The Old School, Brewhouse Hill, Wheathampstead,
Hertfordshire AL4 8AN, UK
www.ufaw.org.uk

Animal Welfare 2014, 23: 95-107
ISSN 0962-7286
doi: 10.7120/09627286.23.1.095

Implementing Welfare Quality® in UK assurance schemes: evaluating the challenges

CAE Heath*, Y Lin, S Mullan, WJ Browne and DCJ Main

School of Veterinary Sciences, University of Bristol, Langford House, Langford BS40 5DU, UK
* Contact for correspondence and requests for reprints: cheryl.heath@bristol.ac.uk

Abstract

This paper presents an account of a Welfare Quality® assessment of 92 dairy farms carried out by seven experienced assessors. The aim was to evaluate the potential of the Welfare Quality® assessment protocol with respect to its uptake by UK farm assurance schemes. Data collection, and measure aggregation were performed according to the Welfare Quality® protocol for dairy cows. This study examined the data itself, by the testing of how hypothetical interventions might be reflected in changes in the aggregated scores, and also investigated human-related aspects, through inter-assessor standardisation sessions to evaluate reliability, and an assessor focus group to collect feedback. Overall, three main 'challenges' were identified. The first challenge related to the large amount of missing data. Unexpectedly, this was such that it was only possible to calculate an overall classification for 7% of farms. The second challenge concerned the way in which aggregated scores did not always reflect hypothetical interventions. The final challenge was inter-assessor reliability, where not all assessors were found to achieve acceptable levels of agreement on a number of outcome measures by the third training session. Suggestions for managing these challenges included, follow-up to assessor training, the use of multiple imputation methods to fill in missing data, and, where applicable, not aggregating the scores. The conclusion of the study was that the protocol provided useful information from which to make an informed selection of measures, but that the challenges, combined with the lengthy assessment time, were too great for its use as a certification tool.

Keywords: *animal welfare, dairy cow, focus group, on-farm assessment, score aggregation, Welfare Quality®*

Introduction

Traditionally, animal welfare assessment schemes have been concerned with the measurement of inputs into the husbandry system, such as the provision of resources and aspects of the farm management system. These 'input-based' measurements have the advantage of being easy to measure and stable over time. Behavioural and physical observations of the animals can be understood to represent the outcome of the husbandry system and are referred to as outcome-based measurements. The Farm Animal Welfare Council (2005) has recommended the inclusion of outcome measures in farm assurance certification schemes. An advantage of an outcome-based assessment is that the animal is the focus of the assessment, and this allows comparisons across farming systems. It is for this reason that animal-based measurements are now considered to provide a more direct account of welfare, reflecting the experience of the animal (Webster *et al* 2004; Welfare Quality® 2009a). However, one of the criticisms of outcome-based measures is that they involve a degree of subjective interpretation, and the scoring is at risk of assessor bias. For this reason, standardisation of scores and adequate training are vital to minimise individual differences.

The question of what should be measured to assess welfare is often debated. While science can provide answers as to how things should be measured, the issue of what is considered important for animal welfare represents more of an ethical decision. Animal welfare is a multi-dimensional concept (Fraser 1995), encompassing both mental and physical health (Dawkins 2003) and, as such, it might be expected that a welfare assessment will reflect this in both the measures that are collected, and the manner in which they are aggregated. While an aim of welfare assessment may be to be as comprehensive as possible, outcome-based measures are inherently time consuming to collect, especially those relating to certain behavioural observations, given their infrequent displays. Whether the substantial amount of time required to collect the measures is warranted may depend on requirements and time restriction of the specific application of the protocol.

The Welfare Quality® assessment protocols present an extensive, scientifically robust, outcome-based account of welfare (Blokhuis *et al* 2010), whose conceptual underpinnings reflect the opinions of stakeholders from numerous backgrounds, including scientists, social scientists and the general public (Miele *et al* 2011). The protocols describe a

three-level system of score aggregation where measures collected at the farm are aggregated, firstly, into 12 criteria scores, which are then aggregated into four principle scores, (*Good feeding*, *Good housing*, *Appropriate behaviour* and *Good health*) and, finally, the principle scores are combined into an overall welfare classification, (*not classified*, *acceptable*, *enhanced* or *excellent*). Potential applications for the protocol include those for legislative purposes, voluntary certification, for on-farm management, or for use in research (Main *et al* 2003; Botreau *et al* 2009).

It is its potential use as a tool used in certification that is of interest here. Although membership of farm assurance schemes is voluntary, membership is often essential for producers to market their products. Thus, the potential role of farm assurance schemes as a force for welfare improvement should not be underestimated. Farm animal welfare assessments have an important role to play, however, this has not been without difficulties, in particular, in carrying out an assessment which is scientifically robust within time constraints. It has been acknowledged that the length of time needed to carry out a Welfare Quality® assessment limits feasibility (Knierim & Winckler 2009), and this could therefore compromise its potential use in farm assurance. However, whether the length of time for an assessment would be considered to be a barrier to implementation would depend very much on the specific needs of individual farm assurance schemes, and which of the three main uses it was intended for: whether it was used as a surveillance tool to identify potential scheme improvements, or whether it was to be used at the farm level as a discussion tool to stimulate welfare improvements, or as a certification tool where farms are assessed for compliance to a particular scheme's standards. The Welfare Quality® protocol is not currently used in certification, however it was the aim of this study to evaluate its potential with respect to uptake by certification schemes in the UK. A similar study by de Vries *et al* (2013) has investigated reducing the length of time for a Welfare Quality® assessment of dairy cows, where the assessors were from a range of backgrounds, and included several researchers (M de Vries, personal communication 2012). This present study investigated the full Welfare Quality® assessment using assessors from certification schemes.

Seven employed professional assessors from the RSPCA Freedom Food or Soil Association Certification with at least one year's experience in carrying out farm animal welfare assessments on commercial farms were used. This study consists of three main areas of investigation, the Welfare Quality® assessment of dairy cows, the standardisation of the assessor scores, and feedback on the assessments from the assessors.

Materials and methods

Recruitment

A full Welfare Quality® assessment was 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. Each assurance scheme recruited farmers independently. RSPCA Freedom Food telephoned all of their assured farms and recruited 30 out of approximately 40 farms. The Soil Association Certification scheme applied a two-stage process. Firstly, a subset of Soil Association Certification farms was selected proportionally from different geographic regions, then Soil Association Certification Officers telephoned the farmers in alphabetical order, until 31 farms were recruited. A total of 31 Red Tractor Farm Assurance Dairy scheme farms were recruited through two UK milk companies.

Data collection

Data were collected by seven assessors, four employed by Soil Association Certification, two by RSPCA Freedom Food, and one was an RSPCA Farm Livestock Officer. All were experienced in carrying out welfare assessments on dairy farms. The assessors received standardised training in the Welfare Quality® assessment protocols (Welfare Quality® 2009b) given by researchers involved in the Welfare Quality® consortium. Training consisted of both classroom and on-farm instruction over two consecutive days. The assessors were trained on the 67 Welfare Quality® measures, 57 of which are associated with herd level data, and ten measures which relate to resources and management practices and which are collected at the group level. Between January and August 2011, each Welfare Quality® assessment was carried out by an individual assessor during a single farm visit, where individual assessors collected data for several farms. The measures, categorised into sections that are collected together, such as avoidance distance measures, Qualitative Behaviour Assessment (QBA), behavioural observations, clinical scoring, and management questionnaire. The order in which they were collected are shown in Tables 1 and 2. Sample sizes in the protocol for clinical scoring measures depend on herd size and in the data collected range from a minimum of 30 to a maximum of 73 cows. The data were collected according to the protocol, which stipulates an amount of time for the collection of each section of measures.

Analysis of Welfare Quality® assessment

The data were analysed according to the Welfare Quality® protocol for dairy cows (Welfare Quality® 2009a). The protocol provides formulae or decision trees for the calculation of 11 criteria scores formed by combining the collected measures. The 11 criteria scores are then aggre-

https://doi.org/10.7120/09627286.23.1.095 Published online by Cambridge University Press

Table 1 Summary of the 45 outcome and 13 resource and management-based quantitative measures of the Welfare Quality® assessment of 92 UK dairy farms.

Section† (time taken according to protocol	Measure code	Description	Min	Q1	Median	Mean	Q3	Max	Absent (%)
		Number of groups	1.00	1.00	1.00	1.93	2.00	9.00	1 (1)
		Number of lactating cows	35.00	95.00	160.00	189.00	237.80	909.00	0 (0)
		Number of cows in group‡	6.00	54.50	90.00	108.40	128.50	865.00	8 (5)
AD (1 min per cow)	1	% cows that can be touched	0	20.23	32.00	30.80	41.27	71.43	13 (14)
	2	% cows that can be approached by 50 cm, but not touched	0	37.07	50.00	47.19	56.77	100.00	13 (14)
	3	% cows that can be approached between 50 cm and 1 m	0	6.16	12.64	15.11	20.16	48.44	13 (14)
	4	% cows that cannot be approached	0	0	2.38	6.87	6.62	97.67	13 (14)
QBA (25 min)	5–24	Calculated score	–3.36	–0.57	0.37	0.00	0.78	1.44	5 (5)
BO (150 min)	25	Duration of lying down movements	3.00	4.24	5.12	5.21	6.15	8.07	30 (33)
	26	% lying down movements with collisions	0	0	20.00	26.50	46.01	100.00	38 (41)
	27	% lying cows which lie partly outside the lying area	0	0	0	2.25	1.65	32.41	29 (32)
	28	Frequency of butts per cow per hour	0	0.28	0.50	0.59	0.81	2.27	20 (22)
	29	Frequency of displacements per cow per hour	0	0.14	0.29	0.37	0.53	1.38	20 (22)
	30	Frequency of coughing per cow per 15 min	0	0.05	0.09	0.09	0.12	0.27	20 (22)
CS (3 min per cow)	31	% very lean cows	0	1.32	3.49	5.72	7.38	30.30	9 (10)
	32	% cows with dirty udder	2.00	13.00	24.00	32.26	48.50	98.00	1 (1)
	33	% cows with dirty flank and upper legs	5.00	39.50	54.00	55.48	73.00	100.00	1 (1)
	34	% cows with dirty lower legs	15.00	69.00	90.00	80.64	97.00	100.00	1 (1)
	35	% not lame cows	23.81	73.39	87.76	81.93	94.50	100.00	1 (1)
	36	% moderately lame cows	0	4.59	10.00	13.23	19.09	53.45	1 (1)
	37	% severely lame cows	0	0	1.75	4.92	6.25	47.62	1 (1)
	38	% cows with at least one hairless patch, and no lesion	4.65	26.34	42.17	39.82	52.78	78.79	1 (1)
	39	% cows with at least one lesion	0	11.79	21.59	29.77	44.53	88.10	1 (1)
	40	% cows with no lesion	11.90	55.47	78.40	70.29	88.22	100.00	1 (1)
	41	% cows with nasal discharge	0	0.00	1.56	3.69	4.94	64.62	1 (1)
	42	% cows with ocular discharge	0	1.62	4.62	6.20	8.63	30.14	1 (1)
	43	% cows with increased respiratory rate	0	0	0	0.47	0	6.15	1 (1)
	44	% cows with diarrhoea	0	0	0	2.22	2.63	24.66	1 (1)
	45	% cows with vulvar discharge	0	0	0	0.95	1.49	14.75	1 (1)
RC (15 min)	46	Number of water troughs‡	1.00	2.00	3.00	3.17	4.00	11.00	13 (8)
	47	Total length of water troughs‡	69.00	382.80	555.00	716.50	828.80	3,502.00	17 (11)
MQ (15 min)	54	Number of days on pasture per year‡	0	180.00	200.00	185.70	215.00	365.00	34 (22)
	55	Number of hours on pasture per day‡	0	17.00	19.00	16.95	20.00	24.00	47 (30)
	56	% de-horned cows	25.00	100.00	100.00	95.46	100.00	100.00	5 (5)
	64	% mastitis	0	8.28	11.82	15.49	19.45	89.00	25 (27)
	65	% mortality during the last 12 months	0	1.75	3.00	4.27	5.45	15.79	5 (5)
	66	% dystocia	0	1.78	4.44	5.29	7.89	18.75	5 (5)
	67	% downer cows	0	1.19	2.50	3.45	4.30	18.20	5 (5)

† AD = Avoidance distance, QBA = Qualitative Behaviour Assessment, BO = Behavioural observations, CS = Clinical scoring, RC = Resources checklist, MQ = management questionnaire/ farm records; Milking group data, n = 155.

Table 2 Summary of the qualitative measures of the Welfare Quality® assessment of 92 UK dairy farms.

Section[†]	Measure code	Description	Coding and count				Absent (%)
			0	1	2	3	
RC	49	Cleanliness of water points[‡]*	Clean or partly (128)	–	Not clean (4)	–	23 (15)
MQ	57	Method used for de-horning*	No de-horning (0)	Chemical (2)	Thermal (82)	Surgery (0)	8 (9)
	58	Use of anaesthetics for de-horning*	Anaesthetics (88)	–	No anaesthetics (0)	–	4 (4)
	59	Use of analgesics for de-horning*	Analgesics (5)	–	No analgesics (20)	–	67 (73)

[†] RC= Resources checklist, MQ = management questionnaire/farm records.
* Categorical data; [‡] Milking group data, n = 155.

Table 3 Composition of criteria and principle scores from the Welfare Quality® protocol for dairy cows, and percentage of farms able to calculate each aggregated score.

Criteria (measures)	% of farms	Principle	% of farms	% of farms with overall score
Absence of prolonged hunger (31)	90	Good feeding	78	7
Absence of prolonged thirst (46–50)	87			
Comfort around resting (25–27, 32–34)	55	Good housing	55	
Ease of movement (51–55)	100			
Absence of injuries (36–39)	99	Good health	11	
Absence of disease (30, 41–45, 64–67)	53			
Absence of pain induced by management procedures (56–63)	26			
Expression of social behaviours (28, 29)	78	Appropriate behaviour	52	
Expression of other behaviours (54, 55)	67			
Good human-animal relationship (1–4)	86			
Positive emotional state (5–24)	95			

gated into four principle scores by applying Choquet integrals, which both minimise compensation between measures and place greater emphasis on the lowest scores (Botreau *et al* 2007). Finally, an overall score is calculated based on thresholds applied to the four principle scores. Table 3 shows the measures used in the construction of the criteria and principle scores. Updated versions of the formulae for the criteria scores were accessed from the Welfare Quality® website (Welfare Quality® 2012). Due to some errors in the printed version of the protocol, further revisions were applied (I Veissier, personal communication April and May 2012). Although Welfare Quality® offers the services of its own software (Welfare Quality® 2009a), in order to compute the criteria, principle and overall scores for the dataset used in this study, 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® (version 2007), and SPSS® (SPSS® version 18).

To illustrate the impact of a targeted welfare intervention, hypothetical interventions were performed on what were considered to be measures, or groups of measures which were (biologically) independent. It was hypothesised that an intervention on those measures on each farm might, in theory, move the measure(s) to the maximum (best) value seen across the sample of farms in the study for that measure. The improved criteria and principle scores as a result of such interventions were then examined, according to interventions based on both maximum observed values, and the maximum theoretical values, while still corresponding to UK farming practices.

https://doi.org/10.7120/09627286.23.1.095 Published online by Cambridge University Press

Table 4 Assessment completion rate for seven assessors for each section of the Welfare Quality® protocol for dairy cows.

Section	Completion rate by each assessor in each section (%)†							Total missing (%)
	Assessor (number of assessments made)							
	A (9)	B (14)	C (7)	D (15)	E (9)	F (22)	G (16)	
Avoidance distance	77.78	71.43	85.71	100	100	100	62.50	14.65
QBA	100	92.86	71.43	100	100	100	87.50	6.89
Behavioural observations	87.04	35.71	100	73.33	90.74	77.27	61.46	24.92
Clinical score	100	100	93.33	100	100	99.70	93.75	1.89
Resource checklist	96.30	85.71	95.24	93.33	100	74.24	58.33	13.83
Management questionnaire	94.44	59.52	85.71	85.00	85.19	76.89	76.04	19.60

† All seven body condition scores for assessor C were excluded on the basis of poor inter-assessor reliability.

Standardisation tests

Three inter-assessor standardisation tests were performed: i) 13 days before; ii) during; and iii) 6 months after the on-farm Welfare Quality® assessments were carried out on the recruited farms. The first standardisation test was carried out with the seven assessors as part of the training course and before the assessors had been to assess any of the recruited farms. The farm used for the training session was not used again. However, the farm used for both the second and third standardisation tests was among the farms recruited for this study, and was assessed ten days after the second assessor standardisation session. During the first test, each assessor's scores were rated against the experienced trainer who was considered the 'gold standard'. The second test was held on a farm with six assessors, one month and eleven days after the first training session. The third test was held eleven months and 13 days after the second test, on the day of the feedback meeting, after the assessors had finished all their farm assessments, and at which five assessors were present. In the absence of the trainer (gold standard), the individual ratings from sessions 2 and 3 were compared to the mode of the assessors. This use of the gold standard or modal value is in line with Mullan *et al* (2011) who suggest comparison with the gold standard where there is inequality in experience level between assessors, otherwise, agreement can be assessed by comparison to the mode. The outcome measures included in the three tests were 'lameness', 'body condition score', 'cleanliness', 'hair loss, lesions, and swellings', and 'sign of diseases'. The percentage of agreement for the assessors with either the gold standard or the mode for each cow assessed was analysed to give the range of percentage agreements per measure of all the assessors. In addition, the prevalence range of each of the different outcome measures for the assessors was compared to the gold standard or mode. Twenty cows were assessed for each assessment.

Focus group

Five out of the seven assessors attended a focus group discussion which was held at the end of the data collection period. The remaining two assessors were unable to attend due to other commitments. The focus group was carried out in three stages. To begin with, the assessors discussed the whole assessment in more general terms. The interviewer then asked the assessors to discuss the advantages and disadvantages, usefulness and practicality of each measure. Follow-up questions were asked when the meaning was uncertain (Roe *et al* 2011; Anneberg *et al* 2012). The final section of the focus group concerned whether assessors had talked to farmers and, if so, what the farmers' opinions were of the Welfare Quality® assessment. The initial group discussion was tape-recorded and transcribed *verbatim* for qualitative research and analysis. The transcript was studied to identify important themes and sub-themes. The themes were then categorised and coded (DeSantis & Ugarriza 2000; Braun & Clarke 2006). Segments of *verbatim* text were categorised into sub-themes and brief summaries of the interviewees' opinions are shown in the *Results*.

Results

Data completion

The Welfare Quality® protocol was carried out on 92 dairy farms with the number of lactating cows ranging from 35 to 909 with a mean number of 189 cows. The measures collected from the 92 farms, are summarised in Tables 1 and 2. The measures are listed in the order they were collected, grouped into sections, (avoidance distance, Qualitative Behaviour Assessment (QBA), behavioural observations, clinical scoring, resource checklist, and management questionnaire), and the length of time taken to collect each section of data is shown. Tables 1 and 2 show that the measures had varying amounts of missing data. Nine

Table 5 Percentage of farms with scores at the criteria and principle level following an 'intervention'[†].

Principle	Criteria: selected measures (measure code)	Percentage of farms at each score interval post-intervention (percentages of farms that moved)									
		Criteria level					Principle level				
		0.0–5.0	5.1–15.0	15.1–50.0	50.1–75.0	> 75.1	0.0–5.0	5.1–15.0	15.1–50.0	50.1–75.0	> 75.1
Good feeding	Absence of prolonged thirst (46, 47, 49)				1 (–32)	99 (+56)			4 (–9)	37 (–7)	59 (+33)
	Absence of prolonged hunger (31)					100 (+51)		16	6 (–7)	34 (–10)	44 (+18)
Good housing	Comfort around resting: duration of lying down movements (25)			49 (–8)	47 (+8)	4 (+2)			4 (–14)	78 (+5)	18 (+8)
	Comfort around resting: % lying down movements with collisions (26)			49 (–8)	47 (+8)	4 (+2)			4 (–14)	86 (+13)	10
	Comfort around resting: % lying cows which lie partly outside the lying area (27)			55 (–2)	43 (+4)	2			12 (–6)	76 (+3)	12 (+2)
	Comfort around resting: % cows with dirty udder; % cows with dirty flank and upper legs; % cows with dirty lower legs (32–34)			37 (–20)	31 (–8)	31 (+29)			2 (–16)	49 (–24)	49 (+39)
Good health	Absence of injuries: % not lame cows; % moderately lame cows; % severely lame cows (35–37)				58 (+16)	42 (+28)			70 (–20)	30 (+20)	
	Absence of injuries: % cows with at least one hairless patch and no lesion; % cows with at least one lesion; % cows with no lesion (38–40)			12 (–31)	36 (–6)	52 (+38)			70 (–20)	30 (+20)	
	Absence of disease: frequency of couging per cow per 15 min (30)			67	31	2			10	90	
	Absence of disease: % cows with nasal discharge (41)			64 (–4)	33 (+2)	4 (+2)			10	90	
	Absence of disease: % cows with ocular discharge (42)			57 (–10)	39 (+8)	4 (+2)			10	90	
	Absence of disease: % cows with increased repiratory rate (43)			65 (–2)	33 (+2)	2			10	90	
	Absence of disease: % cows with diarrhoea (44)			63 (–4)	33 (+2)	4 (+2)			10	90	
	Absence of disease: % cows with vulvar discharge (45)			63 (–4)	35 (+4)	2			10	90	
	Absence of disease: % mastitis (64)			59 (–8)	31	10 (+8)			10	90	
	Absence of disease: % mortality during the last 12 months (65)			51 (–16)	47 (+16)	2			10	90	
	Absence of disease: % dystocia (66)			44 (–23)	47 (+16)	8 (+6)			10	90	
	Absence of disease: % downer cows (67)			51 (–16)	47 (+16)	2			10	90	
	Absence of pain induced by management procedures: method used for de-horning; use of anaesthetics for de-horning; use of analgesics for de-horning (57–59)				100 (+4)				70 (–20)	30 (+20)	
Appropriate behaviour	Expression of social behaviours (28–29)					100 (+57)			46 (–10)	50 (+6)	4 (+4)
	Expression of other behaviours (54–55)					100 (+38)			48 (–8)	50 (+6)	2 (+2)
	Good human-animal relationship (1–4)					100 (+89)			42 (–14)	56 (+12)	2 (+2)
	Positive emotional state (5–24)					100 (+93)			17 (–39)	73 (+29)	10 (+10)

[†] To carry out a hypothetical 'intervention', the highest values observed in this study for measures associated with specific targeted on-farm improvements were substituted for all 92 farms. Criteria and principle scores were then calculated and the percentage of farms at each category of score is shown, together with the percentage of farms that have moved in that particular score category.

https://doi.org/10.7120/09627286.23.1.095 Published online by Cambridge University Press

Table 6 Percentage agreement with the gold standard and the mode for different outcome measures from three inter-assessor standardisation sessions.

Measures	Session[†]	Range prevalence (%)	Gold standard (trainer) prevalence (%)	Mode prevalence (%)	% agreement[‡] with either gold standard or mode (%)
% very lean cows	1	0–5	0	–	95–100
	2	0–11	–	5	89–100
	3	6–38	–	11	80–100
% cows with dirty lower legs	1	21–95	50	–	55–85
	2	100–100	–	100	100–100
	3	87–100	–	100	87–100
% cows with dirty udder	1	5–29	20	–	82–95
	2	28–95	–	65	54–100
	3	20–89	–	58	57–100
% cows with dirty flank and upper legs	1	25–60	45	–	53–80
	2	42–89	–	60	76–100
	3	28–79	–	53	72–87
% cows with nasal discharge	1	0–20	15	–	65–94
	2	0–11	–	0	89–100
	3	0–32	–	0	77–100
% cows with increased respiratory rate	1	0–0	0	–	100–100
	2	0–0	–	0	100–100
	3	0–5	–	0	95–100
% cows with ocular discharge	1	0–0	0	–	100–100
	2	0–0	–	0	100–100
	3	11–29	–	16	79–94
% cows with diarrhoea	1	0–6	0	–	94–100
	2	0–0	–	0	100–100
	3	0–0	–	0	100–100
% cows with vulvar discharge	1	0–6	0	–	94–100
	2	0–5	–	0	95–100
	3	0–24	–	0	87–100
% cows with at least one hairless patch and no lesion	1	15–100	85	–	50–75
	2	95–100	–	95	89–95
	3	75–100	–	84	71–94
% cows with at least one lesion	1	15–90	60	–	10–55
	2	40–95	–	70	56–79
	3	42–95	–	47	53–89
% moderately lame cows	1	22–60	60	–	44–60
	2	17–47	–	20	76–89
	3	25–63	–	26	64–100

[†] Only 19 cows were used in session 3 as one cow escaped assessment, and on occasion some cows were not assessed by some assessors.
[‡] To calculate the percentage agreement, a pair-wise comparison was made for each outcome score per cow between individual assessors and either the gold standard or mode. The percentage agreement for individual assessors was calculated from the sum of the agreements and non-agreements, based on this pair-wise comparison, and this provided the range of percentage agreement shown.

measures were excluded from analysis as they were not applicable to dairy farming practices in the UK. These were, measures 60–63, which relate to the routine practice of tail docking (routine tail docking is against UK legislation); measures 51–53 which relate to the practice of tethering (tethering is not a common husbandry practice in the UK), measure 48, *Number of water bowls* (water bowls are not normally used in the UK), and measure 50, *Water flow* (water flow tests are not required for troughs with large reservoirs, and instead, a default value was applied). In order to ascertain whether it was possible to attribute the missing data to either the assessors or the practicality of collecting the different sections of data, the relative completion of each section was looked at.

Table 4 shows that individual assessors varied in terms of completion rate of each section. The body condition scores for all seven of one particular assessor's on-farm assessments were excluded on the basis of poor inter-assessor reliability. All assessors had data missing from at least two sections of the assessment, and one of the assessors failed to complete all of even one section. Clinical scoring, which had 15 measures, had the smallest variation in completion rate between assessors, ranging from 93.33 to 100%, whereas behavioural observations, with six measurements, represented the largest variation amongst assessors, with completion raates between 35.71 and 100% completion rates.

Table 3 shows that due to missing data an overall score was only calculable on 7% of farms, and only on 11% of farms were we able to calculate a score for *Good health*. Due to the way in which measures are aggregated, a missing value from a single measure prevents the calculation of the criteria score it belongs to, and hence the principle score, and finally prevents an overall score from being calculated.

The number of measures, and criteria which compose different principles' scores varies. Theoretically, the principle *Good housing*, is composed of the criteria *Ease of movement*, *Comfort around resting*, and *Thermal comfort*. However, as no measure has yet been developed for *Thermal comfort*, for the Welfare Quality® protocol for dairy cows, the value attributed to this score equates to the greatest score out of *Ease of movement* and *Comfort around resting*. As the criterion *Ease of movement* relates to the practice of tethering, and as this practice is uncommon in the UK, all British farms will normally achieve the maximum score for this criteria, and as a consequence the only variation of the principle *Good housing*, is found in the criteria for *Comfort around resting*.

Intervention

Through carrying out hypothetical interventions based on maximum observed values, it was possible to look at how independent on-farm improvements were reflected in the scoring system. The interventions were intended to represent potential changes which could be carried out on-farm, irrespective of which level of aggregation the associated measures could be found.

Table 5 shows the improved criteria and principle scores as a result of a number of hypothetical interventions. With the exception of the intervention for 'Avoidance distance', interventions based on the theoretical maximum values, (not shown), produced the same improvement in principle scores as interventions based on the maximum observed values.

The justification for carrying out independent interventions was based on the fact that only three pairs of measures were found to correlate with a coefficient greater than or equal to 0.5 (excluding correlations between descriptive terms used to calculate *Positive emotional state* [QBA]). These correlations were considered to be biologically unrelated, and included *Positive emotional state* and *Percentage of cows with at least one lesion*, r_s (84) = –0.52, $P < 0.01$; *Duration of lying down movements* and *Percentage of cows with dystocia*, r_s (57) = 0.50, $P < 0.01$, and *Percentage of cows with dirty lower legs* and *Percentage of lying down movements with collisions*, r_s (52) = 0.53, $P < 0.01$ (the differing degrees of freedom for the tests are due to missing data). The impact that the interventions had on the criteria scores varied greatly, with some interventions to measures associated with *Absence of prolonged hunger*, *Absence of prolonged thirst*, *Comfort around resting*, 'Cleanliness', and *Positive emotional state*, being able to achieve a greater impact on the criteria and principle scores than, for example, individual interventions associated with *Absence of disease* which had no impact on the score at the principle level whatsoever.

Standardisation test

Table 6 illustrates the range of prevalences of 12 outcome measures for 20 cows. Agreement with the gold standard was calculated as the percentage of times that an observer scored the same as the gold standard for an individual cow, whereas, agreement with the mode was the percentage of times that an observer scored the same as the most frequent (modal) answer. If there was not a clear mode, the data from that animal were discarded. This may have occurred due to an equal split between an even number of assessors, where it was not possible to calculate a modal value, or in the case of missing data. Higher levels of agreement were associated with low prevalences. There was at least 80% agreement in all sessions for the following four measures: 'percentage of very lean cows', 'percentage of cows with increased respiratory rate', 'percentage of cows with diarrhoea', and 'percentage of cows with vulvar discharge'.

Focus group discussion

Five of the assessors that participated in the Welfare Quality® project shared their thoughts about the dairy cattle protocol at a feedback meeting. The major themes of this study were about ease of use and any difficulties observed with carrying out the protocols and completing a report. Considering the amount of missing data, it is surprising that interviewees failed to discuss completing reports as a difficulty. Most of them held more positive than negative views. In general, they thought that the main principle of Welfare Quality®, focusing on outcome- rather than resource-based assessment, was good for animal welfare.

Assessors described their experience as follows:

https://doi.org/10.7120/09627286.23.1.095 Published online by Cambridge University Press

> I mean actually doing the test [Welfare Quality® dairy cow assessment] itself is quite simple (Assessor F).
>
> I think it [QBA] was quite easy, well it's quite, from a self, personally speaking, it was quite quick to pick it up and is a surprisingly good indicator (Assessor B).
>
> They [the farmers] all thought that actually doing more observation with the animals was good for farm assurance (Assessor E).
>
> It [the Welfare Quality® protocol] gives an assessor a lot more credibility (Assessor B).

For most of the outcome measures (avoidance distance, QBA, behavioural observations, and management questionnaire), assessors felt that the concepts were simple to understand. However, clinical scoring was considered to be a more complex concept. With respect to the actual performance of the measures, some difficulties were identified: the first issue, and the most common one, was the duration of the inspection. The assessors claimed that the whole assessment took them a minimum of five hours to complete. For example, comments from the assessors included:

> From a practical point of view, you haven't got two hours to stand there monitoring behaviour (Assessor B).

The second issue with regard to the performance of the measurements was the variation in the type and layout of farms. For example, one assessor mentioned that the avoidance distance test was difficult to accomplish due to the difference in farm resources.

> Where you've got ring feeders or feeding from a face, you'd almost give up doing it, you wouldn't get anything sensible. If you've got mixtures of ages, the first couple of cows you go up to, you get a huge reaction, then three or four cows down, if they were sort of the old stagers, they were inquisitive, alright he's not coming to hurt us [sic], so you could go up and touch their nose. If you've got a flighty one, you know, a heifer that would react to it, that would affect the next two or three (Assessor B).

The third issue raised by the assessors was the effect of the time of year on the measures. When inspecting in winter, it is more likely that cattle are held indoors, whereas in the summer the cattle may be out grazing, which may have an effect on the score for cleanliness. For example:

> You will get a different score on cleanliness for a grazing animal to an animal that's been in the house on for example integuments, cleanliness. If you're going to look at a herd which is always inspected in February, and a herd which is always inspected in August, you're going to get a different result (Assessor F).

The fourth topic was the difficulty of working with the cows. Here, one assessor brought up the issue that the cows are continuously moving, meaning that they are not standing in the correct way for examination. Another assessor suggested the potential for bias in selection:

> You know, you could spend five minutes trying to get good, one side of the cow and then give up, whereas another one would just stand there and you'd get everything done really quickly (Assessor E).
>
> Particularly when it's is finding the right cows and then getting them up and making them walk and all that kind of thing, and then you're obviously surrounded by, [sic] so that's what's going to hamper the whole thing (Assessor D).
>
> The other thing, I think, would cause an issue, it's very subjective the cow to pick out... (Assessor G).

The last issue highlighted some possibly unclear definitions within the assessment protocol. It was felt that some words used in QBA or clinical scoring had abstract meaning, causing confusion when both explaining to the farmers and doing the assessment. For example, one assessor said:

> There's quite a lot of farmers [sic] didn't know what I was talking about (Assessor F).

With regards to the farmers themselves, the assessors' views were that the farmers' opinions of the Welfare Quality® scheme ranged hugely. Some viewed it as vindication of good welfare management to justify the management practice, but others looked upon it as a financial penalty. However, most farmers thought increased observation of the animals is good for farm assurance schemes, and increases the credibility of assessors.

Overall, the assessors' opinions indicated that the outcome measures were considered useful for identifying the welfare status of the farm. In comparison with the current farm assurance schemes, some assessors suggested that Welfare Quality® is more scientific, but that practicality must also be considered. However, as for being used in current farm assurance schemes, the assessors were less than positive about the practical application, in particular with respect to behavioural observations, although some assessors recommended that parts of Welfare Quality® could be included in the current farm assurance schemes giving evidence to the standards.

Discussion

The results reported in this paper present challenges to the Welfare Quality® protocol. These challenges are associated with three main areas: missing data, the way in which the measures are aggregated, and the reliability of the data.

Missing data

Perhaps the most unexpected finding of this study was the amount of data, that the assessors failed to collect, as shown in Tables 1 and 2. The implications of these missing data are shown in Table 3, where an overall classification was only able to be calculated for 7% of farms. This was because, in order to use the Welfare Quality® scoring system, it is necessary to have a complete data set, so that even a single missing value from an assessment prevents the calculation of the overall classification. All body condition scores for one assessor were excluded on the basis of poor inter-assessor reliability. Beyond this, specific reasons why data were missing in this study are unclear, as no explanations were recorded at the time of collection. Table 4 shows that completion rate varied by both section of the assessment and by assessor. While some assessors completed all measures in some sections, they also showed the lowest completion rate in other sections, and not one of the assessors completed all the assessments. One possible reason for this is that beyond the initial training, there was no follow-up for non-completed assessments. This is not

likely to have been the case in a research setting where data entry is monitored frequently, and where supervision is normally available, and this may explain the difference between the amount of missing data in this study and that of de Vries *et al* (2013). Given that missing data invalidates the Welfare Quality® protocol in terms of calculating aggregated scores, if the assessments are to be conducted in a non-academic setting, a robust system of verification and follow up would be recommended.

Further to the data collection itself, there is no provision in the protocol for handling missing data. In a research setting, in order to analyse datasets, it is important for contingencies to be in place for handling missing data. The methods most accessible to researchers, such as mean substitution, are inadequate, and distort vital parameters associated with the data, (Schafer & Graham 2002). As a reliable means of approaching the issue of missing data does not exist, it would be important to investigate other methods such as those of multiple imputation to this end.

Missing data, in this study, can be understood to be an issue associated with both the protocol failing to accommodate missing values, and the management of assessors. With respect to the use of the protocol in farm assurance, in order for the standards set by certification to be feasible, a recommendation might be to refrain from carrying out interventions based on the criteria or principle scores, or at the overall level of classification, and instead to examine the measures individually. Additionally, by ensuring higher assessor compliance, it is anticipated that the amount of missing data would be reduced.

The structure of the protocol

The Welfare Quality® protocol for dairy cows comprises 67 measures. Hence, if a full assessment is to be carried out, it might be expected that the most efficient means of managing these data is to aggregate the measures according to the protocol. The way in which the measures are represented at the different levels of the scoring system is therefore of practical importance.

The hypothetical 'interventions' analysed in this study were selected to reflect improvements which could be carried out on-farm. The aim was to explore the effects of optimising the scores of certain measures on calculated values at higher levels of the assessment (criteria and principle scores). Unfortunately, due to missing data, it was not possible to look at the effect of these interventions on the overall classification. The interventions that had the greatest impact in this study were those associated with *Good feeding*, *Good housing* and *Positive emotional state*, while the independent single measure interventions for *Absence of disease* achieved little or no impact at the criteria level and no effect at the principle level.

Mastitis and lameness have been identified as two of the most financially costly endemic diseases affecting dairy cows in the UK (Bennett *et al* 1999), and cause substantial problems for the UK dairy industry (Whitaker *et al* 2004). They also have significant consequences for health and welfare (FAWC 1997). However, it is clear that interventions on mastitis or lameness are not reflected, to any large extent, in the aggregated scores, in spite of the welfare implications of these diseases. An *enhanced* farm in this study had 89% *mastitis*, and an *acceptable* farm had 30% *severely lame cows*. A consequence of the fact that the construction of the overall classification is not sensitive to these measures, is that it has the potential to create a dilemma for farmers between, on the one hand, addressing important economic welfare concerns and, on the other, effecting changes that would strategically improve the aggregated scores. This could shift the emphasis away from outcome-based measures, such as those associated with lameness and mastitis, towards resource-based measures in order to have a greater effect on the aggregated scores. Given the importance of lameness and mastitis for welfare, and that these conditions are obscured by the aggregated scores, it is proposed that the success of a certification scheme should not reasonably be measured in terms of the overall Welfare Quality® classification, or even the principle scores. This is because, as shown above, a high level of overall classification may obscure important welfare issues associated with mastitis or lameness, and at the scheme level, this could not only have negative consequences for large numbers of farms, but could also mislead stakeholders into believing that with a high level of overall classification, there are no such welfare concerns, which may not necessarily be the case.

Reliability

The measures included in the Welfare Quality® protocol have been tested for inter- and intra- assessor agreement by researchers (Blokhuis *et al* 2010). This study tested inter-assessor reliability amongst certification scheme assessors on different outcome measures over three sessions.

It is difficult to comment on any improvement shown as the sample of assessors was different in each training session. While some studies have shown training to increase levels of inter-assessor agreement (Kristensen *et al* 2006; Brenninkmeyer *et al* 2007; March *et al* 2007; D'Eath 2012), the effect of training has also been shown to have an inconsistent effect on different assessors (Engel *et al* 2003; Thomsen *et al* 2008; Mullan *et al* 2011).

The third session in this study showed high levels of agreement between some assessors with the mode, but on two measures, ('percentage of cows with dirty udders' and 'percentage of cows with at least one lesion') it could reasonably be argued that the levels of agreement for all the assessors were not close enough to achieve the consistency needed by farm assurance schemes. These findings are in line with Mullan *et al* (2011), who in looking at the scoring of outcome measures in pigs, also found insufficient levels of agreement after three sessions. Gibbons *et al* (2012), on the other hand, were able to report high inter-assessor agreement for injury scoring of dairy cows following a similar training programme to that used in this study and by applying a pass/fail policy for individual measures. As was the case for the assessor whose scores for 'body condition' were omitted in this study, this approach is not appropriate for the Welfare Quality® protocol where it is necessary for assessors to collect all measures.

https://doi.org/10.7120/09627286.23.1.095 Published online by Cambridge University Press

For certification, consistency of assessment is fundamental for the credibility of any farm assurance scheme, and Mullan *et al* (2011) have raised a couple of points with regards to this. Firstly, assessors tend to work in specific geographic areas over several years, so the potential for over or underscoring a farm could have long-term consequences. For farms which are consistently over-scored, there is a potential for welfare issues to be overlooked, whereas for those that are under-scored, these farms may appear to be performing more poorly than they are and, as a result, may potentially suffer financial penalties. Secondly, while some degree of variation is inevitable, the importance of scoring with sufficiently high levels of agreement is important for providing feedback to producers, benchmarking, and delivering a pass/fail judgment. Assessing compliance with a set of standards is very different from making a scientific evaluation of the welfare state of the animals on a farm, requiring a very different level of discrimination. The consequences of measurement error inherent in the scoring of welfare outcomes suggest that it may not be possible to pass or fail individual farms based on assessment results, however the results can be used as a basis for further investigation. The difference with making a scientific evaluation is that this will often encompass large numbers of farms where measurement error can be accounted for, and the area of interest will lie not in a pass/fail binary outcome but in incremental responses to treatment effects. The challenge here is in finding ways that satisfactory levels of agreement can be managed.

Other challenges

In the qualitative feedback session some practical issues were discussed concerning the feasibility of data collection. Firstly, the avoidance distance required the presence of a food bunk. The assessors reported that not all farms had feed bunks, and that alternative feed set-ups failed to provide any meaningful results. This raises a specific issue for carrying out Welfare Quality® assessments in the UK, as the Red Tractor Farm Assurance Dairy Scheme, which certifies approximately 85% of dairy farms in the UK reports that only approximately three-quarters of its 11,100 farms have this type of feed system (D Kennedy, personal communication 2013). Furthermore, there was concern over the use of the measures themselves, as they were reported to be influenced by the behaviour of adjacent cows, as well as by the age of the cows. The guidance for carrying out the avoidance distance test stipulates that:

> Neighbouring animals that react to an animal being tested should be tested later on. In order to reduce the risk of influencing the neighbour's result, every second animal can be chosen. (Welfare Quality® 2009a).

While this addresses the concern relating to the behaviour of adjacent cows, no guidance is provided with regards to the age of the cows that are being assessed. Haskell *et al* (2012) report increasing approachability of young dairy cows with age and as a consequence of this recommend that cows of different age groups are included in samples used to assess on-farm animal welfare. A further concern with the avoidance distance test is that shorter avoidance may not always be representative of good welfare, as shorter distances have been reported as being associated with lameness in dairy cows (Spinka *et al* 2005) and in broilers (de Jong *et al* 2011).

A second concern was that the assessors remarked that time of year of the assessment affected several measures including those relating to cleanliness, avoidance distance, *Duration of lying down movements*, and *Frequency of coughing per cow per 15 min*. A single reference is made to this in the protocol on p 13:

> The protocols for cattle have been developed for intensive housing systems (Welfare Quality® 2009a).

As intensive housing systems are in the minority in the UK, this suggests that perhaps the relevance of the Welfare Quality® protocol in the UK may be limited, and hence also its potential as a certification tool.

Finally, the assessors also commented on the length of time taken to carry out an assessment, specifically, the time taken to carry out behavioural observations, locate the cows, and move into position to observe the appropriate behaviour or take the appropriate measures. Although the lengthy assessment time is an issue which has been recognised in the literature (Knierim & Winckler 2009), a study by de Vries *et al* (2013) found:

> ...little scope for reduction of on-farm assessment time of the Welfare Quality® protocol for dairy cattle.

Feedback from the assessors suggested that, in general, although the outcome measures discussed in the meeting were considered to be useful, they were not practical to include as part of farm assessment. However, it must be acknowledged that leaving out these measures would inevitably result in an assessment which would be considerably less comprehensive.

Conclusion

The Welfare Quality® protocols provide a scientifically robust, mainly outcome-based means of welfare assessment. While there is a growing body of literature concerning the science behind Welfare Quality®, there is little on how it performs outside of a research setting. This paper has evaluated a Welfare Quality® assessment carried out by experienced assessors, and identified three main challenges: missing data, the structure of the protocol and consistency of scoring by assessors. Suggestions for managing these challenges include monitoring the assessors, refraining from aggregating the measures and instead, using the measures independently, and the development of statistical approaches to accommodate missing data. The aim of this study was to evaluate the potential for uptake of the Welfare Quality® protocol by farm assurance schemes. As a result of this study the data from the individual measures have informed the selection of measures for inclusion in the AssureWel dairy cow welfare assessment protocol for use by the Freedom Food and Soil Association Certification farm assurance schemes (Main *et al* 2012). However, there are no plans for further uptake beyond this role as a surveillance tool by these two schemes.

Acknowledgements

This project was funded by DairyCo 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. We are

very grateful for the contributions made to this study by 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. Furthermore, we also thank Dr Isabelle Veissier, for help with inaccuracies in the printed protocol, Dr Jenny Gibbons for help with drafting and the two anonymous referees for their constructive comments as part of the peer review process.

References

Anneberg I, Vaarst M, and Sørensen JT 2012 The experience of animal welfare inspections as perceived by Danish livestock farmers: a qualitative research approach. *Livestock Science 147(1-3)*: 49-58. http://dx.doi.org/10.1016/j.livsci.2012.03.018

Bennett R, Christiansen K, Clifton-Hadley R and Barker CE 1999 Preliminary estimates of the direct costs associated with endemic diseases of livestock in Great Britain. *Preventive Veterinary Medicine 39*: 155-171. http://dx.doi.org/10.1016/S0167-5877(99)00003-3

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

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

Botreau R, Veissier I and Perny P 2009 Overall assessment of animal welfare: strategy adopted in Welfare Quality® . *Animal Welfare 18*: 363-370

Braun V and Clarke V 2006 Using thematic analysis in psychology. *Qualitative Research in Psychology 3(2)*: 77-101. http://dx.doi.org/10.1191/1478088706qp063oa

Brenninkmeyer C, Dippel S, March S, Brinkmann J, Winckler C and Knierim U 2007 Reliability of a subjective lameness scoring system for dairy cows. *Animal Welfare 16*: 127-129

Dawkins MS 2003 Behaviour as a tool in the assessment of animal welfare. *Zoology 106*: 383-387. http://dx.doi.org/10.1078/0944-2006-00122

D'Eath RB 2012 Repeated locomotion scoring of a sow herd to measure lameness: consistency over time, the effect of sow characteristics and inter-observer reliability. *Animal Welfare 21*: 219-231. http://dx.doi.org/10.7120/09627286.21.2.219

de Jong IC, Moya TP, Gunnink H, van den Heuvel H, Hindle VA, Mul M and van Reenen K 2011 *Simplifying the Welfare Quality assessment protocol for broilers Report 533*. Wageningen UR Livestock Research: Lelystad, The Netherlands

DeSantis and Ugarriza DN 2000 The concept of theme as used in qualitative nursing research. *Western Journal of Nursing Research 22(3)*: 351-372

de Vries M, Engel B, den Uijl I, van Schaik G, Dijkstra T, de Boer IJM and Bokkers EAM 2013 Assessment time of the Welfare Quality® protocol for dairy cattle. *Animal Welfare 22*: 85-93. http://dx.doi.org/10.7120/09627286.22.1.085

Engel B, Bruin G, Andre G and Buist W 2003 Assessment of observer performance in a subjective scoring system: visual classification of the gait of cows**.** *The Journal of Agricultural Science 140*: 317-333. http://dx.doi.org/10.1017/S0021859603002983

Farm Animal Welfare Council 1997 *Report on the Welfare of Dairy Cattle*. FAWC: London, UK

Fraser D 1995 Science, values and animal welfare: exploring the inextricable connection. *Animal Welfare 5*: 103-117

Gibbons J, Vasseur E, Rushen J and de Passillé AM 2012 A training programme to ensure high repeatability of injury scoring of dairy cows. *Animal Welfare 21*: 379-388. http://dx.doi.org/10.7120/09627286.21.3.379

Haskell MJ, Bell DJ and Gibbons JM 2012 Is the response to humans consistent over productive life in dairy cows? *Animal Welfare 21*: 319-324. http://dx.doi.org/10.7120/09627286.21.3.319

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

Kristensen E, Dueholm L, Vink D, Andersen JE, Jakobsen EB, Illum-Nielsen S, Petersen FA and Enevoldsen C 2006 Within- and across-person uniformity of body condition scoring in Danish Holstein cattle. *Journal of Dairy Science 89*: 3721-3728. http://dx.doi.org/10.3168/jds.S0022-0302(06)72413-4

March S, Brinkmann J and Winkler C 2007 Effect of training on the inter-observer reliability of lameness scoring in dairy cattle. *Animal Welfare 16*: 131-133

Main DCJ, Kent JP, Wemelsfelder F, Ofner E and Tuyttens FAM 2003 Applications for methods of on-farm welfare assessment. *Animal Welfare 12*: 523-528

Main DCJ, Rogerson I, Crawley MC, Avizenius J, Fraser A and Mullan S 2012 Welfare outcomes assessment in dairy farm assurance schemes. *Cattle Practice 20(2)*: 142-145

Miele M, Veissier I, Evans A and Botreau R 2011 Animal welfare: establishing a dialogue between science and society. *Animal Welfare 20*: 103-117

Mullan S, Edwards SA, Butterworth A, Whay HR and Main DCJ 2011 Inter-observer reliability testing of pig welfare outcome measures proposed for inclusion within farm assurance schemes. *The Veterinary Journal 190*: e100-e109. http://dx.doi.org/10.1016/j.tvjl.2011.01.012

Roe E, Buller H and Bull J 2011 The performance of farm animal assessment. *Animal Welfare 20(1)*: 69-78

Schafer JL and Graham JW 2002 Missing data: our view of the State of the Art. *Psychological Methods 7(2)*: 147-177. http://dx.doi.org/10.1037/1082-989X.7.2.147

Špinka M, Dembele I, Panamá J and Stěhulová I 2005 Lame dairy cows have shorter avoidance distances. *Proceedings of the 39th International Congress of the International Society for Applied Ethology*. 20-24 August 2005, Sagamihara, Japan

Thomsen PT, Munksgaard L and Tøgersen FA 2008 Evaluation of a lameness scoring system for dairy cows *91(1)*: 119-126

Webster AJF, Main DCJ and Whay HR 2004 Welfare assessment: indices from clinical observation. *Animal Welfare 13*: S93-98

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

https://doi.org/10.7120/09627286.23.1.095 Published online by Cambridge University Press

Welfare Quality® 2009b *Training in the Welfare Quality® Assessment Protocols*. https://www.google.co.uk/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CD0QFjAA&url=http%3A%2F%2Fwww.welfarequality.net%2Fdownloadattachment%2F43148%2F20260%2F2009-Nov_adapted_leaflet%2520WQ%2520Training%2520Assessment%2520Protocols.pdf&ei=3SQuUbKfGo2V0QXbnoHoDg&usg=AFQjCNGwrRlKVrcpdOOHcV5mlUlwc0thvA&bvm=bv.42965579,d.d2k

Welfare Quality® 2012 *Dairy cow on farm parameters*. http://www1.clermont.inra.fr/wq/pdf/WQ_Dairy%20cows%20on%20farm_parameters.pdf

Whitaker DA, Macrae AI and Burrough E 2004 Disposal and disease rates in British dairy herds between April 1998 and March 2002. *Veterinary Record 155*: 43-47. http://dx.doi.org/10.1136/vr.155.2.43