JSAP JOURNAL OF SMALL ANIMAL PRACTICE

BSAVA BRITISH SMALL ANIMAL VETERINARY ASSOCIATION

PAPER

Shortening of an existing generic online health-related quality of life instrument for dogs

J. Reid*,[1], L. Wiseman-Orr† and M. Scott†

*School of Veterinary Medicine, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow G12 8QQ, UK
†School of Mathematics and Statistics, College of Science and Engineering, University of Glasgow, Glasgow G12 8QQ, UK

[1]Corresponding author email: jacqueline.reid@glasgow.ac.uk

Objective: **Development, initial validation and reliability testing of a shortened version of a web-based questionnaire instrument to measure generic health-related quality of life in companion dogs, to facilitate smartphone and online use.**

Materials and Methods: **The original 46 items were reduced using expert judgment and factor analysis. Items were removed on the basis of item loadings and communalities on factors identified through factor analysis of responses from owners of healthy and unwell dogs, intrafactor item correlations, readability of items in the UK, USA and Australia and ability of individual items to discriminate between healthy and unwell dogs. Validity was assessed through factor analysis and a field trial using a "known groups" approach. Test–retest reliability was assessed using intraclass correlation coefficients.**

Results: **The new instrument comprises 22 items, each of which was rated by dog owners using a 7-point Likert scale. Factor analysis revealed a structure with four health-related quality of life domains (energetic/enthusiastic, happy/content, active/comfortable, and calm/relaxed) accounting for 72% of the variability in the data compared with 64% for the original instrument. The field test involving 153 healthy and unwell dogs demonstrated good discriminative properties and high intraclass correlation coefficients.**

Clinical Significance: **The 22-item shortened form is superior to the original instrument and can be accessed *via* a mobile phone app. This is likely to increase the acceptability to dog owners as a routine wellness measure in health care packages and as a therapeutic monitoring tool.**

Journal of Small Animal Practice (2017)
DOI: 10.1111/jsap.12772

Accepted: 5 September 2017

INTRODUCTION

The measurement of health-related quality of life (HRQL) plays an increasingly important part in human medicine to detect disease (discriminative purposes) and to measure change in health status over time (evaluative purposes) (Fayers & Machin 2013). Structured questionnaires to measure HRQL of people are developed and tested using well-established psychometric methodology (Streiner & Norman 2008, Abell *et al.* 2009, Brod *et al.* 2009). These instruments are designed for self-report by the subject but, where self-report is not possible (e.g. infants and the cognitively impaired), they are completed by an observer who knows the subject well. Instruments can be disease-specific, or they can be generic, designed to be used in a variety of circumstances. Instruments to measure HRQL in companion animals generally consist of questions for the pet owner, who is well placed to report upon the, often subtle, changes in behaviour, attitude and demeanour that occur with chronic disease (Wiseman *et al.*, 2001). The majority of HRQL instruments that have been developed for companion animals are disease-specific [Freeman *et al.* 2005 (cardiac disease), Yazbek & Fantoni 2005 (cancer), Budke *et al.* 2008 (spinal cord injuries), Favrot *et al.* 2010 (atopic dermatitis), Lynch *et al.* 2011 (cancer), Noli *et al.* 2011 (skin disease), Niessen *et al.* 2012 (diabetes mellitus). However, here we describe the shortening of a 46-item generic instrument (VetMetrica, www.newmetrica.com) that measures the impact of chronic pain and non-painful physical chronic diseases on quality of life (QOL) (Reid *et al.* 2013) in dogs. Disease-specific instruments may be more responsive to clinical change, but generic instruments can be valuable indicators of a range of impacts associated with disease and its treatment, and may be the only option when a patient is suffering from more than one condition, as is often the case in older companion animals. There is substantial potential for use of this type of information within veterinary practice, including raising the profile of preventive veterinary medicine. For instance, between routine visits for vaccination, when it may enhance communication with clients and establish stronger bonds with them as partners in health care. Furthermore, it improves disease detection, including chronic disease that is often unrecognised and unreported. The validity of this specific methodology and initial results are encouraging (Yam *et al.* 2016) and users report that being able to demonstrate a deteriorating QOL to owners would help to facilitate end-of-life decision-making for individual dogs. Other than the subject of this article, only two generic methods to measure HRQL in dogs have been published (Wojciechowska *et al.* 2005a,b, Lavan 2013), but one was shown not to distinguish healthy from sick dogs (Wojciechowska *et al.* 2005a,b) and the other (Lavan 2013) was restricted to use in healthy dogs.

Instrument development is an iterative process, in which they are refined and re-tested with new populations in new contexts, and their usefulness may be improved by being shortened, for example, shortening of the generic SF36 item HRQL instrument for people to the SF12version (Cheak-Zamora *et al.* 2009). Guidelines for shortening existing composite scales such as those designed to measure HRQL are scarce. Coste *et al.* (1997) reported that the process of scale shortening lacked rigorous methodology and this was confirmed by Stanton *et al.* (2002). Criteria used previously to select items for retention include expert judgment (Osse *et al.* 2007), the identification of items that discriminate best (Reid *et al.* 2013) and statistical techniques such as factor analysis (FA) (Las Hayas *et al.* 2010, Reid *et al.* 2013). The FA examines the relationships between variables and clusters them into a small number of homogenous groups, which can then be used for analysis. Groupings of variables revealed by such analysis, which in respect of health care are also related to clinical or other grounds, are termed factors and the association between a variable and factor is expressed as a factor loading of the variable (values between 0 and 1), where the higher the loading the closer the association. The communality of a variable is the portion of the variance of that variable that is accounted for by the common factors (DeVellis 2012). Although FA is capable of providing any number of factor models for a given data set, there are established methods that can be used to identify how many factors could sensibly be extracted, including the scree test and the Kaiser criterion (Coste *et al.* 2005). However, it is up to the instrument developer to decide upon the most satisfactory factor model, the number of factors it contains, and name these according to the interpretation of their associated items. Importantly, a good factor model is one in which the derived factors are readily interpretable and which accounts for a reasonable amount of the variance in the data set from which it was created (StatSoft Inc. 2003). The most common type of FA is exploratory factor analysis (EFA), in which the factor loadings of each item are used to determine the factor structure of a data set collected for the purpose of instrument development. Confirmatory factor analysis (CFA) determines whether analysis of a new data set performs in the same way with items loading, as predicted, on the expected number of factors, thus testing the validity of the factor solutions obtained from the EFA (Floyd & Widaman 1995).

In addition to its use in instrument development and shortening, FA is one of the most commonly used procedures in the validation of psychological measures (Nunnally & Bernstein 1994, Floyd & Widaman 1995). Validity (criterion, content and construct) provides evidence that the instrument measures what it was designed to measure. Criterion validity is the agreement of a new instrument with some existing "gold standard". Content validity ensures the appropriateness and completeness of the items within the instrument and is established during its construction (Fayers & Machin 2013). There are many approaches to examine the construct validity of a new questionnaire, including factorial validity and known-groups validity. Factorial validity is demonstrated if, after FA, an interpretable factor structure fits the construct that the instrument was designed to measure (Johnston 1998), a construct being something that is not directly observable or measurable, such as "happiness". In the context of this article the construct was HRQL – the subjective evaluation by an individual of circumstances that include an altered health state and related interventions (Wiseman-Orr *et al.* 2006). In the known groups approach to determine the construct validity of an instrument, predictions are made about how scores obtained with the instrument will differ between groups and these predictions are then tested. For example, an instrument should be able to distinguish correctly between groups that would be expected to have quite different scores, such as healthy and unwell animals (Reid *et al.* 2013). In addition to testing an instrument's validity, evidence should also be sought for its reliability – which is necessary (although not sufficient) for validity. A reliable instrument is one that will produce the same score when an unchanging subject is measured at two time points by the same observer (repeatability/intrarater reliability), or when two people measure the same subject at one time (reproducibility/inter-rater reliability) (Streiner & Norman 2008).

In addition to validity and reliability, to be useful in a clinical setting an instrument must also have utility – it must be acceptably quick, easy to understand and simple to use (Teasdale & Jennett 1974). In terms of speed and ease of use, electronic technologies provide an acceptable, or preferable, alternative to paper, regardless of user age and previous experience of computers (Greenwood *et al.*, 2006). Access to such assessment instruments may be online, and preferably compatible with mobile devices. In human health care, mobile health (m-health) applications are increasing and used in a diverse range of practices (Boulos *et al.* 2014). However, these apps must be carefully designed to retain the utility of the instrument they deliver. Krebs & Duncan (2015) surveyed US health app users and "too much time to enter data" with consequent loss of interest was reported by 44·5% respondents as the reason for discontinuing use of the app. An instrument for providing canine HRQL measurement has been developed and evaluated (Reid *et al.* 2013) but, from a practical perspective, had too many items for presentation via an app. This article reports the process undertaken to shorten the 46-item instrument and to develop an app designed for owner completion.

MATERIALS AND METHODS

Original instrument

A 46-item long-form online questionnaire to measure canine HRQL had been developed from an original, novel, paper-based canine HRQL instrument, and both the original and web-based version had previously been validated (Wiseman-Orr *et al.* 2004, Wiseman-Orr *et al.* 2006, Reid *et al.* 2013). This generic instrument consisted of 46 questions for the dog owner, each of which comprised a descriptor (*e.g.* "active") with a 7-point Likert rating scale, 0 to 6 (with 0 meaning "not at all" and 6 meaning "could not be more"). During development of the 46-item instrument owner responses to the items were used to generate an HRQL profile comprising scores in four QOL domains – named as vitality, pain, distress and anxiety in accordance with the items loading onto them. The 46-item instrument was shown to have high utility, was easy to use, taking around 5 minutes to complete online, and with automatic and instantaneous transformation of responses into the scores profile.

Confirmatory factor analysis

Previously, as part of a field test to determine the known groups validity and reliability of the 46-item questionnaire instrument, owners of unwell dogs attending Glasgow University Small Animal Hospital (GUSAH) and selected Vets Now clinics and owners of dogs recruited from clinical, non-clinical, nursing and administrative staff members at GUSAH deemed to be generally healthy by author JR on the basis of history and lack of clinical signs completed at least one online assessment using the 46-item instrument between January 2011 and April 2012. The only inclusion criterion for unwell dogs was that the dog was suffering from a non-acute condition that was expected to affect its QOL. Owners of unwell dogs were recruited from the daily case load by a senior nurse in GUSAH and the attending clinician in Vets Now clinics when it was logistically possible with no attempt made to control selection bias. Accordingly, the sampling was best described as cluster. Ethics approval was granted by the Glasgow University Veterinary School Ethics Committee and written consent was obtained from all owners.

FA was carried out (Minitab v.16) on the first questionnaire completed by each owner. A principal components method of FA with a varimax rotation was performed. Input variables were all item ratings. Loadings were sorted, and items with loadings of less than 0·3 were excluded (Floyd & Widaman 1995). Guided by a scree test and the Kaiser criterion, the interpretability of a range of factor models was examined. A factor model was sought that accounted for an acceptable amount of the variability in the data, was readily interpretable, and did not include any factors containing only one or two items (Norman & Streiner 1994).

Reduction of items

Using the results of the CFA, any item with a loading onto any factor less than 0·5 was removed (Shevlin & Miles 1998). Thereafter any item with a communality of less than 0·5 was removed (Velicer *et al.* 1982, MacCallum *et al.* 1999). The remaining items were considered for removal on the basis of their correlation with other domain items, readability and ability to discriminate between healthy and unwell dogs. A Pearson coefficient was calculated for each item and those with a correlation of ≥0·80 or ≤0·20 with other item(s) in the same domain were considered for removal on the basis that they were too similar to others in the domain and therefore extraneous (≥0·80) or not related to the underlying construct of the domain (≤0·20) (Boyle 1991, Coste *et al.* 1995). To ensure that the instrument could be used in other English-speaking countries, in which some words might have slightly different meanings from in the UK, the suitability of the items was tested by means of two small surveys, one in the USA (n=9) and one in Australia (n=15). These asked adult respondents to identify items that they considered not to relate to dogs: respondents were dog owners identified and contacted via email by authors JR and LWO or by veterinary surgeons abroad. In addition, in the absence of established readability metrics for individual words (rather than continuous prose), a number of novel approaches were used to test that items would be readily understood by most adults in the UK. For example, items were reviewed by a class of 9-year-old schoolchildren and by a group of adult literacy tutors, and their inclusion in two dictionaries for children aged 9 to 12 years was checked. All decisions to remove items were made primarily on the basis of the surveys in the USA and in Australia, but the UK studies showed that many of those items would also cause difficulty to some UK readers, and showed that none of the remaining items would cause such difficulty. To identify items that could discriminate well between unwell dogs and healthy dogs, histograms of item responses for each item for healthy and for unwell dogs, each plotted on a single graph, were constructed, and those considered by the authors not to discriminate well, on the grounds that the two histograms looked very similar, were removed. Initial screening was carried out by two authors (JR and LWO) and where there was disagreement regarding discrimination a final judgment was made by the third author, a statistical expert.

FA of the items retained for the shortened instrument

Using the same data set as was used for the CFA and selection of items, those items retained for the shortened instrument were extracted and subjected to FA as for the CFA, with the exception that items with loadings of less than 0·5 were excluded. Two, three- and four-factor models were explored to determine the optimum structure for the shortened instrument and an algorithm, based on the item–factor associations of the selected factor model, was derived in order to generate a domain score for each of the resultant factors/domains.

Field test of the shortened instrument

A different group of owners of healthy dogs and owners of unwell dogs recruited at GUSAH and according to the clinical judgment of three vets in general practice and a pharmaceutical veterinary adviser completed at least one online assessment using the shortened instrument between February and April 2014. No attempt was made to control selection bias, but each owner was confirmed as the primary caregiver to each dog. Using the first assessment completed for each dog, descriptive statistics were used to identify differences between the healthy and unwell groups, and followed by formal statistical analysis using non-parametric Mann–Whitney tests due to the non-normality of the data. Linear discriminant analysis was carried out to determine the accuracy of the instrument in differentiating healthy from unwell dogs. The same owner of a number of healthy dogs completed two assessments, 2 weeks apart, and test–retest reliability was assessed using the intraclass correlation coefficient (ICC). A one-way random model was assumed where the subjects are assumed random (Shrout & Fleiss 1979).

RESULTS

Confirmatory factor analysis

FA was carried out on responses from owners of 88 unwell dogs and 34 healthy dogs (Table 1) who participated in the field test to determine the validity of the original 46-item instrument. Forty-seven dog breeds were represented (Table 2). The result of the FA

Table 1. Demographics of the dogs used in different parts of the study

Study details	CFA of the original 46 items, reduction of items and EFA of the 22 retained items	
Health status	Healthy	Unwell
Number of dogs	34	88
Gender	13 females; 21 males	49 females; 39 males
Age	Median 5 years; range 1 to 11	Median 6 years; range 1 to 18
Total number of breeds	46	
Study details	**Field test for the 22-item shortened instrument**	
Health status	Healthy	Unwell
Number of dogs	100	53
Gender	45 females; 55 males	22 females; 31 males
Age	Median 6 years; range 1 to 16	Median 10 years; range 4 to 18
Total number of breeds	42	

CFA Confirmatory factor analysis, EFA Exploratory factor analysis

Table 2. Breeds of dogs used in confirmatory factor analysis of the 46-item instrument and the field test of the final shortened instrument

Breed	Confirmatory factor analysis, n=122	22-item field test, n=153
Airedale terrier		2
Akita		1
American bulldog		2
Basset griffon Vendeen		1
Basset hound	1	
Beagle	2	
Bearded collie	1	5
Belgian shepherd	1	
Bichon frise		2
Border collie	5	9
Border terrier	2	7
Boston terrier	1	1
Boxer	4	5
Bull mastiff	1	
Bull terrier		1
Cairn terrier	2	3
Cavalier King Charles spaniel	3	2
Chihuahua	1	1
Chinese crested	1	
Clumber spaniel	1	
Cocker spaniel	5	8
Crossbreed large	3	
Crossbreed medium	12	27
Crossbreed small	3	
Dachshund	1	
Dalmatian		4
English bulldog	1	
English setter	1	
French bulldog		2
German shepherd dog	1	1
German short-haired pointer	2	5
Golden retriever	5	5
Greyhound	3	
Hungarian viszla	1	1
Irish Setter	3	
Irish terrier	1	1
Jack Russell terrier	4	6
Kerry blue		1
Labradoodle	1	2
Labrador	13	20
Lhasa apso	4	1
Lurcher	3	1
Malamute		1
Miniature Yorkshire terrier	1	
Neopolitan mastiff	1	
Newfoundland	1	
Parson Jack Russell terrier	1	1
Patterdale terrier		1
Poodle	2	1
Pug	3	
Samoyed		1
Scottish deehound	5	
Scottish terrier	1	
Sealyham		1
Sheltie		1
Shihtzu		3
Springer spaniel	2	5
Staffordshire bull terrier	3	1
Swiss mountain dog		1
West Highland white terrier	3	1
Whippet	1	2
Working cocker spaniel	1	
Yorkshire terrier	3	6

gave four factors with similar items loading on to the four factors that had been derived for the 46-item instrument during EFA (Reid *et al.* 2013). The confirmatory factor solution accounted for 63% of the variance in the data, which was similar to that of the EFA (64%).

Reduction of items

Eight items – sluggish, confident, unsociable, contented, alert, obedient, reluctant, and frightened – were removed on the basis of their loading and communality in the CFA, and the remaining 38 items were considered for exclusion because of their correlation with other domain items, their readability and their ability to discriminate between unwell and healthy dogs. Fig. 1 shows the difference between an item that was judged independently by authors JR and LWO to discriminate well between unwell and healthy dogs (A – uncomfortable) and one that they judged did not (B – subdued). Table 3 lists the 24 items removed from the 46-item long form instrument along with the reasons for their removal.

Table 3. List of items removed from the 46-item instrument and the prime reason for their removal

Item	Reason for removal
Alert	Low factor loading and/or communality
Sluggish	Low factor loading and/or communality
Lacklustre	Poor readability
Weary	Poor readability
Miserable	Too highly correlated with other items in factor
Subdued	Poor discrimination healthy and sick
Sorrowful	Too highly correlated with other items in factor
Apathetic	Poor discrimination healthy and sick
Withdrawn	Poor discrimination healthy and sick
Reluctant	Low factor loading and/or communality
Dull	Too highly correlated with other items in factor
Slowed	Poor readability
Fit	Too highly correlated with other items in factor
Contented	Low factor loading and/or communality
Pathetic	Poor discrimination healthy and sick
Confident	Low factor loading and/or communality
Unsociable	Low factor loading and/or communality
Pained	Poor readability
Awkward	Poor discrimination healthy and sick
Moaning	Poor readability
Complaining	Poor discrimination healthy and sick
Obedient	Low factor loading and/or communality
Consistent	Poor discrimination healthy and sick
Frightened	Low factor loading and/or communality

FA of the 22 items retained for the shortened instrument

The four-factor model accounted for more of the variance than the two- and three-factor models and was the most interpretable, with factors very similar in terms of their item loadings to those of the 46-item instrument. These 22 item factors were named as domains of HRQL by the authors in accordance with the items loading onto them [energetic/enthusiastic (E/E), happy/content (H/C), active/comfortable (A/C), calm/relaxed (C/R)]. The four-factor model accounted for 72% of the variance. The scoring algorithm derived for these four domains of HRQL was based on item–factor associations. For ease of interpretation, all domains were named positively and the scoring algorithm provided that higher scores in all domains were associated with better HRQL.

Field test of the shortened instrument

Owners of 53 unwell dogs and 100 healthy dogs (Table 1) completed one assessment and, of these, 49 owners of healthy dogs completed two. Forty-two dog breeds were represented with no breed predominating (Table 2). A comparison of median scores and the interquartile range (IQR) for healthy and unwell dogs for each of the domains (Table 4) showed clear differences for E/E, H/C and A/C, but less so for C/R. However, the results of the Mann–Whitney tests (Table 5) demonstrated a significant difference ($P \leq 0{\cdot}05$) between the scores for healthy and unwell dogs in all domains. The variability, represented by the IQR and the extent of the tails of the distribution, was large in all domains for the unwell dogs compared with that of the healthy dogs, with the exception of C/R where the variability was similar between the groups (Fig. 2). Linear discriminant analysis showed that the 22-item short instrument correctly classified 89% of the healthy dogs and 77% of those that were unwell with an overall misclassification rate of

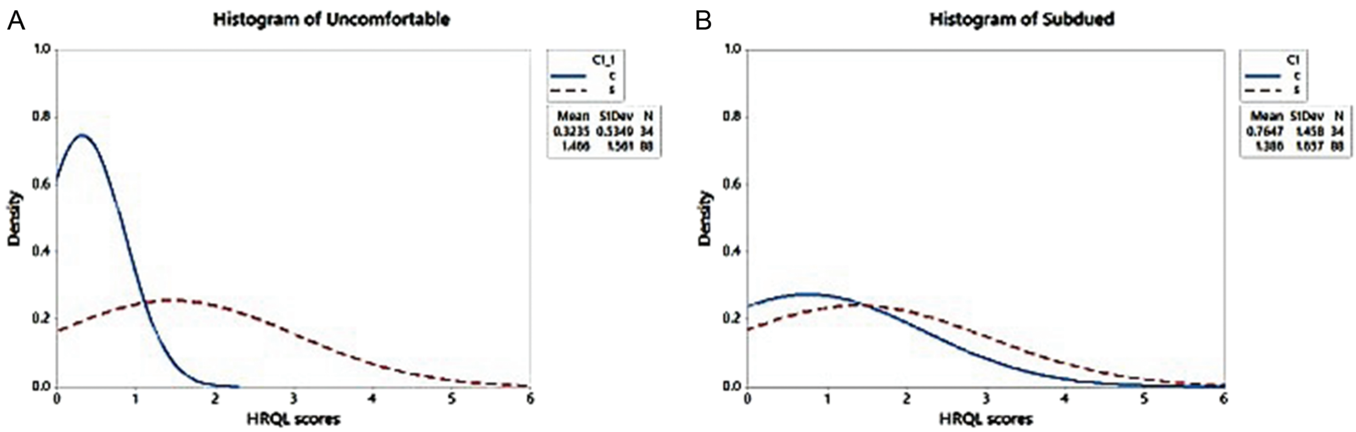

FIG 1. Examples of an item, which discriminates clearly between healthy and unwell dogs: (A) uncomfortable, and one which does not discriminate well: (B) subdued

15%. For those owners who completed two assessments the ICC (95% confidence intervals) for all domains was energetic/enthusiastic 0·75 (0·60 to 0·85); happy/content 0·75 (0·60 to 0·85); active/comfortable 0·75 (0·60 to 0·85); calm/relaxed 0·73 (0·57 to 0·84).

Table 4. Median scores with interquartile ranges (IQR) for healthy and unwell dogs for each of the four domains of HRQL identified through factor analysis

Domain	N	Median score	IQR for score
E/E			
Healthy	100	5·67	0·78
Unwell	53	4·22	1·84
H/C			
Healthy	100	6·00	0·30
Unwell	53	4·86	1·43
A/C			
Healthy	100	5·83	0·67
Unwell	53	3·83	2·09
C/R			
Healthy	100	5·00	1·67
Unwell	53	4·33	1·33

HQRL Health-related quality of life, E/E energetic/enthusiastic, H/C happy/content, A/C active/comfortable, C/R calm/relaxed

DISCUSSION

A review article by Goetz *et al.* in 2013 concluded that item reduction of an existing scale must be based on rigorous methodology if the short-form instrument aims to maintain the validity and other measurement properties of the parent instrument. To that end they highlighted the importance of reporting the validity of the original scale, documenting the reasons for item selection, preserving content validity and the psychometric properties of the original scale and validating the short-form scale in an independent sample. The shortening process described here followed these guidelines.

Construct validity of the original 46-item long form instrument had been demonstrated using factorial validity and a known groups approach in dogs with a variety of chronic conditions, and evidence of its reliability had been obtained (Reid *et al.* 2013). Additionally, these 46 items can generate a valid measure of HRQL in dogs with osteoarthritis (unpublished) and lymphoma (unpublished), for which both disease and aggressive treatment may impact on a dog's QOL, and also in obese dogs (Yam *et al.* 2016), all of which support the validity of the 46-item instrument.

Table 5. The results of Mann–Whitney tests and non-parametric confidence intervals applied to the population median scored for healthy and unwell dogs for each of the four domains of HRQL defined

	Median Difference in (unwell – healthy) populations	P-value	95% confidence interval for population median difference	W
Energetic/enthusiastic	−1·44	<0·001	(−1·67, −1·11)	2088·5
Happy/content	−0·86	<0·001	(−1·14, −0·57)	2447·5
Active/comfortable	−1·67	<0·001	(−2·16, −1·33)	1956·5
Calm/relaxed	−0·66	0·001	(−1·00, −0·33)	3252·5

HQRL Health-related quality of life, W Mann–Whitney test statistic

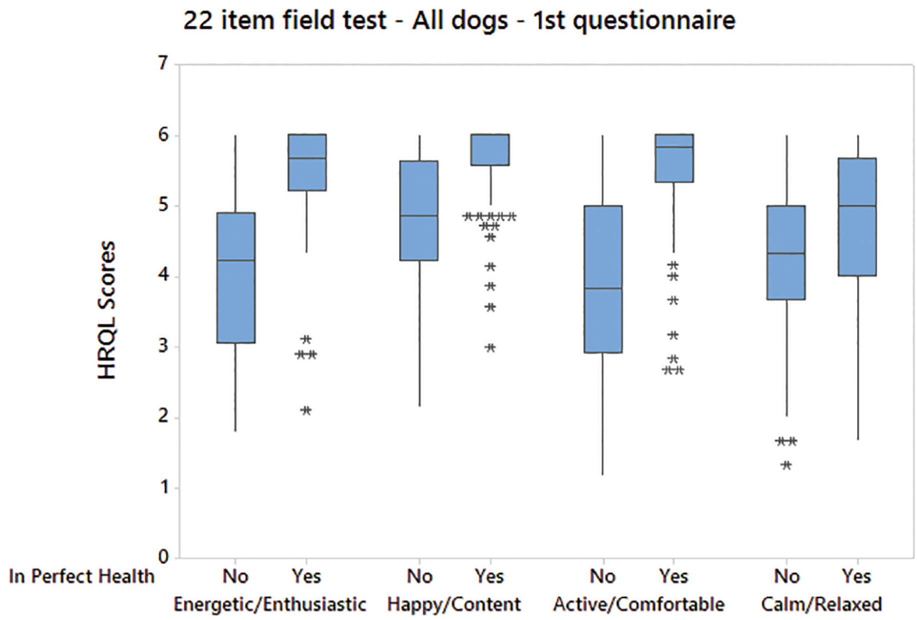

FIG 2. Plots of scores for four domains of health-related quality of life (HRQL) (energetic/enthusiastic, happy/content, active/comfortable, calm/relaxed) generated by the owners of 53 unwell dogs and 100 healthy dogs using the 22-item web-based HRQL instrument, showing the median and interquartile range for all domains

Factor loadings of 0·3, 0·5 and 0·7 are generally considered to be low, medium and high respectively (Shevlin & Miles 1998) with loadings of greater than 0·3 deemed to be the minimum for EFA. Exclusion of items with loadings less than 0·3 is commonly applied in instrument development (Floyd & Widaman 1995). However, loadings of greater than 0·5 are considered to be practically significant and, accordingly, the first step in the item reduction process described here was to exclude items which loaded less than 0·5. It has been suggested that factor structures are improved when both loadings and communalities are higher (Velicer *et al.* 1982, MacCallum *et al.* 1999) and so items with a communality of less than 0·5 were excluded as part of the shortening process. Thirty-three percent (8/24) of the removed items were excluded by this initial process and calculation of the Pearson's coefficient to exclude items that were too similar and therefore extraneous accounted for a further 29% (7/24). Once highly correlated items were identified, the process of choosing which to keep depended on their discrimination and readability. For example, "pained" and "sore" had a correlation of 0·84, both distinguished well between healthy and unwell dogs, but on readability grounds "sore" performed better than "pained", so "pained" was excluded and "sore" retained. The groups used to test readability in the USA and Australia were small (9 and 15 dog owners respectively) and were not representative of the general dog-owning population, which could be seen as a weakness in the study, but all 46 words had been pretested previously in the UK (Reid *et al.* 2013) and the purpose of the Australian and USA tests was purely to identify cultural difference in relation to meaning.

For the purpose of establishing known groups construct validity, expert judgment has been used previously to identify items that could discriminate between unwell and healthy dogs (Reid *et al.* 2013) and that process was repeated here. Although it was considered unlikely that an item that was unable to discriminate healthy from unwell dogs would prove useful in an evaluative context, that possibility cannot be discounted and removed items may be reassessed if the instrument proves not to be responsive to clinical change in further longitudinal studies.

FA to determine the optimum factor model on which to base the HRQL domains of the shortened 22-item instrument was carried out on results from 122 dogs (88 unwell and 34 healthy). The literature includes a range of recommendations regarding the minimum sample size necessary to obtain factor solutions that are adequately stable, including the suggestion that there should be between four and five times as many samples as variables (Floyd & Widaman 1995). On that basis the sample size used here was adequate. However, several workers including Velicer & Fava (1998) found the influence of sample size to be reduced when factor loadings and communalities were high, which was the case in our study where all loadings were greater than 0·5 and 14/25 were greater than 0·7, which is considered high. Similarly, 12/22 communalities were greater than 0·7 and according to MacCallum *et al.* (1999) communalities of 0·6 are considered high. Consequently, the factor model was considered stable.

A useful factor model captures a reasonable amount of the total variance in the data from which it is derived, with higher figures representing better models. A perfect model would account for 100% of the variance in the sample, but this would have the same number of factors as variables and Norman & Streiner (1994) suggested that factors should explain at least 50% of the total variance. The 64% and 62% of the variance captured by EFA and CFA of the 46-item instrument compare well with that accounted for in other proxy instruments to measure the QOL of infants (45%) (Manificat *et al.* 1999), the QOL of older children (62%) (Varni *et al.* 2001), the behaviour and temperament of guide dogs (63%) (Serpell & Hsu 2001) and of pet dogs (57%) (Hsu & Serpell 2003). However, the 22 items comprising the shortened instrument accounted for 72% of the variance compared with 62% for the CFA of the 46 items using the same data set, indicating an improvement in the factor model. This could be a result of the higher loadings (>0·5 *versus* >0·3) representing a closer association of the 22 items with the factors compared with that of the 46 items as a result of the removal of less correlated items that had contributed some measurement "noise". Further to this demonstration of factorial validity, field-testing of the new instrument was designed to confirm that shortening of the instrument had not diminished the psychometric properties of the original. Known group validity was demonstrated by the fact that scores in all four domains of HRQL were significantly different between healthy and unwell dogs. In common with the 46-item instrument, the domain scores in the unwell dog group showed more variation than those for the healthy dogs. The study protocol did not ask clinicians to rate the severity of disease in the unwell dogs, but only specified that cases should be selected on the basis that the condition was likely to affect the QOL. However, the wide interquartile range and the extent of the whiskers in Fig. 2 in the unwell dogs would tend to suggest that there was a wide spread of disease severity with resultant variability in their health status. Subjective evaluation of general behavioural signs such as changes in appetite, activity and sociability has long been reported as changing with ill-health, especially in food animals (Weary *et al.* 2009), but to our knowledge the HRQL domains reported here – E/E, H/C, A/C and C/R – have not been specifically reported in companion animals as likely to change with health status. However, the SF 36 is a generic HRQL instrument for people designed to measure physical and emotional components of health status and it contains the terms "activity", "calm and peaceful", "full of life", "energetic" and "happy" (Ware & Sherbourne 1992). The domain C/R shows more variability in the healthy dogs than was apparent for the other three domains, which is perhaps not surprising given the spectrum of excitability in the healthy dog population. There is also a smaller difference in median scores between the healthy and unwell groups in that domain and more overlap in the interquartile ranges. This may be accounted for by the fact that this domain contains items (*e.g.* "calm") that could reflect relatively stable personality traits, making it more resistant than other domains to change with ill health.

Owners of healthy and unwell dogs for CFA, item reduction and EFA of the 22-item shortened instrument and the field test for the 22-item instrument were drawn in part from a university referral population which may raise some concerns regarding respondent bias. However, we consider that drawing from a variety of sources where possible (primary care and referral) broadens the scope of the recruitment in clinical studies where it is very difficult to control for selection bias. With respect to the use of vets and vet nurses as respondents, who might be influenced by their professional expertise, the questions in the instrument are related to owner-observed behaviours and do not involve any judgment related to health or welfare. Also, the university staff who took part in the study included a mix of administrative and non-veterinary teaching staff in addition to vets and nurses. Because the answers to the questions in the instrument involve an interpretation of behaviour on the part of the owner and that interpretation is best made by the person that knows the dog best, only the primary carer of the dog was recruited to the study. Additionally, if two assessments were carried out they were both completed by the same owner.

Although there was a discrepancy between the types of cases included in each field test, one of which provided the data for the CFA, item reduction and EFA of the shortened instrument (primarily referral), with the other providing the data for the field test of the shortened instrument (primarily first opinion) all dogs were suffering from a condition, usually chronic, likely to affect their QOL. General practitioners regularly treat cases such as osteoarthritis, obesity, diabetes, cardiac failure, chronic skin disease and cancer in primary care practice and, accordingly, we suggest that the impact of any differences is not likely to be significant. The discriminant analysis with cross-validation of the 22-item short form indicated an overall misclassification rate of 15% with 89% of healthy dogs and 77% of unwell dogs classified correctly. These results compare well with those reported for a proxy instrument for pain measurement in communicatively impaired children that correctly classified 92·9% of children with no pain and 71·3% of children in pain with an overall misclassification rate of 13% and which was considered by its developers to have reasonable ability to distinguish between pain and no-pain episodes (Stallard *et al.* 2002). Misclassifications in the study reported here may have been a result of measurement error, or may have occurred because the QOL of some of the healthy dogs was compromised at the time of measurement for reasons other than poor health, or because some of the dogs that were unwell may in any case have been experiencing a good quality of life at the time of measurement.

The ICC values for the domains were greater than 0·7, indicating that test–retest reliability conducted with a two-week interval for the web instrument was good (Rosner 2005). It was assumed that the health status of control dogs would not change over the 2-week period between the completion of questionnaires, and respondents would not remember their previous responses. This result indicated that the reliability of the shortened instrument was improved compared with the original where ICC values were greater than 0·6. However, the current test was carried out with 49 owners compared with 16 for the original and this may have contributed to the improved result (Reid *et al.* 2013).

Best practice in instrument design dictates that when a questionnaire instrument is presented in a new way, for example moving from paper to web, or re-design of presentation, it is recommended that the instrument be re-tested in its new form to ensure that changes in format or design have not altered its measurement properties. For logistical reasons, field testing of the shortened 22-item instrument was carried out using a web-based platform and not via a mobile phone application.

In conclusion, this study has provided evidence for the instrument's ability to detect the HRQL impact of disease. Although this study has provided initial evidence for the reliability and validity of the shortened instrument, it is important to emphasise that validity is not determined by a single statistic, but by a body of research supporting claims that the instrument is valid for particular purposes, with defined populations and in specified contexts (Streiner & Norman 2008). Accordingly, future research will seek to provide such evidence, including evidence for its responsiveness in longitudinal studies.

Acknowledgements

The authors wish to thank all the dog owners who willingly participated in our studies as well as veterinary surgeons Andrew Ash, Grove Lodge, Worthing, Charlie Carnochan, Donview Veterinary Centre, Inverurie, Edwina Gildea, Zoetis, Ann Hutchison, Dalblair Veterinary Surgery, Ayr and the nursing staff and clinicians at the University of Glasgow Small Animal Hospital and Vets Now clinics. Thanks to Zoetis UK for funding this study.

Conflict of interest

Professor JR is a Director of NewMetrica Ltd, which is the developer and supplier of the resulting instrument.

References

Abell, N., Springer, D. W. & Kamata, A. (2009) Developing and Validating Rapid Assessment Instruments. Oxford University Press, New York, NY, USA. p 10

Boulos, M. N. K., Brewer, A. C., Karimkhani, C., *et al.* (2014) Mobile medical and health apps: state of the art, concerns, regulatory control and certification. *Online Journal of Public Health Informatics* **5**, 229

Boyle, G. J. (1991) Does item homogeneity indicate internal consistency or item redundancy in psychometric scales? *Personality and Individual Differences* **12**, 291-294

Brod, M., Tesler, L. E. & Christensen, T. L. (2009) Qualitative research and content validity: developing best practices based on science and experience. *Quality of Life Research* **18**, 1263-1278

Budke, C. M., Levine, J. M., Kerwin, S. C., *et al.* (2008) Evaluation of a questionnaire for obtaining owner-perceived, weighted quality-of-life assessments for dogs with spinal cord injuries. *Journal of the American Veterinary Medical Association* **233**, 925-930

Cheak-Zamora, N. C., Wyrwich, K. W. & McBride, T. D. (2009) Reliability and validity of the SF-12v2 in the medical expenditure panel survey. *Quality of Life Research* **18**, 727-735

Coste, J., Walter, E. & Venot, A. (1995) A new approach to selection and weighting of items in evaluative composite measurement scales. *Statistics in Medicine* **14**, 2565-2580

Coste, J., Guillemin, F., Pouchot, J., *et al.* (1997) Methodological approaches to shortening composite measurement scales. *Journal of Clinical Epidemiology* **50**, 247-252

Coste, J., Bouée, S., Ecosse, E., *et al.* (2005) Methodological issues in determining the dimensionality of composite health measures using principal component analysis: case illustration and suggestions for practice. *Quality of Life Research* **14**, 641-654

DeVellis, R. F. (2012) Factor analysis. In: Scale Development: Theory and Applications, 3rd edn, pp 115-158. Sage, Los Angeles, CA, USA

Favrot, C., Linek, M., Mueller, R., *et al.* (2010) Development of a questionnaire to assess the impact of atopic dermatitis on health-related quality of life of affected dogs and their owners. *Veterinary Dermatology* **21**, 63-69

Fayers, P. M. & Machin, D. (2013) Quality of Life: The Assessment, Analysis and Interpretation of Patient-Reported Outcomes. John Wiley & Sons, New York, NY, USA

Floyd, F. J. & Widaman, K. F. (1995) Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment* **7**, 286-299

Freeman, L. M., Rush, J. E., Farabaugh, A. E., *et al.* (2005) 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
Goetz, C., Coste, J., Lemetayer, F., *et al.* (2013) Item reduction based on rigorous methodological guidelines is necessary to maintain validity when shortening composite measurement scales. *Journal of Clinical Epidemiology* **66**, 710-718
Greenwood, M. C., Hakim, A. J., Carson, E., *et al.* (2006) Touch-screen computer systems in the rheumatology clinic offer a reliable and user-friendly means of collecting quality-of-life and outcome data from patients with rheumatoid arthritis. *Rheumatology* **45**, 66-71
Hsu, Y. & Serpell, J. A. (2003) Development and validation of a questionnaire for measuring behavior and temperament traits in pet dogs. *Journal of the American Veterinary Medical Association* **223**, 1293-1300
Johnston, C. C. (1998) Psychometric issues in the measurement of pain. In: Measurement of Pain in Infants and Children: Progress in Pain Research and Management, Vol. 10. Eds G. A. Finley and P. J. McGrath. IASP Press, Seattle, WA, USA. pp 5-20
Krebs, P. & Duncan, D. T. (2015) Health app use among US mobile phone owners: a national survey. *JMIR mHealth and uHealth* **3**, e101
Las Hayas, C., Quintana, J. M., Padierna, J. A., *et al.* (2010) Use of rasch methodology to develop a short version of the health related quality of life for eating disorders questionnaire: a prospective study. *Health and Quality of Life Outcomes* **8**, 1-12
Lavan, R. P. (2013) Development and validation of a survey for quality of life assessment by owners of healthy dogs. *The Veterinary Journal* **197**, 578-582
Lynch, S., Savary-Bataille, K., Leeuw, B., *et al.* (2011) Development of a questionnaire assessing health-related quality-of-life in dogs and cats with cancer. *Veterinary and Comparative Oncology* **9**, 172-118
MacCallum, R. C., Widaman, K. F., Zhang, S. B., *et al.* (1999) Sample size in factor analysis. *Psychological Methods* **4**, 84-99
Manificat, S., Dazord, A., Langue, J., *et al.* (1999) A new instrument to evaluate infant quality of life. *MAPI Research Institute, Quality of Life Newsletter* **23**, 7-8
Niessen, S. J., Powney, S., Guitian, J., *et al.* (2012) Evaluation of a quality of life tool for dogs with diabetes mellitus. *Journal of Veterinary Internal Medicine* **26**, 953-961
Noli, C., Minafo, G. & Galzerano, M. (2011) Quality of life of dogs with skin diseases and their owners. Part 1: development and validation of a questionnaire. *Veterinary Dermatology* **22**, 335-343
Norman, G. R. & Streiner, D. L. (1994) Principal components and factor analysis. In: Biostatistics, The Bare Essentials. Mosby-Year Book, St. Louis, MO, USA. pp 129-142
Nunnally, J. C. & Bernstein, I. H. (1994) Psychometric Theory. 3rd edn. McGraw-Hill, New York, NY, USA
Osse, B. H. P., Vernooij-Dassen, M., Schade, E., *et al.* (2007) A practical instrument to explore patients' needs in palliative care: the problems and Needs in palliative care questionnaire-short version. *Palliative Medicine* **21**, 391-399
Reid, J., Wiseman-Orr, M. L., Scott, E. M., *et al.* (2013) Development, validation and reliability of a web-based questionnaire to measure health-related quality of life in dogs. *Journal of Small Animal Practice* **54**, 227-233
Rosner, B. (2005) Fundamentals of Biostatistics. Duxbury Press, Belmont, CA, USA
Serpell, J. A. & Hsu, Y. (2001) Development and validation of a novel method for evaluating behavior and temperament in guide dogs. *Applied Animal Behaviour Science* **72**, 347-364
Shevlin, M. & Miles, J. N. V. (1998) Effects of sample size, model specification and factor loadings on the GFI in confirmatory factor analysis. *Personality and Individual Differences* **25**, 85-90
Shrout, P. E. & Fleiss, J. L. (1979) Intraclass correlation: uses in assessing rater reliability. *Psychological Bulletin* **86**, 420-428
Stallard, P., Williams, L., Velleman, R., *et al.* (2002) The development and evaluation of the pain indicator for communicately impaired children (PICIC). *Pain* **2**, 149
Stanton, J. M., Sinar, E. F., Balzer, W. K., *et al.* (2002) Issues and strategies for reducing the length of self-report scales. *Personnel Psychology* **55**, 167-194
StatSoft Inc. (2003). Principal components and factor analysis (1984-2003). http://www.statsoft.com/textbook/stfacan.html. Accessed April 3, 2017.
Streiner, D. L. & Norman, G. R. (2008) Health Measurement Scales: A Practical Guide to their Development and Use. 4th edn. Oxford University Press, Oxford, UK. p 399
Teasdale, G. & Jennett, B. (1974) Assessment of coma and impaired consciousness. A practical scale. *Lancet* **2**, 81-84
Varni, J. W., Seid, M. & Kurtin, P. S. (2001) PedsQL 4·0: reliability and validity of the Pediatric Quality of Life Inventory version 4·0 generic core scales in healthy and patient populations. *Medical Care* **39**, 800-812
Velicer, W. F. & Fava, J. L. (1998) An evaluation of the effects of variable sampling on component, image, and factor analysis. *Psychological Methods* **3**, 231-251
Velicer, W. E., Peacock, A. C. & Jackson, D. N. (1982) A comparison of component and factor patterns: a Monte Carlo approach. *Multivariate Behavioral Research* **17**, 371-388
Ware, J. E. & Sherbourne, C. D. (1992) The MOS 36-Item Short-Form Health Survey (SF-36): I Conceptual framework and item selection. *Medical Care* **30**, 473-483
Weary, D. M., Huzzey, J. M. & von Keyserlingk, M. A. G. (2009) Board-invited review: using behavior to predict and identify ill health in animals. *Journal of Animal Science* **87**, 770-777
Wiseman, M. L., Nolan, A. M., Reid, J., *et al.* (2001) Preliminary study on owner-reported behaviour changes associated with chronic pain in dogs. *Veterinary Record* **149**, 423-524
Wiseman-Orr, M. L., Nolan, A. M., Reid, J., *et al.* (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, M. L., Scott, E. M., Reid, J., *et al.* (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, J. I., Hewson, C. J., Stryhn, H., *et al.* (2005a) Development of a discriminative questionnaire to assess nonphysical aspects of quality of life of dogs. *American Journal of Veterinary Research* **66**, 1453-1460
Wojciechowska, J. I., Hewson, C. J., Stryhn, H., *et al.* (2005b) Evaluation of a questionnaire regarding nonphysical aspects of quality life in sick and healthy dogs. *American Journal of Veterinary Research* **66**, 1460-1467
Yam, P. S., Butowski, C. F., Chitty, J. L., *et al.* (2016) Impact of canine overweight and obesity on health-related quality of life. *Preventive Veterinary Medicine* **127**, 64-69
Yazbek, K. V. & Fantoni, D. T. (2005) Validity of a 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