Animal (2014), **8:3**, pp 461–469 © The Animal Consortium 2013
doi:10.1017/S1751731113002267

On-farm animal welfare assessment in beef bulls: consistency over time of single measures and aggregated Welfare Quality® scores

M. K. Kirchner[1†], H. Schulze Westerath[2a], U. Knierim[2], E. Tessitore[3], G. Cozzi[3] and C. Winckler[1]

[1]*Department of Sustainable Agricultural Systems, University of Natural Resources and Life Sciences (BOKU), Gregor-Mendel-Strasse 33, A–1180 Vienna, Austria;* [2]*Farm Animal Behaviour and Husbandry Section, University of Kassel/Witzenhausen, Nordbahnhofstr. 1a, D-37213 Witzenhausen, Germany;* [3]*Department of Animal Medicine, Production and Health, University of Padova, Agripolis - Viale dell'Università 16, I-35040 Legnaro (PD), Italy*

(Received 24 January 2013; Accepted 18 November 2013; First published online 13 December 2013)

Consistency over time of (on-farm) animal welfare assessment systems forms part of reliability, meaning that results of the assessment should be representative of the longer-term welfare state of the farm as long as the housing and management conditions have not changed considerably. This is especially important if assessments are to be used for certification purposes. It was the aim of the present study to investigate consistency over time of the Welfare Quality® (WQ®) assessment system for fattening cattle at single measure level, aggregated criterion and principle scores, and overall classification across short-term (1 month) and longer-term periods (6 months). We hypothesized that consistency over time of aggregated criterion and principle scores is higher than that of single measures. Consistency was also expected to be lower with longer intervals between assessments. Data were obtained using the WQ® protocol for fattening cattle during three visits (months 0, 1 and 7) on 63 beef farms in Austria, Germany and Italy. Only data from farms where no major changes in housing and management had taken place were considered for analysis. At the single measure level, Spearman rank correlations between visits were >0.7 and variance was lower within farms than between farms for six and two of 19 measures after 1 month and 6 months, respectively. After aggregation of single measures into criterion and principle scores, five and two of 10 criteria and three and one of four principles were found reliable after 1 and 6 months, respectively. At the WQ® principle level, this was the case for three and one of four principles. Seventy-nine per cent and 75% of the farms were allocated to the same overall welfare category after 1 month and 6 months. Possible reasons for a lack of consistency are seasonal effects or short-term fluctuations that occur under normal farm conditions, low prevalence of clinical measures and probably insufficient sample size, whereas poor inter-observer agreement leading to inflation of correlation can be ruled out. At the criterion and principle level, aggregation of information into scores appears to partly smoothen undirected variation at the single measure level without losing sensitivity in terms of welfare evaluation. Reliable on-farm animal welfare assessments should therefore be based on repeated assessments. Further long-term studies are recommended to better understand the factors influencing consistency over time.

Keywords: beef cattle, animal-based measures, on-farm assessment, welfare, consistency

Implication

The Welfare Quality® (WQ®) assessment system for fattening cattle mainly uses animal-based parameters for on-farm evaluation, which can be aggregated stepwise to an overall farm score. This paper investigated WQ® scores resulting from consecutive farm assessments in terms of consistency of animal- and resource-based measures. Consistency over time was partially low and further decreased with longer intervals between assessments. Especially when used for certification and labelling purposes, animal welfare classification should therefore not be based on single assessments. Possible reasons for lack of consistency are discussed and recommendations for future approaches are given.

Introduction

One of the main aims of Welfare Quality® (WQ®) was to develop on-farm welfare assessment systems that focus primarily on animal-based measures and that are scientifically

[a] Present address: Animal Behaviour, Health and Welfare Unit, Institute of Agricultural Sciences, ETH Zürich, Universitätstr. 2, CH-8092 Zürich, Switzerland
[†] E-mail: marlene.kirchner@boku.ac.at

sound and feasible (Blokhuis *et al.*, 2003). The assessment protocols that have been developed for several animal species and categories aim at providing feedback to farm managers on the welfare state of their animals and at translating this information into understandable messages for consumers, for example, for labelling purposes. Following a multidimensional concept of animal welfare, the protocols comprise several animal-based welfare parameters that are complemented with information on farm management and housing structures (Botreau *et al.*, 2007; Blokhuis, 2008; Kirchner *et al.*, submitted). The protocol for fattening cattle is based on 28 measures that are integrated into the so-called criterion and principle scores, finally leading to an overall classification of the farm (Botreau *et al.*, 2009; Welfare Quality®, 2009).

Apart from validity and feasibility, reliability is a central criterion for the selection of measures for (on-farm) welfare assessment (Waiblinger *et al.*, 2001). Reliability relates to the reproducibility or repeatability of results and this can be considered at different levels. Interobserver reliability refers to the agreement between two or more observers assessing the same animals in the same situation. Intraobserver reliability means the extent of agreement when the same observer carries out assessments repeatedly, for example, using video clips or pictures (Martin and Bateson, 2007). Test–retest reliability indicates that repeated tests with the same subjects produce similar data (Windschnurer *et al.*, 2009).

A special case of test–retest reliability is the consistency of outcomes over time (COT). COT is especially important if assessments are intended to be used for certification purposes, meaning that results should be representative of the longer-term farm situation and not too sensitive to changes in the farming conditions or the internal states of the animals as long as the situation has not changed significantly. High levels of consistency have therefore been regarded essential for on-farm welfare measures and assessment systems (Capdeville and Vessier, 2001; Winckler *et al.*, 2007). They will ensure fairness for the farmer and credibility of the system (Knierim and Winckler, 2009; Sørensen and Fraser, 2010). At the same time, an on-farm measure should be sensitive enough to detect variations in welfare state within farms. An additional aspect of COT is the determination of the necessary number of repeated assessments, that is, of farm visits. Indicators that do not change significantly over a long period of time, if farm conditions remain constant, do not require frequent visits to obtain reliable estimates.

COT has been investigated with regard to individual differences in behaviour (reviewed in Bell *et al.*, 2009), for example, for 'behavioural traits' (Spoolder *et al.*, 1996; Kralj-Fišer *et al.*, 2007), but studies on consistency of both animal- and resource-based (on-farm) measures of animal welfare are generally rare (Sundrum and Rubelowski, 2001). In dairy cattle, Winckler *et al.* (2007) investigated the correlation of selected animal-related welfare parameters between consecutive farm visits. Consistency was found to be moderate to good with regard to lameness incidence, skin lesions at the tarsal joint and avoidance distance towards an approaching human. However, variability for measures of social behaviour and animal cleanliness was high. The development of the WQ® assessment protocol for fattening cattle (Welfare Quality®, 2009) also included some work on COT at the measure level, focussing on behavioural measures such as behaviours around resting (e.g. time needed to lie down; Brörkens *et al.*, 2009) or agonistic and socio-positive interactions (Laister *et al.*, 2009; Schulze Westerath *et al.*, 2009). Measures were suggested for inclusion in the protocol only if correlations between farm visits were higher than 0.7 and variance within farms was lower than variance between farms. However, not all measures included in the final protocol were tested for COT and so far such studies have not been carried out at the level of criterion and principle scores. The latter is a key factor for credibility of the welfare judgement and it has been highlighted as an important task and perspective for further investigations (Knierim and Winckler, 2009).

It was therefore the aim of the present study to investigate COT for the single measures used in the WQ® protocol for fattening cattle as well as for the aggregated criterion and principle scores across short-term (1 month) and longer-term periods (6 months). It was hypothesized that the aggregated criterion and principle scores were more consistent over time than the single measures. Furthermore, consistency was expected to be lower with longer intervals between assessments.

Material and methods

Data collection was carried out on 63 bull-fattening farms with alternative housing systems, that is, with straw-bedded lying areas or cubicles with rubber mats. The farms were located in Austria ($n = 21$), Germany ($n = 21$) and Italy ($n = 21$). The average number of animals per farm ranged between 102 (Austria) and 233 (Italy) (for details see Kirchner *et al.*, submitted and Supplementary Table S1). Three assessments were carried out on each farm. An interim assessment was carried out about 1 month after the initial assessment and the final assessment took place about 6 months after the interim assessment. All assessments followed the WQ® assessment protocol for fattening cattle (Welfare Quality®, 2009) and they were carried out by three trained assessors (one per country), who had reached at least satisfying interobserver agreement (Kendall's coefficient of concordance >0.7) during a joint training session.

On the basis of total 28 measures (19 animal, three resource and six management based), WQ® criterion and principle scores were calculated (for an overview on achieved values on all levels, please consider Supplementary Tables S1 to S5). These scores can range from 0 to 100 for each criterion and principle and were calculated for each farm at each assessment separately as described in the WQ® protocol (Welfare Quality®, 2009) using updated formulas and coefficients (Welfare Quality®, 2011).

Statistics

Two methods were used to assess consistency of results over time for the level of measures, criterion and principle scores.

First, Spearman's rank correlations between initial and interim assessment and between interim and final assessment were calculated. Second, between- and within-farm variability was compared using covariance parameter estimates for the farm (random factor) and the residual component. Analysis was based on the following linear mixed effects model:

$$y_{ijkl} = \mu + b_i + \alpha_k + \beta_l + \varepsilon_{ijkl}$$

with the intercept μ, the fixed effects α_k assessment (factor with two levels: initial or interim assessment and interim or final assessment, respectively), β_l country (factor with three levels: AT, DE and IT) and the random effect b_i farm. All models were computed using R 2.13.1 (R Development Core Team, 2008). Residuals were checked for normal distribution (Kolmogorov–Smirnov test; Q–Q, scatter and box plots). Frequencies of the measures '% of animals with severe integument alterations', '% of animals with bloated rumen', '% of animals died or euthanized on farm' and of the resource-based measures '% of groups with dirty water point', and '% of days outdoor loafing area available' were too low for statistical analysis.

For the measures '% of very lean animals', '% of animals with mild integument alterations', '% of animals with hampered respiration' and '% of animals with diarrhoea' as well as the criterion scores regarding 'Absence of prolonged hunger', 'Absence of prolonged thirst' and 'Absence of pain induced by management procedures' normal distribution of residuals was not achieved after transformation. Therefore, a generalized linear mixed model in R 2.13.1 (R Development Core Team, 2008) assuming a Poisson distribution was used; the factors included in the model were the same as described above.

Consistency of measures, criterion scores and principle scores was judged as acceptable if correlation coefficients were equal or higher than 0.7 (Martin and Bateson, 2007) and if the variance explained by the random factor (= variance between farms) was greater than the variance of the residuals describing the variance within farm (thus corresponding to an intra-class correlation coefficient >0.5; Dohoo *et al.*, 2009). This evaluation was only possible for 19 of the 28 measures and 10 of the 11 criteria (see also Tables 1 and 2).

To investigate consistency only in farms that provided rather stable conditions for the animals, farms that showed major deviations throughout the study were excluded from the calculations. 'Major deviations' were defined as alterations of resources or management exceeding changes that can be commonly expected over time. Examples for such management changes are 'substantial increase in amount of litter used', 'increased number of animals bought in' or 'switch to total mixed ration feeding'. Owing to 'major deviations', in total, 16 (AT: 9, DE: 5, IT: 2) of the 189 assessments had to be excluded from statistical analysis. Consequently, 59 farmvisits were included in the analyses comprising initial and interim assessments, and 49 farmvisits for interim and final assessments.

Finally, the proportion of farms in the four WQ® welfare classification categories (Excellent, Enhanced, Acceptable and Not classified) were determined, and the percentage agreement between initial and interim, interim and final, and initial and final assessment was calculated for those farms that had not undergone substantial changes as described above ($n = 48$).

Results

Consistency of results over time (COT) at measure, criterion and principle level

Regarding the initial and interim assessment, covariance parameter estimates for the factor farm (between-farm) were higher than for the residuals (within-farm) and correlations exceeded the 0.7 threshold for six measures (Table 1). These measures were either resource-based ('% of groups with sufficient water points', '% of groups with at least two water points, 'Space allowance') or animal-based ('Duration of lying down movements', '% of animals with ocular discharge', 'Frequency of head butts, displacements, fights and chases per animal and hour'). Measures not fulfilling both COT criteria at least matched the criterion regarding the ratio of variance within and between farms except for '% of dirty animals', '% of animals with mild integument alterations', 'Frequency of social horning and social licking/animal per hour' and the '% of animals with an avoidance distance at the feeding rack (ADF) of 50 to 100 cm' (Table 1). The respective correlation coefficients varied largely ranging from 0.03 for '% of animals with hampered respiration' to 0.68 for the weighted sum score obtained from Qualitative Behaviour Assessment ('Positive emotional state').

Two measures regarding the provision of water showed high consistency between the interim and final assessment. All the other measures did not fulfil at least one of the predefined consistency criteria. Greater between-farm than within-farm variance was found for measures such as '% of very lean animals', '% of animals with diarrhoea' or measures belonging to the principle 'Appropriate behaviour' (e.g. '% of animals with an avoidance distance of >100 cm'); however, correlation coefficients ranged from 0.19 to 0.67 (Table 1).

At the criterion level, the four criteria 'Comfort around resting', 'Ease of movement', 'Absence of disease' and 'Expression of social behaviours' showed larger variance between farms than within farms and additionally correlations between initial and interim assessment exceeded the threshold of 0.70 (Table 2). Partial fulfilment of the COT criteria (comparison of between- and within-farm variance) was achieved for the three criteria 'Absence of injuries', 'Good human–animal relationship' and 'Positive emotional state'. Regarding interim and final assessment, with 6 months in between, two criteria were judged as having good COT: 'Ease of movement' and 'Good human–animal relationship' (Table 2). The variance criterion was fulfilled by only two other welfare criteria: 'Absence of injuries' and 'Positive emotional state'. Variance within farms was smaller than between farms for all principle scores. However, correlation coefficients were above threshold only for three principles (Good feeding, Good housing and Good health) from initial to interim assessments and for one principle from interim to final assessments (Good feeding) (Table 3).

Table 1 *Consistency of the welfare measures from initial to interim assessment and interim to final assessment of the farms without 'major deviations' expressed as correlations between the welfare measures (r_S), as effect of assessment (ASS) and/or country (CO) on the welfare measure and comparison of the variance within and between farms*

		Initial to interim assessment ($n = 59$)			Interim to final assessment ($n = 49$)		
Criteria	Measures	r_S[a]	Effect of assessment and country[b]	Variance within farm < between farms	r_S[a]	Effect of assessment and country[b]	Variance within farm < between farms
Absence of prolonged hunger	% of very lean animals	0.27*	ns	Yes	0.68***	ns	Yes
Absence of prolonged thirst	**% of groups with sufficient water points**	**0.95*****	CO	**Yes**	**0.91*****	CO	**Yes**
	% of groups with dirty water points	Too rare for analysis → n.t.			Too rare for analysis → n.t.		
	% of groups with at least two water points	**0.94*****	ns	**Yes**	**0.94*****	ns	**Yes**
Comfort around resting	**Duration of lying down** (interim to final: $n = 42$)	**0.70*****	CO	**Yes**	0.45**	ASS, CO	No
	% of dirty animals	0.74***	CO	No	0.47***	CO	No
Ease of movement	**Space allowance in m²/700 kg**	**0.80*****	ns	**Yes**	0.52***	ns	No
	% of days outdoor loafing area (or pasture) available (at least 1 h/day)	Too rare for analysis → n.t.			Too rare for analysis → n.t.		
Absence of injuries	% of lame animals	0.36**	CO	Yes	0.39**	CO	No
	% of animals with mild integument alterations	0.72***	ASS, CO	No	0.44**	ASS	No
	% of animals with severe integument alterations	Too rare for analysis → n.t.			Too rare for analysis → n.t.		
Absence of diseases	% of animals with nasal discharge	0.48***	ASS, CO	Yes	0.15	ASS	No
	% of animals with ocular discharge	**0.73*****	ASS, CO	**Yes**	0.69***	ASS, CO	No
	% of animals with hampered respiration	0.03	CO	Yes	0.19	CO	Yes
	number of coughs per animal and hour	0.55***	ns	Yes	0.27	ASS	No
	% of animals with bloated rumen	Too rare for analysis → n.t.			Too rare for analysis → n.t.		
	% of animals with diarrhoea	0.43***	ASS, CO	Yes	0.63***	ASS, CO	Yes
	% of dead animals during a year	Too rare for analysis → n.t.			Too rare for analysis → n.t.		
Absence of pain induced by management procedures	% of disbudded/dehorned animals and disbudding per dehorning procedures	Qualitative measure → n.t.			Qualitative measure → n.t.		
	% of tail-docked animals and tail-docking procedures	Qualitative measure → n.t.			Qualitative measure → n.t.		
	% of castrated animals and castration procedures	Not occurring → n.t.			Not occurring → n.t.		
Expression of social behaviours	**Frequency of head butts, displacements, fights and chases per animal and hour**	**0.74*****	ASS, CO	**Yes**	0.48***	CO	No
	Frequency of social horning and social licking/animal per hour	0.55***	ns	No	0.51***	ns	No
Expression of other behaviour	Access to pasture before fattening in months, % of days pasture available (at least 6 h/day)	Not occurring → n.t.			Not occurring → n.t.		
Good animal-human-relationship	% of animals with an ADF of >100 cm[c]	0.46	ASS, CO	Yes	0.55***	ASS	Yes
	% of animals with an ADF 50 to 100 cm	0.29*	ns	No	0.67***	ASS, CO	Yes
	% of animals with an ADF <50 cm but not be touched	0.54***	CO	Yes	0.47***	CO	No
Positive emotional state	weighted sum score obtained from Qualitative Behaviour Assessment	0.68***	ns	Yes	0.53***	ns	Yes

ns = not significant; n.t. = not tested.
Measurements in bold letters met the COT criteria: correlation coefficients equal or higher than 0.7 and the variance explained by the random factor (= variance between farms) greater than the variance of the residuals describing the within-farm variance.
[a]Spearman's correlation coefficient r_S.
[b]Significant effect of assessment (ASS) and country (CO).
[c]Avoidance distance at feed rack (ADF).
*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$.

Table 2 *Consistency of the welfare criteria scores from initial to interim assessment and interim to final assessment of the farms without 'major deviations', expressed as correlations between the welfare criteria scores (r_S), as effect of assessment and/or country on the criteria scores and comparison of the variance within and between farms*

		Initial to interim assessment ($n = 59$)			Interim to final assessment ($n = 49$)		
Measures	Criteria	r_S[a]	Effect of assessment and country[b]	Variance within farm < between farms	r_S[a]	Effect of assessment and country[b]	Variance within farm < between farms
% of very lean animals	Absence of prolonged hunger	0.27*	ns	No	0.68***	ns	No
% groups with sufficient WP	Absence of prolonged thirst	0.67***	ASS	No	0.73***	ns	No
% of groups with dirty WP							
% of groups with at least two WP							
Duration of lying down	**Comfort around resting**	**0.70*** **	CO	**Yes**	0.43***	CO	No
% of dirty animals							
Space allowance per animal/700 kg	**Ease of movement**	**0.73*** **	ns	**Yes**	**0.70*** **	**ns**	**Yes**
Access to OLA							
% of lame animals	Absence of injuries	0.57***	CO	Yes	0.53***	CO	Yes
% of animals with mild and severe alterations							
% of animals with nasal discharge and ocular discharge	**Absence of diseases**	**0.76*** **	CO	**Yes**	0.55***	CO	No
Number of coughs/animal per hour; % animals with hampered respiration							
% of animals with bloated rumen and diarrhoea							
% of dead animals during a year							
Disbudding/dehorning procedures	Absence of pain induced by management procedures	0.92***	ns	No	0.99***	ns	No
Tail-docking procedures							
Frequency of agonistic and socio-positive interactions/animal per hour	**Expression of social behaviours**	**0.78*** **	CO	**Yes**	0.56***	CO	No
Access to pasture	Expression of other behaviours	n.t.	n.t.	n.t.	n.t.	n.t.	n.t.
ADF >100 cm, 50 to 100 cm and <50 cm	**Good human–animal relationship**	0.61***	ASS, CO	Yes	**0.72*** **	ASS, CO	**Yes**
Weighted sum score obtained from Qualitative Behaviour Assessment	Positive emotional state	0.68***	ns	Yes	0.53***	ns	Yes

WP = water points; OLA = outdoor loafing area; ns = not significant; n.t. = not tested.
Criteria in bold letters met the COT criteria: correlation coefficients equal or higher than 0.7 and the variance explained by the random factor (= variance between farms) greater than the variance of the residuals describing the within-farm variance.
[a]Spearman's correlation coefficient r_S.
[b]Significant effect of assessment (ASS) and country (CO).
*$P < 0.05$, ***$P < 0.001$.

Table 3 *Consistency of the welfare principle scores from initial to interim assessment and interim to final assessment of the farms without 'major deviations', expressed as correlations between the welfare principle scores (r_s), as effect of assessment and/or country on the scores and the comparison of the variance within and between farms*

		Initial to interim assessment ($n = 59$)			Interim to final assessment ($n = 49$)		
Criteria	Principles	r_S[a]	Effect of assessment and country[b]	Variance within farm < between farms	r_S[a]	Effect of assessment and country[b]	Variance within farm < between farms
Absence of prolonged hunger	**Good feeding**	**0.72*****	ns	**Yes**	**0.74*****	ns	**Yes**
Absence of prolonged thirst							
Comfort around resting	**Good housing**	**0.79*****	ns	**Yes**	0.60***	ns	Yes
Ease of movement							
Absence of injuries	**Good health**	**0.82*****	CO	**Yes**	0.52***	CO	Yes
Absence of diseases							
Absence of pain induced by management procedures							
Expression of social behaviours	Appropriate behaviour	0.69***	ns	Yes	0.41**	ns	Yes
Expression of social other behaviour							
Good human-animal relationship							
Positive emotional state							

ns = not significant.
Principles in bold letters met the COT criteria: correlation coefficients equal or higher than 0.7 and the variance explained by the random factor (= variance between farms) greater than the variance of the residuals describing the within-farm variance.
[a]Spearman's correlation coefficient r_S.
[b]Significant effect of assessment (ASS) and country (CO).
$^{**}P < 0.01$, $^{***}P < 0.001$.

Table 4 *Consistency of the welfare classifications from initial to interim assessment, interim to final assessment and initial to final assessment of the farms without 'major deviations' as 'percentage agreement of the welfare classifications' (in bold) and percentages of farms that 'improved' and 'decreased' by one or two categories (no farms switched more than two categories)*

Classification ($n = 48$)	Percentage agreement → farms with same welfare classification	Percentage farms that improved by one category	Percentage farms that improved by two categories	Percentage farms that decreased by one category	Percentage farms that decreased by two categories
Initial to interim assessment	**79**	8	0	10	2
Interim to final assessment	**75**	15	0	10	0
Initial to final assessment	**65**	13	0	10	2

Consistency of results over time concerning the overall welfare classification

Agreement in the overall welfare classification between initial and interim assessment and between interim and final assessment was 79% and 75%, respectively. Sixty-five per cent of the farms maintained the same classification from the first to the last assessment. Farms improved and decreased classification across assessments in approximately equal shares by almost exclusively one category (Table 4).

Discussion

Consistency at measures level

We used two methods to assess consistency over time (COT) of single welfare parameters and criterion/principle scores: the ratio of variance components (between and within farms) and a measure of association (rank correlations between consecutive visits). This may appear rather restrictive, but it recognizes that perfect association can exist, despite absolute differences (Watson and Petrie, 2010). In the WQ® evaluation system, the absolute values, for example, at the level of measures, rather than the relative position of farms to each other, are important. High coefficients of correlation alone may only indicate that farms would be ranked consistently irrespective of actual changes recorded for single measures, and consequently for criterion and/or principle scores. This also means that reaching the variance ratio criterion alone provides a better estimate of consistency as compared with correlations, as it can only be achieved if changes within farms are moderate.

At the measure level, COT was not satisfactory for the majority of the parameters included in the assessment scheme. In line with our expectations, and with results from Bell *et al.* (2009), correlation coefficients mostly decreased when the time interval between assessments was prolonged (6 months *v.* 1 month), whereas the ratio of variance within and between farms increased. Several reasons may account for the rather low level of consistency. Although the farms that had reported changes in housing or management or for which such changes had been identified by the assessors were discarded from the calculation of COT, management routines or handling of the animals may still have changed in some farms and thus may have had an impact on the welfare measures. However, the general impression during farm visits did not indicate that conditions substantially changed throughout the study period. Further but less well-defined factors such as seasonal effects, weather conditions or events such as visits of the veterinarian may have induced changes in the measures of welfare. Season likely played a role for measures related to respiratory disorders (nasal discharge, coughing, ocular discharge), as reflected by a significant effect of number of the assessment in the linear model of these measures. However, even less is known on short-term fluctuations and their possible effects. Behavioural measures may change at very short intervals. For example, Laister (2009) investigated the day-to-day variation in the incidences of agonistic and socio-positive behaviours on 10 Austrian beef bull farms and found correlation coefficients of single measures as low as 0.09 (but also up to 0.90). The lower value refers to horning that pertains to the social behaviours included in the WQ® protocol for fattening cattle. Although these results have to be interpreted carefully because of the small sample size, they give an idea of which day-to-day fluctuation within 'normal' farm conditions may be expected. Assessing clinical parameters in dairy cattle at five bimonthly intervals, Winckler *et al.* (2007) also found that correlations between consecutive visits varied considerably, for example, 0.48 to 0.78 for lameness and 0.05 to 0.37 for skin lesions at the carpal joint.

In the present study, besides the changes that appear to be the result of normal fluctuation of farm routines, some other possible reasons were identified. Low correlations between the initial and interim assessments occurred mainly for measures of health status with low prevalence, that is, up to 1% (principle 'Good health'). In this case, the estimated prevalences often arise from one or two animals showing the symptoms that may be either treated or removed from the fattening pens. In addition, for some clinical measures, it was difficult to achieve the proposed sample sizes, for example, owing to difficulties in approaching the bulls for the assessment. However, the sample sizes proposed in the WQ® assessment protocol (Welfare Quality®, 2009) refer to a prevalence scenario of 0.5, which can be regarded a worst case in terms of sample size (i.e. relatively larger sample sizes than would be necessary to precisely estimate a lower/higher prevalence). Thus, considering that the true prevalences were much lower in the present study, the precision reached is likely to be high. On the other hand, the calculation of sample size relates to the overall number of animals, but does not take into account the pen structure of farms. Variation in prevalence between pens may therefore lead to

a bias if the sample is drawn from selected pens only, especially in larger farms. A relatively large influence of sample size on estimated prevalence at farm level has recently been shown for selected welfare indicators in finishing pig farms using a bootstrapping method (Mullan *et al.*, 2009). Owing to large deviations among pens, even large (and not feasible) sample sizes consisting of up to 80% of the pens were unable to reflect the true situation of the whole farm for measures with low prevalence. Although sampling strategies in the present study with random selection of pens stratified for age class and location in the barn were not completely identical with those of the study on pigs (full random sampling), considering the expected prevalence when calculating appropriate sample sizes as suggested by Mullan *et al.* (2009) also deserves further investigation with regard to beef cattle.

Finally, one may argue that significant differences between countries, as revealed from the analysis of variance for the majority of measures, reflect poor interobserver agreement, thus leading to inflation of correlation. However, all assessors had achieved at least satisfactory interobserver agreement (Kendall's coefficient of concordance >0.7) and there was also no indication of observer drift in the course of the study based on interobserver testing after data collection. Furthermore, measures showing differences between countries were not necessarily highly correlated. Independent from the strength of the correlation, the distribution of data was fairly even (within country and across countries) so that we do not expect effects on the correlations because of clustering of data by country.

Resource- or management-based measures have been described to have a high repeatability (Johnsen *et al.*, 2001). However, to the best of our knowledge, COT of design or management criteria has never been tested before. In our study, the measures of water provision have been found to be consistent, but this was true to a lesser extent for the measurement of space allowance (at the 6-month interval). Stocking density is affected by management decisions depending on, for example, market conditions, the availability of animals and feed, or simply variable strategies in moving animals sooner or later into another pen. This indicates that also management-related measures may change considerably over time and cannot always be assumed as stable as they are often regarded.

COT at the measure level may be less important if the results are used for weak point analysis or for advisory activities carried out shortly after the assessment. In this case, the focus lies more on identifying the most prevalent welfare problems on a given farm. Even if outcomes fluctuate, they refer to the actual state regarding the different components of welfare. In addition, it can be assumed that problems that are not evident in a single assessment will be detected in consecutive assessments if they really represent a consistent welfare problem, affecting a meaningful proportion of animals. However, the present results demonstrate that it is inappropriate to base farm animal welfare evaluation only on single assessments at welfare measure level. In a commercial, for example, welfare-labelling setting, repeated assessments need to take place. Even if this would require a longer baseline assessment period until sufficient reliable data have been gathered, the use of rolling averages to smoothen short-term fluctuations or the verification of assessments that would cause reclassifications of farms may be a potential approach.

Consistency of aggregated scores

Consistency slightly improved with aggregation. At the criterion level, scores for four and two of the 10 criteria were considered consistent for initial to interim and interim to final assessments, respectively. Combining two or more measures into a criterion score may smoothen undirected fluctuations, for example, for the criterion 'Ease of movement'. The criterion scores were consistent even after an interval of 6 months, although the corresponding measure 'Space allowance' did not meet the criteria for consistency over time, whereas '% of days outdoor loafing area available' remained constant in the majority of the farms. Deviations in space allowance between consecutive visits were larger in farms that already provided rather spacious housing conditions (i.e. >7 m^2 per 700 kg live weight), whereas changes were less pronounced when space allowance was lower. Therefore, we do not think that the smoothening effect was at the expense of sensitivity in terms of welfare evaluation.

Similarly, sensitivity is not impaired for criterion scores that are calculated using thresholds, for example, 'Absence of disease'. If in this case prevalence stayed below the 'warning' or 'alarm' thresholds (Welfare Quality®, 2009), even relatively large variation does not lead to changes in the criterion score. On the contrary, rather small oscillations around a given threshold may lead to a substantially lower consistency at the criterion level. This was apparently the case for the criterion 'Absence of prolonged thirst'.

The general effects described for the criterion level also apply to the principle scores. Integrating information based on the multi-criteria evaluation model used by Welfare Quality® (2009) revealed sufficient consistency for three and one of four principles for initial to interim and interim to final assessments, respectively. Again, COT markedly declined over the 6-month period also for the aggregated scores, thus challenging single applications of the tool for certification purposes.

Taking the overall classification into account, 79% and 75% of the farms remained in the same category when comparing initial with interim (1 month) and interim with final assessment (6 months), respectively. Considering the fact that only farms without major changes in housing and management were used for this analysis, it is questionable whether such a high proportion of shifts between categories even within rather short periods truly reflects changes in the welfare state and would be accepted by the farming community. Although none of the farms was 'Not classified' at any assessment, switching between the two categories 'Enhanced' and 'Acceptable' might already be sufficient to lose a certification status, depending on the thresholds set by future welfare-labelling systems. Obtaining a different overall classification may, however, also reflect that not all

criteria and principles were judged consistent and that changes in welfare relevant parameters of individual farms may lead to a different overall evaluation.

Conclusions

The rather low COT at the measure level, especially for many behavioural and health measures, may be considered a minor problem if the information is given to the farmers and subsequently used for advisory purposes. However, our results demonstrate that animal welfare classification should not be based on single assessments. A higher proportion of criterion and principle scores showed promising consistency over time, indicating that the integration to criteria and principles reduces variance within farms. Nevertheless, the repeatability over 6 months was not sufficient for reliable welfare classification as it would probably be used for certification purposes. Repeated assessments appear to be necessary, either to generate rolling averages or to verify assessments that would cause reclassification of farms. Our results also provide first evidence for the differentiation of assessment intervals, depending on the expected COT for different measures. However, further long-term studies on the effects of different sample sizes and feasible sampling strategies on consistency, as well as on appropriate number, interval, schedule and analysis of repeated assessments are recommended.

Acknowledgements

The authors thank the farmers participating in the study for their collaboration and feedback during and after the assessment. The present study was part of the Welfare Quality® research project that has been co-financed by the European Commission, within the 6th Framework Programme, contract No. FOOD-CT-2004-506508. The text represents the authors' views and does not necessarily represent a position of the Commission who will not be liable for the use made of such information.

Supplementary material

To view supplementary material for this article, please visit http://dx.doi.org/10.1017/S1751731113002267.

References

Bell AM, Hankison SJ and Laskowski KL 2009. The repeatability of behaviour: a meta-analysis. Animal Behaviour 77, 771–783.

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

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

Blokhuis HJ 2008. International cooperation in animal welfare: the Welfare Quality® project. Acta Veterinaria Scandinavica 50, S10.

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

Brörkens N, Plesch G, Laister S, Zucca D, Winckler C, Minero M and Knierim U 2009. Reliability testing concerning behaviour around resting in cattle in dairy cows and beef bulls. In Welfare Quality® Reports No. 11 (ed. B, Forkman and LJ, Keeling), pp. 7–23. Welfare Quality® Consortium, Lelystad, the Netherlands.

Capdeville J and Veissier I 2001. A method of assessing welfare in loose housed dairy cows at farm level, focusing on animal observations. Acta Agriculturae Scandinavica, Section A – Animal Science 51, 62–68.

Dohoo I, Martin W and Stryhn H 2009. Intra-class correlation coefficient. In Veterinary epidemiological research (ed. I, Dohoo, W, Martin and H, Stryhn), pp. 478–479. AVC Inc, Charlottetown, Prince Edward Island, Canada.

Johnsen PF, Johannesson T and Sandøe P 2001. Assessment of farm animal welfare at herd level: many goals, many methods. Acta Agriculturae Scandinavica, Section A – Animal Science 51, 26–33.

Kirchner MK, Schulze Westerath H, Knierim U, Tessitore E, Cozzi G, Vogl C and Winckler C. Attitudes and expectations of European beef farmers towards the Welfare Quality® Assessment System. Livestock Science. Submitted.

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.

Kralj-Fiser S, Scheiber IB, Blejec A, Moestl E and Kotrschal K 2007. Individualities in a flock of free-roaming greylag geese: behavioral and physiological consistency over time and across situations. Hormones and Behaviour 51, 239–248.

Laister S 2009. Suitability of selected behavioural parameters for on-farm welfare assessment in dairy and beef cattle. Doctoral thesis, University of Natural Resources and Life Sciences, Vienna (BOKU), Austria.

Laister S, Brörkens N, Lolli S, Zucca D, Knierim U, Minero M, Canali E and Winckler C 2009. Reliability of measures of agonistic behaviour in dairy and beef cattle. In Welfare Quality® Reports No. 11 (ed. B, Forkman and LJ, Keeling), pp. 113–123. Welfare Quality® Consortium, Lelystad, the Netherlands.

Martin P and Bateson P 2007. Measuring behaviour – an introductory guide, 3rd edition Cambridge University Press, The Edinburgh Building, Cambridge, UK.

Mullan S, Browne WJ, Edwards SA, Butterworth A, Whay HR and Main DCJ 2009. The effect of sampling strategy on the estimated prevalence of welfare outcome measures on finishing pig farms. Applied Animal Behaviour Science 119, 39–48.

R Development Core Team 2008. R: a language and environment for statistical computing. ISBN 3-900051-07-0. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org

Schulze Westerath H, Brörkens N, Laister S, MacKintosh N, Winckler C and Knierim U 2009. Reliability of measures of socio-positive and play behaviour in dairy and beef cattle. In Welfare Quality® Reports No. 11 (ed. B, Forkman and LJ, Keeling), pp. 175–188. Welfare Quality® Consortium, Lelystad, the Netherlands.

Sørensen JT and Fraser D 2010. On-farm welfare assessment for regulatory purposes: issues and possible solutions. Livestock Science 131, 1–7.

Spoolder HAM, Burbidge JA, Lawrence AB, Simmins PH and Edwards SA 1996. Individual behavioural differences in pigs: intra-and inter-test consistency. Applied Animal Behaviour Science 49, 185–198.

Sundrum A and Rubelowski I 2001. The meaningfulness of design criteria in relation to the mortality of fattening bulls. Acta Agriculturae Scandinavica, Section A – Animal Science 51, 48–52.

Waiblinger S, Knierim U and Winckler C 2001. The development of an epidemiologically based on-farm Welfare Assessment System for use with dairy cows. Acta Agriculturae Scandinavica, Section A – Animal Science 51, 73–77.

Watson PF and Petrie A 2010. Method agreement analysis: a review of correct methodology. Theriogenology 73, 1167–1179.

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

Welfare Quality® 2011. Welfare Quality® assessment protocol for cattle without veal calves. Welfare Quality® Consortium, Lelystad, the Netherlands. http://www.welfarequalitynetwork.net

Winckler C, Brinkmann J and Glatz J 2007. Long-term consistency of selected animal-related welfare parameters in dairy farms. Animal Welfare 16, 197–199.

Windschnurer I, Boivin X and Waiblinger S 2009. Reliability of an avoidance distance test for the assessment of animals' responsiveness to humans and a preliminary investigation of its association with farmers' attitudes on bull fattening farms. Applied Animal Behaviour Science 117, 117–127.