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

Animal Welfare 2020, 29: 323-337
ISSN 0962-7286
doi: 10.7120/09627286.29.3.323

Interobserver reliability of measures of the Welfare Quality® animal welfare assessment protocol for sows and piglets

L Friedrich*†, J Krieter†, N Kemper‡ and I Czycholl†

† Institute of Animal Breeding and Husbandry, Christian-Albrechts-University, Olshausenstr 40, D-24098 Kiel, Germany
‡ Institute for Animal Hygiene, Animal Welfare and Farm Animal Behaviour, University of Veterinary Medicine Hannover, Foundation, Bischofsholer Damm 15, D-30173 Hannover, Germany
* Contact for correspondence: lfriedrich@tierzucht.uni-kiel.de

Abstract

The aim of this study was to assess the interobserver reliability of the measures forming the Welfare Quality® animal welfare assessment protocol for sows and piglets. The study was carried out at nine farms in Northern Germany. Two trained observers evaluated identical animals simultaneously but independently in 40 joint farm visits. Interobserver reliability was calculated at individual animal level using Cohen's kappa, weighted kappa and the prevalence-adjusted, bias-adjusted kappa (PABAK) and at farm level using Spearman's rank correlation coefficient (RS), the intraclass correlation coefficient (ICC), smallest detectable change (SDC) and limits of agreement (LoA). While a direct comparison of the adjectives of the qualitative behaviour assessment showed poor interobserver reliability, a Principal Component Analysis detected good interobserver reliability. The assessment of social and exploratory behaviours showed acceptable interobserver reliability, while the assessment of stereotypies displayed good interobserver reliability. The human-animal relationship test showed only poor interobserver reliability at individual animal and farm levels. In most cases, measures of health and physical state assessed in sows and piglets exhibited acceptable or good interobserver reliability. In conclusion, after some measures are revised, particularly those examining the human-animal relationship, the Welfare Quality® protocol for sows and piglets will represent a reliable approach in terms of interobserver reliability to assess the welfare of sows and piglets.

Keywords: *animal-based, animal welfare, interobserver reliability, pig, sows, Welfare Quality®*

Introduction

Sustainable food production is becoming increasingly important to consumers and encompasses animal welfare, ethical concerns and concerns about any environmental impact (van Loo *et al* 2014). Regarding animal welfare, Thorslund *et al* (2017) described the importance of and responsibility for market-driven pig welfare in their review and asserted that animal welfare was not only an indicator of meat quality but also eating quality. Furthermore, according to the Eurobarometer (2016), a public opinion survey conducted by the European Commission in EU countries, most Europeans attached importance to the protection of animal welfare and agreed with the notion of paying more for welfare-friendly products. Taking this increasing demand into account, the EU-funded Welfare Quality® project aimed to generate a generally accepted and objective assessment system for animal welfare (Webster 2005). The main outcomes of this project were the science-based Welfare Quality® animal welfare assessment protocols for assessing species-specific animal welfare in cattle, poultry and pigs. In the Welfare Quality® protocols, animal welfare was defined using four main principles: Good feeding, Good housing, Good health and Appropriate behaviour. These four principles were then divided into twelve independent but complementary criteria using a top-down approach, for which approximately 30 mainly animal-based measures were chosen (Botreau *et al* 2007). As the connection between specific resource-based measures and the welfare status of the animal(s) in question is not always entirely understood, one of the intentions of the Welfare Quality® project was to apply animal-based measures wherever feasible (Blokhuis *et al* 2008). For example, within the Welfare Quality® protocol for sows and piglets, the criterion thermal comfort within the principle Good housing is evaluated by assessing the animal-based measures of panting and huddling instead of the ambient temperature. Using this approach, the effect of the environment on the animals is evaluated rather than merely the environment.

Feasibility, validity and reliability are basic requirements of an objective measurement method (Velarde & Geers 2007).

Feasibility assesses the protocol's practical application, ie reliable results are produced at an affordable cost (Velarde & Geers 2007). Validity describes the extent to which a measure actually identifies the characteristic it is designed to assess (Martin & Bateson 2007). Reliability is defined as the extent to which measures are repeatable and consistent, ie the similarity between repeated measurements of the same item. Reliability, on-farm, is generally classified into interobserver and test-retest reliability (de Passillé & Rushen 2005). Therefore, interobserver reliability indicates that different, trained observers should obtain the same results when assessing the same animals at the same time under the same circumstances but independently from each other (Martin & Bateson 2007). The test-retest reliability characterises the consistency of the method over time and, thus, the repeatability of the results (Martin & Bateson 2007; Windschnurer *et al* 2008).

All measures included in the Welfare Quality® protocols were chosen with regard to their feasibility, validity and reliability. However, for the process of developing the protocols, changes were necessary, eg measures with more than three categories, eg the five-point bursitis scale by Lyons *et al* (1995), were reduced to three categories to facilitate their utilisation (Veissier *et al* 2013). Furthermore, the existing preliminary studies performed while developing the protocols mainly focused on video sequences because an on-farm assessment is much more costly and time consuming (Veissier *et al* 2013). Therefore, on-farm studies of pigs to determine the interobserver reliability of the Welfare Quality® system have focused on certain measures (Forkman & Keeling 2009), such as the health measure, lameness (Geverink *et al* 2009). Hence, on-farm studies of the reliability of the entire protocol, ie addressing all measures, are rare, particularly studies assessing the Welfare Quality® protocol for sows and piglets. Additional interobserver reliability studies of growing pigs have been performed by Czycholl *et al* (2016a) and Dalmau *et al* (2010). Likewise, the protocol for growing pigs was tested for its test-retest reliability (Temple *et al* 2013; Czycholl *et al* 2016b). The studies noted shortcomings principally in the criteria of comfort around resting in the principle Good housing, eg the measure of health and physical state related to bursitis. However, the complete volume of the measures of the Welfare Quality® protocol for sows and piglets has not been tested for interobserver reliability. The protocol was used solely in the study by Scott *et al* (2009), who tested a prototype monitoring system to assess animal welfare in sows and piglets. Friedrich *et al* (2019a,b) assessed the test-retest reliability of the measures of this protocol. Consequently, here, the aim was to assess the interobserver reliability of Welfare Quality® measures in the assessment protocol for sows and piglets at piglet-producing farms.

Different statistical parameters have been used to evaluate interobserver reliability. As interpretation of only one parameter can readily lead to misinterpretation, the calculation of various statistical parameters is advised to help offset possible disadvantages, as suggested by de Vet *et al* (2006). For this study, kappa coefficients (Cohen's kappa coefficient [κ], weighted kappa [κ_w] and prevalence-adjusted, bias-adjusted kappa [PABAK]) were calculated. Furthermore, Spearman's rank correlation coefficient (RS) and intraclass correlation coefficient (ICC) were calculated as reliability parameters and smallest detectable change (SDC) and limits of agreement (LoA) were calculated as agreement parameters.

This study's evaluation of interobserver reliability of the measures of the Welfare Quality® protocol for sows and piglets is a first step towards providing an objective assessment tool for animal welfare in sows and piglets, since the practical application of the protocol as a tool to certify welfare is only possible if the measures of the protocol provide reliable information.

Materials and methods

Ethical statement

The authors declare that their study was conducted in strict accordance with international animal welfare guidelines. The study animals were housed either conventionally or according to the EU organic scheme (Council Regulation [EC] No 834/2007 [EC 2007]). Animals were housed according to EU and national law in both cases (German Animal Welfare Act 2006; German Order for the Protection of Production Animals used for Farming Purposes and other Animals kept for the Production of Animal Products 2006). No pain, suffering or injury was inflicted on any animals during the study.

Data collection

The observers of this study (observer 1: female, aged 27, veterinarian with experience in pigs; observer 2: female, aged 26, agricultural science student with experience in pigs) initially participated in a training session for three days provided by two members of the Welfare Quality® consortium. They achieved good interobserver reliability, which was evaluated via assessments of pictures, video sequences and on-farm, since the training was provided until 90% of the assessments were consistent. In addition, a preliminary study with ten joint farm visits was conducted to ensure good training status before the main data collection phase began. Data from the preliminary study have not been included. Re-evaluation and re-training using pictures and video sequences were performed after the first half of the data collection phase to minimise observer drift.

The data collection phase took place between September 2016 and April 2017 at nine farms in Schleswig-Holstein, Germany. The farms' characteristics are shown in Table 1.

The two trained observers jointly performed 40 assessments at these farms. The number of visits was assigned randomly to the farms, ranging from one to eight visits per farm. The number of farm visits was not uniformly distributed across the farms for practical reasons, eg management procedures, such as weaning, which precluded visits from two observers. Therefore, there was an average interval between joint visits of six days,

https://doi.org/10.7120/09627286.29.3.323 Published online by Cambridge University Press

Table 1 Overview of the nine study farms.

Farm	N	PR	F[1]	B	G[2]	BD	W	FR[3,4]	MS[4]	MP[4]
1	400	1	CR, FP	3–4	D	Rubber mats	26	88.8	5.00	19.4
2	120	3	CR	28	D	Straw	28	87.0	1.00	14.9
3	330	1	CR	28	S	None	26	n/a	7.00	11.5
4	80	3	CR	28	D	None	25	79.1	8.00	18.1
5	150	2	CR	36	D	Rubber mats	21	88.9	8.00	15.7
6	810	1	CR	28	S	None	28	85.5	6.00	12.8
7	5,000	1	CR, FP	3–4	S	None	25	83.4	6.00	14.3
8	180	1	CR	28	D	Straw	24	88.4	2.00	10.4
9	240	3	CR	28	S	None	23	89.5	6.00	9.65

N: Herd size; PR: Production rhythm (weeks); F: Farrowing system; B: Period in breeding unit (days); G: Regime in the gestation unit; BD: Type of bedding in the gestation unit; W: Average age at weaning (days); FR: Farrowing rate (%); MS: Mortality rate in sows (%); MP: Mortality rate in piglets (%).
[1] CR: Crates; FP: Free farrowing pens;
[2] D: Dynamic major group; S: Stable groups in pens;
[3] n/a: Not available;
[4] Average for the years 2016 and 2017.

which ranged between one and 27 days. Observers never discussed the outcomes of farm visit so as to avoid any bias in subsequent visits. Thus, the farm visits were treated as independent assessments. Of the 40 assessments, 20 were conducted by applying the entire Welfare Quality® protocol (ie behavioural measures and measures of health and physical state, covering farrowing, breeding and gestation units), while 20 comprised health and physical state measures only, and those assessments were restricted to the gestation unit. In all cases, the same animals were assessed simultaneously but independently by the two observers.

On-farm application of the protocol

The Welfare Quality® animal welfare assessment protocol for sows and piglets is comprised of behavioural measures and measures of health and physical state. The behavioural measures are composed of four parts. First, a qualitative behaviour assessment (QBA), including the different areas of the farm, is performed. Subsequently, an assessment of social and exploratory behaviours, a human-animal relationship test and assessment of stereotypies are performed in the gestation unit. The health and physical state measures are recorded in a sample of animals in farrowing (including samples of piglets), breeding and gestation units. The study was carried out strictly in accordance with protocol specifications (see below) and in the description, emphasis is given to those specific measures arising from the interobserver reliability study. Detailed information regarding the overall methodology can be seen in the Welfare Quality® animal welfare assessment protocol for pigs (Welfare Quality® 2009).

Qualitative behaviour assessment (QBA)

The qualitative animal-based measure QBA is included in the protocol in order to assess the animals' emotional state and was conducted in the barn at four to six observation points, using observations at the group level, for a total observation time of 20 min. The various parts of the farms, eg farrowing, breeding and gestation units, were included in the observations. In the observation time allotted, observers viewed the expressive quality of the activities of all the animals able to be adequately observed from each observation point. Following this, the expressive quality was assessed by rating 20 adjectives on a visual analogue scale of 125 mm ranging from absent (0 mm) to dominant (125 mm) where the sum of the scores for each adjective described the expressive quality of the animals for all the observation points at one particular farm.

Assessment of social and exploratory behaviours

The number of locations used for the assessment of social and exploratory behaviours depended on the group size and varied from one to four pens in the gestation unit. The goal was to attain an overall picture of the animals, similar to the QBA. First, the animals in the pens chosen for the observations were roused by one of the observers as they walked around the pen in question. However, all observations were performed from outside the pen. Sneezing and coughing, which will be linked to measures of health and physical state in subsequent analyses, were counted during a calming period of 5 min. Following this, the assessment of social and exploratory behaviours was performed for a total observation time of 10 min using instantaneous scan sampling with five scan samples at intervals of 2 min. After the assessment, sows were observed at the group level to record measures of health and physical state, such as panting and huddling.

Human-animal relationship test

The human-animal relationship test was performed on a sample of 20 sows in the gestation unit. As the sows were already aware of the observers, the human-animal relationship test was performed following the assessment of social and exploratory behaviours. The human-animal relationship test was always performed by one of the observers strictly in accordance with protocol instructions, while the other observed the animal's reaction from a distance. The role of observers in the human-animal relationship test was randomly assigned for each visit. Each observer was assigned to the active position in one half of the visits and the passive position in the other. The human-animal relationship test was scored on a three-point scale (0 = no fear response, 1 = slight fear response, and 2 = strong fear response). Additionally, observers sought the presence of liquid faeces on the floor while walking around the pen, a factor associated with the measure of health and physical state, scouring.

Assessment of stereotypies

The assessment of stereotypies was performed through observation of a sample of 40 sows in the gestation unit for signs of stereotypical behaviours, such as sham chewing, tongue rolling, teeth grinding, bar, trough or drinker biting and floor licking. Assessments were carried out in the morning when sows were more active but not during feeding. The observers selected each sow sequentially from a distance using the random sampling method as described in the protocol, simultaneously but independently watching the same animals for 15 s to ascertain whether stereotypical behaviours were being performed. If observers were unsure as to whether sows were showing stereotypical behaviour, the period was extended up to 30 s. A binary score (0 = absence or 2 = presence) was used to assess stereotypies.

Measures of health and physical state

A variety of health and physical state measures, such as shoulder sores, metritis and body wounds were assessed in farrowing, breeding and gestation units. Detailed information on scoring protocols can be seen in the Welfare Quality® animal welfare assessment protocol for pigs (Welfare Quality® 2009). Measures of sow health and physical state were assessed on only one side of the pig (the side most clearly visible at time of inspection). Either a three-point (0 = absent, 1 = slight affliction, and 2 = strong affliction) or a binary scale (0 = absence or 2 = presence) was used for scoring. Piglets were first scored individually: a group score was calculated from these scores.

Statistical analysis

All data processing procedures and statistical analyses were performed using SAS® 9.4 statistical software (SAS® Institute Inc 2008).

Data and processing for analyses

Measures with a prevalence less than 0.05% recorded by both observers were excluded from interobserver reliability analysis and not presented here. Thus, measures of health and physical state, including coughing, huddling, scouring, sneezing, constipation, pumping, rectal prolapse, skin condition, mastitis, ruptures and hernias, uterine prolapse and panting category 2 (group level) in sows and measures of neurological disorders, panting category 1, pumping, rectal prolapse and splayed legs assessed in piglets were not included in the analysis.

Evaluation of interobserver reliability at individual animal level

No data processing procedure was applied to the values recorded by the two observers for individual animals in the human-animal relationship test, the assessment of stereotypies and the measures of health and physical state in the analysis of interobserver reliability at individual animal level, but values were compared directly. The QBA and the assessment of social and exploratory behaviours were evaluated solely at the farm level and therefore not analysed at individual animal level.

Evaluation of interobserver reliability at farm level

The results of the QBA were calculated for each adjective by reading out the length (mm) on the visual analogue scale with a ruler. The score was expressed as a percentage of the total scale length. Thus, the dataset contained one score as a percentage for each adjective recorded by both observers for each farm visit (eg farm visit 1, observer 1: happy: 49%, range: 0–100%).

The results of the assessment of social and exploratory behaviours were reported for both observers as the percentage of animals performing a certain behaviour of all animals that were active during each farm visit (eg farm visit 1, observer 1: positive social behaviour: 5%, negative social behaviour: 1%, use of enrichment material: 11%, investigation of the pen: 8%, other active behaviour: 75%, sum: 100%).

For analysis at farm level, the results of the human-animal relationship test, the assessment of stereotypies and the measures of health and physical state were calculated for both observers in each farm visit as the percentage of animals allocated into the corresponding categories (eg farm visit 1, observer 1: wounds on the body category 0: 90%, wounds on the body category 1: 6%, wounds on the body category 2: 4%, sum 100%). The categories were treated as individual variables and therefore compared individually, eg wounds on the body 0, wounds on the body 1, and wounds on the body 2, to examine where differences in interobserver reliability between percentages of animals with wounds had occurred.

Two datasets were included in the analysis at farm level because assessments applying the entire protocol (ie behavioural measures and measures of health and physical state in the farrowing, breeding and gestation units) and assessments focusing on measures of health and physical state in the gestation unit were performed. One dataset contained the 20 observations recorded by each observer for the QBA, the assessment of social and exploratory behaviours, the human-animal relationship test, the assessment of stereotypies and measures of health and physical state assessed in the farrowing and breeding units. The second dataset covered the 40 observations recorded by each observer for measures of health and physical state assessed in the gestation unit.

https://doi.org/10.7120/09627286.29.3.323 Published online by Cambridge University Press

Statistical analysis

Evaluation of interobserver reliability at the individual animal level: The evaluation of interobserver reliability in the human-animal relationship test, the assessment of stereotypies and the measures of health and physical state at individual animal level were performed by applying different kappa coefficients (Cohen's kappa coefficient [κ], weighted kappa [κ_w] and prevalence-adjusted, bias-adjusted kappa [PABAK]). Acceptable interobserver reliability was assigned to the measures when all statistical parameters reached the values defined as acceptable, which are described below.

Cohen's kappa: Cohen's kappa coefficient (κ) is a change-adjusted measure of agreement between observers (Cohen 1960). The scale suggested by Landis and Koch (1977) was adapted for the interpretation, resulting in the following classification: values equal to or greater than 0.40 were assigned as acceptable agreement and values equal to or greater than 0.60 were designated as good agreement.

Weighted kappa: The weighted kappa (κ_w) coefficient is used to reflect the degree of disagreement. Thus, it attributes greater significance to large differences between observers than smaller ones, whereas unweighted kappa treats all disagreements equally (Cohen 1968). Quadratic weights based on agreement were used (Cohen 1968), as they are more sensitive to the number of categories and therefore place more weight on observations that are further apart (Brenner & Kliebsch 1996). In terms of interpretation, values equal to or greater than 0.40 were deemed to show acceptable agreement and values equal to or greater than 0.60 good agreement.

Prevalence-adjusted, bias-adjusted kappa (PABAK): Furthermore, the kappa value may be affected by the presence of bias between the observers and by the distribution of data across the categories used, which is called the prevalence (Feinstein & Cicchetti 1990). Byrt *et al* (1993) introduced measures of prevalence and bias and a formula to adjust the kappa value for these factors. The prevalence-adjusted, bias-adjusted kappa (PABAK) is calculated using the formula:

$$PABAK = 2 \times P_0 - 1$$

where P_0 represents the proportion of observed agreement, which is the sum of the main diagonal cells in a 2 × 2 contingency table for measures with a binary score, or a 3 × 3 contingency table for measures with a three-point scale, respectively. The interpretation was based on an interobserver study conducted by Plesch *et al* (2010), who assigned a PABAK of 0.75 as indicating excellent agreement which, in turn, was rated as good agreement in the present study. Compared with the other statistical parameters, an acceptable agreement was assigned to a PABAK equal to or greater than 0.40.

Evaluation of interobserver reliability at farm level: This was performed by applying the reliability parameters Spearman's rank correlation coefficient (RS) and intraclass correlation coefficient (ICC) and the agreement parameters smallest detectable change (SDC) and limits of agreement (LoA). Thus, the term 'reliability' is used to represent results of reliability parameters, and the term 'agreement' refers to agreement parameters. The term 'interobserver reliability' is used to summarise the results in a final general evaluation. Once again, acceptable interobserver reliability was defined as being when all statistical parameters reached the values defined as acceptable, which are presented below.

Spearman's rank correlation coefficient (RS): RS is a non-parametric measure of rank correlation and is often used in animal welfare science (Gauthier 2001; Dalmau *et al* 2010). In terms of interpretation, an RS equal to or greater than 0.40 was deemed acceptable reliability and RS equal to or greater than 0.70 good reliability (Martin & Bateson 2007).

Intraclass correlation coefficient (ICC): The basis of the ICC is variance. The ICC is defined as the proportion of the variance between study objects (farms) to the variance between study objects plus measurement error (de Vet *et al* 2006).

The following two-way model proposed by Shrout and Fleiss (1979) was used for the fundamental analysis of variance:

$$X_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$

where X_{ijk} represents the measured value, μ represents the general average value, α_i represents the fixed effect of the difference between the study objects (farms), β_j represents the random effect of the observers and ε_{ijk} represents the general error term.

According to the formula of consistency published by de Vet *et al* (2006), the ICC was calculated using the following equation:

$$ICC = \frac{\sigma^2 (objects)}{\sigma^2 (objects) + \sigma^2 (residual)}$$

where σ^2 represents the variance of the study objects and the residual variance, respectively.

By definition, the ICC ranges from 0 to 1. Therefore, a value of 0 indicates the total absence of reliability and a value of 1 indicates perfect reliability. In terms of interpretation, an ICC equal to or greater than 0.40 implied acceptable reliability and an ICC equal to or greater than 0.70 implied good reliability (McGraw & Wong 1996).

Smallest detectable change (SDC): SDC is an expression of the measurement error. The measurement error contains the residual variance. According to de Vet *et al* (2006), the SDC is calculated using the following equation:

$$SDC = 1.96 \times (\sqrt{2} \times \sigma^2 [residual])$$

where σ^2 represents the residual variance.

SDC outputs the smallest change in the score that is detectable, despite the measurement error. The values of the SDC reflect the scale unit of the assessed measures, which are expressed as a percentage displayed as a decimal number in this study. In accordance with the simple agreement coefficient calculated by de Vet *et al* (2006), an SDC less than or equal to 0.10 was interpreted as acceptable agreement, and an SDC less than or equal to 0.05 was defined as good agreement.

Table 2 Mean (± SEM) percentage values for qualitative behaviour assessment adjectives assigned to the two observers and corresponding statistical parameters.

Adjectives	Observer 1	Observer 2	RS	ICC	SDC	LoA
Active	48.9 (± 5.30)	41.5 (± 4.04)	*0.41*	0.37	0.46	–0.39 to 0.54
Relaxed	64.0 (± 5.90)	47.3 (± 5.57)	0.26	0.31	0.57	–0.40 to 0.74
Fearful	11.7 (± 3.25)	24.1 (± 4.04)	**0.87**	*0.63*	0.20	–0.33 to 0.08
Agitated	27.6 (± 5.49)	47.3 (± 5.47)	*0.53*	0.34	0.51	–0.70 to 0.31
Calm	74.4 (± 6.02)	55.7 (± 4.83)	*0.48*	0.36	0.49	–0.31 to 0.68
Content	51.2 (± 5.13)	42.1 (± 4.12)	–0.51	0.00	0.72	–0.63 to 0.81
Tense	25.7 (± 5.95)	40.0 (± 4.81)	*0.41*	*0.41*	0.48	–0.63 to 0.34
Enjoying	44.9 (± 5.97)	40.2 (± 3.83)	0.37	0.34	0.49	–0.46 to 0.55
Frustrated	30.3 (± 4.99)	43.6 (± 4.55)	*0.42*	0.36	0.45	–0.58 to 0.32
Sociable	54.9 (± 5.58)	49.6 (± 4.81)	*0.54*	*0.57*	0.42	–0.37 to 0.47
Bored	34.4 (± 5.68)	49.6 (± 4.68)	*0.59*	*0.46*	0.43	–0.59 to 0.28
Playful	14.3 (± 2.79)	40.4 (± 5.03)	*0.55*	0.19	0.38	–0.66 to 0.14
Positively occupied	48.6 (± 6.08)	50.0 (± 3.95)	0.35	0.26	0.54	–0.56 to 0.54
Listless	13.2 (± 3.39)	37.4 (± 3.77)	0.19	0.01	0.44	–0.68 to 0.20
Lively	8.7 (± 2.86)	49.1 (± 3.92)	0.06	0.00	0.43	–0.84 to 0.04
Indifferent	72.5 (± 5.26)	36.1 (± 3.79)	0.04	0.03	0.54	–0.18 to 0.91
Irritable	9.7 (± 2.91)	38.1 (± 4.75)	0.23	0.10	0.42	–0.71 to 0.15
Aimless	27.7 (± 5.09)	49.5 (± 4.38)	*0.48*	0.36	0.40	–0.62 to 0.18
Happy	50.4 (± 5.31)	46.2 (± 4.88)	0.38	0.37	0.50	–0.46 to 0.55
Distressed	12.7 (± 3.16)	27.5 (± 4.61)	*0.40*	0.28	0.38	–0.54 to 0.24

RS: Spearman's rank correlation coefficient;
ICC: Intraclass correlation coefficient;
SDC: Smallest detectable change;
LoA: Limits of agreement;
Normal type: poor; italics: acceptable; and bold: good interobserver reliability.

Limits of agreement (LoA): LoA were calculated using the method described by Bland and Altman (1986) with the following formula reported by de Vet *et al* (2006):

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

where σ^2 represents the residual variance.

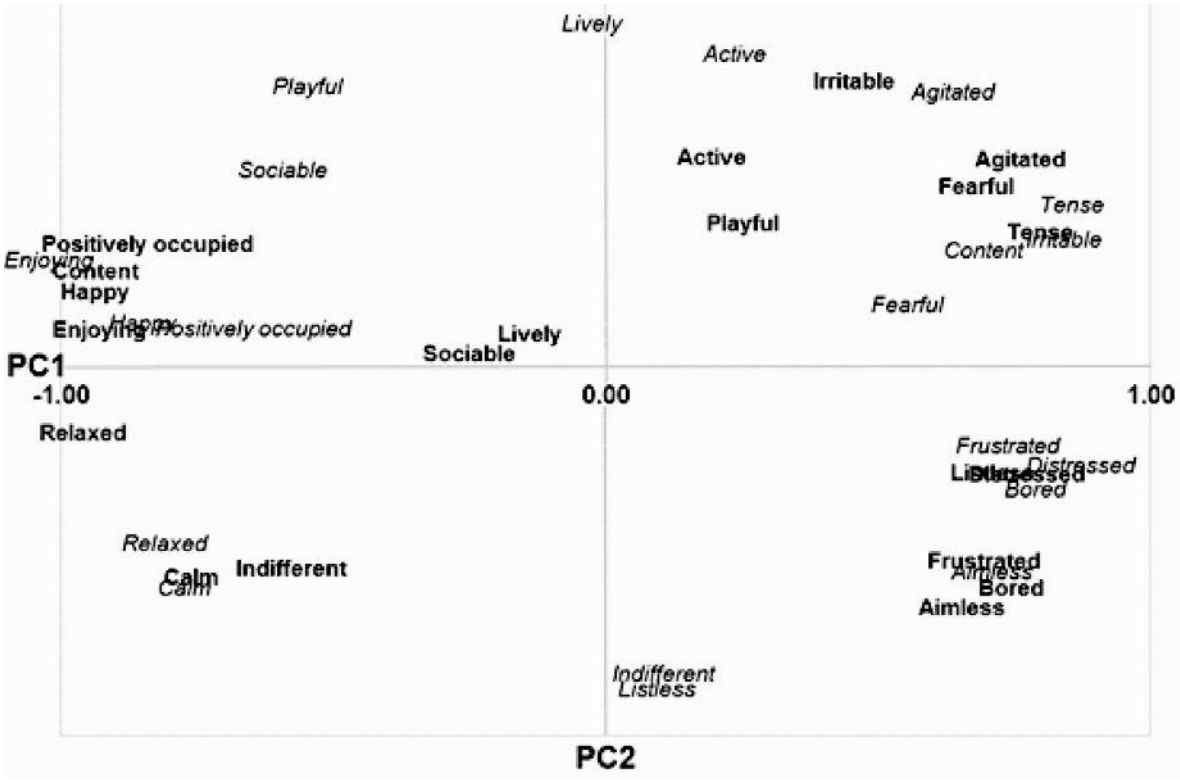
LoA estimate the differences between two sets of measured values which, in this case, was the differences of the measurements obtained by the two observers for each farm visit and the standard deviation of these differences. Most of the differences are expected to be less than two standard deviations. In this study, LoA are expressed as a relative frequency ranging from –1 to 1. The value of –1 indicates that higher values were obtained by observer 2 and the value of 1 indicates that higher values were obtained by observer 1. Moreover, the interpretation was based on the simple agreement coefficient described by de Vet *et al* (2006) and therefore an interval less than or equal to –0.10 to 0.10 indicated acceptable agreement, and an interval less than or equal to –0.05 to 0.05 indicated good agreement.

Principal Component Analysis

The QBA was further analysed using a Principal Component Analysis (PCA). Therefore, the raw data, ie the percentages of the adjectives recorded by each observer on the visual analogue scale, were transformed into a correlation matrix and no rotation was applied because a correlation matrix is a more balanced representation of the adjectives used in the QBA (Temple *et al* 2013). A single PCA was calculated for each of the two observers. The results of these two PCA were then compared. The two principal components (PC; PC1 and PC2) identified with an eigenvalue greater than 1.0 were used. Each adjective achieved a certain factor loading on PC1 and PC2, which is a dimensionless number ranging

https://doi.org/10.7120/09627286.29.3.323 Published online by Cambridge University Press

Figure 1



Word chart plot using the values of the first two principal components (PC1 and PC2) for the farm visits performed by observer 1 (bold) and observer 2 (italics).

from –1 to 1. Finally, RS between factor loadings on PC1 and between factor loadings on PC2 of the two observers were calculated. An RS equal to or greater than 0.40 was interpreted as an acceptable correlation and an RS equal to or greater than 0.70 was defined as a good correlation (Martin & Bateson 2007).

Results

Qualitative behaviour assessment (QBA)

Table 2 shows the mean (± SEM) percentage values for each observer and corresponding statistical parameters for the interobserver reliability of the QBA adjectives. Poor interobserver reliability was observed in direct comparison of the percentages of any of the adjectives, with agreement indicated by the values of SDC and LoA being low for all, although some adjectives showed RS and ICC values equal to or greater than 0.40, such as the term 'fearful'.

As explained above, the QBA was further analysed using a PCA. The factor loadings on the first two components explained 68.6% of the variance for observer 1 and 74.4% for observer 2. The values obtained were plotted in a two-dimensional interpretative word chart, which is shown in Figure 1.

In contrast to the direct comparison of the QBA terms, the interobserver reliability indicated by the PCA showed good reliability. The RS between PC1 of the two observers reached 0.77 and between PC2, 0.76.

Assessment of social and exploratory behaviours

Table 3 lists the mean (± SEM) percentage values obtained for each observer and the corresponding statistical parameters for the interobserver reliability of categories in the assessment of social and exploratory behaviours. The percentages of animals in the prescribed categories assigned by the observers showed a number of discrepancies between observers. Only the category 'negative social behaviour' achieved good interobserver reliability in all statistical parameters. Low reliability was indicated by the values of RS for the categories 'positive social behaviour' and 'investigation of the pen'. Furthermore, the values of SDC and LoA showed low agreement for 'use of enrichment material' and 'other active behaviour'.

Human-animal relationship test, assessment of stereotypies and measures of health and physical state

Table 4 shows the mean (± SEM) values for the two observers together with corresponding statistical parameters for the interobserver reliability analysis of the human-animal relationship test, the assessment of stereotypies and measures of health and physical state in sows and piglets at individual animal level with a prevalence greater than 0.05% for at least one of the observers.

Table 5 presents the mean (± SEM) percentage values of the analysis of interobserver reliability for the human-animal relationship test, the assessment of stereotypies and measures of health

Table 3 Mean (± SEM) percentage values for the categories in the assessment of social and exploratory behaviours assigned to the two observers and corresponding statistical parameters.

Indicator	Observer 1	Observer 2	RS	ICC	SDC	LoA
Positive social behaviour	1.0 (± 0.27)	0.6 (± 0.29)	0.16	0.31	**0.03**	**–0.02 to 0.03**
Negative social behaviour	1.0 (± 0.30)	1.2 (± 0.56)	*0.44*	*0.44*	**0.04**	**–0.04 to 0.04**
Use of enrichment material	11.7 (± 4.09)	17.5 (± 6.03)	**0.85**	**0.78**	0.29	–0.35 to 0.23
Investigation of the pen	1.0 (± 0.27)	0.6 (± 0.29)	0.16	0.31	**0.03**	**–0.02 to 0.03**
Other active behaviour	85.2 (± 4.03)	80.1 (± 5.85)	**0.86**	**0.80**	0.27	–0.22 to 0.32

RS: Spearman's rank correlation coefficient;
ICC: Intraclass correlation coefficient;
SDC: Smallest detectable change;
LoA: Limits of agreement;
Normal type: poor; italics: acceptable; and bold: good interobserver reliability.

and physical state in sows and piglets at farm level as recorded by the two observers and the corresponding statistical parameters. This can be seen in the supplementary material to papers published in *Animal Welfare*; https://www.ufaw.org.uk/the-ufaw-journal/supplementary-material on the UFAW website. Measures with a prevalence less than 0.05% in the assessments of both observers are not presented. Only category 2 is presented if measures were assessed with a binary score (0 = absence or 2 = presence). All categories are presented if measures were assessed with a three-point scale (0 = absent, 1 = slight affliction, and 2 = strong affliction).

Human-animal relationship test

The evaluation of the human-animal relationship test at individual animal level showed good interobserver reliability. However, PABAK indicated effects of prevalence and bias. Good interobserver reliability was achieved in the analysis at farm level for category 0, whereas the agreement parameters, SDC and LoA, indicated low agreement for category 1 and 2.

Assessment of stereotypies

The assessment of stereotypies evaluated at individual animal level showed acceptable to good interobserver reliability for all measures. Interobserver reliability was good when results were adjusted for prevalence and bias. Similarly, the assessment of stereotypies showed overall acceptable to good interobserver reliability in the farm level evaluation. Only the measure 'tongue rolling' achieved low agreement, as indicated by the values of the agreement parameters, SDC and LoA.

Measures of health and physical state

All health and physical state measures with a prevalence greater than 0.05% for at least one of the observers were assigned acceptable to good interobserver reliability in the evaluation at individual animal level, apart from bursitis and shoulder sores in sows and huddling in piglets. In most measures included in the evaluation at farm level, acceptable to good interobserver reliability was detected. In contrast, all statistical parameters indicated poor interobserver reliability for huddling and lameness in piglets.

Discussion

Data collection

To date, reliability studies have varied in terms of experimental design and a standard is yet to be established for studies with this aim (Phythian *et al* 2013a) but a minimum of three observers is recommended (Walter *et al* 1998). The present study was conducted using only two observers since on-farm assessments are more time consuming and therefore more expensive. As described by Walter *et al* (1998), a lower number of observers is counterbalanced by a greater number of observations. Thus, here, up to 40 joint farm visits were carried out with the maximum number of 5,285 sows assessed to increase the validity of the results. Nevertheless, interpretation is limited by the small number of observers, and our report should be considered a case study.

Furthermore, a significant effect of observer bias has been reported in a study assessing reliability (Hewetson *et al* 2006). Therefore, the effects of observers' differing attitudes, backgrounds and character traits were discussed (Bokkers *et al* 2012), and factors influencing assessment, notably observer prejudices or personal interests, an equivocal information base or a subjective method of scoring, were included in a review of available studies (Tuyttens *et al* 2014). Here, possible influences on observers, such as observation of the actual animals and the handling of animals, cleanliness of the farm or medical supplies, were unavoidable because the study took place on-farm (Phythian *et al* 2012, 2013a). The use of trained observers with species-specific experience was recommended as a way of overcoming bias (Phythian *et al* 2016). Thus, our experimental design, utilising trained and experienced observers, gives greater credence to the findings. Furthermore, adequate training (Windschnurer *et al* 2008) and continuous re-evaluation (Gibbons *et al* 2012), as applied in the present study, are advised.

Our study was limited to nine farms, all of which participated voluntarily. The number of visits was randomly assigned and ranged from one to eight per farm, with the interval between the joint visits averaging six days and

https://doi.org/10.7120/09627286.29.3.323 Published online by Cambridge University Press

Table 4 Mean (± SEM) values of the human-animal relationship test, the assessment of stereotypies and measures of physical state and health of sows and piglets by the two observers and corresponding statistical parameters.

Measure	n	Observer 1	Observer 2	κ	κ_w	PABAK
Human-animal relationship test						
Fear of humans	392	0.67 (± 0.04)	0.57 (± 0.04)	**0.72**	**0.87**	*0.68*
Stereotypies						
Sham chewing	741	0.47 (± 0.03)	0.42 (± 0.03)	**0.77**	**0.77**	**0.84**
Tongue rolling	741	0.18 (± 0.02)	0.22 (± 0.02)	*0.49*	*0.49*	**0.82**
Teeth grinding	741	0.02 (± 0.01)	0.01 (± 0.004)	*0.40*	*0.40*	**0.98**
Bar/drinker/trough biting	741	0.03 (± 0.01)	0.04 (± 0.01)	**0.72**	**0.72**	**0.98**
Floor licking	741	0.04 (± 0.01)	0.08 (± 0.01)	*0.51*	*0.51*	**0.94**
Sows						
Wounds on the body	5,240	0.09 (± 0.004)	0.07 (± 0.004)	*0.48*	*0.51*	**0.86**
Vulva lesions	2,721	0.13 (± 0.01)	0.16 (± 0.01)	*0.59*	**0.69**	**0.81**
Body condition score	3,148	0.09 (± 0.01)	0.08 (± 0.01)	*0.43*	*0.47*	**0.83**
Absence of manure on the body	5,285	0.06 (± 0.003)	0.02 (± 0.002)	0.33	0.44	**0.90**
Metritis	559	0.16 (± 0.02)	0.09 (± 0.02)	**0.61**	**0.61**	**0.91**
Lameness	2,636	0.01 (± 0.002)	0.01 (± 0.002)	**0.61**	**0.66**	**0.99**
Local infections	4,858	0.19 (± 0.01)	0.19 (± 0.01)	0.36	*0.46*	*0.65*
Bursitis	5,285	0.43 (± 0.01)	0.54 (± 0.01)	0.34	0.39	0.30
Shoulder sores	140	0.42 (± 0.05)	0.44 (± 0.05)	0.31	*0.47*	0.30
Panting	160	1.19 (± 0.08)	1.41 (± 0.07)	*0.53*	*0.53*	*0.58*
Piglets						
Sneezing	164	0.24 (± 0.05)	0.11 (± 0.04)	*0.59*	*0.59*	**0.87**
Scouring	164	0.11 (± 0.04)	0.02 (± 0.02)	0.35	0.35	**0.91**
Absence of manure on the body	164	0.01 (± 0.01)	0.02 (± 0.01)	0.33	**0.66**	**0.98**
Lameness	164	0.12 (± 0.03)	0.07 (± 0.02)	*0.42*	*0.56*	**0.83**
Huddling	72	0.68 (± 0.11)	0.97 (± 0.12)	0.35	0.38	0.31
Panting	164	0.00 (± 0.00)	0.04 (± 0.02)	a	0.00	**0.95**

κ: Cohen's kappa coefficient;
κ_w: weighted kappa;
PABAK: prevalence-adjusted, bias-adjusted kappa;
Normal type: poor; italics: acceptable; and bold: good interobserver reliability;
[a] Calculation not possible due to lack of variance in one of the observers.

ranging between one and 27 days. Any results were not discussed by observers to avoid introducing bias in subsequent visits, which is why visits were treated as independent. In addition to time and cost considerations, data collection was also limited to nine farms since biosecurity considerations meant a 48-h quarantine between visits to different farms. As regards the question of whether the farm number was sufficient to evaluate reliability, high inter-farm variability and a prevalence of approximately 50% for the measures under assessment are important for an evaluation of reliability (Hoehler 2000; Burn *et al* 2009). A low prevalence may lead to an artificially low reliability (Hoehler 2000; Plesch *et al* 2010). Thus, diversity of farms within a study is of more importance than the number. Requirements were adapted through picking farms with different herd sizes (40 to 5,000 sows) or production rhythms (one-week, two-

week and three-week rhythms). Nevertheless, the prevalence was low in the present study, which limits the interpretation of the results. Further studies are needed that include hospital pens (Mullan *et al* 2011) or international participation to increase the prevalence of the measures and ultimately evaluate the interobserver reliability of the measures of the Welfare Quality® protocol for sows and piglets.

Statistics

The measures included in the Welfare Quality® protocol for sows and piglets were evaluated at individual animal and farm level. The evaluation at individual animal level was performed for the human-animal relationship test, the assessment of stereotypies and for measures of health and physical state. The evaluations of the QBA and the assessment of social and exploratory behaviour were only performed at farm level, as required in the protocol. The additional farm level analysis was carried out because the animals under assessment are normally part of a randomly selected sample (Welfare Quality® 2009). By summarising the data at farm level, small deviations in interobserver reliability are compensated for, which explains why farm level analysis is a decisive evaluation of an assessment tool for animal welfare. Furthermore, evaluation at farm level enables detection of differences between the observers within each measure's specific categories. The categories of the human-animal relationship test, the stereotypies and measures of health and physical state were treated as individual variables and therefore evaluated independently in the analysis conducted at farm level, although animals received scores in either one of the categories, implying that the categories are not independent. Nevertheless, this method helps to illuminate differences between the distinct categories and was also applied in an interobserver study by Czycholl *et al* (2016a), in test-retest studies in sows and piglets (Friedrich *et al* 2019a,b) and in a study of growing pigs (Temple *et al* 2011).

Different statistical parameters were chosen to evaluate interobserver reliability, offset their disadvantages and avoid misinterpretation, as suggested by Byrt *et al* (1993) and de Vet *et al* (2006). Common advantages and disadvantages of the use of kappa coefficients, reliability and agreement parameters and how these were addressed in the present study are discussed below.

Kappa coefficients, which were used to evaluate interobserver reliability at individual animal level, measure the agreement between observers adjusted by chance (Cohen 1960). A requirement for the use of kappa coefficients is the independency of the study objects and the independency of observer ratings (Brennan & Prediger 1981), otherwise kappa values decrease. Observer independence was upheld since each assessed the same animal at the same time without any interaction. Furthermore, the outcomes of farm visits were not discussed during the entire data collection process.

The evaluation of interobserver reliability at farm level was conducted by applying reliability and agreement parameters. Reliability parameters, such as RS and ICC, are correlation coefficients that evaluate the degree of differentiation between study objects despite the measurement error. The parameters are limited by their strong dependency on the total variance of the assessed objects. Thus, reliability parameters achieve higher values if there is large variability between the study objects and smaller values if the study objects do not vary substantially, despite good reliability (de Vet *et al* 2006). Based on our results, the RS was unable to be calculated if no variance occurred in the data recorded by one of the observers. This dependency on the total variance must be taken into account when analysing reliability in order to avoid misinterpretations (Wirtz & Caspar 2002).

Agreement parameters include the SDC and LoA, although the SDC is mathematically derived from the ICC. These parameters assess the extent of similarity of the results of repeated measurements by estimating the measurement error. Even though the parameters depend on the variance of the data, the subjective definition of threshold values remains problematic. Thus, in the present study, the threshold values were oriented towards pre-existing reliability studies (Temple *et al* 2013; Czycholl *et al* 2016a; Friedrich *et al* 2019 a,b).

Qualitative behaviour assessment (QBA)

The QBA was only evaluated at farm level since it was not performed at individual animal level. The percentages obtained by the two observers for each adjective included in the QBA were compared directly. The results showed poor interobserver reliability. For some adjectives, the RS and ICC showed good reliability, but this conclusion was not verified by all statistical parameters (RS, ICC, SDC and LoA). These results are consistent with the findings of Czycholl *et al* (2016a), who tested the Welfare Quality® protocol for growing pigs.

This insufficient interobserver reliability is explained by adjectives not being used in the same manner by observers or observers interpretating the adjectives differently (what does a happy sow look like?), although the evaluation of training showed good interobserver reliability. For example, one observer scored an animal as playful while the other scored the sow's behaviour as active. This discrepancy is explained by a certain redundancy among the adjectives. Furthermore, the visual analogue scale may have been used to varying degrees by observers while assessments were being carried out, eg one observer may have used the scale more in the medium range, while the other may have entered more extreme values.

However, based on the results of the PCA, the observers assigned similar dimensions of behavioural styles to the animals, eg active play behaviour, although the adjectives were used in a different manner. This finding suggests a common trend in terms of the way in which different adjectives relate to each other for the detection of which PCA is a suitable method. PCA is a common measurement tool used to analyse the results of the QBA regarding redundancies between the adjectives (Wemelsfelder *et al* 2000, 2001; Wemelsfelder & Millard 2009). If, for instance, one observer scored the animals as playful and the other as lively, these adjectives will still be scored as agreeing

https://doi.org/10.7120/09627286.29.3.323 Published online by Cambridge University Press

because both observers assessed the active behaviour of the animals. The results of the present study are consistent with findings reported in the literature (Wemelsfelder *et al* 2000, 2009; Napolitano *et al* 2008; Walker *et al* 2010; Phythian *et al* 2013b). However, a large number of these studies were based on video sequences and used a free-choice profiling approach instead of the fixed-rating scale method used for the QBA in the Welfare Quality® protocols. Therefore, the results of the studies are not directly comparable. However, the results are supported by the data presented by Knierim *et al* (2007), who also observed good interobserver reliability in the QBA in laying hens using a fixed-rating scale and analysis applying a PCA. In contrast, Czycholl *et al* (2017), who also performed the fixed-rating scale of the Welfare Quality® protocols on-farm, only identified poor interobserver reliability for the QBA in growing pigs when they analysed their data using a PCA. As animals become less active as they age (Docking *et al* 2008), the QBA might be easier to perform in sows than growing pigs.

One possible approach to further increase interobserver reliability at individual adjective level might be to perform the QBA in sows by analysing individual animals, which are easier to observe than a large group, as described in the AWIN welfare assessment protocol for horses (AWIN 2015). However, this approach requires consideration of an adequate sample size and, more specifically, the feasibility of the protocol. A compromise that accounts for the feasibility would therefore be to score a separate QBA for every observation point rather than to summarise the observation points of the various parts of the farm into a single score. Both approaches must be evaluated to determine their objectivity, including validity, reliability and feasibility.

Assessment of social and exploratory behaviours

In general, the assessment of social and exploratory behaviours showed acceptable interobserver reliability, although the categories 'positive social behaviour' and 'investigation of the pen' showed low reliability, as indicated by the values of RS and ICC. As explained above, reliability parameters evaluate the degree of differentiation between study objects. The reliability of the correlation coefficients depends on the variance of the data. If the values of the study objects are similar, the measurement error affects the ability to distinguish the study objects and the reliability is lower (de Vet *et al* 2006). The behavioural categories 'positive social behaviour' and 'investigation of the pen' were only rarely recorded. This low rate might explain the low reliability of these categories, since the agreement parameters, SDC and LoA, showed good agreement. On the other hand, agreement parameters assess the extent to which the observers assign the same precise value to an object and are not affected by the variance of the data. The behavioural categories, 'use of enrichment material' and 'other active behaviour' showed good reliability but only poor agreement. Observers tended to define both behaviours differently, resulting in poor agreement but good reliability, as these two categories appear to interact.

Czycholl *et al* (2016a) found that in growing pigs observers were able to allocate the animals to the same behavioural categories — consistent with our findings. In contrast, Munsterhjelm *et al* (2015) reported a profound effect of the observers on behavioural observations, indicating that significant subjectivity existed in on-farm assessments. However, since Munsterhjelm *et al* (2015) used a multivariate analysis for their analysis, a comparison between their results and ours is difficult to perform.

Human-animal relationship test

The human-animal relationship test showed good interobserver reliability in evaluation at the level of the individual animal. However, PABAK indicated the effects of prevalence and bias on the sample, since the agreement was lower when the evaluation was adjusted for these factors. In the evaluation at farm level, the human-animal relationship test showed good interobserver reliability for category 0 but only poor reliability for categories 1 and 2. This discrepancy is most likely attributable to insufficient demarcation between category 1 and 2. Although category 0 reached good interobserver reliability, observers appeared to encounter problems when classifying the animals into category 1 or 2. Nevertheless, the simple exclusion of category 1 is not recommended, since three categories provide a higher information value when low severity states are considered. In the study by Scott *et al* (2009), 79.0% of the sows only showed a withdrawal response when the approaching human reached close proximity and tried to touch the animal. An average of 15.1 to 23.0% of the sows withdrew when the observer tried to touch them in the test-retest study by Friedrich *et al* (2019a). The exclusion of category 1 would result in the non-detection of this slight fear response. Therefore, a re-definition of the categories potentially represents a better approach and would lead to a clearer definition and therefore better interobserver reliability. However, before modifications to the scoring system can be implemented, the role of observers in the human-animal relationship test — either performing the test or else watching from a distance — requires further attention. The low interobserver reliability identified in the human-animal relationship test might also be explained by the study design and may not refer to genuinely low interobserver reliability. Pigs and other animals are able to distinguish between different people (Held *et al* 2002). Furthermore, pigs and other mammals react to subtle, potentially subconscious changes in humans' body language (Candland 1993).

Assessment of stereotypies

In the present study, the assessment of stereotypies showed overall acceptable to good interobserver reliability between the observers in the evaluations conducted at both the individual animal and farm level. Only the measure 'tongue rolling' was assigned low agreement based on the agreement parameters, SDC and LoA, in the farm level evaluation. This difference is explained by the fact sows do not always stick their tongues out of their mouths when performing the tongue rolling behaviour. Tongue rolling can occasionally be less visible. Less obvious cases of tongue rolling were difficult to detect by individual observers and may have led to a low interobserver reliability for this

measure. Training should provide a clear explanation of this behaviour to observers. Otherwise, the definition would need to be changed, which would require an evaluation of the feasibility, validity and reliability. Furthermore, alternative methods to assess stereotypies, such as the approaches studied by Friedrich *et al* (2019c), may be implemented.

Measures of health and physical state

The welfare standards of the visited farms were good for most visits, which is why lower numbers of sows were assigned to category 2 for most measures. These results are consistent with those of Scott *et al* (2009). A first evaluation of the measures of health and physical state can be achieved, but the low variance must be considered when determining how they are to be interpreted. Measures of health and physical state with a prevalence less than 0.05% in the assessments of both observers were unable to be assessed because a low prevalence might lead to bias in assumptions about interobserver reliability and, subsequently, to a potentially inaccurate assessment of the interobserver reliability of these measures (Czycholl *et al* 2016a).

The interobserver reliability of measures with a prevalence greater than 0.05% for at least one of the observers was generally acceptable to good in the evaluations performed at individual animal and farm level. In both analyses, observers did not agree on the assessments of bursitis and shoulder sores in sows, as well as huddling in piglets. Poor interobserver reliability was also identified for panting in sows and lameness in piglets in the evaluation conducted at farm level. The assessment of those health and physical state measures with poor interobserver reliability was adversely affected when animals moved themselves, were soiled or when the barns were dark. These circumstances were also noted by Veissier *et al* (2013) as complications of the assessment of welfare measures. Modifications to the assessment of affected measures are discussed below. However, the scoring scales and scoring criteria of the affected measures are, to a degree, inappropriate since a certain effect is derived from the constraints described above. Thus, revision of scoring scales and criteria is advised.

The health and physical state measure addressing bursitis showed only poor interobserver reliability in the present study. Bursitis assessment was impeded when animals moved themselves or when their legs were soiled. According to Dalmau *et al* (2010), a clear view is obstructed when animals move as a group, which is particularly problematic in the gestation unit where sows are housed in groups. Poor interobserver reliability was also revealed for bursitis in the studies performed by Temple *et al* (2013) and Czycholl *et al* (2016a), who used the Welfare Quality® protocol for growing pigs. In contrast, Scott *et al* (2009) suggested good interobserver reliability for this measure. However, because the authors used the five-point scale of bursitis described by Lyons *et al* (1995), their results are not directly comparable to ours since the Welfare Quality® protocols use a three-point scale (Veissier *et al* 2013). Although bursitis category 2 showed better interobserver reliability, the definition of bursitis in the Welfare Quality® protocol for sows and piglets is not appropriate to assess comfort around resting in sows.

The performance of the assessment of shoulder sores in poorly illuminated barns led to difficulties in differentiating between category 0 (no lesions) and category 1 (evidence of scar tissue or a recent healing injury, or reddening of the area without penetration of the tissue), resulting in poor interobserver reliability for this measure of health and physical state. Detecting scar tissue is particularly difficult in an on-farm assessment. As with bursitis, category 2 of shoulder sores showed good interobserver reliability. Thus, categorisation into only two categories might lead to improved interobserver reliability. However, the presence of category 1 and how much information would be lost by excluding category 1 still requires to be evaluated.

An animal's movement also complicates the assessment of panting in sows. While the definition (breathing rapidly in short gasps, > 28 breaths per min) is relatively straightforward, the assessment of breathing frequency is complicated when animals move around.

In piglets, only lameness and huddling showed poor interobserver reliability. As all piglet health and physical state measures are assessed at the group level, moderately lame piglets are particularly difficult to detect. This limitation led to the low interobserver reliabilities of category 0 and 1 in the measure of lameness. Better interobserver reliability was detected for category 2, which is assigned when more than one moderately lame piglet or one severely lame piglet is observed. An assessment of individual piglets might be useful, which would require them to be separated. As the piglets are first scored at individual animal level and the group level score is only calculated from these results, this approach should not reduce the feasibility of the protocol. The difficulty of the on-farm assessment of huddling is attributed to the tendency for piglets to be disturbed very easily, thereby leaving their resting position. In contrast, acceptable to good interobserver reliability was found for the health and physical state measures, absence of manure on the body, coughing and sneezing in piglets, including measures of the principles of Good housing and Good health. Additional studies are needed to determine whether the measures of health and physical state with acceptable to good interobserver reliability are able to compensate for measures showing poor reliability or whether those measures should be revised or replaced. The principle of Good feeding in piglets is assessed by the management-based measure, age of weaning and the resource-based measure, water supply. For this principle, the focus is not on animal-based measures. Potential health and physical state measures to assess the nutritional supply of piglets include face lesions, carpal joint lesions or the presence of undersized animals. However, new measures must be evaluated in specific studies to determine their feasibility, validity and reliability.

Animal welfare implications and conclusion

This study aimed to test the interobserver reliability of measures of the Welfare Quality® animal welfare assessment protocol for sows and piglets in an on-farm study conducted at piglet-producing farms. The interpretation of the results is

https://doi.org/10.7120/09627286.29.3.323 Published online by Cambridge University Press

limited because the study was conducted using only two observers. This approach was preferred since on-farm assessment is considerably more time consuming than simply using pictures and videos. However, acceptable to good interobserver reliability was obtained for the QBA, the assessments of social and exploratory behaviours and stereotypies and most measures of health and physical state, but only questionable interobserver reliability was detected for the human-animal relationship test. Thus, the present study suggests re-categorisation of the human-animal relationship test categories. A better definition of the existing measures of health and physical state or having them replaced by measures with a higher interobserver reliability is needed for measures with low interobserver reliability. This revision is specifically recommended for bursitis as a measure of comfort around resting, which showed low interobserver reliability in previous studies of growing pigs. Continuous re-training and re-evaluation of observers is crucial to prevent observer drift and to maintain good interobserver reliability for all measures. As some health and physical state measures occurred with low prevalence rates or not at all, limited or no conclusions about their interobserver reliability could be drawn. Thus, studies with international participation are needed to evaluate the interobserver reliability of those measures and identify the significance of the measure in an on-farm assessment. In conclusion, the detected interobserver reliability of the measures of the Welfare Quality® protocol for sows and piglets indicate it to be a promising approach to objectively assess animal welfare in sows and piglets, once limitations identified in the present study have been addressed.

Acknowledgements

This study was financially supported by the HW Schaumann Foundation. LF, IC and JK conceived and designed the study. LF performed the data collection. LF and IC analysed and interpreted the data. NK and JK assisted with the interpretation of the results. LF wrote the manuscript and NK did the editing. All authors read and approved the final version of the manuscript. The authors declare no conflicts of interest. The funding agencies had no role in the design of the study; in the collection, analysis, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

AWIN 2015 *AWIN welfare assessment protocol for horses*. http://dx.doi.org/10.13130/AWIN_horses_2015

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

Blokhuis HJ 2008 International cooperation in animal welfare. The Welfare Quality® project. *Acta Agriculturae Scandinavica, Section A 50(S1)*: 10. https://doi.org/10.1186/1751-0147-50-S1-S10

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

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

Brennan RL and Prediger DJ 1981 Coefficient kappa: Some uses, misuses and alternatives. *Educational and Psychological Measurement 41*: 687-699. https://doi.org/10.1177/001316448104100307

Brenner H and Kliebsch U 1996 Dependence of weighted kappa coefficients on the number of categories. *Epidemiology 7*: 199-202. https://doi.org/10.1097/00001648-199603000-00016

Burn CC, Pritchard JC and Whay HR 2009 Reliability of a welfare assessment for working horses and donkeys in developing countries. *Animal Welfare 18*: 177-187

Byrt T, Bishop J and Carlin JB 1993 Bias, prevalence and kappa. *Journal of Clinical Epidemiology 46*: 423-429. https://doi.org/10.1016/0895-4356(93)90018-V

Candland DK 1993 *Feral Children and Clever Children: Reflections on Human Nature*. Oxford University Press: New York, USA

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

Cohen J 1968 Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin 70*: 213-220. https://doi.org/10.1037/h0026256

Czycholl I, Grosse Beilage E, Henning C and Krieter J 2017 Reliability of the qualitative behavior assessment as included in the Welfare Quality® assessment protocol for growing pigs. *Journal of Animal Science 95*: 3445-3454. https://doi.org/10.2527/jas.2017.1525

Czycholl I, Kniese C, Büttner K, Grosse Beilage E, Schrader L and Krieter J 2016a Interobserver reliability of the Welfare Quality® animal welfare assessment protocol for growing pigs. *SpringerPlus 5*: 1114. 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

Dalmau A, Geverink NA, van Nuffel A, van Steenbergen L, van Reenen K, Hautekiet V, Vermeulen K, Velarde A and Tuyttens FAM 2010 Repeatability of lameness, fear and slipping scores to assess animal welfare upon arrival in pig slaughterhouses. *Animal 4*: 804-809. https://doi.org/10.1017/S1751731110000066

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

de Vet HC, 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

Docking CM, van de Weerd HA, Day J and Edwards SA 2008 The influence of age on the use of potential enrichment objects and synchronisation of behaviour of pigs. *Applied Animal Behaviour Science 11*: 244-257. https://doi.org/10.1016/j.applanim.2007.05.004

Eurobarometer 2016 *Attitudes of Europeans towards Animal Welfare: Special Eurobarometer Report 442*. Eurobarometer: Brussels, Belgium

European Commission (EU) 2007 *Council Regulation (EC) No 834/2007 of 28 June 2007 on organic production and labelling of organic products and repealing Regulation (EEC) No 2092/91*. EC: Brussels, Belgium

Feinstein AR and Cicchetti DV 1990 High agreement but low kappa: I the problems of two paradoxes. *Journal of Clinical Epidemiology 43*: 543-549. https://doi.org/10.1016/0895-4356(90)90158-L

Forkman B and Keeling LJ 2009 *Welfare Quality® Reports: Assessment of animal welfare measures for sows, piglets and fattening pigs*. Cardiff University: Cardiff, UK

Friedrich L, Krieter J, Kemper N and Czycholl I 2019a Test-retest reliability of the Welfare Quality® animal welfare assessment protocol for sows and piglets, Part 1. Assessment of the welfare principle of Appropriate Behaviour'. *Animals 9*: 398. https://doi.org/10.3390/ani9070398

Friedrich L, Krieter J, Kemper N and Czycholl I 2019b Test–retest reliability of the Welfare Quality® Assessment protocol for pigs applied to sows and piglets, Part 2. Assessment of the principles good feeding, good housing, and good health. *Journal of Animal Science 97*: 1143-1157. https://doi.org/10.1093/jas/skz018

Friedrich L, Krieter J, Kemper N and Czycholl I 2019c Frothy saliva: A novel indicator to assess stereotypies in sows? *Applied Animal Behaviour Science 222*: 104897. https://doi.org/10.1016/j.applanim.2019.104897

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

German Animal Welfare Act 2006 *Tierschutzgesetz in der Fassung der Bekanntmachung vom 18 Mai 2006 (BGBl I S 1206, 1313), das zuletzt durch Artikel 141 des Gesetzes vom 29 März 2017 (BGBl I S 626) geändert worden ist*. https://www.gesetze-im-internet.de/tierschg/BJNR012770972.html

German Order for the Protection of Production Animals used for Farming Purposes and other Animals kept for the Production of Animal Products 2006 *Tierschutz-Nutztierhaltungsverordnung in der Fassung der Bekanntmachung vom 22 August 2006 (BGBl I S 2043), die durch Artikel 3 Absatz 2 des Gesetzes vom 30 Juni 2017 (BGBl I S 2147) geändert worden ist*. https://www.gesetze-im-internet.de/tierschnutztv/BJNR275800001.html

Geverink N, Meuleman M, van Nuffel A, van Steenbergen L, Hautekiet V and Vermeulen K 2009 Repeatability of a lameness score measured on farm. In: Forkman B and Keeling L (eds) *Welfare Quality® Reports: Assessment of Animal Welfare Measures for Sows, Piglets and Fattening Pigs* pp 73-78. Cardiff University Press: Cardiff, UK

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

Held S, Mendl M, Laughlin K and Byrne RW 2002 Cognition studies with pigs: Livestock cognition and its implication for production. *Journal of Animal Science 80*: E10-E17

Hewetson M, Christley TM, Hunt ID and Voute LC 2006 Investigations of the reliability of the observational gait analysis for the assessment of lameness in horses. *Veterinary Record 158*: 852-858. https://doi.org/10.1136/vr.158.25.852

Hoehler FK 2000 Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. *Journal of Clinical Epidemiology 53*: 499-503. https://doi.org/10.1016/S0895-4356(99)00174-2

Knierim U, Lentfer T, Staack M and Wemelsfelder F 2007 How reliable is a qualitative behavioural assessment of laying hens? *KTBL SCHRIFT 461*: 135

Landis JR and Koch GG 1977 An application of hierarchical Kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics 33*: 362-374. https://doi.org/10.2307/2529786

Lyons C, Bruce JM, Fowler VR and English PR 1995 A comparison of productivity and welfare of growing pigs in four intensive systems. *Livestock Production Science 43*: 265-274. https://doi.org/10.1016/0301-6226(95)00050-U

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

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

Mullan S, Edwards SA, Butterworth A, Whay HR and Main DCJ 2011 Inter-observer reliability testing of pig welfare outcome measures proposed for inclusion within farm assurance schemes. *The Veterinary Journal 190*: 100-109. https://doi.org/10.1016/j.tvjl.2011.01.012

Munsterhjelm C, Heinonen M and Valros A 2015 Application of the Welfare Quality® animal welfare assessment system in Finnish pig production, part II: Associations between animal-based and environmental measures of welfare. *Animal Welfare 24*: 161-172. https://doi.org/10.7120/09627286.24.2.161

Napolitano F, de Rosa G, Braghieri A, Grasso F, Bordi A and Wemelsfelder F 2008 The qualitative assessment of responsiveness to environmental challenge in horses and ponies. *Applied Animal Behaviour Science 109*: 342-354. https://doi.org/10.1016/j.applanim.2007.03.009

Phythian CJ, Cripps PJ, Michalopoulou E, Jones PH, Grove-White D, Clarkson MJ, Winter AC, Stubbings LA and Duncan JS 2012 Reliability of indicators of sheep welfare assessed by a group observation method. *The Veterinary Journal 193*: 257-263. https://doi.org/10.1016/j.tvjl.2011.12.006

Phythian CJ, Michalopoulou E, Cripps PJ, Duncan JS and Wemelsfelder F 2016 On-farm qualitative behaviour assessment in sheep: Repeated measurements across time, and association with physical indicators of flock health and welfare. *Applied Animal Behaviour Science 175*: 23-31. https://doi.org/10.1016/j.applanim.2015.11.013

Phythian CJ, Toft N, Cripps PJ, Michalopoulou E, Winter AC, Jones PH, Grove-White D and Duncan JS 2013a Inter-observer agreement, diagnostic sensitivity and specificity of animal-based indicators of young lamb welfare. *Animal 7*: 1182-1190. https://doi.org/10.1017/S1751731113000487

Plesch G, Broerkens N, Laister S, Winckler C and Knierim U 2010 Reliability and feasibility of selected measures concerning resting behaviour for the on-farm welfare assessment in dairy cows. *Applied Animal Behaviour Science 126*: 19-26. https://doi.org/10.1016/j.applanim.2010.05.003

https://doi.org/10.7120/09627286.29.3.323 Published online by Cambridge University Press

SAS® Institute Inc 2008 *User's Guide (release 9.4)*. SAS: Cary, USA

Scott K, Binnendijk GP, Edwards SA, Guy JH, Kiezebrink MC and Vermeer HM 2009 Preliminary evaluation of a prototype welfare monitoring system for sows and piglets (Welfare Quality® project). *Animal Welfare 18*: 441-449

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

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

Temple D, Manteca X, Velarde A and Dalmau A 2011 Assessment of animal welfare through behavioural parameters in Iberian pigs in intensive and extensive conditions. *Applied Animal Behaviour Science 131*: 29-39. https://doi.org/10.1016/j.applanim.2011.01.013

Thorslund CAH, Aaslyng MD and Lassen J 2017 Perceived importance and responsibility for market-driven pig welfare: Literature review. *Meat Science 125*: 37-45. https://doi.org/10.1016/j.meatsci.2016.11.008

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

van Loo EJ, Caputo V, Nayga RM and Verbeke W 2014 Consumers' valuation of sustainability labels on meat. *Food Policy 49*: 137-150. https://doi.org/10.1016/j.foodpol.2014.07.002

Veissier I, Winckler C, Velarde A, Butterworth A, Dalmau A and Keeling LJ 2013 Development of welfare measures and protocols for the collection of data on farms or at slaughter. In: Blokhuis H, Miele M, Veissier I and Jones B (eds) *Improving Farm Animal Welfare. Science and Society Working Together: The Welfare Quality Approach* pp 115–141. Wageningen Academic Publishers: Wageningen, The Netherlands. https://doi.org/10.3920/978-90-8686-770-7_6

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

Walker J, Dal A, Waran N, Clarke N, Farnworth M and Wemelsfelder F 2010 The assessment of emotional expression in dogs using a Free Choice Profiling methodology. *Animal Welfare 19*: 75-84

Walter SD, Eliasziw M and Donner A 1998 Sample size and optimal designs for reliability studies. *Statistics in Medicine 17*: 101-110. https://doi.org/10.1002/(SICI)1097-0258(19980115)17:1<101::AID-SIM727>3.0.CO;2-E

Webster J 2005 The assessment and implementation of animal welfare: theory into practice. *Scientific and Technical Review of the Office International des Epizooties 24*: 723-734. https://doi.org/10.20506/rst.24.2.1602

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

Wemelsfelder F, Hunter TEA, Mendl MT and Lawrence AB 2001 Assessing the 'whole animal'. A free choice profiling approach. *Animal Behaviour 62*: 209-220. https://doi.org/10.1006/anbe.2001.1741

Wemelsfelder F, Hunter EA, Mendl MT, Lawrence AB, Hunter E, Mendl M and Lawrence A 2000 The spontaneous qualitative assessment of behavioural expressions in pigs: first explorations of a novel methodology for integrative animal welfare measurement. *Applied Animal Behaviour Science 67*: 193-215. https://doi.org/10.1016/S0168-1591(99)00093-3

Wemelsfelder F and Millard F 2009 Qualitative behaviour assessment. In: Forkman B and Keeling LJ (eds) *Welfare Quality® Reports: Assessment of Animal Welfare Measures for Dairy Cattle, Beef Bulls and Veal Calves*. Cardiff University Press: Cardiff, UK

Wemelsfelder F, Nevison I and Lawrence AB 2009 The effect of perceived environmental background on qualitative assessments of pig behaviour. *Animal Behaviour 78*: 477-484. https://doi.org/10.1016/j.anbehav.2009.06.005

Windschnurer I, Schmied C, Boivin X and Waiblinger S 2008 Reliability and inter-test relationship of tests for on-farm assessment of dairy cows' relationship to humans. *Applied Animal Behaviour Science 114*: 37-53. https://doi.org/10.1016/j.applanim.2008.01.017

Wirtz M and Caspar F 2002 *Beurteilerübereinstimmung und Beurteilerreliabilität. Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen*. Hogrefe Verlag für Psychologie: Göttingen, Germany. [Title translation: Observer agreement and observer reliability. Methods for determining and improving the reliability of assessments using category systems and rating scales]