CSIRO PUBLISHING

Animal Production Science, 2016, **56**, 1569–1578
http://dx.doi.org/10.1071/AN15101

Review

The contribution of qualitative behavioural assessment to appraisal of livestock welfare

Patricia A. Fleming[A,B], *Taya Clarke*[A], *Sarah L. Wickham*[A], *Catherine A. Stockman*[A], *Anne L. Barnes*[A], *Teresa Collins*[A] *and David W. Miller*[A]

[A]School of Veterinary and Life Sciences, Murdoch University, WA 6150, Australia.
[B]Corresponding author. Email: t.fleming@murdoch.edu.au

Abstract. Animal welfare is increasingly important for the Australian livestock industries, to maintain social licence to practice as well as ensuring market share overseas. Improvement of animal welfare in the livestock industries requires several important key steps. Paramount among these, objective measures are needed for welfare assessment that will enable comparison and contrast of welfare implications of husbandry procedures or housing options. Such measures need to be versatile (can be applied under a wide range of on- and off-farm situations), relevant (reveal aspects of the animal's affective or physiological state that is relevant to their welfare), reliable (can be repeated with confidence in the results), relatively economic to apply, and they need to have broad acceptance by all stakeholders. Qualitative Behavioural Assessment (QBA) is an integrated measure that characterises behaviour as a dynamic, expressive body language. QBA is a versatile tool requiring little specialist equipment suiting application to *in situ* assessments that enables comparative, hypothesis-driven evaluation of various industry-relevant practices. QBA is being increasingly used as part of animal welfare assessments in Europe, and although most other welfare assessment methods record 'problems' (e.g. lameness, injury scores, and so on), QBA can capture positive aspects of animal welfare (e.g. positively engaged with their environment, playfulness). In this viewpoint, we review the outcomes of recent QBA studies and discuss the potential application of QBA, in combination with other methods, as a welfare assessment tool for the Australian livestock industries.

Additional keywords: animal welfare, consumers, farming, stakeholder.

Received 24 February 2015, accepted 24 December 2015, published online 3 May 2016

What is qualitative behavioural assessment (QBA)?

Most livestock producers would say that they find it reasonably easy to identify a sick sheep in a group. It might be more difficult for them to explain how. It is something about the way that sheep interacts with the rest of the mob, stands, or moves. It is not necessarily *what* the animal is doing, but *how* it is doing it. Such descriptions do not just focus on a part of an animal's body, but the whole animal, and capture qualitative aspects of how the animal responds to and engages with its environment. Scientists call this 'behavioural expression', but we could also talk about 'body language' or 'demeanour' (Wemelsfelder *et al*. 2012). It reflects not only the animal's physical or physiological state, but potentially also its psychological (emotional or affective) state (Boissy *et al*. 2007; Rutherford *et al*. 2012; Murphy *et al*. 2014). Consequently, an animal's body language can reveal important aspects of its physical and mental health, and therefore welfare.

Qualitative Behavioural Assessment is a methodological approach for capturing the body language of animals in numbers that can then be analysed statistically. QBA can be applied under a range of conditions and can identify subtle differences in qualitative behavioural expression. Importantly, because body language is dynamic, QBA allows capture of subtle changes in an animal's body language that can be important for welfare assessment and may otherwise be overlooked when individual behaviours are isolated and quantified (Wemelsfelder 1997, 2007; Meagher 2009; Whitham and Wielebnowski 2009). QBA has been included as one of 12 measures as part of the 2004–2009 European Commission's Welfare Quality® audit (European Union 2011). Importantly, QBA was the only measure that captured positive aspects of animal welfare, such as animals being positively engaged with their environment, being active, and being alert (Keeling *et al*. 2013). QBA can potentially be used as a 'first pass' screening method to identify farms or industry situations where further in-depth assessment may be warranted.

QBA was developed to its present form by Wemelsfelder and colleagues at Scotland's Rural College, who developed the general concept and methodology, and did much of the initial validation with a network of European collaborators, testing inter-observer reliability and correlations with other measures of behaviour for a range of species (Wemelsfelder 1997, 2007; Wemelsfelder *et al*. 2000). Their innovation was to design a formal statistical methodology so that aspects of the animal's

body language, such as 'arousal' and 'engagement', could be quantified and therefore compared objectively (Meagher 2009). This viewpoint describes published QBA studies and explains potential for application of this method in the Australian livestock industries.

How is QBA carried out?

QBA relies on observer assessments of the body language of animals viewed live or as filmed footage using a set of descriptive terms (e.g. 'anxious', 'calm', and so on). The descriptive terms could be either a set of fixed list of terms determined through consultation with experts in the area, or alternatively, observers could be shown a preview of a small number of clips and asked to generate their own descriptive terms (a process called free choice profiling; FCP) (more details in the section '*How do we select descriptive terms to assess the animals*?' below). Observers are then presented with scoring sheets where each descriptive term is presented adjacent to a visual analogue scale and they are asked to score each animal (or group of animals) by placing a mark on the scale at a point between 'minimum' and 'maximum' (Fig. 1) that they believe reflects the intensity of the animal's expression for each descriptive term. If the observers are scoring groups of animals, then they are asked to think about the group as a whole. These marks are converted into numerical scores (between 0 = min. and 100 = max.) that are then compared using Generalised Procrustes Analysis (GPA) to determine common patterns ('consensus dimensions' e.g. the axes of the graph in Fig. 1) in how observers scored individual animals. These consensus dimensions can then be correlated with the scores individual observers ascribed for each of their terms, to determine descriptive terms most strongly correlated with each dimension. The analysis also provides scores for each animal on these dimensions, which can be used to compare between experimental treatments.

What does QBA tell us about animal welfare?

Animal welfare includes both physical and mental aspects of an animal's experience, and therefore both physiological and behavioural indicators are useful in assessment (Duncan 2005). QBA assesses the whole animal (Wemelsfelder *et al*. 2001), and QBA scores are correlated with physiological condition and behaviour (references cited in Table 1). We still cannot know how an animal feels, but QBA can provide an assessment of the animal's whole response to its environment and what is happening to it. QBA therefore measures 'outcomes', and can contribute to welfare assessment because it can capture variation in how animals respond to and deal with their environment at that instant.

Recent research has shown statistically significant correlations between QBA scores and physiological indicators relevant to welfare (Stockman *et al*. 2011, 2013; Wickham *et al*. 2012, 2015). For example, sheep that were described by observers as more *alert*/*curious*/*aware* or more *alert*/*anxious*/*nervous* also had elevated heart rates and body temperatures, as well as other physiological indicators of stress (e.g. changes in red and white blood cell indices) compared with animals that were scored lower on these same terms (Wickham *et al*. 2012, 2015). During transport, the neutrophil : lymphocyte ratio (a typical marker of stress in ruminants stress; Jones and Allison 2007) was elevated in cattle that were described as more *agitated*/*restless*/*anxious* or more *stressed*/*tense*/*alert* (Stockman *et al*. 2011, 2013) (Fig. 2). Cattle in lairage that came through to slaughter at the back of the group were described by observers as more *nervous*/*anxious* (Stockman *et al*. 2012); these cattle had a higher plasma lactate concentration at slaughter, which is a measure of exertion and expenditure of body energy reserves, and is also indicative of a corticosteroid-mediated stress response (Hemsworth and Barnett 2001) and is correlated with flight speed and other temperament measures (Petherick *et al*. 2009).

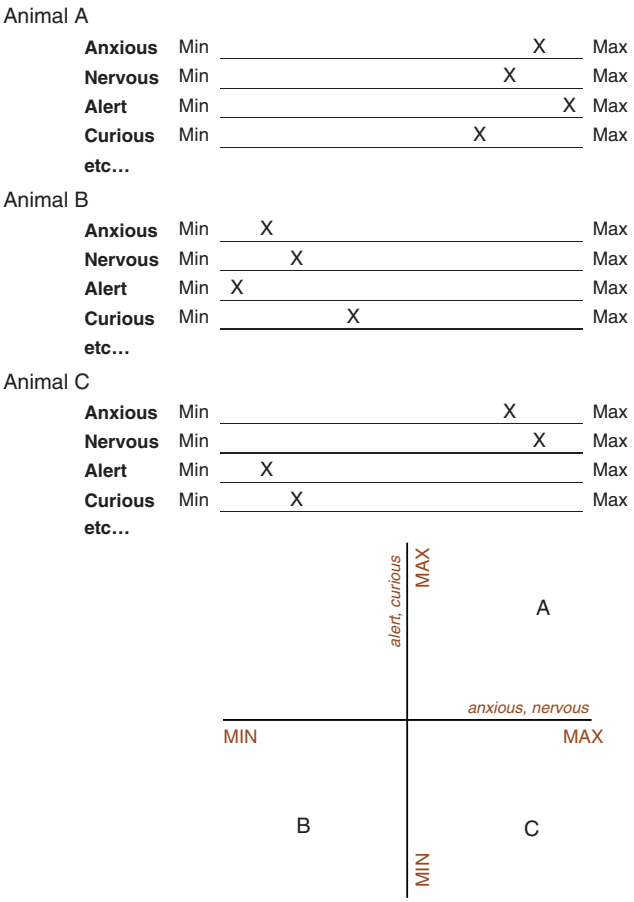

Fig. 1. Qualitative Behavioural Assessment method for scoring three animals (A, B and C). Observers score each animal (or group of animals) on a visual analogue scale (labelled 'minimum' to 'maximum') for a set of descriptive terms. Observers are told to think of the distance between the zero-point (minimum) and their mark on the scale as reflecting the intensity of the animal's expression on each descriptive term. Generalised Procrustes Analysis is a multivariate data reduction technique that then determines the underlying patterns in scores and develops a set of consensus dimensions (the axes in the graph below) that capture this consensus. For example, in this case, animal A received greater scores than animals B for all four descriptive terms, whereas animal C received greater scores for 'anxious' and 'nervous', but lesser scores for 'alert' and 'curious'. GPA identifies differences in the patterns of scoring each term to identify common (consensus) dimensions. Each animal is awarded a score on each consensus dimensions; these scores can be used further, for example to compare between treatments, correlate with physiological measures, or compare with quantitative scores of behaviour.

Table 1. Two Qualitative Behavioural Assessment methods have been applied as an assessment tool across a range of species
Free choice profiling allows observer to use their own descriptive terms to score animals against. Alternatively, the use of Fixed Lists of descriptors have been developed, most notably under the European Union's Welfare Quality® assessment framework

Animal species	Free choice profiling	†	Fixed lists	†
Sheep	Wickham *et al*. (2012, 2015)	P	Phythian *et al*. (2013)	–
	Stockman *et al*. (2014)	B		
	Fleming *et al*. (2015)	B		
Pigs	Wemelsfelder *et al*. (2000, 2001, 2012)	–	Wemelsfelder and Millard (2009)	–
			Wemelsfelder *et al*. (2009*c*)	–
	Rutherford *et al*. (2012)	B	Temple *et al*. (2011*a*, 2011*b*, 2013)	WQ® B
	Morgan *et al*. (2014)	B		
	Lau *et al*. (2015)	B	Duijvesteijn *et al*. (2014)	B
	Clarke (2015)	B		
	Clarke *et al*. (2016)	–	Clarke *et al*. (2016)	–
Cattle	Rousing and Wemelsfelder (2006)	B	Wemelsfelder *et al*. (2009*b*)	–
	Stockman *et al*. (2011, 2012, 2013)	P	Brscic *et al*. (2009)	WQ®
			Bokkers *et al*. (2012)	–
			Andreasen *et al*. (2013)	WQ®
			Sant'Anna and Paranhos da Costa (2013)	B
			Popescu *et al*. (2014)	–
Dairy buffalo	Napolitano *et al*. (2012)	B	Serrapica *et al*. (2014)	–
Horses	Napolitano *et al*. (2008)	B		
	Minero *et al*. (2009)	B		
	Fleming *et al*. (2013)	–		
Poultry			(Wemelsfelder *et al*. 2009*a*)	–
Dogs	Walker *et al*. (2010)	–		

†Studies have compared Qualitative Behavioural Assessment scores with physiological measures (P), quantitative behavioural scores (B), or Welfare Quality® measures (WQ®) as indicated (– indicates no comparison with alternative welfare assessment methods).

QBA scores have also been correlated with various quantitative measures of behaviour (Rousing and Wemelsfelder 2006; Napolitano *et al*. 2008; Minero *et al*. 2009; Sant'Anna and Paranhos da Costa 2013; Morgan *et al*. 2014; Stockman *et al*. 2014; Lau *et al*. 2015). For example, for sheep filmed during a behavioural demand trial (where the animals were required to walk varying distances to receive a food reward), animals that spent more time 'sniffing and looking for more feed' and those that walked a greater distance to obtain food during the trial were described as more *hungry*/*searching*/*excited*. In contrast, those that 'did not walk directly to food reward (stopped along way)', were scored as more *curious*/*intimidated*/*uneasy* (Stockman *et al*. 2014). Rousing and Wemelsfelder (2006) reported significant correlations between social interactions and QBA scores in dairy cows, showing that agonistic behaviour was correlated with an '*aggressive*/*bullying*' demeanour whereas cows that performed social licking were scored as more '*playful*/*sociable*'. Similarly, Napolitano *et al*. (2012) reported that attempts to flee and duration of running in buffalo that had been held in isolation or exposed to a novel chute were associated with an '*agitated*' appearance. Minero *et al*. (2009) found that horses that approached humans, made contact with and nibbled on the clothes of humans were described as '*explorative*/*social*'.

Despite these strong correlations with other welfare-relevant measures (physiology and behaviour), we note that QBA is simply a measuring tool. QBA discerns treatment differences, but we still need other relevant measures to help interpret what the differences indicate about welfare. Therefore, the interpretation of QBA scores and how they relate to overall welfare still requires the judgement of welfare experts. For this reason, QBA has been advocated to be used together with other assessment methods, rather than a stand-alone tool (Wemelsfelder and Mullan 2014).

A recent study comparing 12 quantitative welfare assessment criteria under the Welfare Quality® protocol did not find correlations between QBA and other welfare assessment scores (Andreasen *et al*. 2013). We note, however, that Andreasen *et al*. (2013) relied on a single observer's on-farm assessment of 43 dairy cattle farms. As most of these farms were rated high-welfare ('excellent' under the WQ® framework) and there was little or no variability in many of the welfare criteria, having a greater range of farms may improve the predictive capacity of the tool. Additionally, a recent study (Fleming *et al*. 2013) showed that a small proportion of observers did not score 'tiredness' or 'engagement' in endurance horses, either because they did not perceive such behaviour or they had not generated appropriate descriptive terms (under FCP) to score it. Therefore, relying on a single observer to detect differences between animals (Andreasen *et al*. 2013) may be problematic if that observer is not perceptive to a range of behavioural expression.

Understanding some issues around the use of QBA

'How are observers selected?'

For QBA studies carried out in Australia, observers have largely been university students and people working in various capacities in the livestock industries (principally animal and veterinary scientists) that have volunteered their time. In our

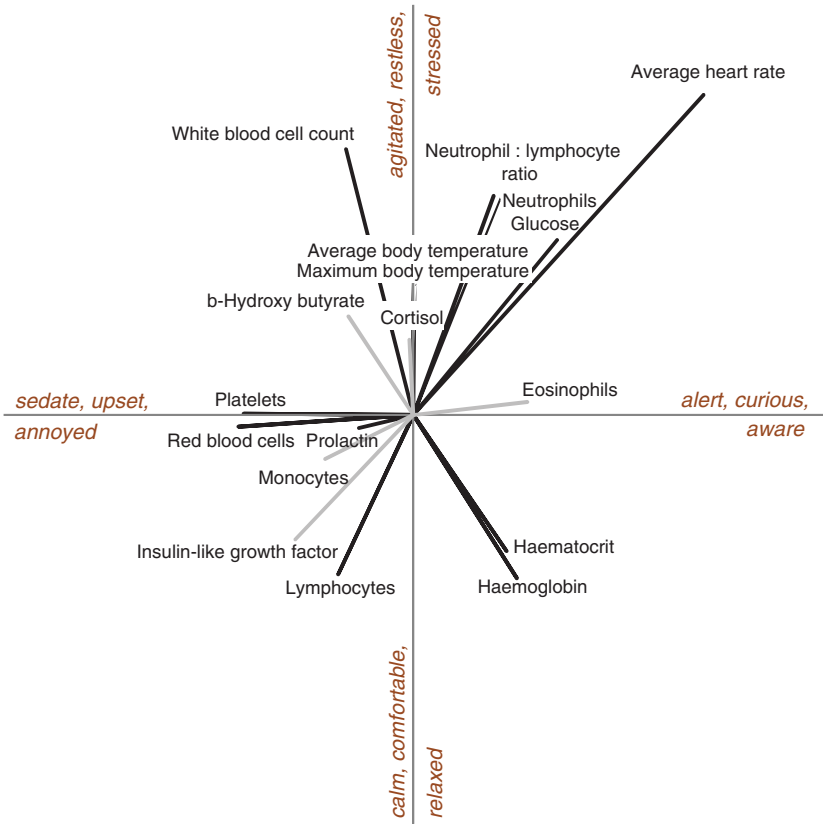

Fig. 2. Correlations between physiological measures and Qualitative Behavioural Assessment axes (described by terms shown in italic) in cattle during transport. Grey lines indicate physiological measures that were not significantly correlated with either Qualitative Behavioural Assessment dimension (*x* or *y*-axis). Data redrawn from Stockman *et al*. (2011).

experience, observers who are willing to take time with their assessments is important, since the process is time consuming and relies on careful observation; engagement with the process means that observers are more willing to carry out and complete scoring in a timely fashion. It is also important for scientific rigour that observers are blinded to the treatments being tested (Fleming *et al*. 2015; and references therein).

Across several international studies, it has become clear that even observers that have little experience with the animal species in question can valuably contribute to qualitative assessments, reaching consensus with other observers in how they score animals. For example, Wemelsfelder *et al*. (2012) reported that, despite different experiences and backgrounds, separate observer groups composed of farmers, veterinarians, or animal rights activists reached consensus (between observer groups) in their assessments of pig behavioural expression. Bokkers *et al*. (2012) found that a group of observers that had little experience with animal welfare assessments (university students who were not familiar with farm animals and had no experience with observing farm animal behaviour) actually reached greater consensus in the way that they interpreted fixed lists of terms for cattle compared with a group of experienced observers who were all familiar with dairy cattle and behavioural assessments. The only difference we have noted between animal-experienced and inexperienced observers is that experienced observers are slightly less likely to use the extremes of their raw visual analogue scales (Box 1); this situation is easily handled by the GPA statistical method (Clarke *et al*. 2016).

The reason that QBA does not necessarily rely on the observer having experience with animals is because the observational and statistical processes are largely independent of subjective interpretation. Observers are asked to focus on the animal and what they see, and the observer's lack of prior experience should not negate their ability to perceive differences in the animal's expression. The mathematical procedures involved in the statistical analyses then identify the common 'dimensions' in their scores. These processes therefore do not rely on subjective judgement to obtain the QBA scores. We note, however, that the interpretation and extrapolation of these scores such as for benchmarking or quality assurance audits, will still require expert opinion and judgement, in the same way that the criteria established for other welfare assessment methods (e.g. maximum incidence of lameness or injuries that trigger reporting requirements) need to be considered carefully.

'Why are there no absolute values associated with QBA assessments?'

All multivariate data reduction methods generate dimensions according to the data that are input into the analysis. GPA

Box 1. Do observers score descriptive terms differently?

Sixty-three observers participated in a study on sheep during land transport (Wickham *et al*. 2012). Various aspects of their demographics as an observer group were recorded during this study, including their gender, age, country of birth, area of study/employment, whether they live in an urban or rural environment, have pets, are vegetarian, and regularity of them witnessing sheep being transported. Forwards stepwise regression comparing these variables with the average range attributed to the raw scores for all descriptive terms indicated that only the observer's age ($P = 0.169$) and their sheep experience ($P = 0.131$) were correlated with how the observers scored their terms. More experienced and older observers were likely to use a smaller range of scores than younger and less experienced observers (Fig. 3). The differences were not significant on their own (i.e. as main effects).

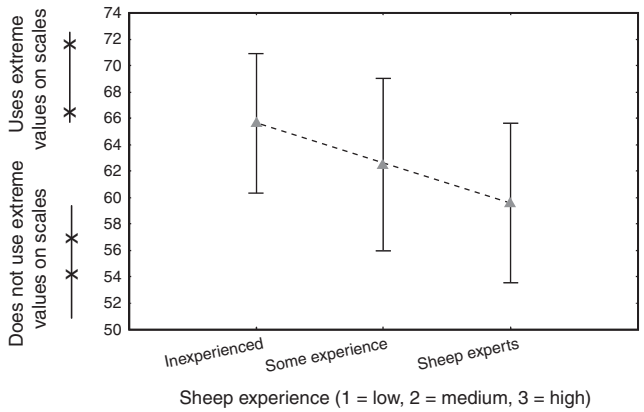

Fig. 3. Average range of visual analogue scale scores attributed by observers compared with their level of experience with sheep, which was interpreted based on answers given for the following questions: 1. Have you ever visited a farm which rears animals (specifically sheep; yes, no); 2. Have you ever visited an abattoir (specifically sheep; yes, no); 3. Currently, how often would you say you come into contact with sheep (daily, once or twice a week, once a fortnight, once a month, once a year, few times a year, never); and 4. How much time have you spent working with sheep in your lifetime (never, a few occasions, a few days, a few weeks, a month or more, a year or more)? The observers classed with a 'high level of sheep experience' had spent more than a month working with sheep, usually came into contact with sheep at least once a month, had visited a sheep farm, and most had visited an abattoir that slaughters sheep. The observers classed as having a 'low level of sheep experience' had spent less than a few days working with sheep in their lifetime, come into contact with sheep less than a few times a year, had not seen sheep at an abattoir, although some had visited a sheep farm. The observers classed as having a 'medium level of sheep experience' fell between these two extremes.

therefore generates consensus dimensions that reflect the particular set of data for an analysis. For example, descriptive terms that capture active behaviour would have a smaller amount of variation for a scoring session that only included pigs held in intensive housing compared with a scoring session that also included animals from outdoor systems. The GPA dimensions generated for each of these scoring sessions would be different, and it is likely that activity would have a greater influence on QBA scores of animals in the second option (indoor and outdoor housing). Because there are no absolute values associated with GPA dimensions, the scores are relative only to the other animals/groups viewed within that session. Comparison between studies is limited by this restriction.

Although there can be some shifts in absolute scores according to the dataset analysed, studies show that the relative positions of animals remain consistent within a dataset and are therefore directly comparable. Wemelsfelder *et al*. (2009*c*) found that observers viewing exactly the same footage of 15 growing pigs interacting with a novel object, but digitally projected onto either an indoor or outdoor background, shifted their responses slightly (scoring pigs as more *confident/content* and less *cautious/nervous* in outdoor than in indoor clips), but this shift did not distort the relative rankings of animals on expressive dimensions. Fleming *et al*. (2015) compared observer scores where they saw the same set of clips of sheep in two sessions, but the clips were juxtaposed with different sets of footage (showing alternative experimental treatments) for each session. For both of these studies, observers shifted their assessments in terms of the absolute numerical values attributed to the animals, but the relative rankings of animals remained similar. Therefore, although observers' value judgements or contrast effects (where observers become more aware of particular aspects when they are highlighted by comparisons) can influence the absolute numerical values attributed to animals during scoring, the pattern of scoring (i.e. the relative scores, which are the important aspect of this method) are not unduly influenced.

Although the QBA scores are all 'normalised' to make the scores of individual observers comparable, by assessing the raw visual analogue scales, we can gain some insight into how different people perceive and score the behaviour of the animals viewed. For example, Duijvesteijn *et al*. (2014) found that, although observers used terms similarly (largely the same set of descriptive terms were correlated with each of the behavioural dimensions) and had the same intra-observer correlations (each observer saw the same video clip twice), farmers tended to score the behavioural expression of animals more positively than urban citizens or animal scientists. Bokkers *et al*. (2012) compared the scores attributed to the same footage by eight experienced observers for each descriptive term over two viewing sessions, separated by 9 months. The authors found that observers' scores

had shifted moderately over this time, and that the shifts varied between descriptive terms. While most terms showed a shift towards lesser values the second time, a few terms (notably '*active*', '*playful*', '*positively occupied*', and '*lively*') scored greater values the second time. This difference in scoring is likely to reflect the accumulation of experiences by the observers because in the interval, they had each visited between 10 and 48 dairy farms. This finding highlights the value of comparisons that have minimal time lapses between them. QBA therefore works best where observers are presented with a range of conditions to compare directly, or requires observers have some exposure to footage that represents the extremes of conditions they are likely to need to score, ideally during structured training sessions where the footage can be discussed in the context of how it would likely be assessed. Exposure to such extremes is more likely to allow observers to apply their scores on a wider knowledge base, although QBA scores should still be thought of as relative only to the other scores within the same session.

'*Is the QBA method sensitive to context?*'

One reason qualitative assessments can be so informative is that they are sensitive to environmental context. Taking environmental clues into account and evaluating the animal's situation allows observers to make a more discerning, and potentially quantitatively more powerful, judgment of an animal's behavioural style (Wemelsfelder *et al*. 2009*c*; Fleming *et al*. 2015). However, sensitivity to context needs careful evaluation and management, since this sensitivity also makes qualitative assessments vulnerable to undesirable bias due to the observers' judgment of that context (Wemelsfelder *et al*. 2009*c*). This is particularly a risk when different contexts might have different moral connotations. For example, Tuyttens *et al*. (2014) found that observers' assessments (using QBA) of laying hens in a conventional commercial aviary was significantly affected by background information; even though the observers were scoring footage of the same hens, they attributed more positive and fewer negative valence scores to the hens if they were told the aviary was on an organic farm compared with when they were told it was a conventional farm. The size of this difference correlated with their pre-recorded opinions on hen welfare in organic versus conventional systems. Tuyttens *et al*. (2014) also demonstrated similar bias for counts of 'negative' and 'positive' interactions between pigs when observers were told that the pigs were animals selected for 'social breeding value' (where pigs with a high social breeding value have a positive effect on the growth of their pen mates), or in subjectively scoring the degree of panting in cattle (recorded on a visual analogue scale), when a coloured bar on the side of the screen for half of the clips indicated an ambient temperature 5°C hotter than in reality.

Observer bias can influence a range of welfare measures, and being aware of this is an important part of designing welfare assessments. All welfare assessments are founded in the experiences of the people judging the situation, an observation that is often ignored for many quantitative measures (Saks *et al.* 2003; Tuyttens *et al.* 2014). Even quantitative measures can be vulnerable to observer bias; for example, Berkson *et al*. (1940) found that comparison of blood counts by person and machine indicated that technicians reported routine blood counts that were within a narrower band of variability than could possibly have existed, whereas observers led to expect a high rate of turns and contractions in planarian worms recorded twice as many head turns and three times as many body contractions as observers who were lead to expect a low incidence (Cordaro and Ison 1963). Lameness is a major welfare problem for dairy animals, inducing pain and discomfort for long periods. Quantifying the degree of lameness is therefore an important welfare consideration, and yet there can be a high degree of heterogeneity in lameness scoring (de Rosa *et al*. 2003). Other quantitative measures may show similar variability that could be accounted for by observer bias.

'*How are descriptive terms to assess the animals selected?*'

QBA was originally developed using the FCP methodology, with observers generating and using their own descriptive terms to score a group of animals, either by all observers simultaneously watching the animals, or observers being shown the same film footage (Wemelsfelder *et al*. 2001). FCP is a powerful tool that has been used in the food and wine industries, since it allows for individuals to express their own perceptions (Arnold and Williams 1985; Oreskovich *et al*. 1991). In the same way that not everyone identifies with 'chocolate' or 'cut grass' in their wine, many observers can be uncomfortable describing sheep as, for example, 'happy' or 'content'. FCP therefore allows greater ownership of the terms being used, since each individual observer develops and uses their own terms to assess the animals (Clarke *et al*. 2016). The constraint of using FCP is that it requires that observers all watch the same animals/footage, because the statistical analysis of the GPA scores relies on pattern recognition of observers' scores to identify terms that have been used in a similar way by different observers.

For practical on-farm welfare assessments, it may be more feasible to use fixed lists of descriptive terms. The value of fixed lists lies in each observer using a common 'scale' to quantify the behaviour of animals being observed, which then can be analysed by Principle Components Analysis. Because observers are all using the same measuring tool to assess animals, the use of fixed lists means that different observers can be sent to monitor different farms. To ensure that observers are scoring QBA descriptive terms in a common manner, observers can be shown some of the same images/farms (showing the extremes of situations) to calibrate their scoring, or multiple observers could view the same images/farms to test concordance of assessments. The training of observers and design of assessments (allowing observers to assess the same animals) is therefore critical to the successful application of fixed lists (Wemelsfelder and Mullan 2014; Clarke *et al*. 2016).

Another valuable aspect of using fixed lists is that the descriptive terms can be selected to capture a breadth of behavioural expression, since some key aspects of behaviour can be missed when observers develop their own lists of terms via FCP (Fleming *et al*. 2013). Fixed lists can be specifically developed for different animal species or industry contexts; for example, different lists of terms have been used to describe dairy cattle, beef cattle, and calves (as applied under the European Union Welfare Quality® audits, e.g. Wemelsfelder and Millard

2009; Wemelsfelder *et al*. 2009*a*, 2009*b*; Temple *et al*. 2011*a*; Andreasen *et al*. 2013). In the same way that grimace scores to assess pain in animals directs people to attend to facial expressions, being able to direct observers towards important behavioural dimensions (e.g. behavioural expression that may reveal fear during pre-slaughter handling, pain during husbandry procedures, or fatigue during sporting events) allows assessments to be tailored towards the measure in question.

'Is the QBA method reliable?'

For our initial example of a farmer being able to discern a sick sheep in a paddock, testing the reliability of QBA ensures that everyone else can also see that the same sick sheep behaves 'differently' from the rest of the flock. The calculation of QBA scores requires the mathematical calculation of a 'consensus' in observer scoring patterns, and provides a statistic that captures the reliability of this consensus (the Procrustes Statistic). A randomisation test is used to measure how reliable this consensus is, where each observer's scores are rearranged randomly 100 times, and GPA scores are calculated for the new permutated data matrices (Dijksterhuis and Heiser 1995). This test provides an indication of how likely it would be to find a consensus in these assessments through chance alone. To date, where the observers have been able to develop their own descriptive terms to use (FCP), good inter-observer agreement has been shown in studies across a range of species (all the studies listed in Table 1 have shown significant consensus in observer scores).

Using fixed lists of descriptive terms may have more problems in achieving observer consensus than using FCP, because individual observers can have different ethical values and understanding of the meaning of the descriptive terms they are provided to score against (Duijvesteijn *et al*. 2014). However, studies using fixed lists have shown that observers can also reach consensus (Brscic *et al*. 2009; Wemelsfelder and Millard 2009; Wemelsfelder *et al*. 2009*a*, 2009*b*; Temple *et al*. 2011*b*). In a comparison between FCP and fixed list methods, Clarke *et al*. (2016) found strong correlations in the outcome (QBA scores) for animals scored by two observer groups watching the same footage of sows and either using their own individual descriptive terms or a fixed list with which they were provided.

'Is the QBA method sensitive enough to detect treatment differences?'

QBA is a relative measure that is capable of detecting extremely subtle differences in the behavioural expression of animals. For example, when applied in blind observer trials, QBA successfully distinguished between different land transport conditions for sheep and cattle which tested novelty of transportation, effects of stop-start driving, and flooring structure (Stockman *et al*. 2011, 2013; Wickham *et al*. 2012, 2015). QBA scores also differed significantly between pigs treated with the neuroleptic drug Azaperone and non-treated pigs (Rutherford *et al*. 2012); between yearling foals assessed before and after having received a month-long handling treatment (Minero *et al*. 2009); between different stages of an horse endurance ride (Fleming *et al*. 2013); for pigs housed under intensive and extensive housing systems (Temple *et al*. 2011*b*); or for sows housed under subtly different group housing systems (Clarke 2015).

Detecting subtle differences between treatments requires that observers can view the contrasting conditions within a short time frame (Fleming *et al*. 2015; and references therein). Consequently, subtle qualitative differences between animals may be lost to observers if a long time frame separates viewing the different conditions (Temple *et al*. 2013) For example, for the Welfare Quality® audits, it takes 8 h to complete the full assessment of each farm, and therefore only one farm can be assessed in a day. For many species, biosecurity issues also restrict the number of farms that can be visited in rapid succession.

It is not only livestock behaviour that can be measured using this tool; the behaviour of people handling stock can also be measured using QBA. Ellingsen *et al*. (2014) applied QBA to both dairy calves and their handlers. They found that stock persons scored as '*calm*/*patient*', who handled their dairy calves patiently, petted and calmly talked to them during handling, had animals with higher levels of 'positive mood', as characterised by high scores on terms like '*friendly*' and '*content*'. Stockpersons with an '*insecure*/*nervous*' handling style, or who were '*dominating*/*aggressive*', had calves that were scored as showing a more 'negative mood' (showing more '*anxious*' or '*apprehensive*' behaviour).

'Is QBA a versatile measure?'

QBA may be carried out for many situations that are not suitable for other methods of welfare assessment due to logistics or because there is a need for a quick, *in situ* method that is capable of capturing dynamic changes in behaviour with little or no equipment. For example, QBA has been used to score the behaviour of animals on-farm, during transport, in sale yards, being handled in chutes *en route* to a slaughterhouse, or under controlled experimental conditions (Table 1). Qualitative assessments are what good stock handlers do every day as part of their business, where they assess the body language of animals in a dynamic manner to make judgements important for husbandry, such as identifying animals that need medical treatment.

'How could QBA be applied in Australia?'

To address the relevance of welfare assessment methods to the Australian livestock industries, we have been validating and applying QBA to key points in the livestock supply chain to test assumptions, limitations, and broad applicability of the method. The validation process encompassed studies on both cattle and sheep exposed to common industry stressors that included road transport, nutritional variation, pre-slaughter handling, isolation, and exposure to novelty. These studies demonstrated that QBA can be reliably and objectively applied to Australian cattle and sheep, and support the suite of studies that have been carried out across the globe, testing the validity and applicability of this method (Table 1). Subsequently, studies examining housing options in pigs (Clarke 2015) and the effects of habituation (handling and adjustment to hand-feeding) in goats (D. Miller, unpubl. data) have applied QBA as an on-

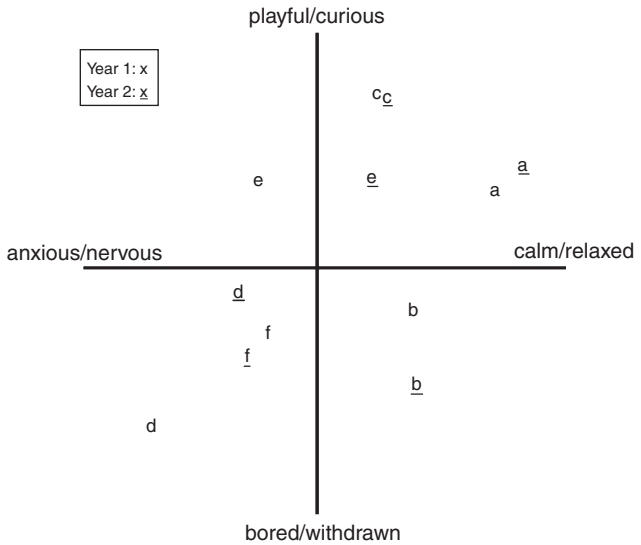

Fig. 4. Hypothetical graphical representation of the behavioural expression of animals for six stakeholders (*a–f*) across two treatments (e.g. for 2 years). Stakeholders can compare themselves with other stakeholders and identify how the behavioural expression of their animals has changed. For example, animals with stakeholder b are more calm/relaxed than animals with stakeholder d, and animals with stakeholder b are more bored and withdrawn in Year 2 than they were in Year 1. Visualising the relative placement of each farm may enable self-evaluation that could contribute to improved welfare.

farm assessment method in Australia, contributing to decisions regarding housing design and husbandry of these animals.

We believe QBA can be used as an initial screening tool to help producers compare husbandry processes or housing options. QBA could be carried out on a repeated basis (e.g. yearly) so that producers could compare outcomes of particular interventions or change in management practices. Capturing footage of animals under conditions to be assessed over time may therefore provide a powerful self-audit method, allowing comparison between similar properties and providing producers with direct feedback.

Development of online assessment tools may be the most accessible way to approach the wide-scale application of welfare benchmarking methods. Rather than relying on individual farm visits which are necessarily spread over time (for logistics and biosecurity reasons), capturing footage for later review and comparison can allow detection of reasonably subtle differences in animal behaviour between farms. Having an interface that allows participants to submit their own footage for review could increase engagement in a benchmarking process and increase ownership of the process by participants. An online system could also allow for direct and immediate feedback to farmers (see Fig. 4 for example), while the process of partaking in a QBA assessment (including viewing footage from other farms) itself could valuably contribute to the observer's stockmanship skills by encouraging time to reflect and better understand patterns of animal behaviour.

A key aspect of any welfare assessment method is that it uses transparent, simple approaches on which all stakeholders agree. QBA uses terms that people can readily relate to and will seek out in their purchase of welfare-friendly products. Because animal welfare requires shared perception and beliefs, tools such as QBA can be very important for communication and learning processes in multi-stakeholder groups. For example, in their study of pig welfare assessments, Duijvesteijn *et al*. (2014) noted that issues involving conflicting framings and polarisation can potentially give rise to misunderstandings or even create a deadlock due to distrust. Shared understanding can be improved through developing a common language, and carrying out and discussing a QBA process effectively stimulated mutual learning among pig farmers, animal scientists, and lay citizens that was necessary to find shared welfare solutions (Duijvesteijn *et al*. 2014). Therefore, in conjunction with other methods, QBA can contribute towards providing the livestock industries with the tools needed to objectively assess animal welfare, and to communicate the high quality standards of the Australian livestock industries to consumers and general public.

Acknowledgements

Our thanks to Francoise Wemelsfelder for her comments and encouragement. The authors thank Meat and Livestock Australia, Beef and Lamb New Zealand, Australian Pork Limited, Pork CRC and Sheep CRC for generous financial support that has enabled the development of this project for the Australian livestock industries.

References

Andreasen SN, 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. doi:10.1016/j.applanim.2012.11.013

Arnold GM, Williams AA (1985) The use of Generalized Procrustes Techniques in sensory analysis. In 'Statistical procedures in food research'. (Ed. JR Piggott) pp. 233–253. (Elsevier Applied Science: London)

Berkson J, Magath TB, Horn M (1940) The error of estimate of the blood cell count as made with the hemocytometer. *American Journal of Physiology* **128**, 309–323.

Boissy A, Manteuffel G, Jensen MB, Moe RO, Spruijt B, Keeling LJ, Winckler C, Forkman B, Dimitrov I, Langbein J, Bakken M, Veissier I, Aubert A (2007) Assessment of positive emotions in animals to improve their welfare. *Physiology & Behavior* **92**, 375–397. doi:10.1016/j.physbeh.2007.02.003

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

Brscic M, Wemelsfelder F, Tessitore E, Gottardo F, Cozzi G, Van Reenen CG (2009) Welfare assessment: correlations and integration between a Qualitative Behavioural Assessment and a clinical/ health protocol applied in veal calves farms. Proceedings of the 18th National Congress ASPA Palermo, Italy. *Italian Journal of Animal Science* **8**, 601–603. doi:10.4081/ijas.2009.s2.601

Clarke T (2015) Qualitative behavioural assessment of sows under different housing conditions. PhD thesis, School of Veterinary and Life Sciences, Murdoch University.

Clarke T, Pluske JR, Fleming PA (2016) Are observer ratings influenced by prescription? A comparison of free choice profiling and fixed list methods of qualitative behavioural assessment. *Applied Animal Behaviour Science* **177**, 77–83.

Cordaro L, Ison JR (1963) Psychology of the scientist: X. Observer bias in classical conditioning of the planarian. *Psychological Reports* **13**, 787–789. doi:10.2466/pr0.1963.13.3.787

de Rosa G, Tripaldi C, Napolitano F, Saltalamacchia F, Grasso F, Bisegnai V, Bordfr A (2003) Repeatability of some animal-related variables in dairy cows and buffaloes. *Animal Welfare* **12**, 625–629.

Dijksterhuis GB, Heiser WJ (1995) The role of permutation tests in exploratory multivariate data analysis. *Food Quality and Preference* **6**, 263–270. doi:10.1016/0950-3293(95)00025-9

Duijvesteijn N, Benard M, Reimert I, Camerlink I (2014) Same pig, different conclusions: stakeholders differ in qualitative behaviour assessment. *Journal of Agricultural & Environmental Ethics* **27**, 1019–1047. doi:10.1007/s10806-014-9513-z

Duncan IJH (2005) Science-based assessment of animal welfare: farm animals. *Revue scientifique et technique-Office international des epizooties* **24**, 483.

Ellingsen K, Coleman GJ, Lund V, Mejdell CM (2014) Using Qualitative Behaviour Assessment to explore the link between stockperson behaviour and dairy calf behaviour. *Applied Animal Behaviour Science* **153**, 10–17. doi:10.1016/j.applanim.2014.01.011

European Union (2011) 'Welfare Quality®: Science and society improving animal welfare in the food quality chain.' Available at http://www.welfarequality.net/everyone [Verified March 2016]

Fleming PA, Paisley C, Barnes AL, Wemelsfelder F (2013) Application of Qualitative Behavioural Assessment to horses during an endurance ride. *Applied Animal Behaviour Science* **144**, 80–88. doi:10.1016/j.applanim.2012.12.001

Fleming PA, Wickham SL, Stockman CA, Verbeek E, Matthews L, Wemelsfelder F (2015) The sensitivity of QBA assessments of sheep behavioural expression to variations in visual or verbal information provided to observers. *Animal* **9**, 878–887. doi:10.1017/S1751731114003164

Hemsworth PH, Barnett JL (2001) Human-animal interactions and animal stress. In 'The biology of animal stress – basic principles and implications for animal welfare'. (Eds GP Moberg, JA Mench) pp. 309–336. (CABI Publishing: Oxon, UK)

Jones M, Allison R (2007) Evaluation of the ruminant complete blood cell count. *Veterinary Clinics of North America: Food Animal Practice* **23**, 377–402.

Keeling L, Evans A, Forkman B, Kjaernes U (2013) 'Welfare Quality principles and criteria.' (Wageningen Publishers: Wageningen, The Netherlands)

Lau YYW, Pluske JR, Fleming PA (2015) Does environmental background (intensive vs. outdoor systems) influence the behaviour of piglets at weaning? *Animal* **9**, 1361–1372. doi:10.1017/S1751731115000531

Meagher RK (2009) Observer ratings: validity and value as a tool for animal welfare research. *Applied Animal Behaviour Science* **119**, 1–14. doi:10.1016/j.applanim.2009.02.026

Minero M, Tosi MV, Canali E, Wemelsfelder F (2009) Quantitative and qualitative assessment of the response of foals to the presence of an unfamiliar human. *Applied Animal Behaviour Science* **116**, 74–81. doi:10.1016/j.applanim.2008.07.001

Morgan T, Pluske JR, Miller DW, Collins T, Barnes AL, Wemelsfelder F, Fleming PA (2014) Socialising piglets in lactation positively affects their post-weaning behaviour. *Applied Animal Behaviour Science* **158**, 23–33. doi:10.1016/j.applanim.2014.06.001

Murphy E, Nordquist RE, van der Staay FJ (2014) A review of behavioural methods to study emotion and mood in pigs, *Sus scrofa*. *Applied Animal Behaviour Science* **159**, 9–28. doi:10.1016/j.applanim.2014.08.002

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. doi:10.1016/j.applanim.2007.03.009

Napolitano F, De Rosa G, Grasso F, Wemelsfelder F (2012) Qualitative behaviour assessment of dairy buffaloes (*Bubalus bubalis*). *Applied Animal Behaviour Science* **141**, 91–100. doi:10.1016/j.applanim.2012.08.002

Oreskovich DC, Klein BP, Sutherland JW (1991) Procrustes Analysis and its applications to free-choice and other sensory profiling. In 'Sensory science: theory and applications in foods'. (Eds HT Lawless, BP Klein) pp. 353–393. (Marcel Dekker: New York)

Petherick CJ, Doogan VJ, Venus BK, Holroyd RG, Olsson P (2009) Quality of handling and holding yard environment, and beef cattle temperament: 2. Consequences for stress and productivity. *Applied Animal Behaviour Science* **120**, 28–38. doi:10.1016/j.applanim.2009.05.009

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. doi:10.1016/j.applanim.2012.11.011

Popescu S, Borda C, Diugan EA, El Mahdy C, Spinu M, Sandru CD (2014) Qualitative behaviour assessment of dairy cows housed in tie-and free stall housing systems. *Bulletin UASVM Veterinary Medicine* **71**, 273–275.

Rousing T, Wemelsfelder F (2006) Qualitative assessment of social behaviour of dairy cows housed in loose housing systems. *Applied Animal Behaviour Science* **101**, 40–53. doi:10.1016/j.applanim.2005.12.009

Rutherford KMD, Donald RD, Lawrence AB, Wemelsfelder F (2012) Qualitative Behavioural Assessment of emotionality in pigs. *Applied Animal Behaviour Science* **139**, 218–224. doi:10.1016/j.applanim.2012.04.004

Saks MJ, Risinger DM, Rosenthal R, Thompson WC (2003) Context effects in forensic science: a review and application of the science of science to crime laboratory practice in the United States. *Science & Justice* **43**, 77–90. doi:10.1016/S1355-0306(03)71747-X

Sant'Anna AC, Paranhos da Costa MJR (2013) Validity and feasibility of qualitative behavior assessment for the evaluation of Nellore cattle temperament. *Livestock Science* **157**, 254–262. doi:10.1016/j.livsci.2013.08.004

Serrapica M, Braghieri A, Riviezzi AM, Bragaglio A, Carlucci A, Napolitano F (2014) Qualitative assessment of buffalo behaviour temporal fluctuations. *Italian Journal of Agronomy* **9**, 157–162. doi:10.4081/ija.2014.612

Stockman CA, Collins T, Barnes AL, Miller DW, Wickham SL, Beatty DT, Blache D, Wemelsfelder F, Fleming PA (2011) Qualitative behavioural assessment of cattle naïve and habituated to road transport. *Animal Production Science* **51**, 240–249. doi:10.1071/AN10122

Stockman CA, McGilchrist P, Collins T, Barnes AL, Miller DW, Wickham SL, Greenwood PL, Cafe LM, Blache D, Wemelsfelder F, Fleming PA (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. doi:10.1016/j.applanim.2012.10.016

Stockman CA, Collins T, Barnes AL, Miller DW, Wickham SL, Beatty DT, Blache D, Wemelsfelder F, Fleming PA (2013) Flooring and driving conditions during road transport influence the behavioural expression of cattle. *Applied Animal Behaviour Science* **143**, 18–30. doi:10.1016/j.applanim.2012.11.003

Stockman CA, Collins T, Barnes AL, Miller DW, Wickham SL, Verbeek E, Matthews L, Ferguson D, Wemelsfelder F, Fleming PA (2014) Qualitative behavioural assessment of the motivation for feed in sheep in response to altered body condition score. *Animal Production Science* **54**, 922–929.

Temple D, Dalmau A, Ruiz de la Torre JL, Manteca X, Velarde A (2011*a*) Application of the Welfare Quality protocol to assess growing pigs kept under intensive conditions in Spain. *Journal of Veterinary Behavior* **6**, 138–149. doi:10.1016/j.jveb.2010.10.003

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

Temple D, Manteca X, Dalmau A, Velarde A (2013) Assessment of test-retest reliability of animal-based measures on growing pig farms. *Livestock Science* **151**, 35–45. doi:10.1016/j.livsci.2012.10.012

Tuyttens FAM, de Graaf S, Heerkens JLT, 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? *Animal Behaviour* **90**, 273–280. doi:10.1016/j.anbehav.2014.02.007

Walker J, Dale A, Waran N, Clarke N, Farnworth M, Wemelsfelder F (2010) The assessment of emotional expression in dogs using a Free Choice Profiling methodology. *Animal Welfare (South Mimms, England)* **19**, 75–84.

Wemelsfelder F (1997) The scientific validity of subjective concepts in models of animal welfare. *Applied Animal Behaviour Science* **53**, 75–88. doi:10.1016/S0168-1591(96)01152-5

Wemelsfelder F (2007) How animals communicate quality of life: the qualitative assessment of behaviour. *Animal Welfare (South Mimms, England)* **16**, 25–31.

Wemelsfelder F, Millard F (2009) Qualitative Behaviour Assessment. In 'Assessment of animal welfare measures for sows, piglets and fattening pigs. Welfare Quality reports No. 10, Sixth Framework Programme'. (Eds B Forkman, L Keeling) pp. 213–219. (University of Cardiff: Cardiff)

Wemelsfelder F, Mullan S (2014) Applying ethological and health indicators to practical animal welfare assessment. *Revue Scientifique et Technique (International Office of Epizootics)* **33**, 111–120.

Wemelsfelder F, Hunter EA, Mendl MT, Lawrence AB (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. doi:10.1016/S0168-1591(99)00093-3

Wemelsfelder F, Hunter TEA, Mendl MT, Lawrence AB (2001) Assessing the 'whole animal': a free choice profiling approach. *Animal Behaviour* **62**, 209–220. doi:10.1006/anbe.2001.1741

Wemelsfelder F, Knierim U, Schulze Westerath H, Lentfer T, Staack M, Sandilands V (2009*a*) Qualitative Behaviour Assessment. In 'Assessment of animal welfare measures for layers and broilers. Welfare Quality reports No. 9, Sixth Framework Programme'. (Eds B Forkman, L Keeling) pp. 113–119. (University of Cardiff: Cardiff)

Wemelsfelder F, Millard F, De Rosa G, Napolitano F (2009*b*) Qualitative Behaviour Assessment. In 'Assessment of animal welfare measures for dairy cattle, beef bulls and veal calves. Welfare Quality reports No. 11, Sixth Framework Programme.' (Eds B Forkman, L Keeling) pp. 215–224. (University of Cardiff: Cardiff)

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

Wemelsfelder F, Hunter AE, Paul ES, Lawrence AB (2012) Assessing pig body language: agreement and consistency between pig farmers, veterinarians and animal activists. *Journal of Animal Science* **90**, 3652–3665. doi:10.2527/jas.2011-4691

Whitham JC, Wielebnowski N (2009) Animal-based welfare monitoring: using keeper ratings as an assessment tool. *Zoo Biology* **28**, 545–560.

Wickham SL, Collins T, Barnes AL, Miller DW, Beatty DT, Stockman CA, Blache D, Wemelsfelder F, Fleming PA (2012) Qualitative behavioral assessment of transport-naïve and transport-habituated sheep. *Journal of Animal Science* **90**, 4523–4535. doi:10.2527/jas.2010-3451

Wickham SL, Collins T, Barnes AL, Miller DW, Beatty DT, Stockman CA, Blache D, Wemelsfelder F, Fleming PA (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. doi:10.1080/10888705.2015.1005302