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

Animal Welfare 2016, 25: 447-459
ISSN 0962-7286
doi: 10.7120/09627286.25.4.447

Test-retest reliability of the Welfare Quality® animal welfare assessment protocol for growing pigs

I Czycholl*†, C Kniese§, K Büttner†, E Grosse Beilage‡, L Schrader§ and J Krieter†

† Institute of Animal Breeding and Husbandry, Christian-Albrechts-University, Olshausenstr 40, D-24098 Kiel, Germany
‡ Field Station for Epidemiology, University of Veterinary Medicine Hannover, Foundation, Buescheler Str 9, D-49456 Bakum, Germany
§ Institute of Animal Welfare and Animal Husbandry, Friedrich Loeffler Institut, Doernbergstr 25/27, D-29223 Celle, Germany
* Contact for correspondence and requests for reprints: iczycholl@tierzucht.uni-kiel.de

Abstract

The aim of this study was to assess the feasibility and test-retest reliability of the Welfare Quality® Animal Welfare Assessment Protocol for Growing Pigs. Twenty-three German pig farms were visited repeatedly by the same trained observers; each farm being visited six times during two fattening periods. The entire protocol assessment was carried out during each farm visit, ie a Qualitative Behaviour Assessment (QBA), behavioural observations (BO), a Human Animal Relationship test (HAR) and different individual parameters (IPs), eg bursitis and tail-biting. Test-retest reliability was evaluated by a Wilcoxon signed rank test (W) and by calculation of the Smallest Detectable Change (SDC) and Limits of Agreement (LoA). The QBA presented non-satisfactory agreement between farm visits. However, good agreement, in general, was found for the BO. For the HAR, no reliability could be detected. Most IPs were of acceptable agreement, with the exception of bursitis and manure on the body. Bursitis showed great differences, which can be explained by difficulties in the assessment when the animals moved around or their legs were dirty. The disagreement in the parameter manure on the body can be explained by seasonal effects. Disagreement was further found concerning the parameters coughing, sneezing, pleuritis, pneumonia and milkspots. Feasibility was good; both observers could be well-trained to fulfil the protocol. Furthermore, the time needed for an assessment did not exceed 6 h. The parts of the protocol that proved to be insufficiently reliable need to be addressed in the future in order to enhance and improve the objective measurement of animal welfare.

Keywords: *animal-based measure, animal welfare assessment, farm, pig, test-retest reliability, Welfare Quality®*

Introduction

The Welfare Quality® Animal Welfare Assessment protocols are considered feasible, valid and reliable measurement methods to determine animal welfare (Velarde 2007). These protocols are based on four main principles — Good feeding, Good housing, Good health and Appropriate behaviour, which is also the constitutional definition of animal welfare. In terms of a top-down process, these principles are divided into twelve subcriteria, which are then measured by a set of approximately 30 predominantly animal-based parameters to be estimated on-farm. After assessment of the parameters on-farm, the measures are usually expressed as percentages of affected animals. A dimensionless number between 0 and 100 is calculated from these percentages utilising different mathematical methods, eg decision trees (Magerman 1995) as well as I-Spline functions (Curry & Schoenberg 1966) and Choquet Integrals (Grabisch & Roubens 2000). This is carried out first at the sub-criteria and then at the principle level. Depending on the numbers obtained (the closer to 100 the better) the farms are scored and labelled as excellent, enhanced, acceptable or not classified (Welfare Quality® 2009).

The basic requirements of all such measurement tools are validity, feasibility and reliability. Validity characterises the amount a given measurement method actually assesses what it is supposed to measure and the relevance of that method. Feasibility describes good cost-effectiveness and applicability (Velarde 2007). Reliability refers to the repeatability of measures under consistent conditions (Carlson *et al* 2009). Reliability consists of inter-observer reliability, intra-observer reliability and test-retest reliability. Inter- and intra-observer reliability can be influenced by the training of the assessors, whereas test-retest reliability refers solely to the method used (Velarde 2007). Inter-observer reliability describes the need to be independent of the results produced by different observers (Wirtz & Caspar 2002). Intra-observer reliability is defined as the extent of agreement that the same observer reaches from the repeated assessment of video clips or pictures, ie the same objects under exactly the same conditions (Martin & Bateson 2007). Due to minor changes, this tends not to be the case when assessing farms or animals repeatedly as, typically, there are other individuals on the farms. However, if the same individuals are being assessed, there may well be individual

changes, such as weight gain or loss or changes in pregnancy status. Therefore, intra-observer reliability in the case of welfare assessment is often thought of as being part of the test-retest reliability (Temple *et al* 2012a), due to the fact that it cannot be assessed under field conditions. For a good test-retest reliability, the method tested should basically attain the same results when the same object is observed despite minor changes (Windschnurer *et al* 2009). However, major changes should be detected by the measurement tool. It becomes obvious that consistency over time is a basic requirement for reliable and feasible measuring.

Prior to inclusion in the Welfare Quality® protocol, most parameters were tested for their feasibility, validity and reliability in pre-studies (Forkmann & Keeling 2009). However, due to the fact that these protocols are relatively new and under continuous improvement and revision, studies on the feasibility, validity and reliability of the entire protocols in their on-farm use are rare. Therefore, the aim of the present study was to analyse the test-retest reliability of the Welfare Quality® Animal Welfare Assessment Protocol for Growing Pigs and, thus, is an important contribution towards the evaluation of reliability and feasibility of the entire protocol. To the authors' knowledge, the only study emphasising the test-retest reliability of the animal-based measures included in the Welfare Quality® Animal Welfare Assessment Protocol for Growing Pigs is that of Temple *et al* (2013), who compared the results of protocol assessments carried out on 15 Spanish, intensive, conventional farms at an interval of approximately 12 months. The present study contributes greatly to the present knowledge since 23 German growing pig farms were visited six times at an interval of less than six months. According to Knierim and Winckler (2009), welfare assessments for certification purposes supposedly need to take place over a period of six months or more due to feasibility. Therefore, results still need to be reliable within this time-frame and should not be sensitive to slight environmental changes. However, they still need to be representative of the situation on the farm (Winckler *et al* 2007). Moreover, as the Welfare Quality® Protocols aim at being valid internationally and since previous studies have revealed a country effect (Temple *et al* 2011c), it is of great importance to study the situation in different countries. Furthermore, in the study by Temple *et al* (2013), some parameters had prevalences close to 0 and low variability in the two repeated visits, which hampered their estimation of reliability. Thus, further studies are needed to evaluate the reliability and importance of these parameters.

Materials and methods

Data collection

Data collection took place between January 2013 and January 2014 on 23 German growing pig farms in Lower Saxony and Schleswig Holstein courtesy of two observers, one visiting the farms in Lower Saxony repeatedly and the other those in Schleswig Holstein repeatedly. The pigs on the farms were housed either conventionally or according to the guidelines of the German animal welfare label 'Tierwohllabel' of the German animal welfare organisation 'Deutscher Tierschutzbund eV' (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 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.

Each of these farms was visited six times by the same observer during two consecutive fattening periods. During each, three assessments took place: the first protocol assessment two weeks after the pigs had entered into the growing barn, at an approximate average weight of 40 kg (farm visit 1, 13 [± 2] weeks of age), the second in the middle of the fattening period at an average weight of 75 kg (farm visit 2, 18 [± 2] weeks of age) and the third assessment two weeks prior to the beginning of sales to the slaughterhouse, at an average weight of 100 kg (farm visit 3, 23 [± 2] weeks of age). While data were collected, no major changes in management on the farms occurred.

The observers had been trained officially by members of the Welfare Quality® project group and reached good agreement during training, whereas training sessions on each parameter were carried out until at least 80% of the observers reached a consensus. Observer agreement was further tested twice during the data collection period by the evaluation of video sequences and pictures. At all times, more than 85% of the pictures and videos were evaluated alike and therefore good agreement attained.

Protocol assessments

The Welfare Quality® Animal Welfare Assessment Protocol for Growing Pigs consists of four different parts: a Qualitative Behaviour Assessment (QBA), behavioural observations (BO), a Human Animal Relationship test (HAR) and the assessment of different individual parameters (IP), which are described in detail below. At each farm visit, an assessment was carried out using the entire Welfare Quality® protocol.

Qualitative Behaviour Assessment (QBA)

In the protocol, the QBA was included as the animal-based measure for the evaluation of positive emotions. This is a free behavioural observation method, meaning that the observer looks at the animals without any restrictions. It was carried out at four to six observation points on the farm for a total surveillance time of 20 min. The number of pigs under surveillance ranged from 80 to 240 animals. A visual analogue scale of 125 mm was assigned to each of 20 given adjectives: 1) active; 2) relaxed; 3) fearful; 4) agitated; 5) calm; 6) content; 7) tense; 8) enjoying; 9) frustrated; 10) bored; 11) playful; 12) positively occupied; 13) listless; 14) lively; 15) indifferent; 16) irritable; 17) aimless; 18) happy; 19) distressed; and 20) sociable). A mark was set as to whether the observer found the term to be absent (0 mm) or dominant (125 mm) for the animals under study. After the

https://doi.org/10.7120/09627286.25.4.447 Published online by Cambridge University Press

assessment on-farm, the length (mm) on the visual analogue scale was measured with a ruler for each of the adjectives. Thus, for each farm visit in our study, one score in mm for each adjective was recorded by each observer.

Behavioural observations (BO)

After the QBA, BO (assessments) in the form of instantaneous scan sampling were performed at three other viewpoints, thus on a total number of pigs ranging from 120 to 180. First, the pigs in the pens under surveillance were chased up and then had 5 min time to calm down. During this time, their coughing and sneezing was counted. The animals were then scanned for a total time of 10 min at each viewpoint. A scan was made every 2 min and the pigs sorted into the categories: positive social behaviour, negative social behaviour, pen investigation, use of enrichment material, other active behaviour or resting.

The results of the BO were expressed as performed behaviour as a percentage of the total active behaviour. Total active behaviour constituted all possible behaviours except resting. Thereby, positive and negative social behaviour were expressed together since total social behaviour and negative social behaviour was also presented individually.

Human Animal Relationship test (HAR)

In the following protocol assessment, ten randomly chosen pens were entered and, initially, the reaction of the animals towards the observer was evaluated using the HAR. The number of pigs under surveillance ranged from 100 to 400. Therefore, after entering the pen and walking around it in one direction, the observer stood still in the middle of the pen for 30 s. Subsequently, he walked around the pen in the other direction and analysed the reaction of the animals, ie whether they fled or showed a panic response. For the HAR, the percentage of pens with a panic response from the total observed pens per farm was taken into account for further analysis.

Individual parameters (IP)

Subsequently, 100 to 150 of the pigs in these pens were scored for a variety of IP, eg wounds, manure on the body, tail-biting and bursitis, whereby only one side of the pigs was considered. The IPs were either scored using a three-point scale (0 = absent, 1 = light impairment, 2 = strong impairment) or else a two-point scale (0 = absent, 2 = present). The complete list of parameters, their definitions and the slotting criteria are presented in Table 1. The mortality rate and the percentages of animals affected by pneumonia, pleurisy, ascites and pericarditis as registered by the slaughterhouse were requested from the farmer as well as whether management procedures such as tail-docking and castration had been carried out.

The IPs were analysed as percentages of animals sorted into the corresponding category (eg bursitis category 0: 50%, bursitis category 1: 40%, bursitis category 2: 10%). In the comparison of the single categories of a parameter (eg bursitis category 0, bursitis category 1 and bursitis category 2) each category was treated as a single variable, ie the agreement of bursitis category 0 between the farm visits was compared independently of the agreement of the categorisation into bursitis category 1 or 2.

Ethical statement

The authors declare that the experiments were carried out in strict accordance with international animal welfare guidelines. The institution to which the authors are affiliated does not have research ethic committees or review boards (in consultation with the animal welfare officer of the Christian-Albrechts-University, Kiel, Germany). Therefore, the study applied the following: 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.

Evaluation of feasibility

Feasibility was defined according to Velarde (2007), Temple *et al* (2011a) and Blokhuis *et al* (2013), ie the overall protocol should require little input from the farmers, should be easy to perform in practical conditions without expert knowledge on the part of the observer and should not be time-consuming.

Statistics and agreement parameters

Results were compared at parameter level without further aggregation to sub-criteria or principle scores. Agreement at criterion or principle score would be untrustworthy without acceptable agreement at the fundamental level of assessment. Furthermore, all results are expressed at total farm level, which is reasonable since the samples of animals are taken randomly to give an overview of the assessed farm (Welfare Quality® 2009). In the case of the BO, this means that independently of the three viewpoints, all behavioural scans made during that farm visit were considered for the calculation of percentages. In the case of the HAR, the percentage of pens reacting with a panic response out of the ten pens evaluated was calculated. For the IP, the percentage of all evaluated animals on that farm was considered. The values of the recorded parameters in percent, respectively, in mm of each farm visit were then compared and evaluated for their test-retest reliability. Thereby, a comparison of the visits at same age classes of the two fattening periods was carried out, ie farm visit 1 of the first fattening period to farm visit 1 of the second fattening period, farm visit 2 of the first fattening period to farm visit 2 of the second fattening period and farm visit 3 of the first fattening period to farm visit 3 of the second fattening period. This was done to take into account that the age of the animals had been considered an influencing factor on many measurements in previous studies (Temple *et al* 2012b, 2013).

For statistical analyses, different agreement parameters were calculated either using the statistic programme SAS 9.2 (SAS Institute 2008) or R (Version 2.11.1) (SAS Institute 2008). A Wilcoxon signed rank test (W) was carried out with the procedure Proc npar1way in SAS 9.2 (SAS Institute 2008) to test for significant differences between the compared farm visits. The agreement parameters Smallest

Table 1 Quantitative animal-based measures with scoring scale and definition (Welfare Quality® 2009).

Animal-based measure	Score	Definition
Body condition	0	Good body condition
	2	Thin: visible spine, hip, pin bones
Bursitis	0	No evidence of bursae/swelling
	1	One/several small bursae on the same leg or one large bursa
	2	Several large bursae on the same leg or one extremely large or eroded bursa
Manure on the body	0	< 20% of body surface soiled with faeces
	1	20–50% of body surface soiled with faeces
	2	> 50% of body surface soiled with faeces
Huddling	0	Pig lying with < 50% of its body on top of another pig
	2	Pig lying with > 50% of its body on top of another pig
Shivering	0	No vibration of any body part
	2	Vibration of any body part
Panting	0	Normal breathing
	2	Rapid breath in short gasp
Wounds	0	< 4 lesions on all zones of the body
	1	4–10 lesions on one or more zones on the body
	2	≥ 10 lesions on two zones or one zone > 15 lesions
Tail-biting	0	No evidence of tail-biting
	2	Evidence of tail-biting
Lameness	0	Normal gait or slight lameness or abnormality in gait
	1	Severely lame, weight-bearing on affected limb
	2	No weight-bearing on one limb or unable to walk
Pumping	0	No evidence of laboured breathing
	2	Laboured breathing
Scouring	0	No liquid manure visible in pen
	1	Some liquid manure in some areas of pen
	2	All faeces visible inside pen are liquid
Skin condition	0	All skin of normal colour and texture
	1	0–10% has abnormal colour or texture
	2	> 10% of skin has abnormal colour or texture
Hernia	0	No hernia/rupture
	1	Small hernia/rupture
	2	Hernia/rupture touching the floor or with bleeding lesion
Twisted snout	0	No evidence of twisted snout
	2	Evidence of a twisted snout
Rectal prolapse	0	No evidence of rectal prolapse
	2	Evidence of rectal prolapse
Coughing	n	Number of coughs per observation point
Sneezing	n	Number of sneezes per observation
Human Animal Relation	0	≤ 60% showing a panic response
	2	> 60% of the pigs fleeing, facing away or huddled in corner of pen
Negative social behaviour	%	Aggressive behaviour or any behaviour with a response from the disturbed animal or any tail in mouth behaviour
Positive social behaviour	%	Sniffing, nosing, licking and moving gently away from the animal without an aggressive or flight reaction from this individual
Pen investigation	%	Sniffing, nosing, licking all features of pen
Use of enrichment material	%	Exploration towards straw and other suitable enrichment material
Resting	%	Non-active behaviour, animals are lying and not performing anything else

https://doi.org/10.7120/09627286.25.4.447 Published online by Cambridge University Press

Detectable Change (SDC) and Limits of Agreement (LoA) were calculated with the IRR package (Gamer *et al* 2012) for R (Version 2.11.1) (Venables & Smith 2010).

Wilcoxon signed rank test (W)

W is a non-parametric paired difference test. As in the present study, carried out with the help of the *W*, testing for statistical significance is a useful tool for reliability assessment (Gelman & Stern 2006) and often used in animal welfare studies (Temple *et al* 2011b, 2013). *W* assesses whether the ranks of the population means differ (Koehler *et al* 1996). In the given study, significant differences in compared farm visits were interpreted as unacceptable agreement as this is a clear indicator of disagreement. However, the reverse conclusion that non-significant differences represent good agreement cannot be automatically drawn as non-significant differences do not necessarily indicate reliable agreement. As significant differences indicate that there is insufficient agreement and as no additional information would be gained by the presentation of *P*-values, only significance or its absence are presented here. A *P*-value ≤ 0.05 was considered as significant.

Agreement parameters

SDC and LoA are both an expression of the measurement error σ^2(error), which is achieved from the simple one-way model according to Shrout and Fleiss (1979):

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

with x_{ijk} being the measured value, μ the general average value, α_i the random effect of the difference between the 23 farms and ε_{ijk} as the general error term.

SDC was calculated according to de Vet *et al* (2006) using the formula:

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

It indicates the smallest change in the score that can be detected with the method above the measurement error (Donoghue & Stokes 2009). The measurement unit of the SDC was in accordance with the measurement unit of the parameters under surveillance. Thus, in the present case, it is expressed in percent. Based on the interpretation of the simple agreement coefficient in de Vet *et al* (2006), an SDC less than or equal to 0.1 was interpreted as acceptable agreement.

The LoA, which was first introduced by Bland and Altman (1986), was also calculated according to de Vet *et al* (2006) by the formula:

$LoA = \text{mean of the differences} \pm 1.96 \times (\sqrt{2} \times \sigma^2[error])$.

The LoA calculates the range of the difference between two sets of measures and in this study was expressed as relative frequency between –1 and 1. The direction of –1 would be differences according to higher values obtained at the farm visit during the second fattening period and the direction of 1 would be due to higher values achieved in the first fattening period. Again, interpretation was based on the simple agreement coefficient of de Vet *et al* (2006) and, thus, an interval between –0.1 to 0.1 was interpreted as acceptable agreement. With the Bland and Altman (1986) plot of the LoA, namely the plot of difference between the means of two measurements against the average prevalence helps to determine the range of errors (de Vet 2005).

Interpretation of W, SDC and LoA

As non-significant differences are not automatically to be equated with agreement (Wirtz & Caspar 2002), acceptable agreement was only assigned if the *W* produced results in accordance with the agreement parameters. Furthermore, for acceptable test-retest reliability, it was expected that *W*, SDC and LoA would be acceptable in all three comparisons of the two fattening periods (farm visit 1, farm visit 2 and farm visit 3).

Results

The average time for a complete protocol assessment was 315 min (5 h and 15 min) with a range of 270 to 360 min. The QBA took a fixed time of 20 min plus the time to walk between 4–6 observation points. The same can be said for the time of the BO, which was fixed at 10 min per observation point plus the initial time between chasing the animals and allowing them to calm down for 5 min, during which coughing and sneezing was assessed. This made a total time of 45 min plus the time for the distance between three observation points. The HAR took about 90 s per pen. The observation time for the IP in the pens varied widely, depending on the dimensions of the pen, the number of animals and their reaction towards the observer, which influenced how fast the animals could be assessed. In our study, the time varied between 10–20 min per pen. Additional time of about 90 s per pen was needed to measure the width and length of the pen and to test the drinkers. Again, there was additional time for the distances between the ten pens, which depended on the layout of the farm. Moreover, especially during the first farm visit, time was needed for a first meeting with the farmer for an interview about general information of management procedures and layout of the farm which took, on average, 25 min with a range of 15–60 min.

Qualitative Behaviour Assessment (QBA)

The average values and the associated standard deviations obtained by each observer are presented in Table 2(a) and the results of the significance testing with the *W* and the corresponding agreement parameters are shown in Table 2(b). None of the adjectives presented agreement in any of the compared farm visits. Even if the comparison of the third farm visit showed non-significant differences, such as for the term active, SDC and LoA indicated non-satisfactory agreement.

Behavioural observations (BO)

On average, similar percentages of animals were sorted into dedicated behavioural categories (Figure 1) during the farm visits. This agreement could also be obtained using the calculation of the agreement parameters for each compared farm visit (Table 3), which achieved acceptable to good values for all behavioural categories, with the exception of the category pen investigation. Although no significant differences were obtained in all three comparisons, the agreement parameters did not suggest acceptable agreement for this category.

Human Animal Relationship test (HAR)

For the comparison of farm visit 1 during the first fattening period, an average of 17.8 (± 19.1)% of the pens on the farms showed a panic response compared to 23.8 (± 17.6)%

Table 2 Mean values (mm) (a) and presentation of significance testing with the Wilcoxon signed rank test (*W*) and agreement parameters Smallest Detectable Change (SDC) and Limits of Agreement (LoA) (b) of farm visits 1, 2 and 3 of the two fattening periods for the adjectives of the Qualitative Behaviour Assessment (QBA).

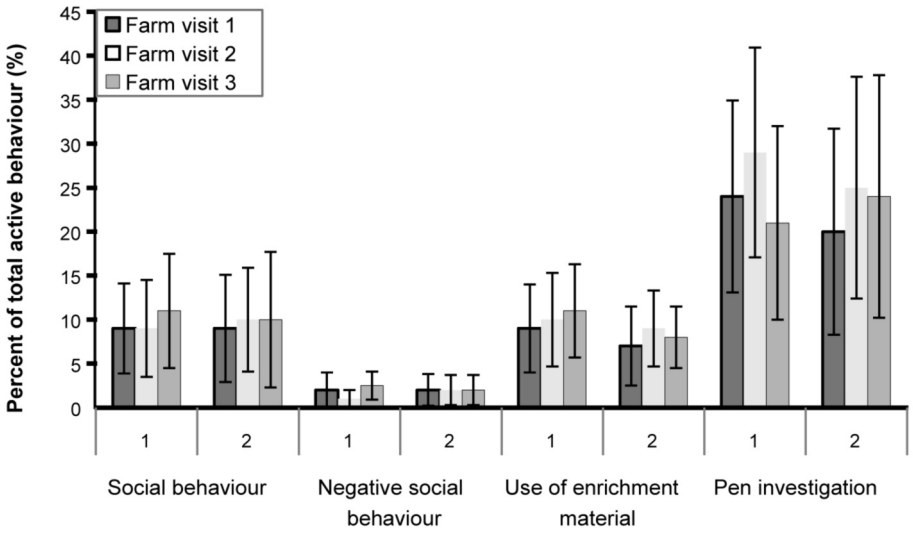
(a)	Fattening period 1			Fattening period 2		
Adjective	**1**	**2**	**3**	**1**	**2**	**3**
Active	69.9 (± 31.0)	71.8 (± 17.4)	64.7 (± 17.1)	73.7 (± 11.2)	82.3 (± 11.6)	66.5 (± 14.1)
Relaxed	58.1 (± 31.6)	60.8 (± 22.4)	58.2 (± 22.0)	54.9 (± 16.7)	49.4 (± 20.3)	54.6 (± 17.2)
Fearful	14.5 (± 5.2)	10.9 (± 4.2)	4.3 (± 6.8)	10.2 (± 5.4)	10.8 (± 4.9)	11.6 (± 2.0)
Agitated	19.5 (± 13.0)	22.5 (± 15.8)	15.5 (± 12.9)	20.8 (± 13.8)	31.8 (± 13.4)	29.8 (± 8.1)
Calm	51.5 (± 23.8)	53.6 (± 19.1)	57.3 (± 20.5)	56.3 (± 17.1)	44.7 (± 19.7)	53.7 (± 16.8)
Content	61.5 (± 18.1)	60.5 (± 15.6)	60.5 (± 13.3)	56.2 (± 7.1)	55.6 (± 14.3)	57.4 (± 7.6)
Tense	22.2 (± 9.8)	23.4 (± 9.1)	19.9 (± 9.0)	17.7 (± 10.7)	25.4 (± 9.0)	26.4 (± 6.8)
Enjoying	37.2 (± 16.0)	30.9 (± 14.7)	31.8 (± 13.3)	37.5 (± 14.8)	33.4 (± 12.2)	37.0 (± 11.6)
Frustrated	23.5 (± 2.5)	21.2 (± 5.5)	20.3 (± 2.7)	10.4 (± 2.0)	14.4 (± 0.8)	19.7 (± 0.8)
Bored	25.9 (± 11.7)	28.7 (± 3.6)	24.7 (± 5.4)	16.5 (± 4.1)	11.2 (± 4.0)	15.2 (± 4.2)
Playful	34.1 (± 18.0)	37.0 (± 15.0)	31.3 (± 13.3)	30.3 (± 15.6)	22.7 (± 9.9)	22.1 (± 10.4)
Positively occupied	55.8 (± 21.4)	51.5 (± 23.7)	51.7 (± 16.2)	53.9 (± 13.1)	61.1 (± 13.1)	59.4 (± 6.9)
Listless	22.5 (± 4.7)	23.4 (± 5.3)	21.9 (± 6.5)	11.3 (± 4.6)	12.8 (± 1.9)	35.1 (± 2.8)
Lively	51.3 (± 17.0)	46.8 (± 16.5)	31.3 (± 15.6)	50.6 (± 16.4)	55.6 (± 14.8)	54.6 (± 13.8)
Indifferent	21.2 (± 16.6)	19.1 (± 6.2)	22.6 (± 6.6)	17.5 (± 5.1)	14.2 (± 4.6)	22.6 (± 7.5)
Irritable	20.1 (± 2.8)	16.7 (± 1.9)	16.8 (± 2.3)	11.5 (± 1.7)	20.7 (± 0.7)	26.6 (± 0.6)
Aimless	26.8 (± 3.4)	28.7 (± 0.6)	27.3 (± 1.3)	24.2 (± 2.2)	18.1 (± 0.5)	26.5 (± 0.5)
Happy	52.3 (± 7.6)	50.4 (± 10.4)	50.6 (± 1.6)	54.8 (± 2.4)	35.1 (± 2.0)	52.7 (± 2.8)
Distressed	4.9 (± 2.4)	5.1 (± 4.0)	3.0 (± 1.4)	3.1 (± 2.1)	3.5 (± 0.6)	6.1 (± 0.5)
Sociable	69.4 (± 21.2)	61.6 (± 20.0)	55.7 (± 18.7)	61.1 (± 12.3)	63.8 (± 14.1)	59.1 (± 16.8)

(b)	*W**			SDC			LoA		
Adjective	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
Active	s	s	ns	0.46	0.34	0.37	–0.50 to 0.46	–0.34 to 0.37	–0.39 to 0.38
Relaxed	ns	s	s	0.55	0.54	0.49	–0.46 to 0.63	–0.46 to 0.63	–0.47 to 0.52
Fearful	s	ns	s	0.17	0.21	0.29	–0.16 to 0.18	–0.20 to 0.23	–0.29 to 0.33
Agitated	ns	s	s	0.35	0.40	0.41	–0.38 to 0.36	–0.46 to 0.40	–0.47 to 0.39
Calm	s	s	s	0.36	0.41	0.51	–0.37 to 0.41	–0.35 to 0.49	–0.49 to 0.54
Content	s	s	s	0.34	0.33	0.24	–0.24 to 0.42	–0.28 to 0.39	–0.20 to 0.26
Tense	s	s	s	0.36	0.39	0.19	–0.38 to 0.38	–0.39 to 0.42	–0.20 to 0.17
Enjoying	ns	s	s	0.25	0.28	0.25	–0.25 to 0.25	–0.30 to 0.28	–0.30 to 0.20
Frustrated	s	s	ns	0.33	0.22	0.24	–0.25 to 0.38	–0.13 to 0.26	–0.22 to 0.26
Bored	s	s	s	0.36	0.36	0.25	–0.23 to 0.42	–0.22 to 0.42	–0.22 to 0.28
Playful	s	s	s	0.36	0.25	0.22	–0.38 to 0.38	–0.27 to 0.24	–0.26 to 0.18
Positively occupied	ns	s	s	0.35	0.41	0.30	–0.39 to 0.35	–0.48 to 0.38	–0.40 to 0.26
Listless	s	s	s	0.31	0.18	0.23	–0.22 to 0.37	–0.11 to 0.21	–0.22 to 0.25
Lively	ns	s	s	0.44	0.39	0.47	–0.49 to 0.46	–0.45 to 0.40	–0.53 to 0.45
Indifferent	s	s	ns	0.44	0.24	0.23	–0.37 to 0.50	–0.20 to 0.26	–0.22 to 0.27
Irritable	s	s	s	0.30	0.22	0.30	–0.28 to 0.37	–0.22 to 0.23	–0.33 to 0.29
Aimless	ns	s	ns	0.35	0.28	0.26	–0.31 to 0.40	–0.23 to 0.33	–0.26 to 0.26
Happy	s	s	ns	0.22	0.17	0.15	–0.24 to 0.22	–0.19 to 0.12	–0.17 to 0.13
Distressed	ns	ns	ns	0.11	0.09	0.11	–0.10 to 0.11	–0.09 to 0.11	–0.12 to 0.10
Sociable	s	ns	s	0.41	0.36	0.28	–0.36 to 0.49	–0.38 to 0.37	–0.28 to 0.32

* Significant differences ($P < 0.05$); s: non-significant differences: ns

https://doi.org/10.7120/09627286.25.4.447 Published online by Cambridge University Press

Figure 1



Behavioural observations (BO): mean percentages of animals sorted into dedicated behavioural categories expressed as a percentage of the total active behaviour in the comparison of two fattening periods (1 and 2) for each of three farm visits (1, 2 and 3).

Table 3 Behavioural observations (BO): results of the Wilcoxon signed rank test (*W*; non-significant: ns; significant: s) and the calculation of the agreement parameters Smallest Detectable Change (SDC) and Limits of Agreement (LoA) in the comparison of two fattening periods with each three farm visits (1, 2 and 3).

Behavioural category	*W*			SDC			LoA		
	1	2	3	1	2	3	1	2	3
Social behaviour	ns	ns	ns	0.08	0.07	0.09	–0.09 to 0.03	–0.07 to 0.04	–0.09 to 0.07
Negative social behaviour	ns	ns	ns	0.02	0.02	0.02	–0.01 to 0.02	–0.01 to 0.02	–0.01 to 0.01
Use of enrichment material	ns	ns	ns	0.07	0.08	0.04	–0.04 to 0.04	–0.07 to 0.07	–0.01 to 0.03
Pen investigation	ns	ns	ns	0.14	0.14	0.12	–0.14 to 0.14	–0.14 to 0.14	–0.13 to 0.11

in the second fattening period. During farm visit 2, panic was seen in 12.1 (± 15.1)% of the pens in fattening period 1 and 14.8 (± 28.6)% of the pens in fattening period 2. For the oldest animals during farm visit 3, the panic reaction decreased to a prevalence of 8.4 (± 14.7)% in the first and 14.8 (± 23.6)% in the second fattening period. This trend of disagreement could also be documented by the agreement parameters, though the differences were tested to be non-significant (farm visit 1: *W*: ns, SDC: 0.48, LoA: –0.52 to 0.46; farm visit 2: *W*: ns; SDC: 0.63 LoA: –0.62 to 0.60; farm visit 3: *W*: ns, SDC: 0.40, LoA: –0.48 to 0.40).

Individual parameters (IP)

The parameters poor body condition, panting, twisted snout and rectal prolapse did not occur at all and the parameters huddling, lameness category 2, shivering, scouring, skin condition category 2 and hernia category 2 were observed only with a prevalence of less than 0.1%. An assumption about the reliability of these parameters was assumed to be meaningless due to their low prevalence and, therefore, reliability parameters were not calculated.

The remaining parameters, however, were recorded with a prevalence of, on average, greater than 0.5%. The mean prevalence of the parameters during the farm visits at different average weights of the pigs are presented in Table 4(a) and the corresponding results of *W*, SDC and LoA for these parameters can be found in Table 4(b). Acceptable to good agreement was detected for most of the parameters. The Bland and Altman (1986) plot of the LoA for hernia category 1 is shown in Figure 2 as an example. Here, the good agreement becomes visible, since the differences between all compared farm visits (each represented by one circle) are values within a small range around 0. However, *W*, SDC and LoA for manure on the body and bursitis of all categories indicated non-satisfactory agreement. This is visualised by the Bland and Altman (1986) plot for bursitis in Figure 3(a) and (b). Bursitis of category 2 proved to be more reliable than category 1 as can be seen by the smaller range of the differences. The same could be said for the category 2 of manure, which was of better agreement than category 1 but still unacceptable. While wounds of category 2 showed good agreement, wounds of category 1 did not present acceptable

Table 4 Mean (± SD) prevalence (%) (a) and results of a Wilcoxon signed rank test (*W*), Smallest Detectable Change (SDC) and Limits of Agreement (LoA) (b) of individual parameters (IPs) that occurred on average with a prevalence of greater than 0.5% in two fattening periods with each of the three farm visits (1, 2 and 3).

(a) Individual parameter	Category	Mean (± SD) prevalence (%) in fattening period 1			Mean prevalence (± SD) (%) in fattening period 2		
		1	2	3	1	2	3
Bursitis	0	67.4 (± 16.3)	50.5 (± 22.3)	43.1 (± 21.4)	52.2 (± 20.5)	47.9 (± 20.7)	43.6 (± 18.5)
	1	30.0 (± 16.1)	46.2 (± 19.9)	50.5 (± 17.1)	45.6 (± 18.3)	46.1 (± 18.4)	55.2 (± 16.0)
	2	0.7 (± 3.5)	3.1 (± 2.5)	6.4 (± 5.6)	2.2 (± 1.8)	6.0 (± 5.1)	5.6 (± 4.7)
Manure	0	93.4 (± 14.3)	89.1 (± 23.5)	80.4 (± 29.5)	77.0 (± 26.6)	74.2 (± 23.7)	67.7 (± 29.3)
	1	5.6 (± 4.5)	8.9 (± 8.4)	14.2 (± 11.1)	27.2 (± 18.9)	20.6 (± 18.2)	24.7 (± 23.7)
	2	0.6 (± 4.1)	1.8 (± 0.7)	3.7 (± 3.1)	3.5 (± 3.3)	5.2 (± 4.0)	7.5 (± 7.3)
Lame	0	99.3 (± 1.7)	99.2 (± 6.5)	99.2 (± 3.0)	99.7 (± 1.7)	99.6 (± 2.3)	99.2 (± 3.6)
	1	0.5 (± 0.4)	0.5 (± 0.3)	0.7 (± 0.6)	0.3 (± 0.1)	0.4 (± 0.3)	0.6 (± 0.3)
Wounds	0	90.6 (± 11.8)	93.7 (± 11.4)	94.0 (± 10.1)	91.5 (± 8.7)	93.3 (± 10.3)	91.9 (± 12.5)
	1	7.6 (± 7.6)	5.5 (± 3.9)	4.8 (± 3.6)	6.7 (± 6.2)	6.3 (± 4.8)	6.9 (± 6.3)
	2	1.6 (± 1.5)	0.5 (± 0.4)	1.1 (± 0.8)	1.0 (± 0.3)	0.4 (± 0.3)	1.1 (± 0.8)
Tail-biting	0	98.2 (± 9.6)	97.7 (± 3.4)	98.2 (± 2.7)	98.4 (± 6.0)	97.2 (± 6.6)	96.7 (± 11.5)
	2	1.8 (± 1.6)	2.1 (± 1.6)	1.7 (± 1.1)	1.6 (± 0.7)	2.8 (± 2.0)	2.8 (± 1.5)
Skin condition	0	99.1 (± 8.3)	99.2 (± 6.6)	99.5 (± 1.5)	99.5 (± 2.8)	99.3 (± 3.9)	99.6 (± 0.4)
	1	0.8 (± 0.2)	0.7 (± 0.7)	0.5 (± 0.1)	0.5 (± 0.3)	0.6 (± 0.3)	0.4 (± 0.4)
Hernia	0	99.5 (± 2.3)	99.3 (± 6.7)	99.6 (± 2.6)	99.4 (± 2.5)	99.4 (± 2.9)	99.5 (± 2.5)
	1	0.5 (± 0.1)	0.6 (± 0.1)	0.4 (± 0.3)	0.5 (± 0.5)	0.6 (± 0.3)	0.5 (± 0.5)
Coughing	n	1.8 (± 1.0)	2.9 (± 1.5)	1.8 (± 1.8)	0.8 (± 0.2)	1.1 (± 1.0)	2.2 (± 2.1)
Sneezing	n	1.1 (± 1.1)	0.7 (± 0.3)	0.6 (± 0.2)	1.1 (± 0.7)	0.5 (± 0.3)	0.5 (± 0.4)
Pneumonia*	%		9.8 (± 9.1)			8.2 (± 8.4)	
Pericarditis*	%		1.7 (± 0.9)			2.1 (± 1.8)	
Pleuritis*	%		3.9 (± 2.9)			3.9 (± 5.5)	
Milkspots*	%		7.2 (± 2.1)			4.7 (± 2.2)	

(b) Individual parameter	Category	*W*†			SDC			LoA		
		1	2	3	1	2	3	1	2	3
Bursitis	0	s	s	s	0.42	0.47	0.37	–0.17 to 0.50	–0.45 to 0.52	–0.42 to 0.38
	1	s	s	s	0.39	0.39	0.24	–0.49 to 0.13	–0.46 to 0.40	–0.24 to 0.26
	2	s	s	ns	0.17	0.17	0.19	–0.09 to 0.03	–0.20 to 0.12	–0.15 to 0.16
Manure	0	s	s	ns	0.40	0.39	0.52	–0.24 to 0.49	–0.21 to 0.49	–0.48 to 0.56
	1	s	s	ns	0.27	0.32	0.36	–0.42 to 0.12	–0.42 to 0.17	–0.45 to 0.32
	2	ns	ns	ns	0.19	0.13	0.20	–0.18 to 0.12	–0.13 to 0.07	–0.22 to 0.21
Lame	0	ns	ns	ns	0.02	0.03	0.03	–0.01 to 0.01	–0.02 to 0.01	–0.02 to 0.02
	1	ns	ns	ns	0.01	0.02	0.03	–0.01 to 0.01	–0.01 to 0.01	–0.01 to 0.02
Wounds	0	ns	ns	ns	0.12	0.12	0.10	–0.12 to 0.12	–0.12 to 0.10	–0.05 to 0.11
	1	ns	ns	ns	0.12	0.11	0.10	–0.11 to 0.12	–0.11 to 0.11	–0.11 to 0.05
	2	ns	ns	ns	0.07	0.03	0.04	–0.05 to 0.05	–0.02 to 0.02	–0.02 to 0.02
Tail-biting	0	ns	ns	ns	0.09	0.06	0.10	–0.09 to 0.07	–0.03 to 0.03	–0.07 to 0.10
	2	ns	ns	ns	0.09	0.06	0.10	–0.07 to 0.09	–0.04 to 0.02	–0.10 to 0.07
Skin condition	0	ns	ns	ns	0.05	0.06	0.01	–0.04 to 0.02	–0.05 to 0.04	–0.01 to 0.01
	1	ns	ns	ns	0.05	0.05	0.01	–0.03 to 0.02	–0.04 to 0.04	–0.01 to 0.01
Hernia	0	ns	ns	ns	0.02	0.05	0.02	–0.01 to 0.01	–0.05 to 0.03	–0.01 to 0.01
	1	ns	ns	ns	0.02	0.03	0.02	–0.01 to 0.01	–0.02 to 0.02	–0.01 to 0.01
Coughing	n	s	s	ns	0.23	0.38	0.45	–0.11 to 0.30	–0.28 to 0.48	–0.48 to 0.49
Sneezing	n	ns	s	ns	0.17	0.14	0.13	–0.13 to 0.17	–0.11 to 0.12	–0.11 to 0.11
Pneumonia*	%		s			0.17			–0.12 to 0.17	
Pericarditis*	%		s			0.04			–0.03 to 0.02	
Pleuritis*	%		ns			0.11			–0.11 to 0.10	
Milkspots*	%		s			0.16			–0.13 to 0.17	

* As these parameters were retrieved from the slaughterhouse, comparison was only possible between the two fattening periods and not between each farm visit. † s = significant ($P < 0.05$); ns = non-significant differences.

https://doi.org/10.7120/09627286.25.4.447 Published online by Cambridge University Press

Figure 2

Bland and Altman (1986) plot of the Limits of Agreement of hernia, category 1 assessed at three visits for each of two fattening periods. White circles indicate the comparison of results of Farm visit 1, black circles of Farm visit 2 and grey circles of Farm visit 3.

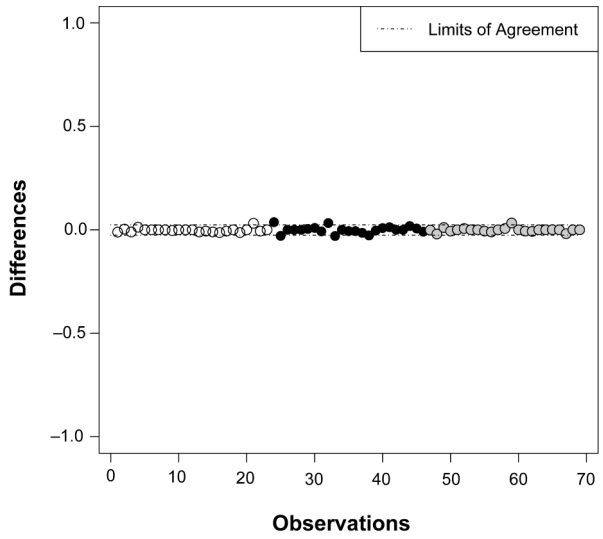
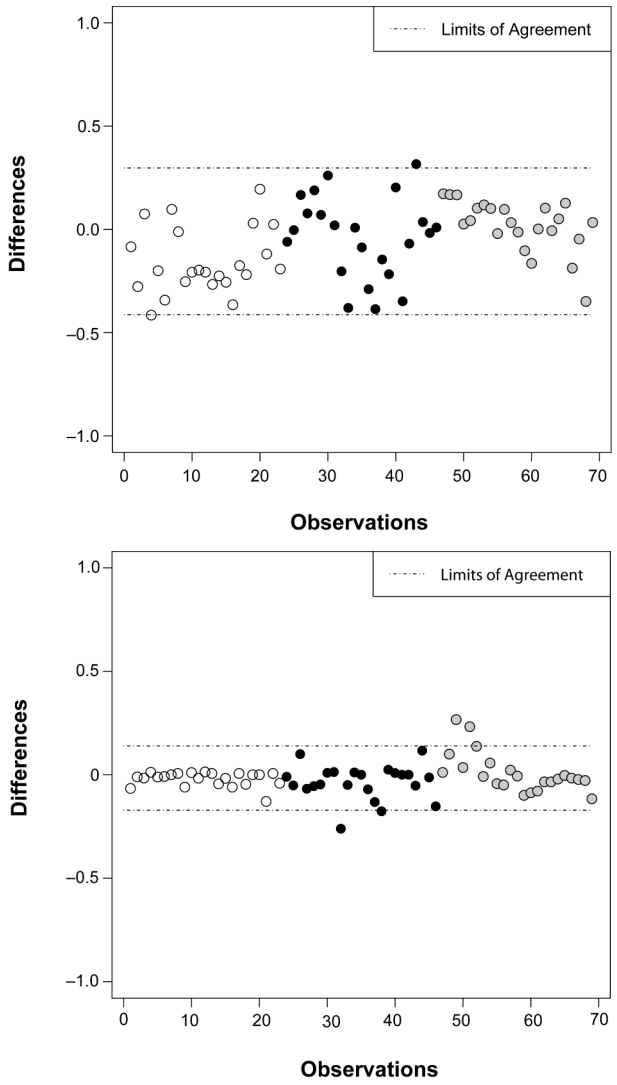


Figure 3

Bland and Altman (1986) plot of the Limits of Agreement of bursitis, category 1 (a) and 2 (b), assessed at three visits for each of two fattening periods. White circles indicate the comparison of results of Farm visit 1, black circles of Farm visit 2 and grey circles of Farm visit 3.



agreement in terms of the fixed limits; however, they did show a clear tendency towards it. The parameters coughing and sneezing also presented unacceptable agreement, with sneezing being closer to the acceptability limit.

The results of pneumonia, pleuritis and milkspots, which were obtained from the slaughterhouse, also presented insufficient reliability.

Discussion

Statistics and agreement parameters

A combination of significance testing (*W*) and two agreement parameters (SDC and LoA) were used for the assessment of reliability. As each statistical method has its own strengths and weaknesses, other studies have advised the use of several parameters and their interpretation together (Dohoo *et al* 2003; de Vet *et al* 2006).

Usually, reliability parameters, such as the Spearman Rank Correlation Coefficient and the Intraclass Correlation Coefficient, should also be used for an interpretation of reliability. The problem with reliability parameters is, however, that they are strongly dependent on the variance of the objects under study (Wirtz & Caspar 2002; de Vet *et al* 2006). This was exactly the case in this study. Therefore, the values and the interpretation of these reliability parameters would have been strongly biased. In further test-retest studies, farm samples with a greater variance would have to be taken into account.

Explanatory power of test-retest reliability

Low test-retest reliability can basically be due to three reasons (Temple *et al* 2013). Firstly, due to real differences between two visits. This was, however, excluded from the study as far as possible since no major changes on the assessed farms occurred. To further minimise possible influencing changes, a control for age effects was implemented when testing reliability by testing it within the same production stage. Still, there can be unknown and unpredictable time effects that cannot be diminished. Secondly, they can be caused by the sensitivity of a measure to routine procedures and minor changes, eg in the case of growing pig farms, changes in the age and weight of pigs. Inconsistency over time was definitely observed in this study since other animals were observed in the consecutive fattening periods. However, according to Knierim and Winckler (2009), the general feasibility of welfare assessment tools has to be carried out at longer time intervals of greater than six months in order to be useful for certification procedures. Therefore, in order to be a useful tool in terms of animal welfare assessment, the method must not be sensitive to minor changes (Temple *et al* 2013). The third main reason for a lack of reliability is variability due to methodological restrictions. The exact reason for the disagreement cannot be determined, but whatever the exact reason for a low test-retest reliability for a parameter, it means that the particular parameter is not suitable for the assessment in its present form.

The statement of Knierim and Winckler (2009) that the time interval should be greater than six months also led to the study design with the assessments in two consecutive fattening periods, since the results needed to be constant during this time in order to provide for feasibility.

Protocol assessments

The prevalence of the different parameters of the Welfare Quality® protocol obtained during the assessments were mostly in accordance with those found in previous studies (Temple *et al* 2011a). This was valid for the mm length of the adjectives of the QBA, for the percentages either of animals sorted into a certain category in terms of the BO or of pens on a farm with a flight reaction (HAR) as well as the percentage of affected animals in terms of the different IP. This statement is also valid for the time needed. The broad range for the time needed for the completion of one protocol assessment was mainly due to the distances between the different observation points and was therefore dependent on the layout of the evaluated farm. Protocol assessments could be fulfilled easily within a day in all cases. Thus, it was possible to carry out the assessments in a feasible manner (Knierim & Winckler 2009). Moreover, the working routine in the farms was not disturbed and the time needed with the farmer was very short. This made implementation under practical conditions easy. These findings are in complete agreement with the results of other studies, eg Kirchner *et al* (2014), who stated a good acceptance of the Welfare Quality® protocols among farmers. Moreover, within three days assessors from different backgrounds could be trained to show accordance in protocol assessments.

Qualitative Behaviour Assessment (QBA)

In general, the adjectives of the QBA did not agree in any of the three comparisons of farm visits. Even if the comparison of farm visit 3 indicated non-significant differences such as for the term active, SDC and LoA clearly indicated disagreement. As mentioned above, the results of *W*, SDC and LoA always have to be interpreted together. Moreover, at all times, significant differences were found for all three comparisons (except for the adjective distressed), which is a clear indicator of disagreement. The only exception to this was the term distressed, which presented non-significant differences in all three statistical criteria and SDC and LoA around the acceptability limit of 0.10. At all times, low millimetre scores on the VAS were assigned to the term distressed. This fact becomes logical when looking in detail at the meaning of distressed which, in terms of the protocol, is interpreted as a pig not only being unhappy but in a crisis situation. Therefore, agreement was probably achieved by the fact that very low scores were awarded at all times, indicating the absence of that term on the assessed farms. A study of farms with a broader variation in this adjective would be needed for a more meaningful assessment of reliability of the term distressed.

The direct comparison of mm length did not indicate acceptable test-retest reliability. The comparison of mm length was also carried out by Wemelsfelder and Millard (2009), who assessed inter-observer reliability of the QBA by the calculation of Kendall's Tau as a reliability parameter and came to the conclusion of good reliability. However, they tested the reliability based on video sequences, which might have provided a better reliability than the assessment on-farm.

https://doi.org/10.7120/09627286.25.4.447 Published online by Cambridge University Press

Behavioural observations (BO)

Acceptable to good agreement was found in the analysis of the percentages assigned to the dedicated behavioural categories in terms of BO. Non-significant differences were found between farm visits for the category pen investigation, but SDC and LoA only narrowly made it into the category of acceptable agreement. The explanation for this lies in practical experience showing behavioural pen investigation and other active behaviour to change in a very short time, thereby making them hard to differentiate. The present study is, to the authors' knowledge, the first to reveal this problem. It was taken into account in the congregation of parameters since pen investigation was not given as much weight as the use of enrichment material when calculating scores for the dedicated criterion (Welfare Quality® 2009). For this reason and due to the fact that acceptability limits were only narrowly exceeded, the general reliability of the BO can still be interpreted as good.

Human Animal Relationship test (HAR)

The HAR revealed non-satisfactory agreement during all compared farm visits. Mean prevalence and observations on the farms during data collection indicated an effect of age, since the prevalence of a panic response was clearly higher in young pigs and then decreased in the following visits with increasing weight and age. However, since the visits were compared at the same average weights, ie the same age classes of pigs, this effect alone cannot have caused the low agreement. Apparently, the outcomes of the HAR are subject to effects by minor changes, which was also stated by de Passillé and Rushen (2005). They not only pointed out this excessive sensitivity towards minor aspects but also questioned whether the HAR really allows differentiation between farms of good and poor animal welfare. To analyse the exact effects on the outcomes of the HAR, the effects should be estimated statistically via appropriate models in further studies. However, the test-retest reliability of this parameter is insufficient.

Individual parameters (IP)

Ambiguous results were found in terms of the different IPs. Some IPs occurred only rarely or not at all, thus making an assumption about their reliability meaningless. To make assumptions on the relevance of these rarely observed IPs, the Welfare Quality® Animal Welfare Assessment Protocol for Growing Pigs should be tested on a broader range of different farms with a greater variation, ideally internationally. If they are proven to be irrelevant due to low prevalence, they can be excluded from the protocol, thereby increasing feasibility as the time needed for the assessments could be shortened (Knierim & Winckler 2009). Further, an assumption about mortality was not possible since, in terms of the protocol, this parameter is assessed as mortality during the last 12 months. In the present study, the value assigned for mortality remained the same over the two consecutive fattening periods.

The test-retest reliability of many IPs, ie lameness, tail-biting, skin condition, hernia, pericarditis and wounds was good with the exception of wounds of category 1. Wounds of category 1 narrowly exceeded the acceptability limit in the comparison of farm visits 2 and 3. This is probably due to the fact that wounds tend to be received at the beginning of the fattening periods by fights in order to establish a rank order in the newly arranged groups (Meese & Ewbank 1973). These fights can be of differing severity and length depending on the aggressiveness of individual group members (D'Eath 2002). Since other individuals and group compositions were assessed during two fattening periods, this might explain the differences observed. Furthermore, the slotting criteria for wounds are somewhat complicated as presented in Table 1. Nevertheless, as wounds of category 2 were well within the ranges of acceptability and category 1 only narrowly exceeded these borders, this parameter can still be interpreted as acceptable. This is especially valid as category 2 implies more relevant constraints in welfare and is thus attributed a greater weight in the final congregation of this parameter (Welfare Quality® 2009).

The IPs bursitis and manure on the body presented insufficient reliability. In terms of bursitis, this can be explained by the fact that it is defined in the Welfare Quality® protocol as a swelling in the region of the joints and assessed mostly visually. In unclear cases of bursitis, it was proposed during training to palpate the legs to elucidate the findings. This was, however, often not possible especially when the animals were moving around quickly. These assessment difficulties that arise as a result of animals moving quickly, being dirty or when buildings are relatively dark were also described by Veissier *et al* (2013). Furthermore, it has to be considered that other causes of swelling in the joint region, eg haematoma or bacterial infection leading to increased synovial fluids in the joints (Plonait *et al* 2004) — which have to be borne in mind as differential diagnoses — cannot be differentiated accurately through mere visual assessment. Temple *et al* (2013) also stated insufficient reliability for this parameter. In contrast, Forkmann and Keeling (2009) found good reliability. However, they used the five-scale scoring system for bursitis of Lyons *et al* (1995) and, therefore, it is not directly comparable with this study which utilised a three-point scale as per the terms of the Welfare Quality® protocol (Veissier *et al* 2013). Although bursitis category 2 was of slightly better agreement, thus indicating a clearer definition, it can be concluded that the parameter is incapable of assessing comfort around resting in its present form. To this end, either clearer slotting criteria or other parameters have to be detected. The insufficient reliability of the parameter manure on the body is most likely caused by seasonal effects. It is a well-known fact that pigs in conventional housing systems tend to wallow in their own dung on hot days in order to cool down. This effect of seasonal changes was also stated by Huynh (2005) and Temple *et al* (2013). Therefore, this parameter is unreliable in its present form and incapable of assessing comfort around resting. More suitable parameters have to be found in the future, especially since both animal-based parameters congregated for the assessment of the criterion 'comfort around resting' have proved to have an insufficient test-retest reliability.

Insufficient agreement was further detected in the parameters coughing and sneezing, whereby sneezing presented slightly better agreement. However, both parameters seem to be subject to minor changes, such as seasonal effects and thus are not sufficiently reliable. A correlation between these parameters and data from the slaughterhouses, such as pneumonia should be subject to further analysis. This could answer the question as to whether the parameter coughing is actually needed for a reliable welfare assessment or whether pathological findings in the lungs and treatment data, ie use of antibiotics during a fattening period, would be a more appropriate way of assessing the health of respiratory organs.

Pneumonia, pleuritis and milkspots were the subject of disagreement between the two farm visits. As the data for these parameters were obtained from the slaughterhouse, only the results of one fattening period could be compared to the other and not the different visits at different ages. Low repeatability could be caused by unclear definitions and slaughterhouse assessments (Olsen *et al* 2007; Hoischen-Taubner *et al* 2011). This would imply that a clear definition, which is internationally valid, could provide an improvement of reliability.

Animal welfare implications

The results of the present study are an important contribution towards the enhancement and advancement of the Welfare Quality® Animal Welfare Assessment Protocol for Growing Pigs and thus of considerable interest in the pursuit of an objective measurement of the welfare status of farm animals.

Conclusion

The aim of the present study was to draw conclusions regarding the feasibility and test-retest reliability of the Welfare Quality® Animal Welfare Assessment Protocol for Growing Pigs. Feasibility of the protocol proved to be satisfactory. As for the test-retest reliability, insufficient reliability was found in terms of the QBA. BO, in the form of instantaneous scan sampling as a parameter for the assessment of social and other behaviour, generally provided good reliability. This is valid despite the fact that the category pen investigation lay, to some extent, outside the defined limits of acceptable reliability, which should be taken into account at interpretation. The HAR was subject to minor changes between farm visits, thus it was not capable of reliably assessing the human animal relationship. The majority of IPs presented good reliability, exceptions were bursitis, manure on the body and coughing and sneezing as well as the parameters retracted from the slaughterhouse. Some IPs occurred only rarely or not at all, thus rendering assumptions on their reliability meaningless. To test the relevance of these IPs, further studies of a farm sample with a greater variance, ideally internationally, are needed. This would have the further advantage of making reliability assessment easier. In general, many welfare parameters included in the 'Welfare Quality® Animal Welfare Assessment Protocol for Growing Pigs' proved to be sufficiently feasible and reliable. However, the present study also revealed there still to be a considerable number of challenges to be addressed in further studies on the Welfare Quality® protocols in order to achieve a constant improvement.

Acknowledgements

This work was supported financially by the German Federal Ministry of Food, Agriculture and Consumer Protection (BMELV) through the Federal Agency for Agriculture and Nutrition (BLE), grant number 2816806711.

References

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

Blokhuis H, Jones B, Veissier I and Miele M 2013 *Improving Farm Animal Welfare*. Wageningen Academic Publishers: Wageningen, The Netherlands. http://dx.doi.org/10.3920/978-90-8686-770-7

Carlson NR, Heth D, Miller H, Donahoe J and Martin GN 2009 *Psychology: The Science of Behavior*. Pearson: Harlow, UK

Curry HB and Schoenberg IJ 1966 On Polyoma frequency functions IV: the fundamental spline functions and their limits. *Journal d'Analyse Mathematique 17*: 71-107. http://dx.doi.org/10.1007/BF02788653

D'Eath RB 2002 Individual aggressiveness measured in a resident-intruder test predicts the persistence of aggressive behaviour and weight gain of young pigs after mixing. *Applied Animal Behaviour Science 77*: 267-283. http://dx.doi.org/10.1016/S0168-1591(02)00077-1

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. http://dx.doi.org/10.1016/j.applanim.2005.05.006

de Vet HCW 2005 *Observer Reliability and Agreement*. Wiley and Sons: Chichester, UK. http://dx.doi.org/10.1002/0470011815.b2a04033

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. http://dx.doi.org/10.1016/j.jclinepi.2005.10.015

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. http://dx.doi.org/10.2340/16501977-0337

Forkmann B and Keeling LJ 2009 *Assessment of Animal Welfare Measures for Sows, Piglets and Fattening Pigs*. Welfare Quality Reports: Cardiff, 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

Gelman A and Stern H 2006 The difference between significant and not significant is not itself statistically significant. *The American Statistician 60*: 328-331. http://dx.doi.org/10.1198/000313006X152649

Grabisch M and Roubens M 2000 Application of the Choquet integral in multicriteria decision making. In: Grabisch M, Murofushi T and Sugeno M (eds) *Fuzzy Measures and Integrals* pp 348-375. Physika Verlag: Heidelberg, Germany

https://doi.org/10.7120/09627286.25.4.447 Published online by Cambridge University Press

Hoischen-Taubner S, Blaha T, Werner C and Sundrum A 2011 Zur Reproduzierbarkeit der Befunderfassung am Schlachthof fuer Merkmale der Tiergesundheit. *Archiv fuer Lebensmittelhygiene 6*: 82-87. [Title translation: Repeatability of anatomical-pathological findings at the abattoir for characteristics of animal health]

Huynh TTT, Aarnink AJA, Gerrits WJJ, Heetkamp MJH, Canh TT, Spoolder HAM, Kemp B and Verstegen MWA 2005 Thermal behaviour of growing pigs in response to high temperature and humidity. *Applied Animal Behaviour Science 91*: 1-16. http://dx.doi.org/10.1016/j.applanim.2004.10.020

Kirchner MK, Westerath-Niklaus HS, Knierim U, Tessitore E, Cozzi G, Vogl C and Winckler C 2014 Attitudes and expectations of beef farmers in Austria, Germany and Italy towards the Welfare Quality® assessment system. *Livestock Science 160*: 102-112. http://dx.doi.org/10.1016/j.livsci.2013.12.004

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.

Koehler W, Schachtel G and Voleske P 1996 *Biostatistik*. Springer: Berlin, Germany. http://dx.doi.org/10.1007/978-3-662-06117-6

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

Magerman DM 1995 Statistical decision-tree models for parsing. *Proceedings of the 33rd annual Meeting on Association for Computational Linguistics* pp 276-283. Association for Computational Linguistics: Cambridge, Massachusetts, USA. http://dx.doi.org/10.3115/981658.981695

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

Meese GB and Ewbank R 1973 The establishment and nature of the dominance hierarchy in the domesticated pig. *Animal Behaviour 21*: 326-334. http://dx.doi.org/10.1016/S0003-3472(73)80074-0

Olsen EV, Candek-Potokar M, Oksama M, Kien S, Lisiak D and Busk H 2007 On-line measurements in pig carcass classification: Repeatability and variation caused by the operator and the copy of instrument. *Meat science 75*: 29-38. http://dx.doi.org/10.1016/j.meatsci.2006.06.011

Plonait H, Bickhardt K and Waldmann K-H 2004 *Lehrbuch der Schweinekrankheiten*. Georg Thieme Verlag: Stuttgart, Germany

SAS Institute 2008 *SAS/STAT 9.2. User's Guide*. SAS Institute Inc: Cary, NC, USA

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

Temple D, Courboulay V, Manteca X, Velarde A and Dalmau A 2011c *The Welfare of Pigs in Five Different Production Systems in France and Spain: Assessment of Behaviour*. Wageningen Academic Publishers: Indianapolis, USA

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

Temple D, Courboulay V, Velarde A, Dalmau A and Manteca X 2012a The welfare of growing pigs in five different production systems in France and Spain: assessment of health. *Animal Welfare 21*: 257-271. http://dx.doi.org/10.7120/09627286.21.2.257

Temple D, Dalmau A, Ruiz de la Torre J, Manteca X and Velarde A 2011a Application of the Welfare Quality protocol to assess growing pigs kept under intensive conditions in Spain. *Journal of Veterinary Behavior: Clinical Applications and Research 6*: 138-149. http://dx.doi.org/10.1016/j.jveb.2010.10.003

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. http://dx.doi.org/10.1016/j.livsci.2012.10.012

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

Tierschutzbund D 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']

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. http://dx.doi.org/10.3920/978-90-8686-770-7_6

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

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

Welfare Quality® 2009 Welfare Quality® applied to growing and finishing pigs. In: Dalmau A, Velarde A, Scott K, Edwards S, Veissier I, Keeling I and Butterworth I (eds) *Welfare Quality® Assessment Protocol for Pigs*. Welfare Quality® Consortium Lelystad: Wageningen, The Netherlands

Wemelsfelder F and Millard F 2009 Qualitative behaviour assessment. In: Forkman B and Keeling, L (eds) *Welfare Quality Reports* pp 213-219. SLU Service/Reproenheten: Uppsala, Sweden

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

Windschnurer I, Boivin X and Waiblinger S 2009 Reliability of an avoidance distance test for the assessment of animal's responsiveness to humans and a preliminary investigation of its association with farmer's attitudes on bull fattening farms. *Applied Animal Behaviour Science 117*: 117-127. http://dx.doi.org/10.1016/j.applanim.2008.12.013

Wirtz M and Caspar F 2002 *Beurteileruebereinstimmung und Beurteilerreliabilitaet*. Hogrefe: Goettingen, Germany. [Title translation: Observer agreement and observer reliability]