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

Animal Welfare 2007, 16(S): 17-24
ISSN 0962-7286

Can we really measure animal quality of life? Methodologies for measuring quality of life in people and other animals

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

* Department of Statistics, University of Glasgow, Glasgow G12 8QW, UK
† Institute of Comparative Medicine, Faculty of Veterinary Medicine, University of Glasgow, Glasgow G61 1QH, UK
* Correspondence: marian@stats.gla.ac.uk

Abstract

Quality of life (QoL) is an abstract construct that has been formally recognised and widely used in human medicine. In recent years, QoL has received increasing attention in animal and veterinary sciences, and the measurement of QoL has been a focus of research in both the human and animal fields. Lord Kelvin said "When you cannot measure it, when you cannot express it in numbers — you have scarcely in your thoughts, advanced to a stage of science, whatever the matter may be" (Lord Kelvin 1893). So are we able to measure animal QoL? The psychometric measurement principles for abstract constructs such as human intelligence have been well rehearsed and researched. Application of traditional and newer psychometric approaches is becoming more widespread as a result of increasing human and animal welfare expectations which have brought a greater emphasis on the individual. In recent decades the field of human medicine has developed valid measures of experienced pain and QoL of individuals, including those who are not capable of self-report. More recently, researchers who are interested in the measurement of animal pain and QoL have begun to use similar methodologies. In this paper, we will consider these methodologies and the opportunities and difficulties they present.

Keywords: *animal quality of life, animal welfare, dog, measurement, psychometric, statistical methods*

What is animal quality of life?

The crucial first step in measurement is to be clear about what is to be measured. So what is quality of life (QoL)? For people, it has been suggested that in the social sciences QoL consists of "objective living conditions and subjective satisfaction with them" while in medicine the same term describes "the health related subjective well-being of the individual" (Birnbacher 1999). In the measurement of animal welfare, the former conceptualisation may be useful for farm animals, where the principal interest is in the effects of standard conditions on groups. The latter conceptualisation may be more appropriate for companion animals, where the focus is on the individual whose circumstances are likely to be unique.

In the field of human health, one frequently referenced definition of QoL devised by a World Health Organisation (WHO) international research group tasked with developing QoL measures for people is: "an individual's perception of their position in life in the context of the culture and value systems in which they live and in relation to their goals, expectations, standards and concerns. It is a broad ranging concept affected in a complex way by the person's physical health, psychological state, personal beliefs, social relationships and their relationship to salient features of their environment" (WHOQOL Group 1995). This definition makes clear that while QoL is *affected by* an individual's physical, psychological, and social state, it *is* 'an individual's perception'. This differentiates QoL from health, and reflects the view that QoL is complex and subjective and can only properly be measured from the individual's perspective.

The term 'welfare' has been used in both the human and animal fields wherever we have responsibility for others. We would define welfare (and similarly well-being) as a complex construct that combines both subjective and objective aspects of the conditions of life for animals (Scott *et al* 2003), for example, as described in the 'Five Freedoms' of the Farm Animal Welfare Council (FAWC 1992). Although the importance to its welfare of how the animal feels is now widely recognised, practical welfare measurement is still most usually concerned with ensuring that minimum standards of care are provided. If we adopt for animals a conceptualisation and definition of QoL that is similar to that for people, then consideration of QoL in veterinary and animal science offers the opportunity to make explicit the importance of the individual's perspective, and to significantly shift the focus in animal welfare, from the avoidance of poor QoL to the attainment of good or even excellent QoL.

Few formal definitions of animal QoL have been published. Some of these are similar or identical to definitions of animal welfare in their inclusion of factors that are presumed to affect QoL, as well as inclusion of the individual's response to those factors. Others (eg McMillan 2005), including our own (Wiseman-Orr *et al* 2006), exclude from the definition that which can affect QoL, leaving only the animal's evaluation and affective response.

https://doi.org/10.1017/S0962728600031687 Published online by Cambridge University Press

Our definition (Wiseman-Orr *et al* 2006) is one that we consider to be applicable to man and to other animals:

> Quality of life is the subjective and dynamic evaluation by the individual of its circumstances (internal and external) and the extent to which these meet its expectations (that may be innate or learned and that may or may not include anticipation of future events), which results in, or includes, an affective (emotional) response to those circumstances (the evaluation may be a conscious or an unconscious process, with a complexity appropriate to the cognitive capacity of the individual).

Why should we measure quality of life?

QoL measurement can tell us how subjects feel about their circumstances. In the past two decades the medical profession has recognised the importance of such measurement, however difficult this might be. A PubMed search matching 'quality of life' and 'questionnaires' returned 137 references in the decade to 1986, and 10 787 references in the decade to 2004 (Emery *et al* 2005). Another bibliographic study of patient-assessed health outcomes recorded 144 reports on development and testing of QoL measures in 1990, and 650 reports in 1999 (Garratt *et al* 2002). The exponential growth in the development and use of QoL measures for people in recent years, in spite of the difficulty of such measurement, is driven by a strongly perceived need for such tools. They are used to aid decision-making for individual patients, and are increasingly used for measuring health outcomes in evaluative research, with funding organisations demanding QoL measurement as a potential endpoint in clinical trials (Fayers *et al* 1997). The WHO has funded the development of QoL measurement tools that are applicable across a diversity of cultures internationally (Skevington *et al* 1997; Skevington 1998). The QoL of people is also measured in non-medical circumstances in which individuals are the responsibility of others, such as in care homes for the elderly. Here too, the focus is shifting from simply measuring the provision of basic care to all residents, to include the experiences of individuals (eg Innes & Surr 2001). Many such individuals may be unable to complete self-report QoL instruments. For them, and for other non-verbal or cognitively impaired people in the care of the medical profession or others, an alternative to self-report must be used.

In the case of suffering caused by pain, it has been argued strongly (Anand & Craig 1996; Cunningham 1999) that those who do not have a voice have a greater need for robust measurement than do those who can more easily make their views known. Animals are, of course, among those who cannot tell us how they feel.

The current emphasis on evidence-based veterinary medicine requires that appropriate measures of clinical impact be developed. The measurement of QoL will facilitate decision-making on the treatment of individual animals, whether active or palliative, including decisions regarding the appropriateness of euthanasia. Such measurement can also be used in clinical trials, to judge the effectiveness of one treatment compared with another, or with none. These are important welfare issues, particularly as treatment options increase for companion animals, some associated with short- or long-term negative impact. Greater choice of treatment options and increased affordability require demanding ethical decision-making in veterinary practice. An instrument that can be used with confidence to monitor clinical change in an individual, and to provide data that will facilitate the selection of treatments with known effectiveness and impact, will reliably inform such decision-making and so benefit patient, client and veterinary practitioner.

There are circumstances other than impaired health that may affect the QoL of companion animals. For example, long-term kennelling of rescue dogs may compromise the QoL of some individuals to an unacceptable degree. The QoL of farm animals may be impacted by husbandry and housing.

How should we measure quality of life?

What are the types of measurement, and what properties do we want them to have?

The role of measurement is to assign numerical values to the attribute of interest or to classify an object on the basis of that attribute — to quantify or categorise the QoL of the animal in such a way that we can have confidence in the derived measure.

Measurement may have nominal, ordinal, interval or ratio scale properties. The least information is provided by a nominal scale, which simply tells into which category a response falls. More information is provided by an ordinal scale, which ranks response options, providing information about how these relate to one another in relative terms. On an interval scale, response options are made on a scale of equal units, and a ratio scale has, in addition, a meaningful zero.

Ordinal and interval level measurements are both practicable and desirable for the assessment of QoL. Ordinal scales have been the most frequently created yet they present difficulties in analysis and interpretation. They may offer the precision required, although their sensitivity and responsiveness to change is compromised if the ordered categories are broad. There must be a careful consideration of the number and definition of ordered levels. Interval level measurement is more demanding to create, but provides more precise measurement; the rank ordering of the individuals is known, and the distance separating the individuals on the attribute is also known.

Measurement of abstract constructs using psychometric methodology

If QoL is conceptualised as an entirely subjective construct, then the goal of its measurement is to access that subjective perception (Stenner *et al* 2003). Consequently, for people, the gold standard method of measuring QoL is the self report, using a structured questionnaire instrument that is subjected to formal assessment.The methods used to

https://doi.org/10.1017/S0962728600031687 Published online by Cambridge University Press

develop such questionnaires were originally established by psychologists and psychiatrists to measure abstract, multiple-attribute constructs such as intelligence and personality. These same approaches have been adopted in the medical field to develop instruments to measure the abstract, subjective construct of QoL.

The processes necessary for the creation of psychometric instruments are well established (Streiner & Norman 1995) and may be described in three phases. Phase 1 involves the specifying of measurement goals (and hence the ideal measurement scale), the identification of the patient population, and the development of a pool of potential items for inclusion in the instrument. In Phase 2, suitable items are selected from the item pool and that selection is subjected to expert validation. The validated collection of items is then incorporated into an instrument, with suitable consideration given to layout, response option(s), instructions to respondent and administration. The resulting prototype is then pretested to ensure that the target respondent can use the instrument correctly. Phase 3 involves field-testing the instrument, in order to evaluate its psychometric properties (Streiner 1993; Streiner & Norman 1995; Juniper *et al* 1996). The important contribution of the psychometric approach to instrument development is widely recognised (Cook *et al* 2003), and has led to the creation of a number of instruments for the valid measurement of QoL in the health field (Garratt *et al* 2002; Matza *et al* 2004).

The psychometric approach requires that instruments demonstrate the psychometric properties of validity, reliability and, usually, responsiveness, before being adopted for clinical use, and offers a range of methods for such evaluation. Criticism has been levelled at instruments developed with insufficient attention paid to such psychometric properties and to clinical utility (Abu-Saad 2001). For example, a review of measures of QoL for children (Eiser & Morse 2001) concluded that, of a total of 43 instruments reviewed, only five "fulfilled very basic psychometric criteria". However, the increasing emphasis on the importance of the scientific development and evaluation of new instruments (Coste *et al* 1995; Landgraf & Abetz 1996) has led to improved reporting of the development of human instruments and their psychometric properties.

Validity

Validity is the most fundamental attribute of an instrument. It provides evidence that the instrument is able to measure the construct(s) that it was designed to measure. Instrument developers should seek evidence for validity of three kinds: criterion validity, content validity (face validity, which is related to content validity, may or may not be sought) and construct validity (Streiner 1993; Coste *et al* 1995; Johnston 1998; Jensen 2003). Criterion validity is the agreement of a new instrument (or parts of it) with some existing 'gold standard'. When no suitable gold standard exists, researchers use validation strategies established by clinical and experimental psychologists to provide evidence for content and construct validity. The content validity of an instrument is the extent to which the attribute(s) of interest are comprehensively sampled by the instrument's items and the appropriateness of each of the items to the measurement of interest: it is largely established through the methodology used to collect and choose the items to be included in an instrument, but often formally assessed by an independent group of 'experts' (Streiner & Norman 1995).

An instrument that has face validity is one in which the items appear 'on the face of it' to be measuring what the instrument is intended to measure. This kind of validity does not improve the psychometric properties of an instrument, but it generally increases the instrument's acceptability to the respondent. However, in circumstances where there is a risk of biased responding, face validity may not be desirable (Streiner 1993).

In psychiatry, the trait that is being measured is usually inferred from a variety of observations. It exists only as a hypothetical construct, which must be tested to provide evidence for the construct validity of the instrument (Streiner 1993). Factorial validity is one kind of construct validity that requires the statistical analysis of correlations between responses to the items of an instrument. Groupings of items revealed by such analysis (that are also related on clinical or other grounds) are termed 'factors' and provide evidence for a factor structure underlying the data generated by the instrument. If this underlying factor structure fits the construct upon which the instrument was developed, then some evidence has been provided for the validity of the instrument and also for that hypothetical construct (Feinstein 1987; Johnston 1998).

Evidence for the construct validity of an instrument is also provided when the scores obtained with that instrument fit the hypothetical construct upon which the instrument was developed: validity is shown by the extent to which the scores for different known groups or within groups over time can be predicted by that construct (Guyatt *et al* 1993; Streiner 1993; Johnston 1998).

Reliability

Reliability is a measure of whether an instrument can measure accurately and repeatedly what it is intended to measure, so that "measurements of individuals on different occasions, or by different observers, or by similar or parallel tests, produce the same or similar results" (Streiner & Norman 1995). If an instrument is to be used by an independent observer, then inter-rater reliability is sought. Alternatively, an instrument's reliability can be estimated by examining the stability of responses when scores are not expected to change between administrations: called 'test–retest reliability'. If an instrument is valid then it is likely also to be reliable, but it may be highly reliable yet lack validity because it is measuring something other than that which it was intended to measure (Fallowfield 1990).

Responsiveness

While reliability is an important attribute of an instrument, it is possible for an instrument to be reliable yet be unresponsive to clinical change. The ability of an instrument to

https://doi.org/10.1017/S0962728600031687 Published online by Cambridge University Press

capture changes that are important (statistically and practically) has been termed its 'responsiveness'. This is an essential requirement of evaluative instruments — those designed principally to measure change over time (Guyatt *et al* 1987). There is a variety of statistical methods with which responsiveness may be evaluated, but none has become standard (Liang *et al* 2002).

Utility

A useful clinical instrument must be not only valid, reliable and responsive but also "practical and easy to administer, score and interpret" (Landgraf & Abetz 1996). Even if a measure is valid and reliable, it will not be used if it requires lengthy training, is time-consuming to administer or if scoring is complex (Streiner 1993). The possibility of self-administration and literacy level required of respondents are also utility considerations (Dijkers 1999).

Choosing response options

Each item in a questionnaire instrument is accompanied by an answer option or answer options. An important consideration is the choice of options to be offered to respondents, which may be dichotomous, categorical, ordinal or even more complex. If item responses are likely to lie on a continuum, it is important to provide the opportunity for respondents to answer in this way to ensure minimum loss of information (Streiner & Norman 1995), since "the finer the distinction that can be made between subjects' responses, the greater the precision of the measure" (Bowling 1991). Different types of scale are commonly used for the direct estimation of continuous variables, including numerical rating scales (NRS), visual analogue scales (VAS), verbal rating scales (VRS; with or without VAS), and Likert scales (where the respondent rates his agreement with a series of statements on an agree–disagree continuum). Where an item in an instrument offers a number of response options, there is evidence that around seven options tends to result in good reliability in scales in which people are asked to discriminate (Cicchetti *et al* 1985; Preston & Coleman 2000). This may be accounted for by the results of a study carried out in 1956 (Miller 1956), which suggested that the human mind has a span of apprehension capable of distinguishing about 7 items (plus or minus 2).

Constructing a composite indicator

In phase 2, the multiple items must be combined in some way to generate the composite indicator. The metrological principles underlying the creation of a composite indicator 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), and these are described in more detail in the following sections.

Scaling models

A scaling model is a technique that allows weights to be devised for the items included in a scale reflecting the level of the attribute of interest (eg pain or QoL) associated with the given item. There are two main types of classical scaling models: direct or subjective estimation techniques, and indirect or discriminant techniques (Nunnally & Bernstein 1994). The direct or subjective estimation techniques are based on the developers' best subjective estimate of the weights that should be assigned to the items. Using a subjective estimate of the weights may place in doubt the validity of the weighting scheme.

In indirect or discriminant techniques, the weight for each item is derived from experimental observations. Two of the most commonly used indirect scaling models, the equally weighted and paired comparison models, are discussed below. Latent trait models (including factor analysis models) and Item Response Theory provide an alternative to these classical scaling models.

The equally weighted model is the simplest of all scaling models. This model assumes an equal weight for each of the items included in the measurement instrument and assigns a score of 1 to each item. The total combined score represents the number of items observed when the assessment is made.

The paired comparison model was derived from the classical law of comparative judgement proposed by LL Thurstone (1927). This scaling model assumes that the items included in an instrument are correlated with the intensity of the attribute of interest (eg pain or QoL) and that the intensity associated with each item follows a normal distribution. Hence, the best estimate of a weight for any item is its associated mean QoL intensity. Since the attribute of interest, eg QoL, cannot be measured directly, the intensity associated with each item can only be judged relative to the other items within each domain.

The underlying theory of the Thurstone method is that for a psychological continuum (such as intelligence or some other attribute), there are items that appear along it representatively. Each item is associated with a range of the attribute. If a large group of subjects is asked to make comparative judgements regarding items j and k then the proportion judging item k as more favourable/more reflective of the attribute than j, suggests that the discriminal difference of items j and k can be considered positive. Further assuming that the variance of each discriminal distribution is constant leads to the simplest and most commonly used form, Case V of the complete form of the Law of Comparative Judgement (Thurstone 1927). The Law has been most widely applied in psychometric test theory and often the items concern attitudes rather than behaviours. The assumptions underlying the Law of Comparative Judgement under Case V are (1) normality of the item distributions, (2) equal standard deviations of the item distributions, and (3) additivity. For use in the veterinary field, the items typically do not concern attitudes but observational and behavioural stimuli and so the validity of such a law to scale pain or QoL intensity, for instance, must be assessed.

Practically, the weights for each item included in the scale are calculated using a panel of judges. Judges assess each pair of items, j *and* k, as to whether item j indicates a greater

https://doi.org/10.1017/S0962728600031687 Published online by Cambridge University Press

level of the attribute in a subject than item k. Results are the observed proportion p_{jk} — proportion of times item k is judged indicative of greater intensity than item j. These proportions (probabilities) are then transformed to z scores from the standard normal distribution, under the assumptions of normality of item response and unidimensionality. For calculating the raw weight for each item, the set of z scores for item j compared to the other items is averaged (Streiner & Norman 1995). The total score for any instrument can then be calculated by adding together scores for the items observed.

Latent trait models and Item Response Theory

Latent trait models (such as factor analysis) assume that the responses on the multiple items can be described in terms of a smaller number of factors. An item should 'load' onto a particular factor, which is hidden or 'latent', ie it should emerge from the patterns of correlations between the different item responses. Each factor consists of clusters of items whose members are correlated amongst themselves. Factor analysis also plays an important role in evaluation of instrument validity (Nunnally & Bernstein 1994).

Item Response Theory (IRT) emerged in the 1960s as a response to the perceived shortcomings of the Classical Test Theory (CTT) prevalent in psychometrics. IRT is a model-based measurement approach where *both* items and subjects are scaled, with respect to some underlying trait value, θ (Embretson & Reise 2000). Whereas in classical psychometrics, tests generally require their validity and reliability to be re-established when using the scale on samples from differing populations, this is not the case with IRT. In fact, different sets of items can be used on differing occasions — so that tests can even be tailor-made for a given subject. In particular, CTT approaches produce scales that are population-specific whereas in IRT the item characteristics are not dependent on the group used to develop the scale (Yen & Edwardson 1999).

IRT is based on two major (and related) assumptions. The first assumption is that the attribute is unidimensional. It is usually recognised that QoL in humans is multidimensional, and this is likely to be the case for animals. However the assumption of unidimensionality may be loosened a little to that of 'essential unidimensionality'. Under this assumption, whilst several traits may be reflected by the items, one single trait should clearly dominate. The second assumption is that of *local* (or *conditional*) *independence* of the items. This means that after taking into consideration the trait level of the subject, no further correlation exists between the items.

Most IRT models are parametric logistic models that relate the latent trait θ to the probability of a certain response for each item. This relationship is graphically represented by the Item Characteristic Curves (ICCs) of each item, which for the logistic models, at least, are usually monotonic and ogive-shaped (that is, s-shaped). The simplest such model is the Rasch model (Fischer & Molenaar 1995).

If the items have multiple response options then there also exist several models that accommodate the different multiple choice formats (known as *polytomous items*), such as the Graded Response Model and the Partial Credit Model. These models naturally make the parameter estimation process more complex. Practically, IRT requires very large sample sizes (of subjects) to estimate parameters.

A good scale should have items that vary (evenly spaced) along the trait continuum, and should be such that the rank order of difficulty does not vary from subject to subject. Thus Item Response Theory provides a model where the level of QoL which is most likely given all the different responses observed (Embretson & Reise 2000) is estimated.

Questionnaire instruments developed to measure quality of life in dogs

A number of instruments have recently been developed for the measurement of QoL and chronic pain in dogs, for completion by a proxy. These have been modelled to some extent on existing QoL instruments for people, or developed using similar processes. The dog is the species in which the crossover of methodology is particularly apparent, perhaps because of our close association with the domestic dog, and perhaps because dogs are increasingly succumbing to the chronic and painful diseases of old age that we suffer from ourselves, with QoL impact from the conditions themselves and also from an increasing choice of treatments.

Freeman and colleagues (2003) developed an instrument to measure health-related QoL in dogs with cardiac disease. The owner-completion questionnaire was modelled on the Minnesota Living with Heart Failure Questionnaire (http://www.mlhfq.org/). It contained 18 items, developed from a review of the veterinary literature and from the authors' clinical experience. Items were intended to assess an owner's perception of the degree to which clinical signs of cardiac disease had affected the dog's comfort or sociability during the preceding seven days. Instrument development and psychometric properties were fully reported, and the authors concluded that the FETCH (Functional Evaluation of Cardiac Health) questionnaire provided a valid and reliable method of assessing health-related QoL in dogs with cardiac disease.

Gingerich and Strobel (2003) designed a questionnaire to assess treatment effects in geriatric, arthritic dogs. Owner-completed questionnaires were included among a range of outcome measures (eg physical examination, daily activities questionnaire, case-specific questionnaire and owner's global assessment). It was reported that the patient-specific questionnaire and the owner's global assessment differentiated treatment and control groups, but that more extensive interviews with owners were required to identify relevant functional impairments. The questionnaires were based on validated human instruments, but no details of origin of questionnaire items were provided.

Hielm-Björkman and colleagues (2003) developed a questionnaire designed to assess pain in dogs with chronic osteoarthritis. This owner-completed questionnaire was

https://doi.org/10.1017/S0962728600031687 Published online by Cambridge University Press

included among a range of outcome measures (clinician-assigned locomotor index, plasma hormone assay, radiographic examination). It contained 25 questions about behaviour and locomotion, with a range of response options, but no details of the origin of questionnaire items were provided. Eleven questions, that were generally applicable and provided scores that were significantly different for clinical group dogs compared with healthy controls, were subsequently included in a 'chronic pain index', but this contained a theoretical scoring 'grey area'.

Wojciechowska and colleagues (2005a,b) developed a questionnaire designed to measure non-physical aspects of QoL in dogs. This questionnaire was based on objective list theory — that optimal QoL results when certain conditions are met: basic physical needs, normal physiologic function, appropriate social interaction and minimal distress. The questionnaire was not able to discriminate between healthy and sick dogs, and the authors suggested that this might be because certain factors were more important than others for individual dogs. The development and testing of the instrument was fully reported. In their discussion the authors acknowledged the importance of the perception of circumstances by the individual.

Yazbek and Fantoni (2005) developed a questionnaire designed to measure health-related QoL in dogs with pain secondary to cancer. The questionnaire contained 12 questions with Likert-type response options. No details were provided about the source of the questionnaire items, which included a global question about disease impact on QoL, and ranged from a question on changes in the dog's mood, to a question about frequency of vomiting. Scores using this questionnaire were significantly lower for dogs with cancer compared with healthy controls. However, the method of recruiting to the clinical group may have contributed to a risk of respondent bias.

Schneider (2005) developed an instrument to measure QoL in dogs. The owner-response 35-item questionnaire was intended to assess four broad dimensions of QoL in companion dogs: physical, psychological, social, and environmental. Details of item generation were not provided other than reference to a review of the human QoL measurement literature. The instrument was tested with a large number of subjects and found to demonstrate construct validity and reliability that varied across different dimensions, from low to high. The author suggested that the QoL tool was able to discriminate between generally healthy and ill companion dogs. In the same study, the author examined the influence of the human–animal bond on owners' health ratings of their ill dogs, and concluded that the bond between dog and owner can influence reports about dog health, and suggested that such influences should be controlled for in circumstances where owners are asked to provide health ratings for their dogs.

Wiseman-Orr and colleagues (2004, 2006) developed the GUVQuest (Glasgow University Veterinary Questionnaire) as a questionnaire instrument designed to measure chronic pain in dogs through its impact on QoL. The owner-completed questionnaire items were generated directly from the respondent population. The instrument was shown to have good discriminative ability, differentiating well between dogs with chronic degenerative joint disease and healthy controls. Preliminary evidence for the evaluative ability of the instrument is also encouraging, and there is evidence for its ability to minimise respondent bias. The extent to which this instrument is generic for chronic pain or QoL compromise of non-painful cause is currently being explored.

These instruments illustrate the direction of research in this area. Although all of these instruments were developed for use with the dog, the techniques should be applicable to any species with which a human observer and reporter has a sufficiently close association.

Discussion and conclusions

The complex and subjective construct of QoL should not be over-simplified in order to measure it. If existing approaches are not appropriate then new ones should be explored, whether these are philosophical, methodological or statistical. The psychometric approach is one such methodology. If taken, the approach should be carefully followed, and the development of the resultant tool and its psychometric properties reported fully before the instrument is used in any subsequent study. Because the content validity of a questionnaire instrument is largely dependent upon the source of its items, any haste during the early stages of instrument development may reduce its validity. If a careful process of item generation has been undertaken, then that process should be fully reported.

An objective list approach to QoL measurement, whether for people or animals, cannot accommodate variation in what is important to different individuals, or even important to one individual at different times. Consequently, a QoL measure that is based on such an approach is likely to offer only a relatively unsophisticated level of measurement. On the other hand, in the case of herd animals with very similar genotype, phenotype and experience, it may be appropriate and certainly practical to adopt an objective list approach based on causal indicators such as housing, husbandry and indicators of herd health.

It is important to be conceptually aware of the distinction between causal and indicator variables for QoL. Some researchers have confused the measurement of health status and the measurement of QoL, and in some cases this has resulted in the inclusion in the instrument of both causal and indicator variables — addressing elements likely to impact upon QoL, as well as elements that reveal QoL. The confounding of causal and indicator variables has important implications for the analysis of data obtained in such studies (Fayers & Hand 2002).

A criticism that can be levelled at some proxy instruments developed for human QoL measurement is that they are often developed from existing self-report instruments, with the proxy respondent required to make a complex judgement that involves 'second guessing' the responses that would be provided by the subject if his or her self-report were

https://doi.org/10.1017/S0962728600031687 Published online by Cambridge University Press

available. An alternative approach, which has been recommended in the human health measurement field, is to focus on what potential respondents actually observe (Theunissen *et al* 1998) — on variables that appear to be indicators for QoL. This approach is obvious to the veterinary instrument developer, who is not distracted, as is the human instrument developer, by the content of self-reports provided by similar but not identical human populations.

In instrument development and use, a decision must be made about what form of report to seek from a proxy respondent. Some justification has been provided for the use of qualitative interpretation of animal behaviour as a means of obtaining information about the mental state of the animal (Wemelsfelder 1997; Wemelsfelder *et al* 2000, 2001). Both human (particularly in the case of the parent as proxy) and animal proxy instrument developers may consider making better use of our ability to communicate using non-verbal behaviour and the confidence with which we interpret the behaviour of certain other species.

The term 'critical anthropomorphism' was coined by Burghardt (1985), who argued that anthropomorphism was a legitimate approach to science if it was used to develop hypotheses that could be rigorously tested, and he proposed that critical anthropomorphism could use various sources of information including our perceptions, intuitions, feelings and identification with the animal in order to generate "ideas that may prove useful in gaining understanding and the ability to predict outcomes of planned (experimental) and unplanned interventions" (Burghardt 1991). The dangers of uncritical anthropomorphic projection are real, particularly where it is skewed by the complexities of the relationship between animal and owner/carer. Such dangers are not restricted to those working on animal QoL measurement — those working with parents as proxies in paediatric medicine face similar problems of bias. In any case, bias is something to be strenuously guarded against, even in self-report instruments. There are ways of making it difficult for respondents to answer in a biased fashion, whether consciously or unconsciously, or to make it possible to identify those that are.

Animal welfare implications

The clear conceptualisation of animal QoL as the individual's perspective offers the animal and veterinary science communities, as well as the public, the opportunity and encouragement to focus on the feelings of the individual in matters of animal welfare. The development of valid, reliable and responsive measures of animal QoL will facilitate decision-making with regard to medical or other interventions at the level of the individual and through improved veterinary medical research and strengthened evidence-based veterinary medicine, or through better care or better resource-allocation decisions in non-medical contexts. Such instrument development will not be easy, but, to paraphrase Albert Einstein, "not everything that counts can [easily] be counted and not everything that can [easily] be counted counts". Without minimising such difficulties, we conclude that yes, we can measure animal quality of life.

References

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

Anand KJS and Craig KD 1996 New perspectives on the definition of pain. *Pain 67*: 3-6

Birnbacher D 1999 Quality of life — evaluation or description? *Ethical Theory and Moral Practice 2*: 25-36

Bowling A 1991 *Measuring Health: A Review of Quality of Life Measurement*. Open University Press: Buckingham, UK

Burghardt GM 1985 *Foundations of Comparative Ethology*. Van Nostrand Reinhold: New York, USA

Burghardt GM 1991 Cognitive ethology and critical anthropomorphism: a snake with two heads and hognose snakes that play dead. In: Ristau CA (ed) *Cognitive Ethology: The Minds of Other Animals* pp 53-90. Erlbaum: San Francisco, USA

Cicchetti DV, Showalter D and Tyrer PJ 1985 The effect of number of rating scale categories on levels of interrater reliability: a Monte Carlo investigation. *Applied Psychological Measurement 9*: 31-36

Cook KF, Monahan PO and McHorney CA 2003 Delicate balance between theory and practice: health status and item response theory (Editorial). *Medical Care 41*: 571-574

Coste J, Fermanian J and Venot A 1995 Methodological and statistical problems in the construction of composite measurement scales: a survey of six medical and epidemiological journals. *Statistics in Medicine 14*: 331-345

Cunningham N 1999 Primary requirements for an ethical definition of pain (Focus). *Pain Forum 8*: 93-99

Dijkers M 1999 Measuring quality of life: methodological issues. *American Journal of Physical Medicine and Rehabilitation 78*: 286-300

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

Embretson SE and Reise SP 2000 *Item Response Theory for Psychologists*. LEA: New Jersey, USA

Emery M-P, Perrier L-L and Acquadro C 2005 Patient-Reported Outcome and Quality of Life Instruments Database (PROQOLID): Frequently asked questions. *Health and Quality of Life Outcomes 3*: 12

Fallowfield L 1990 *The Quality of Life: The Missing Measurement in Health Care*. Souvenir Press: London, UK

Farm Animal Welfare Council 1992 FAWC updates the five freedoms. *Veterinary Record 131*: 357

Fayers PM and Hand DJ 2002 Causal variables, indicator variables and measurements scales: an example from quality of life. *Journal of the Royal Statistical Society 165(2)*: 1-21

Fayers PM, Hopwood P, Harvey A, Girling DJ, Machin D and Stephens R 1997 Quality of life assessment in clinical trials: guidelines and a checklist for protocol writers. The UK Medical Research Council experience. *European Journal of Cancer 33*: 20-28

Feinstein AR 1987 *Clinimetrics* pp 249-253. Yale University Press: Yale, USA

Fischer GH and Molenaar IW 1995 *Rasch Models. Foundations, Recent Developments and Applications*. Springer: Dordrecht, The Netherlands

Freeman LM, Rush JE, Farabaugh AE and Must A 2003 Development and evaluation of a questionnaire for assessing health-related quality of life in dogs with cardiac disease. *Journal of the American Veterinary Medical Association 226*: 1864-1868

Garratt A, Schmidt L, Mackintosh A and Fitzpatrick R 2002 Quality of life measurement: bibliographic study of patient assessed health outcome measures. *British Medical Journal 324*: 1417-1419

https://doi.org/10.1017/S0962728600031687 Published online by Cambridge University Press

Gingerich DA and Strobel JD 2003 Use of client-specific outcome measures to assess treatment effects in geriatric, arthritic dogs: controlled clinical evaluation of a neutraceutical. *Veterinary Therapeutics 4*: 376-386

Guyatt G, Walter S and Norman G 1987 Measuring change over time: assessing the usefulness of evaluative instruments. *Journal of Chronic Diseases 40*: 171-178

Guyatt GH, Feeny DH and Patrick DL 1993 Measuring health-related quality of life. *Annals of Internal Medicine 118*: 622-629

Hielm-Björkman AK, Kuusela E, Lipman A, Markkola A, Saarto E, Huttunen P, Leppäluoto J, Tulamo R-M and Raekallio M 2003 Evaluation of methods for assessment of pain associated with chronic osteoarthritis in dogs. *Journal of the American Veterinary Medical Association 222*: 1552-1558

Innes A and Surr L 2001 Measuring the well-being of people with dementia living in formal care settings: the use of Dementia Care mapping. *Ageing and Mental Health 5*: 258-268

Jensen MP 2003 Questionnaire validation: a brief guide for readers of the research literature. *The Clinical Journal of Pain 19*: 345-352

Johnston CC 1998 Psychometric issues in the measurement of pain. In: Finley GA and McGrath PJ (eds) *Measurement of Pain in Infants and Children, Progress in Pain Research and Management, Volume 10* pp 5-20. IASP Press: Seattle, USA

Juniper EF, Guyatt GH and Jaeschke R 1996 How to develop and validate a new health-related quality of life instrument. In: Spilker B (ed) *Quality of Life and Pharmacoeconomics in Clinical Trials (2nd Edition)* pp 49-56. Lippincott-Raven: Philadelphia, USA

Kelvin 1893 *Popular Lectures and Addresses, Vol 1: Electrical Units of Measurement.* Lecture given to the Institute of Civil Engineers

Landgraf JM and Abetz LN 1996 Measuring health outcomes in pediatric populations: issues in psychometrics and application. In: Spilker B (ed) *Quality of Life and Pharmacoeconomics in Clinical Trials (2nd Edition)* pp 793-802. Lippincott-Raven: Philadelphia, USA

Liang MH, Lew RA, Stucki G, Fortin PR and Daltroy L 2002 Measuring clinically important changes with patient oriented questionnaires. *Medical Care 40(4) (Suppl II)*: II-45–II-51

Matza LS, Swensen AR, Flood EM, Secnik K and Leidy NK 2004 Assessment of health-related quality of life in children: a review of conceptual, methodological, and regulatory issues. *Value in Health 7*: 79-92

McMillan FD 2005 The concept of quality of life in animals. In: McMillan FD (ed) *Mental Health and Well-Being in Animals* p 193. Blackwell: Ames, Iowa, USA

Miller GA 1956 The magical number seven, plus or minus two: some limits on our capacity for processing information. *The Psychological Review 63*: 81-97

Nunnally JC and Bernstein IH 1994 *Psychometric Theory (3rd Edition)*. McGraw Hill: New York, USA

Preston CC and Coleman AM 2000 Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica (Amst) 104*: 1-15

Schneider TR 2005 Methods for assessing companion animal quality of life. Iams Symposium Proceedings at the North American Veterinary Conference, *Improving Our Listening Skills: What Dogs and Cats are Telling Us*, Orlando, FL, USA. The Iams Company: Dayton, OH, USA

Scott EM, Fitzpatrick JL, Nolan AM, Reid J and Wiseman ML 2003 Evaluation of welfare state based on interpretation of multiple indices. *Animal Welfare 12*: 257-268

Skevington SM 1998 Investigating the relationship between pain and discomfort and quality of life, using the WHOQOL. *Pain 76*: 395-406

Skevington SM, MacArthur P and Somerset M 1997 Developing items for the WHOQOL: an investigation of contemporary beliefs about quality of life related to health in Britain. *British Journal of Health Psychology 2*: 55-72

Stenner PHD, Cooper D and Skevington SM 2003 Putting the Q into quality of life: the identification of subjective constructions of health-related quality of life using Q methodology. *Social Science and Medicine 57*: 2161-2172

Streiner DL 1993 Research methods in psychiatry. A checklist for evaluating the usefulness of rating scales. *Canadian Journal of Psychiatry 38*: 140-148

Streiner DL and Norman GR 1995 *Health Measurement Scales: A Practical Guide to Their Development and Use (2nd Edition)* pp 1-180. Oxford University Press: New York, USA

Theunissen NCM, Vogels TGC, Koopman GHW, Zwinderman KAH, Verloove-Vanhorick SP and Wit JM 1998 The proxy problem: child report versus parent report in health-related quality of life research. *Quality of Life Research 7*: 387-397

Thurstone LL 1927 A law of comparative judgement. *Psychological Review 34*: 273-386

Wemelsfelder F 1997 The scientific validity of subjective concepts in models of animal welfare. *Applied Animal Behaviour Science 53(1-2)*: 75-88

Wemelsfelder F, Hunter TEA, Mendl MT and 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(3)*: 193-215

Wemelsfelder F, Hunter TEA, Mendl MT and Lawrence AB 2001 Assessing the 'whole animal': a free choice profiling approach. *Animal Behaviour 62*: 209-220

WHOQOL Group 1995 The World Health Organization quality of life assessment (WHOQOL): position paper from the World Health Organization. *Social Science and Medicine 41*: 1403-1409

Wiseman-Orr ML, Nolan AM, Reid J and Scott EM 2004 Development of a questionnaire to measure the effects of chronic pain on health-related quality of life in dogs. *American Journal of Veterinary Research 65*: 1077-1084

Wiseman-Orr ML, Scott EM, Reid J and Nolan AM 2006 Validation of a structured questionnaire as an instrument to measure chronic pain in dogs on the basis of effects on health-related quality of life. *American Journal of Veterinary Research 67*: 1826-1836

Wojciechowska JI, Hewson CJ, Strhynn H, Guy NC, Patronek GJ and Timmons V 2005a Development of a discriminative questionnaire to assess non-physical aspects of quality of life of dogs. *American Journal of Veterinary Research 66*: 1453-1460

Wojciechowska JI, Hewson CJ, Strhynn H, Guy NC, Patronek GJ and Timmons V 2005b Evaluation of a questionnaire regarding non-physical aspects of quality of life in sick and healthy dogs. *American Journal of Veterinary Research 66*: 1461-1467

Yazbek KVB and Fantoni DT 2005 Validity of health-related quality of life scale for dogs with signs of pain secondary to cancer. *Journal of the American Veterinary Medical Association 226*: 1354-1358

Yen M and Edwardson SR 1999 Item Response Theory approach in scale development. *Nursing Research 48*: 234-238

https://doi.org/10.1017/S0962728600031687 Published online by Cambridge University Press