Repeated locomotion scoring of a sow herd to measure lameness: consistency over time, the effect of sow characteristics and inter-observer reliability

R.B. D'Eath

Animal Behaviour & Welfare, Sustainable Livestock Systems, SAC, West Mains Road, Edinburgh, EH9 3JG

rick.death@sac.ac.uk

Running head: Locomotion scoring of sows

Abstract

Investigating variability of scores between different observers, between animals and over time aids the design of valid sampling methodologies for measuring animal welfare. Locomotion scores (0 to 5 scale) were collected from i) 154 sows in one herd, using 3 to 5 (mean ± s.d. = 3.95 ± 0.65) observers each time, and scoring sows on up to 10 (mean ± s.d. = 4.8 ± 2.7) occasions over a 19 month period, and ii) for 123 of these sows, locomotion scoring also took place prior to farrowing and at weaning.

The distribution of scores was highly skewed towards low scores (0: 84.8%, 1: 9.5%, 2: 4.0%, 3+: 1.7%). Sows showed moderate consistency in their scores over time (W=0.496, p<0.001), and later parity sows had higher scores (χ^2_1 = 3.98, p = 0.049), but there was no effect of stage in the reproductive cycle (days pregnant, pre-farrowing, post-weaning). This suggests that infrequent visits to a farm (e.g. annual) might give a fair picture of the extent of lameness if a representative range of parities was sampled, although a larger study of more farms would be required to investigate this.

The three different types of agreement between observers (absolute differences, matching and association) were assessed as follows: i) Analysis of absolute differences between observers showed that the farm manager scored lower than researchers/technicians (S = 102.6, p < 0.001). ii) Exact matching approaches suggested fair (κ = 0.443) or good (PABAK = 0.692) agreement. Agreement was poorest for mild gait abnormalities (score 1 "stiff"), and agreement improved if scores were combined into 'sound' (0-1) and 'lame' (2-5) categories (κ = 0.653). iii) Measures of association suggested moderate agreement (Kendall's W = 0.692, p<0.001). Inter-observer reliability improved over time until the 5th scoring event. To improve inter-observer agreement, observer training/practice and the use of fewer categories are recommended, and inter-observer agreement should be checked regularly.

Keywords: animal welfare, inter-observer agreement, lameness, locomotion scoring, pigs, sows

Introduction

On-farm measurement of animal health and welfare is an important and current issue, to meet consumers' demands for demonstrably high standards of farm animal welfare (Blokhuis et al 2003; Blokhuis et al 2008). Setting standards and inspecting to ensure that they are met is a goal of government agencies (Gibbens 2008) and of voluntary farm assurance schemes such as the UK Red Tractor scheme, free range, ecological and organic (Main et al 2003; Main et al 2007; Veissier et al 2008). Membership of such schemes generally depends on the producer meeting certain 'design criteria' (Rushen & De Passillé 1992) relating to the housing and resources provided to animals (such as stocking density, drinkers, substrates), management (e.g. weaning age in pigs, age at slaughter) and administration (e.g. keeping accurate records of the use of drugs). Conformance with these criteria is generally assessed in a visit which takes place about once a year and takes less than a day to complete (Main et al 2007).

With a few exceptions, direct assessment of health and welfare by inspecting the animals themselves (animal-based or 'performance criteria', Rushen & De Passillé 1992) has not formed part of these schemes. Recently an EU funded project 'Welfare Quality®' (Blokhuis et al 2003) developed a comprehensive animal-based scoring system for on-farm assessment of animal health and welfare for pigs (Welfare Quality® 2009) and other housed species. The measures adopted were assessed for validity (does the indicator really measure what it should), repeatability (across observers), and feasibility (can it be assessed quickly enough to be included in a short visit). This process has been described in general terms but not in detail (Keeling et al 2009; Knierim & Winckler 2009).

Integration of multiple measures into an overall assessment is a difficult part of Welfare Quality (Botreau et al 2007a; Botreau et al 2007b; Botreau et al 2007c; Botreau et al 2009; Knierim & Winckler 2009) and of similar schemes (e.g. Main et al 2007). Even before reaching this stage though, there are a number of difficulties (Knierim & Winckler 2009). For any single measure, there are already practical constraints: on-farm scoring of animal welfare involves a sampled subset of animals from each age class and housing type, often by one trained observer in an annual visit of less than one day (Mullan et al 2009). Some researchers have attempted to assess the effect of such low 'sampling intensity' on reliability of measures. These include studies of changes over time with repeated visits (Winckler et al 2007), the effect of sampling different numbers of animals at each visit (Mullan et al 2009; Main et al 2010), inter-observer reliability (Brenninkmeyer et al 2007; Bokkers et al 2009) and test-retest reliability (O'Callaghan et al 2003; Flower & Weary 2006; Bokkers et al 2009). These questions are not merely academic: animal-based scoring systems could, and perhaps should (e.g. FAWC 2008) become the basis on which producers are deemed to pass or fail the criteria of an assurance scheme, so certain standards of reliability must be reached (De Passille & Rushen 2005). Using locomotion scoring to detect lameness in sows as an example of an animal-based scoring method, the present study investigated the extent to which individual sows varied over time in their scores, and whether scores were affected by parity and stage of pregnancy or veterinary treatment. Observers had different levels of experience with sows, and were all initially naïve to the scoring system, so the extent to which experience with the system improved consistency in the absence of training in these varied observers was also examined.

Lameness occurs in a variety of captive species, and is of welfare concern as the animal experiences pain, discomfort and reduced mobility (Knowles et al 2008; Flower & Weary 2009). Lameness has production costs too due to costs of treatment and is responsible for premature culling of around 7 – 11 % sows (Lucia et al 2000; Anil et al 2005; Engblom et al 2007; Anil et al 2009; Jensen et al 2010).

Although automated methods of assessing lameness have been investigated (Gonzalez et al 2008; reviewed in dairy cattle by Flower & Weary 2009), simple human observer scoring systems are still the main method used as they are relatively cheap, reliable and easy to apply on farm (Pigs, Main et al 2000; KilBride et al 2009; KilBride et al 2010; Cattle, Winckler & Willen 2001; Flower & Weary 2006; Flower & Weary 2009; Rutherford et al 2009; Chickens, Kestin et al 1992; Garner et al 2002). For sows in most countries, confinement in stalls during pregnancy and crates during farrowing and lactation makes on-farm assessment of lameness difficult, as altered posture is most likely not as sensitive a measure of lameness as locomotion scoring (KilBride et al 2010). Confinement thus potentially obscures the extent of the problem. The move to group housing following the EU ban on individual stalls during pregnancy (Council Directive 2001/88/EC, 2001) in 2013 will both increase the need for good locomotion in sows and make lame sows easier to identify.

In the present study, I applied locomotion scoring to group-housed sows (*Sus scrofa*) on one farm, using multiple observers at each scoring event to measure inter-observer reliability. Agreement between observers was assessed in terms of i) absolute differences, ii) exact matching and iii) association. Although these are conceptually and statistically complementary approaches, they are rarely all used on the same dataset (but see Kaler et al 2009). Because the same sows were scored on several occasions, in addition I investigated consistency of scores over time, and the factors affecting locomotion scores such as sow parity, stage in the reproductive cycle and the application of veterinary medicines

Methods

Animals and Housing

154 Large White × Landrace sows at SAC's research pig unit were the subjects of this study. The study animals included maiden gilts up to parity 7 sows (mean ± s.d. parity = 2.89 ± 1.72), and included dry sows at various stages from soon after weaning (waiting to return to oestrous and be served) through to heavily pregnant sows. They were housed in groups of 1-6 (mean ± s.d. = 4.5 ± 1.6) to a pen. Figure 1 shows the building where the sows were housed during the study. The pens (3.60m × 6.45m) were concrete-floored, with an enclosed straw-bedded area at the rear (3.60m × 2.50m), walled with concrete blocks, with a 2.0m wide opening onto a solid-floored central dunging passage (3.60m × 1.95m), and an access passageway plus six individual feeding stalls side by side at the front (each 0.50m wide, 1.80m long). Sows were fed on a rationed quantity of a commercial sow diet suitable for their size/age and stage of gestation once a day (0800h). At each side of the pen was a barred gate across the width of the dunging passage, which could be swung across to shut the sows into the bedded area (Figure 1). Water was available in each pen via a nipple drinker mounted on one of the gates. The pens were arranged in two back-to-back rows of 9 pens (108 sow places) and an automated natural ventilation system set at 14.5 °C maintained a temperature of 8.1 °C min – 21.6 °C max, mean 16.2 ± 4.1 °C during the study. At one end of the building, there was an empty concrete-floored service pen, with a

small amount of straw and 3 cm deep sawdust (4.8m × 3.1m), used for artificial insemination, positioned adjacent to where two 'teaser' boars were individually housed.

Locomotion Scoring

Sows were locomotion scored in two contexts- 1) systematic dry sow herd scoring, where all sows currently in the dry sow house were scored at one time and 2) scoring of sows due to farrow, or recently weaned sows when they were being moved to or from their farrowing accommodation.

Dry Sow Herd Locomotion Scoring

This took place on 11 occasions in total. The first two scoring events took place 6 months apart, but after that they took place at intervals one to two months (min 27 – max 202 days apart, mean ± s.d. = 56.6 ± 53.9 days) over a period of 566 days between December 2008 and June 2010. There was turnover in the breeding herd (old sows being culled, new gilts coming in) and a proportion of sows were in the farrowing house at any one time. As such, between 47 and 76 (mean ± s.d. = 67.6 ± 8.4) sows were scored at each scoring event, and individual sows were scored on 1- 10 occasions (mean ± s.d. = 4.79 ± 2.66).

Between 3 and 5 observers (mean ± s.d. = 3.95 ± 0.65) were present at each scoring event. Observers B and C attended all 11 scoring events, A and D attended 9 events and observer E attended 4. Each observer independently assigned each sow a locomotion score according to the system shown in table 1 and did not discuss or compare scores with other observers. The scoring system was simplified from the system developed by Main et al. (2000) for growing pigs (see also KilBride et al 2009). Observers were also free to record comments for every sow. For the dry sow herd scoring, observers consisted of a scientist with practical pig experience over a period of 11 years, but now primarily desk-based (A; the first author), whose interaction with these sows was limited to the scoring sessions only, another scientist (B) with 8 ½ years of regular experience with sows in general, and running a project involving these particular sows, the pig unit manager (C) with over 28 years experience working with all ages of pigs every day, and two technicians (D) with 3 ½ years and (E) 18 years experience of regularly working with pigs, including these sows. Before scoring began, the scoring system was explained and discussed by the observers to ensure it was well understood, but no formal training was undertaken.

In the morning before scoring took place, sows were spray marked for ease of identification. Scoring began at 1330h and took approximately 45 – 70 mins. First, all sows on one side of the dry sow house (Figure 1) were shut into their bedded area, and then the group furthest from the service area was let out of their pen and moved to the service pen. Sows either walked or ran, and some required encouragement to walk from a stockperson walking behind them. During scoring, it sometimes became apparent that sows were difficult to move. These were scored by all observers present and then sows which were slow to move (scoring 4 Limping) were only moved a short way before being returned, while those scoring 5 'Downer' were scored in their home pen. Once sows had been moved to the service pen, they were shut in for 3 to 5 minutes, with one person who moved amongst the sows encouraging them to move to facilitate scoring. Sows were then returned to their home pen, where they were no longer shut into the bedded area. Then the next group was moved to the service pen in a similar way, and so on for all the pens of sows until one side of the building was complete, then the procedure was repeated

for the other side of the building. While sows were moving to the service pen, in the service pen and moving back, they were continually observed and scored by the observers. Scorers made sure to focus on each individual sow for at least 10 strides of walking during this period and to give their score once this was complete. Sows which were identified as scoring 3 or higher which had not previously been identified during the course of normal husbandry were examined following scoring and subsequently given appropriate veterinary treatment. This occurred on 13 occasions during the study. Depending on the suspected cause of the lameness, sows were treated with different combinations of drugs. Eight sows were given the non-steroidal anti-inflammatory drug Metacam (5ml) was usually given once, but for two sows this was repeated daily for up to 7 days. A three day course of the antibiotic depocillin (8-12.5ml depending on weight) was given to 8 sows and a broad spectrum antibiotic (Baytril; 8ml) was also used on two occasions.

Farrowing Sow Locomotion Scoring

For 123 of the sows, locomotion scoring took place prior to farrowing and after weaning. When they reached 109 days after service, (mean ± s.d. 4.38 ± 1.84 days before farrowing), they were moved out of the dry sow house, across a concrete outside area, into a separate farrowing building and placed individually in loose-housed (non-crate) farrowing pens (PigSAFE, Baxter et al 2011). They were locomotion scored as they walked. The scoring was all done by observer B (a scientist). Once their piglets were weaned (mean 26.8 ± 3.1 days after farrowing), sows were moved back to the dry sow house, and were again locomotion scored as they walked between buildings.

Statistical Analysis

The nature of the dry sow herd locomotion scoring data led to a number of challenges for analysis. The data were incomplete because not every sow was present at each scoring event: sows were sometimes in the farrowing house, and there was some turnover in the herd (culls and replacements) during the study period. Also, not every observer was able to attend every scoring event. So there were a lot of 'missing' data where a particular sow/ observer/ scoring event combination did not occur.

Non-linear mixed models for ordinal data (using SAS; Gilmour et al 1987; Keen & Engel 1997) were fitted for each event separately to assess inter-observer reliability. An underlying latent continuous variable for locomotion was assumed of the form: $y_{ij} = \beta_0 + u_i + e_{ij}$ where u_i is a sow-level normal random effect and e_{ij} are independent and identically distributed normal errors. y_{ij} was estimated from observer scores. The 0-5 scoring scale corresponds on the latent variable scale to 6 intervals: $(-\infty,0)$, $(0,I1)$, $(I1,I1+I2)$, $(I1+I2,I1+I2+I3)$,($I1+I2+I3,I1+I2+I3+I4$) and $(I1+I2+I3+I4,\infty)$ where I1, I1+I2, I1+I2+I3, I1+I2+I3+I4 are the thresholds for the categories. They were estimated when a threshold model for the latent variable was fitted as a generalized linear mixed model (GLMM) using the NLMIXED procedure in SAS 9.1 (SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513). Inter-observer reliabilities were derived from the variance components. In practice, scoring categories had to be merged on an event-by-event basis to enable model fitting when there were very few animals on specific scores.

A different approach to dealing with the 'missing' data was used to look at consistency over time in sow locomotion scores. A subset of the data was used: 81 sows which were scored on four or more occasions by the four observers that did the most scoring (A-D), were identified. Kendall's coefficient of

concordance was calculated across the first four scoring events for each sow, using the mean of the four observers' scores at each one. Note that the four specific scoring events used could vary between the sows in this subset.

Another challenge for model fitting was that there were a high proportion of zero scores in the data. To model the different factors affecting locomotion scores (sow parity and stage of pregnancy), the scores were first re-coded into 0-1 data (0 → 0, 1-5 → 1) and a Generalized Linear Mixed Model was used to fit a binomial model (with a logit transformation; Genstat 11th Edition), with sow as a random effect, and observer, scoring event, sow parity and days until next farrowing (either actual or estimated from service records) as fixed effects (with parity and days until next farrowing being fitted as covariates). Note that non-pregnant sows and pregnant sows were treated similarly in this analysis: non pregnant sows were simply those with greater than 113 days until next actual or predicted farrowing.

The effect of veterinary treatment on sow locomotion score was analysed by comparing scores for each observer using paired Wilcoxon signed-ranks test. Scores from 0 to 90 (mean ± s.d. = 32.8 ± 24.6) days before treatment began were compared with scores from 6 to 190 days after (mean ± s.d. = 57.3 ± 54.3). One observer (B) was present for all 13 of the relevant scoring events, but other scorers had some missing data (A = 8 events, C = 12, D = 10).

Various non-parametric methods were applied to measure inter-observer reliability. These require that there was no missing data at all, so data from the 498 occasions when four of the observers (A, B, C & D) were present was used (8 scoring events: 1, 3, 4, 6, 7, 8, 9 and 10; which included 137 sows). Using these data, three aspects of agreement were considered:

1) Whether observers differed systematically in the absolute level of scores awarded. This was assessed using Friedman for an overall comparison, and Sign tests and Wilcoxon signed rank tests (Minitab 15, 2006) to compare each pair of observers.

2) Whether scores from different observers matched exactly. Proportion of agreement and kappa were calculated overall, by pairs of observers, and for different scores (Minitab 15, 2006). The prevalence adjusted bias adjusted Kappa (PABAK, Byrt et al 1993) was also calculated as this is preferred by some researchers (Brenninkmeyer et al 2007; Rutherford et al 2009). Weighted kappa (using linear weightings) were also calculated (AgreeStat Excel workbook, Advanced Analytics, 2010). In interpreting Kappa and PABAK, the scale suggested by Byrt (1996) was used: 0 or less no agreement, 0.01-0.20 poor…, 0.21-0.40 slight…, 0.41-0.60 fair…,0.61-0.80 good…, 0.81-0.92 very good…, 0.93-1.00 excellent agreement. Finally, kappa statistics were calculated following a re-coding of the 0-5 data to 0-1 data in two different ways: Either all scores above zero were re-coded as 1 (0 → 0, 1-5 → 1), or 0 and 1 were re-coded as 0, and higher scores as 1 (0 and 1→ 0, 2-5 → 1). This was done for comparison with other researchers who have scored animals in two categories of 'sound' and 'lame' (e.g. Rutherford et al 2009). Two different approaches were chosen because of the difficulty in classifying score '1' animals (see Results).

3) Whether scores from different observers were associated. This was assessed using Kendall's coefficient of concordance (W; Genstat 11, VSN International Ltd, 2008) as a measure of overall

agreement, and Spearman's (ρ) and Kendall's (τ) rank correlation coefficients were calculated between each pair of observers (Genstat 11, VSN International Ltd, 2008).

Finally, farrowing sow locomotion scores from before farrowing and after lactation were compared using Wilcoxon-signed ranks test for paired data (after calculating differences, data were not normally distributed).

Ethical Note

This study was given ethical approval by SAC's Animal Experiments Committee. As detailed above, any Sows which scored 3 or higher (3 Lame, 4 Limping, 5 Downer) were given appropriate veterinary treatment, while those scoring 1 or 2 were investigated and treated if necessary. Since sows were being checked daily as part of their routine husbandry, it was rare for lameness problems to be newly identified as part of the scoring process.

Results

Distribution of scores

The frequency of scores given overall by each of the observers (i.e. for all sows over all scoring events) is shown in figure 2. The vast majority of scoring events resulted in the observer giving the sow a score of 0: Normal (mean % of scores at each level: 0: 84.8%, 1: 9.5%. 2: 4.0%, 3: 1.2%, 4: 0.4%, 5: 0.2%).

Consistency of repeated scores from the same individual sows

To estimate consistency over time of sow scores, the mean locomotion score of observers A-D, from 81 sows at four scoring events were used. These scores for this subset were distributed similarly to the dataset as a whole as follows: 78.1% had mean score 0, 16.4% had a mean score between 0.25 and 1, 4.0% had a mean score between 1.25 and 2, 1.5% scored 2.25 or higher. There was a significant moderate level of association between sows: Kendall's W = 0.496, $p < 0.001$, suggesting that individual sows showed similar scores over repeated scoring events.

Factors affecting locomotion score

GLMM models showed that there were differences in locomotion scores (re-coded to 0-1) due to observer (Wald statistic $\chi^2_4 = 114.22$, $p<0.001$) and scoring event: inspection of estimated means showed that scores were lower at later scoring events ($\chi^2_{10} = 38.02$, $p<0.001$). A histogram showing the distribution of sow parities in the study is shown in figure 3. Older (higher parity) sows had higher locomotion scores ($\chi^2_1 = 3.98$, $p = 0.049$). A histogram showing the distribution of stage of pregnancy is shown in figure 4. There was no effect of stage of pregnancy ($\chi^2_1 = 0.10$, N.S.). Locomotion scores observed before and after sows were given veterinary treatment for lameness did not differ for any observer.

Agreement between observers- do they differ?

From this section forward, all analyses (unless stated otherwise) use data from the 498 occasions when observers A-D were all present, as the mainly non-parametric methods used require no missing data. There were differences between the observers in the proportion of sows scored in each category. Observer C (pig unit manager) recorded a higher proportion of zeroes than the other observers (Figure

2). A Friedman test comparing observers for the same sow on the same occasion showed a highly significant difference between observers, and inspection of the sums of ranks showed this was because observer C's scores were lower than those of the other three observers ($S = 102.6$, $p < 0.001$; sums of ranks A = 1252, B = 1299, C = 1143, D = 1286). When pairs of values were compared with Sign and Wilcoxon tests, these confirmed that observer C was scoring lower than the other three, and observer A gave lower scores than B and D, who did not differ (Table 2).

Agreement between observers- do they match?

Raw proportions of agreement both overall and between pairs of observers were high (Table 3). The proportion of agreement was considerably higher for 0 than for other scores. This is best illustrated by the proportion of occasions on which all 4 observers agreed to give the same score (first row of table 3). The overall kappa statistic of 0.443 is at a level which suggests only 'fair' agreement (Byrt 1996), although the PABAK statistic at 0.692 suggests 'good' agreement.

When broken down by scores (table 3), kappa and PABAK were noticeably lower for score 1, perhaps reflecting the difficulty in identifying and agreeing on the threshold between 0:'normal' and 1: 'stiff', a score used for minor locomotor anomalies.

When broken down by pairs of observers (table 2), values of kappa (and of weighted kappa and kappa following combining of categories) were considerably lower for pairings involving observer C (A-C, B-C and C-D) than for the other three pairings. Pairings which include observer C suggest 'slight' agreement, while pairings including observers A, B and D agreed at 'fair' or 'good' level. Contingency tables of the frequency of scores given by observer pairs B and D (Table 4) and B and C (Table 5) are shown to illustrate good and poor agreement respectively. For observers B and D, scores which differ by 2 are rare, and the greatest number of discordant scores occurs at the 0 vs. 1 level. For table 5 (observers B and C- poor agreement), there are more scores which differ by 2, 3 or even 4. Again, the greatest discrepancy occurs at the 0 vs. 1 level.

Values for weighted kappa, which takes into account the size of the disagreement between observers (table 2) were considerably higher than those of kappa, suggesting that disagreement by a small degree (e.g. by one point of the scoring system) was more common than disagreement by a larger amount.

0-5 scores were converted into 0-1 scores (sound/lame) by combining categories in two different ways (Table 2): i) Scores of 0 were classed as 'sound' and scores of 1-5 were classed as lame. Or ii) scores of 0 or 1 were classed as 'sound' and scores of 2-5 were classed as 'lame'. The Kappa was considerably lower for the first of these methods than for the second, consistent with the suggestion that the observers showed higher levels of agreement about the higher (2+) categories than they did about the distinction between 'Normal (0)' and 'Stiff (1)' pigs.

Agreement between observers- are they associated?

Using all available data, a non-linear mixed model for ordinal data was fitted for each scoring event, to estimate inter-observer reliability. Across the 11 scoring events, reliability was moderate to high, ranging from 0.552 to 0.879. It was evident that reliability improved over time (Figure 5), reaching a plateau by about the 4th or 5th scoring event.

Using data from the 498 occasions when observers A-D were all present, Kendall's coefficient of concordance (W) showed that there was moderate agreement between the observers. Levels of spearman's ρ and kendall's τ (Table 2) for pairwise agreement were moderate to high. Values were again lowest for pairs involving observer C, although the difference was less marked.

Because pairings involving observer C showed a lower level of agreement than other pairings, Measures of agreement between the other observers, A, B and D were also calculated. Kendall's W was 0.784, Fleiss's Kappa was 0.557, and weighted Kappa was 0.684. These values were all higher by around 0.1 than the overall agreement measures (Table 2).

Farrowing sow locomotion scoring

The distribution of sows' locomotion scores before farrowing was 0: 83%, 1: 27%, 2: 12%, 3: 1%, and after weaning was 0: 89%, 1: 25%, 2: 7%, 5: 2%. There were no differences between the locomotion scores before farrowing and after weaning (W=439, n for test ignoring ties = 45, p=0.379).

Discussion

The right-skewed distribution of scores (figure 2) was as expected. This shape of distribution is common in locomotion scoring studies (e.g. KilBride et al 2009). Levels of sow lameness in the present study were comparable to previous reports: KilBride et al (2009) studied 88 UK herds and found that 14.4% of pregnant gilts and 16.9% of pregnant sows had abnormal gait (score of 1 or higher), while 1.0 and 1.8 % had minimal weight bearing on an affected limb (score 3 or higher). Respective figures for the present study of 15.2% and 1.8%. Further studies are needed to assess the welfare significance of these scores (Main et al 2000), but it seems likely that reduced weight bearing on the affected limb is indicative of some discomfort/pain. Approaches to measure the welfare significance of locomotion scores include the use of motivational measures (Weeks et al 2002) or analgesics (Rushen et al 2007; Flower et al 2008; Danbury et al 2000), but these have not as yet been applied to pigs.

The structure of the dataset, with inevitable 'missing' data as sows moved through the system, complicated the analysis of consistency of sow scores. Kendall's coefficient of concordance using the average of four observers showed a moderate level of consistency. This analysis corresponds with studies in dairy cows which reported a degree of consistency over time in locomotion scores (O'Callaghan et al 2003; De Rosa et al 2003). It suggests that sow lameness is often a chronic problem. Although veterinary treatments were only applied on 13 occasions during the study, there was no evidence that scores were improved after treatment compared with before. This result should be viewed with some caution: it was not the primary aim of the study, and sample size was very low. Also there was a large and variable amount of time between the treatment and the before and after scores. The treatments may well have worked well in the short or even long term, but new causes of lameness may have occurred before the next scoring event.

Analysis of predictive variables found that locomotion score (analysed as binomial data, 0 vs 1-5) was affected by parity but not stage of pregnancy. Parity may have affected locomotion score because heavier animals are more likely to become lame (because of the greater pressure on their feet and joints), or because older sows have had longer to pick up an accidental injury which may then take some time to resolve. In dairy cow studies, size, conformation and udder fill affect how cows walk (reviewed

by Flower & Weary 2009), and weight affects broiler locomotion scores (Kestin et al 1992). During this study, observers commented that they had had difficulty with the lowest end of the scoring system. It was felt that the system failed to reflect the diversity of 'normal' and 'abnormal' gaits, particularly in sows differing in age, weight or stages of pregnancy. For example 'swagger of rear end while walking' (2-slight lameness) was quite pronounced in some otherwise normal older sows, so the effect of parity on locomotion score may partly reflect this change in gait of older sows, as any deviation from score 0 would have affected this analysis.

In terms of recommendations for welfare assessment study design and sampling intensity, the variation across parities suggests that the sampling strategy of larger on-farm studies should take this into account, ensuring that a representative cross-section of the range of parities on the farm is sampled. The moderate consistency over time in sow scores suggests that locomotion scores do not change rapidly over time, so infrequent visits should give a fair representation of the typical locomotion score of a given herd. However, these recommendations are tentative. A larger study specifically addressing the question of sampling methods would be desirable (see e.g. Mullan et al 2009; Main et al 2010).

Three aspects of agreement between observers were analysed. Absolute differences between observers will be considered first. Variation of this type is particularly problematic when absolute consistency is required, for example in order to compare the actual level of lameness between different farms or studies. In our study, the farm manager (observer C) used lower scores and more zeroes than other observers and that his scores didn't match or agree as well with those of other observers. While being cautious not to over-interpret this finding from a single observer, it is notable that in dairy cattle, farmers as a group generally underestimate the incidence of lameness in their own herd (Wells et al 1993; Whay et al 2003; Rutherford et al 2009; Leach et al 2010). In a study of lameness in sheep, farmers were asked to estimate the prevalence of lameness and then to carry out direct animal-based scoring. Prevalence recorded was higher from animal-based scoring (than from their initial estimate) but scores were still systematically lower than those of a researcher whilst showing good agreement (correlation) with them (King & Green 2011). It is possible that farmers usually categorise low levels of lameness as 'normal', because the key thing for them is to identify the level at which the threshold for treatment occurs (King & Green 2011). Alternatively, since farmers primarily spend time with their own animals, on a farm with widespread low levels of lameness, this may become the 'new normal'.

A different interpretation of this finding would be that observer C was correct, and that the other less experienced observers over-scored. In particular, the farm manager may have had more experience of the range of 'normal' gaits in older sows (see observers comments above). Farmers and scientists have different perspectives on animal welfare which may affect their interpretation of the same scoring scheme (Hubbard & Scott 2011).

Observer agreement in terms of matching of scores will be considered next. The overall kappa statistic suggested fair agreement, although the PABAK statistic was higher suggesting good agreement (Byrt 1996). This difference probably arises because kappa does not just reflect 'agreement' but can be affected by bias, where there are systematic differences between observers, and prevalence, where the distribution of scores is not uniform (Byrt et al 1993). Both bias and prevalence were clearly issues in this dataset, and the PABAK statistic is intended to adjust for these. Rutherford et al (2009) reported a mean

PABAK of 0.88 (range 0.67 – 0.94) between 3 or more trained observers (after scores were pooled into 'sound' and 'lame' categories) in a study of dairy cattle lameness. The PABAK of 0.692 for the six level scoring system in the present study (Table 3) is quite good in comparison. The level of agreement was comparable to a dairy cow locomotion study by Winckler and Willen (2001) in which the proportion of agreement between three observers was 0.68, while the present study found a slightly higher level of 0.743 between four observers (Table 3).

When analysing agreement at each score level, Kappa and PABAK were lower for score 1, suggesting poorer agreement. Work in sheep (Kaler et al 2009) and cattle (Flower & Weary 2006) has also shown that agreement is better for higher scores distinguishing between the lowest level (normal) and next level up is often the most difficult. Winckler and Willen (2001) found that 62% of disagreements between observers were at this level, and two other studies found much improved agreement after merging the lowest 2 scores and the top 2 (or 3) scores into simpler non-lame and lame categories (Rutherford et al 2009; Brenninkmeyer et al 2007).

Finally, I applied methods of agreement based on association or near matching. Again, levels of agreement found were comparable to other studies Flower and Weary (2006) reported an R^2 from regression of 0.69 between 2 observers, which equates to correlation coefficient (r) of 0.83, higher than the best pairwise agreement measured by spearman's ρ (0.781) or Kendall's τ found in this study (0.771; see table 2). Higher levels of association and matching and lower levels of difference between observers than those found here have been reported in a study of sheep lameness (Kaler et al 2009). This may be in part due to the use of video sequences in this study rather than live scoring.

Mixed models showed that inter-observer reliability improved over time and then plateaued at around event 4 or 5 (Figure 3). This suggests that observers showed better agreement with each other with experience, probably because each showed more internal consistency and perhaps also because after each session observers discussed their experience with the method and how to deal with certain examples or borderline cases. Other studies have shown that locomotion scorers show higher levels of agreement with increasing experience (Main et al 2000) or with more training (March et al 2007; Brenninkmeyer et al 2007; Thomsen et al 2008).

Overall, methods to analyse agreement in terms of absolute scores, exact matching and association or near matching all showed that observers did not agree perfectly, typically showing 'moderate' agreement. The levels of agreement were largely in line with other studies of this type which rely on visual assessment and ordinal scoring systems. Ways in which such locomotion scoring systems could be improved are discussed further below. It was notable that agreement for score 1 was poorest and observers reported difficulties with the lower end of the scale. Improved kappa values were obtained when data were combined into 0 and 1 vs 2+, suggesting that a simple sound/ lame system would be preferable, as agreement is better (Brenninkmeyer et al 2007; Rutherford et al 2009). The 3 levels in the Welfare Quality protocol for pigs are equivalent to scores 0-2 (normal or altered gait but 'still using all legs'), score 3 ('minimum weight-bearing on affected limb') and scores 4-5 ('no weight-bearing on affected limb or unable to walk', Welfare Quality® 2009). Others have advocated more complex scoring systems to pick out specific types of abnormality (O'Callaghan et al 2003; Flower & Weary 2006). These

might be more precise for research purposes or where early diagnosis for intervention is a priority. If simpler systems are more reliable, these might be better for on-farm overall welfare assessment.

What is the ideal locomotion scoring system?

For any locomotion scoring system based on visual scoring by human observers (and indeed for welfare scoring systems in general), the following attributes are desirable: 1) Easy to use and efficient to carry out on a variety of experimental and commercial situations, 2) Objective, unambiguous descriptions of each score (Garner et al 2002), 3) External validation (Knierim & Winckler 2009) for example against foot pathologies (Flower & Weary 2006; KilBride et al 2010; Kaler et al 2011), analgesics (Rushen et al 2007; Flower et al 2008; Danbury et al 2000) or other behavioural measures such as in broiler chickens latency to lie down in shallow water, which is aversive for them, was associated with locomotion scores (Weeks et al 2002). 4) Training before scoring 'in the field', since training and experience increase agreement between observers (March et al 2007; Brenninkmeyer et al 2007). 5) Users (researchers, assurance schemes) should implement ongoing assessment of inter-observer reliability. Another interesting recent idea is to use a modified visual analogue scale, which retains the advantages of ordinal categorical scoring while adding the advantage of capturing some of the variation within categories (Tuyttens et al 2009).

Conclusions and animal welfare implications

Valid animal-based scoring methods to assess welfare are important both for research and for on-farm assurance purposes. This study suggested that the locomotion scoring system used was promising, although it would benefit from external validation. Sows were moderately consistent over time in their locomotion scores, and older sows had higher scores. In terms of animal welfare, this suggests that lameness is a chronic problem and that measures to treat it are not entirely successful. In terms of animal welfare assessment methodologies, it suggests that infrequent visits to a farm might give a fair picture of the extent of lameness, provided that a representative range of sow parities was sampled, although this requires further validation in a larger study.

Inter-observer agreement can be thought of in terms of three complementary approaches: absolute differences, exact matching and association, which are all useful. Of these, the issue of absolute differences is of greatest importance in terms of 'fairness' of welfare assessment (e.g. comparing the prevalence of lameness between individual farms or systems). Inter-observer agreement was moderate and improved with practice, suggesting that training and regular assessment of inter-observer agreement is important to ensure standardisation of data collection methods. Agreement improved when categories were combined: observers found minor locomotor anomalies difficult to classify. As such a simpler system may be preferable for application of welfare assessment on-farm (e.g. Welfare Quality® 2009).

Acknowledgements

SAC is supported by the Research and Science Division of the Scottish Government. Farrowing sow scoring data were collected as part of the Defra-funded project AW0143. Ian Nevison of BioSS gave statistical advice and ran the non-linear mixed models for ordinal data.

References

Anil SS, Anil L and Deen J 2005 Evaluation of patterns of removal and associations among culling because of lameness and sow productivity traits in swine breeding herds. *Javma-Journal of the American Veterinary Medical Association* 226: 956-961

Anil SS, Anil L and Deen J 2009 Effect of lameness on sow longevity. *Javma-Journal of the American Veterinary Medical Association* 235: 734-738

Baxter EM, Lawrence AB and Edwards SA 2011 Alternative farrowing systems: Design criteria for farrowing based on the biological needs of sows and piglets. *Animal* 5: 580-600

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

Blokhuis HJ, Keeling LJ, Gavinelli A and Serratosa J 2008 Animal welfare's impact on the food chain. Trends in Food Science & Technology 19, S79-S87.

Bokkers EAM, Leruste H, Heutinck LFM, Wolthuis-Fillerup M, van der Werf JTN, Lensink BJ and Van Reenen CG 2009 Inter-observer and test-retest reliability of on-farm behavioural observations in veal calves. *Animal Welfare* 18: 381-390

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

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

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

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

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

Byrt T 1996 How good is that agreement? *Epidemiology* 7: 561

Byrt T, Bishop J and Carlin JB 1993 Bias, Prevalence and Kappa. *Journal of Clinical Epidemiology* 46: 423-429

Danbury TC, Weeks CA, Chambers JP, Waterman-Pearson AE and Kestin SC 2000 Self-selection of the analgesic drug carprofen by lame broiler chickens. *Veterinary Record* 146: 307

De Passille 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

De Rosa G, Tripaldi C, Napolitano F, Saltalamacchia F, Grasso F, Bisegna V and Bordi A 2003 Repeatability of some animal-related variables in dairy cows and buffaloes. *Animal Welfare* 12: 625-629

Engblom L, Lundeheim N, Dalin AM and Andersson K 2007 Sow removal in Swedish commercial herds. *Livestock Science* 106: 76-86

FAWC 2008 Opinion on Policy Instruments of Protecting and Improving Farm Animal Welfare. London, Farm Animal Welfare Council.

Flower FC, Sedlbauer M, Carter E, von Keyserlingk MAG, Sanderson DJ and Weary DM 2008 Analgesics improve the gait of lame dairy cattle. *Journal of Dairy Science* 91: 3010-3014

Flower FC and Weary DM 2006 Effect of hoof pathologies on subjective assessments of dairy cow gait. *Journal of Dairy Science* 89: 139-146

Flower FC and Weary DM 2009 Gait assessment in dairy cattle. *Animal* 3: 87-95

Garner JP, Falcone C, Wakenell P, Martin M and Mench JA 2002 Reliability and validity of a modified gait scoring system and its use in assessing tibial dyschondroplasia in broilers. *British Poultry Science* 43: 355-363

Gibbens N 2008 Animal Health 2008 The Report of the Chief Veterinary Officer. London, Department of Food and Rural Affairs.

Gilmour AR, Anderson RD and Rae AL 1987 Variance-Components on An Underlying Scale for Ordered Multiple Threshold Categorical-Data Using A Generalized Linear Mixed Model. *Journal of Animal Breeding and Genetics-Zeitschrift fur Tierzuchtung und Zuchtungsbiologie* 104: 149-155

Gonzalez LA, Tolkamp BJ, Coffey MP, Ferret A and Kyriazakis I 2008 Changes in feeding behavior as possible indicators for the automatic monitoring of health disorders in dairy cows. *Journal of Dairy Science* 91: 1017-1028

Hubbard C and Scott K 2011 Do farmers and scientists differ in their understanding and assessment of farm animal welfare? *Animal Welfare* 20: 79-87

Jensen TB, Bonde MK, Kongsted AG, Toft N and Sorensen JT 2010 The interrelationships between clinical signs and their effect on involuntary culling among pregnant sows in group-housing systems. *Animal* 4: 1922-1928

Kaler J, George TRN and Green LE 2011 Why are sheep lame? Temporal associations between severity of foot lesions and severity of lameness in 60 sheep. *Animal Welfare* 20: 433-438

Kaler J, Wassink GJ and Green LE 2009 The inter- and intra-observer reliability of a locomotion scoring scale for sheep. *Veterinary Journal* 180: 189-194

Keeling L, Forkman B and Veissier I 2009 Towards a Welfare Quality Assessment System. Welfare Quality.

Keen A and Engel B 1997 Analysis of a mixed model for ordinal data by iterative re-weighted REML. *Statistica Neerlandica* 51: 129-144

Kestin SC, Knowles TG, Tinch AE and Gregory NG 1992 Prevalence of Leg Weakness in Broiler-Chickens and Its Relationship with Genotype. *Veterinary Record* 131: 190-194

KilBride AL, Gillman CE and Green LE 2009 A cross-sectional study of the prevalence of lameness in finishing pigs, gilts and pregnant sows and associations with limb lesions and floor types on commercial farms in England. *Animal Welfare* 18: 215-224

KilBride AL, Gillman CE and Green LE 2010 A cross-sectional study of prevalence and risk factors for foot lesions and abnormal posture in lactating sows on commercial farms in England. *Animal Welfare* 19: 473-480

King EM and Green LE 2011 Assessment of farmer recognition and reporting of lameness in adults in 35 lowland sheep flocks in England. *Animal Welfare* 20: 321-328

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 (R) approach. *Animal Welfare* 18: 451-458

Knowles TG, Kestin SC, Haslam SM, Brown SN, Green LE, Butterworth A, Pope SJ, Pfeiffer D and Nicol CJ 2008 Leg Disorders in Broiler Chickens: Prevalence, Risk Factors and Prevention. *PLoS ONE* 3:

Leach KA, Whay HR, Maggs CM, Barker ZE, Paul ES, Bell AK and Main DCJ 2010 Working towards a reduction in cattle lameness: 1. Understanding barriers to lameness control on dairy farms. *Research in Veterinary Science* 89: 311-317

Lucia T, Dial GD and Marsh WE 2000 Lifetime reproductive performance in female pigs having distinct reasons for removal. *Livestock Production Science* 63: 213-222

Main DCJ, Barker ZE, Leach KA, Bell NJ, Whay HR and Browne WJ 2010 Sampling strategies for monitoring lameness in dairy cattle. *Journal of Dairy Science* 93: 1970-1978

Main DCJ, Clegg J, Spatz A and Green LE 2000 Repeatability of a lameness scoring system for finishing pigs. *Veterinary Record* 147: 574-576

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

Main DCJ, Whay HR, Lee C and Webster AJF 2007 Formal animal-based welfare assessment in UK certification schemes. *Animal Welfare* 16: 233-236

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

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

O'Callaghan KA, Cripps PJ, Downham DY and Murray RD 2003 Subjective and objective assessment of pain and discomfort due to lameness in dairy cattle. *Animal Welfare* 12: 605-610

Rushen J and De Passillé AMB 1992 The scientific assessment of the impact of housing on animal welfare: A critical review. *Canadian Journal of Animal Science* 72: 721-743

Rushen J, Pombourcq E and De Passille AM 2007 Validation of two measures of lameness in dairy cows. *Applied Animal Behaviour Science* 106: 173-177

Rutherford KMD, Langford FM, Jack MC, Sherwood L, Lawrence AB and Haskell MJ 2009 Lameness prevalence and risk factors in organic and non-organic dairy herds in the United Kingdom. *Veterinary Journal* 180: 95-105

The Council of The European Union 2001 Council Directive 2001/88/EC of 23rd October 2001 amending Directive 91/630/EEC laying down minimum standards for the protection of pigs.

Thomsen PT, Munksgaard L and Togersen FA 2008 Evaluation of a lameness scoring system for dairy cows. *Journal of Dairy Science* 91: 119-126

Tuyttens FAM, Sprenger M, van Nuffel A, Maertens W and Van Dongen S 2009 Reliability of categorical versus continuous scoring of welfare indicators: lameness in cows as a case study. *Animal Welfare* 18: 399-405

Veissier I, Butterworth A, Bock B and Roe E 2008 European approaches to ensure good animal welfare. *Applied Animal Behaviour Science* 113: 279-297

Weeks CA, Knowles TG, Gordon RG, Kerr AE, Peyton ST and Tilbrook NT 2002 New method for objectively assessing lameness in broiler chickens. *Veterinary Record* 151: 762-764

Welfare Quality® 2009 *Welfare Quality® Assessment Protocol for Pigs*. Welfare Quality® Consortium: Lelystad, Netherlands

Wells SJ, Trent AM, Marsh WE and Robinson RA 1993 Prevalence and Severity of Lameness in Lactating Dairy-Cows in A Sample of Minnesota and Wisconsin Herds. *Journal of the American Veterinary Medical Association* 202: 78-82

Whay HR, Main DCJ, Green LE and Webster AJF 2003 Assessment of the welfare of dairy cattle using animal-based measurements: direct observations and investigation of farm records. *Veterinary Record* 153: 197-202

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

Winckler C and Willen S 2001 The reliability and repeatability of a lameness scoring system for use as an indicator of welfare in dairy cattle. *Acta Agriculturae Scandinavica Section A-Animal Science* 103-107

Table 1 The lameness scoring system used in this study. Observers used the integer scores as instructed, on all but 4 occasions when an intermediate score (e.g. 1.5) was recorded.

Score	Label	Description
0	Normal	Even strides, rear end sways slightly while walking, pig is able to accelerate and change direction rapidly. Stands normally.
1	Stiff	Abnormal stride length, movements no longer fluent, pig appears stiff, Pig still able to accelerate and change direction. Stands normally.
2	Slight lameness	Shortened stride, lameness detected, Swagger of rear end while walking, no hindrance in pig's agility. Uneven posture while standing.
3	Lame	Pig slow to get up (may dog-sit), Shortened stride, Minimum weight-bearing on affected limb (standing on toes), Swagger of rear end while walking. May still trot and gallop.
4	Limping	Pig reluctant to get up, holds limb off floor while standing, avoids placing affected limb on the floor while moving
5	Downer	Pig unresponsive- does not move and struggles to stand when encouraged to do so.

Table 2 Inter-observer agreement statistics are shown as follows: *Differences* between observers' scores. Friedman test for effect of the 4 observers on locomotion score, blocked by sow-event. For pairwise comparisons using Sign and Wilcoxon signed-ranks tests: The score for the second observer was subtracted from the score for the first observer (e.g. for A-B difference = A minus B), so +ve difference means the first score was higher (e.g. A's score higher than B's). *Association* between pairs of observers' scores as calculated by Spearman's Rho (ρ) correlation coefficient and Kendall's Tau (τ) correlation coefficient, Kendall's Coefficient of concordance (W) between multiple observers is also given. *Agreement (exact matching)* between observers' scores given by Fleiss's Kappa (κ) for 4 observers and Cohen's Kappa (κ) for pairs of observers. Weighted Kappa is also shown (using a linear decline in weightings further from the diagonal). Finally, Kappa (κ) calculated after re-coding the data to 1-0 form using two different methods is shown. In the first method, all scores of 1-5 were re-coded as 1 ('lame'), and in the second method, scores of 0 and 1 were re-coded as 0 ('sound'), while scores of 2-5 were re-coded as 1 ('lame'). † = p<0.1 * = p<0.05, ** = p<0.01, *** = p<0.001.

	Sign test (-ve : +ve)	Friedman or Wilcoxon test	Spearman's (ρ)	Kendall's W or τ	Kappa (κ)	Weighted Kappa (κ_w)	Kappa (κ) after combining categories (sound = 0, lame = 1+)	Kappa (κ) after combining categories (sound = 0 or 1, lame = 2+)
Overall (A-B-C-D)		102.6***		0.692***	0.443***	0.582	0.532	0.653
A-B	53 : 28**	1199.5*	0.604***	0.591***	0.482***	0.629	0.549	0.725
A-C	8 : 64***	2352.5***	0.536***	0.524***	0.289***	0.457	0.440	0.478
A-D	46 : 28*	1047.5†	0.643***	0.626***	0.498***	0.642	0.599	0.704
B-C	6 : 85***	3863.0***	0.457***	0.448***	0.249***	0.416	0.335	0.574
B-D	24 : 29	754.0	0.781***	0.771***	0.679***	0.777	0.750	0.852
C-D	74 : 5***	160.0***	0.555***	0.541***	0.290***	0.459	0.439	0.488

Table 3 Raw proportions of agreement between observers broken down into overall agreement (all 4 observers agree) and agreement between pairs. Proportions are worked out using n = 498 as the denominator for calculations of overall agreement. For separate score proportions (for 0, 1, 2, 3, 4 & 5), the denominator was worked out as the average number of scores awarded at that level, either overall (scores in **bold**) or by that pair of observers. The Kappa statistic (κ) and the Prevalence Adjusted Bias Adjusted Kappa (PABAK) are also given *in italics* overall and broken down by score.

Agreement		Proportion of Agreement						
	(n / 498)	Overall	0	1	2	3	4	5
Kappa (κ)		*0.443*	*0.532*	*0.279*	*0.486*	*0.458*	*0.379*	*1.000*
PABAK		*0.692*	*0.834*	*-0.176*	*0.040*	*0.004*	*-0.20*	*1.000*
Overall (A-B-C-D)	370	**0.743**	**0.862**	**0.020**	**0.200**	**0.170**	**0.000**	**1.000**
A-B	417	**0.837**	0.918	0.370	0.667	0.727	0.500	1.000
A-C	426	**0.855**	0.937	0.097	0.229	0.600	0.000	1.000
A-D	424	**0.851**	0.932	0.364	0.612	0.429	0.500	1.000
B-C	407	**0.817**	0.906	0.198	0.333	0.444	0.000	1.000
B-D	445	**0.894**	0.951	0.641	0.727	0.462	0.667	1.000
C-D	419	**0.841**	0.929	0.169	0.294	0.167	0.000	1.000

Table 4 Contingency table of scores for observer B and D to illustrate good agreement. Exact matches (on the diagonal, in **bold**) were summed to give the agreement, which was then expressed as a proportion (of 498; see Table 3).

		Observer D Sow locomotion scores						
		0	1	2	3	4	5	*Totals*
Observer B Sow locomotion scores	0	**382**	12	1				*395*
	1	26	**41**	6	1			*74*
	2		1	**16**	3			*20*
	3			1	***3***	1		*5*
	4				1	***2***		*3*
	5						**1**	*1*
	Totals	*408*	*54*	*24*	*8*	*3*	*1*	*498*

Table 5 Contingency table of scores for observer B and C to illustrate poor agreement. Exact matches (on the diagonal, in **bold**) were summed to give the agreement, which was then expressed as a proportion (of 498; see Table 3).

		Observer C Sow locomotion scores						
		0	1	2	3	4	5	*Totals*
Observer B Sow locomotion scores	0	**390**	3	2				*395*
	1	65	**9**					*74*
	2	9	5	**5**	1			*20*
	3	1		2	**2**			*5*
	4	1		1	1	**0**		*3*
	5						**1**	*1*
	Totals	*466*	*17*	*10*	*4*	*0*	*1*	*498*

Figure 1 Diagram of dry sow house where the study was carried out, including dimensions (not to scale). Dashed lines indicate barred gates, shown in their normal position. The top left pen indicates the arc of gate swing to show how gates can be temporarily closed to shut sows into the straw bedded area for mucking out and for locomotion scoring. Sows were observed while being moved from one of the sow pens, along the dunging passage, into the service pen and back again.

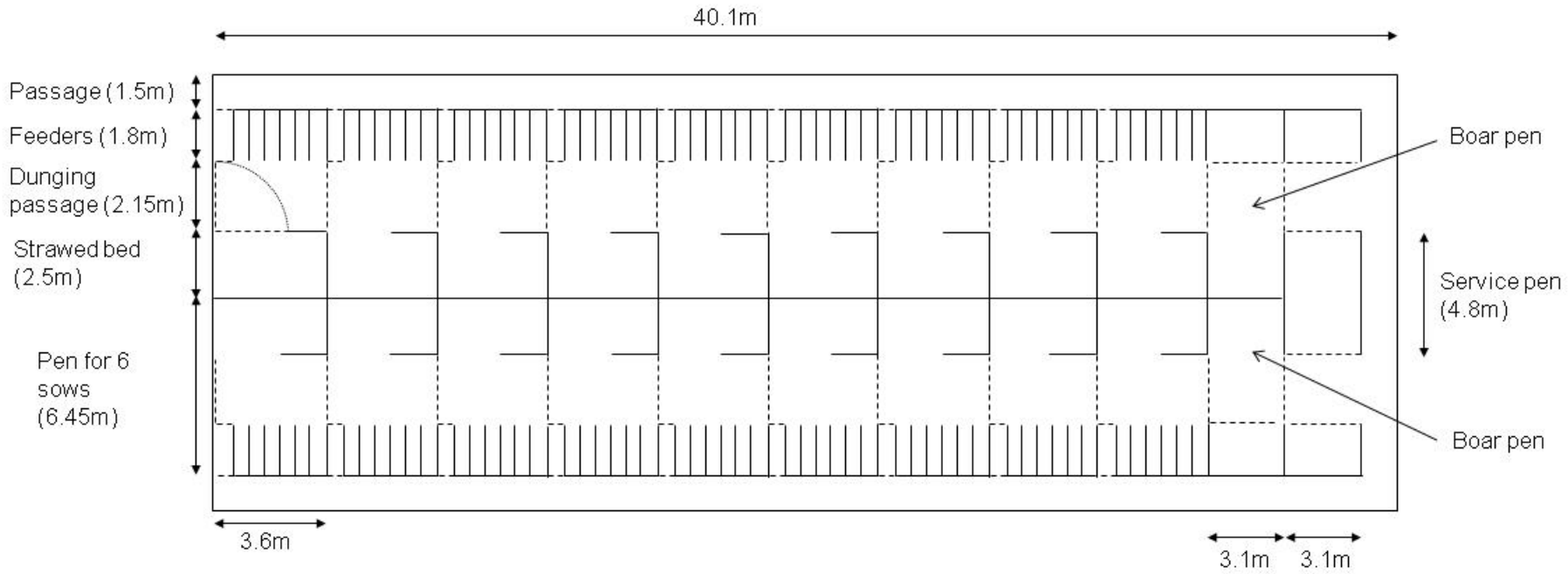

Figure 2 Frequency of locomotion scores given overall by four different observers (A, B, C and D). This data is all scores generated by these 4 observers used in the study (i.e. for all sows at all scoring events), so includes repeat observations of the same sows. Note the broken scale on the Y axis.

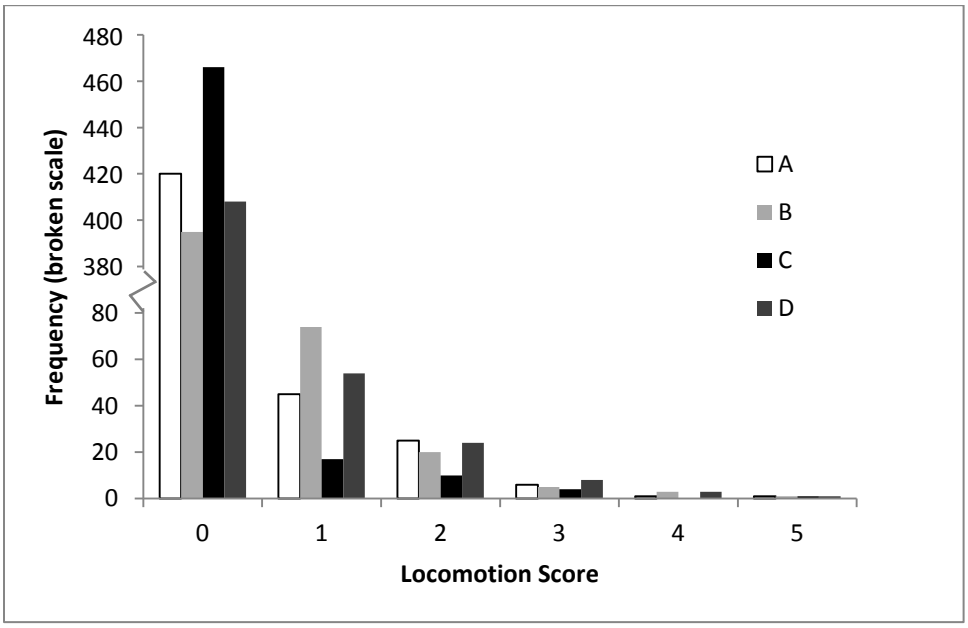

Figure 3 Histogram showing the distribution of sow parities at the end of the study. Sows which have not yet been served are shown as parity 0, parity 1 sows have produced one litter etc. (n=113, data not available for 41 sows).

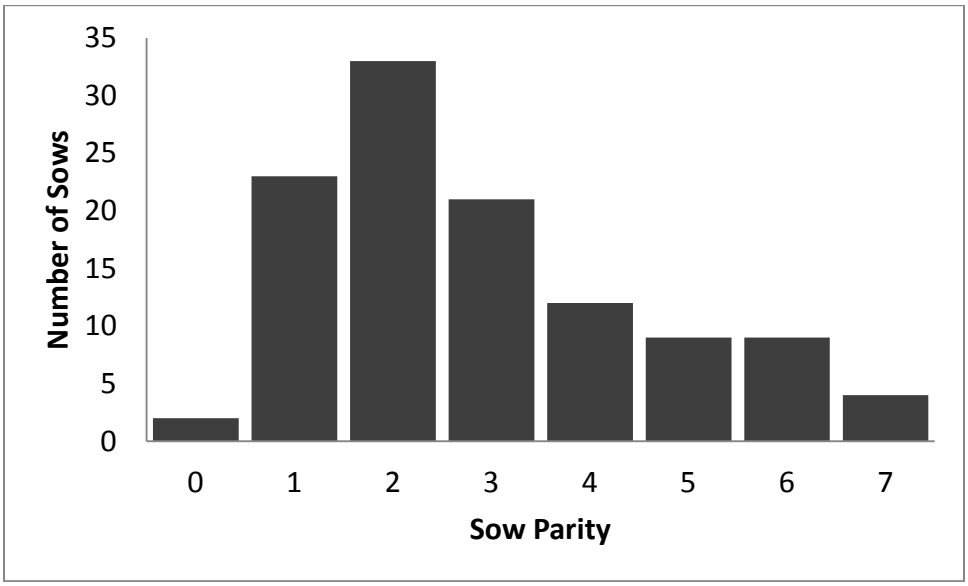

Figure 4 Histogram showing the number of sows at each stage of pregnancy. Sows were usually moved into the farrowing house 109 days after service. N = 563 for this graph as individual sows can appear repeatedly.

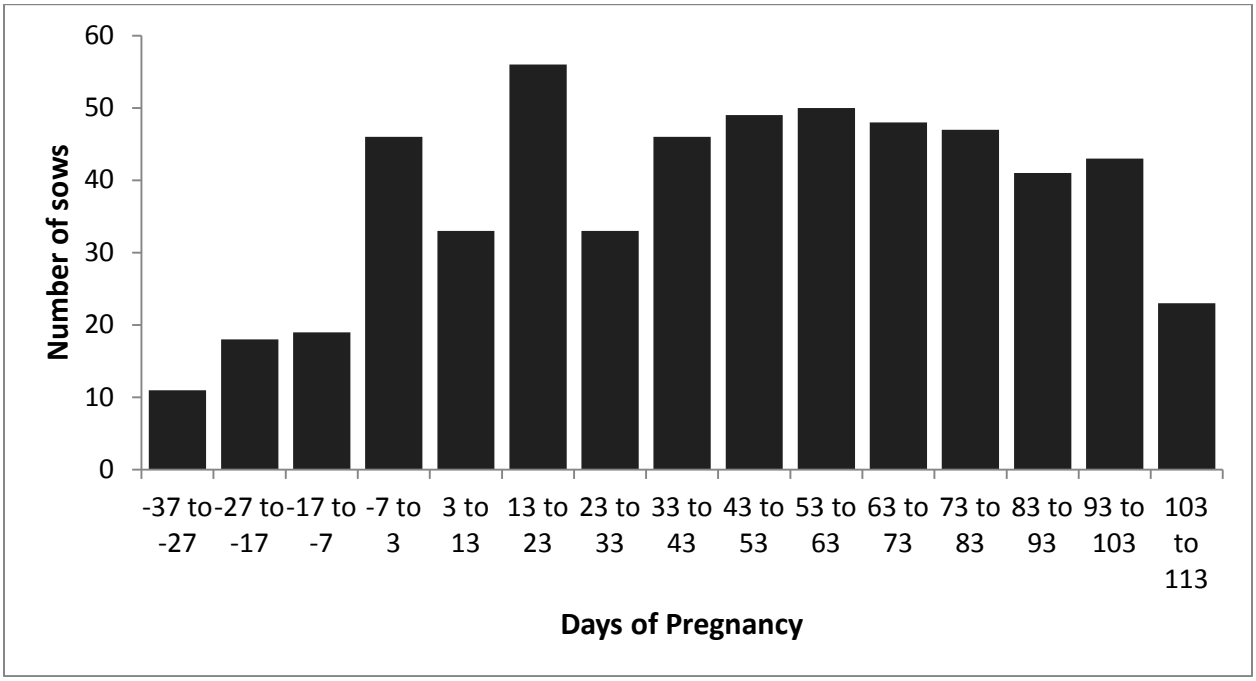

Figure 5 Repeatability of scores across different observers for the 11 scoring events, using data for all observers present at each event. Repeatability estimates were obtained using non-linear mixed models for ordinal data. A separate model was run for each scoring event.