Development and validation of a quality of life assessment tool for use in kennelled dogs (*Canis familiaris*)

Jenna L. Kiddie[a], Lisa M. Collins[b],*

[a] *Veterinary Clinical Studies, Royal Veterinary College, Hawkshead Lane, North Mymms, Hatfield, Herts AL9 7TA, UK*
[b] *School of Life Sciences, University of Lincoln, Brayford Pool, Lincoln LN6 7TS, UK*

ARTICLE INFO

ABSTRACT

There is currently no objective and validated quality of life assessment tool available to assess the quality of life of domestic dogs in kennels. This study aimed to develop a validated scoring system to assess the quality of life of dogs kennelled in rehoming centres. Objective animal-based measures of welfare and quality of life, identified from the scientific literature, were included in the scoring system to indicate negative and positive quality of life. Each item was scored using a binary system to minimise subjectivity in scoring. Dogs were recruited from 13 rehoming centres into four different treatment groups: group NS consisted of dogs that had newly arrived at the shelter and received a standard husbandry routine; group NE consisted of dogs that had newly arrived at the shelter and received an enrichment programme; group LS consisted of dogs that had been in the centre for more than 30 days and received a standard husbandry routine; and group LE consisted of dogs that had been in the shelter for more than 30 days and received an enrichment programme. Inter-observer reliability of each item was established and internal consistency of the entire score was assessed: seven unreliable items were removed and the scoring system was refined. 26.7% (ICC1 = 0.267) of the variation in quality of life scores was explained by rehoming centre group membership, and rehoming centre groups could reliably be differentiated in terms of quality of life scores (ICC2 = 0.832). Therefore, quality of life scores varied between rehoming centres but other factors must exist to explain the remaining 73.3% of the variation in quality of life scores. The addition of an enrichment programme and later recruitment onto the study increased quality of life scores by 0.035 ± 0.027 (SE) and 0.086 ± 0.027 (SE), respectively. These increases represent an 8.27% and a 20.33% improvement on the overall (across all treatment groups) mean QoL score (0.423). Evidence of good content, construct and criterion validity was established, however, internal consistency was found to be poor, indicating that the reliability of the score could be improved.

1. Introduction

Quality of life (QoL) can be defined as:

"*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 unconscious process, with a complexity appropriate to the cognitive capacity of the individual)*" (Wiseman-Orr et al., 2006).

---

* Corresponding author.
E-mail addresses: jenna.kiddie@anglia.ac.uk (J.L. Kiddie),
lcollins@lincoln.ac.uk (L.M. Collins).

Many QoL measures have been developed for humans in recent years (e.g. Auquier et al., 2003; Boyer et al., 2010; Fitzsimmons et al., 1999; Gómez-Gallego et al., 2012; Maillé et al., 1997; Patrick et al., 1999; Pérez-Campos et al., 2011; Rajmil et al., 2011; Suzukamo et al., 2006). These usually employ self-report via Likert-type rating scales (e.g. Auquier et al., 2003; Boyer et al., 2010; Fitzsimmons et al., 1999; Maillé et al., 1997; Patrick et al., 1999; Pérez-Campos et al., 2011), which is regarded as the gold standard method of measuring QoL in humans (Bateson and Matheson, 2007; Boissy et al., 2007a; Scott et al., 2007). Questionnaires developed for use by caregivers acting as proxies for humans unable to speak for themselves, such as infants (e.g. Rajmil et al., 2011; Rautava et al., 2009) or the cognitively impaired (e.g. Kuo et al., 2010; León-Salas et al., 2011) can be adapted to assess the emotional state of animals (Wojciechowska et al., 2005). These can be adapted for completion by owners, veterinarians, or others familiar with the individual animal in question (Hewson et al., 2007; Taylor and Mills, 2007a,b). These questionnaires are often based on objective list theory (e.g. Wojciechowska et al., 2005) related to feeding, activity level, environment, social interaction, behaviour, and feelings of the animal as assessed by the owner. However, this method focuses largely on input-based measures and does not take into account individual differences in preference. Traditionally, QoL and welfare assessments have focussed on input-based measures (Whay et al., 2003). This is partly because physical resources tend to remain constant and can be measured objectively. Input-based assessments can include measures of stockmanship, environmental and animal factors that affect the welfare of the animal; for example, stockman competency and handling skills, suitable housing and diet, and the genetic suitability of animals to their management system (Whay et al., 2003). However, providing good management and recommended environmental resources does not guarantee that an animal is healthy and experiencing a high standard of welfare (Whay et al., 2003; Whitham and Wielebnowski, 2009). It may be particularly difficult to include causal indicators in companion animal QoL assessments (as compared with farm animal assessments), as there is currently no general agreement on the basic needs of companion animals (Taylor and Mills, 2007b). Studies of companion animal QoL can attempt to avoid this problem of a lack of information regarding needs by asking owners about their pets' individual preferences. However, this is unlikely to be possible where animals are kept in establishments such as rescue or rehoming centres where previous preferences are not known (Taylor and Mills, 2007b). Therefore, without the evidence of behavioural tests to establish individual preferences, QoL should rely wholly on animal-based measures, i.e. health and behaviour, as emphasised in the Welfare Quality assessments (Temple et al., 2011).

Many animal-based measures rely upon subjective assessments (e.g. Kessler and Turner, 1997; Wemelsfelder et al., 2009) and require the owner to interpret their animal's behaviour; however, many owners are not accurate in their interpretation of their pet dogs, and may therefore introduce anthropomorphic or biased results (Kiddie, 2004, unpublished data; Timmins et al., 2007). Main et al.

(2000) suggest that there is no inherent problem with subjective assessment as long as the measures are repeatable. However, these evaluations may be subject to the value system and aesthetic sensibilities of the evaluator (Timmins et al., 2007). Therefore, questionnaires should be designed so that owners acting as proxies may assess their pets using objective methods (Taylor and Mills, 2007b).

When developing welfare assessment tools, both reliability and validity must be considered in order to assess how well the tool measures the construct for which it was designed (FAWC, 2009; Martin and Bateson, 2007). Reliability measures the extent to which a measurement is repeatable and consistent (Carmines and Zeller, 1979). Reliability can be measured within the observer (intra-observer), between observers (inter-observer), within the subject (test–retest), and within components of measures designed to assess the same construct (internal consistency) (Litwin, 1995; Taylor and Mills, 2006). Validity assesses the extent to which the measurement does in fact measure the construct it was designed for (Carmines and Zeller, 1979; Martin and Bateson, 2007; Petrie and Watson, 2006; Wiseman-Orr et al., 2011, 2006). This is the most fundamental attribute of a measurement tool (Wiseman-Orr et al., 2011).

Two issues must be addressed before reliability and validity can be assessed: the purpose of the assessment and the standardisation of the measurement process (Taylor and Mills, 2006). What the tool is designed to measure must be clearly identified for it to be valid (Groth-Marnat, 2009; Taylor and Mills, 2006) and standardisation is necessary for the tool to be reliable (Taylor and Mills, 2006). Additionally, when developing welfare assessments, consideration must be given to what measures are relevant but also which are practical and feasible (FAWC, 2009; Martin and Bateson, 2007; Taylor and Mills, 2006; Whay et al., 2003). Measurements that are time-consuming or difficult to implement are unlikely to be performed accurately or reliably (Taylor and Mills, 2006). Similarly, the quality of the measure is important to consider, for example the quality of data derived from computerised veterinary medical record databases relies on personnel's accuracy and completeness when filling out the medical records and transferring data from original paper records to computerised copies (Pollari et al., 1996). Many tools will need to be refined in order to make them more reliable and valid but also more practical to use in the field. Decisions on how to reduce the number of items in the measurement tool may be based on poor intra-observer, inter-observer and test–retest reliability. However, validity of the refined tool must not be assumed and therefore must be retested using the same procedure (Taylor and Mills, 2006).

The aim of the study reported here was to develop and validate a novel tool to assess the QoL of kennelled dogs at a moment in time, that could be used to track QoL over an extended period. The reliability and validity of this scoring system was tested through its use in Dogs Trust kennels by Dogs Trust staff. This was achieved through the following objectives:

- Analysing inter-observer reliability of the score's individual items.

- Analysing internal consistency of the score.
- Establishing validation by way of:

- Content evidence, provided by the review and approval, by an expert panel, of items generated by a literature review.
- Construct evidence, provided by response evidence.
- Criterion evidence, provided by testing for differences of QoL scores between four groups of dogs:

- 16 dogs in each of 13 participating Dogs Trust centres were divided into four treatment groups as follows:

- Group NS: dogs that were newly admitted to the centre and received standard husbandry.
- Group NE: dogs that were newly admitted to the centre and received an additional human-interaction enrichment programme.
- Group LS: dogs that were in the centre for at least 30 days and received standard husbandry.
- Group LE: dogs that were in the centre for at least 30 days and received an additional human-interaction enrichment programme.

The enriched dogs were hypothesised to have higher (better) QoL scores than dogs receiving a standard husbandry routine. Long stay dogs were also hypothesised to score higher for QoL due to habituation to the kennel environment. Therefore, the groups of dogs were expected to have differing QoL scores, with scores increasing for each group in the listed order.

2. Materials and methods

2.1. Creating the score (content validity)

Content evidence evaluates whether the content of the instrument accurately represents the concept it was designed to measure (Morgan et al., 2006; Wiseman-Orr et al., 2011). Establishing content evidence of validity may involve a panel of experts reviewing items generated from a literature search of the defined concept being measured (Morgan et al., 2006). A literature search was performed by one researcher to create the score with good content validity rather than generating new items that had not previously been found to be potentially useful indicators of welfare or quality of life. Two further experts then reviewed and approved the list of items generated from the literature search.

2.1.1. Behavioural measures: assessing negative emotions

Stephen and Ledger (2005) compiled an ethogram of 15 behaviours that were associated with stress: 10 were taken from previous studies and five were identified following consultation with shelter staff. These 15 behaviours were: repetitive pacing; wall bouncing; tail chasing; circling; play bouncing; chewing bedding; self-licking; polydipsia; panting; lack of appetite; listlessness; escape attempts; hiding; chewing bars; and excessive vocalisations. The behaviours in this study, with the exception of excessive vocalisations, were entered as items into the current quality of life (QoL) score (Table 1 and Supplementary data). Excessive vocalisation was excluded as the definition used in Stephen and Ledger's paper was "dog barks for prolonged period (>1 min) in the visual and auditory absence of people and other dogs": dogs kennelled in Dogs Trust rehoming centres are very rarely in auditory or visual absence of other dogs, therefore this measure would not be applicable to the research context within this study. Additional behaviours, which previous studies have found to be associated with poor welfare were also added to the score (Table 1 and Supplementary data): a low posture, coprophagy (Beerda et al., 1999), paw lifting (Beerda et al., 1999, 2000), changing from one state of locomotion to another, sitting, standing, nosing (Beerda et al., 2000), whining (Walker et al., 2009), aggression to conspecifics (Beerda et al., 1999), startling (Hiby et al., 2006), and walking (Beerda et al., 2000; Hiby et al., 2006). All of these behaviours were included as indicators of negative emotional state, were measured from a distance and were therefore unprovoked by the observers' presence. Sitting and standing are not immediately obvious behaviours to choose to represent negative emotional states, but Beerda et al. (2000) found that these behaviours were performed for longer periods of time in conditions thought to be compromising to dog welfare, compared with conditions that were thought to promote better welfare.

Behaviours performed by stressed dogs in a human social context were also included (Table 1 and Supplementary data). The following behaviours were recorded as present or absent in response to the approach and interaction of the observers: increased oral behaviours (Beerda et al., 1998), ambivalent postures, nosing, sitting and aggression (Beerda et al., 2000).

2.1.2. Behavioural measures: assessing positive emotions

Assessors were asked to assess dogs in the few minutes before the dogs were given their morning feed as anticipatory behaviours are thought to indicate positive emotions (Boissy et al., 2007b; Van der Harst and Spruijt, 2007) (Table 1 and Supplementary data). The anticipatory behaviours included in this part of the score were: increased locomotor activity; frequent behavioural transitions; grooming; alertness; scanning; exploration; orientation towards the place where food is presented; play markers (in this case the play bow); erect ears; higher body posture; and more time at the front of the cage (Boissy et al., 2007b) (Table 1 and Supplementary data). Grunting (Bleicher, 1963), lying down (Beerda et al., 1998), play, and affiliative behaviours (Boissy et al., 2007b) – licking the other dog's muzzle (Bauer et al., 2009), initiating physical contact, tail wagging, and body shaking (Rehn, 2011) – were also included as indicators of positive emotion (Table 1 and Supplementary data).

Behaviours performed by relaxed dogs in a human social context were also included: tail wagging, lying down in a rested position (Beerda et al., 2000), making physical contact, body shaking (Rehn, 2011) and playing (Boissy et al., 2007b) (Table 1 and Supplementary data).

Table 1
Ethogram of behavioural items comprising the quality of life score.

Behaviour	Description	Reference
Indicator of negative emotional state: unprovoked behaviours		
Pace repetitively	Dog repeatedly (>3 times) paces around kennel in a fixed route	Stephen and Ledger (2005)
Wall bouncing	Dog repeatedly (>3 times) jumps up kennel wall from one side to another	
Tail chasing	Dog chases its tail repeatedly (>3 times)	
Circling	Dog repeatedly walks around in small circle (>3 times)	
Play bouncing	Dog repeatedly displays the play bow posture (>3 times)	
Chewing bedding	Dog chews its own bedding	
Drink excessively	Dog drinks large volumes of water, in excess of what is normal	
Panting	Dog pants for reasons unrelated to physical exertion or warm ambient temperature (only record if temperature <25 °C)	
Lack of appetite	Did the dog eat over half it's food?	
Listless	Dog is withdrawn and unresponsive to commands	
Escape attempt	Dog attempts to escape kennel in a forceful manner whenever the kennel door is opened	
Hide	Dog is obscured from view of kennel staff, behind its bed or other kennel furniture for prolonged periods when not asleep (>2mins)	
Chewing bars	Dog repeatedly chews and bites at the bars of the kennel (>20secs)	
Low posture	Tail is lowered, ears are back and legs are bent	Beerda et al. (1998)
Coprophagy	Did the dog eat its own or another dog's faeces?	
Paw lift	A forepaw is lifted off the ground and held there	Beerda et al. (1999) and Beerda et al. (2000)
Change movement	Change from one type of locomotion to another, e.g. changing from walking to sitting, or sitting to standing	Beerda et al. (2000)
Sit	Hindquarters and at least foot two pads in contact with ground, front legs extended	
Stand	Positioned with four feet in contact with ground and legs almost or fully extended	
Nose	The nose is held close to or touching a surface, and/or sniffing the surface	
Whine	High pitched vocalisation	Walker et al. (2009)
Aggression	Any lip lifting, growling, snapping, or biting	Beerda et al. (1999)
Startle	Legs flex briefly, and body and head quickly and briefly move back, usually in response to a sudden noise, or dog quickly moves backwards a few paces	Hiby et al. (2006)
Walk	Travels forward without obviously investigating its environment	Beerda et al. (2000) and Hiby et al. (2006)
Indicators of negative emotional state: provoked behaviours (in response to observer approach or interaction)		
Oral behaviours	Includes tongue out; tip of tongue briefly extended; snout licking; lip licking; swallowing, lip smacking	Beerda et al. (1998)
Ambivalent posture	A crouched body posture accompanied by a position that is higher than the breed-specific position; or a high body posture accompanied by a position of the tail that is below normal	Beerda et al. (2000)
Nose	The nose is held close to or touching a surface, and/or sniffing the surface	
Sit	Hindquarters and at least foot two pads in contact with ground, front legs extended	
Aggression	Any lip lifting, growling, snapping, or biting	
Indicators of positive emotional state: unprovoked behaviours		
High level of activity	Increased levels of any locomotion or movement	Boissy et al. (2007b)
Changing from one behaviour to another	The dog changes from one behaviour to another during the observation period	
Groom	Behaviours directed to own body, including licking, stretching, scratching	
Alert	Generally inactive but with eyes open, and head and ears moving, can be lying down, sitting or standing	
Scan	Eyes continuously move to view the environment	
Explore	Walks with nose close to surfaces or sniffing objects	
Orientation towards where food is presented	Did the dog face or approach the place where its food is placed?	
Play bow	Forequarters are lowered to the ground, with rump raised	
Erect ears	Ears held forward	
High body posture	Breed specific posture shown by dogs under neutral conditions, but with a higher tail or head elevated and ears forwards, or dog standing extremely erect	
Time at front of kennel	Time spent in the half of the kennel closest to the external wall/door	
Grunt	Isolated intense expiration (breathing out)	Bleicher (1963)
Lie down	Most of body in contact with ground	Beerda et al. (1998)
Object play	Any vigorous or galloping gaited behaviour directed towards a toy or other object, including chewing, biting, shaking it from side to side, batting it with a paw	Boissy et al. (2007b)
Play with other dog	Leaps onto another dog, with body relaxed, stands on hind legs and paws at other dog, places mouth around muzzle, head, neck, or legs of other dog with little pressure, pats another dog with a forepaw, lifting both front paws off the ground rapidly to bounce up and down, done in front of and orientated towards another dog, races in pursuit of another dog or play bows to the other dog by lowering forequarters and raising hindquarters	

Table 1 (*Continued*)

Behaviour	Description	Reference
Lick other dog's muzzle	The dog licks the muzzle of its kennelmate	Bauer
Lick	The dog licks any part of the assessor's body	et al.
Tail wagging	Repetitive wagging movements of the tail	(2009)
Body shake	Dog shakes its whole body briefly as if drying itself	(2011)
Indicator of positive emotional state: provoked behaviours (in response to observer approach or interaction)		
Tail wagging	Repetitive wagging movements of the tail	Beerda et al. (2000)
Lie down	Most of body in contact with ground	
Initiate physical contact	Dog starts an interaction with the assessor or kennelmate	Rehn (2011)
Body shake	Dog shakes its whole body briefly as if drying itself	
Play with assessor	Play bows to assessor, engages in tug of war with the ragger	Boissy et al. (2007b)

2.1.3. Behavioural measures: internal consistency

The assessors filled in the score while observing the dogs rather than from memory after the session finished. This method was used to increase accuracy. However, one consequence of this approach is that items early in the score may have needed reassessment by the end of the observation session. For example, if the assessor completed the score sheet factor by factor, then behavioural measures listed at the start of the score sheet may be left empty if the dog did not perform this behaviour within the initial period of observation, but may have done so by the end of the 2 min session. In this case, the assessor would need to go back to this item and amend their response. Therefore, the assessors were instructed to quickly go through the score a second time after the 2 min observation session ended in order to make any such amendments. As a result some memory recall was required and could have affected internal consistency.

Internal consistency measures the correlation between different items in a questionnaire that are designed to measure the same construct and is therefore a measure of reliability. The two items in the questionnaire – "changing from one behaviour to another" and "changing type of movement" – should agree with items that are their component parts. "Changing from one type of behaviour to another" should be assessed as present if a dog is observed doing more than one of: play bowing; wall bouncing; chasing its tail; circling; walking; standing; sitting; startling; lying down; and changing type of movement. "Changing type of movement" should be assessed as present if a dog is observed doing more than one of: chasing tail; circling; walking; sitting; standing; and wall bouncing. These component items were chosen as they are mutually exclusive of each other, therefore the observer should have recorded a shift from one to another as a change, either "changing from one behaviour to another" or "changing type of movement".

2.1.4. Physical measures

Three physical measures were included in the score: body condition score (Brinkmann et al., 2013; Ellegaard et al., 2010); presence of scurf (Buske-Kirschbaum et al., 2006; Hendrix and Peters, 2007); and eye discharge (Varnell et al., 1995; Wu and Ariyasu, 1999). Body condition score (BCS) was assessed on a standard 9-point scale from physical handling and observation of the dog (Nestlé Purina Petcare Centre, 2014). Scurf, or skin flakes, in the hair along the back of the dog was categorised as present or absent, as was eye discharge. As physical measures change state at a much slower rate than behavioural measures, they did not require amendment after completion of the observation session.

2.1.5. The QoL score

The score consists of five sections. The first three sections require the assessor to observe the dogs from a distance in order to assess unprovoked behaviours. Section 1 relates to the morning feed, including questions about anticipatory behaviours and the dog's appetite, i.e. whether the dog ate over of half its food. These are indicators of positive emotional state. Section 2 assesses the dog's general behaviour in the kennel and includes indicators of both positive and negative emotional states. Section 3 relates to the dog's interactions with its kennel mate and includes affiliation, play and aggression, therefore measuring both valences of emotional state. Section 4 assesses the dog's reaction to being approached and handled by the assessor, both positively and negatively. Therefore, after the assessor has answered the previous sections where they observe the dog from a distance, they must approach the dog, enter its kennel and make contact with it, if safe to do so, i.e. the dog is not showing any aggressive or appeasement behaviours. This section also includes questions regarding affiliative and aggressive behaviours. The assessor is also required to engage the dog in play using a ragger toy in order to assess the dog's willingness to play. In Section 5, the assessor gives the dog a brief physical examination in order to assess BCS, the presence of scurf and eye discharge. The QoL score is a binary 1/0 scoring system to keep it as objective as possible, therefore, each item in the score was recorded as being observed or not observed.

2.2. Testing the score (response and convergent validity)

2.2.1. Study site

Thirteen Dogs Trust centres, distributed across Great Britain, recruited dogs from their existing population of kennelled dogs between 1st September 2011 and 31st November 2011. Each centre was asked to complete the QoL scores for their recruited dogs within two months of the QoL score training day held for staff in that centre. Nine of the 13 centres were situated in a quiet location; two were situated in a quiet location but experienced intermittent noise; and two were in a noisy location. All centres bar one had line block kennels: seven had glass fronted

kennels; nine had bar fronted kennels; two had entirely indoor line block kennels; and one had line block kennels that only had one compartment, rather than two. All centres pair housed dogs unless the dog had to be singly housed for behaviour reasons. However, eight centres also habitually housed dogs singly and two centres also housed their dogs in small groups. All centres provided the majority of dogs with beds and blankets; however two gave some dogs blankets only. Four centres provided bunk or raised beds for their dogs and three centres provided armchairs or sofas. Nine centres fed their dogs twice daily, and four centres fed their dogs twice or three times daily, depending on the individual dog's requirements. All centres fed their dogs the same brand of dry complete food. All centres lead-walked their dogs and allowed free running exercise.

2.2.2. Study animals

Each centre was asked to recruit (i) eight newly admitted dogs, and (ii) eight dogs that had been in the centre for at least 30 days. Dogs were recruited alphabetically to avoid recruitment based on appearance or demeanour. The 13 Dogs Trust centres partaking in this study applied the treatment groups as stated. Of these, six recruited the correct number of dogs for each treatment group. The remaining seven centres recruited differing numbers of dogs per group, due to time restraints, giving a total actual sample size of $n = 202$, rather than the planned $n = 224$. 102 (55.8%) were male (sex was not known for 5.5% dogs). No information was collected regarding the dogs' age as it is very difficult to accurately estimate a dog's age and not all dogs were relinquished to the centre by an owner who could give their dog's age. Similarly, no information was collected regarding the dogs rearing history as this would not have been known in many cases.

The recruited dogs were then randomly allocated to a further two groups: (i) four dogs from the newly admitted group and four dogs from the long stay group were allocated to a standard treatment group, where they received the standard husbandry routine for that centre; (ii) four dogs from the newly admitted group and four dogs from the long stay group were allocated to an enriched treatment group. Thus the sample of 16 dogs per centre was divided into four treatment groups as follows, with a planned total sample size (across all participating rehoming centres) of $n = 56$ per group: group NS – dogs that were newly admitted to the centre and received standard husbandry (actual $n = 53$); group NE: dogs that were newly admitted to the centre and received an additional human-interaction enrichment programme (actual $n = 48$); group LS: dogs that were in the centre for at least 30 days and received standard husbandry (actual $n = 52$); group LE: dogs that were in the centre for at least 30 days and received an additional human-interaction enrichment programme (actual $n = 49$).

Manipulating the groups of dogs to have different levels of QoL allowed convergent evidence (criterion validity) to be established, i.e. if the tool is shown to agree with similar constructs that are hypothesised to agree. A disagreement suggests that either the tool is not measuring the intended

construct or the hypothesised link between the constructs is not correct (Suen and Rzasa, 2004).

2.2.3. Enrichment treatment

For dogs in groups NE and LE, the enrichment programme was carried out away from the kennel for 12 min on six consecutive days, starting from the day of recruitment. The 12 min of enrichment consisted of four stages: 3 min of the handler sitting on the floor and encouraging the dog to make body contact, for example, leaning against, sitting, or lying down next to the handler. The handler used slow hand motions to massage the shoulders, neck, back and hindquarters of the dog and spoke to it in a soothing voice. 2 min of the handler using a soft brush to groom the dog. 5 min of the handler using clicker training to teach basic commands, such as sit or lie down. 2 min of relaxed massage as above. These specific enrichment activities were implemented as they have been found to be enriching to dogs in previous studies (Hennessy et al., 1998; Hubrecht, 1993; Valsecchi et al., 2007).

2.2.4. Staff training

Response evidence evaluates the extent to which the types of participant responses match the intended construct (Morgan et al., 2006; Suen and Rzasa, 2004) and also the extent to which raters are influenced by irrelevant factors when responding (Morgan et al., 2006). Threats to this type of construct validity can be reduced by providing the raters with training; descriptions of rater training forms part of the evidence (Suen and Rzasa, 2004). Raters should attend training sessions to ensure that they all receive the same information thus preventing multiple methods of rating (Suen and Rzasa, 2004).

In order to encourage standardised implementation of the enrichment programme and the QoL score assessments across centres and to promote response evidence of construct validity, a staff training day was conducted at each of the 13 centres for all relevant staff. At least two of the members of staff that were to assess and enrich the dogs were presented with a detailed explanation of the scoring system and the experimental protocol. Question and answer sessions were held to encourage staff to ask any questions they had about the procedure.

2.2.5. Conducting the QoL assessment

All dogs from each of the four treatment groups were assessed on the 7th day after recruitment to the study. Each dog was assessed in its kennel by two different members of staff to allow inter-observer reliability to be calculated. The assessors were blinded to the dogs' treatment group to avoid any bias in their scoring. The assessors were allowed to assess each dog simultaneously (but still independently of each other in order to avoid consensual drift) on the first three assessment observation visits but were asked to assess each dog separately on the fourth visit as they would be interacting with the dog and therefore the dog may act differently in the presence of both assessors than when only interacting with one. Questions relating to the morning feed (Section 1) could only be answered on the first assessment observation, directly before the morning feed was given (between 08:00 h and 09:00 h). Similarly,

the assessor was only required to interact with the dog during assessment observation four, which was completed before 13:00 h. Each assessment observation lasted for two minutes, with the exception of the fourth which allowed an additional two minutes for the assessor to engage in play and give the brief physical examination of the dog. In total, there were 74 assessors, assessing between one and 16 dogs each.

2.2.6. Calculating the QoL score

A per-dog QoL score average was calculated over the four different assessment observations during the course of one morning (between 08:00 h and 13:00 h). To calculate the score, proportions of the positive indicators present and the negative indicators present were calculated. The proportion of the negative indicators present was then subtracted from the proportion of positive indicators present to produce a QoL score. The two scores, one provided by each of the two observers, were then averaged to produce one score for each dog.

2.3. Statistical analysis

All analyses were performed in R statistical programming language v3.0.1. (R Core Team, 2013), with the exception of the inter-observer reliability analysis which was performed using Randolph's online kappa calculator (Randolph, 2008) and internal consistency which was analysed using SPSS (IBM SPSS Statistics 20).

2.3.1. QoL scores

Testing the score was based on a 2×2 factorial design. There were two fixed factors: treatment group (standard and enriched); and time of recruitment (newly admitted dogs and long stay dogs). The QoL scores were logit transformed to achieve a normal distribution. Variances of transformed data were checked to ensure homogeneity across centres (Bartlett's K-squared$(12) = 5.9925$, $p = 0.9165$). Variance in QoL scores was examined using a linear mixed-effect model (lmer), fitted using the maximum likelihood procedure. Transformed QoL scores were fitted as the dependent variable; rehoming centre was fitted as a random factor; while the fixed factors fitted were time of recruitment and treatment, with an interaction between the two. The full lmer model was then compared to a null model (without the fixed factors), and an intermediate model (with the fixed factors but without an interaction), to assess goodness of fit, using the log likelihood ratio test. Sample size calculations were conducted a priori using Mead's (1988) (Eq. (1a)) resource equation. This was based on data collection from one rehoming centre (Eq. (1b)), i.e. not using a blocked design, as not all centres asked to take part in the study were guaranteed to do so at the time of experimental design. As 13 centres took part the resulting sample size was much larger than necessary, as would have been calculated if the 13 centres taking part had been guaranteed (Eqs. (2a) and (2b)) and a blocked design was used.

$$E = N - T \tag{1a}$$

where E is the error degrees of freedom, N is the total number of experimental units and T is the number of treatment combinations

$$E = 16 - 4 = 12 \tag{1b}$$

$$E = N - T - B \tag{2a}$$

where E is the error degrees of freedom, N is the total degrees of freedom (total number of experimental units minus one), T is the treatments degrees of freedom (number of treatment combinations minus one), and B is the blocks degrees of freedom (number of blocks minus one)

$$E = 207 - 3 - 13 = 192 \tag{2b}$$

The resource equation method was chosen over power analysis as it does not require an estimate of effect size or standard deviation (Festing et al., 2002), both of which were unavailable at the time of experimental design due to a lack of relevant information in current literature. Additionally, the QoL scores calculated in this study were also used to investigate different environmental and management factors that may affect QoL (described elsewhere), consequently increasing the complexity of the study and the number of hypotheses involved. One drawback of using this method is that it may be more appropriate for studies where relatively large effect sizes are likely. It is not known what effect size is clinically important in QoL studies, therefore important yet subtle differences in QoL may be hidden. However, the larger than necessary sample size resulting from more than one rehoming centre taking part in the study will allow smaller effect sizes to be estimated more accurately.

2.3.2. Inter-observer-reliability

Intra-observer reliability could not be assessed in this study for logistical reasons. Inter-observer reliability for each item in the QoL score was analysed using Randolph's online kappa calculator (Randolph, 2008). Testing binary components of a score allows unreliable items to be rejected from the final score. Inter-observer reliability is reported as a percentage agreement and a corresponding kappa value. The kappa coefficient is an index of inter-observer reliability that takes account of chance agreements (Martin and Bateson, 2007). Randolph's free-marginal multi-rater kappa was used (Randolph, 2005).

Good agreement is indicated by percentage agreements close to 100 and kappa values close to 1, whereas poor agreement is indicated by percentage agreements and kappa values close to 0 (Burn and Weir, 2011). The minimum thresholds of percentage agreement and kappa considered to be clinically useful by some authors is 75% and 0.4, respectively (Burn et al., 2009), whereas a cut-off point for kappa has also been stated as 0.7. As the assessors in this study were not necessarily assessing individual dogs at the same time, the minimum percentage agreement was lowered to 50% as the dogs were not expected to perform the same behaviours on every assessment visit and therefore agreement would be less likely to be perfect. The kappa was set at the lower recommendation of 0.4, therefore any item with a value of kappa below this was considered unreliable, regardless of the percentage agreement. Therefore,

any item with a percentage agreement less than 50% or a kappa value less than 0.4 was removed from the final QoL score.

Analysing inter-observer agreements also provided response evidence of validity by checking for observer drift. However, this would not detect multiple raters drifting simultaneously – consensual drift. Consensual drift was therefore reduced by preventing raters from seeing each other rating thus reducing the likelihood of copying or otherwise being influenced by each other.

2.3.3. Internal consistency

Agreement between "changing from one behaviour to another" and "changing type of movement", and their component parts, was assessed using the McNemar's test and the kappa coefficient.

2.3.4. Treatment group associations with individual QoL score items

It is possible that some of the QoL score items may have been biased towards dogs that had been in the centre for longer. The first section regarding anticipatory behaviours to the morning feed may only be applicable to dogs who have been in the centre for long enough to learn that they get fed at a specific time in the morning. Thus, it could be hypothesised that only the longer stay dogs would show anticipatory behaviours to the morning feed. Therefore, effects of treatment group were also investigated for the individual items using chi-square tests.

2.4. Ethical note

The procedures and data collection were conducted in accordance with the ethical guidelines laid down by the Royal Veterinary College; none required licensing by the UK Home Office.

3. Results

3.1. Inter-observer reliability

The inter-observer reliability of each QoL score item was calculated for each centre and then averaged across centres to produce average inter-observer reliability scores per item. There was good inter-observer reliability for all but five items: "Did the dog sit when unprovoked?" (67%, $\kappa = 0.335$); "Did the dog nose when the assessors approached on the fourth visit?" (56%, $\kappa = 0.150$); "Did the dog sit when the assessor approached on the fourth visit?" (63%, $\kappa = 0.252$); "Did the dog lick the assessor on the fourth visit?" (63%, $\kappa = 0.258$); and "Did the dog play with the assessor on the fourth visit of the assessment?" (46%, $\kappa = -0.087$). These items were removed from the score and therefore QoL scores were analysed from the refined score sheet, thus producing response evidence of validity. After removal f these five items the mean percentage agreement for all items was 87% and the mean kappa for all items was 0.733.

3.2. Internal consistency

There was low agreement between "changing from one behaviour to another" and "changing type of movement", and their component part items. "Changing from one behaviour to another" was not reliably recorded as present when more than one of play bowing, wall bouncing, chasing its tail, circling, walking, standing, sitting, startling, lying down, and changing type of movement were recorded (McNemar's $p < 0.001$, $\kappa = 0.201$, $N = 147$). "Changing type of movement" was not reliably recorded as present when more than one of chasing tail, circling, walking, sitting, standing, or wall bouncing were recorded (McNemar's $p < 0.001$, $\kappa = 0.176$, $N = 147$). These two items were therefore removed from the QoL score and the scores recalculated.

3.3. QoL score associations with treatment group

The intermediate lmer model ($X^2(2) = 11.533$, $p = 0.00313$) and the full fitted lmer model ($X^2(3) = 11.644$, $p = 0.008707$) were both significantly different to the null model, indicating that the addition of fixed factors and their interaction to the model produces a significantly better fit than the null model. However, as there was no significant difference between the intermediate and the full model ($X^2(1) = 0.1107$, $p = 0.7394$), indicating that the addition of the interaction term is adding very little information, the intermediate model was chosen as the best fit model. Calculation of the intraclass correlation coefficients indicate that only 26.7% (ICC1 = 0.267) of the variation in QoL scores is explained by centre group membership, but that centre groups can reliably be differentiated in terms of QoL scores (ICC2 = 0.832). The addition of an enrichment programme and later recruitment onto the study increased QoL scores by $0.035) \pm 0.027$ (SE) and 0.086 ± 0.027 (SE) respectively. The hypotheses that enriched dogs would have higher (better) QoL scores than dogs receiving a standard husbandry routine and that long stay dogs would have higher (better) QoL due to habituation to the kennel environment than newly admitted dogs was supported by the findings: enriched, long stay dogs had a mean QoL score of 0.477; long stay dogs with a standard routine had a mean QoL score of 0.453; newly admitted dogs with an enriched routine had a mean QoL score of 0.399; and newly admitted dogs with a standard routine had a mean QoL score of 0.362 (Fig. 1).

3.4. Treatment group associations with individual QoL score items

The only items that were significantly associated with treatment group were tail wagging in response to a kennelmate ($\chi^2(3) = 8.496$, $p = 0.037$) and drinking excessively ($\chi^2 = 5.042$, $p = 0.043$). The presence of tail wagging was associated with newly admitted dogs, regardless of treatment (Table 2). Drinking excessively was associated with long stay dogs that received a standard husbandry routine (Table 2).
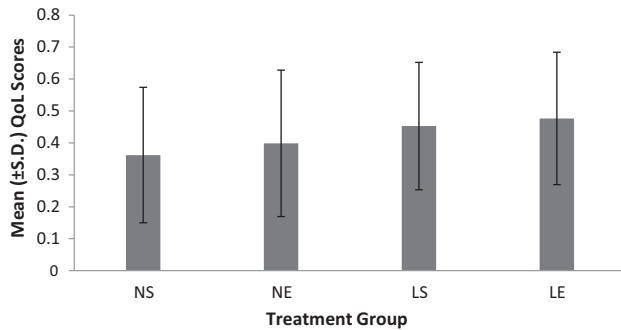
Fig. 1. Mean quality of life scores of kennelled dogs assigned to four different treatment groups. NS – newly admitted dogs with a standard husbandry routine; NE – newly admitted dogs with an enriched routine; LS – long stay dogs with a standard routine; LE – long stay dogs with an enriched routine.

Table 2
Observed and expected frequencies of two behaviours (tail wagging and drinking excessively) within four treatment groups of kennelled dogs.

		NS	NE	LS	LE
No tail wagging	Count	24	16	25	26
	Expected count	*25.7*	*22.8*	*22.8*	*19.8*
Tail wagging	Count	28	30	21	14
	Expected count	*26.3*	*23.3*	*23.3*	*20.2*
Did not drink excessively	Count	53	49	49	49
	Expected count	*52.2*	*48.3*	*51.2*	*48.3*
Drank excessively	Count	0	0	3	0
	Expected count	*0.8*	*0.7*	*0.8*	*0.7*

NS, newly admitted dogs with a standard husbandry routine; NE, newly admitted dogs with an enriched routine; LS, long stay dogs with a standard routine; LE, long stay dogs with an enriched routine.

4. Discussion

This study aimed to develop and validate a scoring system to assess the quality of life of kennelled dogs at a moment in time, which could be used to track QoL within an individual over an extended period. The reliability and validity of this scoring system was tested through its use in Dogs Trust kennels by Dogs Trust staff. This was achieved by analysing inter-observer reliability of the score's individual items and establishing test validation by way of content, construct and criterion evidence.

Five items in the score sheet had low inter-observer reliability. Several factors can affect how well a category of behaviour, or an item on the score sheet, is measured, for example, experience, frequency of occurrence, observer fatigue, observer drift, and accuracy of definition. The more experienced the observer is at using the measurement technique the better their recordings will be. If a behaviour occurs very rapidly then it may be difficult to record each occurrence reliably. Similarly, rare behaviours may be missed completely if observation sessions are too short. Observer fatigue and loss of concentration may occur if the observation session is too long or if the staff are trying to multi-task, resulting in a reduced ability to record accurately. Each category of behaviour, or other measurement, needs to be defined clearly in order to be recorded reliably. However, observer drift may still occur in long studies where the observer becomes familiar with the behaviour

and unconsciously 'improves' or refines the definitions (Martin and Bateson, 2007).

The majority of these explanations do not seem to apply to one of the unreliable items found in this score – "Did the dog play with the assessor on the fourth visit?" Frequency of occurrence seems an unlikely explanation, as at this point the observer was looking specifically for play signals from the dog and therefore unlikely to miss them. Additionally, the definition of play did not include behaviours that are generally performed so rapidly as to be easily missed. Similarly these behaviours were unlikely to be missed due to short observation sessions as they were provoked by the observer. Observers were unlikely to lose concentration due to observer fatigue as the sessions were very short; the relevant session was only 4 min long. A clear ethogram was provided for each assessor to refer to as they were assessing each dog, thus a clear definition of play with observer was provided. However, observer experience or observer drift may have influenced the reliability of this item. The majority of the observers taking part in this study will have been experienced dog handlers: although Dogs Trust centres often have volunteers helping who might be less familiar with dog body language the centre managers were not keen to use volunteers for this study. Additionally, the clear definition provided in the ethogram should have allowed even people with very little experience of dogs to recognise the stated behaviours. A likely explanation is that some of the staff were simply better at eliciting play behaviour, or were associated with play based on past experiences. However, observer drift should also be considered. Because the staff are experienced with dogs they might have preconceived ideas of what "play" is and applied these definitions instead of the one provided. One way to combat this is to periodically check inter-observer reliability scores (Suen and Rzasa, 2004). However, this would only point to definitions changing over time, rather than personal definitions being used from the start of the study.

The item "Did the dog sit when unprovoked?" (i.e. when watched from a distance) may have had low reliability due to being missed by one of the two observers, either because the behaviour was rare or performed briefly. The remaining items that were found to be unreliable – "Did the dog nose when the assessor approached on the fourth visit?"; "Did the dog sit when the assessor approached on the fourth visit?"; and "Did the dog lick the assessor on the fourth visit?" – may also have been missed by an assessor because they were performed very rapidly or rarely. However, these behaviours were expected to be performed in reaction to the approaching assessor. Something about each assessor's demeanour may have provoked different reactions in the dogs they were assessing, such as perceived friendly or threatening body language or because of the dog's past associations with the appearance of the assessor (e.g. gender, height, weight, facial hair, spectacles). Thus different reactions to different assessors would give the impression of low reliability but it may reflect the different behaviours provoked by their approach. Because only one assessor was asked to approach a dog at a time and this was not recorded by video there was no way to test this post hoc.

There was low agreement between the items "changing from one behaviour to another" and "changing type of

movement" and their component part items. These items were originally included in the assessment tool in order to assess internal consistency – agreement between items intended to measure the same construct. Although the lack of agreement between them and their component parts suggests a lack of internal consistency it may reflect that their inclusion made the tool too burdensome. Not only did the assessor have to go back over the tool at the end of each assessment session in order to ensure they had noted all occurrences of the item behaviours, they would also have had to mentally calculate whether the dog had changed behaviour or movement. This required more attention to detail than simply noting the occurrence of a behaviour and may explain the lack of agreement. Additionally, only mutually exclusive behaviours were considered in the analysis but the observers may have been using non-mutually exclusive behaviours when recording these two terms, which would lead to a discrepancy between their records and the formal analysis.

After calculating the overall QoL scores based on the 51 reliable items, following removal of the seven unreliable items, fitting the lmer model indicated that giving dogs an additional programme of enrichment improved QoL. Dogs that had been in their respective centres for 30 days or more also had better QoL. These results are in agreement with previous studies, which have suggested that dogs adapt to the kennel environment over time (e.g. Hennessy et al., 1997; Rooney et al., 2007; Stephen and Ledger, 2006) and that environmental enrichment helps animals to cope with their environments (e.g. Graham et al., 2005; Hetts et al., 1992; Hubrecht, 1993; Schipper et al., 2008; Valsecchi et al., 2007; Wells and Hepper, 1992). This result supports the use of the QoL score as a tool for assessing quality of life as it differentiated, as hypothesised, between four groups of dogs that were manipulated to have different levels of QoL. However, some of the variation in QoL scores was also attributed to centre group membership. The effect of centre level factors on these QoL scores is discussed elsewhere (Kiddie and Collins, in preparation).

Some of the individual QoL score items were found to be associated with treatment group. The presence of tail wagging in response to a kennelmate was associated with newly admitted dogs, regardless of treatment. Dogs wag their tails for many reasons: motivations include play, aggression, and appeasement. Without considering the quality of the movement of the tail and other body signals it is difficult to infer what the underlying emotional state is (Paz and Escobedo, 2011) and can therefore only indicate the arousal of the dog. As there was no difference in levels of play or aggression it may be that newly admitted dogs experienced more arousal as a result of uncertainty of emotional conflict (Bradshaw, 2011) in the presence of unfamiliar dogs and surroundings than in the presence of kennelmates and surroundings that they have had time to become familiar with and could therefore be a sign of stress (Beerda et al., 1999). Drinking excessively was associated with long stay dogs that received a standard husbandry routine. Previous studies have suggested that polydipsia is also a sign of stress (Stephen and Ledger, 2005) and may indicate that dogs that have been in kennels for a long time without additional stimulation or enrichment may become frustrated. Although both of these behaviours can be associated with stress, they are associated with different groups of dogs in this study. This suggests that these behaviours reflect different types of stress.

Anticipatory behaviours towards the morning feed might only have applied to those dogs that had been in the centre long enough to learn the predictive cues of the morning feed. However, there was no significant difference of anticipatory behaviours to the morning feed between newly admitted dogs and longer stay dogs. This suggests that there is no bias towards longer stay dogs as a result of learning. It may be that the majority of dogs are fed in the morning in a home setting, and therefore expect this on entering a kennel environment as well.

This study aimed to develop a validated scoring system to assess the QoL of kennelled dogs. Validation was achieved through establishing good inter-observer reliability and providing evidence of content, response and convergent validity. Inter-observer reliability of the score's individual items was established, allowing the scoring system to be refined through removal of one unreliable item. After another two items were removed because of general underscoring, the refined scoring system was investigated for validation.

Content evidence evaluates whether the content of the instrument accurately represents the concept it was designed to measure, in this case, QoL. Providing content evidence of this score involved the generation of a list of items from a literature search of QoL. Content evidence would have been strengthened by a peer review of the generated list using a larger expert panel. Although there were time restraints preventing this, sections of the list had been peer reviewed separately elsewhere: the behaviours taken from Stephen and Ledger (2005) had been through this process as had the physical presence of scaling, eye discharge, and body condition score (Kiddie and Collins, in preparation).

The descriptions of assessor training provide construct validation in the form of response evidence (Suen and Rzasa, 2004). At least two assessors from each centre were given a presentation explaining in more detail the research procedure. Assessors were also provided with a letter detailing the research procedure and an ethogram to refer to when assessing the dogs. Although it is preferable that all assessors attend training sessions to ensure that they all receive the same information, thus preventing multiple methods of scoring (Suen and Rzasa, 2004), this was not possible as some centres could not spare the staff for the training sessions. Therefore, the members of staff that took part in the study but missed the training session had to rely on the written instructions and information relayed from members of staff that did attend the training session. Providing the detailed ethogram will also have reduced observer bias. Observer drift did not appear to occur in the refined tool, as evidenced by the good inter-observer reliability, although it may have occurred for some of the less reliable items. Consensual drift was reduced by asking assessors not to confer with each other while assessing each dog thus reducing the likelihood of copying or otherwise being influenced by each other.

Concurrent criterion evidence was established as the tool demonstrated interaction between the QoL scores and the classification variables, i.e. treatment group and recruitment time (Canadian Psychological Association, 1996). There may be some controversy over whether it is possible to provide evidence of criterion validity for this tool as Wojciechowska et al. (2005) state that there is no established test or gold standard for QoL assessment in dogs.

5. Conclusion

This is the first published and peer-reviewed QoL scoring system that we know of for use in kennelled dogs in rehoming kennels. In addition to assessing and monitoring quality of life over time, this tool can also be used to assess the impact of housing and husbandry changes as well as veterinary interventions on the quality of life of kennelled dogs. The use of this tool to identify centre-level factors associated with QoL scores is described elsewhere (Kiddie and Collins, in preparation). The results can thus guide kennel staff in future decision-making processes, either on a day-to-day basis for individuals, or at centre or organisation level.

Authors' contributions

JLK conceived and designed the study, co-ordinated collection of the data, performed the data analysis and drafted the manuscript. LMC was principal supervisor on the project, and advised on the study design, data analysis, interpretation of results and assisted with preparation of the final manuscript. Both authors read and approved the final manuscript.

Acknowledgments

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.applanim.2014.05.008.

References

Auquier, P., Simeoni, M.C., Sapin, C., Reine, G., Aghababian, V., Cramer, J., Lançon, C., 2003. Development and validation of a patient-based health-related quality of life questionnaire in schizophrenia: the S-QoL. Schizophr. Res. 63, 137–149.

Bateson, M., Matheson, S.M., 2007. Performance on a categorisation task suggests that removal of environmental enrichment induces 'pessimism' in captive European starlings (*Sturnus vulgaris*). Anim. Welf. 16, 33–36.

Bauer, E., Ward, C., Smuts, B., 2009. Play like a puppy, play like a dog. J. Vet. Behav. 4, 68–69.

Beerda, B., Schilder, M.B.H., Van Hooff, J., De Vries, H.W., Mol, J.A., 1999. Chronic stress in dogs subjected to social and spatial restriction. I. Behavioral responses. Physiol. Behav. 66, 233–242.

Beerda, B., Schilder, M.B.H., van Hooff, J.A.R.A.M., de Vries, H.W., Mol, J.A., 1998. Behavioural, saliva cortisol and heart rate responses to different types of stimuli in dogs. Appl. Anim. Behav. Sci. 58, 365–381.

Beerda, B., Schilder, M.B.H., van Hooff, J.A.R.A.M., de Vries, H.W., Mol, J.A., 2000. Behavioural and hormonal indicators of enduring environmental stress in dogs. Anim. Welf. 9, 49–62.

Bleicher, N., 1963. Physical and behavioural analysis of dog vocalisations. Am. J. Vet. Res. 24, 415–427.

Boissy, A., Arnould, C., Chaillou, E., sir, L., Duvaux-Ponter, C., Greiveldinger, L., Leterrier, C., Richard, S., Roussel, S., Saint-Dizier, H., Meunier, S., n, M.C., Valance, D., Veissier, I., 2007a. Emotions and cognition: a new approach to animal welfare. Anim. Welf. 16, 37–43.

Boissy, A., Manteuffel, G., Jensen, M.B., Moe, R.O., Spruijt, B., Keeling, L.J., Winckler, C., Forkman, B., Dimitrov, I., Langbein, J., Bakken, M., Veissier, I., Aubert, A., 2007b. Assessment of positive emotions in animals to improve their welfare. Physiol. Behav. 92, 375–397.

Boyer, L., Simeoni, M.-C., Loundou, A., D'Amato, T., Reine, G., Lancon, C., Auquier, P., 2010. The development of the S-QoL 18: a shortened quality of life questionnaire for patients with schizophrenia. Schizophr. Res. 121, 241–250.

Bradshaw, J.W.S., 2011. Dog Sense: How the New Science of Dog Behavior Can Make You a Better Friend to Your Pet. Basic Books, New York.

Brinkmann, L., Gerken, M., Riek, A., 2013. Effect of long-term feed restriction on the health status and welfare of a robust horse breed, the Shetland pony (*Equus ferus caballus*). Res. Vet. Sci. 94 (3), 826–831.

Burn, C.C., Pritchard, J.C., Whay, H.R., 2009. Observer reliability for working equine welfare assessment: problems with high prevalences of certain results. Anim. Welf. 18, 177–187.

Burn, C.C., Weir, A.A.S., 2011. Using prevalence indices to aid interpretation and comparison of agreement ratings between two or more observers. Vet. J. 188, 166–170.

Buske-Kirschbaum, A., Ebrecht, M., Kern, S., Hellhammer, D.H., 2006. Endocrine stress responses in TH1-mediated chronic inflammatory skin disease (*Psoriasis vulgaris*) – do they parallel stress-induced endocrine changes in TH2-mediated inflammatory dermatoses (atopic dermatitis)? Psychoneuroendocrinology 31, 439–446.

Canadian Psychological Association, 1996. Guidelines for Educational and Psychological Testing. Canadian Psychological Association, Ottawa.

Carmines, E.G., Zeller, R.A., 1979. Reliability and Validity Assessment. Sage Publications, Inc., pp. 11.

Ellegaard, L., Cunningham, A., Edwards, S., Grand, N., Nevalainen, T., Prescott, M., Schuurman, T., 2010. Welfare of the minipig with special reference to use in regulatory toxicology studies. J. Pharmacol. Toxicol. Methods 62, 167–183.

FAWC, 2009. Farm Animal Welfare in Great Britain: Past, Present and Future. Farm Animal Welfare Council.

Festing, M.F.W., Overend, P., Gaines Das, R., Cotina Borja, M., Berdoy, M., 2002. The Design of Animal Experiments: Reducing the Use of Animals in Research Through Better Experimental Design. The Royal Society of Medicine Press Limited, London.

Fitzsimmons, D., Johnson, C.D., George, S., Payne, S., Sandberg, A.A., Bassi, C., Beger, H.G., Birk, D., Büchler, M.W., Dervenis, C., Fernandez Cruz, L., Friess, H., Grahm, A.L., Jeekel, J., Laugier, R., Meyer, D., Singer, M.W., Tihanyi, T., 1999. Development of a disease specific quality of life (QoL) questionnaire module to supplement the EORTC core cancer QoL questionnaire, the QLQ-C30 in patients with pancreatic cancer. Eur. J. Cancer 35, 939–941.

Gómez-Gallego, M., Gómez-Amor, J., Gómez-García, J., 2012. Validation of the Spanish version of the QoL-AD scale in Alzheimer disease patients, their carers, and health professionals. Neurología (English Edition) 27, 4–10.

Graham, L., Wells, D.L., Hepper, P.G., 2005. The influence of olfactory stimulation on the behaviour of dogs housed in a rescue shelter. Appl. Anim. Behav. Sci. 91 (1–2), 143–153.

Groth-Marnat, G., 2009. Handbook of Psychological Assessment. John Wiley & Sons, Inc., Hoboken, NJ, pp. 17.

Hendrix, S., Peters, E.M.J., 2007. Neuronal plasticity and neuroregeneration in the skin – the role of inflammation. J. Neuroimmunol. 184, 113–126.

Hennessy, M.B., Davis, H.N., Williams, M.T., Mellott, C., Douglas, C.W., 1997. Plasma cortisol levels of dogs at a county animal shelter. Physiol. Behav. 62 (3), 485–490.

Hennessy, M.B.T., Williams, M., Miller, D.D., Douglas, C.W., Voith, V.L., 1998. Influence of male and female petters on plasma cortisol and

behaviour: can human interaction reduce the stress of dogs in a public animal shelter? Appl. Anim. Behav. Sci. 61, 63–77.

Hetts, S., Derrell Clark, J., Calpin, J.P., Arnold, C.E., Mateo, J.M., 1992. Influence of housing conditions on beagle behaviour. Appl. Anim Behav. Sci. 34 (1–2), 137–155.

Hewson, C.J., Hiby, E.F., Bradshaw, J.W.S., 2007. Assessing quality of life in companion and kennelled dogs: a critical review. Anim. Welf. 16, 89–95.

Hiby, E.F., Rooney, N.J., Bradshaw, J.W.S., 2006. Behavioural and physiological responses of dogs entering re-homing kennels. Physiol. Behav. 89, 385–391.

Hubrecht, R.C., 1993. A comparison of social and environmental enrichment methods for laboratory housed dogs. Appl. Anim. Behav. Sci. 37 (4), 345–361.

Kessler, M.R., Turner, D.C., 1997. Stress and adaptation of cats (*Felis Silvestris Catus*) housed singly, in pairs and in groups in boarding catteries. Anim. Welf. 6, 243–254.

Kiddie, J., (MSc) 2004. The Interpretation of the Intra-specific Signalling of the Domestic Dog (*Canis familiaris*) by Owners. University of Edinburgh, Edinburgh.

Kuo, Y.-C., Lan, C.-F., Chen, L.-K., Lan, V.M., 2010. Dementia care costs and the patient's quality of life (QoL) in Taiwan: home versus institutional care services. Arch. Gerontol. Geriatr. 51, 159–163.

León-Salas, B., Olazarán, J., Muñiz, R., González-Salvador, M.T., Martínez-Martín, P., 2011. Caregivers' estimation of patients' quality of life (QoL) in Alzheimer's disease (AD): an approach using the ADRQL. Arch. Gerontol. Geriatr. 53, 13–18.

Litwin, M.S., 1995. How to Measure Survey Reliability and Validity. Sage Publications, Inc., Thousand Oaks, California, pp. 8–30.

Maillé, A.R., Koning, C.J.M., Zwinderman, A.H., Willems, L.N.A., Dijkman, J.H., Kaptein, A.A., 1997. The development of the 'Quality-of-Life for Respiratory Illness Questionnaire (QOL-RIQ)': a disease-specific quality-of-life questionnaire for patients with mild to moderate chronic non-specific lung disease. Respir. Med. 91, 297–309.

Main, D.C.J., Clegg, J., Spatz, A., Green, L.E., 2000. Repeatability of a lameness scoring system for finishing pigs. Vet. Rec. 147, 574–576.

Martin, P., Bateson, P., 2007. Measuring Behaviour: An Introductory Guide. Cambridge University Press, pp. 114, 116, 120 and 122.

Mead, R., 1988. The Design of Experiments. Cambridge University Press, Cambridge, New York.

Morgan, G.A., Gliner, J.A., Harmon, R.J., 2006. Understanding and Evaluating Research in Applied Clinical Settings. Lawrence Erlbaum Associates, Inc., Mahwah, NJ, pp. 52.

Nestlé Purina Petcare Centre, 2014. http://www.purina.co.uk/content/docs/downloadable-pdfs/bodycondsytempdf.pdf (accessed 18.02.14).

Patrick, D.L., Martin, M.L., Bushnell, D.M., Yalcin, I., Wagner, T.H., Buesching, D.P., 1999. Quality of life of women with urinary incontinence: further development of the incontinence quality of life instrument (I-QOL). Urology 53, 71–76.

Paz, B., Escobedo, R., 2011. Happy tail wagging: a laboratory artifact? Lateral tail wagging in the field. J. Vet. Behav. 6, 94–95.

Pérez-Campos, E., Dueñas, J.L., de la Viuda, E., Gómez, M.Á., Lertxundi, R., Sánchez-Borrego, R., Canals, I., Bermejo, R., Arbat, A., Badia, X., Perulero, N., Lete, L.I., 2011. Development and validation of the SEC-QOL Questionnaire in women using contraceptive methods. Value Health 14, 892–899.

Petrie, A., Watson, P.F., 2006. Statistics for Veterinary and Animal Science. Blackwell Publishing Ltd., Ames, Iowa, pp. 366.

Pollari, F.L., Bonnett, B.N., Allen, D.G., Bamsey, S.C., Martin, S.W., 1996. Quality of computerized medical record abstract data at a veterinary teaching hospital. Prev. Vet. Med. 27, 141–154.

Rajmil, L., Abad, S., Sardon, O., Morera, G., Pérez-Yarza, E.G., Moreno, A., Detmar, S., Fekkes, M., Herdman, M., Alonso, J., 2011. Reliability and validity of the Spanish version of the TAPQOL: a health-related quality of life (HRQOL) instrument for 1- to 5-year-old children. Int. J. Nurs. Stud. 48, 549–556.

Randolph, J.J., 2005. Free-marginal multirater kappa (multirater κfree): an alternative to Fleiss' fixed-marginal multirater kappa. In: Joensuu Learning and Instruction Symposium, Joensuu, Finland.

Randolph, J.J., 2008. Online kappa calculator.

Rautava, L., Häkkinen, U., Korvenranta, E., Andersson, S., Gissler, M., Hallman, M., Korvenranta, H., Leipälä, J., Linna, M., Peltola, M., Tammela, O., Lehtonen, L., 2009. Health-related quality of life in 5-year-old very low birth weight infants. J. Pediatr. 155, pp. 338–343.e1–e3.

Rehn, T., 2011. The Role of the Emotional Relationship with Humans on Dog Welfare. Swedish University of Agricultural Sciences, Uppsala.

Rooney, N.J., Gaines, S.A., Bradshaw, J.W.S., 2007. Behavioural and glucocorticoid responses of dogs (*Canis familiaris*) to kennelling: investigating mitigation of stress by prior habituation. Physiol. Behav. 92 (5), 847–854.

Schipper, L.L., Vinke, C.M., Schilder, M.B.H., Spruijt, B.M., 2008. The effect of feeding enrichment toys on the behaviour of kennelled dogs (*Canis familiaris*). Appl. Anim Behav. Sci. 114 (1–2), 182–195.

Scott, E.M., Nolan, A.M., Reid, J., Wiseman-Orr, M.L., 2007. Can we really measure animal quality of life? Methodologies for measuring quality of life in people and other animals. Anim. Welf. 16, 17–24.

Stephen, J.M., Ledger, R.A., 2005. An audit of behavioural indicators of poor welfare in kennelled dogs in the United Kingdom. J. Appl. Anim. Welf. Sci. 8, 79–95.

Stephen, J.M., Ledger, R.A., 2006. A longitudinal evaluation of urinary cortisol in kennelled dogs, *Canis familiaris*. Physiol. Behav. 87 (5), 911–916.

Suen, H.K., Rzasa, S.E., 2004. Psychometric foundations of behavioral assessment. In: Hersen, M. (Ed.), Comprehensive Handbook of Psychological Assessment: Behavioral Assessment. John Wiley & Sons, Inc., Hoboken, NJ (Chapter 3).

Suzukamo, Y., Noguchi, H., Takahashi, N., Shimatsu, A., Chihara, K., Green, J., Fukuhara, S., 2006. Validation of the Japanese version of the Quality of Life-Assessment of Growth Hormone Deficiency in Adults (QoL-AGHDA). Growth Horm. IGF Res. 16, 340–347.

Taylor, K.D., Mills, D.S., 2006. The development and assessment of temperament tests for adult companion dogs. J. Vet. Behav. 1, 94–108.

Taylor, K.D., Mills, D.S., 2007a. The effect of the kennel environment on canine welfare: a critical review of experimental studies. Anim. Welf. 16, 435–447.

Taylor, K.D., Mills, D.S., 2007b. Is quality of life a useful concept for companion animals? Anim. Welf. 16, 55–65.

Temple, D., Dalmau, A., Ruiz de la Torre, J.L., Manteca, X., Velarde, A., 2011. Application of the Welfare Quality® protocol to assess growing pigs kept under intensive conditions in Spain. J. Vet. Behav. 6, 138–149.

Timmins, R.P., Cliff, K.D., Day, C.T., Hart, B.L., Hart, L.A., Hubrecht, R.C., Hurley, K.F., Phillips, C.J.C., Rand, J.S., Rochlitz, I., Serpell, J.A., Zawistowski, S.L., 2007. Enhancing quality of life for dogs and cats in confined situations. Anim. Welf. 16, 83–87.

Valsecchi, P., Pattacini, O., Beretta, V., Bertozzi, J., Zannoni, S., Viggiani, R., Accorsi, P.A., 2007. 13 effects of a human social enrichment program on behavior and welfare of sheltered dogs. J. Vet. Behav. 2, 88–89.

Van der Harst, J.E., Spruijt, B.M., 2007. Tools to measure and improve animal welfare: reward-related behaviour. Anim. Welf. 16, 67–73.

Varnell, E.D., Kaufman, H.E., Hill, J.M., Thompson, H.W., 1995. Cold stress-induced recurrences of herpetic keratitis in the squirrel monkey. Invest. Ophthalmol. Vis. Sci. 36, 1181–1183.

Walker, J.K., Dale, A.R., D'Eath, R.B., Wemelsfelder, F., 2009. Welfare assessment in dogs: reducing the invasiveness of welfare assessment methodology through trial and validation of a qualitative, behavioral-based approach. J. Vet. Behav. 4, 102–104.

Wells, D., Hepper, P.G., 1992. The behaviour of dogs in a rescue shelter. Anim. Welf. 1, 171–186.

Wemelsfelder, F., Nevison, I., Lawrence, A.B., 2009. The effect of perceived environmental background on qualitative assessments of pig behaviour. Anim. Behav. 78, 477–484.

Whay, H.R., Main, D.C.J., Greent, L.E., Webster, A.J.F., 2003. Animal-based measures for the assessment of welfare state of dairy cattle, pigs and laying hens: consensus of expert opinion. Anim. Welf. 12, 205–217.

Whitham, J.C., Wielebnowski, N., 2009. Animal-based welfare monitoring: using keeper ratings as an assessment tool. Zoo Biol. 28, 545–560.

Wiseman-Orr, M.L., Scott, E.M., Nolan, A.M., 2011. Development and testing of a novel instrument to measure health-related quality of life (HRQL) of farmed pigs and promote welfare enhancement (Part 2). Anim. Welf. 20, 549–558.

Wiseman-Orr, M.L., Scott, E.M., Reid, J., Nolan, A.M., 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. Am. J. Vet. Res. 67, 1826–1836.

Wojciechowska, J.I., Hewson, C.J., Stryhn, H., Guy, N.C., Patronek, G.J., Timmons, V., 2005. Development of a discriminative questionnaire to assess nonphysical aspects of quality of life of dogs. Am. J. Vet. Res. 66, 1453–1460.

Wu, W., Ariyasu, R.G., 1999. Cornea and external disease. In: Lee, D.A., Higginbotham, E.J. (Eds.), Clinical Guide to Comprehensive Ophthalmology. Thieme Medical Publishers Inc., New York, p. 199.