© 2018 Universities Federation for Animal Welfare
The Old School, Brewhouse Hill, Wheathampstead,
Hertfordshire AL4 8AN, UK
www.ufaw.org.uk

Animal Welfare 2018, 27: 147-156
ISSN 0962-7286
doi: 10.7120/09627286.27.2.147

How reliable is the multi-criteria evaluation system of the Welfare Quality® protocol for growing pigs?

I Czycholl*[†], C Kniese[‡], L Schrader[‡] and J Krieter[†]

[†] Institute of Animal Breeding and Husbandry, Christian-Albrechts-University, Olshausenstrasse 40, D-24098 Kiel, Germany
[‡] Institute of Animal Welfare and Animal Husbandry, Friedrich Loeffler Institute, Doernbergstrasse 25/27, D-29223 Celle, Germany
* Contact for correspondence and requests for reprints: iczycholl@tierzucht.uni-kiel.de

Abstract

This paper focuses on the reliability of the multi-criteria evaluation model included in the Welfare Quality® protocol for growing pigs to aggregate the animal-based indicators, first to criteria, then to principle level and finally to an overall welfare score. This assessment was carried out in a practical application study on a sample of 24 farms in Germany. Altogether, 102 protocol assessments were carried out in repeated visits to these farms in order to evaluate the inter-observer and test-retest repeatability of the overall scores calculated by the multi-criteria evaluation system. Reliability is then assessed by the calculation of different reliability and agreement parameters: Spearman Rank Correlation Coefficients (RS), Intraclass Correlation Coefficients (ICC), Smallest Detectable Changes (SDC) and Limits of Agreement (LoA). Inter-observer repeatability was insufficient for the criteria comfort around resting, absence of injuries, expression of social behaviours, expression of other behaviours, good human-animal relationship and positive emotional state as well as for the principles good housing and appropriate behaviour. This is probably due in the main to insufficient repeatability of the underlying indicators that have been revealed in previous studies. Test-retest repeatability is predominantly insufficient. Overall, the present results highlight the importance of absolutely reliable indicators at the baseline level. Furthermore, it could be shown that the calculation procedure is partly incorrect and consequently needs correction. Therefore, this study is an important contribution to the future progression of the Welfare Quality® protocols and animal welfare assessment tools in general.

Keywords: *animal welfare, multi-criteria evaluation, pigs, reliability, repeatability, Welfare Quality®*

Introduction

Animal welfare has become an important topic within public and political discussion over recent decades (Hobbs *et al* 2002). There is, therefore, a growing need for stakeholders to comply with and increase animal welfare standards (Vapnek & Chapman 2010). In order to meet consumers' concerns in the form of, eg animal welfare certification schemes, it is necessary to carry out an objective evaluation of the welfare status on-farm (Blokhuis *et al* 2013).

Animal welfare comprises different aspects, such as the absence of thirst, hunger, discomfort, disease, pain, injuries and stress as well as the possibility to express normal behaviour (FAWC 1993). Thus, it is a multi-dimensional concept. A welfare assessment system has to take into account all different aspects in order to gain general acceptance. Different approaches have been proposed in recent years (Czycholl *et al* 2015). As such, the Welfare Quality® protocols are very promising as a systematic welfare evaluation tool since all the different dimensions of animal welfare are addressed and focus is given to animal-based indicators. According to Blokhuis *et al* (2013), animal-based parameters are the only parameters which assess the true value with regard to animal welfare.

In detail, in the Welfare Quality® protocols, the implementation of the multi-dimensionality of the concept of animal welfare takes place in the form of four principles, which are good feeding, good housing, good health and appropriate behaviour. These principles are divided into twelve criteria. These are then measured, on-farm, by a set of approximately 30 predominantly animal-based indicators. After the on-farm assessment, the measures are usually expressed as percentages of affected animals. These percentages are standardised into a dimensionless number between 0 and 100 by a multi-criteria evaluation system. Depending on the scores reached on principle level, as an overall assessment the farms are labelled as excellent, enhanced, acceptable or not classified (Welfare Quality® 2009). This multi-criteria evaluation system again places particular emphasis on the multi-dimensionality of the concept of animal welfare. Moreover, the aggregation of the on-farm measures for the total evaluation of the farm is especially required for labelling purposes (Botreau *et al* 2007).

https://doi.org/10.7120/09627286.27.2.147 Published online by Cambridge University Press

For a welfare assessment tool, reliability is one of the basic requirements (Velarde 2007). While the included animal-based indicators have already been thoroughly examined as regards their reliability, eg Czycholl *et al* (2016a,b) and Temple *et al* (2013), hitherto, there has been no study concerned with the reliability of the multi-criteria evaluation system. This, however, is of special importance as, for instance, Courboulay *et al* (2009) hypothesised that although there might be a degree of observer effect concerning some indicators assessed on-farm, this would not lessen reliability in general, ie it would not cause observers to rank farms differently based on the overall assessment. However, to the authors' knowledge, it has never been studied in detail as to whether the ranking of farms by the overall assessment is indeed reliable.

Concentrating on growing pigs, the present study aims therefore at evaluating the reliability, ie the inter-observer and the test-retest repeatability of the multi-criteria evaluation system of the Welfare Quality® protocol for growing pigs. Thus, in detail, the reliability of the calculated criteria and principle scores as well as the overall assessment were studied. Whether the ranking of farms does indeed stay the same, ie whether the hypothesis of Courboulay *et al* (2009) holds in the reliability assessment of the overall score was also evaluated.

Materials and methods

Data collection

Between January and August 2013, the data were recorded on 24 growing pig farms in Northern Germany. The pigs on the farms were housed either conventionally or according to the guidelines of the animal welfare label 'Tierschutzlabel' of the German animal welfare organisation 'Deutscher Tierschutzbund eV.' (Deutscher Tierschutzbund 2013). The size of the farms ranged from 250 to 1,500 pigs per farm. Pigs were fed *ad libitum* and kept indoors on fully or partially slatted floors except for two farms where outdoor access to a fully slatted area was provided. The number of pigs per pen ranged from 9 up to 100, the average space allowance at the finishing stage was 1.05 m² per pig, ranging from 0.8 to 1.35 m² per pig. The animals were crossbred, including German Landrace, Large White, Danish Landrace, Danish Yorkshire, Duroc and Pietrain, the exact lineage varied from farm-to-farm.

Three trained observers (A, B and C) carried out the complete Welfare Quality® protocols repeatedly on these farms. The training of observers taking part in this study was carried out by members of the Welfare Quality® project group in November 2012 for a group of 12 persons in total. Training sessions on each single measure of the protocol were continued until at least 80% of the participants had reached a consensus, ie matched their evaluations of an animal. Outliers among the participants were re-trained individually until the measure had been assessed correctly.

For the evaluation of inter-observer repeatability, a total of 30 protocol assessments were carried out on these farms: observers A and B fulfilled 20 combined assessments whereas observers A and C fulfilled ten assessments. Thus, the observers assessed the same animals at the same time, but completely independently of each other.

For the evaluation of test-retest repeatability, the 24 farms were evaluated repeatedly, meaning eight farms were visited repeatedly by observer A and 16 repeatedly by observer B. Each of the farms was visited six times during two consecutive fattening periods (batches). During each fattening period, three assessments took place: the first protocol assessment at the beginning of the fattening period at an average pig weight of 40 kg (farm visit 1), the second in the middle of the fattening period at an average weight of 75 kg (farm visit 2) and the third assessment at the end of the fattening period at an average weight of 100 kg (farm visit 3). Thus, the average time between farm visit 1 and 2 was 45 days and the average time between farm visit 2 and 3 was 40 days. This difference was due to practical conditions in the study as a gap of 48 h had to be left between visiting different farms to minimise the risk of disease transmission. The farmers were advised and agreed not to make any major changes in management during data collection, such as new mixing, new feed composition, new piglet suppliers or new treatments. Thus, common farm practices were maintained. Remixing did not occur on any of the farms during the fattening period. Immunisation and de-worming strategies remained unchanged. Feed management was practiced as two- or three-phase fattening on the different farms but not changed in between the two growing periods.

Protocol assessments

The entire Welfare Quality® protocol for growing pigs (Welfare Quality® 2009) was carried out during each farm visit. A detailed list of all assessed indicators can be found in Table 1. In the sty, this divides into a Qualitative Behaviour Assessment (QBA), behavioural observations (BO), a human animal relationship test (HAR) and the assessment of a variety of individual animal-based indicators (II). Protocol assessments started with a short farmer interview to collect information about management practices, such as castration and tail-docking procedures. Data on the prevalence of pneumonia, pleurisy, ascites and pericarditis were collected from the previous 12 months' slaughterhouse records of routine inspections, in accordance with Directive EC 854/2004. The observer then randomly selected the observation points for the QBA and the BO and ten pens for the HAR and the II. Following the Welfare Quality® protocol, four to six observation points were chosen for the QBA, three others for the BO and for the II, 150 individual pigs were assessed independent of the exact size of the farm. If more than 15 pigs were present in each of the ten randomly selected pens, 15 pigs in each of the pens were selected randomly. If pigs of different ages were present on a farm, all age categories were included. Hospital pens were excluded.

The QBA was carried out at four to six randomly chosen observation points for a total surveillance time of 20 min.

https://doi.org/10.7120/09627286.27.2.147 Published online by Cambridge University Press

Table 1 Principles, criteria and indicators of the Welfare Quality® protocol for growing pigs.

Principle	Criteria	Indicators
Good feeding	1 Absence of prolonged hunger	• Body condition score
	2 Absence of prolonged thirst	• Number of drinking places
		• Functioning of drinkers
		• Cleanliness of drinkers
Good housing	3 Comfort around resting	• Bursitis
		• Manure on the body
	4 Thermal comfort	• Huddling
		• Shivering
		• Panting
	5 Ease of movement	• Space allowance
Good health	6 Absence of injuries	• Lameness
		• Wounds on the body
		• Tail-biting
	7 Absence of disease	• Mortality
		• Coughing
		• Sneezing
		• Pumping
		• Twisted snouts
		• Rectal prolapse
		• Scouring
		• Skin condition
		• Hernias
		• Pneumonia
		• Pleurisy
		• Pericarditis
		• White spots
	8 Absence of pain induced by management procedures	• Castration
		• Tail-docking
Appropriate behaviour	9 Expression of social behaviour	• Social behaviour
	10 Expression of other behaviour	• Exploratory behaviour
	11 Good human-animal relationship	• Fear of humans
	12 Positive emotional state	• Qualitative Behaviour Assessment

After this time, the group of animals under surveillance was rated on a 125-mm visual analogue scale with the following 20 adjectives: active, relaxed, fearful, agitated, calm, content, tense, enjoying, frustrated, bored, playful, positively occupied, listless, lively, indifferent, irritable, aimless, happy, distressed and sociable. A mark was set on the scale to record whether the observer found that term to be rather absent (0 mm) or dominant (125 mm) for the animals under study. The length (mm) on the visual analogue scale was measured with a ruler for each of the adjectives. Thus, one score in millimetres for each adjective was obtained at farm level.

https://doi.org/10.7120/09627286.27.2.147 Published online by Cambridge University Press

Following QBA, BO were performed by instantaneous scan sampling at three other randomly chosen observation points. For the BO, all pigs in the pens had to stand up. If necessary, hands were clapped before starting the observation 5 min later. During this time, coughing and sneezing were counted. Afterwards, 40–60 pigs were scan sampled in five scans every 2 min at each observation point. For each scan, each of the animals was sorted into one of the following behavioural categories: positive social behaviour, negative social behaviour, pen investigation, use of enrichment material, other active behaviour or resting; the exact ethogram can be found in the Welfare Quality® protocol for growing pigs (Welfare Quality® 2009). The results of the BO were expressed as performed behaviour in percent of the total active behaviour. Thus, positive and negative social behaviour were expressed together as total social behaviour and negative social behaviour was also presented individually.

Prior to entering the ten previously randomly selected pens for the HAR and the evaluation of II, shivering, panting, and huddling were scored from outside these pens. Huddling was assessed only in resting animals.

Afterwards, the pens were entered. First, the observer walked around it in one direction and then waited in the middle of the pen for 30 s. While walking around the pen in the other direction, it was assessed whether more than 60% of the animals in the pen showed a panic response or not. While performing the HAR, it was also assessed whether scouring was present in the pen. The percentage of pens with a panic response from the total observed pens per farm was taken into account for further investigation.

Inside these ten entered pens, the pigs were scored individually for body condition, bursitis, manure on the body, wounds, tail-biting, lameness, laboured breathing, twisted snouts, rectal prolapse, skin condition and hernias. This meant only one randomly chosen side of the pig was assessed for wounds and manure on the body, skin condition and bursitis. The II were scored on a three- (0 = absent, 1 = light affection, 2 = strong affection) or a two-point scale (0 = absent, 2 = present). Furthermore, some resource-based parameters were recorded, such as the number, functioning and cleanliness of the drinkers. The pen size was measured, and the average weight of the animals was estimated. The IIs were analysed as the percentage of animals sorted into the corresponding category (eg bursitis category 0: 50%, bursitis category 1: 40%, bursitis category 2: 10%).

The collected data were aggregated into criteria and principle scores using the algorithm of the Welfare Quality® protocol (Welfare Quality® Network 2009). First, a dimensionless score ranging from 0–100 was calculated for the twelve criteria and then for the four main principles originating from the data collected on-farm, ie the indicator level. Table 1 shows a detailed overview of the allocation of indicators, criteria and principles according to Welfare Quality®. As a result, the individual criteria within a particular principle do not compensate for each other, thus a high score in one criterion will not compensate for a low score in another. Zero represents the worst and 100 the best possible welfare state. Depending on the scores of the four principles, the farms were rated overall as excellent, enhanced, acceptable or not classified. An overview of the corresponding slotting criteria can be found in Table 2.

Table 2 Slotting criteria for the overall assessment of farms based on the outcomes at principle level.

	Minimum values of all principles	Further required conditions
Excellent	> 55	> 80 in two of the principles
Enhanced	> 20	> 55 in two of the principles
Acceptable	> 10	> 20 in three of the principles*

* If these minimum requirements are not met, the farm is scored as not classified.

Ethical statement

The authors declare that the experiments were carried out in strict adherence to international animal welfare guidelines. The animals were kept and handled according to the 'German Animal Welfare Act' (German designation: TierSchG), the 'German Order for the Protection of Animals used for Experimental Purposes and other Scientific Purposes' (German designation: TierSchVersV) and the 'German Order for the Protection of Production Animals used for Farming Purposes and other Animals kept for the Production of Animal Products' (German designation: TierSchNutztV). No pain, suffering or injury was inflicted on the animals during the experiments.

Statistical analysis

After calculating the criteria and principle scores, these obtained scores were compared between the observers and repeated farm visits, respectively. The 20 combined protocol assessments of observers A and B as well as the ten combined protocol assessments of observers A and C were compared for the evaluation of inter-observer repeatability. Test-retest repeatability was estimated for the eight repeatedly visited farms of observer A and the 16 of observer B. As a result, farm visits at the same weight, ie age classes, of the two consecutive fattening periods (batches) were compared: thus, farm visit 1 of the first fattening period to farm visit 1 of the following second fattening period, farm visit 2 of the first fattening period to farm visit 2 of the following second fattening period and farm visit 3 of the first fattening period to farm visit 3 of the following second fattening period. This was done as the age of the animals had been an influencing factor on the protocol outcomes in previous studies (Temple *et al* 2012, 2013) and preliminary studies of the dataset of this study also revealed an age effect.

Different reliability and agreement parameters Spearman Rank Correlation Coefficient (RS), Intraclass Correlation Coefficient (ICC), Smallest Detectable Change (SDC), Limits of Agreement (LoA) were calculated for statistical analysis using the statistic programme R (Version 2.11.1)

https://doi.org/10.7120/09627286.27.2.147 Published online by Cambridge University Press

(Venables & Smith 2010). The IRR package (Gamer *et al* 2012) for R (Version 2.11.1) was used for calculation of the ICC, SDC and the LoA. This combination of reliability and agreement parameters is advised by de Vet *et al* (2006) and was also used in the studies of Czycholl *et al* (2016a,b) and Temple *et al* (2013). To determine criteria or principles as reliable, acceptable reliability thresholds should be met in all four parameters.

Spearman Rank Correlation Coefficient (RS)

The RS evaluates the degree of linear correlation between two variables (Gauthier 2001) by comparing the rank order instead of the directly obtained values (Dohoo *et al* 2003). The values of the RS ranges between –1 to 1, whereas correlation is better the closer the value is to 1. Negative values indicate negative correlations. Following the suggestion of Martin and Bateson (2007), an RS equal to or greater than 0.4 is interpreted as acceptable correlation and equal to or greater than 0.7 as good correlation.

Intraclass Correlation Coefficient (ICC)

The ICC is based on an analysis of variance. It then assesses the reliability by putting into proportion the variance of the same subject, ie observers and farm visits, respectively, to the total variance of all measurements and subjects (Bartko 1966).

For the determination of inter-observer repeatability, the ICC was calculated based on the following two-way model according to Shrout and Fleiss (Shrout & Fleiss 1979):

$x_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$,

with x_{ijk} being the measured value, μ the general average value, α_i the fixed effect of the difference between the measurement objects (farms), β_j the random effect of the observers, $(\alpha\beta)_{ij}$ the interaction effect between observers and objects and ε_{ijk} as the general error term.

ICC was then calculated according to the formula of agreement (Shrout & Fleiss 1979):

$$\mathrm{ICC} = \frac{\sigma^2_{\text{(farms)}}}{\sigma^2_{\text{(farms)}} + \sigma^2_{\text{(observers)}} + \sigma^2_{\text{(residual)}}}$$

with σ^2 describing the variance of the study objects, the observers or the residual variance, respectively.

For the evaluation of test-retest repeatability, the fundamental analysis of variance was carried out by the following one-way model:

$x_{ijk} = \mu + \alpha_i + \varepsilon_{ijk}$,

with x_{ijk} being the measured value, μ the general average value, α_i the random effect of the difference among the 24 farms and ε_{ijk} as the general error term.

The according formula for the calculation of ICC consequently was:

$$\mathrm{ICC} = \frac{\sigma^2_{\text{(farms)}}}{\sigma^2_{\text{(farms)}} + \sigma^2_{\text{(residual)}}}$$

ICC can range between 0 and 1. As proposed by McGraw and Wong (1996), values equal to or greater than 0.4 were interpreted as acceptable and equal to or greater than 0.7 as good reliability.

Smallest Detectable Change (SDC)

SDC is an expression of the measurement error $\sigma^2_{\text{(error)}}$. SDC was calculated according to de Vet *et al* (2006) by the formula:

$\mathrm{SDC} = 1.96 \times \sqrt{2}(\sigma^2_{\text{[error]}})$

It indicates the smallest change in the score that can be detected with the measurement instrument above the measurement error (Donoghue & Stokes 2009). The measurement unit of the SDC is in accordance with the measurement unit of the parameters under surveillance. Thus, in the present case, it is to be understood as a dimensionless number ranging from 0–100. Based on the interpretation of the simple agreement coefficient in de Vet *et al* (2006), an SDC lesser than or equal to 10, which corresponds to 10% variation, was interpreted as acceptable agreement.

Limits of Agreement (LoA)

LoA was also calculated according to de Vet *et al* (2006) by the formula:

$\mathrm{LoA} = \text{mean} \pm 1.96\ (\sqrt{2} \times \sigma^2_{\text{[error]}})$

The LoA, which was first introduced by Bland and Altman (1986), calculates the range of the difference between two sets of measurement values and, in this study, is expressed as the relative frequency between –100 and 100. The direction of –100 indicates differences according to higher values obtained by observer B/C or fattening period 1, respectively, and the direction of 100 due to higher values achieved by observer A or fattening period 2, respectively. Interpretation of the LoA was also based on the simple agreement coefficient of de Vet *et al* (2006) and, thus, an interval lesser than or equal to –10 to 10, which corresponds to a variance of 10%, was interpreted as acceptable agreement.

Results

Overall assessment

Of all protocol assessments, 0.5% were scored as excellent in the overall classification, 89.9% were labelled as enhanced and 9.6% as acceptable. No protocol assessment led to a classification of a farm as not classified.

Inter-observer repeatability

Mean values, as well as reliability and agreement parameters of criteria, principle and overall welfare scores of the inter-observer repeatability study are presented in Table 3 (see the Appendix in the supplementary material to papers published in *Animal Welfare*; https://www.ufaw.org.uk/the-ufaw-journal/supplementary-material). The criteria absence of prolonged hunger, absence of prolonged thirst, thermal comfort, ease of movement, absence of disease and absence of pain induced by management procedures were of good inter-observer repeatability in both comparisons, according to our definition. The other criteria did not achieve suffi-

https://doi.org/10.7120/09627286.27.2.147 Published online by Cambridge University Press

cient inter-observer repeatability, as not all four calculated parameters were of acceptable values. The principle good feeding, consisting of the criteria absence of prolonged hunger and absence of prolonged thirst, was of good repeatability. The same can be said for the principle good health, which is made up of the criteria absence of injuries, absence of disease and absence of pain induced by management procedures. The principles good housing and appropriate behaviour were of insufficient repeatability. The overall assessment was of good repeatability in the comparison of observers A and C and of acceptable repeatability in the comparison of observers A and B.

Test-retest repeatability

Mean values, as well as reliability and agreement parameters of criteria, principle and overall welfare scores of the test-retest repeatability study are presented in Tables 4 and 5 (see the Appendix in the supplementary material to papers published in *Animal Welfare*; https://www.ufaw.org.uk/the-ufaw-journal/supplementary-material). Test-retest repeatability of the three farm visits in two consecutive fattening periods of the eight farms of observer A is presented in Table 4 (https://www.ufaw.org.uk/the-ufaw-journal/supplementary-material) and the test-retest repeatability of the repeated farm visits of observer B is presented in Table 5 (https://www.ufaw.org.uk/the-ufaw-journal/supplementary-material). Test-retest repeatability, as defined in the present study, in all comparisons of both observers was only present in the criterion absence of pain induced by management procedures. Furthermore, neither any of the principles nor the overall assessment demonstrated repeatability congruently in all three comparisons.

However, there were also some cases in which interpretation was not straightforward, which are discussed in detail later on. For example, for the criterion absence of hunger, the comparison of farm visits 2 and 3 of observer A as well as the comparison of farm visit 1 of observer B were of good repeatability. For the criterion absence of prolonged thirst, only the comparison of farm visit 3 of observer B led to the assumption of good repeatability. Repeatability was further present for the criterion thermal comfort in the comparison of farm visits 2 and 3 of observer A as well as in the comparison of farm visit 3 of observer B. The criterion thermal comfort was of good repeatability in farm visits 1 of both observers. Furthermore, repeatability was detected in the comparison of farm visit 3 of observer A for the criterion expression of other behaviours. While only one of the principles in one farm visit (good feeding, farm visit 3) was of good test-retest repeatability in the repeated farm visits of observer B, the principle good housing was of good repeatability in the assessment of observer A in the comparison of farm visit 1 and the principle good health in all of the comparisons. The overall assessment was of sufficient test-retest repeatability for the comparison of farm visit 1 of observer A and farm visits 2 and 3 of observer B.

Discussion

Assessment of reliability

Reliability implies that different people need to agree in their assessments (inter-observer repeatability) and that a certain consistency of the results over time is needed (test-retest repeatability) (de Passillé & Rushen 2005). In the case of a certification tool for labelling purposes, it is obvious that the results may not be dependent on the person carrying out the certification. Furthermore, certification audits will be carried out at longer time intervals. Therefore, the present study assessed the inter-observer as well as the test-retest repeatability of the Welfare Quality® protocol for growing pigs focusing on the overall assessment.

Different reliability and agreement parameters were calculated for the assessment of reliability: the Spearman Rank Correlation Coefficient (RS), the Intraclass Correlation Coefficient (ICC), the Smallest Detectable Change (SDC) as well as the Limits of Agreement (LoA). This procedure is recommended by de Vet *et al* (2006) and was also mainly followed in the studies of Czycholl *et al* (2016a,b) and Temple *et al* (2013). As each parameter has its own weaknesses and benefits, there is not one single parameter capable of satisfactorily assessing reliability (Dohoo *et al* 2003). For this reason, it is advised to calculate a range of different reliability and agreement parameters and interpret the reliability of the measured objects based on all statistical coefficients (Dohoo *et al* 2003; de Vet *et al* 2006; Temple *et al* 2013). In the interpretation, one must take into account certain specific benefits and disadvantages of the statistical parameters: on the one hand, the limits for acceptability for the agreement parameters Smallest Detectable Change (SDC) and Limits of Agreement (LoA) remain to a certain extent subjective (de Vet *et al* 2006). On the other, the reliability parameters, ie Spearman Rank Correlation Coefficient (RS) and Intraclass Correlation Coefficient (ICC) are strongly dependent on the variance amongst study objects (Dohoo *et al* 2003), which was rather small in the present study. Therefore, this should be considered in the interpretation of the reliability parameters.

Inter-observer repeatability

Inter-observer repeatability was especially favourable for those criteria made up of indicators rarely observed. The observers agreed in the absence of these indicators. For example, no lean animals were observed in the inter-observer repeatability study. This should be somewhat normal for fattening pig units. Similarly, no farms in the study had problems regarding the thermoregulation of the pigs and many of the diseases that were scored were not observed in the study at all. Furthermore, observers showed a high level of agreement in those criteria including management-based parameters, eg in the evaluation of the parameters tail-docking and castration, which are the indicators from which the criterion absence of pain induced by management procedures is calculated. Observers did not agree in regards to the evaluation of comfort around resting.

https://doi.org/10.7120/09627286.27.2.147 Published online by Cambridge University Press

This is perhaps unsurprising given previous studies have revealed that the parameter bursitis, which is included in the calculation of this criterion, cannot be reliably assessed (Temple *et al* 2013; Czycholl *et al* 2016a,b). This is probably due to the fact that the visual evaluation becomes complicated if the legs are dirty or the pigs move around quickly (Czycholl *et al* 2016a). The reason for the insufficient inter-observer repeatability of the criterion positive emotional state is probably that the underlying indicator Qualitative Behaviour Assessment (QBA) is of insufficient reliability, which has been revealed in several previous studies (Bokkers *et al* 2012; Tuyttens *et al* 2014; Czycholl *et al* 2016a,b). Basically, the same can be said for the low repeatability of the criterion human animal relationship consisting of the indicator human animal relationship test, which was of insufficient repeatability in other studies (Czycholl *et al* 2016a). This was most probably due to the study design as observers entered the pens one after the other to avoid mutual interference later on in the assessment. However, this probably affected the reaction of the animals towards the second intruder. Therefore, repeatability of this indicator should be re-checked with adapted study designs in future studies. The criterion absence of injuries is also of insufficient repeatability, according to the definition that all four statistical parameters need to be within the prescribed limits. However, the underlying indicators lameness, tail-biting and wounds on the body were proven to be of acceptable to good reliability (Czycholl *et al* 2016a). But this insufficiency regarding the criterion absence of injuries is only due to the statistical parameter Limits of Agreement (LoA). The LoA demonstrates insufficient agreement in both observer comparisons. However, the limits for acceptability are exceeded only narrowly. It should be borne in mind that these limits constitute a degree of arbitrary determination (de Vet *et al* 2006). Thus, this minor discrepancy should not be overestimated. However, as previously stated, in earlier studies the underlying indicators have been proven to be of good reliability. Therefore, this inconsistency must be caused by the calculation procedure. The behavioural observations which provide the basis for the criteria expression of social behaviours and other behaviours were actually of sufficient reliability in the previous study of Czycholl *et al* (2016a). Therefore, as the reason cannot be insufficient reliability in the underlying indicators, in this case the calculation procedure must be questioned and checked. Especially, as pointed out by Botreau *et al* (2013), the weights of the I-Spline functions assigned by experts should be re-checked as well as some of the Shapley values of the Choquet integrals.

On the level of principles, inter-observer repeatability was good for the principles good feeding and good health, but not good housing and appropriate behaviour. In the case of good feeding, this may be expected as the two criteria making up this principle were of good inter-observer repeatability. For the principle good health, however, the criterion absence of injuries which was of insufficient repeatability also has an influence. This obviously did not impede the reliable calculation of the principle score. This can probably be explained by the fact that repeatability of the criterion absence of injuries was only marginally unreliable as only the parameter LoA suggested insufficient reliability, as described above. The low repeatability of the principles good housing and appropriate behaviour is not surprising, as they consist of criteria that were proven to be insufficiently reliable and these again consisted of indicators that were not reliable (ie bursitis and QBA). Regarding these criteria and principles in the present study, this means that the hypothesis of Courboulay *et al* (2009) which states that although there are observer effects concerning some parameters assessed on-farm, these would not lessen reliability since they would not cause observers to rank farms differently, needs to be reconsidered. As can be seen from the RS, the rank order of farms did not stay the same.

Test-retest repeatability

In the present study, two consecutive, but different, fattening periods were compared. The first one at a weight of 40 kg, the second at a weight of 75 kg and the third at a weight of 100 kg. For the comparison of test-retest reliability, we always compared the first, second and third farm visit, respectively, of the two fattening periods. Thus, different animals were compared. However, in terms of welfare assessment, as Knierim and Winckler (2009) stated, it is assumed that regarding feasibility, results on a farm should stay the same in between a period of six months. Thus, one can argue that for test-retest reliability, despite the fact that different animals are present in the two growing periods, the welfare status should not change much if no major changes in management occurred. Moreover, analysing exactly whether results in between these stated six months actually are consistent assumes greater importance. As previous studies revealed age effects in the basic indicators (Temple *et al* 2012, 2013), it was decided to compare same age classes in our study.

In the assessment of test-retest repeatability, the only criterion presenting good repeatability for both observers in all comparisons was absence of pain induced by management procedures. This is not surprising as this criterion is made up of the two management-based criteria tail-docking and castration. As major changes in the management were excluded during the studies, results had to stay the same, otherwise this prerequisite of the study design would not have been met. Although lean animals were rarely observed on the fattening pig farms, the test-retest repeatability of this criterion was only good in half of the comparisons. This is probably due to the rather low variability observed on the farms under study, which led to an overvaluation of outliers. Despite this fact, test-retest repeatability proved to be inacceptable for criteria and principle scores in the present study. This is especially of interest as the chosen time interval for comparison was relatively small (two consecutive fattening periods). According to Knierim and Winckler (2009), assessments probably need to be carried out in time intervals of more than six months in order to be feasible. The present results show that given the current multi-criteria evaluation system this will not be reliably possible. As demonstrated by the results here, it is essential to carefully revise the indicators in question which were detected as insufficiently reliable in previous studies, eg bursitis (Temple *et al* 2013; Czycholl *et al* 2016a,b). However, these studies demonstrated good reliability for most of the indicators, therefore it comes as a surprise to see the criteria and principle scores had even lower reliability than on the indicator scale. This can be most probably explained by incongruities in the calculation procedure. For example, Czycholl *et al* (2017) revealed influences in the calculation procedure for which a control was necessary. Specifically, in the calculation procedure, the aim was to avoid the double counting of and compensation between indicators and criteria, respectively (Botreau *et al* 2013). Czycholl *et al* (2017), however, showed that double counting and compensation occurs in the calculation procedure. These might also be responsible, then, for the subsequent deterioration in repeatability since indicators and criteria, respectively, of lower repeatability might also have had an influence, thereby lowering the general reliability.

To summarise, the present results support the conclusion of de Vries *et al* (2013) that the multi-criteria evaluation system of Welfare Quality® needs revision. Most of all, one should concentrate on using absolutely reliable indicators. Thus, those indicators proven to be of insufficient reliability in previous studies (Bokkers *et al* 2012; Temple *et al* 2013; Tuyttens *et al* 2014; Czycholl *et al* 2016a,b) should be revised or replaced accordingly. Moreover, as already suggested by Czycholl *et al* (2017), alternative aggregation systems providing more flexibility, as proposed by Martin *et al* (2017a,b) should be considered and validated further in order to revise the current aggregation system of the Welfare Quality® protocol for growing pigs.

Moreover, it should be noticed that a maximum score of 97 was reached in the criterion absence of pain induced by management procedures, even if no castration and no tail-docking had been carried out at all which, per the definition of Welfare Quality®, is the best possible option in terms of welfare. This needs correction, because the best possible option should reach the maximum score of 100.

Overall assessment

The majority of protocol assessments (89.9%) led to a classification as enhanced. Most of the other assessments (9.6%) came to the result acceptable. The narrow-scale utilisation in this study shows that there were only small variations in the protocol assessments. It is not surprising that no protocol assessment led to labelling as not classified, because all farms in the study met legal requirements, which was defined as baseline for the score acceptable (Botreau *et al* 2009). Nevertheless, farms using differing housing systems were analysed so a larger variety had been previously expected. The question is whether there was a true, low variation between farms or whether the Welfare Quality® protocol lacked the sensitivity to detect small variations between the farms. Further studies with a wider sample of farms, ideally from differing nations, are needed to clarify this aspect. The Welfare Quality® protocols have been criticised in the past as several studies have revealed that the European public expects higher standards of welfare (Evans & Miele 2007; Miele *et al* 2011), ie that the classification as enhanced of almost all farms might be too good. Moreover, in preliminary application studies regarding the Welfare Quality® project, there was no success in scoring a single farm as excellent (Botreau *et al* 2013). Thus, the gradient should definitely become stricter for the aggregation from principle level to the overall assessment. However, as discussed earlier, the current aggregation system requires a thorough revision. It is probably more advisable to also use Choquet integrals for this last aggregation step, which might have the potential to minimise compensation between the principles and thus lead to a more accurate classification of farms. However, this needs to be verified in further studies that thoroughly revise the current aggregation system.

Animal welfare implications

While an aggregation is essential in terms of welfare assessment, this study demonstrates that the current aggregation system of the Welfare Quality® protocol for growing pigs is not reliable. Suggestions for future improvements are made, contributing to a future reliable and broadly implemented objective welfare assessment system. A reliable, objective welfare assessment tool will also help to further improve the welfare status of the animals.

Conclusion

The aim of the present study was to assess the reliability of the multi-criteria evaluation system in the overall assessment of the Welfare Quality® protocol for growing pigs in its practical application. Several limitations and challenges were detected regarding the test-retest repeatability of criteria and principle scores, in particular. Sufficient reliability, ie inter-observer and test-retest repeatability could only be detected for the criterion absence of pain induced by management procedures. The present results highlight the importance of absolutely reliable indicators at the baseline level as it could be refuted that, despite some minor irregularities, the ranking of farms would remain the same and therefore reliable. Furthermore, it could be shown that the calculation procedure is partly incorrect and consequently needs correction. In particular, this study demonstrates that the simple use of the aggregation system to answer welfare questions should, at present, be interpreted with caution. However, this study remains an important contribution to the future progression of the Welfare Quality® protocols and animal welfare assessment tools in general.

Acknowledgements

This work was financially supported by the German Federal Ministry of Food, Agriculture and Consumer Protection (BMELV) through the Federal Agency for Agriculture and Nutrition (BLE), grant number 2817200913.

References

Bartko JJ 1966 The intraclass correlation coefficient as a measure of reliability. *Psychological Reports 19*: 3-11. https://doi.org/10.2466/pr0.1966.19.1.3

Bland MJ and Altman DG 1986 Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet 327*: 307-310. https://doi.org/10.1016/S0140-6736(86)90837-8

Blokhuis H, Veissier I, Jones B and Miele M 2013 Improving farm animal welfare - science and society working together: the Welfare Quality® approach. In: Blokhuis H, Miele M, Veissier I and Jones B (eds) *Improving Farm Animal Welfare - Science and Society Working Together: The Welfare Quality Approach*. Wageningen Academic Publishers: Wageningen, Gelderland, The Netherlands. https://doi.org/10.3920/978-90-8686-770-7

Bokkers EAM, de Vries M, Antonissen I and de Boer IJM 2012 Inter- and intra-observer reliability of experienced and inexperienced observers for the Qualitative Behaviour Assessment in dairy cattle. *Animal Welfare 21*: 307-318. https://doi.org/10.7120/09627286.21.3.307

Botreau R, Bonde M, Butterworth A, Perny P, Bracke MBM, Capdeville J and Veissier I 2007 Aggregation of measures to produce an overall assessment of animal welfare. Part 1: a review of existing methods. *Animal 1*: 1179-1187. https://doi.org/10.1017/S1751731107000535

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, Winckler C, Velarde A, Butterworth A, Dalmau A, Keeling LJ and Veissier I 2013 Integration of data collected on farms or at slaughter to generate an overall assessment of animal welfare. In: Blokhuis H, Miele M, Veissier I and Jones B (eds) *Improving Farm Animal Welfare-Science and Society Working Together: The Welfare Quality® Approach*. Wageningen Academic Publishers: Wageningen, Gelderland, The Netherlands. https://doi.org/10.3920/978-90-8686-770-7_7

Courboulay V, Meunier-Salaun MC, Edwards SA, Guy JH and Scott K 2009b Repeatability of abnormal behaviour. In: Forkman B and Keeling LJ (eds) *Welfare Quality® Reports* pp 131-141. Cardiff University: Cardiff, UK

Czycholl I, Büttner K, Grosse Beilage E and Krieter J 2015 Review of the assessment of animal welfare with special emphasis on the Welfare Quality® animal welfare assessment protocol for growing pigs. *Archive for Animal Breeding 58*: 237-249. https://doi.org/10.5194/aab-58-237-2015

Czycholl I, Kniese C, Büttner K, Grosse Beilage E, Schrader L and Krieter J 2016a Inter-observer reliability of the Welfare Quality® Animal Welfare Assessment Protocol for Growing Pigs. *Springer Plus 5*: 1-13. https://doi.org/10.1186/s40064-016-2785-1

Czycholl I, Kniese C, Büttner K, Grosse Beilage E, Schrader L and Krieter J 2016b Test-retest reliability of the Welfare Quality® Animal Welfare Assessment Protocol for Growing Pigs. *Animal Welfare 25*: 447-459. https://doi.org/10.7120/09627286.25.4.447

Czycholl I, Kniese C, Schrader L and Krieter J 2017 Assessment of the multi-criteria evaluation system of the Welfare Quality® protocol for growing pigs. *Animal 9*: 1-8. https://doi.org/10.1017/S1751731117000210

de Passillé AM and Rushen J 2005 Can we measure human animal interactions in on-farm animal welfare assessment? Some unresolved issues. *Applied Animal Behaviour Science 92*: 193-209. https://doi.org/10.1016/j.applanim.2005.05.006

Deutscher Tierschutzbund 2013 *Kriterienkatalog für eine tiergerechte Haltung und Behandlung von Mastschweinen im Rahmen des Tierschutzlabels "Für mehr Tierschutz"*. Deutscher Tierschutzbund ev: Bonn, Germany. [Title translation: Criteria catalogue for an animal-friendly husbandry and handling of growing pigs within the German animal welfare label 'increasing animal welfare']

de Vet HCW, Terwee CB, Knol DL and Bouter LM 2006 When to use agreement versus reliability measures. *Journal of Clinical Epidemiology 59*: 1033-1039. https://doi.org/10.1016/j.jclinepi.2005.10.015

de Vries M, Bokkers EAM, van Schaik G, Botreau R, Engel B, Dijkstra T and de Boer IJM 2013 Evaluating results of the Welfare Quality® multi-criteria evaluation model for classification of dairy cattle welfare at the herd level. *Journal of Dairy Science 96*: 6264-6273. https://doi.org/10.3168/jds.2012-6129

Dohoo I, Martin W and Stryhn H 2003 Screening and diagnostic tests. *Veterinary Epidemiologic Research 1*: 85-120

Donoghue D and Stokes EK 2009 How much change is true change? The minimum detectable change of the Berg Balance Scale in elderly people. *Journal of Rehabilitation Medicine 41*: 343-346. https://doi.org/10.2340/16501977-0337

Evans A and Miele M 2007 *Consumers' Views about Farm Animal Welfare*. Cardiff University: Cardiff, UK

FAWC 1993 *Second report on priorities for research and development in farm animal welfare*. Defra: London, UK

Gamer M, Lemon J, Fellows I and Singh P 2012 *Irr: various coefficients of interrater reliability and agreement (R package version 0.83)*. http://CRAN.R-project.org/package=irr

Gauthier TD 2001 Detecting trends using Spearman's Rank Correlation Coefficient. *Environmental Forensics 2*: 359-362. https://doi.org/10.1006/enfo.2001.0061

Hobbs AL, Hobbs JE, Isaac GE and Kerr WA 2002 Ethics, domestic food policy and trade law: assessing the EU animal welfare proposal to the WTO. *Food Policy 27*: 437-454. https://doi.org/10.1016/S0306-9192(02)00048-9

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

Martin P and Bateson P 2007 *Measuring Behaviour: An Introductory Guide*. University of Cambridge: Cambridge, UK. https://doi.org/10.1017/CBO9780511810893

Martin P, Czycholl I, Buxadé C and Krieter J 2017 Validation of a multi-criteria evaluation model for animal welfare. *Animal 11*: 650-660. https://doi.org/10.1017/S1751731116001737

Martin P, Traulsen I, Buxadé C and Krieter J 2017 Development of a multi-criteria evaluation system to assess growing pig welfare. *Animal 11*: 466-477. https://doi.org/10.1017/S1751731116001464

https://doi.org/10.7120/09627286.27.2.147 Published online by Cambridge University Press

McGraw KO and Wong SP 1996 Forming inferences about some intraclass correlation coefficients. *Psychological Methods 1*: 30-46. https://doi.org/10.1037/1082-989X.1.1.30

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

Shrout PE and Fleiss JL 1979 Intra-class correlations: uses in assessing rater reliability. *Psychological Bulletin 86*: 420-428. https://doi.org/10.1037/0033-2909.86.2.420

Temple D, Courboulay V, Manteca X, Velarde A and Dalmau A 2012 The welfare of growing pigs in five different production systems: assessment of feeding and housing. *Animal 6*: 656-667. https://doi.org/10.1017/S1751731111001868

Temple D, Manteca X, Dalmau A and Velarde A 2013 Assessment of test-retest reliability of animal-based measures on growing pig farms. *Livestock Science 151*: 35-45. https://doi.org/10.1016/j.livsci.2012.10.012

Tuyttens FAM, de Graaf S, Heerkens JLT, Jacobs L, Nalon E, Ott S, Stadig L, Van Laer E and Ampe B 2014 Observer bias in animal behaviour research: can we believe what we score, if we score what we believe? *Animal Behaviour 90*: 273-280. https://doi.org/10.1016/j.anbehav.2014.02.007

Vapnek J and Chapman M 2010 Legislative and regulatory options for animal welfare. *FAO Legislative Study, No 104*. https://ssrn.com/abstract=2898362

Velarde AG 2007 *On farm Monitoring of Pig Welfare*. Wageningen Academic Publishers: Gelderland, The Netherlands. https://doi.org/10.3920/978-90-8686-591-8

Venables WN and Smith DM 2010 The R development core team (2004), an introduction to R. *The R Development Core Team 2*: 1-90

Welfare Quality® 2009 *Welfare Quality*® *Assessment Protocol for Pigs*. Wageningen Academic Publishers: Wageningen, Gelderland, The Netherlands

Welfare Quality® **Network** 2009 *Online Calculator*. http://www1.clermont.inra.fr/wq/index.php?id=simul&new=1&situation=FPF

https://doi.org/10.7120/09627286.27.2.147 Published online by Cambridge University Press