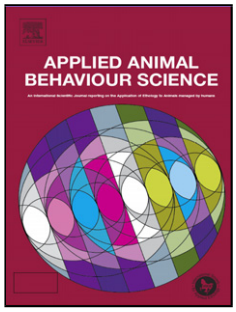

Title: Are observer ratings influenced by prescription? A comparison of free choice profiling and fixed list methods of qualitative behavioural assessment

Author: Taya Clarke John R. Pluske Patricia A. Fleming

PII: S0168-1591(16)30016-8
DOI: http://dx.doi.org/doi:10.1016/j.applanim.2016.01.022
Reference: APPLAN 4200

To appear in: *APPLAN*

Received date: 27-3-2015
Revised date: 19-1-2016
Accepted date: 20-1-2016

Please cite this article as: Clarke, Taya, Pluske, John R., Fleming, Patricia A., Are observer ratings influenced by prescription? A comparison of free choice profiling and fixed list methods of qualitative behavioural assessment.Applied Animal Behaviour Science http://dx.doi.org/10.1016/j.applanim.2016.01.022

Applied Animal Behaviour Science

Are observer ratings influenced by prescription? A comparison of free choice profiling and fixed list methods of qualitative behavioural assessment

Taya Clarke[a], John R. Pluske[a], Patricia A. Fleming[a]*

[a] Animal Production, Health and Welfare, School of Veterinary & Life Science, Murdoch University, WA 6150, Australia

* Trish Fleming *t.fleming@murdoch.edu.au* Tel: +61 8 93606577

Running head: comparing two methods of rater scales for behaviour assessment

Word count: 5,300 words

Highlights:

- Reliability of rater assessments depends on understanding how observers apply descriptive terms.
- We compared two methodologies, Fixed List (FL) and Free Choice Profiling.
- Observers reached consensus using either FL or FCP methods.
- There were correlations in scores attributed to groups of sows between FL and FCP.
- Training is an important aspect of reliability of rater assessments.

Abstract. Qualitative methods of behavioural assessment use observer rating scales to score the overall demeanour or body language of animals. Establishing the reliability of such holistic approaches requires test and validation of the methods used. Here, we compare two methodologies used in Qualitative Behavioural Assessment (QBA): Fixed-Lists (FL) and Free-Choice Profiling (FCP). A laboratory class of 27 students was separated into two groups of 17 and 10 students (FL and FCP respectively). The FL group were given a list of 20 descriptive terms (used by the European Union's Welfare Quality® program), shown videos of group-housed sows, and as a group discussed how they would apply the descriptive terms in an assessment. The FCP group were shown the same footage but individually generated their own descriptive terms to describe body language of the animals. Both groups were then shown 18 video clips of group-housed sows and scored each clip using a visual analogue scale (VAS) system. We analysed the VAS scores using Generalised Procrustes Analysis (GPA) for each observer group separately, which indicated high inter-observer reliability for both groups (FL: 71.1% of scoring variation explained, and FCP: 63.5%). There were significant correlations between FL and FCP scores (GPA dimension 1: r_{16}=0.946, P<0.001, GPA dimension 2: r_{16}=0.477, P=0.045). Additional analysis of the

raw VAS scores for the FL group by Principal Component Analysis (PCA) produced four factors; PC1 scores were correlated with GPA1 (r_{16}=0.984, P<0.001) and PC3 scores correlated with GPA2 (r_{16}=0.880, P<0.001). Kendall's coefficient of concordance (a measure of observer agreement) of the VAS scores indicated statistically significant agreement in use of the 20 descriptive terms (W range 0.37–0.64; all significant at P<0.001, although a value of W >0.7 is usually accepted to show strong agreement). This study demonstrates that, regardless of whether they are given their terms or are allowed to generate their own, observers score sow body language in a similar way. Strengths and weaknesses within the two methods were identified, which highlight the importance of providing thorough and consistent training of observers, including providing good quality training footage so that the full repertoire of demeanours can be identified.

Keywords: Sows; QBA; Free Choice Profiling, Generalised Procrustes Analysis (GPA), Fixed Lists, Visual Analogue Scale

Introduction

Behaviour is the outward expression of physiological changes, and since it incorporates aspects of animal perception, cognition and emotions, inclusion of behavioural assessment has been widely recognised as important in the development of future welfare measures (Mellor, 2012; Fleming et al., 2015a). Qualitative Behavioural Assessment (QBA) makes use of observers to score animals against descriptive terms of the animals' qualitative behavioural expression using visual analogue scales (VAS). Behavioural expression is not *what* the animals are doing, but *how* it goes about doing what it is doing (Wemelsfelder et al., 2000; 2001). QBA therefore captures the body language or demeanour of an individual in a dynamic, integrated measure (Stevenson-Hinde, 1983; Feaver et al., 1986; Wemelsfelder, 1997; 2007). QBA has been validated for a variety of animal

species across a range of experimental and on-farm treatments (reviewed by Fleming et al., 2015a). Two main procedures have been widely used: Free-Choice Profiling (FCP) and the use of Fixed Lists (FL) of terms. Each of these methods has benefits and limitations in terms of how they can be applied to welfare assessments, and careful review of the methods will help to clarify their use (Fleming et al., 2015a).

The FCP methodology allows multiple observers to generate their own unique terms to describe behavioural expression, and then use these terms to score the behaviour of a group of animals that all observers watch in common (either watching the animals at the same time, or being shown the same footage). These data are analysed by Generalised Procrustes Analysis (GPA), which identifies common patterns in the use of descriptive terms, followed by a Principal Components Analysis (PCA) to simplify the data into two or three main dimensions. QBA scores generated through the FCP methodology have been validated against physiological parameters (Stockman et al., 2011; 2012; Wickham et al., 2012; 2015) as well as quantitative behavioural measurements (Napolitano et al., 2008; Stockman et al., 2014).

An alternative approach is to have all observers use the same set of descriptive terms. Species-specific lists of terms (FL) have been developed through a process of experimentation and consultation with expert opinion to reach consensus regarding behavioural terms that are relevant for particular housing systems (e.g. Wemelsfelder et al., 2009a; Wemelsfelder and Millard, 2009; Wemelsfelder et al., 2009b). These FL should therefore ideally allow different inspectors, working independently, to use the same scales of qualitative behavioural expression in their assessments of different farms. Using FLs, QBA has been included as one of 13 measures as part of the 2004-2009 European Commission's Welfare Quality® audit (Temple et al., 2011a; Andreasen et al., 2013). Importantly, QBA was the only measure which captured positive welfare. Analysis of FL scores is

carried out by PCA, and therefore relies on observers having similar interpretation of the descriptive terms.

Recent reviews suggest that observer-based methods can be robust and perform a useful task in scientific investigations (Meagher, 2009; Whitham and Wielebnowski, 2009), although other authors have questioned the validity of rater assessments due to a lack of inter-observer agreement (Bokkers et al., 2012). Wemelsfelder et al. (2012) demonstrated consistency and agreement in how pig farmers, veterinarians and animal activists scored pig behaviour, but by contrast, Duijvesteijn et al. (2014) reported that pig farmers observed the behaviour of pigs in a more positive way compared to animal scientists and urban citizens. Duijvesteijn et al. (2014) used a FL methodology (the same descriptive terms as the present study) while Wemelsfelder et al. (2012) used FCP. It is argued that these methodological differences accounted for different outcomes of these two studies (Duijvesteijn et al., 2014).

To investigate and develop the role of QBA as an objective measure of animal welfare, validation in terms of observer reliability, cross-validation against other methods, and understanding sensitivity of QBA to experimental treatment is necessary (Meagher, 2009; Wemelsfelder et al., 2009c; Bokkers et al., 2012; Tuyttens et al., 2014; Fleming et al., 2015b). The aim of the present study was to compare the outcomes for observers using the FCP or FL methods. A direct comparison of the inter-observer reliability of these two methodologies has yet to be undertaken, and this study is therefore important in understanding the value and caveats of each experimental method.

Methods

Animals and experimental design

This study was carried out at a commercial breeder farm in Western Australia, made up of 1,200 parity 1-9 sows of Landrace/Large White breed. Sows were moved into group pens of n=10 sows at 5 days after their last service, and then 3 weeks later were moved again into larger pens of approximately n=15-18 individuals, where they were kept until 1 week prior to their expected farrow date. All sow tag numbers and parity were recorded, although due to the nature of the recording system at the piggery, we were unable to determine whether any of the sows had been housed together previously.

Each of the group pens was retrofitted from old mating/gestation stalls. The group pens were 1.6 x 0.55 m (length x width) in dimension and provided an overall space allocation of ~1.8m^2 per sow. Water was provided *ad libitum* (in the stalls) via a trough that ran the length of the pen, and food was deposited by an automatic feeder once a day (although no food was provided at the time of video recording).

For 18 groups re-mixed at 3 weeks post-mating, continuous video footage (15 fps; Panasonic SDR-H250 camcorders, Panasonic Belrose, NSW, Australia) was collected for the first 90 minutes from the time of mixing. Half of these sow groups had been mixed into pens with a concrete partition (a short wall, 2 m long and 1.6 m high) and the other half into pens without the partition. Our previous study (Clarke et al., In review) revealed the greatest differences in demeanour between these two treatment groups at 50-60 min post-mixing; footage from this time window was therefore chosen for the present study. Footage was edited (Adobe Premier Pro CS3 and Adobe After Effects CS3, Chatswood, NSW, Australia) to produce video clips of 1-min duration to show to observers for QBA.

A series of 12 clips from the same farm were used as the training video. The 12 clips showed a repertoire of behaviour and were designed to capture and familiarise observers with a wide range of sow demeanours. The treatment video was made up of 18 video clips (one per sow group), with clips from the two housing treatments (n=9 partition and n=9 no partition) randomly ordered. Observers were not made aware of these treatment differences and the footage did not explicitly focus on the partition.

Twenty-seven observers were recruited for this study as part of a laboratory session teaching Animal Science (n=22) and Veterinary Science (n=5) students about assessment methods for pig behaviour. These observers were all naïve to the QBA assessment method. Within the group there were 22 females and five males. Twenty-four of the 27 participants had been to farms that reared livestock and 11 had visited a pig farm. Twenty-four had pets of their own. Three rated their knowledge of animal welfare issues in Australia as strong, seven as good, 16 as average, and one as poor. Twenty-two students thought that animal welfare could be improved in Australia, one did not, and four were unsure.

The class was randomly split into two groups (which differed in numbers due to logistics – we had to exclude some observers because they had not completed their assessment). Seventeen students used the Fixed List (FL) method and 10 the Free Choice Profiling (FCP) method.

Phase 1: Term generation and training

Fixed list (FL). The FL method requires all observers to use the same set of descriptive terms, which requires that they are trained to recognise these terms in common. The 17 students were provided with a list of the 20 terms used by Welfare Quality® for assessment of pigs (Wemelsfelder and Millard, 2009). The terms were arranged to ensure that contrasting terms were placed adjacent to each other and were (in order of appearance): active, relaxed, fearful, agitated, calm,

content, tense, enjoying, frustrated, sociable, bored, playful, positively-occupied, listless, lively, indifferent, irritable, aimless, happy and distressed. The observers watched the training video and were asked to verbalise which term described the behaviour they saw. If a term from the list had not been used by half way through this session, it was identified and observers were asked to specifically look for that behaviour in the clips. Observers were then asked to discuss how they interpreted each of the terms so that all observers had a common understanding within the group. These descriptive terms were presented on an Excel spreadsheet (Microsoft Excel 2003, North Ryde, NSW, Australia) where each of the 20 terms was placed adjacent to a 100-mm visual analogue scale (VAS) for scoring in the subsequent quantification session.

Free Choice Profiling (FCP). The FCP methodology allows observers to generate and use their own descriptive terms. For this experiment, these 10 students were shown the training video, and at the end of each clip, they were given as much time as they needed to write down descriptive terms they felt suitably described the expressive qualities of the observed animals. These descriptive terms were copied into lists (unique lists for each individual observer) and the duplicates were removed. Their respective terms were copied into an Excel spreadsheet where each term was placed adjacent to a 100-mm VAS for scoring in the subsequent quantification session. Due to time-constraints, we did not have time to edit any unsuitable terms that the observers may have developed (e.g. hungry) but these were eliminated before the data were analysed.

Phase 2: Quantification

All observers were brought back into one room after a short break. They were then instructed to score the 18 treatment video clips using their respective rating scales; they were told to think of the distance between the zero-point and their mark on the VAS as reflecting the intensity of the sows expression on each descriptive term.

Statistical Analysis of FL and FCP scores

Fixed list and FCP scores: Both the FL data and the FCP data were each analysed separately using Generalised Procrustes Analysis (GPA) (Wemelsfelder, 2007). GPA is a multidimensional analysis that re-scales data that are scored on different dimensions so that they can be directly compared. Because each observer scores the same footage, GPA captures the similarity in scoring patterns between observers, and outputs a 'best fit' profile or consensus between observer assessments. The level of consensus (i.e. the percentage of variation between observers explained) achieved is expressed as the Procrustes Statistic. Whether this consensus is a significant feature of the data set (or an artefact of the Procrustean calculation procedures) is determined by comparison of the Procrustes Statistic with a randomisation test (Dijksterhuis and Heiser, 1995). This procedure rearranges at random each observer's scores and produces new permutated data matrices. By applying GPA to these permutated matrices, a 'randomised' profile is calculated. This procedure is repeated 100 times, providing a distribution of the Procrustes Statistic revealing how likely it is to find an observer consensus based on chance alone. Subsequently a one-sample t-test was used to determine whether the actual observer consensus profile falls significantly outside the distribution of randomised profiles. The number of dimensions of the GPA consensus profile is reduced to several main dimensions (usually two or three) through Principal Components Analysis (PCA). Interpretation of these consensus dimensions is made possible by selecting terms for each observer that correlated strongly with the consensus dimensions (>75% of the highest absolute correlation coefficient value, Mardia et al., 1979). Each clip received a quantitative score on each of the first two GPA consensus dimensions.

Fixed list scores only: To determine inter-observer reliability, the VAS scores for the FL group were correlated using Kendall's coefficient of concordance (*W*-coefficient). This test is used for expressing inter-rater agreement amongst judges who are seeing the same thing, and has been

used previously to compare the ratings of observers for other QBA studies (Wemelsfelder and Millard, 2009; Bokkers et al., 2012; Phythian et al., 2013). A Principal Component Analysis (PCA) was carried out for the VAS scores for each individual clip scored by each observer. We then carried out mixed-model ANOVA of BoxCox-transformed PCA scores for each of the four PCA dimensions that had an Eigenvalue >1, with observer ID as a random factor and clip (animal number) as a fixed factor. For the FL group data, the BoxCox-transformed PCA factor scores were compared with the GPA dimension scores by Pearson's correlation matrix.

A Box-Cox transformation was applied to all GPA and PCA dimension scores to meet the assumption of a normal distribution (Kolmogorov-Smirnov test). GPA analyses were carried out in Genstat 10.2 (VSN International, Hemel Hempstead, Hertfordshire, UK); all other analyses were carried out in Statistica 8.0 (StatSoft-Inc, Tulsa, OK, USA). Statistical significance was considered where $\alpha<0.05$.

Results

Fixed Lists

Kendall's coefficient of concordance (W) tests indicated statistically significant concordance in the use of all 20 of the FL descriptive terms ($P<0.001$) (Table 1). The Kendall's W values ranged from 0.37 ('sociable') to 0.64 ('happy').

Table 1 describes how each of the fixed list terms was weighted on the PCA components. The PCA generated four main factors that had Eigenvalues >1.0. Terms that weighted most strongly for the first Principal Component (PC1; 37.58% of the variation in the groups' behavioural expression data) were 'calm', 'relaxed' and 'content' at the low end of the dimension axis, and 'tense', 'agitated' and 'irritable' at the high end of the dimension axis. PC3 (14.03% of data variation) ranged from 'bored', 'listless', and 'aimless' on the low end of the axis, and 'active',

'lively', and 'sociable' on the high end of the axis. PC4 (5.64% of data variation) consisted of terms 'enjoying' and 'happy' on the low end of the axis and 'sociable' and 'playful' on the high end of the scale. All of these dimensions showed significant differences in scores (mixed-model ANOVA) between clips ($p<0.001$) in addition to significant differences between observers ($p<0.001$). By contrast, PC2 (17.89% of data variation) was positively correlated with all descriptive terms. There were significant differences in PCA Factor 2 scores between observers ($p<0.001$) in terms of how they used their scoring sheets, but there were no significant differences between clips (animals) on this dimension ($p=0.339$). PCA Factor 2 therefore largely captured differences between observers in respect to how they used the visual analogue scales: whether their scores were generally clustered around lower, mid, or upper values, or observers used the whole range of scores.

The GPA consensus profile for the FL group explained 71.14% of the variation in the observer scoring patterns, which were significantly different from the mean randomised profile ($t_{99}= 48.33$, $P<0.001$) (Table 2). Two main GPA dimensions explained a cumulative total of 63.49%, with GPA dimension 1 contributing 57.4% and GPA dimension 2 contributing 9.9% (Table 4). The terms correlated with each of the dimensions are summarised in Table 3. For GPA dimension 1, terms 'calm', 'content', and 'relaxed' were correlated with low values of the axis and 'agitated', 'tense', and 'irritable' with high values. Terms associated with GPA dimension 2 were 'listless', 'indifferent', and 'aimless' and on the low end of the axis and 'active', 'positively_occupied', and 'sociable' and on the high end of the axis.

There was a strong correlation between the GPA dimension 1 scores and PC1 ($r_{16}=0.984$, $P<0.001$). These two measures also weighted the terms 'agitated/tense/irritable' vs. 'calm/content /relaxed' in the same way (Table 3). GPA dimension 2 scores correlated strongly with PC3 ($r_{16}=0.880$, $P<0.001$) and also weighted the terms 'listless/indifferent' vs. 'active/sociable' in the same way. There was no significant correlation with GPA dimension 2 and PC4 ($r_{16}=0.309$, $P>0.05$).

Free Choice Profiling

The GPA consensus profile for the FCP group explained 63.49% of the variation in observer scoring patterns, which differed significantly from the mean randomised profile (t_{99} = 17.14, P<0.001; Table 2). Two main GPA dimensions explained 65.3% of the variation in scores attributed to individual clips of sows; 49.8% of variation in the data was attributed to GPA dimension 1 and 15.5% to GPA dimension 2. The terms correlated with each of the dimensions are summarised in Table 3. GPA dimension 1 was characterised by terms such as 'calm', 'relaxed' and 'docile' on the low end of the axis, and 'aggressive', 'agitated' and 'restless' at the high end. GPA dimension 2 (14.5% of the variation) was characterised by terms such as 'lethargic' and 'tired' at the low end of the axis and 'calm' and 'happy' at the high end.

Comparison between FL and FCP scores

Pearson's correlation showed a significant positive relationship between the FL and FCP BoxCox-transformed scores attributed to the 18 individual video clips on GPA dimension 1 (r_{16}=0.946, P<0.001) and GPA dimension 2 (r_{16}=0.477, P=0.045) (Fig. 1).

Discussion

Free Choice Profiling (FCP) is a method of qualitative behavioural assessment that allows observers to measure the body language of animals in an expressive way using their own terms. However, FCP may not necessarily be appropriate for on-farm welfare assessment since each observer uses a different set of terms to score (Wemelsfelder and Millard, 2009; Fleming et al., 2015a). Alternatively, Fixed Lists (FL) of descriptive terms have the advantage of providing a standardised list of descriptive terms and assessment can therefore be performed by trained assessors that work independently. We found strong positive correlations in qualitative behavioural assessment scores for observer groups that simultaneously scored footage of sows

using either FL or FCP methods, indicating that both approaches enabled our observer groups to adequately capture and score the body language of group-housed sows. Both methods therefore have value for use in animal welfare assessment.

There were a number of terms developed by observers using FCP that overlapped with the FL of terms derived from the Welfare Quality® audit for assessment of pigs (Wemelsfelder and Millard, 2009). Where these descriptive terms were strongly correlated with the GPA dimension axes, they appeared in the same ends of the dimension axes as they had for the FL observers' data. GPA dimension 1 generally demonstrates a valence of mood with 'calm/relaxed' vs. 'agitated' common to both FL and FCP methods. GPA dimension 2 also had semantically-consistent terms for both FL and FCP, and was reflective of general energy level. Valence and arousal have similarly been reported on the first two dimensions of QBA analyses in other species (e.g. pigs, Temple et al., 2011b; sheep, Phythian et al., 2013).

The 20 fixed-list terms used in this study were designed in conjunction with pig experts to identify a range of body language expression (Wemelsfelder and Millard, 2009). It is possible that sets of terms may have to be generated for each type of application. The same issue needs to be considered for the FCP methodology, where it is important to provide observers with a wide range of footage/experiences during the term generation training to ensure that they generate terms that will be relevant for particular aspects of behaviour during the quantification sessions. For example, in a study of horses during a 160km long endurance ride, some observers did not generate terms that would enable them to capture tiredness ('tired/lazy/sleepy' vs. 'alert/curious/excited') in the horses (Fleming et al., 2013). The range of descriptive terms given to observers is therefore likely to influence whether they can detect reasonably subtle behavioural patterns. The appropriateness of descriptive terms should also be considered. For example, the term 'playful' is relevant to piglets

and grower pigs, but can have has less value for intensively-housed adult sows, where it could simply lead to noise in observer scoring data.

A perceived issue with FL assessments is that observers given a prescribed set of descriptive terms may not interpret each term with the same meaning as the next person. The training session is important to adequately discuss and refine a person's perception of the terms to ensure that they are using the descriptive terms in the same manner. The quality of the training method as well as the quality of the training/term generation footage may also impact on how the observers perform in this respect. Wemelsfelder et al. (2009b) initially reported poor agreement for terms, and therefore re-ran their assessment using different video footage; this subsequently improved the inter-observer reliability. The potential for discrepancy in scoring behaviour for the FL observers is best avoided by checking the inter-reliability during the training session and addressing any discrepancies accordingly by ensuring that all observers understand the terms (Wemelsfelder et al., 2009b). Despite these potential pitfalls, we report similar degrees of inter-observer reliability ('consensus') for FL and FCP groups (where consensus dimensions accounted for 71.14% and 63.49% of variation in the datasets, respectively) and strong positive correlations in scoring patterns for the FL and FCP groups, suggesting that observers can reach consensus, regardless of whether they were prescribed terms or they were allowed to generate their own.

The common method of analysing FL scores is using PCA (Meagher, 2009). However, PCA requires that the data are measured on the same scales, and therefore we need to make the assumption that individual observers are interpreting and using their FL set of descriptive terms in the same manner. By contrast, GPA seeks common elements of scoring patterns and therefore does not assume that the scores for each descriptive term have the same scale or range. GPA is not commonly used for data analysis in the FL methodology, but we compared PCA and GPA handling of the same QBA data to compare the outcomes of these statistical analyses. We found anomalous

results for PCA Factor 2, and investigation indicated that this dimension captured differences in scoring patterns between observers – where some observers generally clustered their VAS marks at the low, middle or high end of the scales, while others used the entire VAS range. PCA Factor 2 therefore revealed differences in observer scoring patterns that GPA implicitly dealt with through its re-scaling calculations. GPA analysis of FL datasets offers the advantage of not having to assume that observers are using their FL terms in the same way, as well as dealing with these observer-related scaling artefacts.

The Kendall's coefficient of concordance performed on the FL observer scores produced W values that indicated reasonable degree of inter-observer reliability. Although none of the terms had W>0.7, terms scored the most consistently among the observers (W>0.5) were 'happy', 'enjoying', 'indifferent', 'aimless', 'fearful', 'frustrated', and 'bored'. There was only one term with W <0.40, that being 'sociable'. The ways terms are scored should generally improve with greater understanding of the meaning of the word for individual observers. Bokkers et al. (2012) reported large intra-observer variation, although further analysis of the data presented in their study actually indicates consistency in scoring over time: if an observer group scored the terms low in their first session, then they also scored them low in the second session, or conversely, those observers that scored terms highly for session 1 also scored highly for session 2. In addition, the Kendall's W doubled for the terms 'content', 'distressed' and 'positively_occupied' from the first session to the second session for an experienced observer group (Bokkers et al., 2012). This suggests that there can be improvement in intra-observer reliability for more experienced observers, again highlighting the importance of thorough and consistent training of observers.

Several prior investigations using the FL method have similarly reported significant inter-observer reliability in sheep (Phythian et al., 2013) dairy cattle (Andreasen et al., 2013), pigs (Wemelsfelder and Millard, 2009) and cattle (Sant'Anna and Paranhos da Costa, 2013). Conversely,

Bokkers et al. (2012) reported low inter-observer reliability between trained and inexperienced observers scoring dairy cattle behaviour. Our Kendall's W scores are not vastly different from those of Bokkers el al. (2012) and Wemelsfelder and Millard (2009), suggesting that it is not the terms themselves that affect the reliability of the FL methodology, but rather the way the VAS scoring data are handled (e.g. PCA vs. GPA) and interpreted. Wemelsfelder et al. (2009b) showed a lack of agreement for separate descriptors using Kendall's W, but PCA identified commonality in observer scoring behaviours. This supports the holistic QBA approach where welfare assessment does not rely on single behaviours or descriptive terms on their own, but rather through identifying common scoring patterns and relationships in whole-body demeanour and qualitative expression.

Conclusion

In conclusion, we found little evidence that observer ratings of sow behaviour were influenced by use of free choice profiling or fixed list methods. We found similar significant inter-observer reliability for FL and FCP groups and significant agreement in GPA dimension 1 and 2 scores between the two observer groups, suggesting that observers in each group scored the body language of the sows in a similar way. Clearly, therefore, selection of either FL or FCP methodology cannot explain discrepancies in conclusions about reliability of observer scores between previous QBA studies (Wemelsfelder et al., 2012; Duijvesteijn et al., 2014). Greater understanding of the methodological and statistical procedures will help elucidate the benefits and potential issues associated with using either FL or FCP methods for qualitative behavioural assessment.

Acknowledgements

This project was funded by Australian Pork Limited and Murdoch University. Appreciation is extended to the students who participated in this study, to the managers and staff on the

commercially-run piggery, and to Dr Sarah Wickham for assistance in the session. Research was approved by the Murdoch University Animal Ethics Committee (permit number O2441/11) and the Murdoch University Human Ethics committee (Permit number 2008/021).

References

Andreasen, S.N., Wemelsfelder, F., Sandøe, P., Forkman, B., 2013. The correlation of Qualitative Behavior Assessments with Welfare Quality® protocol outcomes in on-farm welfare assessment of dairy cattle. Applied Animal Behaviour Science 143, 9-17.

Bokkers, E.A.M., de Vries, M., Antonissen, I.C.M.A., de Boer, I.J.M., 2012. Inter-and intra-observer reliability of experienced and inexperienced observers for the Qualitative Behaviour Assessment in dairy cattle. Animal Welfare 21, 307-318.

Clarke, T., Bryant, G.L., Pluske, J.R., Collins, T., Miller, D.W., Fleming, P.A., In review. A quantitative and qualitative approach to the assessment of behaviour of sows upon mixing into group pens with or without a partition. Animal Production Science Submitted 11-Mar-2015.

Dijksterhuis, G.B., Heiser, W.J., 1995. The role of permutation tests in exploratory multivariate data analysis. Food Quality and Preference 6, 263-270.

Duijvesteijn, N., Benard, M., Reimert, I., Camerlink, I., 2014. Same pig, different conclusions: stakeholders differ in qualitative behaviour assessment. Journal of Agricultural and Environmental Ethics 27, 1019-1047.

Feaver, J., Mendl, M., Bateson, P., 1986. A method for rating the individual distinctiveness of domestic cats. Anim. Behav. 34, 1016-1025.

Fleming, P.A., Clarke, T., Wickham, S.L., Stockman, C.A., Barnes, A.L., Collins, T., Miller, D.W., 2015a. Qualitative behavioural assessment as a welfare assessment method for Australian livestock industries. Animal Production Science In Press, Accepted Oct 2015.

Fleming, P.A., Paisley, C., Barnes, A.L., Wemelsfelder, F., 2013. Application of Qualitative Behavioural Assessment to horses during an endurance ride. Applied Animal Behaviour Science 144, 80-88.

Fleming, P.A., Wickham, S.L., Stockman, C.A., Verbeek, E., Matthews, L., Wemelsfelder, F., 2015b. The sensitivity of QBA assessments of sheep behavioural expression to variations in visual or verbal information provided to observers. Animal 9, 878-887.

Mardia, K.V., Kent, J.T., Bibby, J.M., 1979. Multivariate Analysis. Academic Press, London.

Meagher, R.K., 2009. Observer ratings: Validity and value as a tool for animal welfare research. Applied Animal Behaviour Science 119 1-14.

Mellor, D.J., 2012. Animal emotions, behaviour and the promotion of positive welfare states. New Zealand Veterinary Journal 60, 1-8.

Napolitano, F., De Rosa, G., Braghieri, A., Grasso, F., Bordi, A., Wemelsfelder, F., 2008. The qualitative assessment of responsiveness to environmental challenge in horses and ponies. Applied Animal Behaviour Science 109, 342-354.

Phythian, C., Michalopoulou, E., Duncan, J., Wemelsfelder, F., 2013. Inter-observer reliability of Qualitative Behavioural Assessments of sheep. Applied Animal Behaviour Science 144 73-79.

Sant'Anna, A.C., Paranhos da Costa, M.J.R., 2013. Validity and feasibility of qualitative behavior assessment for the evaluation of Nellore cattle temperament. Livestock Science 157, 254-262.

Stevenson-Hinde, J., 1983. Individual characteristics: a statement of the problem, in: Hinde, R.A. (Ed.), Primate Social Relationships: An Integrated Approach, Blackwell Scientific Publications, Oxford, UK, pp. 28-34.

Stockman, C.A., Collins, T., Barnes, A.L., Miller, D.W., Wickham, S.L., Beatty, D.T., Blache, D., Wemelsfelder, F., Fleming, P.A., 2011. Qualitative behavioural assessment of cattle naïve and habituated to road transport. Animal Production Science 51, 240-249.

Stockman, C.A., Collins, T., Barnes, A.L., Miller, D.W., Wickham, S.L., Verbeek, E., Matthews, L., Ferguson, D., Wemelsfelder, F., Fleming, P.A., 2014. Qualitative behavioural assessment of the

motivation for feed in sheep in response to altered body condition score. Animal Production Science 54, 922-929.

Stockman, C.A., McGilchrist, P., Collins, T., Barnes, A.L., Miller, D.W., Wickham, S.L., Greenwood, P.L., Cafe, L.M., Blache, D., Wemelsfelder, F., Fleming, P.A., 2012. Qualitative behavioural assessment of cattle pre-slaughter and relationship with cattle temperament and physiological responses to the slaughter process. Applied Animal Behaviour Science 142, 125-133.

Temple, D., Dalmau, A., Ruiz de la Torre, J.L., Manteca, X., 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.

Temple, D., Manteca, X., Velarde, A., 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.

Tuyttens, F.A.M., de Graaf, S., Heerkens, J.L.T., Jacobs, L., Nalon, E., Ott, S., Stadig, L., Van Laer, E., Ampe, B., 2014. Observer bias in animal behaviour research: can we believe what we score, if we score what we believe? Anim. Behav. 90, 273-280.

Wemelsfelder, F., 1997. The scientific validity of subjective concepts in models of animal welfare. Applied Animal Behaviour Science 53, 75-88.

Wemelsfelder, F., 2007. How animals communicate quality of life: the qualitative assessment of behaviour. Animal Welfare 16, 25-31.

Wemelsfelder, F., Hunter, A.E., Paul, E.S., Lawrence, A.B., 2012. Assessing pig body language: agreement and consistency between pig farmers, veterinarians and animal activists. J. An. Sci. 90 3652-3665.

Wemelsfelder, F., Hunter, E.A., Mendl, M.T., Lawrence, A.B., 2000. The spontaneous qualitative assessment of behavioural expressions in pigs: first explorations of a novel methodology for integrative animal welfare measurement Applied Animal Behaviour Science 67, 193-215.

Wemelsfelder, F., Hunter, T.E.A., Mendl, M.T., Lawrence, A.B., 2001. Assessing the 'whole animal': a free choice profiling approach. Anim. Behav. 62, 209-220.

Wemelsfelder, F., Knierim, U., Schulze Westerath, H., Lentfer, T., Staack, M., Sandilands, V., 2009a. Qualitative Behaviour Assessment, in: Forkman, B., Keeling, L. (Eds.), Assessment of Animal Welfare Measures for Layers and Broilers. Welfare Quality® reports No. 9, Sixth Framework Programme, University of Cardiff, Cardiff, pp. 113-119.

Wemelsfelder, F., Millard, F., 2009. Qualitative Behaviour Assessment, in: Forkman, B., Keeling, L. (Eds.), Assessment of Animal Welfare Measures for Sows, Piglets and Fattening Pigs. Welfare Quality® reports No. 10, Sixth Framework Programme, University of Cardiff, Cardiff, pp. 213-219.

Wemelsfelder, F., Millard, F., De Rosa, G., Napolitano, F., 2009b. Qualitative Behaviour Assessment, in: Forkman, B., Keeling, L. (Eds.), Assessment of Animal Welfare Measures for Dairy Cattle, Beef Bulls and Veal Calves. Welfare Quality® reports No. 11, Sixth Framework Programme, University of Cardiff, Cardiff, pp. 215-224.

Wemelsfelder, F., Nevison, I., Lawrence, A.B., 2009c. The effect of perceived environmental background on qualitative assessments of pig behaviour. Anim. Behav. 78, 477-484.

Whitham, J.C., Wielebnowski, N., 2009. Animal-based welfare monitoring: Using keeper ratings as an assessment tool. Zoo Biol. 28, 545-560.

Wickham, S.L., Collins, T., Barnes, A.L., Miller, D.W., Beatty, D.T., Stockman, C.A., Blache, D., Wemelsfelder, F., Fleming, P.A., 2012. Qualitative behavioral assessment of transport-naïve and transport-habituated sheep. Journal of Animal Science 90, 4523-4535.

Wickham, S.L., Collins, T., Barnes, A.L., Miller, D.W., Beatty, D.T., Stockman, C.A., Blache, D., Wemelsfelder, F., Fleming, P.A., 2015. Validating the use of Qualitative Behavioural Assessment as a measure of the welfare of sheep during transport. Journal of Applied Animal Welfare Science 18, 269-286.

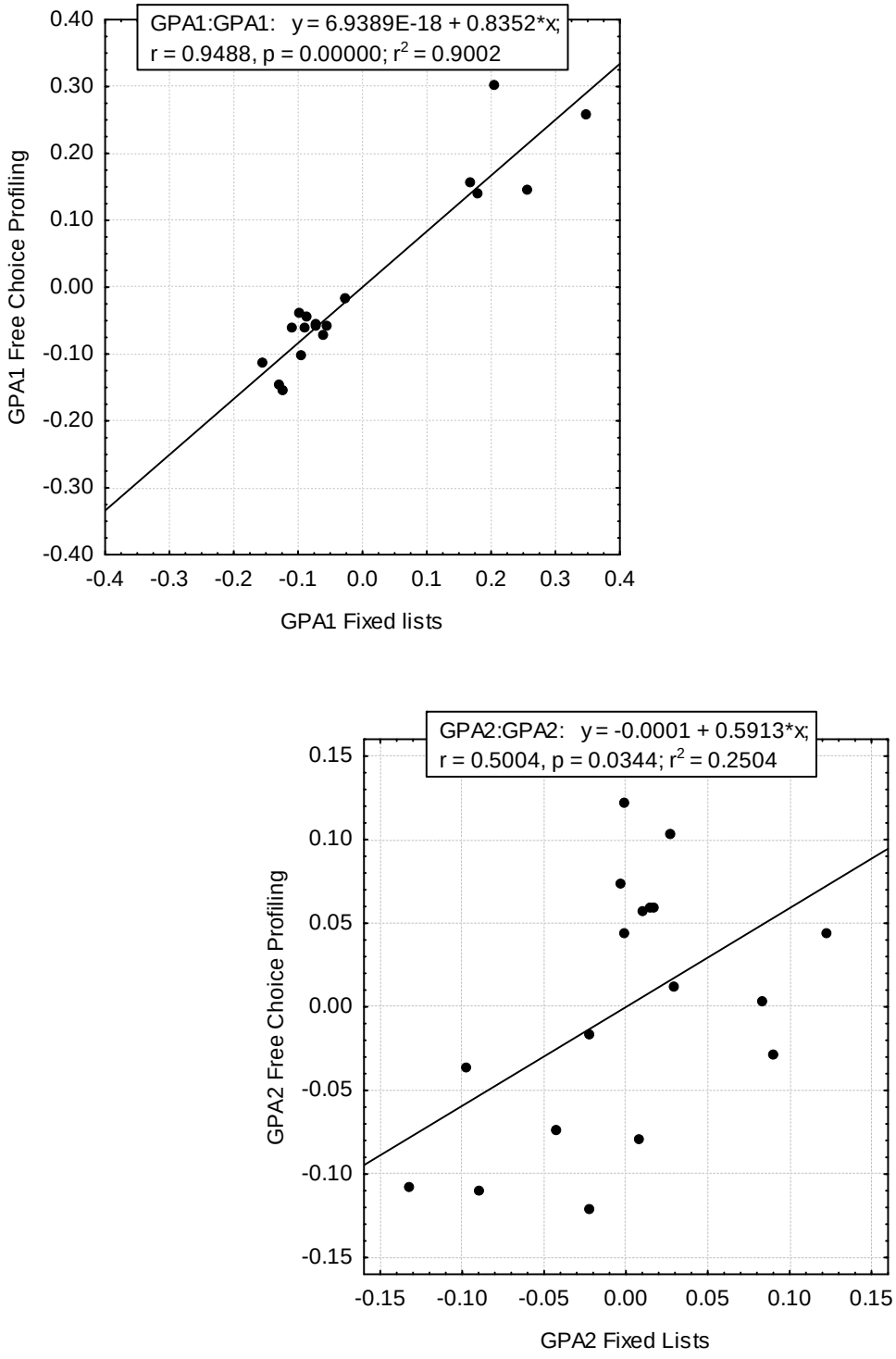

Fig. 1: Comparison between Free Choice Profiling and Fixed List methods for Generalised Procrustes Analysis (GPA) dimensions 1 and 2. Values are the GPA scores attributed by the two observer groups to video clips of n=18 groups of sows.

Table 1: Summary of the fixed list method results, comparing absolute values for observer scores on the 20 terms (Kendall's; *** p<0.001), a Principle Components Analysis (PCA) of the raw scores, and a Generalised Procrustes Analysis (GPA) of these data. Highlighted cells are those terms that strongly loaded on each behavioural dimension (>75% of the highest absolute correlation coefficient value).

Descriptive term	Absolute values for observer scores				Principle Components Analysis				Generalised Procrustes Analysis	
	Average	±SD	Kendall's W		PC Factor 1	PC Factor 2†	PC Factor 3	PC Factor 4	Average GPA1	Average GPA2
Active	39.18	25.43	0.46	***	0.33	0.45	0.53	0.31	0.51	0.39
Agitated	23.68	24.13	0.40	***	0.88	0.21	0.11	-0.12	0.81	-0.06
Aimless	38.24	25.89	0.53	***	0.03	0.45	-0.68	0.18	-0.19	-0.21
Bored	36.19	24.61	0.51	***	0.25	0.39	-0.70	0.21	0.09	-0.35
Calm	59.97	25.17	0.45	***	-0.84	0.34	-0.17	-0.11	-0.82	-0.01
Content	59.13	24.31	0.46	***	-0.78	0.41	-0.04	-0.21	-0.70	0.12
Distressed	16.10	21.01	0.47	***	0.84	0.28	0.00	-0.20	0.73	-0.14
Enjoying	44.97	23.33	0.55	***	-0.62	0.43	0.31	-0.33	-0.47	0.22
Fearful	14.92	17.53	0.53	***	0.82	0.26	-0.01	-0.20	0.69	-0.10
Frustrated	18.49	22.47	0.51	***	0.85	0.26	-0.02	-0.23	0.76	-0.18
Happy	39.89	25.86	0.64	***	-0.58	0.51	0.29	-0.32	-0.50	0.22
Indifferent	40.43	26.02	0.54	***	-0.09	0.55	-0.60	0.06	-0.19	-0.24
Irritable	20.48	24.68	0.41	***	0.87	0.31	0.02	-0.20	0.79	-0.12
Listless	28.65	25.07	0.46	***	0.03	0.41	-0.69	0.02	-0.10	-0.33
Lively	31.31	22.85	0.40	***	0.22	0.63	0.46	0.15	0.25	0.41
Playful	20.13	17.51	0.43	***	0.00	0.63	0.26	0.41	0.08	0.18
Positively_occupied	44.43	27.46	0.46	***	-0.39	0.53	0.19	-0.21	-0.38	0.20
Relaxed	55.03	26.26	0.41	***	-0.80	0.30	-0.13	-0.20	-0.71	-0.04
Sociable	40.91	22.87	0.37	***	-0.20	0.48	0.42	0.42	-0.16	0.27
Tense	22.52	25.45	0.46	***	0.89	0.25	0.03	-0.20	0.84	-0.10
Eigenvalue					7.52	3.58	2.81	1.13		
% total variance					37.58	17.89	14.03	5.64	57.4	9.9

† Mixed-model ANOVA indicated that there were significant differences in PCA Factor 2 scores between observers (p<0.001), but there were no significant differences between clips (animals) on this dimension (p=0.394); by contrast, all other PCA components showed significant differences between clips (p<0.001) in addition to significant differences between observers (p<0.001).

Table 2: Summary of the results for the Generalised Procrustes Analyses (GPA) for Fixed List (FL) and Free Choice Profiling (FCP) methods of Qualitative Behavioural Assessment (QBA). Numbers in brackets indicates are the number of participants in each observer group.

	Method	
	Fixed list (n=17)	FCP (n=9)
% Variation explained by GPA consensus	71.14	63.49%
t test results	48.33***	17.14 ***
% Variation explained by GPA1	57.4	49.8
% Variation explained by GPA2	9.9	15.5

*** $P<0.001$.

Table 3: Terms used by two groups of observers using either Free Choice Profiling (FCP) or Fixed Lists (FL) or terms to describe behavioural expression of sows filmed under group housing. Terms that correlate strongly with the Generalised Procrustes Analysis (GPA) consensus or Principle Components Analysis (PCA) dimensions (>75% of the highest absolute correlation coefficient value) are listed and term order is determined firstly by the number of observers to use each term (in brackets if > than 1), and secondly by weighting of each term (i.e. correlation with the GPA dimension). The last column shows a summary comparison between the Fixed List (FL) and PCA methods (Pearson's correlations between BoxCox-transformed values).

Observer group	Dimension (% of variation explained) minimum correlation	Low Values	High values	Correlation between FL GPA and PC (*r*)	P value
FCP	GPA1 (49.8%)	Calm (3), Relaxed (2), Docile, Tired	Aggressive (7), Agitated (4), Restless (3), Competitive (2), Alert, Annoyed, Angry, Aggravated, Irritated, Forceful, Anxious, Grumpy, Energetic, Dominant, Defensive,		
FL	GPA1 (57.4%)	Calm (13), Content (9), Relaxed (7), Happy (5), Positively_occupied (2), Enjoying (2), Aimless (2), Indifferent	Agitated (14), Tense (14), Irritable (12), Frustrated (10), Fearful (9), Distressed (9), Active (3), Listless, Lively, Playful, Bored		
	PCA Factor 1	Calm, relaxed, content	Tense, agitated, irritable, frustrated, distressed, fearful	0.984	<0.001
FCP	GPA2 (15.5%)	Lethargic (2), Tired (2), Interested, Playful, Lazy, Sleepy	Calm (2), Happy (2), Antisocial, Curious, Content, Relaxed, Stimulated, Fidgety, Active, Restless, Alert		
FL	GPA2 (9.9%)	Listless (3), Indifferent (2), Aimless (2)	Active (2), Positively_occupied (2), Sociable (2), Playful, Happy		
	PCA Factor 2†	-	All terms were positively correlated	0.536†	<0.05
	PCA Factor 3	Bored, Listless, Aimless, Indifferent	Active, Lively, Sociable	0.880	<0.001
	PCA Factor 4	Enjoying, Happy	Sociable, Playful	0.309	NS

† Mixed-model ANOVA indicated that there were significant differences in PCA Factor 2 scores between observers ($p<0.001$), but there were no significant differences between clips (animals) on this dimension ($p=0.394$); by contrast, all other PCA components showed significant differences between clips ($p<0.001$) in addition to significant differences between observers ($p<0.001$).