ANIMAL BEHAVIOUR, 2001, **62**, 209–220
doi:10.1006/anbe.2001.1741, available online at http://www.idealibrary.com on IDEAL®

Assessing the 'whole animal': a free choice profiling approach

FRANÇOISE WEMELSFELDER*, TONY E. A. HUNTER†, MICHAEL T. MENDL‡ & ALISTAIR B. LAWRENCE*
*Animal Biology Division, Scottish Agricultural College Edinburgh
†Biomathematics and Statistics Scotland, University of Edinburgh
‡Department of Clinical Veterinary Science, University of Bristol

*(Received 10 April 2000; initial acceptance 25 May 2000;
final acceptance 9 March 2001; MS. number: 6552)*

The qualitative assessment of animal behaviour summarizes the different aspects of an animal's dynamic style of interaction with the environment, using descriptors such as 'confident', 'nervous', 'calm' or 'excitable'. Scientists frequently use such terms in studies of animal personality and temperament, but, wary of anthropomorphism, are reluctant to do so in studies of animal welfare. We hypothesize that qualitative behaviour assessment, in describing behaviour as an expressive process, may have a stronger observational foundation than is currently recognized, and may be of use as an integrative welfare assessment tool. To test this hypothesis, we investigated the inter- and intraobserver reliability of spontanous qualitative assessments of pig, *Sus scrofa*, behaviour provided by nine naïve observers. We used an experimental methodology called 'free choice profiling' (FCP), which gives observers complete freedom to choose their own descriptive terms. Data were analysed with generalized Procrustes analysis (GPA), a multivariate statistical technique associated with FCP. Observers achieved significant agreement in their assessments of pig behavioural expression in four separate tests, and could accurately repeat attributing expressive scores to individual pigs across these tests. Thus the spontaneous qualitative assessment of pig behaviour showed strong internal validity under our controlled experimental conditions. In conclusion we suggest that qualitative behaviour assessment reflects a 'whole animal' level of organization, which may guide the intepretation of behavioural and physiological measurements in terms of an animal's overall welfare state.

The qualitative assessment of behaviour is based upon the integration by the observer of many pieces of information that in conventional quantitative approaches are recorded separately, or are not recorded at all. This may include incidental behavioural events, subtle details of movement and posture, and aspects of the context in which behaviour occurs. In summarizing such details of behaviour, qualitative assessment specifies not so much what an animal does, but how it does it, that is, its dynamic style of interaction with the environment (Stevenson-Hinde et al. 1980; Stevenson-Hinde 1983; Feaver et al. 1986; Fagen et al. 1997; Wemelsfelder 1997a, b). This form of assessment is widely used in the study of animal temperament and personality, and in this context tends to be retrospective. Observers recollect how individual animals have behaved over previous time periods, and sum up the behavioural style of these animals as, for example, 'bold/shy', or 'friendly/hostile'. Such qualifications designate 'traits' of one or more underlying dimensions of temperament or personality, which in turn are regarded as 'predispositions' for certain response styles. Thus 'boldness' and 'friendliness' are not primarily regarded as concrete behavioural phenomena immediately present for observation, but rather as abstract 'intervening variables' in causal mechanistic accounts of behaviour. In this view one does not see boldness, one infers it from behaviour (Plutchik 1980; Rutter 1987; Mendl & Harcourt 1988; Boissy 1995; Clarke & Boinski 1995; Gosling 2001).

We suggest however that qualitative assessments of behaviour may have a stronger observational (i.e. empirical) foundation than is currently recognized, and may have an important, but as yet unexplored, potential as an integrative welfare measurement tool. In summarizing aspects of an animal's behavioural style we describe behaviour as an expressive process, not only in retrospective abstraction, but also in direct observation of an animal at any given moment in time (Hebb 1946; Fagen et al. 1997; Wemelsfelder 1997b). For example, a ewe

Correspondence: F. Wemelsfelder, Animal Biology Division, SAC, Bush Estate, Penicuik EH26 0PO, Midlothian, U.K. (email: f.wemelsfelder@ed.sac.ac.uk). T. E. A. Hunter is at Biomathematics and Statistics Scotland, West Mains Road, Edinburgh EH9 3JZ, U.K. M. T. Mendl is at the Department of Clinical Veterinary Science, University of Bristol, Langford House, Langford, Bristol BS40 5DU, U.K.

separated from her lamb in the hills may walk about with her ears pricked up, looking out and bleating loudly, all the while appearing 'agitated', 'anxious' and 'distressed'. These terms characterize the animal's behavioural style as an observed expressive state, that is as a state apparently reflecting the animal's experience of the situation it finds itself in. The animal's expressive state may change: when the ewe finds her lamb, she will become calm and more relaxed. Thus qualitative assessments can capture fluctuations in behavioural expression that are otherwise hard to record, and seem naturally suited to summarize how a given condition may affect an animal's welfare state. Behavioural scientists are traditionally wary of such assessments, fearing they may be anthropomorphic judgments of uncertain validity (Kennedy 1992; Heyes 1993). In theory, however, it is possible that assessments of animal behavioural expression are based on observable aspects of behavioural organization, and are amenable to scientific analysis.

A first step in testing this hypothesis is to investigate the inter- and intraobserver reliability of qualitative assessments of animal behavioural expression. In a previous publication we proposed an experimental methodology specifically suited to this end (Wemelsfelder et al. 2000). This methodology, generally known as 'free choice profiling' (FCP), is widely used in food science, but has to our knowledge not been used in studies of animal behaviour before. Crucially, FCP allows observers complete freedom to generate their own descriptive terminology, rather than asking them to complete a predetermined rating scale. This spontaneity safeguards the integrative nature of the assessment, making it possible to determine whether individual observers have similar ways of integrating perceived behaviour into qualitative descriptors. FCP is associated with a multivariate statistical technique called generalized Procrustes analysis (GPA), which calculates observer agreement independently of fixed descriptors (see Methods for a detailed description). Our first exploratory FCP-based study found that 18 naïve observers showed significant consensus in their spontaneous qualitative assessments of pig, *Sus scrofa*, behavioural expressions (Wemelsfelder et al. 2000). However, in this study assessments were based on a 1/0 scoring system (i.e. the presence or absence of a perceived expressive trait) and did not allow quantitative measurement of perceived expressive traits. As a result the data resolution provided by observer terminologies was rather poor.

Our aim in the present study was to improve our FCP-based experimental procedures by fitting individual observer terminologies with a quantitative visual-analogue scale. We hypothesized that this would significantly increase data resolution, and provide us with a dependable basis to investigate the inter- and intraobserver reliability of spontaneous qualitative behaviour assessment. We instructed naïve observers to assess the behavioural expressions of growing female pigs with our improved FCP procedures. Observers repeated their assessment from two videorecordings, one showing pigs in the same order as the live session and one in a different order. These live and video assessments provided multiple tests of the inter- and intraobserver reliability of spontaneous qualitative behaviour assessment, allowing us to evaluate the internal validity of this methodology. If good internal validity were found, the proposed method of qualitative assessment could be used to investigate the biological basis of animal behavioural expression, and open up new avenues of behaviour and welfare research.

METHODS

Animals and Housing

Our experiment consisted of two phases, each phase using a different set of pigs of similar age and weight, and housed under identical conditions.

Each set of pigs consisted of 10 Large White × Landrace growing female pigs of around 17 weeks of age, and with weights ranging from 37 to 48 kg (average 43 kg) at the start of each stage of the experiment. The pigs were housed in an enclosure consisting of two identical, directly adjacent pens of 4 × 4 m. Both pens had solid walls and were visually isolated from each other by a solid partition 2 m high. A door in this partition allowed us to move pigs between pens. Each pen contained a layer of straw with some fresh branches, a drinker bowl and a food trough. To achieve maximum habituation to both pens and to being moved between pens, we housed pigs in the two pens on alternate days.

Throughout the experiment we gave pigs an ad libitum supply of food (Growlean E. R. Pellets, BOCM Pauls Ltd, Ipswich, U.K.) appropriate to their age. Food was provided at 0830 hours each day and pens were cleaned at ca. 0900 hours. We removed soiled straw and branches from the pens as necessary and replaced them with fresh material. Room temperature was maintained between 17 and 20 °C throughout the experiment.

Experimental Procedures

Animals

In the first week of phase 1 of the experiment we left the pigs to adjust to their housing regime as described above. In the second week they were trained daily to be separated from their penmates and to spend 7 min alone in the test pen directly adjacent to the home pen. During training the test pen was surrounded by wooden observation screens. Training schedules were balanced for treatment and time of day.

Testing took place in the third week on 2 consecutive days. On day 1 we let pigs from their home pen into the test pen singly and in random order. On day 2 we repeated this, but let pigs into the test pen in a different order from day 1. On both days, each pig had the opportunity to interact for 7 min with a human interactor sitting in the centre of the test pen. This human interactor was familiar to the pigs from the previous training sessions, but she had never sat down to interact with the pigs for any length of time, and the experimental situation was therefore in that sense new to them. The interactor consistently responded only to interactions elicited by the pig. If the pig looked at her or approached

she would extend a hand towards it. If the pig remained close and initiated further interaction, she would pat its nose, head or back, or extend her face towards the pig. If the pig became aggressive and inclined to bite she would push it off and remain passive until the pig again initiated interaction.

Three weeks later, in phase 2 of the experiment, we repeated this procedure with a new set of 10 pigs.

Observers

The observers were nine graduate students (six male and three female, all British except one female Canadian student), most of whom had experience with the observation of animals, but not with observing pigs. At the start of the study these observers were given ca. 1 h of instruction in a room adjacent to the experimental area. We told them that the experiment was part of a research programme aimed at developing a methodology for the assessment of behavioural expression in pigs. Behavioural expression was defined as style of interaction, that is, how an animal behaves as opposed to what it does, and observers were given a few examples to underline the behavioural character of this definition. We explained that an essential characteristic of the adopted methodology was that observers would generate their own descriptive terminology to score the behavioural expressions of the pigs. The FCP procedures used to facilitate the different stages of this process (described below) were outlined to observers in detail. They then proceeded into the experimental area, where they were seated around the test pen behind wooden screens to observe and describe the pigs as instructed. To ensure independence of assessment, we told them to refrain from discussing terms throughout the experiment.

Free choice profiling

The FCP procedures used in this study for the assessment of pig behavioural expression follow conventional FCP methods as used in food science (Arnold & Williams 1985). These methods consist of two phases. The first phase allows observers to generate their individual terminologies. The experimenter then provides these individual terminologies with a visual analogue scale, and subsequently in the second phase observers are instructed to use their personal terminologies as a quantitative measurement tool.

In line with this procedure, the assessment of pig behavioural expression consisted of two phases. In phase 1 (consisting of 2 days), observers generated their own set of descriptive terms while observing the first set of 10 individually presented pigs. On day 1 observers first watched each pig for 4 min and then (after a signal) used the following 3 min to write down terms that in their view best summed up the expressive qualities of that pig's behaviour. We pointed out that they were entirely free to choose new terms for each new pig, or use terms chosen for previous pigs, but that they should concentrate on choosing the best terms for each individual pig. Thus each observer compiled a set of terms describing the expressive repertoire of the 10 pigs. On day 2 observers practised using their individual terminologies as measurement tools. The 10 pigs observed were the same as day 1, but presented in a different order. Observers could not identify individual pigs, but on both days assessed the pigs in order of appearance. The experimenter had added a smooth line of 12.5 cm ranging from 'minimum' to 'maximum' to each observer term. Observers were told that (after the signal) they should score each pig on each of the terms of their personalized rating scales, by ticking the line at an appropriate point between 'minimum' and 'maximum'. They could add new terms to their terminology if they wished. At the end of day 2 observers were given 30 min to edit their terminologies (e.g. by removing synonyms) and make a final selection of terms. The quantitative scores generated during this practice were not used for further analysis.

In phase 2 of the experiment (which took place 3 weeks after phase 1, and also consisted of 2 days) observers were presented with a new set of 10 pigs. These pigs were housed and trained under identical conditions, and were the same age and weight as the first set of pigs. In this phase, observers used their previously generated personalized rating scales to provide the actual data for further analysis. On day 1 they were told as before to watch the pigs for 4 min, and then after a signal to score each pig on each of the terms of their rating scale. On day 2 the same 10 pigs were presented to observers in a different order, and this procedure was repeated. Thus in the second phase observers twice used their individual terminologies to measure the behavioural expressions of the same 10 pigs.

Video repeat

On day 1 of phase 2 (called 'Live1'), we used a digital Panasonic NV-DX1E camcorder to record the behaviour of the 10 pigs. The camcorder was mounted on a tripod at eye level and positioned at the side of the test pen where the majority of observers were seated, so that the video's perspective closely matched that of the observers during the live session. A microphone was suspended above the interactor's head to provide the recordings with sound. The behaviour of the pigs on day 2 of phase 2 (called 'Live2') was not recorded.

The digital tape of Live1 was edited in a professional studio, to produce two high-resolution S-VHS tapes to show to observers. One tape showed the pigs in the same order of appearance as Live1 ('Video1.1'), the other showed the pigs (again from day 1) in a different order ('Video1.2'). In both videos a 4-min signal was digitally imposed on the footage of each pig.

In phase 3 of the experiment these two videotapes were shown to observers on 2 consecutive days, 1 week after the Live sessions had come to an end. On both days observers were divided into two groups, each group watching the video on a widescreen TV monitor. We instructed them to observe the behaviour of each pig until a 4-min signal appeared on the screen, and then to use their personalized rating scales (the same as used in phase 2) to score the pigs' behavioural expressions. Thus in this third phase observers assessed twice from video the

behaviour of the 10 pigs they had previously assessed during 'Live1'.

Questionnaires

When the experiment came to a close at the end of phase 3, we gave the observers two questionnaires. In the first we asked them to evaluate the efficacy of video assessment against that of live assessment, while in the second we asked them to define each of the terms they had used in their personalized rating scale. We explained that these definitions should describe the criteria that had guided them in their use of terms during the various phases of assessment. What kind of criteria these should be was left unspecified; the purpose of this questionnaire was to find out what it is that observers perceive when describing the pigs' behavioural expressions (e.g. a bold pig is a pig that . . .).

Method of Analysis

General outline

At the end of the experiment observers had used their personalized rating scales to produce four sets of scores (Live1, Live2, Video1.1, Video1.2), all for the same pigs. We determined the score for each pig on each observer term by measuring the distance (mm) between the left 'minimum' point of the scale and the point where the observer's tick crossed the line. The four sets of scores thus obtained were entered into four sets of nine data matrices (one for each observer), with each matrix defined by the number of pigs (10), and the number of terms used by a particular observer.

To analyse these data matrices a multivariate statistical technique that does not rely on fixed variables was required. Generalized Procrustes analysis (GPA) is such a technique (Gower 1975; Arnold & Williams 1985; Gower & Dijksterhuis 1994; Wemelsfelder et al. 2000). GPA can be thought of as a pattern-matching mechanism, and is based on the assumption that even if observers use different variables (terms) for measurement, the distances between samples (pigs) will be comparable, because the samples are the same. In other words, GPA takes for granted that measurement patterns that deal with the same samples will converge, and is designed to compute the coordinates of the convergent configuration (the so-called 'consensus profile'). Thus GPA detects the level of consensus between observer assessment patterns not on the basis of fixed reference points (terms), but on the basis of the (multidimensional) intersample distances specified by each observer (i.e. how each observer uses his/her terms to score pigs). A second step is then to evaluate the statistical significance of the consensus profile through a randomization test (see below).

Statistical procedures

Interobserver reliability: comparison of nine data matrices within one set. To find the consensus of data matrices within one set, GPA assesses each matrix as a multidimensional configuration. Each matrix has as many dimensions as it has terms, and the position of the 10 pigs in this multidimensional space is defined by their scores. Columns of zeros are added to individual matrices, so that all observer configurations acquire equal dimensionality. The nine configurations thus obtained are then matched to each other by GPA through a series of iterative transformations (translation, rotation/reflection and scaling), while the relative intersample relationships within each configuration are maintained (Arnold & Williams 1985; Oreskovich et al. 1991). The mean of the transformed individual configurations is then taken, and is thought of as the 'consensus profile', or the 'best possible fit', of these configurations.

Precisely how well individual observer configurations fit the consensus profile is quantified by the Procrustes statistic. This statistic reflects the degree of similarity (as regards projected geometric distances between pigs) between transformed observer configurations and the consensus profile. The larger the Procrustes statistic, the more the observers agree about the configuration of pigs (but not necessarily about descriptive terms, see below). Wemelsfelder et al. (2000) provided an appendix in which these successive steps of GPA transformation and testing are illustrated with a simple example, in which two observers each use two terms to describe four pigs.

As indicated above, GPA is designed to find a consensus between a given set of matrices, regardless of how variable the data are. (Hence its name: Procrustes was a Greek innkeeper in Attica who managed to fit his guests into his one-size beds by tying them to the ironwork and either cutting or stretching their legs as necessary, Oreskovich et al. 1991.) Thus the danger exists that the attained consensus profile could be an artefact of the statistical technique rather than a significant feature of the data set.

Evaluation of the significance of a computed consensus profile and its 'goodness of fit' is possible through a randomization test (Wakeling et al. 1992). By analysing the original data in randomized form a large number of times, GPA derives a 'goodness-of-fit' statistic for a random association between matrices. The significance of the 'goodness of fit' of the original data sets can then be evaluated against this 'random association' statistic, using a Student's t test (one tailed). We took a probability of <0.001 to indicate that the consensus profile was a meaningful feature of the data set and not a statistical artefact.

The Procrustes statistic also provides information on the similarity/disparity of individual observer configurations relative to the final consensus. GPA provides a Procrustes statistic for each pair of observers, and this measure can be thought of as a measure of the distance between these observers. Thus a triangular table can be formed, giving the distances between each possible pair of observers (cf. tables of distances between towns in a road atlas). Principal coordinate analysis (PCO) of these relative observer distances then makes it possible to map the observers in preferably two (sometimes more) dimensions (the 'observer plot').

Using robust methods (i.e. methods that are not influenced by outliers) it is possible to estimate the centre of distributions of observers together with a standard

deviation, and thus to draw a 95% confidence region. Observers lying outside this region are potentially outliers, that is, in some sense they may differ from the other observers in their assessment of the samples (Williams & Langron 1984; Arnold & Williams 1985; Gains & Thompson 1990). If valid reasons exist for excluding an outlier from analysis (e.g. an observer may be of a different professional background), GPA can be rerun and the new consensus profile can be assessed for significance.

GPA thus transforms the nine different pig/term configurations into one multidimensional consensus profile, entirely independently of any interpretative judgment by the experimenter. This consensus profile is defined purely in terms of its geometrical properties, has as yet no semantic connotations attached to it, and serves as the basis for further statistical transformation to make interpretation possible.

A first step towards interpretation is to reduce the number of dimensions of the consensus profile through conventional principal component analysis (PCA; see for example Stevenson-Hinde et al. 1980). PCA determines which are the principal axes of the consensus profile, and how much of the variation between pigs each of these axes explains. This information is reflected in one or more two-dimensional 'sample plots' which show the distribution of the 10 pigs along the principal axes of the consensus profile. A standard error ellipse can be drawn indicating the reliability of each pig's position on the two axes. Again at this point these axes are still defined purely in terms of their geometrical properties and bear as yet no relationship to semantic referents.

The second step, however, is to confer semantic meaning on to the principal axes of the consensus profile. This is done not by somehow pooling the terms of individual observers into one common terminology, but by calculating how the coordinates of the consensus profile correlate with the coordinates of each of the nine original individual data matrices (created by the nine observers). This analysis results in nine two-dimensional 'word charts' (one for each observer). In each chart, all terms of a particular observer are correlated with the first two (or the third and fourth) principal axes of the consensus profile. The higher a term's correlation with an axis, the more weight it has as a descriptor for that axis. Thus, nine independent word charts for the description of the consensus profile are obtained for comparison and interpretation.

The degree of semantic convergence between these charts indicates the extent to which individual observers concur in their descriptions of the pigs' expressions. For example, in one observer's chart the terms confident/playful may show the highest correlation with the consensus profile's main axis, while in another observer's chart the terms assertive/boisterous take this place. Even though these are different terms, they have similar meaning, and the two observers seem to agree about what they saw. However, if another observer described the main axis in terms of uncertain/restless, disagreement has obviously occurred. In principle it is possible to find a valid consensus profile for which observers show poor semantic agreement, and which therefore makes little sense. An important second measure of observer agreement (in addition to the Procrustes statistic) is therefore whether the word charts of individual observers converge in semantic structure and tone.

As a third and final step of interpretation, the experimenter can summarize any apparent convergence between individual word charts, to interpret the variation between pigs as reflected in the 'sample plots'. This active role of the experimenter, however, is entirely post hoc, and plays no role in the computation of the consensus profile. The strength of GPA is that it preserves semantic information as part of the analysis of object-based data sets, independently of the experimenter's interpretation of that information. This makes it possible to investigate whether observers apply their qualitative vocabulary in similar ways to characterize a group of animals.

If observer assessments show significant convergence, then the consensus profile can be used to appraise qualitative differences between individual animals. These differences, defined by the position of individual animals on the plot, are entirely relative to the group of animals observed and can be interpreted semantically with the help of the individual observer word charts. Although we can conclude from the pigs' distribution along the main axes that one pig is 'more' or 'less' confident than another, the scores defining a pig's position in a 'sample plot' have no meaning outside that particular plot. GPA is designed to compare qualitative assessments of particular sample sets, and as such seems perfectly suited to the aims of our study.

Intraobserver reliability: comparison of four different sets of data matrices. As indicated above, the behavioural expression of each pig is defined by its scores on the main axes of the consensus profile. In this study there are four repeated assessments of the same 10 pigs. By comparing the pigs' scores obtained in each of these four sessions, we can determine the repeatability, or intraobserver reliability, of qualitative assessments of behaviour.

However, as pointed out above, the pig scores within a given 'sample plot' have meaning only relative to that particular plot. It is inappropriate therefore to correlate pig scores produced by separate live and video analyses. To make comparison possible, data from live and video assessments must be merged before analysis. This is done by pasting the live and video assessments of each observer into one large data matrix attributed to that observer. For the sake of the GPA programme, repeated assessments of the same pig within this large matrix have to be numbered differently. GPA analysis can then proceed with nine merged data matrices (one for each observer). This analysis will produce pig scores for live and video assessments as part of one and the same 'sample plot', and so these scores can now be compared and correlated on the same numerical scale.

We merged the four data sets in two ways. First, data matrices of Live1, Video1.1 and Video1.2 were merged (Merged assessment 1). These three sessions are identical repeats (i.e. of the same pigs on the same occasion) and therefore allow direct testing of the repeatability of qualitative behaviour assessments. The nine data matrices of

Table 1. Procrustes statistics of live, video and merged assessments

Procrustes statistic	Live1	Live2	Video1.1	Video1.2	Merged assessment 1	Merged assessment 2
Consensus profile	81.1	79.7	85.3	85.3	74.5	72.6
Mean randomized profile±SD	61.6±1.5	67.0±0.7	59.6±1.2	61.1±1.3	40.3±0.4	51.0±0.6
t_{99}	15.75*	15.53*	23.62*	20.96*	52.68*	27.67*

See text for details.
*$P<0.001$.

each of the three sessions were merged into nine amalgamated data matrices (one for each observer) of $3 \times 10=30$ pigs. The actual number of pigs in each data matrix, however, was 31 and not 30, as we inadvertently pasted footage of one pig twice consecutively while editing Video1.1. We decided not to correct this mistake, as it presented an interesting 'mini' repeatability test. When asked, observers were convincing in indicating that they had not noticed the footage repetition.

Second, we merged data matrices of Live1 and Live2 (Merged assessment 2). These two sessions presented the same 10 pigs to observers on 2 consecutive days, and therefore allow us to test the repeatability of live assessments of pig behaviour. However, a confounding factor is that the pigs may behave differently on the 2 days. Any difference in sample plot scores could thus be behaviour based rather than resulting from poor repeatability. If Merged assessment 2 shows good observer agreement, and if analysis of Merged assessment 1 establishes good repeatability, then any difference between Live1 and Live2 scores apparent in Merged assessment 2 may be interpreted as pig based rather than observer based.

RESULTS AND DISCUSSION

The methods of analysis section above discusses sample plots, word charts and observer plots in the order followed by conventional statistical expositions of GPA. However, the central theme of this paper is the inter- and intraobserver reliability of spontaneous assessments of behavioural expression, and it therefore seems more suitable to present observer plots first, followed by word charts and sample plots.

Observer Plots

In all cases the consensus profile explained a significantly higher percentage of the variation between observer matrices than the mean of 100 randomized profiles (Table 1). The observer plots for live, video and merged assessments (Fig. 1) indicate good consensus, as the majority of observers fell within the 95% confidence region. Some observers were marginal outliers. Depending on a study's aim it may be of interest to take a closer look at such outliers and investigate whether valid grounds exist for excluding them from analysis; this may increase the significance of the consensus profile. However, as the consensus profiles in this study are highly significant, such close scrutiny is not needed.

These results show that none of the six consensus profiles we analysed are artefacts of GPA procedures. In each of the separate live and video assessments, and in both merged assessments, observers showed significant agreement in their qualitative assessment of the pigs' behavioural expressions. Thus spontaneous qualitative assessments of pig behaviour, made by naïve observers given complete freedom to create their own vocabularies, consistently show high interobserver reliability.

Below, we use the merged assessments as the basis for discussion of word charts and sample plots. The separate live and video assessments do not differ essentially from the merged assessments and would not contribute useful additional information.

Word Charts

Figure 2 shows, as examples, the word charts of observers 4 and 8 (see Fig. 1) for Merged assessment 1. The charts of these observers for Merged assessment 2 are very similar and are therefore not shown. The axes of these word charts reflect the first two principal axes of the consensus profile, and indicate which of each particular observer's terms best correlate with those axes. Thus observer 4 described axis 1 as ranging from 'confident/persistent/inquisitive/relaxed' to 'timid/cautious/defensive/anxious'. Observer 8 gave a similar characterization of this axis as ranging from 'bold/confident/pushy/playful' to 'tentative/guarded/puzzled/nervous'. The second axis was characterized by observer 4 as ranging from 'relaxed/slow' to 'excitable/restless', while observer 8 described it as ranging from 'steady/relaxed' to 'excitable/pushy'. These terms all correlate strongly with the axes of the consensus profile (r values between 0.5 and 0.9), and thus describe these axes reliably.

Table 2 and Fig. 3 summarize how similar or different the other observer word charts were from these examples. Table 2 lists for all nine observers, in both Merged assessments 1 and 2, which two terms of their vocabulary showed the highest positive and negative correlations with axes 1 and 2. For example, for seven observers in Merged assessment 1 'confident' best described the positive end of axis 1, while for four observers 'excitable' was the best positive descriptor of axis 2. This list of terms

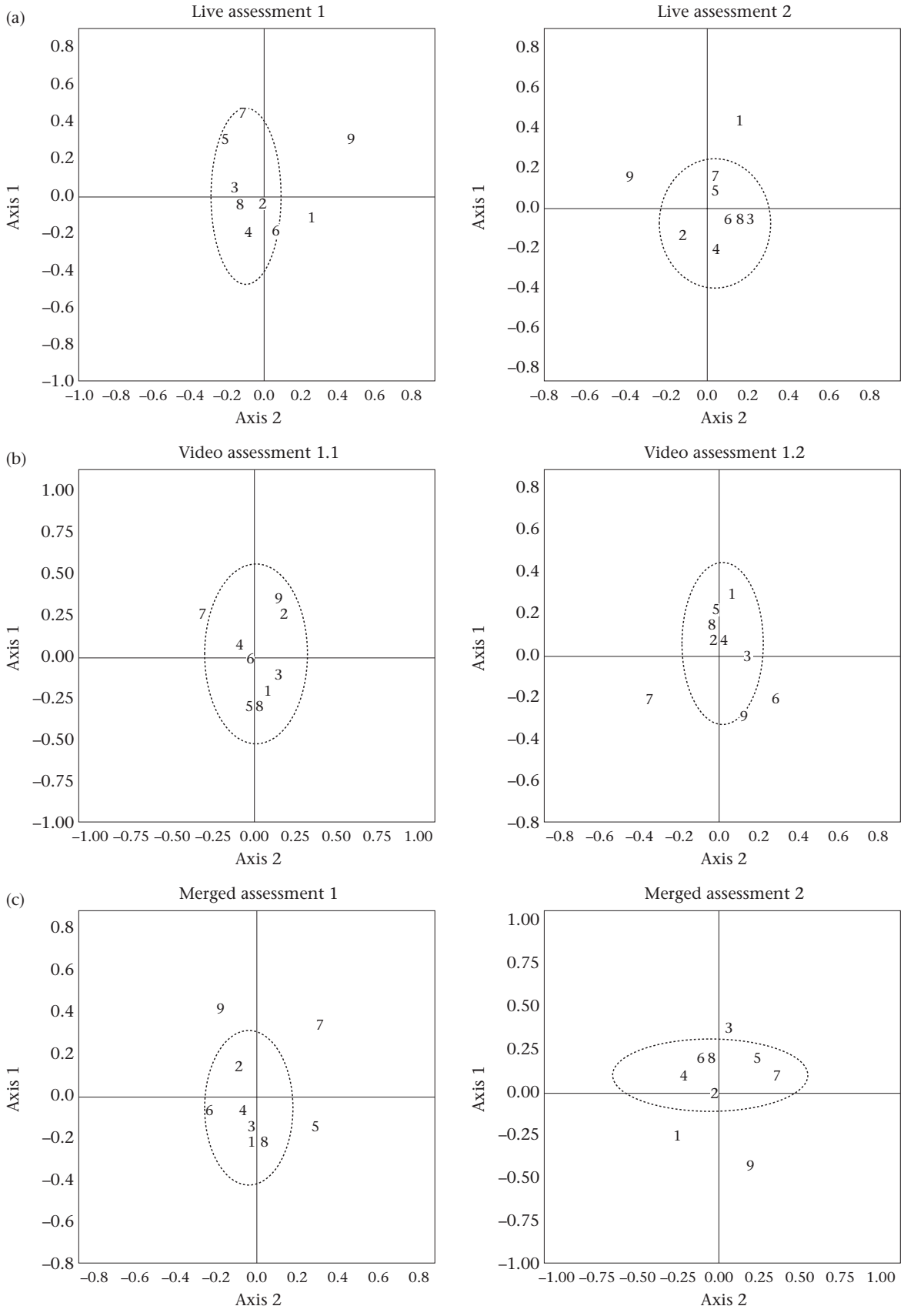

Figure 1. Observer plots. (a) Live assessments 1 and 2; (b) video assessments 1.1 and 1.2; (c) merged assessments 1 and 2. Axes reflect GPA scaling values for relative observer distance. Numbers represent individual observers. The dotted ellipse reflects a 95% confidence region for what may be considered the 'normal population' of observers. See text for details.

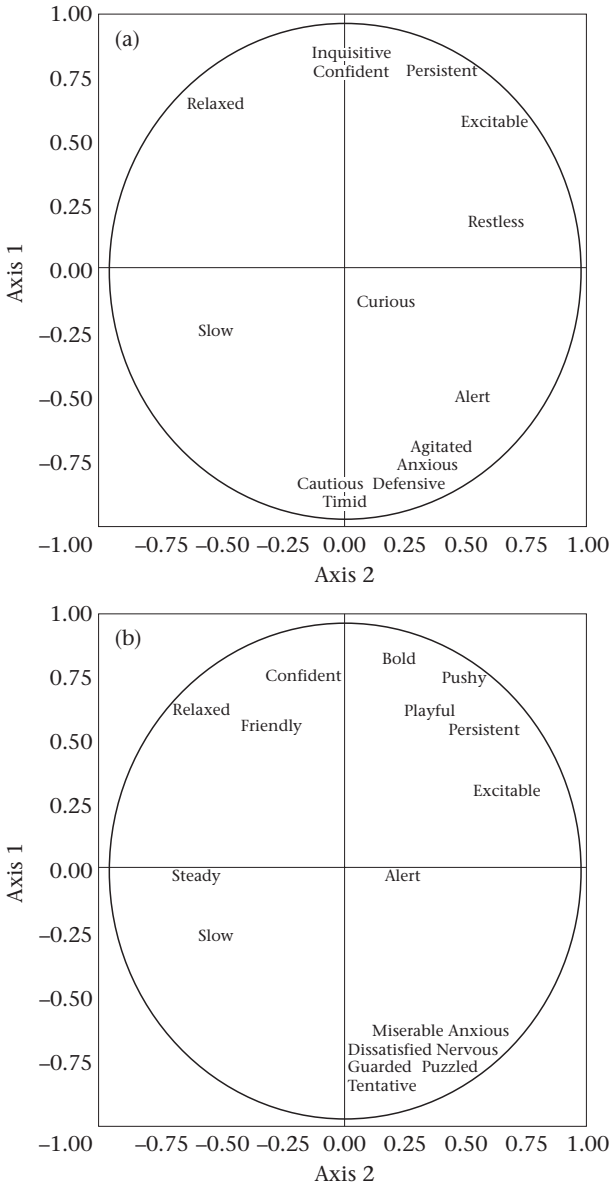

Figure 2. Word charts of (a) observer 4 and (b) observer 8 for Merged assessment 1. Axes reflect the correlation of an observer's terms with axes 1 and 2 of the consensus profile. See text for details.

shows, first, that a considerable number of observers chose identical terms to describe axes 1 and 2 (e.g. 'confident', 'timid', 'excitable', 'relaxed'). Second, where different terms were chosen the meanings of these terms were either very close (e.g. 'confident'/'bold', 'excitable'/'lively'/'energetic', 'uncertain'/'tentative'/'unsure', 'relaxed'/'calm'/'steady'), or they reflected complementary aspects of the expressive repertoire. 'Confident' is not the same as 'playful' or 'domineering', nor does 'timid' have the same meaning as 'tense', yet a playful or domineering pig is likely to appear confident, while timid pigs are likely to appear tense rather than relaxed. Thus the terms on both ends of axes 1 and 2 converge in general semantic tone, and provide a transparent characterization of the dimensions of pig behavioural expression.

Figure 3 indicates the strength of correlation of these main descriptors with axes 1 and 2. The descriptors of axis 1 (including both Merged assessments 1 and 2) are all highly correlated with this axis, and thus characterize it well. Descriptors of axis 2 in both merged assessments show a lower but still acceptable level of correlation, with the majority of r values between 0.4 and 0.8.

These results demonstrate that naïve observers consistently used their self-generated terminologies as coherent semantic frameworks for the assessment of pig behavioural expression. Given that the observers had received no prior training, their individual word charts could well have been an unstructured jumble of terms, with little overlap between them. However, throughout the live and video assessments (and therefore in the merged assessments) observer terminologies showed strong semantic convergence in their characterization of pig expression. This indicates that the observers based their assessments on commonly perceived and systematically applied criteria of behavioural expression.

Sample Plots

Figure 4 shows the 'sample plots' of Merged assessments 1 and 2. In both assessments the standard error of the position of individual pigs on the plot is small, and so this position reliably characterizes the coordinates of pigs on axes 1 and 2. The pigs are distributed reasonably evenly over the plots, indicating that the two axes provide good resolution as independent dimensions of pig behavioural expression.

In Merged assessment 1, axis 1 explains 63.0%, and axis 2 14.6%, of the variation between pigs, giving a total of 77.6% of the variation between pigs explained. In Merged assessment 2, these figures are 51.2, 17.5 and 68.7%, respectively. This variation in behavioural expression can be interpreted semantically with the word charts as discussed above. For example, in the terms of observer 8 (Fig. 2b) pig 5 was seen as tentative, guarded and puzzled, pig 7 as bold, pushy and playful, and pig 3 as steady, relaxed and friendly.

The repeatability of these assessments can be determined by comparing and correlating the Live1, Video1.1 and Video1.2 pig scores in Merged assessment 1, and the Live1 and Live2 scores in Merged assessment 2. In Merged assessment 1, for each pig these three scores are all close or very close to each other (Fig. 4a). For pigs 5, 8, 3, 4 and 7, these scores basically occupy the same position on the plot, including the inadvertent repeat of pig 8 on Video1.1 (pig 8bb). For pigs 1, 2, 6, 9 and 10, the three scores are spread slightly wider but still appear close to each other. This impression is confirmed by the correlations between the repeated pig scores in Merged assessment 1 which range from 0.88 to 0.99 and are all highly significant (Table 3). Furthermore, the mean values of Live1, Video1.1 and Video1.2 scores do not differ significantly on either axis. Thus the pig scores of Live1, Video1.1 and Video1.2 assessments show excellent repeatability on both axes 1 and 2 of the consensus profile.

Table 2. Merged assessments 1 and 2: terms (two for each of the nine observers) that showed the highest positive and negative correlations with axes 1 and 2 of the consensus profile

	Positive correlations	Negative correlations
Axis 1		
Merged assessment 1	Confident (7), playful (3), domineering (2), confrontational (1), bold (1), persistent (1), interactive (1), relaxed (1), affectionate (1).	Timid (3), wary (2), apprehensive (2), cautious (2), uncertain (1), tentative (1), unsure (1), suspicious (1), restrained (1), tense (1), nervous (1), anxious (1), scared (1).
Merged assessment 2	Confident (4), bold (2), interactive (2), assertive (1), domineering (1), confrontational (1), aggressive (1), persistent (1), inquisitive (1), playful (1), brisk (1), relaxed (1), affectionate (1).	Timid (4), wary (2), tense (2), cautious (1), uncertain (1), tentative (1), unsure (1), suspicious (1), evasive (1), avoiding (1), nervous (1), frightened (1) defensive (1).
Axis 2		
Merged assessment 1	Excitable (4), persistent (3), alert (2), pushy (2), lively (1), active (1), energetic (1), confrontational (1), aggressive (1), restless (1), frustrated (1).	Relaxed (6), calm (3), confident (2), friendly (1), comfortable (1), unconcerned (1), slow (1), steady (1), distracted (1), vocal (1).
Merged assessment 2	Excitable (4), alert (3), lively (1), frisky (1), brisk (1), energetic (1), persistent (1), active (1), aggressive (1), pushy (1), domineering (1), confrontational (1), compulsive (1).	Relaxed (4), calm (3), friendly (2), confident (1), comfortable (1), unconcerned (1), slow (1), steady (1), timid (1), vocal (1), interactive (1), agitated (1).

Values in parentheses give the number of observers using that term. See text for details.

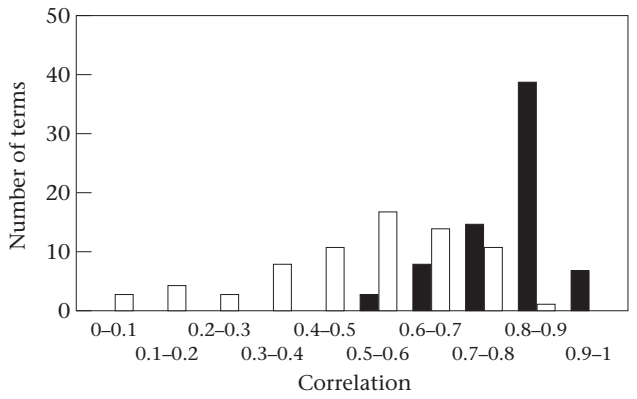

Figure 3. Correlation of observer terms with axes 1 (■) and 2 (□). Merged assessments 1 and 2 are included. See text for details.

In Merged assessment 2, however, the distribution of pigs along axes 1 and 2 differs from that in Merged assessment 1 (Fig. 4b). Live1 scores are the same as in Merged assessment 1, but Live2 scores (unlike Video1.1/ Video1.2 scores) appear at some distance from Live1 scores. The correlation between Live1 and Live2 scores is 0.82 (N=10, P<0.01) on axis 1, and 0.35 (N=10, NS) on axis 2. The means of Live1 and Live2 scores do not differ significantly, although Live2 scores on axis 1 show some tendency to be higher than Live1 scores (two-tailed paired Student's t test: $t_9 = -1.81$, P<0.10), indicating that on day 2 pigs appeared more confident than on day 1.

The question is whether these differences are observer based, indicating poor repeatability of live assessments, or whether they are pig based, describing real differences in the behaviour of pigs on the 2 consecutive days of live assessment. Taking into account the very high levels of observer consistency found in all other parts of this study, we suggest that the latter is the case. First, as indicated above, pigs appeared slightly more confident on day 2 which may be due to habituation to the test situation and a concomitant decrease in fear. Second, on day 2 pigs shifted in different directions along axis 2; some were more lively and excited on day 2, while others were calmer than before. These individual differences are unrelated to the pigs' confidence scores on either day 1 or day 2, so it remains unclear why they arise. Together these results indicate that pigs responded to the repeated live testing with consistent levels of confidence/timidity, but with unpredictable levels of excitability/calmness.

In summary the results from these two merged assessments show that observers can repeat their assessment of the behavioural expressions of individual pigs with great accuracy. Given this ability, observers were also able to detect and measure shifts in the behavioural expressions of individual pigs. These results support the notion that spontaneous qualitative assessments of pig behavioural expressions consistently show high intraobserver reliability.

GENERAL DISCUSSION

Our results show that spontaneous qualitative assessments of pig behavioural expression, as facilitated by free choice profiling methodology, show very high inter- and intraobserver reliability. Nine naïve observers, on four separate occasions, achieved significant agreement in their spontaneous assessment of pig behavioural expression, while merged analyses of these four sessions showed that observers could repeat their assessment with great accuracy. Levels of data resolution were high throughout the different parts of the study, with the consensus profiles of merged analyses explaining 78 and 69% of the variation between pigs. Accordingly observer terminologies correlated strongly with the principal dimensions

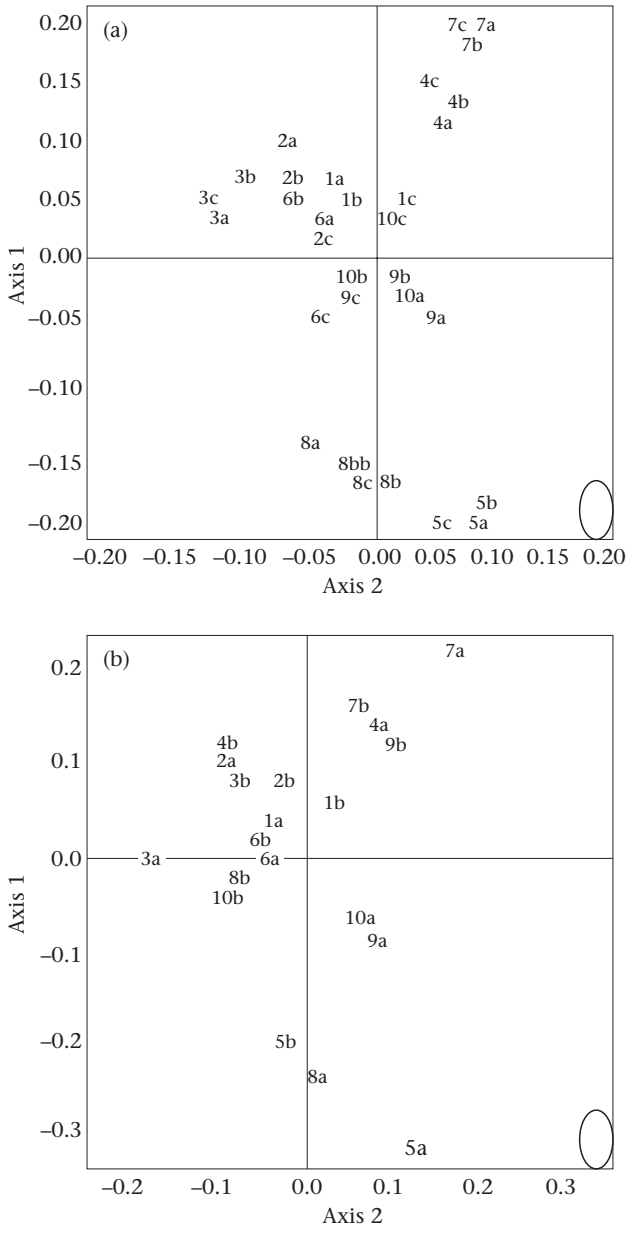

Figure 4. Pig plots for (a) Merged assessment 1 and (b) Merged assessment 2. Axes reflect GPA scaling values for relative sample (pig) distance on axes 1 and 2 of the consensus profile. Numbers represent individual pigs. In Merged assessment 1 suffixes a, b and c represent Live1, Video1.1 and Video1.2 assessments of each pig, respectively. Suffix bb in pig 8bb reflects the inadvertent repeated assessment of pig 8 on Video1.1. In Merged assessment 2 suffixes a and b reflect Live1 and Live2 assessments of each pig, respectively. The ellipses in the right bottom corner of a plot reflect the standard error for each pig's position in that plot. See text for details.

of these consensus profiles, and thus were transformed from a loose collection of terms into a structured and meaningful framework for characterizing the behavioural expressions of individual pigs.

These results provide strong evidence of the internal validity of spontaneous qualitative assessments of pig behavioural expression. The implication is that these assessments were based on commonly perceived and

Table 3. Pearson correlations between pig scores of Live and Video assessments

	Axis 1 (confident–timid)	Axis 2 (excitable–relaxed)
Live1–Video1.1	0.99*	0.94*
Live1–Video1.2	0.96*	0.88*
Video1.1–Video1.2	0.98*	0.90*

See text for details. df=8.
*P<0.001.

systematically applied criteria, and so the question arises what the nature of these criteria may be. Good internal validity does not in itself demonstrate that criteria reflect observable aspects of behavioural organization; it is possible to train observers to project criteria collectively on to reality. However, an important feature of this study was that observers were not collectively instructed to use predetermined categories, but naïvely and freely assessed pigs with their self-created terms. Thus observers actively discerned criteria of behavioural expression through direct observation of behaving pigs, in a coherent and systematic way. It is this special feature that in our view suggests that the observed criteria of behavioural expression were pig based, and biologically 'real'.

Other studies of individual differences in pigs support the biological relevance of the two dimensions of behavioural expression ('confident/timid' and 'excitable/relaxed') reported here. These studies indicate that pigs differ in their response to a novel environment or to human handling, as measured for example by latency of approach, time spent in interaction or speed and degree of movement (Spoolder et al. 1996; Hemsworth & Coleman 1998; Erhard & Mendl 1999; Erhard et al. 1999; Thodberg et al. 1999; Andersen et al. 2000). The various authors interpret these differences in terms of fearfulness (Spoolder et al. 1996; Hemsworth & Coleman 1998; Andersen et al. 2000), restlessness (Spoolder et al. 1996), indifference/curiosity (Thodberg et al. 1999), walking hesitantly/freely (Erhard et al. 1999), and relaxed/tense posture (Erhard et al. 1999). Furthermore Špinka et al. (2000) mentioned 'calmness' as an important dimension of sow maternal behaviour. Such qualitative interpretations of individual behavioural differences generally support our results. However, in most studies qualitative terminologies were applied post hoc as interpretations of acquired quantitative data, whereas in our study they were generated unmediated by quantitative measurement, in direct observation of the pigs. Our study thus suggests that direct, unmediated qualitative behaviour assessment also rests on reliable empirical ground. The question is on which aspects of behavioural organization such assessment may be based.

We propose that given its integrative nature, qualitative behaviour assessment reflects an integrative, 'whole-animal' level of organization (Webster & Goodwin 1996; Ho 1997; Goodwin 1999). On this level we perceive not merely a string of separate 'behaviours', but the unity of

those behaviours, their focal point of origin, which is the 'behaver', the animal. This behaver does not just emerge from the sum total of observed separate behaviours; it executes these behaviours in a certain manner, and it is this instrumental relationship that gives the animal's movement its expressive character (Hacker 1993; Smit 1995; Wemelsfelder 1997a; Sheets-Johnstone 1999a; Wilkerson 1999). Thus it is not the 'walking', 'biting' or 'vocalizing', or a combination of these behaviours, that express fear or confidence: the animal expresses these qualities through its behaviours. This principle, that animals are expressive 'behavers' rather than assembled strings of 'behaviour', is known in the wider psychological and philosophical literature as 'agency' (Nagel 1986; Goodwin 1999; Sheets-Johnstone 1999b). The animal science literature currently gives this concept little heed; however, in discussions of the integrative aspects of behavioural organization it should play an important role (Wemelsfelder & Birke 1997). Thus we propose that descriptions of behavioural expression are descriptions of agency, and reflect the 'whole animal' as a dynamic focus for observation.

The notion of a whole animal level of organization is compatible with current behaviour theory. In the guise of 'comparator mechanisms', 'general information processors', or 'decision-making rules', the notion of integration features in virtually every model of behaviour (McFarland 1989). As these concepts indicate, behavioural integration serves to compare and evaluate alternative routes of behaviour: 'The evidence . . . tends to support the view that animals are capable of consistently evaluating alternatives and of maximising their combined value' (McFarland 1977, page 15). However, current behaviour theory conceives of integration as an implicit systemic feature shaped by natural selection, not as an overt behavioural activity in its own right. Overt behaviour is regarded merely as an aggregate of physical movements (e.g. sit, walk, stand, bite, sniff), caused by underlying factors of an environmental, genetic, or cognitive/emotional kind. Animals (and humans, for that matter) thus emerge as passively driven assembled physical forms (Dennett 1991).

Our results do not fit this model of behavioural organization. Descriptors of behavioural expression (e.g. 'confident', 'anxious') should not, given their integrative character, be seen as causal states of any (physical or mental) kind. They do not describe feelings, genetic predispositions, cognitions, or any factor in isolation, but all of these factors as inclusive aspects of one integrated (psychophysical) state. The reductionist quest to determine whether 'confidence' is either a behaviour, a feeling, or a cognitive state would fragment, and thereby eliminate, the phenomenon it is seeking to explain. Rather than explain the whole animal away, we should accept that it is there and benefit from the expressive information it provides. In analysing such information, the study of animal experience should gain empirical ground, and become less, not more, vulnerable to the risk of anthropomorphic projection (Wemelsfelder 2001). Descriptions of behavioural expression may, for example, facilitate the integration of separate behavioural and physiological measurements into specific, context-sensitive assessments of an organism's welfare state (e.g. as fearful, tense or depressed). Informally and behind the scenes, scientists already use such whole animal perceptions to guide their interpretation of behavioural and physiological results (Davis & Balfour 1992). Formal recognition and systematic analysis of these perceptions could enhance models of animal behaviour, and open up new avenues of behaviour and welfare research.

Acknowledgments

We thank Hans Erhard for valuable help with the experimental design, Hans Erhard and two anonymous referees for helpful comments on the manuscript, and Joan Chirnside, Sheena Calvert, Dave Anderson and Terry McHale for taking care of the animals and their help with experimental procedures. This research was financially supported by the UK Ministry of Agriculture, Fisheries and Food, and by the Scottish Office Agriculture Fisheries and Food Department.

References

Andersen, I. L., Bøe, K. E., Foerevik, G., Janczak, A. M. & Bakken, M. 2000. Behavioural evaluation of methods for assessing fear responses in weaned pigs. *Applied Animal Behavioural Science*, **69,** 227–240.

Arnold, G. M. & Williams, A. A. 1985. The use of Generalized Procrustes Techniques in sensory analysis. In: *Statistical Procedures in Food Research* (Ed. by J. R. Piggott), pp. 233–253. London: Elsevier Applied Science.

Boissy, A. 1995. Fear and fearfulness in animals. *Quarterly Review of Biolology*, **70,** 165–191.

Clarke, A. S. & Boinski, S. 1995. Temperament in non-human primates. *American Journal of Primatology*, **37,** 103–125.

Davis, H. & Balfour, D. 1992. *The Inevitable Bond: Examining Scientist–Animal Interactions*. New York: Cambridge University Press.

Dennett, D. C. 1991. *Consciousness Explained*. Boston: Little, Brown & Company.

Erhard, H. W. & Mendl, M. 1999. Tonic immobility and emergence time in pigs: more evidence for behavioural strategies. *Applied Animal Behavioural Science*, **61,** 227–237.

Erhard, H. W., Mendl, M. & Christiansen, S. B. 1999. Individual differences in tonic immobility may reflect behavioural strategies. *Applied Animal Behavioural Science*, **64,** 31–46.

Fagen, R., Conitz, J. & Kunibe, E. 1997. Observing behavioral qualities. *International Journal of Comparative Psychology*, **10,** 167–179.

Feaver, J., Mendl, M. & Bateson, P. 1986. A method for rating the individual distinctiveness of domestic cats. *Animal Behaviour*, **34,** 1016–1025.

Gains, N. & Thomson, D. M. H. 1990. Contextual evaluation of canned lagers using repertory grid method. *International Journal of Food Science and Technology*, **25,** 699–705.

Goodwin, B. 1999. Reclaiming a life of quality. *Journal of Consciousness Studies*, **6,** 229–235.

Gosling, S. D. 2001. From mice to men: what can we learn about personality from animal research? *Psychological Bulletin*, **127,** 45–86.

Gower, J. C. 1975. Generalized Procrustes Analysis. *Psychometrica*, **40,** 35–51.

Gower, J. C. & Dijksterhuis, G. B. 1994. Multivariate analysis of coffee images: a study in the simultaneous display of multivariate quantitative and qualitative variables for several assessors. *Quality and Quantity*, **28,** 165–184.

Hacker, P. M. S. 1993. *Wittgenstein: Meaning and Mind. Part I, Essays*. Oxford: Blackwell.

Hebb, D. O. 1946. Emotion in man and animal: an analysis of the intuitive processes of recognition. *Psychological Review*, **53,** 88–106.

Hemsworth, P. H. & Coleman, G. J. 1998. *Human–Livestock Interactions*. Wallingford: CAB International.

Heyes, C. 1993. Anecdotes, training, trapping and triangulating: do animals attribute mental states? *Animal Behaviour*, **46,** 177–188.

Ho, M. W. 1997. Towards a theory of the organism. *Integrative Physiological and Behavioral Science*, **32,** 343–363.

Kennedy, J. S. 1992. *The New Anthropomorphism*. Cambridge: Cambridge University Press.

McFarland, D. J. 1977. Decision-making in animals. *Nature*, **269,** 15–21.

McFarland, D. 1989. *Problems of Animal Behaviour*. Burnt Mill: Longman.

Mendl, M. & Harcourt, R. 1988. Individuality in the domestic cat. In: *The Domestic Cat: The Biology of its Behaviour* (Ed. by D. C. Turner & P. Bateson), pp. 41–54. Cambridge: Cambridge University Press.

Nagel, T. 1986. *The View from Nowhere*. Oxford: Oxford University Press.

Oreskovich, D. C., Klein, B. P. & Sutherland, J. W. 1991. Procrustes Analysis and its applications to free-choice and other sensory profiling. In: *Sensory Science: Theory and Aplications in Foods* (Ed. by H. T. Lawless & B. P. Klein), pp. 353–393. New York: Marcel Dekker.

Plutchik, R. 1980. A general psychoevolutionary theory of emotion. In: *Emotion, Theory, Research and Experience* (Ed. by R. Plutchik & H. Kellerman), pp. 3–33. New York: Academic Press.

Rutter, M. 1987. Temperament, personality and personality disorder. *British Journal of Psychiatry*, **150,** 443–458.

Sheets-Johnstone, M. 1999a. Emotion and movement: a beginning empirical-phenomenological analysis of their relationship. *Journal of Consciousness Studies*, **6,** 259–277.

Sheets-Johnstone, M. 1999b. Phenomenology and Agency: methodological and theoretical issues in Strawson's 'The Self'. *Journal of Consciousness Studies*, **6,** 48–69.

Smit, H. 1995. Are animal displays bodily movements or manifestations of the animal's mind? *Behavior and Philosophy*, **23,** 13–19.

Špinka, M., Illmann, G., de Jonge, F., Andersson, M., Schuurman, T. & Jensen, P. 2000. Dimensions of maternal behaviour characteristics in domestic and wild×domestic crossbred sows. *Applied Animal Behavioural Science*, **70,** 99–114.

Spoolder, H. A. M., Burbridge, J. A., Lawrence, A. B., Simmins, P. H. & Edwards, S. A. 1996. Individual behavioural differences in pigs: intra- and inter-test consistency. *Applied Animal Behavioural Science*, **49,** 185–198.

Stevenson-Hinde, J. 1983. Individual characteristics: a statement of the problem. In: *Primate Social Relationships: an Integrated Approach* (Ed. by R. A. Hinde), pp. 28–34. Oxford: Blackwell Scientific.

Stevenson-Hinde, J., Stillwell-Barnes, R. & Zunz, M. 1980. Subjective assessment of rhesus monkeys over four successive years. *Primates*, **21,** 66–82.

Thodberg, K., Jensen, K. H. & Herskin, M. S. 1999. A general reaction pattern across situations in prepubertal gilts. *Applied Animal Behaviour Science*, **63,** 103–119.

Wakeling, I. N., Raats, M. M. & MacFie, H. J. H. 1992. A comparison of consensus tests for Generalized Procrustes Analysis. *Journal of Sensory Studies*, **7,** 91–96.

Webster, G. & Goodwin, B. 1996. *Form and Transformation; Generative and Relational Principles in Biology*. Cambridge: Cambridge University Press.

Wemelsfelder, F. 1997a. Investigating the animal's point of view; an inquiry into a subject-based method of measurement in the field of animal welfare. In: *Animal Consciousness and Animal Ethics* (Ed. by M. Dol, S. Kasanmoentalib, S. Lijmbach, E. Rivas & R. Van den Bos), pp. 73–89. Assen: Van Gorcum.

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

Wemelsfelder, F. 2001. The inside and outside aspects of consciousness: complementary approaches to the study of animal emotion. *Animal Welfare*, **10,** S129–139.

Wemelsfelder, F. & Birke, L. I. A. 1997. Environmental challenge. In: *Animal Welfare* (Ed. by M. C. Appleby & B. O. Hughes), pp. 35–47. Wallingford: CAB International.

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.

Wilkerson, W. S. 1999. From bodily motions to bodily intentions: the perception of bodily activity. *Philosophical Psychology*, **12,** 61–77.

Williams, A. A. & Langron, S. P. 1984. The use of Free-Choice Profiling for the evaluation of commercial ports. *Journal of the Science of Food and Agriculture*, **35,** 558–568.