© 2013 Universities Federation for Animal Welfare
The Old School, Brewhouse Hill, Wheathampstead,
Hertfordshire AL4 8AN, UK
www.ufaw.org.uk

Animal Welfare 2013, 22: 85-93
ISSN 0962-7286
doi: 10.7120/09627286.22.1.085

Assessment time of the Welfare Quality® protocol for dairy cattle

M de Vries*[†], B Engel[‡], I den Uijl[§], G van Schaik[§], T Dijkstra[§], IJM de Boer[†] and EAM Bokkers[†]

[†] Animal Production Systems Group, Wageningen University, PO Box 338, 6700 AH Wageningen, The Netherlands
[‡] Biometris, Wageningen University, Wageningen, The Netherlands
[§] GD Animal Health Service, Deventer, The Netherlands
* Contact for correspondence and requests for reprints: marion.devries@wur.nl

Abstract

The Welfare Quality® (WQ) protocols are increasingly used for assessing welfare of farm animals. These protocols are time consuming (about one day per farm) and, therefore, costly. Our aim was to assess the scope for reduction of on-farm assessment time of the WQ protocol for dairy cattle. Seven trained observers quantified animal-based indicators of the WQ protocol in 181 loose-housed and 13 tied Dutch dairy herds (herd size from 10 to 211 cows). Four assessment methods were used: avoidance distance at the feeding rack (ADF, 44 min); qualitative behaviour assessment (QBA, 25 min); behavioural observations (BO, 150 min); and clinical observations (CO, 132 min). To simulate reduction of on-farm assessment time, a set of WQ indicators belonging to one assessment method was omitted from the protocol. Observed values of omitted indicators were replaced by predictions based on WQ indicators of the remaining three assessment methods, resources checklist, and interview, thus mimicking the performance of the full WQ protocol. Agreement between predicted and observed values of WQ indicators, however, was low for ADF, moderate for QBA, slight to moderate for BO, and poor to moderate for CO. It was concluded that replacing animal-based WQ indicators by predictions based on remaining WQ indicators shows little scope for reduction of on-farm assessment time of the Welfare Quality® protocol for dairy cattle. Other ways to reduce on-farm assessment time of the WQ protocol for dairy cattle, such as the use of additional data or automated monitoring systems, should be investigated.

Keywords: *animal welfare, dairy cows, on-farm assessment, prediction, protocol, Welfare Quality®*

Introduction

The use of animal-based indicators is gaining increased preference over resource- and management-based indicators in farm animal welfare assessment schemes. Animal-based indicators, which measure the state of the animal rather than its environment, are assumed to possess a higher validity than resource- and management-based indicators because they are more closely linked to the actual welfare state of animals (Webster *et al* 2004; Blokhuis *et al* 2010). Duration of assessing animal-based indicators on-farm, however, is a main constraint with regard to feasibility (Mülleder *et al* 2007; Knierim & Winckler 2009; Blokhuis *et al* 2010). In the Welfare Quality® (WQ) protocol for dairy cattle, for example, 60% of the indicators are animal-based, but take about 90% of the total on-farm assessment time (depending on herd size; Welfare Quality® 2009). Consequently, on-farm assessment time of the WQ protocol ranges from about 4.4 to 7.7 h for herds of 25 to 200 cows (Welfare Quality® 2009). Assessment time and associated costs of on-farm assessments may hamper the practical implementation of the WQ protocol in welfare audit programmes (Knierim & Winckler 2009).

Various studies have shown associations between indicators of dairy cattle welfare. Lame cows, for instance, were associated with a lower body condition and changes in lying behaviour (Bowell *et al* 2003; Ito *et al* 2010; Blackie *et al* 2011). Also, a higher frequency of agonistic behaviour in dairy herds was associated with larger avoidance distances towards cows (Waiblinger *et al* 2003). Although these associations may not always involve causal relationships, it suggests that animal-based indicators may have potential to predict other animal-based indicators. Such predictions could replace on-farm observations, and reduce on-farm assessment time of the WQ protocol. So far, mainly resource- and/or management-based indicators have been considered for prediction of animal-based indicators (eg Mülleder *et al* 2007).

Two out of four assessment methods in the WQ protocol contain more than one animal-based indicator (Welfare Quality® 2009): behavioural observations (BO; six indicators), and clinical observations (CO; 13 indicators). When an indicator belonging to one of these assessment methods is replaced, cows still need to be observed to

Table 1 Descriptive statistics of Welfare Quality® indicators collected using a resources checklist or interview.

Assessment method	Resource- and management-based indicators (categorical)	Category (n herds)
Resources checklist	Type of housing	Loose (181), tied (13)
	Sufficient number of drinkers	Yes (97), partly (64), no (33)
	Clean drinkers[1]	Yes (192), no (2)
	At least two drinkers per cow	Yes (177), no (17)
Interview	Access to pasture (with at least 6 h per day)	Yes (145), no (49)
	Releasing cows from tie stalls for at least 1 h per day in winter[1]	Yes (0), no (13)
	Dehorning young stock (in at least 15% of animals)	Yes (181), no (13)
	Method of dehorning[1]	Chemical (1), thermal (180)
	Use of analgesics[1]	Yes (3), no (178)
	Use of anaesthetics[1]	Yes (173), no (8)
	Dehorning adult cattle (in at least 15% of animals)[1]	Yes (0), no (194)
	Use of analgesics[1]	NA[2]
	Use of anaesthetics[1]	NA[2]
	Tail-docking (in at least 15% of animals)[1]	Yes (0), no (194)
	Method of tail docking[1]	NA[2]
	Use of analgesics[1]	NA[2]
	Use of anaesthetics[1]	NA[2]
	Animal-based indicators (continuous)	
Interview	% on-farm mortality	0.6 (0, 3.1)
	% cows with SCC > 400,000	11.0 (0, 36.3)
	% dystocia	5.0 (0, 50)

[1] Indicator excluded from predictions due to observed prevalence < 5%.
[2] NA: not applicable.

collect data for the other WQ indicators, which takes an equal (BO), or only slightly less (CO) amount of time. Hence, all indicators of an assessment method should be considered together in order to reduce assessment time.

Our aim was to evaluate the performance of a reduced protocol, in which a set of WQ indicators belonging to one assessment method is replaced by predictions based on remaining animal-, resource- and management-based indicators, in order to assess the scope for reduction of on-farm assessment time of the WQ protocol for dairy cattle.

Materials and methods

Herd selection

To properly assess the scope for prediction of animal-based WQ indicators, we aimed for data from herds that span a wide range of levels of animal welfare. Therefore, herds were selected based on a composite health score. From 5,000 Dutch herds participating in a health scheme of a Dutch dairy co-operative, a composite health score between 0 (worst) and 50 (best) was determined over the period January 2008 to June 2009. This score consisted of five parameters that have been shown to correlate with different WQ indicators (De Vries *et al* 2011): cow and young stock mortality, bulk tank milk somatic cell count (SCC), new udder infections, and fluctuations in standardised milk production. Herds were attributed zero points per parameter when the parameter value was among the 10% worst, and 10 points when it was among the 90% best values of all dairy herds in 2004.

To ensure a minimum sample of 100 herds from the 5% lowest composite health scores and 100 herds from the rest of the population, 250 herds were randomly selected from each of these respective categories. Of the selected herds, 163 farmers responded positively, 75 negatively and 262 failed to respond. Due to the insufficient positive response rate, non-responders were further contacted by telephone. Finally, 196 farmers agreed to participate: 90 from the 5% lowest composite health scores, and 106 from the rest of the population. Composite health scores of the participating herds (median = 40, 95% range = 27.5 to 50) were similar to the original selection of 500 herds (median = 35, 95% range = 27.5 to 50).

https://doi.org/10.7120/09627286.22.1.085 Published online by Cambridge University Press

Table 2(a) Observed and predicted prevalence and agreement (Cohen's kappa, positive [PR] and negative rate [NR] with 95% confidence intervals [CI]) between observed and predicted values of categorical animal-based indicators assessed in behavioural observations (BO), and clinical observations (CO).

Method	Indicator	Problems (n herds)						Agreement[2]		
		Observed			Predicted[1]			κ	PR	NR
		Minor	Moderate	Severe	Minor	Moderate	Severe		95% (CI)	95% (CI)
BO	Mean time to lie down (s)	41	75	78	10	92	88	0.14	97% (92–99)	12% (4–26)
	% cows colliding with stall components	81	23	90	90	0	102	0.44	72% (62–80)	73% (62–83)
	% cows lying outside lying area	152	17	25	183	0	10	0.19	15% (6–29)	97% (93–99)
	Frequency coughing per cow per 15 min[3]	194	0	0	–	–	–	–	–	–
CO	% cows with dirty hind legs	15	28	151	0	0	194	0.00	100% (98–100)	0% (0–22)
	% cows with dirty udder	80	45	69	132	0	60	0.25	41% (32–51)	83% (72–90)
	% cows with dirty hindquarters	28	24	142	1	0	193	0.07	100% (98–100)	4% (0–18)
	% cows with ocular discharge	170	16	8	194	0	0	0.00	0% (0–7)	100% (97–100)
	% cows with nasal discharge	145	27	22	193	0	0	0.00	0% (0–14)	100% (98–100)
	% cows with diarrhoea	126	20	48	191	0	2	–0.03	0% (0–5)	98% (94–100)
	% cows with vulvar discharge	149	31	14	192	0	2	0.03	2% (0–12)	99% (96–100)
	% cows with hampered respiration[3]	190	4	0	–	–	–	–	–	–

[1] Some herds excluded because highest predicted odds were equal for two or more categories.
[2] Results based on two classes: 'minor problem' and 'moderate or severe problem'.
[3] Indicator excluded from predictions due to observed prevalence < 5%.

Farm visits

Seven observers, all with previous experience in dairy production and handling, were trained to use the Welfare Quality® assessment protocol for dairy cattle (Welfare Quality® 2009) in a three-day course given by delegates of the Welfare Quality® consortium. Observers visited 14 to 48 herds during the winter months of November 2009 through to March 2010 when the cows had been denied access to pasture for at least two weeks. During a farm visit, observers collected data for 17 resource- and management-based (Table 1) and 24 animal-based (Tables 1 and 2[a], [b]) WQ indicators in six assessment methods. Assessment methods, which were executed in a fixed order, are described briefly (details can be found in Welfare Quality® [2009]) below.

For avoidance distance at the feeding rack (ADF), which was measured on a pre-defined sample of lactating and dry cows (Welfare Quality® 2009), individual cows were approached from a distance of 2 m on the feed bunk. The avoidance distance was estimated at the moment the cow moved back, turned, or pulled back the head, and was categorised in one of four categories: > 100 cm, 100 to > 50 cm, 50 to > 0 cm, or touched. For the Qualitative Behaviour Assessment (QBA), cows were observed in segments of the barn for 20 min, regardless of the number of cows in the herd or in a segment. After this observation, 20 descriptors were scored on a visual analogue scale between 0 (expressive quality of the descriptor was entirely absent in any of the animals) and 125 mm (dominant across all observed animals). For BO, lying behaviour, agonistic behaviour, and coughing was recorded in segments (with a maximum of approximately 25 lactating cows) using continuous behaviour sampling (Martin & Bateson 1993). For CO, 13 health indicators (Table 2[a], [b]) were assessed for a pre-defined sample of lactating and dry cows. Body condition was scored on a five-point scale, and grouped into classes 'very lean' (score 1) and 'not very lean' (score ≥ 2). Locomotion was scored on a five-point scale, and grouped into classes 'not lame' (scores 1 and 2), 'lame' (score 3) and 'severely lame' (scores 4 and 5). Assessment details of other indicators of CO can be found in the WQ protocol (2009). Besides this, four resource-based, 13 management-based, and three animal-based indicators (Table 1) were collected using a resources checklist and an interview. Identical indicators were used for cattle in loose housing and tie stalls, except for lameness. Cows in tie stalls were categorised into two lameness classes (not lame or lame), instead of three (not lame, lame or severely lame).

Table 2(b) Difference (y-$\hat{y}$) and Spearman rank correlation (r_s) between observed (y) and predicted ($\hat{y}$) values of continuous animal-based indicators assessed in the avoidance distance at the feeding rack (ADF), Qualitative Behaviour Assessment (QBA), behavioural observations (BO) and clinical observations (CO).

Method	Indicator	y (median [95% range])	$\hat{y}$ (median [95% range])	y–$\hat{y}$ (median [95% range])	r_s
ADF	ADF index	68 (25.6, 92.3)	67.9 (54.7, 76.2)	2.2 (–33.9, 24.2)	0.31
QBA	QBA index	–1.0 (–8.8, 4.6)	–1.2 (–3.8, 2.8)	0.4 (–6.1, 4.1)	0.54
BO	Frequency of head butts per cow per h	0.7 (0.1, 2.8)	0.8 (0.4, 1.4)	–0.1 (–0.8, 1.6)	0.38
	Frequency of displacements per cow per h	0.3 (0, 1.5)	0.4 (0.0, 0.8)	–0.0 (–0.5, 0.8)	0.46
CO	% very lean cows	2.4 (0, 20.0)	3.8 (0.9, 12.0)	–1.2 (–7.6, 15.5)	0.43
	% moderately lame cows	24.1 (3.6, 51.4)	24.1 (14.6, 36.3)	–0.43 (–21.6, 24.0)	0.39
	% severely lame cows[1]	6.0 (0, 28.9)	6.9 (1.8, 24.1)	–1.8 (–11.9, 17.0)	0.50
	% cows with hairless patches	33.3 (3.3, 61.5)	32.8 (21.8, 42.3)	–0.1 (–26.2, 29.9)	0.33
	% cows with lesions or swellings	35.3 (4.6, 94.7)	39.4 (24.3, 72.6)	–4.4 (–30.7, 43.4)	0.49

[1] Prediction concerns only loose-housing systems because severe lameness was not assessed in tie stalls.

Table 3 Threshold values for categorical indicators representing a minor, moderate or severe problem (adapted from Welfare Quality® 2009).

Indicator	Minor problem	Moderate problem	Severe problem
Mean time to lie down (s)	≤ 5.2	< 5.2 and ≤ 6.3	> 6.3
% cows colliding with components of the stall	≤ 20	< 20 and ≤ 30	> 30
% cows lying outside lying area	≤ 3	< 3 and ≤ 5	> 5
% cows with dirty hind legs	≤ 20	< 20 and ≤ 50	> 50
% cows with dirty udder	≤ 10	< 10 and ≤ 19	> 19
% cows with dirty hindquarters	≤ 10	< 10 and ≤ 19	> 19
% cows with ocular discharge	≤ 3	< 3 and ≤ 6	> 6
% cows with nasal discharge	≤ 5	< 5 and ≤ 10	> 10
% cows with diarrhoea	≤ 3.25	< 3.25 and ≤ 6.5	> 6.5
% cows with vulvar discharge	≤ 2.25	< 2.25 and ≤ 4.5	> 4.5
% cows with hampered respiration	≤ 3.25	< 3.25 and ≤ 6.5	> 6.5
Average frequency of coughing per 100 cows and 15 min	≤ 3	< 3 and ≤ 6	> 6

Time needed per assessment method and total assessment time per herd were not recorded during the farm visits, but were estimated based on the information given in the WQ protocol (Welfare Quality® 2009). For this study, on-farm assessment time was estimated for an average Dutch dairy herd (78 lactating cows; LEI 2008). Total estimated assessment time, therefore, was 381 min: 44 for ADF (1 min per animal), 25 for QBA, 150 for BO, 132 for CO (3 min per animal), 15 for the resources checklist, and 15 for the interview.

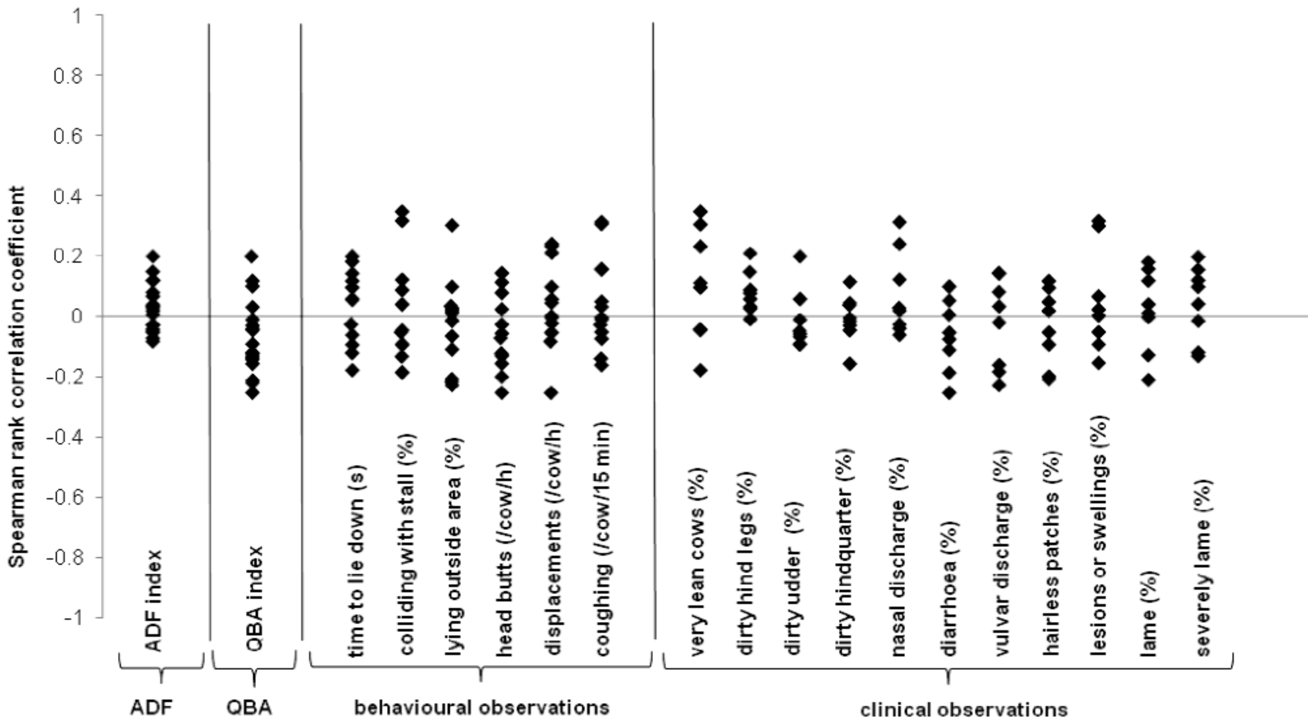
Data processing

Data collected from the herds were expressed as 'WQ indicators' at the herd level, using weights for the aggregation of ADF categories and QBA descriptors, and threshold values for the conversion into ordinal indicators as described in the WQ protocol (2009). The percentage of cows in each ADF category was weighted and aggregated into an 'ADF index' ranging from 0 (worst) to 100 (best). For QBA, the 20 descriptors were weighted and aggregated into a 'QBA index' ranging from –10 (worst) to 7 (best). Data related to lying behaviour, cleanliness, and disease were converted to an ordinal scale representing a minor, moderate, or severe problem (Table 3).

WQ indicators were not included in the statistical analyses when the standard deviation was zero or the prevalence was less than 5%. Because ignorance of missing values can lead to reduced power (Donders *et al* 2006; Dohoo *et al* 2009), multiple imputation (MI) was used to replace missing values. MI is a technique in which a missing value is replaced by a value that was drawn from an estimate of the distribution of this variable (Donders *et al* 2006).

https://doi.org/10.7120/09627286.22.1.085 Published online by Cambridge University Press

Figure 1



Spearman rank correlation coefficients per animal-based indicator when compared with indicators belonging to different assessment methods (avoidance distance at the feeding rack [ADF], Qualitative Behaviour Assessment [QBA], behavioural observations, or clinical observations).

Statistical analysis

Spearman rank correlations between animal-based WQ indicators were calculated. They were preferred over Pearson correlations, because a number of variables could not be assumed to be (approximately) normally distributed. Subsequently, individual animal-based WQ indicators of each of the four assessment methods were predicted, using WQ indicators of the remaining three assessment methods, resources checklist, and interview as potential predictors. For example, to predict an indicator of BO (the 'outcome indicator'), indicators of ADF, QBA, CO, resources checklist, and interview were used as potential predictors. In a first univariate screening, each predictor variable was selected in turn to judge its potential for prediction. A multinomial distribution with a logit-link function was used when the outcome indicator involved categorical data, a binomial distribution with a logit-link function for binary data, and a Poisson distribution with a log-link function and a multiplicative overdispersion parameter for count data (all models were generalised linear models [McCullagh & Nelder 1989]).

Subsequently, the outcome indicator was predicted using multiple predictors that were selected (*P*-value of Wald test < 0.20) in the first screening. The final prediction model was selected based on the lowest value for Akaike's Information Criterion (AIC). For categorical indicators, herds were assigned to the category with the highest predicted odds. The level of agreement between observed and predicted values of continuous WQ indicators was shown by their absolute difference and Spearman rank correlation (r_s). The latter correlation was interpreted by an informal classification system as suggested by Martin and Bateson (1993) for a Pearson correlation: slight ($r_s \leq 0.2$), low ($r_s > 0.2$ to 0.4), moderate ($r_s > 0.4$ to 0.7), high ($r_s > 0.7$ to 0.9), and very high ($r_s > 0.9$ to 1.0). For categorical WQ indicators, agreement between observed and predicted values was assessed by Cohen's kappa coefficient (κ; Cohen 1960). This coefficient was interpreted by an informal classification system as described by Landis and Koch (1977): poor ($\kappa \leq 0$), slight ($\kappa > 0$ to 0.2), low ($\kappa > 0.2$ to 0.4), moderate ($\kappa > 0.4$ to 0.6), high ($\kappa > 0.6$ to 0.8), and very high ($\kappa > 0.8$ to 1.0). In addition, positive (PR) and negative (NR) rates (which are similar to sensitivity and specificity of a diagnostic test) were calculated. To that end, observed and predicted values were grouped into classes 'minor problem' and 'moderate or severe problem'. The PR is defined as the probability for a 'moderate or severe problem' being predicted, given a 'moderate or severe problem' being observed. The NR is similarly defined for the 'minor problem' class. All calculations were performed with GenStat (GenStat for Windows 2011).

Results

The WQ protocol was executed in 196 dairy herds. Data from two herds were excluded because the protocol could not be executed correctly in these herds. In the remaining 194 herds,

https://doi.org/10.7120/09627286.22.1.085 Published online by Cambridge University Press

with herd size ranging between 10 and 211 lactating cows, cows were loose-housed on 181 farms, and tied on 13 farms. On 145 farms, cows had access to pasture in summer.

Twelve resource- and management-based (Table 1) and two animal-based WQ indicators (Table 2[a]) showed a prevalence of less than 5% and were therefore excluded from the statistical analyses. Missing values were replaced using MI in eight indicators: the number of days with access to pasture (missing in three herds), percentage of cows with lesions and swellings (one herd), with hairless patches (one herd), with SCC > 400,000 (seven herds), with dystocia (one herd), and ADF (could not be executed in six herds).

Correlations between animal-based indicators

Correlations between animal-based WQ indicators ranged from –0.51 (percentage of cows with hairless patches versus lesions) to 0.75 (percentage of cows with dirty udder versus dirty hindquarter). When animal-based WQ indicators belonging to different assessment methods were compared, correlations ranged from –0.26 (frequency of displacements versus QBA index) to 0.35 (percentage of very lean cows versus percentage of cows colliding with components of the stall while lying down; Figure 1).

Predicting ADF

The correlation between observed and predicted values for the ADF index was 0.31, which was interpreted as a low agreement. The difference between the observed and predicted values for the index ranged between –33.9 and 24.2 (95% range; Table 2[b]), which is comparable to an over- and underestimation of 33.9 and 24.2%, respectively, of cows that could be not be approached closer than 100 cm. The final prediction model for the ADF index comprised percentage of cows with dirty hind legs, lame, lying outside the supposed lying area, and QBA index as predictors (see *Appendix* [Available at the supplementary material to papers published in *Animal Welfare* section at the UFAW website; http://www.ufaw.org.uk/supplementarymaterial.php]).

Predicting QBA

Prediction of the QBA index resulted in a correlation of 0.54 between observed and predicted values. This was interpreted as a moderate agreement. The difference between the observed and predicted values ranged from –7.0 to 6.5 (95% range; Table 2[b]). The difference at the index level is hard to interpret at the level of descriptors due to the large number of terms in the QBA index. The final prediction model comprised percentage of cows with vulvar discharge, SCC > 400,000, lying outside the lying area, lame, severely lame, frequency of displacements, sufficient number of drinkers, ADF index, and herd size as predictors (see *Appendix*).

Predicting BO

The correlation between observed and predicted values was 0.38 for frequency of head butts and 0.46 for displacements, which was interpreted as a low and a moderate correlation. The difference between the observed and predicted values ranged from –0.8 to 1.6 head butts and –0.5 to 0.8 displacements per cow per hour (95% range; Table 2[b]). The final prediction model for frequency of head butts comprised percentage of cows with dirty hind legs, dirty hindquarters, diarrhoea, hairless patches, mortality, and lameness as predictors. For frequency of displacements, the final prediction model comprised percentage of cows that were very lean, dirty hind legs, nasal discharge, vulvar discharge, type of housing, and QBA index as predictors (see *Appendix*).

For the indicators of lying behaviour, κ ranged from 0.14 (mean time to lie down) to 0.44 (percentage of cows colliding with components of the stall; Table 2[a]). This was interpreted as a low to moderate agreement. NR was 12% for the mean time to lie down (Table 2[a]), which indicates that the probability for predicting a minor problem for this indicator, given a minor problem being observed, was low. PR was 15% for the percentage of cows lying outside the lying area, which indicates that the probability for predicting a moderate or severe problem, given a moderate or severe problem being observed, was low. The final prediction models for the indicators of lying behaviour comprised indicators relating to type of housing, lesions, lameness, body condition, diarrhoea, ocular discharge, cleanliness, and QBA index as predictors (see *Appendix*).

Predicting CO

For the continuous indicators of CO, correlation between observed and predicted values ranged from 0.33 (percentage of cows with hairless patches) to 0.50 (percentage of severely lame cows; Table 2[b]). This was interpreted as a low to moderate agreement. The largest difference (based on a 95% range) between observed and predicted values ranged from 15.5% for the percentage of very lean cows to 43.4% for the percentage of cows with lesions or swellings.

For the categorical indicators, κ ranged from –0.03 (percentage of cows with diarrhoea) to 0.07 (percentage of cows with dirty hindquarters), except for the percentage of cows with dirty udder, which showed a κ of 0.25 (Table 2[a]). This was interpreted as a poor to low agreement. NR was 0 and 4% for dirty hind legs and hindquarters, respectively, whereas PR ranged from 0 to 2% for the percentage of cows with diarrhoea, ocular, nasal, and vulvar discharge (Table 2[a]). None of the herds were assigned to a 'moderate problem', although a substantial number of herds were observed in this category.

The final prediction model for the percentage of very lean cows comprised herd size, the percentage of cows colliding with components of the stall while lying down, dehorning, and frequency of displacements as predictors (see *Appendix*). For the percentage of lame and severely lame cows, final prediction models were rather similar, comprising indicators relating to drinkers, mean time to lie down, frequency of head butts, ADF index, and QBA index as predictors. In addition, the model for the percentage of severely lame cows included herd size, access to pasture, frequency of coughing, the percentage of cows with SCC > 400,000, and mortality as predictors. Final prediction models for the percentage of cows with hairless patches and with lesions or swellings comprised indicators relating to drinkers, lying behaviour, agonistic behaviour, mortality, access to pasture, ADF index, and QBA index as

https://doi.org/10.7120/09627286.22.1.085 Published online by Cambridge University Press

predictors. With regard to indicators relating to cleanliness, final prediction models comprised indicators relating to lying behaviour, SCC, agonistic behaviour, type of housing, access to pasture, and ADF index as predictors. For indicators relating to disease (diarrhoea, ocular, nasal and vulvar discharge), final prediction models comprised indicators relating to drinkers, lying behaviour, agonistic behaviour, access to pasture, and coughing as predictors.

Discussion

Our aim was to assess the scope for reduction of on-farm assessment time of the WQ protocol for dairy cattle. To this end, performance was evaluated of a reduced protocol, in which a set of WQ indicators belonging to one assessment method was omitted and replaced by predictions based on remaining animal-, resource- and management-based indicators. Omitting indicators belonging to BO and CO from the protocol were estimated to result in the highest time gain: 150 and 132 min. Omitting indicators of ADF and QBA were estimated to result in 44 and 25 min time gain.

Herds in this study were selected on the basis of a composite health score to achieve more variation in the level of animal welfare. At the same time, this may have resulted in a better agreement between observed and predicted values. Consequently, a lower level of agreement might be found when herds are selected randomly. To avoid reduced power due to missing values (Donders *et al* 2006, 2009), multiple imputation was used to replace missing values. The percentage of missing values in our study was less than 1%. This technique has shown to be an appropriate method to deal with much larger proportions of missing values (Schafer & Olsen 1998). Therefore, the use of multiple imputation is not expected to have affected the results of this study to the extent of practical relevance.

More than one-third of the 41 indicators in the WQ protocol showed a prevalence of less than 5%. Because the majority of these indicators were resource- or management-based, exclusion of these indicators from the WQ protocol would result in approximately 15 min time gain only. With the exception of five indicators that were related to issues regulated by Dutch law (tail docking and use of anaesthetics for dehorning young stock), exclusion of these indicators is not recommended because prevalence may change over time and space, and herds that participated in this study may not be indicative for future populations.

Agreement between observed and predicted values was poor to moderate. The fact that WQ indicators provided little predictive value for other WQ indicators may reflect the aim of the Welfare Quality® project to select a minimum set of welfare criteria (Botreau *et al* 2007). On the other hand, factors inherent to the quality of the WQ monitoring system may have influenced predictive value. For example, the level of agreement between predicted and observed values is likely to be negatively affected by low inter-observer reliability (IOR) of indicators. This effect can be illustrated as follows: when indicator 'A' has a high IOR (ie little variation among different observers) and indicator 'B' has a low IOR (ie large variation among different observers), a low association between indicators 'A' and 'B' can be expected. Hence, a low IOR of 'B' negatively affects the prediction of 'A' by 'B'. A high IOR, for example, has been shown for the lameness scoring method used in our study (Winckler & Willen 2001), whereas IOR was found to be low for QBA (Kendall's *W* between 0.14 and 0.62; Bokkers *et al* 2012). If two observers, assessing lameness and QBA on the same farms, find similar percentages of severely lame cows but different scores for the QBA index, prediction of lameness by QBA (and *vice versa*) will be negatively affected. Obviously, the level of agreement deteriorates even more if IOR of both outcome and predictor are low.

Another possible reason for poor agreement between observed and predicted values, was that the observed classification was rather skewed for categorical indicators. Half of the indicators of BO and CO were categorical, whereas QBA and ADF contained no categorical indicators. For six of the twelve categorical indicators, more than two-thirds of the herds were in the 'minor problem' category. For two other indicators, more than two-thirds of the herds were in the 'severe problem' category. Prediction models assigned nearly all herds to the most frequent category. Consequently, herds with problems were overlooked (poor PR), or herds with proper welfare were incorrectly assumed to have a problem (poor NR).

Six indicators showed a moderate agreement between observed and predicted values; percentage of cows colliding with stall components, very lean, severely lame, with lesions or swellings, QBA index, and frequency of displacements. However, only omission of the QBA index from the WQ protocol would imply a reduction of on-farm assessment time because, contrary to the other indicators, the assessment method (QBA) contains only one indicator. Despite its low IOR (Bokkers *et al* 2012), the QBA index showed the highest agreement (r_s = 0.54) between observed and predicted values. The QBA index was predicted by frequency of displacements, amongst others, for which a correlation was also found in another study (Rousing & Wemelsfelder 2006). The ADF index was another important predictor for the QBA index. However, since ADF was assessed before QBA during the farm visit, the QBA scoring might have been influenced by the observations on the cows during the ADF.

The 'moderate' agreement between observed and predicted values for six indicators in the WQ protocol suggests that these observations and predictions were not completely unrelated. However, it also means that less than 30% of the observed variance was explained by the prediction models. This lack of predictive value was also illustrated by the large absolute differences between observed and predicted values. Therefore, it is not recommended to use these predictions as a replacement for omitted indicators in the WQ protocol.

In order to enhance the use of the WQ protocol in welfare audit programmes, other ways to reduce on-farm assessment time should be investigated. For example, few herd health records and resource- and management-based variables were used to predict WQ indicators in this study, whereas such variables have been shown to correlate with a large number of WQ indicators (eg Mülleder *et al* 2007; Sandgren *et al* 2009). Compared to animal-based WQ indicators, collecting herd health records and data for resource- and management-based variables is less time consuming and costly. Prediction of WQ indicators based on a larger share of herd health records and resource- and management-based variables, therefore, should be further investigated. Because in many countries herd health records are available in national databases, these could even be used for a first estimate of the level of animal welfare before an on-farm assessment is performed (Sandgren *et al* 2009; De Vries *et al* 2011). Besides the use of additional data, automated monitoring systems show the potential to reduce on-farm assessment time of the WQ protocol. Mainly for the assessment methods BO and CO, animal activity sensors or video recordings could replace direct visual observations for monitoring of, for example, lying behaviour or lameness (eg Flower *et al* 2005; Bewley *et al* 2010; Pluk *et al* 2012).

Conclusion

Replacing a set of animal-based WQ indicators belonging to one assessment method with predictions based on remaining WQ indicators showed little scope for a reduction of on-farm assessment time of the WQ protocol for dairy cattle. Therefore, except for indicators regulated by law, it is not recommended to omit indicators of the WQ protocol for dairy cattle. Other ways to reduce on-farm assessment time of the WQ protocol, such as the use of additional data or automated monitoring systems, should be investigated.

Acknowledgements

The authors gratefully acknowledge the farmers participating in this study, and Kees van Reenen, Wim Swart, Jac Thissen, and Pieter Vereijken for their stimulating discussions and ideas.

References

Bewley JM, Boyce RE, Hockin J, Munksgaard L, Eicher SD, Einstein ME and Schutz MM 2010 Influence of milk yield, stage of lactation, and body condition on dairy cattle lying behaviour measured using an automated activity monitoring sensor. *Journal of Dairy Research 77*: 1-6. http://dx.doi.org/10.1017/S00220 29909990227

Blackie N, Amory J, Bleach E and Scaife J 2011 The effect of lameness on lying behaviour of zero grazed Holstein dairy cattle. *Applied Animal Behaviour Science 134*: 85-91. http://dx.doi.org/10.1016/j.applanim.2011.08.004

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

Bokkers EAM, de Vries M, Antonissen I 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. http://dx.doi.org/10.7120/09627286.21.3.307

Botreau R, Bracke MBM, Perny P, Butterworth A, Capdeville J, Van Reenen CG and Veissier I 2007 Aggregation of measures to produce an overall assessment of animal welfare. Part 2: analysis of constraints. *Animal 1*: 1188-1197

Bowell VA, Rennie LJ, Tierney G, Lawrence AB and Haskell MJ 2003 Relationships between building design, management system and dairy cow welfare. *Animal Welfare 12*: 547-552

Cohen J 1960 A coefficient of agreement for nominal scales. *Educational and Psychological Measurement 20*: 37-46. http://dx.doi.org/10.1177/001316446002000104

de Vries M, Bokkers EAM, Dijkstra T, van Schaik G and de Boer IJM 2011 Invited review: associations between variables of routine herd data and dairy cattle welfare indicators. *Journal of Dairy Science 94*: 3213-3228. http://dx.doi.org/10.3168/jds.2011-4169

Dohoo IR, Martin SW and Stryhn H 2009 *Veterinary Epidemiologic Research*. VER, Inc: Charlottetown, Canada

Donders ART, van der Heijden GJMG, Stijnen T and Moons KGM 2006 Review: a gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology 59*: 1087-1091. http://dx.doi.org/10.1016/j.jclinepi.2006.01.014

Flower FC, Sanderson DJ and Weary DM 2005 Hoof pathologies influence kinematic measures of dairy cow gait. *Journal of Dairy Science 88*: 3166-3173. http://dx.doi.org/10.3168/jds.S0022-0302(05)73000-9

GenStat for Windows 2011 *GenStat for Windows Release 14*. VSN International Ltd: Hemel Hempstead, UK

Ito K, von Keyserlingk MAG, LeBlanc SJ and Weary DM 2010 Lying behavior as an indicator of lameness in dairy cows. *Journal of Dairy Science 93*: 3553-3560. http://dx.doi.org/10.3168/jds.2009-2951

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

Landis JR and Koch GG 1977 The measurement of observer agreement for categorical data. *Biometrics 33*: 159-174. http://dx.doi.org/10.2307/2529310

LEI 2008 *Farm Accountancy Data Network 2009*. LEI: The Hague, The Netherlands. http://www.lei.wur.nl/UK/statistics/Binternet/

Martin P and Bateson P 1993 *Measuring Behaviour. An Introductory Guide*. Cambridge University Press: Cambridge, UK. http://dx.doi.org/10.1017/CBO9781139168342

McCullagh P and Nelder JA 1989 *Generalized Linear Models*. Chapman and Hall: London, UK

Mülleder C, Troxler J, Laaha G and Waiblinger S 2007 Can environmental variables replace some animal-based parameters in welfare assessment of dairy cows? *Animal Welfare 16*: 153-156

Pluk A, Bahr C, Poursaberi A, Maertens W, van Nuffel A and Berckmans D 2012 Automatic measurement of touch and release angles of the fetlock joint for lameness detection in dairy cattle using vision techniques. *Journal of Dairy Science 95*: 1738-1748. http://dx.doi.org/10.3168/jds.2011-4547

Rousing T and Wemelsfelder F 2006 Qualitative assessment of social behaviour of dairy cows housed in loose housing systems. *Applied Animal Behaviour Science 101*: 40-53. http://dx.doi.org/10.1016/j.applanim.2005.12.009

Sandgren CH, Lindberg A and Keeling LJ 2009 Using a national dairy database to identify herds with poor welfare. *Animal Welfare 18*: 523-532

Schafer JL and Olsen MK 1998 Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research 33*: 545-571. http://dx.doi.org /10.1207/s15327906mbr3304_5

https://doi.org/10.7120/09627286.22.1.085 Published online by Cambridge University Press

Waiblinger S, Menke C and Folsch DW 2003 Influences on the avoidance and approach behaviour of dairy cows towards humans on 35 farms. *Applied Animal Behaviour Science 84*: 23-39. http://dx.doi.org/10.1016/S0168-1591(03)00148-5

Webster AJF, Main DCJ and Whay HR 2004 Welfare assessment: indices from clinical observation. *Animal Welfare 13(S)*: S93-S98

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

Winckler C and Willen S 2001 The reliability and repeatability of a lameness scoring system for use as an indicator of welfare in dairy cattle. *Acta Agriculturae Scandinavica, Section A - Animal Science 51*: 103-107