The Old School, Brewhouse Hill, Wheathampstead,
Hertfordshire AL4 8AN, UK

Animal Welfare 2011, 20: 549-558
ISSN 0962-7286

Development and testing of a novel instrument to measure health-related quality of life (HRQL) of farmed pigs and promote welfare enhancement (Part 2)

ML Wiseman-Orr†, EM Scott*† and AM Nolan‡

† School of Mathematics and Statistics, 15 University Gardens, University of Glasgow G12 8QW, UK
‡ College of Medical, Veterinary and Life Sciences, University of Glasgow G12 8QQ, UK
* Contact for correspondence and requests for reprints: Marian.Scott@glasgow.ac.uk

Abstract

The development of a novel structured questionnaire instrument to measure health-related quality of life (HRQL) in individual farmed pigs was described previously (companion paper). The instrument embraces the measurement of positive welfare, and was developed with farmers and stockpersons, for use by them on-farm. This paper describes the development of a scoring methodology for the instrument and provides evidence for its construct validity. Field testing on four commercial farm units indicated that scores for health and affect correctly allocated 88.7% of pigs to known treatment groups and strongly predicted previously defined intervention levels. The tool was also used in an experimental study alongside other measures to identify the impact of early-life challenges (mixing of pregnant gilts and tail-docking neonatal pigs) on subsequent pig welfare, and identified long-term changes in HRQL of prenatally stressed piglets, a finding supported by other measures. This work describes a novel approach to farm-level welfare assessment in which entirely animal-based HRQL measurement can provide a measure of welfare at the herd level while retaining information about individuals within the herd and about aspects of provision that can be targets of intervention to improve welfare, and promotes a move from welfare assurance to welfare enhancement.

Keywords: *animal welfare, farmed pigs, health-related quality of life, measurement, validity, welfare enhancement*

Introduction

As part of a research programme to investigate the welfare consequences for pigs of early-life adverse experiences, health-related quality of life (HRQL) was identified as an appropriate focus of integrative welfare measurement for farmed pigs, and one that would embrace the measurement of positive welfare. A novel structured questionnaire instrument was developed using a psychometric approach: evidence for content validity was presented and the instrument was found to have high utility when pre-tested on commercial farm units (Wiseman-Orr *et al* 2011). The instrument's items consist of animal-based observations commonly made by experienced farmers and stockpersons, addressing provision of the Five Freedoms (FAWC 2010). The instrument also addresses the direct assessment of affect, including positive affect. The instrument comprises 128 items, 98 that can be related to the Five Freedoms and considered to be causal for HRQL (such as 'coughing', 'tail bitten' and 'scouring'), and 30 variables that can be considered to be indicator variables for HRQL (such as 'lively', 'curious' and 'frightened'), assessing affect.

The psychometric approach, which was adopted to develop the pig HRQL instrument (Wiseman-Orr *et al* 2011), requires that instruments demonstrate properties of validity and reliability before being adopted for measurement purposes, and offers a range of approaches to such evaluation. Criticism has been levelled at instruments developed with insufficient attention paid to such properties and to utility (Abu-Saad 2001; Eiser & Morse 2001). Validity, evidence that the instrument is able to measure the construct that it was intended to measure, is the most fundamental attribute of a measurement instrument. Validation of any HRQL instrument is an iterative process, as new information is revealed for its use with new populations, in new contexts and for new purposes. In this process, instrument developers should seek evidence for validity of three kinds: content validity, criterion (or convergent or concurrent) validity, and construct validity (Fayers & Machin 2007; Streiner & Norman 2008).

Content validity, a measure of the extent to which an instrument's items are relevant and adequate for its purpose, was ensured for the instrument described in this paper by the method of generation, selection and scaling of the instrument items and by the results of pre-testing the prototype instrument which are described elsewhere in this issue (Wiseman-Orr *et al* 2011). Evidence for criterion validity is

provided when a novel measure can be compared with an existing gold standard or standards, where these exist. Where they do not, an important type of validity sought by instrument developers is construct validity. This can be established through a process of making and testing predictions about how the instrument will perform with particular groups or in particular circumstances. For example, do scores co-vary in predicted ways, and can the instrument discriminate between groups that are known to be different with respect to the construct of interest.

Any instrument is required to generate a numerical score or scores, and it is usual to generate them by combining multiple item responses to generate a composite indicator in such a way as to increase the reliability or precision of measurement. The metrological principles underlying the creation of such a composite score formed from sets of distinct, observable, behavioural components are found in the choice of the scaling model. A variety of scaling models exists, including direct or indirect estimation models (from Classical Test Theory [CTT]) and Item Response Theory [IRT]). All scaling models are techniques that allow weights to be devised for the items included in an instrument, reflecting the level of the construct of interest (eg HRQL) associated with the given item (Nunnally & Bernstein 1994). The way in which item responses are weighted and also combined should vary according to the relationship of the item to the construct being measured. In the case of indicator variables, which can be considered to be parallel measures of a single attribute (how the subject feels), it is usual to generate a score by computing the mean of individual item responses. However, most causal variables are likely to combine and interact in various ways, and it is therefore difficult and usually inappropriate to combine them in a simple model. In this case, it is recommended that an approach be taken in which deliberate and justified choices are made about how individual item responses should be combined to generate scores (Fayers & Machin 2007).

In this paper, the development of the scoring algorithm for a novel instrument to measure HRQL in farmed pigs is described and evidence for construct validity is presented. An approach for combining scores obtained with the instrument to form a measure of welfare at farm level is proposed.

Materials and methods

Scoring methodology

In the course of instrument development an impact assessment exercise was carried out which provided a measure of expert agreement on HRQL impact associated with each instrument item (Wiseman-Orr *et al* 2011). Briefly, an expert group (n = 29) of experienced farmers and stockpersons, pig veterinary specialists and welfare scientists indicated QOL impact associated with each item on a 100-mm Visual Analogue Scale anchored with 'quality of life could not be worse' (0) and 'quality of life could not be better' (100). The median score awarded by a group of farmers, veterinarians and welfare scientists provided an HRQL score for each causal or indicator item for instrument-scoring purposes. The item scores were then used to generate a profile of scores that addressed the Five Freedoms (using causal items) and affect (using indicator items). Scores were derived for one domain for each of Freedoms 1, 2, 4 and 5 (FF1-2 and FF4-5) and for eight sub-domains for Freedom 3 (FF3), which were arranged to accommodate items that described particular areas of the body or of health, and are presented in Table 1. The score for a domain or sub-domain was the lowest scoring item selected in that domain or sub-domain, or, if no items were selected with scores in the lower half of the range, the highest scoring item in that domain or sub-domain was chosen. In this way, good scores for one domain did not cancel poor scores for another. It was also important that a pig showing multiple negative signs (items with scores of 50 and below) within a domain was not given a less poor score than the worst-scoring item, which would happen if a number of poor scores were averaged. The reverse would be true for positive scores, whereby a pig showing a range of positive signs (being items with scores > 50), one of which had the highest score of any sign in that domain, would score lower than a pig showing only that one highly positive sign. An illustration of the algorithm applied to the sub-domain of Skin for six different pigs is shown in Table 2.

For the domain of affect, within which items were considered to represent parallel measurements of a single attribute, the domain score was generated by averaging the scores for all indicator items selected.

Construct validity

Use on commercial farms

Construct validity of the prototype HRQL tool was sought by means of field-testing the instrument on a sample of pigs on four commercial farm units that were contractor farms for a production company. All pigs were kept in straw-bedded large pens, with *ad libitum* feeding. Numbers of pigs per pen ranged from 51 to 244, with a median of 100. Two farms had three pens (with a total of 409, A, and 448 pigs, B), one had four pens (total of 521 pigs, C) and one, D, had seven pens (total of 786 pigs).

The sampling protocol, designed to reflect routine welfare assessment practice was as follows: the stockperson assessed every pig that was identified as warranting closer examination following a scan of all pigs on a walk though the house. From among all remaining pigs, a random sample of 20 pigs per farm unit (24 on the larger unit D) was assessed, allocated *pro rata* to each pen. These numbers were determined on the basis of the time available to conduct the assessments and were designed to minimise fatigue on the part of the assessor. Each pig assessed was spray marked to ensure that it was assessed only once, and remained at all times in the home pen.

Evidence was sought for the construct validity of the instrument, by testing the following hypotheses: (1) Expected relationships (poor scores for causal domains/sub-domains would be associated with poor scores for affect) would be observed between within-pig causal domain/sub-domain scores and affect scores; (2) The profile of scores generated

Table 1 Domains and sub-domains included in final instrument, with associated items.

Domains/sub-domains and associated items
FF1: Freedom from hunger and thirst (Causal)
Excessive queuing/fighting to eat
Excessive queuing/fighting to drink
FF2: Freedom from discomfort (Causal)
Sleeping with legs stretched out/lying on side
Looks comfortable
Panting
Huddled with other pigs
Hair standing up
Curled up tight, looking cold
Shivering
FF3: Freedom from pain, injury and disease (Causal)
Skin
Shine on skin
Good skin colour
Skin dull
Pale colour
Hairier than pen-mates
Anaemic
Respiration
Quiet, relaxed, regular breathing
Coughing
Panting
Breathing heavily
Breathing laboured
Mobility
Lively
Running about
Gets up fast
Slow to move
Not walking properly/stiff
Lame
Holding/saving a leg
Not weight-bearing on leg
Lethargic
Not getting up
Listless
Not moving
Discharges
Clear nose, no mucous
Mucous from the nose
Discharge from eyes/tear staining
Discharge from vulva
Blood discharge from wound or other orifice
Body condition
Thriving
Good body condition
Growing well
Looking hollow/empty
Slower growing (than others in group)
Unthrifty
Pot-bellied
Not growing at all
Losing condition
Injuries
No injuries
Abscess
Ear bitten
Ear swollen/haematoma
Damaged by navel sucking
Tail bitten
Flank bitten
Sore leg
Vulva bitten
Prolapsed
Hernia
Fresh (light/heavy) scratching
Visible wounds
Joint swollen
Sore foot
Injured
Appetite/digestion
Eating well
Interested in food
Off its food
Dehydrated
Faeces (normal/very hard/slightly loose)
Dung smeared over back end/on tail
Scouring
Vomiting
Blood in the scour
Miscellaneous
Looks well
Grinding teeth
Looks ill
Shaking
Lying on side, paddling legs
Normal posture
Holding head to one side
Clear, bright eyes
Sunken/deep-set eyes
FF4: Freedom to express normal behaviour (Causal)
Interacting with other pigs
Nosey
Rooting about normally
Grunting
Mounting
Fighting excessively
Abnormal biting/chewing/other behaviour
Outcast from group
Having insufficient room to move

Table 1 (cont)

Domains/sub-domains and associated items
FF5: Freedom from fear and distress (Causal)
Relaxed
Unafraid of stockperson
Squealing
Isolated (by penmates)
Bullied
Affect (indicator)
Lying contented and comfortable
Not looking right
Not eating well
Eating well
Lively
Listless
Alert
Running around happily
Lying on its own
Back is hunched
Squealing
Contented
Playing
Reluctant to get up
Curious
Slow moving
Hanging back
Interacting with other pigs
Nosing about
Grinding teeth
Friendly
Not interested in feed
Frightened
Inquisitive
Relaxed
Head is down
Bright
Nervous
Not interested in surroundings
Communicating with low grunting

(scores for causal domains and for affect) would discriminate between pigs that were not of concern and those that were of concern (known groups); and (3) Expected relationships would be observed between the profile of scores generated and the stockperson's choice of appropriate intervention (selected from among five predefined options, ranging from 'having no concerns: check daily', to 'having specific concerns: take action immediately').

Use in experimental study

With existing evidence for validity (face, content and construct), the HRQL instrument was used in a study designed to investigate the potential impact of early life challenges on the long-term health and welfare of pigs. The measure of HRQL was one of a series of measures used to monitor pigs following pre-natal stress (dams exposed to mixing during second trimester of pregnancy) and post-natal pain (tail docking) and their interaction.

Pre-natal stress: Stress-treatment gilts were exposed to social mixing during the second trimester of pregnancy. This involved moving the gilts into a new pen with three older multiparous sows on two different occasions between 39–45 and 59–65 days of pregnancy. This procedure causes profound social defeat, and as a consequence is highly stressful (Jarvis *et al* 2006). In the intervening period between mixing, the sows were returned to their home pens. Control gilts remained in their home pen undisturbed throughout the 114-day gestation period.

Tail docking/sham docking: Half of each sex group within each litter from the two maternal treatment groups (mixed [stressed] [M]/unmixed control [C]) were tail docked (D) 2–4 days post farrowing. Approximately one-half (3 cm) of the tail was removed using sterile surgical cutters. Typically, the tail was cut around the fifth/sixth coccygeal vertebra. Sham-docked pigs (I) were handled in the same way as the docked pigs, but without surgical amputation of a portion of the tail.

Piglets were farrowed into individual commercial farrowing crates in which they were able to move around the gilt and had access to a sheltered creep area providing additional heat. Approximately 28 days after farrowing, the sow was removed from the farrowing crate. The newly weaned piglets were kept in this environment for 2–3 days before being transferred to growing pens for the remainder of the experimental period.

At approximately eight weeks of age a small number of piglets per litter were subjected to nociceptive threshold testing (Sandercock *et al* 2009), and soon after that the litter was moved, within the same site, to commercial farm accommodation consisting of straw-bedded small pens designed to hold 20 pigs, with *ad libitum* feeding. Larger litters were not mixed, but on moving onto the farm, smaller litters were mixed with other litters that were in the same mixed or control treatment group. The number of pigs accommodated in each pen on farm ranged from 7–17 (median of 13).

Pigs in the four treatment groups were measured using the novel HRQL instrument, on farm, when aged 14–16 weeks. Assessments were carried out by one experienced female stockperson who was blinded to whether pigs were in the stressed or sham group but could not be blinded to whether or not pigs were tail docked. Pigs were individually assessed in their home pens, in accordance with routine welfare assessment practices on farm. One hundred and sixty-four pigs in four treatment groups were measured once: stressed/docked (MD) (23 pigs), stressed/sham (MI) (26 pigs), control/docked (CD) (36 pigs) and control/sham (CI) (79 pigs).

Table 2 Sample illustration of how scoring algorithm would be applied to instrument responses for six different pigs to generate a sub-domain score for skin for each pig.

Pig	Shine on skin (item score 79.5)	Good skin colour (item score 75)	Skin dull (item score 39)	Pale colour (item score 27.5)	Hairier than pen-mates (item score 24)	Anaemic (item score 17.5)	Sub-domain score for skin
A	↑						79.5
B	↑	↑					79.5
C		↑					75
D			↑		↑		24
E					↑		24
F			↑				39

Statistical analysis

Analysis of the data was carried out using Minitab v 15 and the same statistical methods were used in both studies. Regression analysis was used to examine the relationship between causal domain scores and affect, and a best subsets' regression analysis was conducted to examine which variables best predicted the affect score. Discriminant analysis (with cross-validation) was used to test whether a profile of scores for causal domains and affect could discriminate known groups. An ordinal logistic regression was used to examine the relationship between the profile of scores and the intervention level for each pig assessed.

Table 3 Mean, minimum, median, maximum and quartile 1 and 3 scores for skin, mobility, body condition, injuries, miscellaneous health signs, and affect, for all pigs assessed on commercial farm units (n = 157).

	Mean	Min	Q1	Median	Q3	Max
Skin	72.73	17.5	75.0	79.5	79.5	79.5
Mobility	67.3	13.5	76.5	79.5	79.5	79.5
Body condition	71.45	10.0	47.0	88.5	88.5	88.5
Injuries	73.48	20.0	66.5	83.0	83.0	83.0
Miscellaneous health signs	83.13	14.0	86.0	86.0	86.0	86.0
Affect	77.64	23.5	83.17	83.55	83.75	85.17

Results

Scoring methodology

Scores for each causal item ranged from 3.5 to 88.5, with a median score of 27. The scores for positive indicator items ranged from 67.5 to 88 and the scores for negative items ranged from 13 to 38.

During instrument development, items were selected to provide adequate and relevant content and some items were deemed by expert respondents to be appropriate to both causal and indicator domains. However, preliminary analysis of data indicated that the scores for the FF3 sub-domain of Mobility and for the domains FF4 and FF5 had a particularly strong relationship with the overall score for affect computed from the instrument's 30 indicator items, which may have resulted from the inclusion of the same items within those causal domains and the affect domain thus inflating any relationships between them. It was therefore decided to exclude such items from the generation of scores for the relevant causal domains and sub-domain. As a result, there were insufficient items remaining within FF4 (two items) and FF5 (three items) to generate a valid score for those domains.

FF1 contained only negative items (queuing/fighting to eat or drink) and this was not noted or not recorded for any pig so scores for FF1 were not able to be included in the analysis. (However, the relevant items and responses to supplementary — unscored — questions about availability of feed and water revealed that all assessed pigs had freedom from hunger and thirst). Analysis used scores for the domains or sub-domains (out of 10 remaining) with fewest missing values. In the case of data generated during commercial farm testing, these were: skin, mobility, body condition, injuries and miscellaneous health signs, and affect. Using data generated for experimental pigs, these were: respiration, skin, mobility, body condition, injuries, appetite/digestion, miscellaneous health signs, and affect.

Construct validity

Use on commercial farms

A total of 157 pigs across the four farm units were assessed by one experienced stockperson. All pigs that were identified by the stockperson as 'of concern' were assessed: 16, 11, 15 and 23 pigs on farms A, B, C and D, respectively. Twenty pigs were randomly selected from the remaining pigs which were 'not of concern' on farms A, B and C while 24 such pigs were assessed on farm D.

The computed (indicator items) or otherwise derived (causal items) scores for domains and sub-domains of skin, mobility, body condition, injuries, miscellaneous health signs, and affect (Table 3) were used to explore hypothesised relationships that would provide evidence for the construct validity of the instrument.

Hypothesis 1 — Expected relationships would be observed between within-pig causal domain scores and affect scores

Regression analysis used best subset regression to examine the relationship between multiple causal sub-domain scores (for skin, mobility, body condition, injuries and miscellaneous health signs) and for affect. This showed (with highest R-Sq [adj] of 50.1%, and lowest Mallows Cp of 0.7) that the best model to predict affect using available data was one containing skin and mobility.

Table 4 Summary of classification with cross-validation in known groups analysis using data from pigs assessed on commercial farm units (using scores profiles for 97 pigs).

	True group	
Put into group	**1**	**2**
1	22	0
2	11	64
Total n	33	64
N correct	22	64
Proportion	0.667	1.000

Hypothesis 2 — The profile of scores generated (for causal domains and for affect) would discriminate pigs that were randomly sampled from the population that were not of concern and a selected sample of pigs that the stockperson wanted to take a closer look at.

Discriminant analysis was used to test whether a profile of scores for causal sub-domains (skin, mobility, body condition, injuries and miscellaneous health signs) and affect could discriminate a selected sample of pigs that the stockperson wanted to take a closer look at (33 pigs: Group 1) and pigs that were randomly sampled from the population that were not of concern (64 pigs: Group 2). A summary of classification with cross-validation is shown in Table 4. Scores for those five causal sub-domains and the affect domain correctly allocated 66.7% of the pigs that were looked at more closely and 100% of the pigs that were randomly sampled, with an overall correct classification rate of 88.7%. This provided further evidence for the construct validity of the instrument.

Hypothesis 3 — Expected relationships will be observed between the profile of scores generated and the intervention level and global assessment of QOL

Ordinal logistic regression analysis was carried out using the scores for skin, mobility, body condition, injuries and miscellaneous signs of health, and the score for affect, to predict intervention level from 0 representing least intervention to 4 representing most intervention as determined by the stockperson using pre-defined options (Table 5). This provided evidence of statistically significant relationships for body condition and affect with intervention level. The *P*-value for the tests that all slopes are zero was 0.000 (Log-likelihood = –17.741; Test that all slopes are zero: $G = 66.874$, df = 6, *P*-value = 0.000). The Sumers' D and Goodman-Kruskal Gamma measures were both 0.96, indicating strong predictive ability (these measures lie between 0 and 1 with higher values meaning more predictive ability). This expected relationship between the domain and sub-domain scores and the intervention level required provides further evidence for the construct validity of the instrument.

Use in experimental study

Using the scores for nine domains and sub-domains (including respiration, skin, mobility, body condition, injuries, appetite/digestion, miscellaneous health signs, and affect), the proportion of animals that could be correctly allocated to the control (C) or mixed (M) groups was 65%, and to the docked (D) or sham docked (I) groups was 57%. Allocation to four known groups MD/MI/CD/CI was only 30% (raising the possibility of a complex interaction effect between the two treatments).

The HRQL domains and sub-domains performed extremely well (88.5% correctly classified) when used to classify the pig as needing intervention or not. The score for affect accounted for ~70% of this intervention proportion correctly classified, and indeed pigs identified as belonging to the 'intervention' group, had a mean affect score lying between 7.7 and 15.2 lower than the pigs identified as not requiring intervention.

Discussion

Two principal types of item in HRQL measurement have been proposed — items capturing causal variables and those capturing indicator variables. According to Fayers and Hand (2002), indicator variables do not alter or influence the underlying concept, they are merely indicators of its magnitude. Indicator variables are assumed to be multiple parallel tests measuring a single construct with an underlying continuum that is uni-dimensional: affect, or how the pig feels. All indicator items are assumed to be contributing equally to that measurement so a score for affect may be computed by averaging item responses, which should reduce the effect of individual errors and thereby increase the reliability of the measurement.

Causal variables, on the other hand, which reflect living conditions, including the physical and social environment, and symptoms of illness or injury, are likely to have an effect — either positive or negative — on HRQL. These variables are not measures of QOL *per se* but their presence or magnitude tells us about circumstances that are likely to have an impact upon QOL. Sometimes a low score on just one potential causal variable can be sufficient to produce the outcome value of an HRQL instrument — it can be a sufficient component cause of poor welfare (Fayers & Machin 2007). Consequently, a QOL instrument should be designed so that a poor score on any one of the causal variables leads to a low QOL score. Since most causal variables will interact with HRQL in a complex way, clear protocols must be determined with regard to generating scores from responses to causal items. For example, in most cases, assessed pigs showed either positive or negative signs (not both) for each domain or sub-domain. Where incompatible positive and negative signs were selected, respondent error was assumed and only the negative sign was considered, to ensure that any potential welfare compromise was not overlooked. This kind of explicit justification for combining item responses to generate scores has been recommended for development of measures of human QOL (Fayers & Machin 2007).

Table 5 Results of ordinal logistic regression to examine relationship between profile of domain scores and intervention level for individual pigs (n = 97) assessed on commercial farm units.

Predictor	Coefficient	SE	Coefficient Z	*P*-value	Odds ratio	Lower	Upper
Const(1)	–19.0014	6.11441	–3.11	0.002			
Const(2)	–12.8750	5.14588	–2.50	0.012			
FF3 skin	–0.0352568	0.0570783	–0.62	0.537	0.97	0.86	1.08
FF3 mobility	0.0354613	0.0300024	1.18	0.237	1.04	0.98	1.10
FF3 body condition	0.0945916	0.0314547	3.01	0.003*	1.10	1.03	1.17
FF3 injuries	–0.0018421	0.0290111	–0.06	0.949	1.00	0.94	1.06
FF3 miscellaneous	0.0310720	0.0559584	0.56	0.579	1.03	0.92	1.15
Affect	0.159980	0.0466825	3.43	0.001*	1.17	1.07	1.29

* $P < 0.05$.

It has been suggested that causal item scores should be weighted (Fayers & Machin 2007), the weights derived from an expert rating of the importance and severity of each potential QOL impact. Such ratings were obtained in this study using an expert group and these impact assessments contributed to the scoring mechanism for the instrument. This represents an advance in farm animal welfare assessment which, to-date, has depended upon simple counts of subjects (such as pigs requiring hospitalisation or demonstrating oral behaviours) or scores of single attributes (such as body condition or lameness scoring). An important advance made recently by the Welfare Quality® project was to devise a scoring methodology which transposed observations in each relevant area onto a 0–100 scale, with 0 corresponding to a situation where welfare cannot be lower and 100 corresponding to a situation where welfare cannot be improved further. This process of calibration was achieved by consultation with five or six animal scientists who were involved in the choice and development of the Welfare Quality® measures, but future recalibration involving more experts is intended (Botreau *et al* 2008).

The instrument provided good discrimination between known groups (88.7% correct allocation), with results that are comparable to those for an observational instrument to measure pain in communicatively impaired children which was able to correctly classify 87.4% of pain/no pain episodes (Stallard *et al* 2002). While 100% of pigs of no concern were correctly allocated, only 66.7% of pigs selected for closer examination were correctly allocated using instrument scores. For use on-farm to identify pigs with poor welfare, better sensitivity would be required of the instrument. However, it should be noted that a number of the pigs were placed in the latter group only because they were smaller than other pigs in the pen and closer inspection of three of those pigs, and one other in the group, revealed no health or welfare problems. This is reflected in the scores for those pigs and accounts for four of the eleven misallocations, increasing to 75.9% the number of pigs of concern that could be correctly allocated to that group using instrument scores. Furthermore, not all domains and sub-domains were able to be included in the analysis due to missing data: it is expected that a more complete profile of scores would demonstrate better discrimination.

In the case of the experimental pigs, evidence from a range of other measurements made on the same group of pigs indicates that pre-natal stress but not tail docking is associated with long-term risks to pig welfare (Rutherford *et al* 2011) so the 'incorrect' allocation to known groups reported in the results of this study is likely in some part to be accounted for by a lack of difference in HRQL of pigs in the tail-docked and sham-handled groups. HRQL is also likely to be influenced by other factors, such as mixing of some litters when moving onto the farm accommodation, different numbers of pigs in each farm pen, and the social influences of each group, thereby confounding the results. Agreement between different measures provides some evidence for the criterion validity of the novel instrument.

It was not possible to include all domain scores in this analysis because of missing values for some domains. While it had been intended that a domain or sub-domain score would be produced for every sub-scale of the instrument, many pigs had missing values within some domains or sub-domains. This was a result of the design of the prototype instrument which meant that the respondent was not obliged to select at least one item for every domain. Consequently, not all domains and sub-domains could be included in the planned analysis for data from commercial farm units or from experimental data. Furthermore, it was necessary, following preliminary analysis, to remove some items from FF4 and FF5 domains, which left too few items in those domains to compute a valid score. Future development of the instrument should include new causal items for those domains because it is essential that all relevant causal domains are sampled during measurement. The availability of a more complete profile of scores for all pigs will permit an examination of the relationships between scores for all causal domains and for affect, and of the ability of the more complete profile of scores to discriminate known groups.

The problematical free-choice approach to item selection also meant that some error was introduced when the non-

selection of an item resulted from an oversight and not because that observation had not been made. A forced-choice format which requires respondents to select at least one item from all items addressing a particular domain is unlikely to decrease utility and would ensure that a score for each domain is available for all pigs assessed. Refining the instrument in this way will be an immediate consideration. Such an approach would require that each domain or subdomain include appropriate items from which a selection could be made (eg FF1 would require the addition of at least one positive item that could be selected if negative items were not applicable).

The evidence that some domain scores (eg body condition and affect) can be used to predict intervention level as judged by an experienced stockperson links those two outcomes in an important, practical sense and highlights the opportunity to use the instrument as an educational as well as a measurement tool, to guide the inexperienced stockperson to take appropriate action with regard to welfare

The limitations imposed on the field-testing protocol by practical considerations are recognised; these meant that intervention level and allocation to known groups were choices made by the instrument respondent. Future testing will be designed so that pigs are independently allocated to known groups using agreed criteria for high and low HRQL, and intervention levels for each pig are similarly independently selected by an experienced stockperson. In the absence of an existing gold standard measure for pig HRQL, criteria for independently allocating pigs to groups with high or low HRQL might include relevant existing measures (eg body condition score, lameness score) which would also serve to provide some evidence for the criterion validity of the novel instrument.

Although it would be likely to reduce utility for on-farm use, for research purposes future refinement of the instrument might include extending the rating response options to appropriate indicator variables which would facilitate the use of statistical methods such as factor analysis and IRT.

Future testing of the instrument should include testing of its reliability when used by different operators assessing the same pig or pigs, and when used by the same operator assessing the same pig on different occasions when its HRQL is unchanged (which can be achieved using high-quality video recordings). Finally, instrument utility would be significantly improved by electronic and mobile delivery, scoring and storage of data.

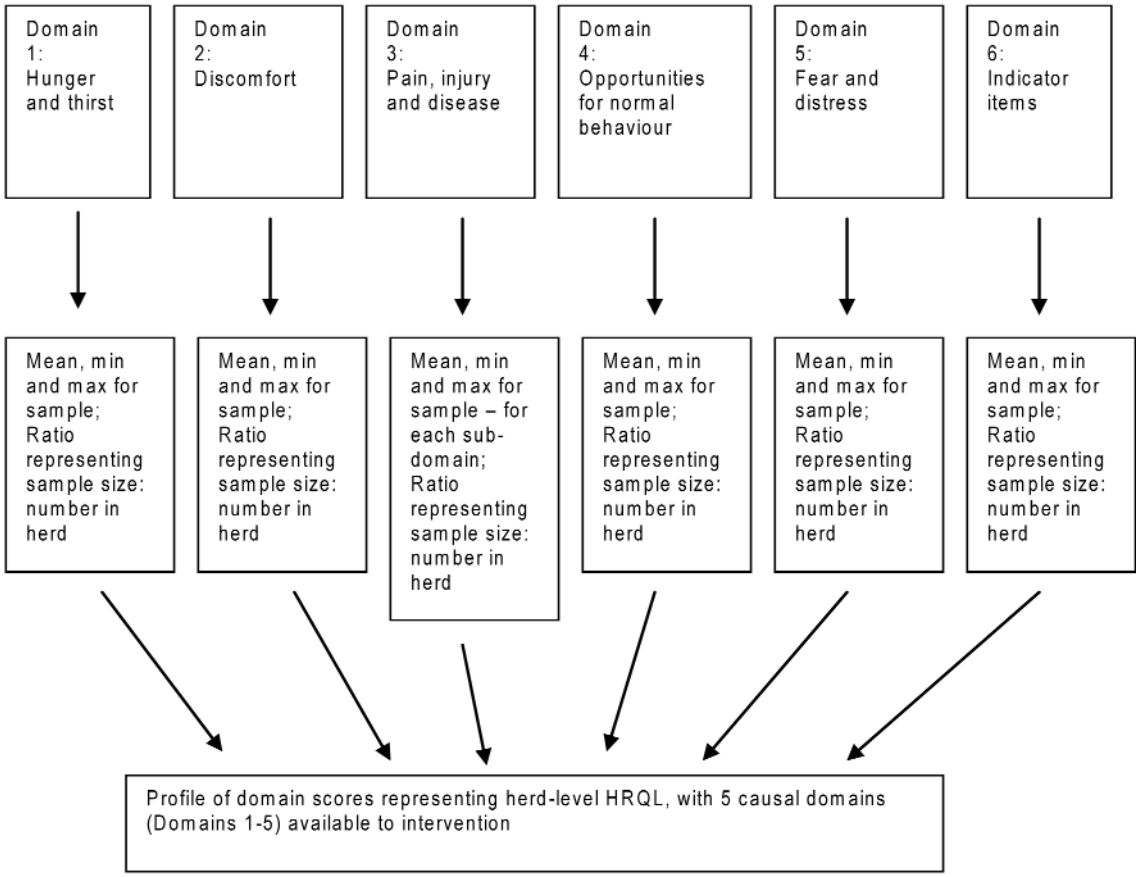
Since HRQL is an individual's experience of its circumstances, measurement of HRQL must always be made at the level of the individual (animal-based measurement). Scores for individuals can then be used to achieve a group measure while retaining the variability in individual scores. This information is important since it has been argued that the welfare of a population of animals should be measured by the welfare of the one animal in the population with the poorest welfare (Dewey 2008). The information provided by scores distributions for domains and sub-domains can not only allow an assessment of welfare at the farm level but can also serve to direct attention to specific issues in relation to which changes in environmental conditions or management practices could improve commercial performance as well as welfare on low-performance farms. The usefulness of such management information emerging from welfare assessment is recognised (Main 2009). One way in which presentation of individual-level measurement could provide a group-level measure without losing individual- or domain-level information, would be to use a summary measure, broadly based upon the approach of social scientists to developing indices of multiple deprivation. Such a measure is illustrated in Figure 1.

Designing a tool for use by those individuals most closely engaged in caring for pigs was a specific strategy adopted in the development of this tool. This was driven by a belief that moving from audit-based assurance-type approaches to one of engaging farmers and stockpersons in welfare enhancement would improve use of the tool. In the field of education there has been a cultural shift in recent years from quality assurance to quality improvement, and further to 'quality enhancement' in which a 'quality culture' is embedded and owned (Quality Assurance Agency for Higher Education [QAA] 2003; Learning and Skills Network 2009) and which recognises that further improvement is always possible. The gains in animal welfare that were seen in the latter part of the 20th century and the very early years of the 21st century can be likened to the first step in this process — from quality assurance (a safety net below which standards cannot be allowed to fall) to quality improvement (which seeks higher than minimum welfare standards). We propose that the evolution towards quality enhancement can be seen as a continuation of this improvement process, and requires that contributors to the process take ownership of it. That fundamental emphasis was the tenet for the design of this novel instrument; it was built upon the knowledge and expertise of farmers and stockpersons, made possible by their interest in pig welfare, and designed to be used by them in the context in which they work. The initial phase of testing reported here provides preliminary evidence for the validity of the instrument, and also for this approach. With refinement and further testing of validity, and with evidence for reliability, it will provide an opportunity for farmers and stockpersons to 'own' formal welfare assessment, and the improvement and welfare enhancement that society is beginning to demand.

Animal welfare implications and conclusion

The HRQL instrument described in this paper was designed for the measurement of individual pigs on-farm as a robust, routine welfare measure for a range of purposes. These include education and guidance of inexperienced stockpersons, welfare benchmarking, identification of problems at the individual and herd level, and assessment of intervention impact. The results of pre-testing and field-testing to-date have provided some evidence for its validity for these purposes, and have directed attention to areas in which slight modification should further improve the measurement properties of the prototype. With further testing and instru-

Figure 1



Representation of a summary measure in which HRQL scores for individual animals may be combined to create a measure of herd-level welfare, while retaining information at the individual and at the domain/sub-domain level.

ment refinement, and development of electronic presentation and output, the instrument can be used by farmers and stockpersons to guide welfare-related decision-making on a day-to-day basis, to contribute to group-level welfare measurement at single time-points and over time, and to compare the welfare impact of management practices and targeted interventions. Routine use of the instrument for educational purposes may itself have a positive impact upon animal welfare, and industry involvement throughout its development makes more likely its adoption and success as a welfare enhancement tool.

Acknowledgements

We are grateful to the farmers, stockpersons, veterinarians and welfare scientists who contributed to this project, and to staff at the Scottish Agricultural College, Edinburgh, UK, for their invaluable assistance. This research was supported by the Biotechnology and Biological Sciences Research Council as part of a project titled 'Perinatal programming of stress responses, nociceptive mechanisms and the welfare consequences in pigs (grant reference number BBSRC BB/C518965/1).

References

Abu-Saad HH 2001 Commentary. *Archives of Disease in Childhood Fetal and Neonatal Edition 85*: F40-F41

Botreau R, Capdeville J, Perny P and Veissier I 2008 Multicriteria evaluation of animal welfare at farm level: an application of MCDA methodologies. *Foundations of Computing and Decision Sciences 33(4)*: 287-316

Dewey CE 2008 Assessing the health status of populations of animals in relation to welfare. *Book of Abstracts of the 4th International Workshop on the Assessment of Animal Welfare at Farm and Group Level*. WAFL: Ghent, Belgium

Eiser C and Morse R 2001 A review of measures of quality of life for children with chromic illness. *Archives of Disease in Childhood 84*: 205-211

FAWC 2010 *Five Freedoms*. Available at http://www.fawc.org.uk/freedoms.htm. (Accessed 28 May 2010)

Fayers P and Hand DJ 2002 Causal variables, indicator variables and measurement scales: an example from quality of life. *Journal of the Royal Statistical Society 165*: 1-22

Fayers PM and Machin D 2007 *Quality of Life: The Assessment, Analysis and Interpretation of Patient-Reported Outcomes, Second Edition*. Wiley: Chichester, UK

Jarvis S, Moinard C, Robson SK, Baxter E, Ormandy E, Douglas AJ, Seckl JR, Russell JA and Lawrence AB 2006 Programming the offspring of the pig by prenatal social stress: neuroendocrine activity and behaviour. *Hormones and Behaviour 49*: 68-80

Learning and Skills Network 2009 *Evaluation of the Scottish Funding Council's strategy for quality enhancement in the college sector; Annual Report, year one.* Learning and Skills Network: London, UK

Main DCJ 2009 Application of welfare assessment to commercial livestock production. *Journal of Applied Animal Welfare Science 12*: 97-104

Nunally JC and Bernstein IH 1994 *Psychometric Theory*. McGraw Hill: New York, USA

Quality Assurance Agency for Higher Education (QAA) 2003 *Handbook for Enhancement-Led Institutional Review: Scotland*. Available at http://www.qaa.ac.uk/reviews/elir/handbook/scottish_hbook_preface.asp. (Accessed 17 August 2010)

Rutherford KMD 2011 Early influences on welfare in the domestic pig. *Oral presentation at BBSRC Animal Welfare Programme Dissemination Workshop*. 22 February 2011, London, UK

Sandercock DA, Gibson IF, Brash HM, Rutherford KMD, Scott EM and Nolan AM 2009 Development of a mechanical stimulator and force measurement system for the assessment of nociceptive thresholds in pigs. *Journal of Neuroscience Methods 182*: 64-70

Stallard P, Williams L, Vellman R, Lenton S, McGrath PJ and Taylor G 2002 The development and validation of the pain indicator for cognitively impaired children (PICIC) *Pain 98*: 145-149

Streiner DL and Norman GR 2008 *Health measurement Scales: A Practical Guide to Their Development and Use, Fourth Edition.* OUP: Oxford, UK

Wiseman-Orr ML, Scott EM and Nolan AM 2011 Development and testing of a novel instrument to measure health-related quality of life (HRQL) of farmed pigs and promote welfare enhancement (Part 1). *Animal Welfare 20*: 535-548

Health surveys urban health health status indicators quality of life environmental health socioeconomic factors health behavior europe.
Â Pilot: it is always necessary to test the questionnaire, methodology and analysis of a survey on a small sample before undertaking the main survey. The pilot assists in anticipating problems, for example, if the questionnaire is too long, the respondents not easy to contact, the results difficult to ana-lyse and allows alternative choices to be made.