Animal (2014), **8:12**, pp 1978–1986 © The Animal Consortium 2014
doi:10.1017/S1751731114002018

animal

Navigating the iceberg: reducing the number of parameters within the Welfare Quality® assessment protocol for dairy cows

C. A. E. Heath†, W. J. Browne, S. Mullan and D. C. J. Main

School of Veterinary Sciences, University of Bristol, Langford House, Langford, Bristol, BS40 5DU, UK

(Received 6 January 2014; Accepted 27 June 2014; First published online 27 August 2014)

The Welfare Quality® protocols provide a multidimensional assessment of welfare, which is lengthy, and hence limited in terms of practicality. The aim of this study was to investigate potential 'iceberg indicators' which could reliably predict the overall classification as a means of reducing the length of time for an assessment and so increase the feasibility of the Welfare Quality® protocol as a multidimensional assessment of welfare. Full Welfare Quality® assessments were carried out on 92 dairy farms in England and Wales. The farms were all classified as Acceptable or Enhanced. Logistic regression models with cross validation were used to compare model fit for the overall classification on farms. 'Absence of prolonged thirst', on its own, was found to correctly classify farms 88% of the time. More generally, the inclusion of more measures in the models was not associated with greater predictive ability for the overall classification. Absence of prolonged thirst could thus, in theory, be considered to be an iceberg indicator for the Welfare Quality® protocol, and could reduce the length of time for a farm assessment to 15 min. Previous work has shown that the parameters within the Welfare Quality® protocol are important and relevant for welfare assessment. However, it is argued that the credibility of the published aggregation system is compromised by the finding that one resource measure (Absence of prolonged thirst) is a major driver for the overall classification. It is therefore suggested that the prominence of Absence of prolonged thirst in this role may be better understood as an unintended consequence of the published measure aggregation system rather than as reflecting a realistic iceberg indicator.

Keywords: dairy cows, animal welfare, classification, aggregation, models

Implications

The measures which comprise the Welfare Quality® assessment protocol for dairy cows are important and relevant for welfare assessment. The measure aggregation system provides a way of integrating these measures into an overall classification. The finding that the score for *Absence of prolonged thirst*, which represents only a single aspect of welfare, can reliably predict the overall classification, suggests that the measure aggregation system is not fulfilling the original aim of the protocol, by synthesising appropriately the multidimensional nature of welfare into a balanced summary measure. Understanding the nature of the measure aggregation system in this light is important for those considering using it.

Introduction

Scientifically robust and feasible welfare assessment protocols are vital to the role farm assurance schemes play in safeguarding farm animal welfare. Where previously, welfare assessment protocols have been concerned with the provision of resources, animal-based assessments are now considered to provide a more direct account of welfare (Webster *et al.*, 2004), and are recommended to be included in welfare assessments (Farm Animal Welfare Council, 2005). Animal-based measures include behavioural and physical observations and have the potential to assess welfare according to a definition which incorporates both physical and mental health (Dawkins, 2003).

The holistic nature of the Welfare Quality® protocol reflects current understanding of welfare as a multidimensional construct (Fraser and Broom, 1990), however, the time taken to carry out such an extensive assessment has been cited as limiting the potential for its practical application (Knierim and Winckler, 2009; Blokhuis *et al.*,. 2010; de Vries *et al.*, 2013c). According to the timings provided in the protocol, the time taken for a Welfare Quality® assessment of an average UK herd of 125 dairy cows (DairyCo, 2013) would be 7 h (Welfare Quality®, 2009). This is currently beyond the scope of farm assurance schemes, therefore any means

† E-mail: Cheryl.heath@bristol.ac.uk

of reducing the length of time taken for an assessment while maintaining accuracy would be beneficial. Previous work aimed at reducing the length of the assessment has demonstrated that, 'Replacing a set of animal-based Welfare Quality® indicators belonging to one assessment method with predictions based on remaining Welfare Quality® indicators showed little scope for reduction of on-farm assessment time of the Welfare Quality® protocol for dairy cattle' (de Vries *et al.*, 2013c).

As an alternative line of investigation, a theoretical notion of animal-based 'iceberg indicators' has been proposed. 'An "iceberg" indicator provides an overall assessment of welfare, just as the protruding tip of an iceberg signals its submerged bulk beneath the water's surface' (Farm Animal Welfare Council, 2009). If these proposed indicators could be found to reliably provide information about the welfare state of animals, assessments could be limited to those specific measures, and the time taken for an assessment could be reduced accordingly. Although work in this area remains theoretical, a potential candidate for an iceberg indicator might be the Qualitative Behaviour Assessment (QBA). Wemelsfelder *et al.* (2001) suggest that QBA '…reflects a "whole-animal" level of organisation, which may guide the interpretation of behavioural and physiological measurements in terms of the animal's overall welfare state'. This connection between QBA and the overall welfare state is proposed to be due to the integrative nature of QBA, whereby the behavioural descriptors which comprise the assessment can be understood as expressing the qualitative nature of the animal's experience (Wemelsfelder *et al.*, 2001; Wemelsfelder, 2007). With QBA in mind then, the aim of this study was to use modelling approaches that would predict the overall classification based on a subset of the measures, with a view to potentially finding a means of reducing the time taken to carry out an assessment.

Material and methods

Data collection

The collection of the data used in this paper, along with a list of the measures and any missing values associated with those measures, has been previously described in Heath *et al.* (2014). Briefly, full Welfare Quality® assessments were carried out on 92 dairy farms located in England and Wales. The farms were voluntary participants from three farm welfare assurance schemes, RSPCA Freedom Food, Soil Association Certification, and the Red Tractor Farm Assurance Dairy Scheme, and each assurance scheme recruited farmers independently. Data were collected by seven employed assessors who had received standardised training in the Welfare Quality® assessment protocols. Inter-assessor reliability testing of the assessors was conducted, and is described in Heath *et al.* (2014). The assessments took place between January and August 2011, each by an individual assessor during a single farm visit, immediately after morning milking, according to the protocol guidelines. Measures are defined, in this study, as independent units of data as required by Welfare Quality®, from which the aggregated scores are calculated. The data consists of 58 measures associated with herd level data, and 4 measures which relate to the provision of drinkers and access to pasture which were collected at the group level, where a herd may be made up of a number of groups of cows. For the purpose of this study, these group level measures have been considered as their aggregated herd level criteria scores, *Absence of prolonged thirst* and *Expression of other behaviours*. Similarly, as the QBA terms are not intended to be considered individually, the aggregated score of *Positive emotional state*, which refers to the version of QBA included in the Welfare Quality® protocol, has been used throughout the analysis instead.

Analysis of Welfare Quality® assessment

The data were analysed according to the Welfare Quality® protocol for dairy cows (Welfare Quality®, 2009). The protocol provides formulae or decision trees for the calculation of 12 criteria scores formed by combining the collected measures. The 12 criteria scores are then aggregated into four principle scores by applying the mathematical technique of Choquet integrals. Finally, an overall classification is calculated based on thresholds applied to the four principle scores. Owing to some errors in the printed version of the protocol, further revisions were applied (Veissier, I., personal communications, April and May 2012); Up to date versions of the calculations were programmed using the program R (R version 2.14.1) by the first author. All other analyses were carried out using Microsoft Excel (2010).

Missing data

Owing to the hierarchical nature of the scoring system, a single missing value from an assessment prevents the calculation of the overall classification of a farm. For this reason, it was necessary to fill in the missing data. For continuous measures, missing data were filled in using the observed mean value of the respective farm assurance scheme. For categorical data the observed modal value for the respective farm assurance scheme was used. With two exceptions, a similar method was applied to group level data using the modal or mean value for the farm, depending on whether the data were categorical or continuous. Firstly, when this was not available, for categorical data, the mode value for the respective farm assurance scheme was used. Secondly, for *Number of water troughs*, the number used to fill in the missing data corresponded to the ratio of the number of troughs to the number of cows in the group, based on other groups on the farm. In accordance with what would normally be expected on UK farms, data relating to measures which were universally omitted were filled in to indicate no routine tail docking, no tethering, no water bowls, and *Water flow* was also filled in as the water flow test had not been carried out, on the basis that, 'In the case of troughs with a large reservoir, this test does not have to be carried out. Water flow is then set to 20 l/min' (Welfare Quality®, 2009, p. 79).

Table 1 *Summary of measures included in models used to predict scores/classifications at higher levels of aggregation of the Welfare Quality® protocol for dairy cows*

Predicted aggregated score (LR/LDA)	Variables from which different combinations were selected	Number of models
Overall score (LR)	Absence of prolonged thirst	255
	Positive emotional state	
	% severely lame cows	
	% cows colliding with housing equipment	
	% cows with at least one hairless patch no lesion/swelling	
	% very lean cows	
	% moderately lame cows	
	Use of analgesics for dehorning	
Good behaviour (LR)	Positive emotional state	255
	Average frequency of butts per cow per hour	
	Average frequency of displacements per cow per hour	
	% cows that can be touched % cows that can be approached closer than 50cm but not touched	
	% cows that can be approached as closely as 100 to 50 cm	
	% cows that cannot be approached as closely as 100 cm	
	Other behaviour	
Good feeding (LDA)	% very lean cows	3
	Absence of prolonged thirst	
Good housing (LDA)	Mean time to lie down	31
	% of cows colliding (with housing equipment)	
	% cows lying outside lying area	
	% cows with dirty hind legs	
	% cows with dirty flank	
Good health (LDA)	% moderately lame cows	511
	% severely lame cows	
	% cows with at least one lesion/swelling	
	% cows with at least one hairless patch, no lesion/swelling	
	% cows with ocular discharge	
	% mastitis	
	% dystocia	
	% downer cows	
	% mortality	

LDA = linear discriminant analysis; LR = logistic regression.

Statistical analyses

To assess how well individual indicators or groups of indicators were able to predict higher levels of measure aggregation, at the principle or overall classification, a randomly drawn sample of half the data set of 92 farms was drawn. This was used as a 'training' set from which the coefficients for the model variables were established, and the remaining 46 farms were set aside as a 'test' set. By excluding the test set from the model building process, this allowed the models to be tested on hitherto 'unseen' data, thus providing a means of cross-validation. For each model the process of splitting the data set into 'training' and 'test' sets was repeated 1000 times. The accuracy of the models was defined as the proportion of times that the farms from the 'test' set were predicted as having the same classification as calculated from their actual Welfare Quality® assessment.

For the overall classification, only two categories were observed in our data, *acceptable* and *enhanced*, and it was therefore possible to build logistic regression models to relate groups of welfare indicators to the overall Welfare Quality® classification. In the initial stage of selecting which variables to include, a correlation threshold with the overall classification was applied (0.200). Then, additional variables were excluded on the basis of co-linearity. Finally, further variables were removed when the models failed to execute on account of limited variability associated with those variables. A summary of the models is included in Table 1. All possible values for the threshold were tested for each 'training' set. The value for the threshold that was associated with the greatest proportion of correctly classified farms within the 'training' set was then used on the 'test' set. As well as the accuracy of the models, the sensitivity and specificity were also calculated based on the 'test' set.

In terms of the principle level scores, for *Appropriate behaviour*, only two categories were observed in our data, and it was possible to carry out the same methodology as applied for predicting the overall classification, but this time including all the welfare indicators associated with *Appropriate behaviour.* For the remaining principles, where more than two classifications were observed, models based

on linear discriminant analysis were used. For the principle *Good health* insufficient variation of a number of welfare indicators resulted in their exclusion from the models. This related largely to a low prevalence associated with certain health measures, that is, most of the data was 0 as opposed to missing, and this has been reported elsewhere with regard Welfare Quality® (e.g. de Vries *et al.*, 2013b).

Results

The percentage of farms with missing data for criteria level scores ranged from 0% to 74% (for *Absence of pain induced by management procedures procedures*). This was excluding measures associated with routine tail docking, presence of tethering, presence of water bowls, and water flow. 5% of farms had missing data for the Welfare Quality® criteria *Positive emotional state*, and 18% of farms had missing data for the Welfare Quality® criteria, *Absence of prolonged thirst.* In our data set, all farms were classified either as *enhanced* (57 farms) or *acceptable* (35 farms). A χ^2 test was performed, and no association was found between overall classification and whether the cows were at pasture or housed indoors χ^2 (2, $n = 92$) = 0.04, $P = 0.84$. The distribution across the classification thresholds for the principle scores is shown in Table 2. The composition of *Absence of prolonged thirst* is shown in Table 3.

Figure 1 shows that, the predictive ability of the models ranged from 56% to 90%. *Absence of prolonged thirst* alone was able to correctly classify farms 88% of the time. *Positive emotional state*, the term given to the version of QBA included in the Welfare Quality® protocol, was less effective, and on its own, it only achieved 67% predictive accuracy. When included with *Absence of prolonged thirst,* it did not improve the predictive ability of the model. Intuitively, one might expect a longer assessment time to result in a higher degree of accuracy, however, as shown in Figure 2 this was not the case. Table 4 shows the best performing models including and excluding *Absence of prolonged thirst* for correctly predicting overall classification. The models including *Absence of prolonged thirst* show both good sensitivity and specificity. The relatively lower specificity suggests that there is more chance of farms being underscored and classed as *acceptable* when they are *enhanced*, than farms being overscored and being classed as *enhanced* when they are in fact *acceptable.* The sensitivity and specificity of the models depends on the threshold.

In response to the original research question, the model including only *Absence of prolonged thirst* would provide the shortest assessment time with a high degree of accuracy. In terms of the costs associated with misclassification, of the most accurate models, this model has the smallest risk of over-scoring, thus maximising the chances of correctly

Table 2 *Distribution of Welfare Quality® principle scores according to thresholds for overall classification from an assessment of 92 farms in England and Wales*[1]

	Principle score			
Welfare Quality® principle	⩽15	>15 and ⩽50	>50 and ⩽75	>75 and ⩽100
Good feeding	23	12	35	22
Good housing	0	10	75	7
Good health	1	68	23	0
Appropriate behaviour	0	48	44	0

[1]Welfare Quality® Overall classifications are calculated using the following rules: excellent: score >50 on all principles, and >75 on two; enhanced: score >15 on all principles, and >50 on two; acceptable: score >5 on all principles, and >15 on two; not classified: failure to meet the above criteria.

Table 3 *Decision-tree responses used to calculate the Welfare Quality® criteria 'Absence of prolonged thirst' for 92 dairy farms in England and Wales*[1]

Decision tree responses according to number of farms				
Sufficient number of drinkers? (yes/no/partly)	Drinkers clean? (yes/no)	At least 2 drinkers per cow? (yes/no)	Final score for Absence of prolonged thirst	Number of farms
44 (Yes)	43 (Yes)	38 (Yes)	100	38
		5 (No)	60	5
	1 (No)	na	32	1
25 (Partly)	24 (Yes)	22 (Yes)	60	22
		2 (No)	40	2
	1 (No)	na	20	1
23 (No)	na	na	3	23

[1]According to the Welfare Quality protocol (Welfare Quality® 2009), *Absence of prolonged thirst* is assessed at the group level, whereby the lowest group score is recorded, provided that that group contains at least 15% of the cows in the herd, otherwise the group with the next lowest score is considered. For trough drinkers, a sufficient number of drinkers is defined as having at least 6 cm of trough per cow, partly sufficient is defined as having at least 4 cm of trough per cow.

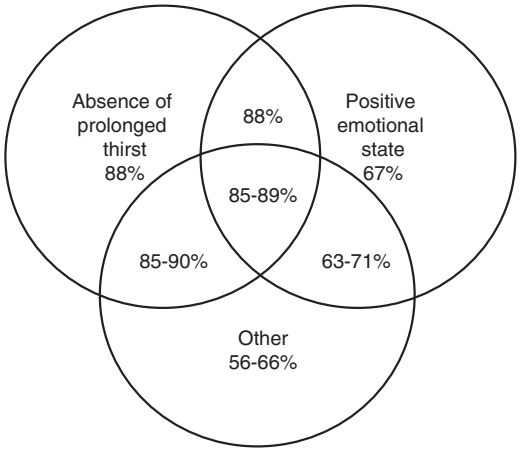

Figure 1 Accuracy of logistic regression models for predicting the overall Welfare Quality® classification of 46 'unseen' randomly sampled dairy farms in England and Wales, highlighting models which include the variables *Absence of prolonged thirst* and *Positive emotional state*.

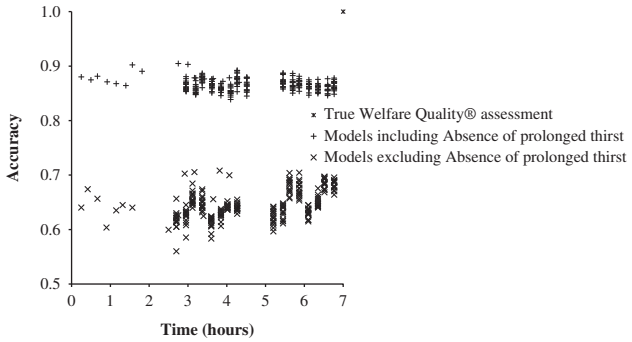

Figure 2 Accuracy of models, including and excluding the variable *Absence of prolonged thirst*, in predicting the overall Welfare Quality® classification tested on 46 randomly sampled, 'unseen' farms in England and Wales, relative to the true Welfare Quality® assessment.

identifying farms with lower levels of welfare. However, several farms shown in Table 5 were consistently misclassified. The mean misclassification rate of the remaining farms was very small (0.0004).

In addition to modelling the overall classification, it is interesting to examine which measures are important in driving the principle level scores. Table 6 shows that all principles could be modelled effectively. In contrast to what was observed at the overall level, at the principle level, *Positive emotional state*, does appear to be important in determining *Appropriate behaviour*, with its removal from the models being associated with a decrease in predictive ability from 0.824 to 0.549 (not shown). In turn, Table 7 shows *Good feeding* is associated with the most accurate models for predicting the overall classification. These results strongly emphasise the role of *Absence of prolonged thirst* in the Welfare Quality® overall classification.

Discussion

The aim of this study was to investigate potential 'iceberg indicators', with the overall Welfare Quality® classification as the gold standard, as a means of reducing the length of time taken to carry out a Welfare Quality® assessment on dairy cows. As with a previous study by de Vries *et al.* (2013a), no *excellent* farms were found in the sample, and the analysis here was limited to *acceptable* and *enhanced* farms. The results show that by only measuring *Absence of prolonged thirst* (Welfare Quality® criteria score) for an assessment time of 15 min, the correct welfare classification can be achieved 88% of the time. While individual parameters within the Welfare Quality® protocol have been shown to be relevant to animal welfare, including those associated with *Absence of prolonged thirst*, the finding that a

Table 4 *Most accurate models including and excluding the Welfare Quality® criteria score 'Absence of prolonged thirst' for predicting the overall Welfare Quality® classification*

	Training[1]	Test[2]		
Most accurate models for predicting overall Welfare Quality® classification	Acc	Acc	Se	Sp
Including Absence of prolonged thirst				
Absence of prolonged thirst + % of cows colliding[3]	0.934	0.905	0.981	0.830
Absence of prolonged thirst + % of cows colliding + use of analgesics for dehorning	0.943	0.903	0.976	0.836
Absence of prolonged thirst + % of cows colliding + % cows with at least one hairless patch no lesion/swelling	0.931	0.887	0.953	0.829
Absence of prolonged thirst + Positive emotional state + % severely lame cows	0.948	0.887	0.946	0.838
Absence of prolonged thirst + % severely lame cows + % of cows colliding	0.946	0.887	0.945	0.842
Excluding Absence of prolonged thirst				
Positive emotional state + % of cows colliding + use of analgesics for dehorning	0.784	0.705	0.878	0.451
Positive emotional state + % of severely lame cows + % of cows colliding + use of analgesics for dehorning	0.798	0.704	0.861	0.479
Positive emotional state + % severely lame cows + % of cows colliding	0.790	0.704	0.853	0.489
Positive emotional state + % of cows colliding	0.773	0.703	0.861	0.469
Positive emotional state + % severely lame cows + % of cows colliding + % moderately lame cows	0.793	0.695	0.843	0.477

Acc = accuracy, the proportion of times that, as part of the repeated process of building and testing the models on different random draws of 46 farms, the calculated Welfare Quality® classification is predicted; Se = sensitivity; Sp = specificity.
[1]Training = the training set corresponds to a random draw of 46 farms which were used to build the models.
[2]Test = the test set corresponds to the remainder 46 farms excluded from the training set.
[3]% of cows colliding with housing equipment.

Table 5 *Summary of individual 'unseen' farms that were consistently misclassified at a rate of >0.50 when predicting the overall classification using only 'Absence of prolonged thirst' as a single variable in the model*

	Calculated Welfare Quality® aggregated scores					
	Good feeding criteria scores		Welfare Quality® principle scores			
Misclassification rate: Modelled under (−) or over (+) classification	Absence of prolonged thirst	Absence of prolonged hunger	Good feeding	Good health	Appropriate behaviour	Good housing
1 (+)	100	82	87	43	48	47
1 (+)	60	75	62	34	17	47
1 (+)	100	70	78	48	41	50
1 (+)	60	44	48	35	32	59
1 (+)	60	76	62	31	26	37
1 (+)	60	30	38	23	42	74
1 (+)	60	44	49	53	37	47
1 (+)	100	31	50	14	40	59
0.998 (+)	60	38	44	42	35	75
0.801 (−)	32	78	38	52	75	54
0.731 (+)	40	72	44	23	42	54
0.728 (+)	40	32	34	29	47	49
0.515 (+)	20	77	27	29	41	71

Table 6 *Summary of most accurate models for predicting Welfare Quality® principle-level scores*

	Accuracy[1]	
Most accurate models for predicting Welfare Quality® principle-level scores (LDA/LR)	Training[2]	Test[3]
Appropriate behaviour (LR)		
Positive emotional state + Mean frequency of displacements[4] + % cows that cannot be approached[5] + other behaviour[6]	0.972	0.902
Positive emotional state + butts[7] + mean frequency of displacements + % cows that cannot be approached + other behaviour	0.992	0.900
Positive emotional state + % cows that cannot be approached + other behaviour	0.956	0.899
Good feeding (LDA)		
% very lean cows + Absence of prolonged thirst	0.910	0.874
Absence of prolonged thirst	0.729	0.688
% very lean cows	0.514	0.439
Good housing (LDA)		
Mean time to lie down + % of cows with dirty lower legs + % of cows with dirty flank	0.897	0.845
Mean time to lie down + % of cows colliding[8] + % of cows lying outside the lying area + % of cows with dirty flank	0.921	0.842
Mean time to lie down + % of cows with dirty lower legs	0.879	0.839
Good health (LDA)		
% Moderately lame cows + % severely lame cows + % cows with at least one lesion/swelling + % cows with ocular discharge + % mastitis + % dystocia + % downer cows + % mortality	0.928	0.845
% Moderately lame cows + % severely lame cows + % cows with at least one lesion/swelling + % cows with ocular discharge + % dystocia + % downer cows	0.907	0.843
% Moderately lame cows + % severely lame cows + % cows with at least one lesion/swelling + % cows with ocular discharge + % dystocia + % downer cows + % mortality	0.915	0.843

LDA = linear discriminant analysis; LR = logistic regression.
[1]Accuracy is the proportion of times that, as part of the repeated process of building and testing the models on different random draws of 46 farms, the calculated Welfare Quality® classification is predicted.
[2]Training = the training set corresponds to a random draw of 46 farms which were used to build the models.
[3]Test = the test set corresponds to the remainder 46 farms excluded from the training set.
[4]Average frequency of displacements per cow per hour.
[5]% cows that cannot be approached as closely as 100 cm.
[6]Expression of other behaviour.
[7]Average frequency of butts per cow per hour.
[8]% of cows colliding with housing equipment.

Table 7 *Principle-level models for predicting the Welfare Quality® overall classification*

	Training[1]	Test[2]		
Models Welfare Quality® overall classification	Acc	Acc	Se	Sp
Good feeding + appropriate behaviour + good health	0.969	0.915	0.955	0.901
Good feeding + appropriate behaviour + good housing + good health	0.950	0.914	0.946	0.915
Good feeding + appropriate behaviour + good housing	0.957	0.900	0.944	0.879
Good feeding + appropriate behaviour	0.965	0.899	0.943	0.873
Good feeding + good housing + good health	0.917	0.878	0.927	0.842
Good feeding + good health	0.923	0.874	0.943	0.802
Good feeding + good housing	0.920	0.862	0.926	0.803
Good feeding	0.927	0.861	0.921	0.812
Appropriate behaviour + good housing + good health	0.744	0.653	0.789	0.448
Appropriate behaviour + good housing	0.749	0.640	0.783	0.422
Appropriate behaviour + good health	0.735	0.638	0.797	0.388
Appropriate behaviour	0.720	0.633	0.772	0.410
Good housing + good health	0.679	0.610	0.873	0.191
Good health	0.684	0.609	0.853	0.211
Good housing	0.706	0.603	0.812	0.272

Acc = accuracy, the proportion of times that, as part of the repeated process of building and testing the models on different random draws of 46 farms, the calculated Welfare Quality® classification is predicted; Se = sensitivity; Sp = specificity.
[1]Training = the training set corresponds to a random draw of 46 farms which were used to build the models.
[2]Test = the test set corresponds to the remainder 46 farms excluded from the training set.

single, resource-based score can predict well the overall result is a significant challenge to the published Welfare Quality® aggregation system. Concern over the poor discriminatory ability of the overall Welfare Quality® classification has recently led Eerdenburg (2013) to propose an adapted version of the measure aggregation system which included, among other changes, the recalibration of the score for *Absence of prolonged thirst* for what was considered to be an improvement in the overall classification of farms.

Missing data were an issue in this study, and are indeed an issue for the Welfare Quality® measure aggregation system in general (Heath *et al*., 2014). The methods used here to fill in the missing values were often fairly simple (mean imputation) and could potentially therefore have reduced the variation in some measures between farms. For example, in the case of the measure *Length of troughs*: where the measure was available for other groups on the farm, then the mean of those groups was used, however, where the measure was not available for any group on the farm, then the mean for the farm assurance scheme was used. Filling in a measure as the same for all groups on a farm (which occurred for 14% of farms for this measure), clearly reduces variability within farm, and is a deficiency of this method of imputation, although it should be noted that 68% of farms only had one group and so were unaffected by this imputation. In order to account for missing data at both the herd and group level, multi-level multiple imputation would be a more sophisticated alternative (see e.g., Goldstein *et al*., 2014) which is not so susceptible to the same issues of reduction in variation. However, it was considered that the application of such a method to fill in the missing data would not have added anything to the illustrative purposes of this study and may have instead lost some of the readership through its added complexity.

Returning to the original research question, as an iceberg indicator for Welfare Quality®, *Absence of prolonged thirst* could be argued to have a multi-dimensional impact on welfare, albeit as a resource-based measure. Water is the most important nutrient for dairy cows, essential for health and productivity, (National Research Council, 2001), and water deprivation is associated with increased aggression and less time spent lying down (Little *et al*., 1980). Given that water intake may be influenced by the dimensions of the drinkers (Machado Filho *et al*., 2004; Teixeira *et al*., 2006), and water flow rate (Andersson *et al*., 1984), the measures included for *Absence of prolonged thirst*, reflect a fundamental aspect of welfare. Furthermore, *Absence of prolonged thirst* could also be considered to be a proxy measure of management standards, such that the care that a farmer might show in providing a sufficient number of clean drinkers might be expected to be found in other areas relating to welfare. So, while adequate water provision is an essential element in preventing poor welfare, as a resource-based measure, it is unable to provide an actual account of the cow's thirst, and therefore, might be better considered as part of a risk-based assessment. As thirst is influenced by other factors including climate, milk yield, dry matter intake (Cardot *et al*., 2008), and social hierarchy (Andersson *et al*., 1984), the measure also risks either being overly penalising or overly lenient for farms at the extreme ends of these variables. This is increased by the method of score calculation which assigns discrete scores by means of a decision tree. Furthermore, the validity of this measure has been questioned by Tuyttens *et al*. (2013) who reported that the Welfare Quality® criteria score for *Absence of prolonged thirst* in broilers was not found to be associated with a corresponding animal-based measure.

Positive emotional state, on the other hand, is a holistic, animal-based measure, which might have been expected to have a greater ability to discriminate farms than it did. QBA can take one of two forms, the free-choice profiling format where descriptive terms are elicited from the assessors themselves, or the fixed term format where designated terms are rated by assessors according to what they find, and is the form found in the Welfare Quality® protocols. Compared to the free-choice profiling format, where QBA has been associated with physiological welfare indicators in cattle and steers (Stockman *et al.*, 2011 and 2012), and induced emotional states in pigs (Rutherford *et al.*, 2012), QBA as part of the Welfare Quality® protocol has not been found to be associated with other measures in veal calves (Brscic *et al.*, 2009), or beef cattle (Kirchner *et al.*, 2012). In terms of inter-assessor agreement, the evidence is also mixed for the fixed term format with high levels of inter-observer agreement having been reported in sheep (Phythian *et al.*, 2013), but only 'slight to moderate' agreement for Welfare Quality® QBA in dairy cows (Bokkers *et al.*, 2012). In addition, QBA has been found to be influenced by time of day (Schwed, 2013), though perhaps this can be understood as relating to changes in the environment, which have been found to be reflected by QBA in veal calves (Brscic *et al.*, 2009) and similarly, by pigs in enriched environments (Mullan *et al.*, 2011).

Nevertheless, *Positive emotional state* is the only measure of positive welfare included in the Welfare Quality® protocol. Recent advances propose that welfare should move beyond the focusing on the absence of negative states to also encompass positive experiences or positive welfare (Boissy *et al.*, 2007; Yeates and Main, 2008; Farm Animal Welfare Council, 2009). If welfare is thought of as a continuum from negative welfare at one end to positive welfare at the other, and if *Absence of prolonged thirst* is understood to be only associated with thirst and is the driver for the overall classification, the implication is that *excellent* farms can only be considered to reflect a neutral state of welfare, as an absence of suffering does not signify positive welfare (Yeates and Main, 2008). However, if *Absence of prolonged thirst* is instead considered to be a proxy measure for management, then it may indeed be suggestive of higher levels of welfare. *Positive emotional state*, in theory, had the potential to extend the scope of the Welfare Quality® protocol to include positive welfare, however in practice this was not observed, either because of the weightings assigned to the measure, because of issues arising from inter observer reliability, or because of limited variability. While this study did not look at the individual QBA descriptors, interestingly, de Vries *et al.* (2013a) found *acceptable* farms had lower scores than *enhanced* farms for the terms 'happy' and 'relaxed'. With an absence of any further feasible animal-based measures of positive welfare, behavioural opportunities have been proposed by the Farm Animal Council as being important for a 'good life' for farm animals (Farm Animal Welfare Council, 2009). This recommendation is reflected in the 'citizen juries' conducted as part of the Welfare Quality® project, where members of the public considered that the classification *excellent* should only be applied to farms which had extensive systems of production with outdoor access (Miele *et al.*, 2010). A measure of outdoor access is included in the Welfare Quality® protocol, in the form of the criteria *Expression of other behaviour*, however in this study, this criteria was not correlated highly enough with the overall classification for inclusion in the models. This suggests that the potential for this measure to differentiate farms according to levels of positive welfare is being underutilised due to the way in which the measures have been aggregated.

The aim of this paper was to investigate possible iceberg indicators as a means of reducing the amount of time taken for an assessment. Using the Welfare Quality® overall classification as a gold standard, *Absence of prolonged thirst* (Welfare Quality® criteria) was able to discriminate between farms to a high degree of accuracy and reduce the amount of time taken for an assessment to 15 min, suggestive of its role as an iceberg indicator. However, by using the overall classification as the gold standard: the outcome of the measure aggregation system, the findings only reflect the relative weightings that have been assigned to different measures as a proxy for overall welfare state. That *Absence of prolonged thirst* has been shown in this study, to have such a deterministic role in the overall classification, suggests that the outcome-based, multi-dimensional assessment of welfare that Welfare Quality® aimed to provide, may be compromised by the system of measure aggregation. While certain 'challenges' have been identified with using the measure aggregation system (Heath *et al.*, 2014), the associated weightings, which derive from both expert opinion and the application of mathematical techniques, have been criticized by de Vries *et al.* (2013a). Therefore, it is the opinion of the authors that rather than evidence of a credible iceberg indicator, the prominent role of *Absence of prolonged thirst*, in driving the overall classification, represents an unintended consequence of the measure aggregation system, which could potentially be improved with further development.

Acknowledgements

Funding for this project was provided by the Agriculture and Horticulture Development Board DairyCo division and AssureWel. The AssureWel project is a collaboration between the RSPCA, the Soil Association and the University of Bristol, which is supported financially by the Tubney Charitable Trust. The authors are very grateful for the assistance provided by the following individuals for their part in data collection: Professor Christoph Winckler; Iain Rogerson; Alison Bond; Anna Fraser; the farm assurance assessors; Milk Link Ltd; Dairy Crest Ltd; PAI, and indeed the farmers who took part. They would also like to acknowledge Dr Isabelle Veissier for her help with inaccuracies associated with the calculations in the printed protocol, and Dr Jenny Gibbons for help with drafting.

References

Andersson M, Schaar J and Wiktorsson H 1984. Effects of drinking water flow rates and social rank on performance drinking behaviour of tied-up dairy cows. Livestock Production Science 11, 599–610.

Blokhuis HJ, Veissier I, Miele M and Jones B 2010. The Welfare Quality® project and beyond: Safeguarding farm animal well-being. Acta Agriculturae Scandinavica, Section A – Animal Science 60, 129–140.

Boissy A, Manteuffel G, Jensen MB, Moe RO, Spruijt B, Keeling LJ, Winckler C, Forkman B, Dimirov I, Langbein J, Bakken M, Veissier I and Aubert A 2007. Assessment of positive emotions in animals to improve their welfare. Physiology and Behaviour 92, 375–397.

Bokkers EAM, de Vries M, Antonissen ICMA and de Boer IJM 2012. Inter- and intra-observer reliability of experienced and inexperienced observers for the Qualitative Behaviour Assessment in dairy cattle. Animal Welfare 21, 307–318.

Brscic M, Wemelsfelder F, Tessitore E, Gottardo F, Cozzi G and Van Reenen CG 2009. Welfare assessment: correlations and integration between a Qualitative Behavioural Assessment and a clinical/health protocol applied in veal calves farms. Italian Journal of Animal Science 8, 601–603.

Cardot V, Le Roux Y and Jurjanz S 2008. Drinking behavior of lactating dairy cows and prediction of their water intake. Journal of Dairy Science 91, 2257–2264.

DairyCo 2013. Dairy statistics an insiders guide 2013. The Agriculture and Horticulture Development Board, Kenilworth, UK.

Dawkins MS 2003. Behaviour as a tool in the assessment of animal welfare. Zoology 106, 383–387.

de Vries M, Bokkers EAM, van Schaik G, Botreau B, Engel B, Dijkstra T and de Boer IJM 2013a. 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.

de Vries M, Bokkers EAM, van Schaik G, Engel B and de Boer IJM 2013b. Exploring the value of routinely collected herd data for estimating dairy cattle welfare. Journal of Dairy Science 97, 715–730.

de Vries M, Engel B, den Uijl I, van Schaik G, Dijkstra T, de Boer IJM and Bokkers EAM 2013c. Assessment time of the Welfare Quality® protocol for dairy cattle. Animal Welfare 13, 85–93.

Eerdenburg 2013. On-farm comparison of the Welfare Quality® resource-based versus an animal-based measure of thirst in broiler chickens. In Abstracts for WQ Network Workshop 11th December, Lille, France, 10pp.

Farm Animal Welfare Council 2005. Report on the Welfare Implications of Farm Assurance Schemes. FAWC, London, UK.

Farm Animal Welfare Council 2009. Farm animal welfare in Great Britain: past, present and future. FAWC, London, UK.

Fraser AF and Broom DM 1990. Farm Animal Behaviour and Welfare. Saunders, New York, USA.

Goldstein H, Carpenter J and Browne WJ 2014. Fitting multilevel multivariate models with missing data in responses and covariates that may include interactions and nonlinear terms. Journal of Royal Statistical Society, Series A. 177, 553–564.

Heath C, Lin Y, Browne WJ, Mullan S and Main DCJ 2014. Implementing Welfare Quality® in UK assurance schemes: evaluating the challenges. Animal Welfare 23, 95–107.

Kirchner MK, Schulze Westerath-Nicklaus H, Gutmann A, Pfeiffer C, Elena T, Giulio C, Knierim U and Winckler C 2012. Qualitative Behaviour Assessment is independent from other parameters used in the Welfare Quality® assessment system for beef cattle. In Proceedings of the 46th Congress of the International Society for Applied Ethology, 31st July to 4th August, Vienna, Austria, 79pp.

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

Little W, Collis KA, Gleed PT, Sansom BF, Allen WM and Quick AJ 1980. Effect of reduced water intake by lactating dairy cows on behaviour, milk yield and blood composition. Veterinary Record 106, 547–551.

Machado Filho LCP, Teixeira DL, Von Keyserlingk MAG, Weary DM and Hötzel MJ 2004. Designing better water troughs: dairy cows prefer and drink more from larger troughs. Applied Animal Behaviour Science 89, 185–193.

Miele M, Evans A and Higgin M (ed.) 2010. Dialogue between citizens and experts regarding farm animal welfare: citizen juries in the UK, Norway, the Netherlands and Italy, Welfare Quality Reports, 16 vols. Cardiff School of City and Regional Planning, Cardiff, UK.

Mullan S, Edwards SA, Butterworth A, Whay HR and Main DCJ 2011. A pilot investigation of possible positive system descriptors in finishing pigs. Animal Welfare 20, 439–449.

National Research Council 2001. Nutrient Requirements of Dairy Cattle, 7th edition. National Academy Press, Washington, DC, USA.

Phythian C, Nichalopoulou E, Duncan J and Wemeldsfelder F 2013. Inter-observer reliabilty of Qualitative Behavioural Assessments of sheep. Applied Animal Behaviour Science 144 (s 1–2), 73–79.

Rutherford KMD, Donald RD, Lawrence AB and Wemelsfelder F 2012. Qualitative Behavioural Assessment of emotionality in pigs. Applied Animal Behaviour Science 139, 218–224.

Schwed B 2013. Intra-day variation of Qualitative Behaviour Assesement outcomes in dairy cattle. In Proceedings of UFAW International Animal Welfare Science Symposium, 4th to 5th July 2013, Universitat Autonoma de Barcelona, Barcelona, Spain, 91pp.

Stockman CA, Collins T, Barnes AL, Miller D, Wickham SL, Beatty DT, Blache D, Wemelsfelder F and Fleming PA 2011. Qualitative Behavioural Assessment and quantitative physiological measurement of cattle naïve and habituated to road transport. Animal Production Science 51, 240–249.

Stockman CA, McGilchrist P, Collins T, Barnes AL, Miller D, Wickhama SL, Greenwood PL, Cafe LM, Blache D, Wemelsfelder F and Fleming PA 2012. Qualitative Behavioural Assessment of Angus steers during pre-slaughter handling and relationship with temperament and physiological responses. Applied Animal Behaviour 142, 125–133.

Teixeira DL, Hötzel MJ and Machado Filho LCP 2006. Designing better water troughs: 2. Surface area and height, but not depth influence dairy cows' preference. Applied Animal Behaviour Science 96, 169–175.

Tuyttens FAM, Vanderhasselt RF, Federici JF, Sans ECO, Molento CFM, Goethals K, Buijs S and Duchateau L 2013. On-farm comparison of the Welfare Quality® resource-based versus an animal-based measure of thirst in broiler chickens. In Abstracts for WQ Network Workshop, 11th December, Lille, France, 4p.

Webster AJF, Main DCJ and Whay HR 2004. Welfare assessment: indices from clinical observation. Animal Welfare 13, 93–98.

Welfare Quality® 2009. Welfare Quality® assessment protocol for cattle. Welfare Quality® Consortium, Lelystad, The Netherlands.

Wemelsfelder F 2007. How animals communicate quality of life: the qualitative assessment of behaviour. Animal Welfare 16, 25–31.

Wemelsfelder F, Hunter TEA, Mendl MT and Lawrence AB 2001. Assessing the 'whole animal': a free choice profiling approach. Animal Behaviour 62, 209–220.

Yeates JW and Main DCJ 2008. Assessment of positive welfare: a review. The Veterinary Journal 175, 293–300.