OfTroughsPerCow}; 60\right) & \text{if less than 2 drinkers per cow} \\ 15 * cm_{OfTroughsPerCow} & \text{if at least 2 drinkers and } cm_{OfTroughsPerCow} < 4 \\ \min\left(60 + \left(cm_{OfTroughPerCow} - 4\right) * 20; 100\right) & \text{else} \end{cases}$$

Note that in the particular case of tied cows, when there is 1 bowl for 2 cows, each cow has access in theory to half a bowl, corresponding 30 cm equivalent trough length (60 cm divided by two). In the worst case (insufficient water flow and only one bowl for two cows), the average drinkers' length per cow is 15 cm (30 cm divided by two), which remains above the recommendation of 6 cm per cow (cf. equations and Fig. 1). This results in a drinkers' availability score of 60, which is the best score when cows have access to only one drinker.

4. We calculate a 'drinker dirtiness' score as the average dirtiness of drinkers (a clean drinker scored 1, a partially dirty 2, and a dirty one 3)
5. The score for 'Absence of prolonged thirst' is then the 'drinkers' availability' score divided by the 'drinker dirtiness' score.

Absence of disease

For each health disorder, we asked nine animal welfare scientists (seven of them authors of de Graaf et al. 2017, some of them being also Veterinarians by training, and all of them being experts of the model) to give us their expert scores (on a [0–100] scale as in Welfare Quality) for four prevalences: the alarm threshold defined in the original Welfare Quality scoring model and ¾, ½ and ¼ of this threshold (e.g., for nasal discharge, we asked experts to attribute a score on the 0–100 scale to 10, 7.5, 5 and 2.5% of cows affected, 10% corresponding to the alarm threshold). In addition, we asked them the lowest prevalence to which they would attribute a score of 0. We regressed an I-spline curve to model the experts' score according to the prevalence of each health disorder, I-spline curves being used in Welfare Quality to account for non-linearity between prevalence and experts' score (Welfare Quality®, 2009). Spline calculations were performed with R 3.6 (R Core Team, 2019) with the help of the 'spline2' package (Wang and Yan, 2020), so as to minimise the sum of square errors between scores given by experts and the calculated ones. As in Welfare Quality (Welfare Quality®, 2009), splines were interpolated by a piecewise polynomial of degree 3, in order to be easily manipulated.

We chose to keep the three lowest scores obtained (from the I-spline curves) across health disorders. We then aggregate them with a Choquet integral to produce the score for 'Absence of disease':

$$S_a + (S_b - S_a)\mu_{bc} + (S_c - S_b)\mu_c$$

where S_a, S_b, S_c are the three lowest scores, sorted such as $S_a \leq S_b \leq S_c$. We use 0.3 for μ_c and 0.155 for μ_{bc}; these values corresponding to averaged values that were used to calculate the 'Good health' principle score for dairy and beef cattle in the Welfare Quality protocol.

Dataset

The dataset from de Graaf et al. (2017 and 2018) was used for the sensitivity analysis. It contains data from 491 dairy cattle farms (460 with loose-housing and 31 with tied stalls) assessed using the Welfare Quality protocol. The farms originate from 10 European countries: Belgium (140 farms), France (128), The Netherlands (60), Austria (63), Denmark (42), Scotland (16), Macedonia (12), Romania (10), Northern Ireland (10), and Spain (10). This dataset is considered to reflect the current range of variation across European farms.

Expert opinion had been previously collected on two subsets of this dataset. We used these subsets to check the consistency of the models' results with experts' opinion:

Subset 1

From the 491 above-mentioned dairy cattle farms, data from 44 farms (25 loose-housing and 19 tied stalls; 20 from Denmark and 24 from Austria) were assessed within the Welfare Quality project, by four animal scientists from the project team who attributed an overall welfare score to each farm on a visual analogue scale of 120 mm (thus leading to a score from 0 to 120).

Subset 2

From the 491 above-mentioned dairy cattle farms, data from 60 dairy cattle loose-housing farms from The Netherlands were used. Veterinary practitioners expressed their opinion on the overall farm welfare on a 3-point scale: 1 – weak welfare, 2 – sufficient welfare or 3 – good welfare. The veterinarians came from four large veterinary practices spread out over The Netherlands. The

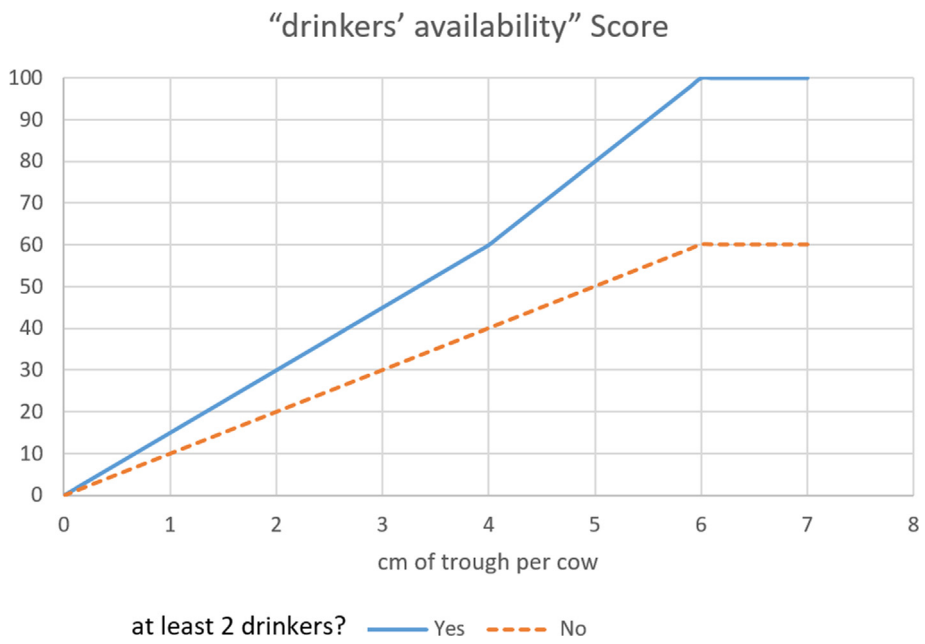

Fig. 1. 'Drinkers' availability' score depending on the presence of at least two drinkers accessible per animal, and the cumulated cm of trough (or equivalent) per cow.

classification was made by consensus of all the veterinarians (n > 5) that visited the dairy farms on a regular basis. See van Eerdenburg et al. (2021) for more details.

Sensitivity analysis of the original and alternative models

We used the Morris method (Morris, 1991) modified by Campolongo et al. (2007) to perform sensitivity analyses of both the original Welfare Quality scoring model and the alternative model including the modifications for 'Absence of prolonged thirst' and 'Absence of disease'. We performed the analysis for loose-house farms and tied-stall farms separately because Welfare Quality measures slightly differ between the two systems.

The Morris method allows the identification of the important inputs of a model, including those involved in interactions. The Morris method is used when the number of model inputs (i.e. measures in our case) is too important and thus, testing all combinations is too expensive from a computational point of view. The method is based on a 'One-factor-At-a-Time' (**OAT**) design of experiments. In brief, within the input space (consisting of all combinations of possible values for each input), an initial point (e.g. a farm with 20% too lean cows and 10% lame cows and 2% mortality and etc.) is randomly selected; a next point is defined by increasing or decreasing the value for only one input by an elementary shift. The difference in model output produced by the two points is calculated. This is repeated until all inputs have varied once. Thereafter, the whole process is repeated starting from another initial point, until the convergence of order of influence of inputs. The Morris method calculates elementary effects (R_i) due to each input using the equation:

$$R_i(x_1.\cdots.x_n.\Delta) = \frac{y(x_1.\cdots.x_{i-1}.(x_i+\Delta).x_{i+1}.\cdots.x_n) - y(x_1.\cdots.x_{i-1}.x_i.x_{i+1}.\cdots.x_n)}{\Delta}$$

where $y(X)$ is the output. $X = (x_1. x_2. \ldots. x_n)$ is the n-dimensional vector of inputs studied. Δ is the elementary increment (decrease or increase) of the OAT.

For each input, we obtain two sensitivity indices calculated from of all the R_i obtained for the input (Saltelli et al., 2004):

- The absolute mean (μ^*) of R_i, estimating the overall influence of the input i on the output. For instance, a μ^*-value of 50 for an input means that an increase (=Δ) in the input of 0.2 of its distribution, i.e. an increase of 20 percentile thus considering the initial value and the distribution of the input values (e.g. uniform *vs.* observed distribution), increases the score by 10 points – 50 * 0.2 = 10);
- The SD of R_i, estimating higher order effects, i.e. non-linear effects (e.g. threshold effects) or interactions with other inputs.

The most influencing inputs are those with high values for both μ^* and *SD*.

As OAT is subject to randomness, the exploration of the input space was improved by using Latin Hypercube Sampling (**LHS**) (e.g. Van Griensven et al., 2002; Francos et al., 2003), which maximises the 'maximin' criterion (Johnson et al., 1990) and so ensures that initial positions of points are well distributed.

We expect the 'overall' output to be little sensitive to each input because there are many inputs and only four overall categories. To be able to discriminate influent inputs, we used 500 OAT (allowing to calculate 500 R_i for each input). Moreover, we checked the convergence of the method (i.e. same ranking of inputs according to their influence and changes in μ^* or *SD* < 0.01 for all inputs).

We decided on an elementary increment (Δ) corresponding to 1/5 of its distribution. In a first step, we used a Uniform distribution of inputs, which is common when one wants to understand model behaviour (Monod et al., 2006). In that case, the increment corresponds to 1/5 of the range (e.g., 20 if the input is expressed on a 0–100 scale). In a second step, we considered the distributions observed in the whole dataset (n = 491 farms). In that case, the increment corresponds to a 20 percentile. We thus could check if the influence of the inputs was similar when tested on the two types of distributions.

In order to avoid unit effects and to facilitate the sensitivity analysis interpretation, we rescaled input variables between zero and one, before calculating the two sensitivity analysis indices.

Calculations for the sensitivity analysis were performed with R (version 3.6) (R Core Team, 2019). The 'sensitivity' package (Iooss et al., 2020) was used for the calculation of sensitivity indices.

Consistency of the original and alternative models with experts' opinion

We evaluated the consistency of the original and the alternative model with experts' opinion:

- In Subset 1, experts expressed their opinion on each of the 44 farms by providing an overall welfare score on a continuous scale. We used mixed linear regression to relate experts' opinion (the variable to be explained) and the welfare category ('not classified', 'acceptable', 'excellent', or 'enhanced') produced by the original or the alternative scoring model (explanatory variable), with the expert as random factor.
- In Subset 2, experts expressed their opinion on each of the 60 farms using a 3-point scale. We used an ordinal regression (also known as cumulative link model) to relate experts' opinion and the category produced by the original or the alternative scoring model.

We applied the Vuong test for non-nested models (Vuong, 1989) to check if the alternative model better fits experts' opinion than the original model or inversely.

Calculations were performed with R 3.6 (R Core Team, 2019), with the use of the package 'lme4' (Bates et al., 2015) for mixed linear regression, the 'car' package (Fox and Weisberg, 2019) for the Levene Test, the 'ordinal' package (Christensen, 2019) for calculation of ordinal model, and the package 'nonnest2' (Merkle and You, 2020) for the Vuong tests.

Polynomial approximation of the I-spline curves designed to score 'Absence of disease' are given in Supplementary Table S2. The measure 'Frequency of coughing per cow per 15 min' was not kept, because it never reached the level of warning threshold in the database, and partners from the Welfare Quality Network agreed to remove it from the protocol because it cannot be measured accurately at animal level.

We only detail results of the sensitivity analysis for criteria and principles affected by the changes in calculations (Criteria 'Absence of prolonged thirst' and 'Absence of disease'; Principles 'Good feeding' and 'Good health'; and overall assessment), for the Uniform distribution, and for the case of loose housing. The detailed results for other criteria and principles (which are not impacted by the updated protocol), or calculated with the observed distribution or Uniform distribution in tied farms are given in Supplementary Tables S3–S13 for criteria and Supplementary Tables S14–S17 for principles.

Sensitivity of the new calculations for 'Absence of prolonged thirst'

The five inputs involved in the criterion 'Absence of prolonged thirst' were all influential (i.e. μ^* and $SD \geq 0.1$) at criterion level in the case of loose housing (Table 1): the number and the cumulated length of the water troughs, the number of water bowls, the cleanliness of the water elements and the answer to the question 'Is the water flow enough? (True/False)'. In the original model, the mean effects (μ^*) ranged from 14.7, for the cumulated length of water troughs, to 27.2 for water flow, except for the cleanliness of water points, which mean effect was equal to 109.8. As the input 'cleanliness of water points' is a Boolean (true/false), when applying the sensitivity analysis, the shift over the distribution necessarily implied a change from true to false (or inversely). In the model with thirst improvement, the mean effects (μ^*) ranged from about 9.5, for the cumulated length of water troughs, to 20.4 for the number of water bowl, except for the cleanliness of water points, which mean effect was 109.8. From the original model to the alternative model with thirst improvement, the *SD* of the elementary effects were reduced by about 50% (from 40% for the number of water troughs to 66% for the number of water bowls).

In the case of tied housing, only the number and the cleanliness of the water bowl were influential (Supplementary Table S4). There is no water trough in tied-stall barns, so that the 'Total length of water troughs' and the 'Number of water troughs' always equal 0, that explains the absence of influence in the sensitivity analysis. Theoretically, the water flow would influence this criterion. However, as described previously, even if all the water bowls present an insufficient water flow, the drinkers' length remains far above the recommendation, so that the minimum score (i.e. only one drinker available per cow) was 60 in both models when water was clean, and was 32 and 20 in the original model and the alternative model, respectively, when the water was not clean.

The principle 'Good feeding' was sensitive to the five inputs of 'Absence of prolonged thirst' and sensitive to the input '% of very lean cows' (Table 2). The new model slightly increased the influence of the '% of very lean cows' and of the 'cleanliness of water points', by about 9%, while reducing the influence of the other inputs by 33% (for the number of water bowls) to 51% (for the water flow). Between the original model and the model with thirst improvement, the *SD* of the elementary effects for the five inputs linked to thirst were reduced by about 55% (from 27% for the cleanliness to 75% for the number of water bowls), whereas the SD for '% of lean cows' (linked to hunger) was slightly increased.

Sensitivity of the new calculations for 'Absence of disease'

In the original model, there were 10 influential inputs that constitute the Criterion 'Absence of disease' (Table 3). The mean effects (μ^*) of the 10 inputs used in the Criterion 'Absence of disease' ranged from 5.5 to 9.9. When the new calculations for 'Absence of disease' were used, the mean effects (μ^*) of inputs ranged from 1.1 to 15.4, with only nine inputs considered since the 'Frequency of coughing per cow per 15 min' has been removed. Results for tied stalls, with Uniform distribution, were similar to those obtained for loose housing. With the observed distribution, the "Frequency of coughing per cow per 15 min" had no influence, as no observation in the database reached the warning threshold (three coughs/-cow/15 min, while the maximum observed was 1.07) (Supplementary Table S8). The measure '% cows with increased

Table 1
Mean (μ^*) and SD of the elementary effects associated with Criterion 'Absence of prolonged thirst' of the Welfare Quality scoring model for dairy cows. Results were scaled with a Uniform distribution for loose housing (detailed in Supplementary Table S1). The higher the μ^* and SD, the more influence the input has.

Item	Original model		Model with health and thirst improvement	
	μ^*	SD	μ^*	SD
Cleanliness of water points	109.81	46.90	119.37	26.60
Water flow	27.17	61.08	15.65	33.24
Number of water bowls	23.72	129.87	20.40	44.58
Number of water troughs	22.49	60.88	14.38	36.74
Total length of water troughs	14.72	54.15	9.47	24.88

Table 2
Mean (μ^*) and SD of the elementary effects associated with Principle 'Good feeding' of the Welfare Quality scoring model for dairy cows. Results were scaled with a Uniform distribution for loose housing (detailed in Supplementary Table S1). The higher the μ^* and SD, the more influence the input has.

Item	Original model		Model with health and thirst improvement	
	μ^*	SD	μ^*	SD
% of very lean cows	34.39	35.57	37.48	36.05
Cleanliness of water points	32.12	16.80	35.05	12.26
Water flow	9.48	21.78	4.66	9.84
Number of water bowls	8.82	50.15	5.94	12.47
Number of water troughs	7.82	21.41	4.22	10.50
Total length of water troughs	5.38	21.26	2.89	7.18

Table 3
Mean (μ^*) and SD of the elementary effects associated with Criterion 'Absence of disease' of the Welfare Quality scoring model for dairy cows. Results were scaled with a Uniform distribution for loose housing (detailed in Supplementary Table S1). The higher the μ^* and SD, the more influence the input has. 'NA' implies here that the 'frequency of coughing' input was removed from the model.

Item	Original model		Model with health and thirst improvement	
	μ^*	SD	μ^*	SD
% cows with vulvar discharge	9.92	11.86	6.54	9.90
% cows with diarrhoea	9.82	11.86	4.64	8.86
% mastitis (milk somatic cell count > 400 000)	9.80	11.97	11.62	14.67
% downer cows	9.80	11.79	6.17	9.79
% mortality during the last 12 months	9.69	11.70	8.91	12.11
% dystocia	9.38	11.39	1.15	2.97
Frequency of coughing per cow per 15 min	6.87	12.16	NA	NA
% cows with nasal discharge	6.11	11.32	2.48	5.97
% cows with increased respiratory rate	5.69	11.15	15.36	16.13
% cows with ocular discharge	5.52	10.81	1.20	3.69

respiratory rate' had very little influence due to the low prevalence in the database. With the observed distribution for loose housing, the mean effects (μ^*) of the eight other inputs used in the Criterion 'Absence of disease' ranged from 10.20 to 18.05. When the new calculations for 'Absence of disease' were used, the mean effects (μ^*) of these inputs increased from 11 to 37% (Supplementary Table S8).

The principle 'Good health' was sensitive to the 10 inputs of Criterion 'Absence of disease' and to the 14 inputs of Criteria 'Absence of injuries' and 'Absence of pain induced by management procedures' (Table 4). The influence of several diseases was increased in the alternative model compared to the original one: the influence of '% cows with increased respiratory rate'', +262%; '% cows with mastitis', +54%; '% of mortality', +14%. The influence of other 'Absence of disease' inputs was reduced (−15 to −86%). With the observed distribution for loose housing, the influence of all diseases increased (from +60 to +165%, Supplementary Table S16), except for '"% of cows with vulvar discharge" which decreased by 9%. The influence of the inputs from Criteria 'Absence of injuries' and 'Absence of pain due to management procedures' was also reduced (−8 to −29%). Within the criterion 'Absence of injuries', lameness was 66% more influential than skin alterations/lesions.

Within the principle 'Good health', using a uniform distribution, lameness ('% of not lame cows') which was the second most influential input (after the type of method used for dehorning cows) with the original model is the third most influential with the alternative model.

Sensitivity of the overall scoring and consistency with experts' opinion

In both the original and the alternative models, the overall score showed very low levels of sensitivity (the maximum μ^* for an input is 0.39) with high levels of interaction effects (*SD* is an average 10 times higher than μ^*) (Table 5).

Within Subset 1, the consistency to experts' opinion was slightly better for the alternative model compared to the original one ($Z = -1.548$, $p_{original\_better} = 0.939$, $p_{new\_better} = 0.061$, with $p_{original\_better}$ the probability to reject the hypothesis that the original model match better with the expert opinion than the new one and p_{new_better} the probability to reject the hypothesis that the new model match better with the expert opinion than the original one).

Within Subset 2, the consistency to experts' opinion was similar between the original and the alternative model ($Z = 0.371$, $p_{original\_better} = 0.355$, $p_{new\_better} = 0.645$).

Discussion

We applied two modifications of the Welfare Quality scoring model for dairy cows. We modified the calculation of the score for criterion 'absence of prolonged thirst' to avoid threshold effects and to more precisely take into account the cleanliness of drinkers. We modified the calculation of the 'absence of diseases' criterion by calculating an elementary score for each health disorder and computing the three lowest of them to limit compensation between health disorders. An alternative scoring model is proposed that incorporates these two modifications, which aim to change the sensitivity of the model to measures and to better match experts' opinion.

The modified calculation of the score for 'Absence of prolonged thirst' reduced the overly large influence of the resource-based measures that the original model has been criticised for. This overly large influence in the original scoring model was partly due to the fact that this criterion was based on a decision tree, which by nature implies the definition of several thresholds and thus threshold effect. For instance, adding a drinker in a pen resulted in a large improvement of the score for 'Absence of prolonged thirst' if it allowed a farm to switch from one branch to another in the decision tree (e.g., switching from the branch 'only one drinker' to the branch 'at least two drinkers'). The modifications proposed for the calculation of the 'Absence of prolonged

Table 4
Mean (μ^*) and SD of the elementary effects associated with Principle 'Good health' of the Welfare Quality scoring model for dairy cows. Results were scaled with a Uniform distribution for loose housing (detailed in Supplementary Table S1). The higher the μ^* and SD, the more influence the input has. 'NA' implies here that the 'frequency of coughing' input was removed from the model.

Item	Original model		Model with health and thirst improvement	
	μ^*	SD	μ^*	SD
Criterion Absence of injuries				
% not lame cows	9.03	6.75	7.53	5.85
% severely lame cows	6.15	7.02	4.35	4.03
% moderately lame cows	1.74	1.86	1.34	1.34
% cows with no lesion	6.63	5.96	5.19	4.6
% cows with at least one lesion	5.16	5.9	3.64	3.99
% cows with at least one hairless patch and no lesion	1.02	1.2	0.74	0.85
Criterion Absence of pain induced by management procedures				
Method used for dehorning	10.06	14.64	8.45	12.3
% dehorned cows	4.53	14.69	4.15	13.33
Use of anaesthetics for dehorning	3.58	4.92	3.18	4.27
Use of analgesics for dehorning	3.35	4.74	2.95	4.06
Use of analgesics for tail docking	4.71	5.78	4.08	4.88
Method used for tail docking	4.69	5.42	4.03	4.51
% tail-docked cows	4.33	12.57	3.80	10.8
Use of anaesthetics for tail docking	4.03	5.30	3.50	4.51
Criterion Absence of disease				
% downer cows	4.03	5.65	2.95	5.06
% cows with diarrhoea	3.92	5.54	2.29	4.8
% mastitis (milk somatic cell count > 400 000)	3.91	5.56	6.03	8.79
% cows with vulvar discharge	3.86	5.38	3.29	5.46
% mortality during the last 12 months	3.83	5.39	4.38	6.53
% dystocia	3.50	5.19	0.48	1.35
Frequency of coughing per cow per 15 min	2.63	5.13	NA	NA
% cows with nasal discharge	2.36	4.78	1.1	2.38
% cows with increased respiratory rate	2.30	4.94	8.32	9.53
% cows with ocular discharge	2.08	4.51	0.53	2.07

thirst' score avoid most of the threshold effects because they include continuous equations instead of thresholds. This is confirmed by the sensitivity analysis, which shows a reduction of the SD of the effects of each individual measure on Criterion 'Absence of prolonged thirst' (SD of the elementary effects: from an average of 70.6 in the original model to 33.2 in the alternative model). There may be a problem for tied stall where the minimum score is 60. We did not put our efforts into refining the calculation for cows in tied stalls because they are likely to disappear, at least in Europe (EFSA Panel on Animal Health and Animal Welfare et al., 2023).

The new calculations for criterion 'Absence of disease' significantly reduce potential compensations between health disorders. Indeed, by considering the three lowest scores associated with a health disorder, we avoid the compensation by the six others. Moreover, by using Choquet integral on the three lowest scores, we can allow only partial compensation between these three scores. For instance, contrary to the original scoring model, a shift from 0% of mortality to an extreme value e.g. 100% results in a very low score for 'Absence of disease' with the alternative model. These expected effects were confirmed by the sensitivity analysis. In the original model, with the Uniform distribution, all inputs had a similar influence, in terms of mean levels (μ^*) and interactions (SD). With the alternative model, the influence of each health disorder varies. For example, mortality, which was found by de Graaf et al. (2018) as not influential enough in the original model, is now the third most influent input for 'Absence of disease'. Another positive effect of the new model is the reduction of threshold effects, due to the use of alarm and warning thresholds. With the use of spline curves, thresholds are not used anymore and each health disorder can vary continuously from 0 to 100. However, by reducing the compensation between health disorders, the average value of the 'Absence of disease' score is lower in the alternative model than in the original model. Because in Welfare Quality low values are more influential than high ones (Botreau et al., 2008), the influence of 'Absence of disease' on the principle 'Good health' is increased and as a consequence, the influence of the other criteria ('Absence of injuries' and 'Absence of pain due to management procedures') and their related measures is reduced.

The overall assessment is not very sensitive to each input. Indeed, there are 49 inputs and only four categories for the overall assessment and these inputs are rather independent of each other, so it is expected that a change in one input only rarely modifies the overall assessment and only a combination of inputs changes can modify the overall assessment.

The number of measures to be aggregated varies between criteria. The more numerous the measures aggregated into a criterion or a principle, the lower the influence of each measure. Indeed, when calculating the average of two or four data points, the influence of each data is ½ (respectively ¼) when two (respectively four) data points are averaged. The aggregation method is more complex than an average but the impact on the number of measures to be combined is similar. Alternatively, one could make groups of measures of the same size to build criteria and groups of criteria of the same size to build principles; for instance, from a set of 45 measures one could make nine criteria of five measures each then three principles of three criteria each, before aggregating the three resulting principles into an overall score. This would result in the same mathematical expectation for the influence of each individual measure but will certainly weaken the biological meaning of criteria. The question lies in whether we consider that welfare is composed of criteria that can be measured in different ways or of measures with each measure representing an aspect of welfare. Welfare Quality rather identified the criteria that are meaningful for animal welfare and represent separate aspects of welfare, then grouped the criteria into functional principles (feeding, housing, health, behaviour). The fact that the number of measures varies with criteria, inevitably results in varying influence of

Table 5
Mean (μ^*) and SD of the elementary effects associated with the 'overall' score of the Welfare Quality scoring model for dairy cows. Results were scaled with a Uniform distribution for loose housing (detailed in Supplementary Table S1). The higher the μ^* and SD, the more influence the input has. 'NA' implies here that the 'frequency of coughing' input was removed from the model, and '–' implies below 0.01.

Item	Original model		Model with health and thirst improvement	
	μ^*	SD	μ^*	SD
Principle Good feeding				
Criterion Absence of prolonged hunger				
% of very lean cows	0.34	1.26	0.36	1.29
Criterion Absence of prolonged thirst				
Cleanliness of water points	0.33	0.74	0.39	0.79
Number of water bowls	0.25	2.12	-	-
Number of water troughs	0.22	0.78	0.03	0.30
Water flow	0.17	0.70	0.02	0.27
Total length of water troughs (cm)	0.10	0.70	0.01	0.22
Principle Good housing				
Criterion Comfort around resting				
% of lying down movements with collisions	0.04	0.45	0.03	0.39
Duration of lying down movements (s)	0.01	0.22	0.01	0.22
% of lying cows which lie partly outside lying area	0.01	0.22	0.01	0.22
Criterion Ease of movement				
Number of hours on pasture per day	0.17	0.91	0.17	0.91
Number of days on pasture per year	0.13	0.8	0.12	0.77
Number of days with access to outdoor loafing area per year	0.07	0.59	0.08	0.63
Principle Good health				
Criterion Absence of injuries				
% of severely lame cows	0.10	0.70	0.06	0.54
% of not lame cows	0.09	0.67	0.11	0.73
% of moderately lame cows	0.07	0.59	0.03	0.39
% of cows with no lesion	0.08	0.63	0.08	0.63
% of cows with at least one lesion	0.05	0.50	0.11	0.73
% of cows with at least one hairless patch and no lesion	0.01	0.22	0.03	0.39
Criterion Absence of pain induced by management procedures				
Method used for dehorning	0.14	0.63	0.14	0.63
Use of analgesics for dehorning	0.07	0.37	0.08	0.4
Use of anaesthetics for dehorning	0.06	0.35	0.10	0.43
% of dehorned cows	0.05	0.50	0.06	0.54
% of tail-docked cows	0.06	0.54	0.04	0.45
Method used for tail docking	0.04	0.29	0.05	0.32
Use of analgesics for tail docking	0.03	0.25	0.07	0.36
Use of anaesthetics for tail docking	0.02	0.22	0.06	0.35
Criterion Absence of disease				
% of mastitis (milk somatic cell count > 400 000)	0.06	0.54	0.13	0.80
% of cows with diarrhoea	0.04	0.45	0.01	0.22
% of cows with vulvar discharge	0.02	0.32	0.03	0.39
% of mortality during the last 12 months	0.02	0.32	0.02	0.32
Frequency of coughing per cow per 15 min	0.02	0.32	NA	NA
% of cows with nasal discharge	0.02	0.32	0.02	0.32
% of downer cows	0.01	0.22	0.05	0.50
% of dystocia	0.01	0.22	-	-
% of cows with increased respiratory rate	0.01	0.22	0.18	0.93
% of cows with ocular discharge	0.01	0.22	-	-
Principle 'Appropriate behaviour'				
Criterion Expression of social behaviour				
Frequency of other aggressive events per animal per hour	0.13	0.80	0.16	0.88
Frequency of butts per cow per hour	0.02	0.32	0.03	0.39
Criterion Expression of other behaviour				
Number of hours on pasture per day	0.17	0.91	0.17	0.91
Number of days on pasture per year	0.13	0.80	0.12	0.77
Number of days with access to outdoor loafing area per year	0.07	0.59	0.08	0.63
Criterion Good human-animal relationship				
% of cows that can be touched	0.19	0.96	0.21	1
% of cows that cannot be approached	0.05	0.50	0.05	0.50
% of cows that can be approached between 50 cm and 1 m	0.03	0.39	0.03	0.39
% of cows that can be approached by 50 cm but not touched	0.02	0.32	0.02	0.32
Criterion Positive emotional state				
Qualitative Behaviour Assessment	0.33	1.24	0.39	1.34

each individual measure across criteria. Therefore, the influence of measures should be interpreted keeping in mind the number of measures in a criterion.

The alternative model is more aligned with experts' opinion than the original model (in particular in terms of sensitivity). Assessing animal welfare is a value-based exercise (Fraser, 1995). The Welfare Quality scoring model was, therefore, based on expert opinions and, by definition, cannot fit all expert opinions, because values vary between experts. For example, a farm can be considered 'Excellent' (best) by an expert and 'not classified' (worst) by another (e.g. de Graaf et al., 2017, Fig. 3, Herd 2). The model should nevertheless fit consulted experts' opinion. Here, we were able to compare the overall assessment produced by the original and the alternative models to that attributed by experts. The alternative

model better matches experts' opinion than the original model, at least when these opinions are expressed on the same scale as that of the model (four categories of welfare: Excellent, Enhanced, Acceptable, Not classified).

There is definitively room for further improvements of the Welfare Quality scoring model. The way the measures are aggregated within a criterion can still be improved. For example, we could change the way to aggregate integument alterations as proposed in van Eerdenburg et al. (2018), or we could consider the distribution of individuals' problems (e.g. an animal affected by two disorders may have more welfare consequences than two animals affected each by a single disorder) as proposed by Sandøe et al. (2019). This would require a specific consultation of experts.

Conclusions

The alternative model, proposed in this paper to improve the Welfare Quality scoring, performs better than the original one. Compared to the original model, the alternative one significantly reduces the influence and 'threshold effects' of measures related to drinkers, changes the influence of each health disorder reducing compensation between them. The Welfare Quality Network is updating the welfare Quality protocol for dairy cows by including the alternative scoring model proposed in this paper.

Supplementary material

Supplementary material to this article can be found online at https://doi.org/10.1016/j.animal.2023.101018.

Ethics approval

Not applicable.

Data and model availability statement

Welfare Quality protocol are freely available at https://www.welfarequalitynetwork.net/ and Welfare Quality simulator is freely accessible at https://www1.clermont.inrae.fr/wq/.

The data that support the study findings were not deposited in an official repository, but are available from the authors upon request.

Declaration of Generative AI and AI-assisted technologies in the writing process

The authors did not use any artificial intelligence-assisted technologies in the writing process.

Author ORCIDs

R. Lardy: https://orcid.org/0000-0003-1338-8553.
R. Botreau: https://orcid.org/0000-0002-9599-4313.
A. de Boyer des Roches: https://orcid.org/0000-0002-6903-7456.
F. van Eerdenburg: https://orcid.org/0000-0002-6024-6905.
S. de Graaf: https://orcid.org/0000-0002-1676-8925.
M. Haskell: https://orcid.org/0000-0001-9373-0624.
M.K. Kirchner: https://orcid.org/0000-0003-1321-4342.
L. Mounier: https://orcid.org/0000-0003-1083-8563.
Miroslav Kjosevski: https://orcid.org/0000-0001-5725-4384.
F. Tuyttens: https://orcid.org/0000-0002-1348-218.
I. Veissier: https://orcid.org/0000-0002-8497-5395.

Author contributions

R. Lardy: Conceptualisation, Methodology, Software, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualisation. **R. Botreau**: Formal analysis, Investigation, Writing – review & editing. **A. de Boyer des Roches**: Investigation, Resources, Writing – review & editing. **F.J.C.M. van Eerdenburg**: Investigation, Resources, Writing – review & editing. **S. de Graaf**: Investigation, Resources, Writing – review & editing. **M. Haskell**: Investigation, Resources, Writing – review & editing. **M. K. Kirchner**: Investigation, Resources, Writing – review & editing. **L. Mounier**: Investigation, Resources, Writing – review & editing. **M. Kjosevski**: Investigation, Resources, Writing – review & editing. **F. Tuyttens**: Investigation, Resources, Writing – review & editing. **I. Veissier**: Conceptualisation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Supervision, Funding acquisition.

Declaration of interest

None.

Acknowledgements

The authors thank Metaform Langues for their editorial support and English reviewing. We thank the experts consulted to build the spline curves for the health symptoms.

We thank the Welfare Quality Network partners for their contribution in the choice in the new scores calculation. We thank S. N. Andreasen and C. Winckler for his help in obtaining the Welfare Quality database.

Financial support statement

This work received financial support from the French government's IDEX-ISITE initiative 16-IDEX-0001 (CAP 20-25).

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. Livestock Production Science 61, 179–192.
Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. Journal of Statistical Software 67, 1–48.
Botreau, R., Capdeville, J., Perny, P., Veissier, I., 2008. Multicriteria evaluation of animal welfare at farm level: an application of MCDA methodologies. Foundations of Computing and Decision Sciences 33, 287–316.
Burow, E., Thomsen, P.T., Rousing, T., Sørensen, J.T., 2013. Daily grazing time as a risk factor for alterations at the hock joint integument in dairy cows. Animal 7, 160–166.
Campolongo, F., Cariboni, J., Saltelli, A., 2007. An effective screening design for sensitivity analysis of large models. Environmental Modelling & Software 22, 1509–1518.
Christensen, R.H.B., 2019. ordinal—Regression Models for Ordinal Data. " R package Retrieved on 09 October 2023 from http://www.cran.r-project.org/package=ordinal/.
de Graaf, S., Ampe, B., Winckler, C., Radeski, M., Mounier, L., Kirchner, M.K., Haskell, M.J., van Eerdenburg, F.J.C.M., des Roches, A. de B., Andreasen, S.N., Bijttebier, J., Lauwers, L., Verbeke, W., Tuyttens, F.A.M., 2017. Trained-user opinion about Welfare Quality measures and integrated scoring of dairy cattle welfare. Journal of Dairy Science 100, 6376–6388.
de Graaf, S., Ampe, B., Buijs, S., Andreasen, S., Roches, A.D.B.D., van Eerdenburg, F., Haskell, M., 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, 157–166.
de Vries, M., Bokkers, E.A.M., van Schaik, G., Botreau, R., Engel, B., Dijkstra, T., de Boer, I.J.M., 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.
EFSA Panel on Animal Health and Animal Welfare (AHAW), Nielsen, S.S., Alvarez, J., Bicout, D.J., Calistri, P., Canali, E., Drewe, J.A., Garin-Bastuji, B., Gonzales Rojas, J. L., Gortázar Schmidt, C., Herskin, M., Michel, V., Miranda Chueca, M.Á., Padalino, B., Roberts, H.C., Spoolder, H., Stahl, K., Velarde, A., Viltrop, A., De Boyer, A., des

Roches, Jensen, M.B., Mee, J., Green, M., Thulke, H.-H., Bailly-Caumette, E., Candiani, D., Lima, E., Van der Stede, Y., Winckler, C., 2020. Welfare of dairy cows. EFSA Journal 21, e07993.
Fox, J., Weisberg, S., 2019. An R Companion to Applied Regression. Sage, Thousand Oaks CA, USA.
Francos, A., Elorza, F.J., Bouraoui, F., Bidoglio, G., Galbiati, L., 2003. Sensitivity analysis of distributed environmental simulation models: understanding the model behaviour in hydrological studies at the catchment scale. Reliability Engineering & System Safety 79, 205–218.
Fraser, D., 1995. Science, values and animal welfare: exploring the 'inextricable connection'. Animal Welfare 4, 103–117.
Fraser, D., 2003. Assessing animal welfare at the farm and group level: the interplay of science and values. Animal Welfare 12, 433–443.
Heath, C.A.E., Browne, W.J., Mullan, S., Main, D.C.J., 2014. Navigating the iceberg: reducing the number of parameters within the Welfare Quality® assessment protocol for dairy cows. Animal 8, 1978–1986.
Iooss, B., Janon, A., Pujol, G., Broto with contributions from B., Boumhaout, K., Veiga, S.D., Delage, T., Fruth, J., Gilquin, L., Guillaume, J., Gratiet, L.L., Lemaitre, P., Marrel, A., Meynaoui, A., Nelson, B.L., Monari, F., Oomen, R., Rakovec, O., Ramos, B., Roustant, O., Song, E., Staum, J., Sueur, R., Touati, T., Weber, F., 2020. Sensitivity: Global sensitivity analysis of model outputs. R package retrieved on 09 October 2023 from https://cran.r-project.org/web/packages/sensitivity/index.html.
Iooss, B., Lemaître, P., 2015. A review on global sensitivity analysis methods. In: Dellino, G., Meloni, C. (Eds.), Uncertainty Management in Simulation-Optimization of Complex Systems. Operations Research/Computer Science Interfaces Series. Springer US, Boston, MA, USA, pp. 101–122.
Johnson, M.E., Moore, L.M., Ylvisaker, D., 1990. Minimax and maximin distance designs. Journal of Statistical Planning and Inference 26, 131–148.
Knierim, U., Winckler, C., Mounier, L., Veissier, I., 2021. Developing effective welfare measures for cattle. Understanding the behaviour and improving the welfare of dairy cattle. Burleigh Dodds Science Publishing, Cambridge, UK.
Merkle, E., You, D., 2020. nonnest2: Tests of Non-Nested Models. R package retrieved on 09 October 2023 from https://cran.r-project.org/web/packages/nonnest2/index.html.
Monod, H., Naud, C., Makowski, D., 2006. Uncertainty and sensitivity analysis for crop models. In: Wallack, D., Makowski, D., Jones, J.W. (Eds.), Working with Dynamic Crop Models. Elsevier, Amsterdam, The Netherlands, pp. 55–100.
Morris, M.D., 1991. Factorial sampling plans for preliminary computational experiments. Technometrics 33, 161–174.
R Core Team, 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Saltelli, A., Annoni, P., 2010. How to avoid a perfunctory sensitivity analysis. Environmental Modelling & Software 25, 1508–1517.
Saltelli, A., Tarantola, S., Campolongo, F., Ratto, M., 2004. Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models. John Wiley & Sons, Hoboken, NJ, USA.
Saltelli, A., Ratto, M., Tarantola, S., Campolongo, F., 2006. Sensitivity analysis practices: Strategies for model-based inference. Reliability Engineering & System Safety 91, 1109–1125.
Sandøe, P., Corr, S., Lund, T., Forkman, B., 2019. Aggregating animal welfare indicators: can it be done in a transparent and ethically robust way? Animal Welfare 28, 67–76.
Spoolder, H., De Rosa, G., Horning, B., Waiblinger, S., Wemelsfelder, F., 2003. Integrating parameters to assess on-farm welfare. Animal Welfare 12, 529–534.
Stomp, M., Demont, C., Veissier, I., 2023. Pratiques actuelles d'évaluation du bien-être animal des filières volailles et bovins. Innovations Agronomiques 87, 19–28.
Tuyttens, F.A.M., de Graaf, S., Andreasen, S.N., de Boyer des Roches, A., van Eerdenburg, F.J.C.M., Haskell, M.J., Kirchner, M.K., Luc, Mounier, Kjosevski, M., Bijttebier, J., Lauwers, L., Verbeke, W., Ampe, B., 2021. Using expert elicitation to abridge the Welfare Quality® protocol for monitoring the most adverse dairy cattle welfare impairments. Frontiers in Veterinary Science 8, 634470.
van Eerdenburg, F.J.C.M., Hulsen, J., Snel, B., van den Broek, J., Stegeman, A., 2018. A proposal for three modifications for the Welfare Quality© protocol for dairy cattle. In: van Eerdenburg, F.J.C.M. (Ed.), Bienestar animal en la práctica, en producciones lecheras, desde la perspectiva europea. Utrecht University Repository, Utrecht, the Netherlands, pp. 46–54.
van Eerdenburg, F.J.C.M., Hof, T., Doeve, B., Ravesloot, L., Zeinstra, E.C., Nordquist, R. E., van der Staay, F.J., 2021. The relation between hair-cortisol concentration and various welfare assessments of Dutch dairy farms. Animals 11, 821.
Van Griensven, A., Francos, A., Bauwens, W., 2002. Sensitivity analysis and auto-calibration of an integral dynamic model for river water quality. Water Science & Technology 45, 325–332.
Veissier, I., Jensen, K., Botreau, R., Sandøe, P., 2011. Highlighting ethical decisions underlying the scoring of animal welfare in the Welfare Quality® scheme. Animal Welfare 20, 14.
Vuong, Q.H., 1989. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica 57, 307.
Wang, W., Yan, J., 2020. splines2: Regression Spline Functions and Classes. R package Retrieved on 09 October 2023 from https://cran.r-project.org/web/packages/splines2/index.html.
Welfare Quality®,, 2009. Welfare Quality® Assessment Protocol for Cattle. Welfare Quality®Consortium, Lelystad, Netherlands.