The Measurement of Pain in Dogs

Lynne Louise Holton

Thesis submitted for the degree of Ph.D.

University of Glasgow

Department of Pre-Clinical Studies, Faculty of Veterinary Medicine

March 2000

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

Abstract

The ability to measure pain is a key issue in veterinary medicine for two reasons. Firstly, adequate pain management can only be provided if the animal's pain can be recorded accurately. Secondly, pain research demands a reliable and valid measurement method if the mechanisms of pain and analgesics are to be explored scientifically. The development of pain measurement scales used in veterinary medicine has followed a similar path to that seen in human medicine. However, this is limited by the lack of effective communication between the patient and care provider. The work undertaken in this thesis is aimed at exploring the pain measurement scales commonly used in veterinary medicine and developing a novel composite measurement pain scale specifically designed for use in dogs, in a clinical setting.

The consistency of the visual analogue (VAS), numerical rating (NRS) and simple descriptive (SDS) scales when used by a number of observers and over time, and the relationships between the scales were explored. The results indicated that the VAS and NRS were not adequately generalizable when used by more than one observer (generalizability coefficients between 0.27 and 0.53). The generalizability over a long period, i.e. from the day of surgery to the following day, was also low (generalizability between 0.42 and 0.45), however, the generalizability within a relatively short time period was reasonable (generalizability coefficient between 0.69 to 0.73). When using the SDS the agreement between the observers was not acceptable (Kappa statistics between 0.23 and 0.37). Thus, pain measurements made using the VAS, NRS and SDS were not consistent when used by more then one observer or over time. Investigation of the correspondence between the VAS and NRS demonstrated that a strong relationship existed, but this was dependent on the observer. The relationship of the VAS and NRS to the SDS was shown to be consistent across observers, although each category of the SDS corresponded to a wide range of NRS and VAS scores. Thus, when used in a clinical setting, these scales should not be used interchangeably since there is no unique relationship between them.

Pain measurement in human medicine has progressed from the simple subjective rating scales to composite measurement scales such as the McGill Pain Questionnaire (MPQ). Since the subjective rating scales used in veterinary medicine have shown inadequate generalizability when used in a clinical setting, a composite measurement pain scale (CMPS) was developed. The construction of this scale followed the methods used in the development of the MPQ. A bank of behaviours and physiological signs were gathered

from a group of practicing veterinary surgeons. This information was then rationalised and categorised to form a list of 47 behaviours and signs that were allocated into 8 behaviour categories and one category of physiological signs. Information on the pain intensity associated with each item was collected by consulting another group of veterinary surgeons. Each person assigned a pain intensity score to each item using a VAS (100mm). This information allowed the relationships between the items to be explored and the results of this investigation were reported to a focus group. This group of experts in animal pain refined the items and defined an examination procedure, which constituted the CMPS. Before the scale could be further explored, the items included were defined by consulting with a panel of 16 veterinary surgeons with specific interest in the measurement of pain. Weights were then assigned to each item using Thurstone's paired comparisons model, which provides interval level measurement.

The performance of the CMPS was assessed by carrying out two studies. In the first study 5 veterinary surgeons examined 4 groups of 20 animals each. The group to which an animal belonged was defined by the reason for hospitalisation (orthopaedic surgery, soft tissue surgery and medical cases) a fourth group of clinically sound animals was also included, this comprised the control group. The pain measurement scores collected were used to investigate the validity and reliability of the CMPS. The validity of the scale was supported as there were significant differences in the pain scores assigned to animals that had and had not undergone surgery (median scores of 2.4 and 0.9 respectively, p-value<0.01). Significant differences were also seen between the study groups (medians of 3.0, 2.0, 1.2, and 0.9 for orthopaedic, soft tissue, medical and control groups respectively, p-value<0.01). A positive relationship between pain score and perceived severity of pain associated with the animal's condition was observed, although this was not shown to be significant (p-value>0.3). The reliability of the CMPS over the observers was low (reliability coefficient between 0.33 and 0.51), although this improved when the coefficients were adjusted to account for multiple observers (reliability coefficient between 0.56 and 0.78). These results indicated that there was large variability between the observers when using the CMPS.

The second study involved 4 veterinary surgeons who had not previously used the CMPS. Each person watched a vide recording of examinations carried out on 12 dogs and assessed the pain each animal was experiencing, using the CMPS and a NRS. This exercise was repeated between two and four weeks later and the generalizability of the pain scores over observers and time was examined. The generalizability of the CMPS scores over the

observers was improved compared to the previous study and was comparable to the NRS (generalizability coefficients of 0.61 and 0.66 for the CMPS and NRS respectively). The generalizability over time for the NRS was slightly better than the CMPS (generalizability coefficients of 0.52 and 0.68 for CMPS and NRS respectively). Following the study, discussion with the participants highlighted a number of issues in the use of the scale, such as the provision of training and the use of the item definitions. These indicated that the CMPS required further exploration and development to allow it to realise its full potential.

Acknowledgements

I would like to thank the Clinical Studies Trust Fund (now Petsavers) for providing funding for the initial part of work. Thanks also go to Dr Anne Lennon of AstraZeneca Pharmaceuticals for the support that allowed the work to continue, and to PharmaPart UK for the final push.

I would like to give my utmost thanks to Prof. Marian Scott for all the supervision, encouragement and support she has given over the years and to Prof. Andrea Nolan and Prof. Jacky Reid for their invaluable input and expert guidance.

Much of the work presented here would not have been possible without the hard work of a large number of people from Glasgow University Veterinary School. Firstly I would like to thank all of the theatre and nursing staff of the veterinary hospital who were involved in the work and undoubtedly made life easier. I am also grateful to Pat Pawson for her work on the collection and collation of the original bank of words that formed the foundation of the scale and to Derek Flaherty, Dr Elizabeth Welsh, Prof. Andrea Nolan and Prof. Jacky Reid for their involvement in the focus group. The people (post-graduate students at the time) who gave up their precious time to help with my studies also deserve my thanks; they are Hafid Benchoui, Ian Scott, Pat Pawson, David Argyle, Susan Grant, Liz Norman, Jill Price, Martin Owen and Thierry Beths. For the video study, I am grateful to Alan Reid for his skills behind the camera and to Colin Brierley of the media services department for editing the video footage. Thanks also go to the 16 veterinary surgeons who responded to my plea for help with the item definitions and scaling model via the American College of Veterinary Anaesthetists email list.

The route I have taken to get to this stage has been, without doubt, the long way round, and many people deserve my thanks for helping me get here. Thanks go to my Mum, Dad and Allan for their encouragement and enthusiasm and to my Aunt Rene for making me promise to finish. Thanks also go to my friends for listening and to my colleagues in the division of Veterinary Pharmacology.

Lastly, thanks to Stuart for his expert proof reading, for understanding and for believing in me.

Declaration

I declare that the work presented in this thesis is my own unless otherwise stated and acknowledged.

Lynne Holton

Articles accepted for publication taken from the work presented in this thesis are:

REID, J. HOLTON, L., NOLAN, A.M., WELSH, E., PAWSON, P. (1997) The development of a multidimensional scale to assess pain in dogs. In Proceedings of the 6th International Congress of Veterinary Anaesthesiology, Halkidiki, Greece.

HOLTON, L.L., SCOTT, E.M., NOLAN, A.M., REID, J., WELSH, E., PAWSON, P. (1997) The development of a composite measure scale to assess pain in dogs. In Proceeding of the 7th European Association for Veterinary Pharmacology and Toxicology, Madrid, Spain.

HOLTON, L.L., SCOTT, E.M., NOLAN, A.M., REID, J., & WELSH, E.M. (1998) Investigation of the relationship between physiological factors and clinical pain in dogs scored using a numerical rating scale. *Journal of Small Animal Practice* **39**, 469-474.

HOLTON, L.L., SCOTT, E.M., NOLAN, A.M., REID, J., WELSH, E.M., & FLAHERTY, D. (1998) Comparison of three methods used for assessment of pain in dogs. *Journal of the American Veterinary Medical Association* **212** (1), 61-66.

A further article submitted for publication taken from the work presented in this thesis is:

HOLTON, L.L., REID, J., SCOTT, E.M., PAWSON, P., NOLAN, A.M. (1999) The development of a behavioural based pain scale to measure acute pain in dogs

To Stuart, to my family and to Rene

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. General Introduction

The importance of pain in medicine, both human and veterinary, is undisputed. Indeed the control of pain is fundamental to patient welfare and recovery (McGrath and Hillier, 1989; Bonica 1992). To ensure that pain can be controlled, it is crucial that each patient's pain can be quantified and recorded accurately. Over the last 50 years, the importance of measuring pain has been realised and has led to a large body of research in human medicine. The measurement of pain in veterinary medicine has not been studied as extensively. In the main, the methodology used in veterinary medicine has been based on that developed in human medicine although there are clear differences between the two areas, including communication between the patients and carers (Chapman *et al*, 1985).

1.1 Definition of pain and issues in pain measurement

The physiological mechanisms involved in the sensory phenomenon of pain have been examined in detail (Yaksh and Hammond, 1982). If the experience of pain were a straightforward translation of nerve activity into sensation, the task of defining and measuring it would be simple. However, the perception of pain is complex and so too is the measurement of that experience. An individual's pain is dependent not only on the level of nerve activity, but also on previous pain experiences and a variety of emotional and sociocultural influences (Sternbach, 1983). Hence, pain is a truly unique, sensory and emotional experience.

Before pain can be measured, it must be defined, though this is by no means simple. Thomas Lewis (1942) declined to define pain saying, 'I am so far from being able satisfactorily to define pain, that the attempt could serve no useful purpose.' Nevertheless, some years later a definition was agreed by a Taxonomy Committee set up by The International Association for the Study of Pain (IASP; Merskey and Bogduk, 1994). This group defined pain as 'an unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage' with the subscript, 'Pain is always subjective.' This definition treats pain as a human experience and is not appropriate for animals as it may be unreasonable to assume that an animal could express 'an unpleasant sensory and emotional experience' (Wall, 1992). The task of deriving such a definition for animals is difficult and solutions have been offered ranging from the anthropomorphic view that human beings can relate to animals only through analogy to the human experience and therefore pain in animals can only be defined in terms of human

pain, given by Lewis (1942), to the opinion of Wall (1992) that defining pain serves no real purpose. Wall was of the view that the phrase 'pain in animals' was meaningless. He went on to say that it was of greater importance to observe animals to learn how they deal with and react to pain. Kitchell *et al* (1987) proposed that pain in animals was 'an aversive sensory and emotional experience (a perception), which elicited protective motor actions, resulting in learned avoidance, and possibly modifying species-specific traits of behaviour.' Although the viewpoints expressed by Lewis, Wall and Kitchell seem to differ, they do have one point in common, which is that pain is recognised by observation of an animal's behaviour.

The measurement of pain is one of the most difficult tasks undertaken by health professionals in all areas of medical and surgical care (Taylor, 1985; Banos *et al*, 1989; Beyer and Wells, 1989). Critics have said that the measurement of pain is impossible, and while it is accepted that pain has differing intensities, the internal and personal nature of the experience brings such critics to believe that assigning a number to pain intensity is meaningless (Savage, 1970; Chapman 1976). However, the measurement of pain is important, since inadequate treatment of pain can have detrimental effects on a patient (McGrath and Hillier, 1989; Bonica, 1992). Thus, the conservative views expressed by some have not deterred workers who have investigated different methods of pain measurement.

1.2 The theory of measurement

Lord Kelvin declared that measurement was the key to understanding when he said, 'When you cannot measure it, when you cannot express it in numbers - you have scarcely in your thoughts, advanced to the stage of Science, whatever the matter may be' (Chapman, 1976). This is true in all branches of science, not least in pain research.

The concept of measurement is a simple one: according to a set of rules, numbers that quantify an attribute are assigned to objects or events. The requirement for definitions to be laid down regarding how these numbers are assigned is an issue for debate. It has been argued that without clear definitions detailing how the attribute is to be measured, it is impossible to obtain empirically verifiable scores (Chapman, 1976). However, in everyday life we encounter measurements that are made without operational definitions, for example measurements of weight and height are accepted without explicit instructions being laid down.

The difference between measuring attributes such as weight or height and pain lies in how they are manifest. Weight is a physical phenomenon that can be measured directly, whereas pain is a purely internal and personal experience. Consequently, no single entity or event provides objective measurement of a patient's pain. Pain can only be measured indirectly via the patient's reports or behaviours (Chapman *et al*, 1985). Thus, a key issue in pain measurement is to identify and define items and events that provide information on the patient's pain experience. The need for definitions, which allow indirect assessments and so provide information on the pain experience, is crucial. The validity of the items and events included in a measurement scale, and thus the scale itself, can be judged by the assumptions made in their construction (Chapman, 1976). The fewer assumptions made, and the more the events, items and definitions are demonstrated empirically, the greater the confidence that can be placed in the method. Hence, an ideal pain measurement scale should contain items with intuitively appealing definitions, which can be shown empirically to be related to pain.

In Psychology, a great deal of research has been undertaken to explore the construction and properties of measurement scales, the principles of which can be applied to the measurement of pain. In 1985, the American Psychological Association published a manual detailing a comprehensive set of standards that should be met by measurement methods used in Psychology and Education. The criteria defined have been accepted as providing benchmarks against which the performance of any measurement scale can be judged, i.e. health measurement scales as well as psychological tools (Streiner and Norman, 1995).

When investigating a measurement scale the main concerns are the performance of the scale and the properties of the scores derived. The performance of a scale can be explored by examining the validity and reliability of the method. Both validity and reliability are familiar ideas, but within measurement and psychometric theory, these properties have specific definitions. The level of measurement of a scale is also important; it does not provide information about the performance of a scale although it does give an insight into the information contained within the scores.

1.2.1 Level of measurement

Four levels of measurement exist: nominal, ordinal, interval and ratio measurement. Nominal measurement occurs when values are assigned from categories that have no inherent ordering, for example, race or sex. The categories serve merely as labels for the

observations made and do not quantify them. Ordinal measurement constitutes responses that are ordered and categorical in nature, such as mild, moderate, severe. This level of measurement provides no information on the differences between the categories, only on their relative ordering. Interval level measurement applies to continuous measurement where the difference between an observed response or score and a constant is known and the spacing between points within the scale is consistent. For example, on an interval scale, the difference between scores of 8 and 10 has the same interpretation as the difference between scores of 2 and 4. Ratio level measurement is similar to interval level but with one added condition: the zero point is absolute. Thus, a score of zero on a ratio scale indicates that the attribute of interest is not present whereas the zero point on an interval measurement scale is arbitrary (Streiner and Norman, 1995).

1.2.2 Validity

The basic concept of validity in any measurement scale is a simple one: validity is ascribed when the scale is shown to measure the property for which it was developed (Kline, 1993; Streiner and Norman, 1995). In the case of pain measurement scales, validity is investigated by assessing to what extent a scale actually measures pain. One interpretation of this concept is that the validity of a scale reflects the confidence that can be placed in any decisions made based on the scores observed (Nunnally, 1978; Streiner and Norman, 1995; Cohen *et al*, 1996).

Examination of the validity of measurement scales concerned with attributes that cannot be measured directly is crucial to understanding the performance of the scale. In cases where the attribute cannot be observed directly, it is possible that individuals will define the attribute in different ways, and measurement of the attribute may therefore be dependent on the definition and hence the individual. Investigation of the validity of the measurement method is required in these circumstances as it provides an insight into whether observed measurements consistently reflect the attribute of interest (Streiner and Norman, 1995). The definition of pain and the way in which pain is manifest in animals have been debated within the veterinary community (Chapman, 1989; Bateson, 1991; Sanford *et al*, 1986; Ericksson and Kitchell, 1984; Bonica, 1992). Hence, the need to investigate the validity of any pain measurement method for use in animals is clear since pain cannot be measured directly and its definition is much debated.

Validity can be separated into four different types: face, content, criterion and construct validity (Guion, 1977; Landy, 1986). These can be regarded as individual attributes and

investigated on an individual basis (Landy, 1986). However, this strict partitioning has been criticised and is no longer accepted without question (Cohen *et al,* 1996; Streiner and Norman, 1995). An alternative view is to regard each type of validity as contributing to the overall validity of the scale.

1.2.2.1 Face and Content Validity

The face validity of a scale is the simplest type of validity. It is a subjective judgement that the scale is thought to be valid by one or more experts, i.e. it indicates whether the users think the scale looks valid.

Content validity is slightly more complex. It was developed in the field of achievement testing and is also known as content relevance or content coverage (Messick, 1980; Streiner and Norman, 1995). The content validity of a scale examines the scope of the items included in the scale.

Content validity addresses the question of whether one person with a higher score than another has more of the attribute of interest. This is achieved by investigating whether the scale items tap into all the factors relevant to that attribute (Kline, 1993; Streiner and Norman, 1995). If the items included in a scale are representative of all the factors related to the attribute of interest, then the scale has good content validity (Cohen *et al*, 1996). If some aspect of the attribute is not addressed by the items in the scale then incorrect inferences could be made since the scale does not provide all the information required.

Content validity is an intrinsic property of any measurement scale and is determined during development. Ideally, the items contained in a scale should encompass as many factors relating to the attribute as possible. Once a measurement method is used in practice, the level of content validity becomes apparent from discussion with those using the scale or by consulting experts in the field (Kline, 1993). An example of a pain scale where content validity is satisfied is the McGill Pain Questionnaire (MPQ; Melazack, 1975). The MPQ has been used in clinical practice for many years and patients have indicated that it provides them with a meaningful method of communicating their pain experience (Reading, 1983). Conversely, the content validity of subjective rating scales such as the visual analogue scale (VAS) cannot be appraised as these scales do not specify the factors which should be incorporated into the measurement process. However, the face validity of scale such as the VAS can be investigated as they simply indicate the belief that the users have in the ability to measure pain.

1.2.2.2 Criterion Validity

Criterion validity explores the relationship between the new measurement scale of interest and some existing measure or gold standard, the criterion method (Streiner and Norman, 1995; Cohen *et al*, 1996). A new method may be required even though a gold standard is already available in situations where a cheaper, safer or quicker alternative would be an advantage. The criterion method against which a new measurement scale is tested can take almost any form, however it must provide acceptable measurement of the attribute (Cohen *et al*, 1996).

Two types of criterion validity exist, namely concurrent validity and predictive validity. The concurrent validity of a scale is examined by applying the new method and the criterion method at the same time and comparing the results (Kline, 1993; Loewenthal, 1996). A strong relationship indicates that the new measurement method has high concurrent validity. Predictive validity looks at the relationship between the scores on the new measurement scale and the outcome that is observed some time later (Kline, 1993; Cohen *et al* 1996; Streiner and Norman, 1995; Loewenthal, 1996). The predictive validity of a measurement scale is only informative if the scale itself is to be used to predict an outcome in advance.

The criterion validity of the pain scales in current use in veterinary medicine would be difficult to ratify as there is no globally accepted gold standard, or indeed any valid and reliable method, of measuring pain in animals.

1.2.2.3 Construct Validity

The theory behind construct validation of a scale is slightly more complex than content and criterion validity. A construct is a hypothetical concept that is developed or 'constructed' to explain relationships between attributes. It cannot be observed directly since it exists only as a theoretical relationship between various factors such as behaviours, attitudes, performance on tests, etc (Cohen *et al*, 1996). Pain is a construct since it cannot be observed directly, but can only be measured through factors such as patient reports, medication request, behaviour, etc. Construct validity is investigated by first formulating hypotheses about the relationships between the construct under investigation and some other variables or constructs. These hypotheses can then be explored, and if the expected relationships are upheld then the construct validity of the measurement scale is demonstrated (Cohen *et al*, 1996; Streiner and Norman, 1995).

For any measurement scale, no single study can 'prove' construct validity; typically, a number of validation studies are carried out. If an investigation does not support construct validity it does not immediately mean that the scale is invalid (Cohen *et al*, 1996; Streiner and Norman 1995). In such investigations both the validity of the scale and the theory underlying the construct are being examined, therefore the interpretation of a study which does not support the validity of the scale is not straightforward. The result may be unsupportive for any of the following reasons: the measurement scale may be valid, but the theory behind the study wrong, or the theory may be correct, but the scale invalid or finally, the scale may be invalid and the theory incorrect. Further validation studies are required before the reason for the result can be identified.

In summary, the methods used to explore the types of validity differ although the fundamental concepts have the same basis. It has been said that both content and criterion validity are merely alternative forms of construct validity (Guion, 1977). This clouding of the distinctions between the types of validity has led to a reduction in the need to identify and examine them independently. Studies may be simply labelled as investigating the general validity of a measurement scale.

1.2.3 Reliability

The concept of reliability is familiar, however when investigating a measurement scale it has a specific meaning that must be fully understood before its implications can be appreciated. The reliability of a scale gives an indication of how much of the variability in the scores is due to errors in the measurement method. The reliability coefficient is the ratio of the between subjects variability to the total variability and can range from 0 to 1. The value of the reliability coefficient indicates how much of the variability is due to real differences between the individuals, and how much is due to random error caused by inaccuracies in the measurement method.

$$Reliability = \frac{Subject\ Variability}{Total\ Variability} \qquad \text{Equation 1.1}$$

A value of 0 indicates that the variability is not attributable to differences between the subjects and is purely due to measurement error, therefore the observed scores do not differentiate between the subjects, and the scale is not reliable. A reliability of 1 indicates that all of the observed variability is due to real differences between the subjects and none

is due to measurement error (Cronbach, 1970; Nunnally, 1978). Hence, the measurement scale can be said to be reliable.

Intuitively it would be expected that a measurement scale with a high reliability would imply high agreement between subjects. However, this is not necessarily the case. Where all subjects are given the same score the subject variability would be zero, and hence the reliability would be zero (Streiner and Norman, 1995). In addition, reliability of a measurement scale is not a fixed property, but is dependent on the sample in which it is tested (Nunnally, 1978; Streiner and Norman, 1995; Cohen *et al*, 1996). If the measurement scale were used on a sample of subjects who are very heterogeneous, it would be expected that the subject variability would be large relative to any measurement error. Thus, the resulting reliability coefficient would be large. However, if the reliability of the same scale were to be examined in a more homogeneous group it is likely that the reliability coefficient would be smaller since the subject variability would be reduced but the error variability would be unchanged.

Two main approaches for the investigation of reliability have been developed: classical test theory and generalizability theory (Streiner and Norman, 1995; Cohen *et al*, 1996).

1.2.3.1 Classical Test Theory

The original methodology that led to the formulation of a reliability coefficient is known as classical test theory. This assumes that any measurement can be broken down into two component parts, namely the underlying true score for the subject and the measurement error associated with that observation.

$$X_{ij} = \tau_i + \varepsilon_{ij} \qquad \text{Equation 1.2}$$

Where X_{ij} : observed score for subject i at measurement j

τ_i : true score for subject i

ε_{ij} : measurement error associated with subject i at measurement j

Using this model the classical definition of a reliability coefficient is the ratio of the subject variability to the total variability observed, where the total comprises subject plus measurement error variability. This coefficient is known as the intra-class correlation coefficient (ICC) and it provides an overall estimate of the reliability of the measurement scale (Nunnally, 1978; Streiner and Norman, 1995).

The intra-class correlation coefficient (ICC) or overall reliability coefficient, assumes that all variability other than between subjects is due to measurement error; in practice, this assumption may not hold. According to Cohen *et al* (1996), variability may be caused by a number of sources such as multiple observers or changes over time. Coefficients to examine the error caused by these sources have been developed, for example, inter-observer reliability, test-retest reliability, and internal consistency, also known as Cronbach's α (Cronbach, 1970; Nunnally, 1978; Streiner and Norman, 1995).

This classical model is based on the premise that each individual has an underlying true score and is the most widely used model when investigating reliability (Cohen *et al*, 1996). However, alternative theories such as generalizability theory that have moved away from the concept of an absolute true score have been developed and have gained acceptance as an extension to classical theory (Cronbach, 1970).

1.2.3.2 Generalizability Theory

Generalizability theory was developed by L. J. Cronbach in the 1970s (Cronbach, 1970; Cronbach *et al*, 1972). The theory was based on the idea that a subject's scores may vary because of the circumstances under which they are tested. Generalizability theory provides a mechanism for exploring these factors in a manner that may not be possible using classical test theory (Streiner and Norman, 1995).

The first step in designing a generalizability study is to identify the sources of variability most likely to influence a subject's score. These multiple factors comprise the '*universe*' under which the measurements are made. The subjects in such a study are known as the '*facets of differentiation*' since the aim of the measurement scale is to differentiate between them. The other sources of variability are known as '*fixed facets*' or '*facets of generalization*' (Cronbach, 1970). The '*facet of generalization*' is the facet that the researcher wishes to explore. It should be noted that the term '*fixed facets*' is used to identify the facets which are held constant when calculating a particular generalizability coefficients, i.e. all facets other than the facet of differentiation and generalization relating to that coefficient. The facets identified as '*fixed facets*' or '*facets of generalization*' is dependent on the generalizability coefficient which is being calculated, not on the design of the study and therefore is not influenced by the statistical model used to calculate the components of variance.

An example of such a universe would be if pain were measured in 16 dogs by 4 veterinary surgeons on 4 occasions. In such an experiment the researcher may be interested in the effect of different observers using the scale and hence whether the scores can be generalized over observers. The facet of generalization is the observers, the facet of differentiation is the dogs, and time is the fixed facet. When considering this model there is no concept of a subject's true score, the average score over all the possible factors is considered an unbiased estimate of their score within that universe. This is not the same as their true score since the universe may not include all factors that contribute to error and therefore the average may not be the subject's true value.

This definition of the universe identifies the facets included in the generalizability model and allows them to be explored using statistical modelling. The statistical models used in such studies are general linear models, specifically random or mixed effects models. An example of the type of model that may be fitted to the study described above is shown in Equation 1.3.

$$X_{ijk} = \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \varepsilon_{ijk} \qquad \text{Equation 1.3}$$

Where X_{ij} : Score allocated by observer *i* to subject *j* on occasion k

α_i : Random effect of observer *i*, distributed $N(0,\sigma^2_{obs})$

β_j : Random effect of subject *j*, distributed $N(0,\sigma^2_{dog})$

γ_k : Random effect of time *k*, distributed $N(0,\sigma^2_{time})$

$\alpha\beta_{ij}$: Random interaction between observer and subject

$\alpha\gamma_{ik}$: Random interaction between observer and time

$\beta\gamma_{jk}$: Random interaction between subject and time

ε_{ijk} : Random error effect, distributed $N(0,\sigma^2_{\varepsilon})$

In this model the observer and time effect may be fitted at fixed or random effects, as appropriate, depending of the design of the study. The choice between fixed or random main effects will not influence the definition of the generalizability coefficients, as the interaction between subject and the two effects will be random. From the random effects model, the observed mean squares for each factor can be decomposed into the relevant components of variance. The methodology of mean square decomposition and components of variance are discussed by Glass and Stanley (1970) and by Snedecor and Cochran (1980). The components of variance are used to calculate the generalizability coefficients in a similar way to the reliability coefficients previously described. However, since a

potentially infinite number of factors could be incorporated as facets in the generalizability study, a potentially infinite number of generalizability coefficients can be calculated. These coefficients are constructed by following a similar format and can be used to examine any facet of generalization (Streiner and Norman, 1995).

A generalizability coefficient identifies and quantifies the variability due to the corresponding facet of generalization (Streiner and Norman, 1995). To isolate this variability the coefficient is calculated as the ratio. The numerator of the ratio contains the variability in scores due the facet of differentiation (for example the subjects) and any interaction between the facet of differentiation and the fixed facets. This incorporates all factors that contribute to the variability in the subject's scores, other than the facet of generalization. The denominator comprises the variability due to the facet of differentiation, any interactions with this facet and the error variability. This incorporates all factors that contribute to the variability in the subjects' scores. Therefore, the ratio highlights how much of the variability in the subjects' scores is due to the facet of generalisation. A coefficient close to 1 indicates that the variability due to the facet of generalization is small and hence the scale is generalizable over that facet (Nunnally and Bernstein, 1994; Guilford, 1986; Streiner and Norman, 1995). In the example described the generalizability coefficient over the observers is as shown in Equation 1.4.

$$G = \frac{\sigma^2_{subject} + \sigma^2_{subject*time}}{\sigma^2_{subject} + \sigma^2_{subject*observer} + \sigma^2_{subject*time} + \sigma^2_{\varepsilon}} \qquad \text{Equation 1.4}$$

In this example, the factors that contribute to the variability in the subjects' scores are variability between the subjects, variability between subjects over observers, variability between subjects over time and any residual error variability. These factors constitute the denominator of the generalizability coefficient. Since, equation 1.4 is concerned with the generalizability over observers the numerator comprises the factors that contribute to the subjects' variability, excluding the observers, i.e. subject variability and variability between subjects over time. Thus, the ratio indicates what proportion of the total variability in the subjects' scores was due to the observers.

Similarly, the coefficient indicating generalizability over time could be derived. This would take a similar form to that shown in Equation 1.4, as the denominator would be identical. However, the aim would be to isolate and quantify how time contributes to the

total variability therefore, the numerator would comprise the variability due to subjects, and between subjects over observers.

A number of the generalizability coefficients are equivalent to reliability coefficients used in classical test theory. For example, Equation 1.4 is equivalent to the classical inter-observer reliability. However, the wider scope of generalizability theory means that the coefficients can be calculated to address a broader range of situations than would be possible using classical test theory (Streiner and Norman, 1995).

1.2.4 Relationship between validity and reliability

Validity and reliability are intrinsically linked. Low reliability in a scale implies a great deal of measurement error is present in observed scores, therefore it would be impossible to show perfect validity. The random variability in the measurement error would cloud the relationship between the score and validity criteria being used, thus the validity of a scale is limited by its reliability (Nunnally, 1978; Guilford, 1986; Streiner and Norman, 1995).

1.2.5 Measurement theory and the measurement of pain

The methodology described for examining the performance of measurement scales was first developed in Psychology, where interest lies primarily in attributes that can only be measured indirectly. However, these ideas are just as applicable to measurement methods in other areas, particularly in the measurement of health where a variety of conditions and diseases require indirect measurement (Guyatt *et al*, 1992; Streiner and Norman, 1995).

1.3 Measurement of pain in adults

The methods used for measuring clinical pain in adults fall into three main categories, those based on physiological signs, self-reporting scales and observational scales.

1.3.1 Physiological Signs

The physiological signs most commonly cited as indicating pain include pulse rate, temperature, respiratory rate and skin conductance. It has been shown that these variables change in response to pain, but that these changes lessen over time as the patients become familiar with the pain (Chapman *et al*, 1985). Where the pain experience has a longer duration, such as post-surgical or chronic pain, the utility of such methods is limited.

A more complex approach using physiological signs is to directly measure the peripheral nerve activity that forms the basis of pain itself (Culp *et al*, 1982). However, pain is also an emotional experience and does not simply reflect the level of nerve activity. Fors *et al* (1984) demonstrated that the relationship between nerve activity and pain is complex and that this method does not provide reliable measurement. Further, Wolf *et al* (1982) proposed that muscle tension measured via electromyography could be used to measure pain, but that relationship was also found to be unreliable (Kravitz *et al*, 1981). Thus, in adults, no single physiological sign has been identified as a reliable measure of pain. The majority of methods used in the assessment of human pain rely on patient self-assessment, via simple subjective rating scales or using more sophisticated composite measurement scales.

1.3.2 Self-Reporting Scales

1.3.2.1 Subjective Rating Scales

The simplest self-reporting pain tools are the subjective rating scales, such as the simple descriptive scale (SDS), the numerical rating scale (NRS) and the visual analogue scale (VAS). These scales are widely used because of their simplicity, though their usefulness in the measurement of pain has been questioned (Revill *et al*, 1976; Linton, 1983). Whereas pain is acknowledged to be a multidimensional experience that varies in temporal, spatial and affective dimensions, as well as in intensity, the subjective rating scales address only intensity (Melzack, 1975; Chapman, 1976). The scores obtained when using the subjective scales can be liable to response bias because patients are forced to express the entirety of their experience on an artificially small continuum (Gracely, 1980). In addition, it is often assumed that interval level measurement scores are produced using such scales. This is not always the case and this assumption can lead to the use of inappropriate statistical methods and unreliable results (Chapman, 1976).

More specifically, simple descriptive scales have been shown to be sensitive to age and ethnic differences (Kaiko *et al*, 1983). Visual analogue scales have been shown to be unreliable in the assessment of chronic pain (Carlsson, 1983) and have poor sensitivity when used to measure the effects of analgesics (Atkinson *et al*, 1982).

1.3.2.2 Composite Measurement Scales

The limitations of the subjective rating scales have stimulated the development of more complex multidimensional composite measurement scales for use in pain research. These are aimed at addressing the different dimensions of the pain experience (Guyatt *et al*, 1992).

The most widely used and thoroughly explored composite measurement scale for human pain is the McGill Pain Questionnaire (MPQ). The background development of the questionnaire was first published by Melzack and Torgerson in 1971, but it was 1975 before it was published as a formal measurement method (Melzack, 1975). The questionnaire itself consists of 102 expressions that describe differing aspects of pain. These expressions are divided into 3 major classes (sensory, affective and evaluative), which contained 16 subclasses in total. For example, in one subclass patients are asked to indicate whether their pain is Sharp, Cutting or Lacerating, in another subclass the descriptors are Hot, Burning, Scalding and Searing. When using the MPQ, each patient is asked to pick the expressions from each subclass that best describe their pain, although they need not choose expressions from every group. Thus, when the questionnaire has been completed, the patient will have identified a list of expressions that describe the sensory, affective and evaluative aspects of their pain. When the MPQ was first developed, the overall pain score was defined in a number of ways. These were the sum of weights allocated to each expression chosen, the total number of expressions chosen by the patient and the patient's subjective assessment of their overall present pain intensity (Melzack, 1975). Since the initial development of the scale, the numbers of words chosen and sum of weighted expressions have become the most commonly used scoring methods.

The structure of the MPQ was defined to allow three different dimensions of pain to be assessed, i.e. the sensory, affective and evaluative dimensions (Melzack, 1975). Independent investigations have shown that the MPQ does contain more than one dimension of pain and hence is multidimensional in nature. However, the dimensions demonstrated were not always consistent with those originally proposed (Prieto and Geisinger, 1983; Doctor, 1995). The scale has been shown to be useful in the measurement of acute and chronic pain although the dimensions were less distinct when measuring chronic pain since the pain descriptors fall into 6 factors rather than the 3 proposed by Melzack (Reading, 1983). The MPQ has been translated into a number of different languages including Norwegian and Finnish (Ketovuori and Pontinen, 1981; Strand and Wisnes, 1991; Kim *et al*, 1995).

The reliability, face, construct and criterion validities of the MPQ have been investigated (Reading, 1983). In his appraisal of the performance of the MPQ, Reading (1983) discussed research papers where the MPQ had been shown to fulfil these criteria and concluded that the MPQ was a valuable tool in the measurement of pain in humans. However, he noted that further research would shed more light on the relationships between the MPQ and other pain scales in current use.

The MPQ is not without limitations as it is more complex than the subjective rating scales, requires more time to complete, and therefore cannot be used as widely. Its use is questionable in some patient groups such as the very ill and the elderly. In addition, the MPQ makes use of some complex vocabulary and so the degree to which the patient comprehends the scale cannot be guaranteed. This could lead to problems when comparing pain scores across cultural and social groups.

1.3.3 Observational Scales

In adults, the use of pain measurement scales based on observation of the patient's behaviour has been less common than self-reporting scales (Fordyce, 1983). A number of behaviours have been reported as useful indicators of pain, such as activity levels, alteration in sleep patterns, medication demand etc (Chapman, 1985). Some examples of behaviours that are used in observational tools for the assessment of chronic pain are guarded movement, rubbing and bracing (Keefe and Block, 1985). Other, more specific behaviours such as facial expression and the frequency of discussion relating to pain have also been included in such scales (Le Resche, 1982).

Discrepancies between a patient's self-reported pain and pain scores based on observed behaviour have been highlighted. This may be due to inaccuracies in the patient's self-report or complications in the use of behaviour as a predictor of pain (Fordyce, 1983). While observation of the patient can be informative, the validity of scales based exclusively on this has been questioned. Verbal or written communication between patient and carer, about the internal experience of pain, is optimal for effective pain management. However, situations often arise where verbal communication between the patient and carer is not possible, such as in very young children.

1.4 Measurement of pain in children

The measurement of pain in children is one of the most challenging problems faced by clinicians. Despite the difficulties, research in this area has grown rapidly over the last 10 years (Finley and McGrath, 1998). Testimony to this was the first International Symposium on Paediatric Pain held in 1996 and the formation of the first special interest group by the International Association for the Study of Pain, dedicated to examining pain in children.

It has been acknowledged that under-prescription and under-administration of analgesics has been commonplace in clinical practice (McGrath and Brigham, 1992). This, coupled with the fact that inadequate treatment of pain in young children can have profound developmental effects, has provided the impetus for research into pain measurement in this patient group, an overview of which is provided by Finley and McGrath (1998).

1.4.1 Physiological Signs

The use of physiological variables to measure pain is attractive as it could provide a simple measure, independent of a child's cognitive development. However, the validity and reliability of the physiological parameters have been difficult to establish (Erickson, 1990; Hester, 1993; McGrath, 1996). The physiological signs that have been used to indicate the presence of pain include heart rate, respiratory rate, blood pressure, oxygen saturation and palmar sweating (Harpin and Rutter, 1990; Johnston and Strada, 1986; Maxwell *et al*, 1987; Durnad *et al*, 1989; Howard *et al*, 1994). While all of these parameters have been shown to change in response to painful stimuli, such as heelstick or circumcision, their performance when used in other types of pain is unknown. These signs have only been explored when indicating the presence of pain, but do not allow pain to be quantified (Hester, 1993). Thus, in their review of the literature Sweet and McGrath (1998) found no single physiological sign that provided acceptable measurement of pain.

1.4.2 Self-Reporting Scales

From the age of approximately 3 years, children can understand and communicate varying degrees of pain intensity (Beyer and Wells, 1989). Therefore, self-reporting tools can be used in children who can communicate their pain using some abstract mechanism (Champion *et al*, 1998). A number of self-reporting tools have been developed for use in children of differing age groups.

One simple example is a verbal rating scale which contains 5 pain categories ranging from 'none' to 'very severe'. The patient picks an expression that best represents his or her pain (Frank *et al*, 1982). Other more abstract scales include the Poker Chip tool (Hester, 1979) where the patient assigns the number of 'pieces of hurt' according to how much pain he or she is experiencing. The Oucher (Beyer, 1984) comprises 6 photographs of young children's faces in varying degrees of pain. The photographs are arranged in ascending order of intensity and the patient picks the photograph that best indicates their pain. Similar facial scales have been developed based on photographs and drawings of children in varying degrees of pain. It has been shown that the children's ability to use these scales is dependent on their cognitive development, and no single self-reporting scale is completely satisfactory in all age groups (Dworkin & Whitney, 1992).

Multidimensional self-reporting tools for pain measurement have not been developed for use in children to the same extent as in adults (Champion *et al*, 1998). Nevertheless, self-reporting methods currently provide the best available assessment of children's pain experiences (Beyer and Wells, 1989) but, as mentioned previously, these are not feasible in children aged 3 years and under.

1.4.3 Observational Scales

Since self-reporting tools cannot be used in children under 3 years old and physiological parameters are uninformative in quantifying pain the only options for pain measurement in very young children lie in tools based on observation of the patient's behaviour.

Altered behaviours may be a first indication to a carer that a child is in pain. Behaviours indicating pain in infants include torso and limb movements, facial expressions and crying patterns (McGrath, 1987). In children from birth to 4 years, changes in torso and limb movements in response to a painful stimulus were found to be dependent on age and cognitive development (McGraw, 1945). However, these claims have been challenged and the change in response with age is under debate (McGrath, 1987). An infant's facial expressions and crying patterns have been shown to change in response to pain (Grunau and Craig, 1987; McGrath, 1987). These behaviours do not provide an unequivocal measure of an infant's pain when examined individually.

Scales based on observation of a patient's behaviour have been developed and their psychometric properties investigated. One of the most frequently used is the Children's Hospital of Eastern Ontario Pain Scale (CHEOPS, McGrath *et al*, 1985). A trained

observer uses the CHEOPS to record the child's crying, facial expressions, verbal communication, torso movement, response to touch and leg movement. Each behaviour has a pre-specified weight and the sum of these weights constitutes the total score. The validity and reliability of the scale have been established. However, CHEOPS was originally developed and validated in children aged between 1 and 7 years old as they emerged from anaesthesia following surgery, so the application of the tool in other groups and circumstances may be questionable (Beyer *et al*, 1990).

The Observer Scale is similar in aim to the CHEOPS, but has a simpler scoring system (Krane *et al*, 1987). This scale simply describes 5 different behavioural states and each patient is allocated a score according to the description of their behaviour. It has not been validated formally, but it was thought to be useful as it was designed to parallel the decision-making process in post-operative wards with respect to patient care (Tyler *et al*, 1993).

The Observational Scale of Behavioural Distress (OSBD) and the Procedural Behavioural Rating Scale-revised (PBRS-r) were developed to assess pain in children with cancer during lumbar puncture and bone marrow aspiration procedures (Katz *et al*, 1980; Jay *et al*, 1983). The OSBD examines both pain and anxiety, as it can be difficult to differentiate between these in a clinical setting (McGrath, 1987). Both scales examine 11 behaviours that are observed over a specified period. Each behaviour is allocated a score of between 1 and 4, where 4 indicates the highest intensity of pain. The weights assigned were defined by clinical personnel familiar with the procedures and overall scores are obtained by adding the individual scores of the 11 behaviours. While both the OSBD and PBRS-r have been shown to be reliable when used to measure pain and distress in the procedures detailed above, it has not been established if it is appropriate to apply these methods to patients undergoing different procedures (McGrath, 1987).

In summary, the use of behavioural measures in paediatric pain has been restricted to measurement of pain in acute situations, for example, following surgery or painful procedures such as lumbar puncture. The use of these methods has been shown to give a reliable and valid assessment of a patient's distress. However, behaviours can be influenced by many external factors including the presence of parents, and whether the observations can be refined to ensure that only pain is assessed and other emotional factors such as fear do not influence the results has been questioned (McGrath, 1987).

1.5 Measurement of Pain in Animals

Pain measurement in animals has generated a great deal of interest in recent years. One major difficulty in the measurement of pain in animals is the lack of an effective means of communication between the patient and carer. There is an obvious parallel between pain measurement in very young children and animals, where the only means of communication between patient and carer is through observation of the patient's behaviour.

The problem of pain recognition in animals is further complicated by the huge range of species-specific behavioural traits. People may be confident in their ability to recognise pain in one species but not in others. Despite these difficulties, there is a wealth of information available regarding the measurement of pain in animals and the tools that are currently in use.

1.5.1 Physiological Signs

A number of physiological markers have been studied in relation to pain in animals. Physiological responses that have been quoted as being indicative of pain include tachypnoea or panting, sinus tachycardia, hyperglycaemia, hypotension or hypertension, dilated pupils and pallor (Pain *et al*, 1986; Crane, 1987; Haskins, 1987; Kitchell, 1987; Spinelli and Markowitz, 1987; Willis and Chung, 1987; Hellyer and Gaynor, 1998). The use of these signs to quantify pain, rather than merely indicate its presence, has not been established and further study is required. The relationships between pain and three parameters thought to be indicative of pain (heart rate, respiratory rate and pupil dilation) have been shown to be tenuous (Conzemius *et al*, 1997). These issues are discussed further in Chapter 5 of this thesis.

Other work has shown that, in dogs, plasma cortisol levels are elevated above normal levels following surgery (Fox *et al*, 1994). However, cortisol is also released in non-painful situations, so cannot be used as a reliable measure of pain (Mason, 1968). In cats, plasma catecholamine concentration following surgery changed according to the use of analgesic agents, suggesting that catecholamine levels could be used as a marker for pain (Benson *et al*, 1989). However, conflicting evidence indicates that catecholamine levels were no different during recovery from surgery compared to pre-surgery levels (Rawlings *et al*, 1989).

1.5.2 Observational Scales

A number of behaviours that are thought to be indicative of pain and distress in animals have been described (Ericksson and Kitchell, 1984; Sanford *et al,* 1986; Chapman, 1989; Bateson, 1991; Johnson, 1991; Sackman, 1991; Heavner, 1992; Light *et al*, 1993). However, there is little discussion of how these behaviours could be used to formally measure pain intensity in animals as part of a measurement scale.

1.5.2.1 Subjective Rating Scales

The simplest and most frequently used scales based on the subjective rating of animal behaviour are the SDS, NRS and VAS (Reid and Nolan, 1991; Nolan and Reid, 1993; Welsh *et al*, 1993; Lascelles *et al*, 1994). The VAS and NRS have demonstrated good agreement between two trained observers when used in a post-operative setting (Reid and Nolan, 1991). This finding was supported when the VAS and NRS were used in the measurement of lameness in sheep (Welsh *et al*, 1993). The validity of the scales has not been formally investigated, although they have been shown to be sensitive enough to detect the effects of analgesics, which supports their validity (Reid and Nolan, 1991; Nolan and Reid, 1993; Lascelles *et al*, 1994).

1.5.2.2 Composite Measurement Scales

Composite measure scales for the measurement of abstract attributes such as pain or intelligence have been developed by the human medical and psychological communities (Guyatt *et al*, 1992; Streiner and Norman, 1995). However, it has taken some time for the utility of such methods to be recognised in the veterinary literature.

A composite measurement type approach for use in quantifying animal pain was first proposed by Morton and Griffiths in 1985. They published 'Guidelines for the recognition of pain, distress, and discomfort in experimental animals and a hypothesis for assessment'. The paper detailed a number of specific behaviours including posture, vocalisation, temperament, food and water intake, and locomotion as being potentially indicative of pain. Scores from 0 to 3 were allocated for specific changes in an animal's bodyweight, appearance, relevant clinical signs, unprovoked behaviour and response to stimuli. The overall pain score was the sum of these ratings and could range from 0 to 24. The paper also defined what interpretation should be applied to certain ranges of scores: 0-4 was regarded as normal; 5-9 indicated the animal should be monitored carefully; 10-14 implied relief from suffering or termination of the experiment should be considered; any score

greater than 14 and the experiment should be terminated. However, the method was not designed for use as a pain measurement tool. Rather, the purpose of the guidelines was to ensure consistent practice between laboratories and to provide an objective set of criteria to safeguard animal welfare.

In the year following the publication of Morton and Griffith's guidelines, Sanford *et al* published further 'Guidelines for the recognition and assessment of pain in animals' (1986). The aim of this document was to provide an effective and uniform set of criteria to control the suffering endured by experimental animals. The behaviours listed as indicating pain were posture, facial expression, gait, acceptance of handling, vocalisation and overall mental status. These guidelines also suggested that a clinical examination should be carried out with particular attention being paid to physiological signs such as pupil dilation, changes in blood pressure, increased heart rate and increased body temperature.

More recently, a tool that has been used in the measurement of post-operative pain in dogs is the numerical rating scale (Conzemius *et al*, 1997). Despite having the same name, this scale differs from the subjective rating scale described previously and explored in Chapter 2. The method details three types of behaviours, namely vocalisation, movement and agitation. The observer allocates a score from 0 to 2 for the first two behaviours and 0 to 3 for the third, according to pre-defined rules. The total score is the sum of scores for each behaviour and gives a possible total score of between 0 and 7.

In their 1998 paper, Hellyer and Gaynor acknowledge that for pain to be assessed reliably a well-defined pain scale is required, hence the behaviour-based Colorado State University Veterinary Teaching Hospital Pain Scale was developed. The scale consists of 8 categories: comfort, movement, appearance, unprovoked behaviour, interactive behaviour, vocalisation, heart rate and respiratory rate. Each type of behaviour is assessed and assigned a score of between 0 and 4 according to pre-defined criteria. The aim of the tool was to examine patient's requirement for analgesia; it was not intended to be a formal research tool.

The most recent composite measurement scale developed to evaluate post-operative pain in dogs, based on assessment of behaviours and physiological signs, is the University of Melbourne Pain Scale (UMPS; Firth and Haldane, 1999). The items included in UMPS were derived from a review of the pain measurement literature relating to dogs. It contains 6 categories including physiological variables, response to palpation, activity, mental status, posture and vocalisation. Weights were assigned to the items within each category

subjectively, according to the developers' perception of how much pain they implied. The assessments of an animal's mental status, heart rate and respiratory rate were based on the change from the pre-surgery status therefore these variables are assessed before and after surgery. The degree of change in heart rate, respiratory rate and mental status is defined and a score allocated according the extent of the change. When using this tool, the assessor indicates which one item from each category best describes the dog being observed. The weights were assigned to the items subjectively according to the developers' perception of how much pain they implied. This method was similar to that used in the development of CHEOPS (section 1.4.3).

1.6 Development of pain measurement in veterinary medicine

This brief review of the pain measurement scales used in human and veterinary medicine illustrates the type of work that has been undertaken in this area. In particular, the literature demonstrates that measurement properties of the pain scales used in human medicine have been explored much more thoroughly than in veterinary medicine. Only a handful of papers explicitly explore the validity or reliability of scales used in veterinary medicine (Reid and Nolan, 1991; Nolan and Reid, 1993; Welsh *et al*, 1993; Conzemius, 1997; Firth and Haldane, 1999). In an effort to further explore pain measurement in animals, the work reported in this thesis provides an investigation of the current pain measurement scales and the development of a novel approach to pain measurement in animals.

In Chapter 2, three methods of pain measurement commonly used in veterinary medicine are examined. These three scales are also used in human medicine and are among the simplest tools available. Their generalizability over time and between observers and the relationship between them is explored. This work constitutes an exploration of the pain measurement methods currently used in veterinary medicine. The next step was to explore measurement methods that may provide some improvement over the scales currently available.

Chapters 3 and 4 describe the development of a composite measurement pain scale designed for use in animals, specifically dogs. The distinction between the measurement scale discussed in this thesis and others that have been developed in veterinary medicine lies in the methodology utilised in their construction. The work described uses the theory of measurement and psychometrics to construct and investigate a composite measurement scale aimed at measuring pain in dogs, in a clinical setting.

The work detailed in Chapter 5 explores the relationship between three physiological parameters and pain and the performance of the new scale developed in the preceding chapters in comparison to one of the subjective scales investigated in Chapter 2. This provides an insight into the relative merits of the two approaches and hence allows an appraisal of the new composite measurement pain scale. Finally, Chapter 6 provides a summary of the observations made, explores the implications of these and discusses possible avenues for further investigations in this field.

2. Performance of the Visual Analogue Scale, Numerical Rating Scale and Simple Descriptive Scale when used to measure pain in dogs

2.1 Introduction

The review of pain measurement tools in Chapter 1 illustrates the importance of being able to measure pain reliably and accurately. The pain measurement scales currently in use in veterinary medicine are primarily scales developed for use in humans. The three most commonly used scales are the visual analogue scale (VAS; Reid and Nolan, 1991; Nolan and Reid, 1993; Lascelles *et al*, 1994), the numerical rating scale (NRS; Taylor and Houlton, 1983; Taylor and Herrtage, 1986) and the simple descriptive scale (SDS; Taylor and Houlton, 1984; Waterman and Kalthum, 1988). In human medicine, the scales are used by patients, as self-reporting tools to record pain intensity. However, in veterinary medicine observers use the scales to record the intensity of pain they believe an animal to be enduring, based on observation of the animal's behaviours (Morton and Griffiths, 1985).

The VAS typically consists of a 100mm horizontal line with 10mm vertical lines at each end. The ends of the scale are anchored with expressions relating to extremes in pain intensity, e.g. 'no pain' and 'pain could not be worse' (Huskisson, 1974; Langley and Sheppard, 1985). When using the scale the observer places a mark on the line corresponding to pain intensity. The distance measured between the 'no pain' end and the observer's mark corresponds to the animal's pain score, which can be treated as an interval level measurement.

When using the NRS, the observer is asked to assign a numerical score, generally between 0 and 10, or 0 and 100 rather than placing a mark on a line. The endpoints of the NRS are defined by extremes in pain intensity, similar to the VAS. The NRS scores are also commonly treated as interval level measurement.

The SDS is the simplest of the scales and typically consists of 4 or 5 expressions used to describe increasing pain intensities (Chapman *et al*, 1985). The observer picks the expressions they believe best represents the animal's pain. Each expression is allocated a numerical score (e.g. no pain = 0, mild pain = 1, moderate pain = 2 and severe pain = 3)

and the pain score is the value associated with the expression chosen, thus the SDS provides ordinal measurement (Downie *et al*, 1978).

It is worth noting that the VAS, NRS and SDS are not exclusively pain measurement scales and have been used as measurement methods in a number of areas (Aitken 1969; Welsh *et al*, 1993). The scales can be thought of simply as providing the observer with a means of recording their subjective impression of the attribute of interest. The performance of these scales is of critical importance when used to measure an attribute such as pain.

The VAS has been reported to be simple to use and understand in pain measurement (Huskisson, 1974). However, when trying to reproduce points on the scale, subjects tend to underestimate points below 60mm and overestimate points above 60mm (Dixon and Bird, 1981). This indicates that the scale's performance is to some extent dependent on the visual acuity, or hand-eye co-ordination of those people using it (Revill *et al*, 1976). The validity of the VAS has been both supported (Boeckstyns and Backer, 1989) and challenged (Langley and Sheppard, 1985). However, it is in such widespread use in pain measurement throughout human medicine that it can be assumed that any fundamental problems with its validity would have been identified and reported. The reliability of the VAS in pain measurement has been investigated and supported in a number of studies by examining the correlation between consecutive pain scores allocated by patients (Revill *et al*, 1976; Nyrén *et al*, 1987; Boeckstyns and Backer, 1989).

The SDS is also widely used in the measurement of pain, despite criticism that it lacks sensitivity and is unable to detect small changes in pain (Joyce *et al*, 1975; Seymour, 1982). Conversely, Linton and Götestam (1983) contended that the SDS may be advantageous as it would be subject to less error when used by patients, although the benefits of this point may be debated.

Investigations of the NRS for pain measurement in human medicine have been less prevalent than for the VAS and SDS, since it is not as widely used. Price *et al* (1994) reported that the NRS was valid when measuring both experimental and clinical pain. Yet, Seymour (1982) demonstrated that the NRS was less sensitive to demonstrating the effects of analgesics than the VAS. However, Downie *et al* (1978) indicated that the NRS may be more sensitive than the SDS and simpler to use than the VAS, and consequently proposed that it was a suitable compromise between the VAS and SDS. Clearly, each measurement scale has advantages and disadvantages, which should be considered when they are used to measure pain in humans.

Few studies have examined the performance of the VAS, NRS and SDS in veterinary medicine. The VAS has been used to assess pain and sedation in dogs (Reid and Nolan, 1991; Nolan and Reid, 1993; Lascelles *et al*, 1994). These studies have shown that the VAS was sensitive enough to demonstrate the analgesic and sedative effects of four different drug treatments. Nolan and Reid (1991) also indicated that there were no significant differences between two observers when using the VAS and hence the scale was deemed reliable.

Both the VAS and NRS have been used to assess lameness in sheep (Welsh *et al*, 1993). The study found no significant difference between two observers for either of the scales, consequently both were said to be reliable. An investigation of the relationship between the scales in the same study indicated that they should not be used interchangeably as the relationship between them was not unique.

These investigations of the simple unidimensional scales currently used to measure pain in animals suggest that their performance warrants further investigation. The study undertaken here was designed to formally examine the performance of the VAS, NRS and SDS when used to measure clinical pain in dogs. In addition, the relationship between the three scales was investigated.

The study discussed had three main objectives. These were to examine:

- the generalizability of the VAS and NRS, particularly when used across time and by different observers,
- the inter-observer agreement when using the SDS,
- the relationship between the VAS, NRS and SDS.

2.2 Materials and Methods

This study was carried out between October 1994 and March 1995. The conditions of the study were designed to reflect clinical practice as closely as possible and thus represent the use of the scales in practice.

2.2.1 Pain measurement scales

Each observer used the VAS, NRS and SDS to record the pain they believed the animals to be experiencing. When using the SDS, the observers picked the term that best described

the animal's pain; 'No pain', 'Mild pain', 'Moderate pain', or 'Severe pain'. These terms were then translated into a numerical score of 0 to 3 where 'No pain' = 0 and 'Severe pain' = 3. When using the NRS the observers were asked to choose a number between 0 and 10 that best represented the pain experience of the animal being examined, where 0 was defined as 'No Pain' and 10 as 'Pain could not be worse'. The VAS used was a 100mm horizontal line bounded with two vertical 10mm lines, labelled as 'No Pain' on the left hand side and 'Pain could not be worse' on the right hand side. When using the VAS, the observer placed a mark on the horizontal line at a point that they believed represented the animal's pain intensity. The pain score was the distance from the 'No Pain' end and the observer's mark, measured to the nearest 1mm.

2.2.2 Animals

A total of 50 animals were included in the study. All had undergone surgery at Glasgow University Veterinary Hospital. No restrictions were placed on the age, breed, sex or surgical procedures undergone by the animals and all received analgesics according to standard clinical practice. Twenty-five animals were assessed 1 hour after the end of surgery (Group 1). Sixteen of these animals plus an additional 25 animals were assessed between 21 and 27 hours after the end of surgery, giving a total of 41 (Group 2). Thus, 16 animals were assessed on both occasions, i.e. immediately following surgery and again on the following day (Group 3).

2.2.3 Observers

The four observers who took part in this study were all qualified veterinary surgeons working at the Glasgow University Veterinary School. All observers had postgraduate qualifications in veterinary anaesthesia and were experienced in the management of pain in animals. Three of the observers assessed pain in the animals in the immediate post-surgical period and on the day following surgery (Groups 1, 2 and 3). The fourth observer carried out assessments only on the day following surgery (Group 2 only).

2.2.4 Examination procedure

On each day, each animal was examined four times over the course of 1 hour (at 20 minute intervals). The observers were informed of the surgical procedure the animal had undergone and when the surgery had taken place. Each animal was examined using the same procedure, which was carried out by one investigator (LH) and was watched by the

observers. Firstly, the animal was called by name and beckoned from outside the kennel. The kennel door was then opened and where possible, the investigator entered the kennel. The animal was approached and where practical, led out of the kennel on a leash. Once outside the kennel the animal was walked for a distance of approximately 10m up and down the hospital ward. The animal was encouraged to stand still while gentle even pressure was applied to the area approximately 4cm around the surgical wound. The animal was then returned to the kennel and asked to sit or lie down. In cases where the animal could not be led out and walked the examination procedure was carried out within the kennel. The procedure lasted between 4 and 5 minutes. The animal's behaviour was watched closely throughout by the observers who then allocated pain scores using the VAS, NRS and SDS. The three scales were presented to the observers in random order and, to minimise possible carry-over between the scales, the observers were asked to use the scales in the order presented. If an observer believed an animal to be suffering an unacceptable degree of pain, the veterinary nursing staff were notified, and additional analgesic therapy was administered. All such cases were excluded from any statistical analyses.

2.2.5 Statistical methods

The statistical analysis of all data collected in this study was carried out using SAS for Windows version 6.12 and Minitab for Windows version 10.

The variability in the pain scores was explored using summary statistics and graphical methods. These investigations provided an initial picture of the variability in the pain scores both between observers and over time. Graphical methods were also used to explore the relationships between the three scales.

The three groups of animals included in this study provided information pertaining to the performance of the VAS, NRS and SDS in three different situations: specifically in the immediate post-operative period, some time after surgery (i.e. the following day) and over an extended period (i.e. between the day of surgery and following day). The scores obtained using the VAS and NRS were treated as continuous, and the generalizability of the scales was explored for each group. The performance of the SDS was explored by assessing the agreement between observers when assessing pain in the same animal.

2.2.5.1 Generalizability of the VAS and NRS

To calculate the generalizability coefficients it is necessary to first calculate the variance associated with each of the random factors thought to contribute to variability in the subjects' scores and therefore included in the generalizability study. The components of variance for each random factor in the generalizability study were calculated by fitting an appropriate random effects model to the observed data. From this model the expected means squares could be decomposed into the appropriate components of variance and then used to calculate the generalizability coefficients. The effects included in the random effects models reflect the facets explored in the generalizability study, i.e. the facets of differentiation, generalizability and the fixed facets, as appropriate.

Random effects models with the appropriate factors were fitted to the NRS and VAS scores. For groups 1 and 2 the facet of differentiation was the dogs. The facet of generalization was the time or the observers depending on which of these factors was being examined. Consequently, the fixed facet was either time or observer, dependent on which was treated as the facet of generalizability. The model fitted to the scores observed in groups 1 and 2 is shown in Equation 2.1.

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{jk} + \beta\gamma_{jk} + \varepsilon_{ijk} \qquad \text{Equation 2.1}$$

where X_{ijk} : Score allocated by observer *i*, to dog *j*, at time *k*

α_i : Random effect of observer *i* distributed $N(0,\sigma^2_{obs})$

β_j : Random effect of dog *j*, distributed $N(0,\sigma^2_{dog})$

γ_k : Random effect of time *k*, distributed $N(0,\sigma^2_{time})$

$\alpha\beta_{ij}$: Random interaction between observer *i* and dog *j*, distributed $N(0,\sigma^2_{obs*dog})$

$\alpha\gamma_{ik}$: Random interaction between observer *i* and time *k*, distributed $N(0,\sigma^2_{obs*time})$

$\beta\gamma_{jk}$: Random interaction between dog *j* and time *k*, distributed $N(0,\sigma^2_{dog*time})$

ε_{ijk} : Random error effect, distributed $N(0,\sigma^2_{\varepsilon})$

i=1 to 3 (group 1) or 4 (group 2), *j*=1 to 25 (group 1) or 41 (group2), and *k*=1 to 4

From the random effects model the expected mean squares were decomposed into the appropriate components of variance (Table 2.1). These coefficients were used to calculate the generalizability of the VAS and NRS both over observers and time (Equation 2.2 and Equation 2.3).

Table 2.1: Decomposition of expected mean squares into components of variance from the random effects model. Factors fitted in the model included, dog effect, observer effect, time of assessment and appropriate interactions. The model was fitted to pain measurement scores for two groups of dogs assessed by four observers at four time points.

Source	Degrees of Freedom	Expected Mean Square
Observer	I-1	$\sigma_\varepsilon^2 + K\sigma^2_{obs*dog} + J\sigma^2_{obs*time} + JK\sigma^2_{obs}$
Dog	J-1	$\sigma_\varepsilon^2 + K\sigma^2_{obs*dog} + I\sigma^2_{dog*time} + IK\sigma^2_{dog}$
Time	K-1	$\sigma_\varepsilon^2 + I\sigma^2_{dog*time} + J\sigma^2_{obs*time} + IJ\sigma^2_{time}$
Observer*Dog	(I-1)(J-1)	$\sigma_\varepsilon^2 + K\sigma^2_{obs*dog}$
Observer*Time	(I-1)(K-1)	$\sigma_\varepsilon^2 + J\sigma^2_{obs*time}$
Dog*Time	(J-1)(K-1)	$\sigma_\varepsilon^2 + I\sigma^2_{dog*time}$
Error	(I-1)(J-1)(K-1)	σ_ε^2

Equation 2.2: Generalizability over observers

$$G_{obs} = \frac{\sigma^2_{dog} + \sigma^2_{dog*time}}{\sigma^2_{dog} + \sigma^2_{dog*obs} + \sigma^2_{dog*time} + \sigma^2_{\varepsilon}}$$

Equation 2.3: Generalizability over time

$$G_{time} = \frac{\sigma^2_{dog} + \sigma^2_{dog*obs}}{\sigma^2_{dog} + \sigma^2_{dog*obs} + \sigma^2_{dog*time} + \sigma^2_{\varepsilon}}$$

The mixed effects model fitted to group 3 included an additional factor to indicate the day on which the assessment was carried out, i.e. day of surgery or the following day. The factor day was included in the model as a random effect and the interactions between the factor day and other factors (subject, observer and time) were treated as random interactions (Equation 2.4).

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + \alpha\beta_{ij} + \alpha\gamma_{jk} + \alpha\delta_{il} + \beta\gamma_{jk} + \beta\delta_{jl} + \gamma\delta_{kl} + \varepsilon_{ijk} \quad \text{Equation 2.4}$$

where X_{ijk}: Score allocated by observer i, to dog j, at time k

α_i : Random effect of observer i distributed $N(0,\sigma^2_{obs})$

β_j : Random effect of dog j, distributed $N(0,\sigma^2_{dog})$

γ_k : Random effect of time k, distributed $N(0,\sigma^2_{time})$

δ_l : Random effect of day l, distributed $N(0,\sigma^2_{day})$

$\alpha\beta_{ij}$: Random interaction between observer i and dog j, distributed $N(0,\sigma^2_{obs*dog})$

$\alpha\gamma_{ik}$: Random interaction between observer i and time k, distributed $N(0,\sigma^2_{obs*time})$

$\alpha\delta_{kl}$: Random interaction between observer i and day l, distributed $N(0,\sigma^2_{obs*day})$

$\beta\gamma_{jk}$: Random interaction between dog j and time k, distributed $N(0,\sigma^2_{dog*time})$

$\beta\delta_{jl}$: Random interaction between dog j and day l, distributed $N(0,\sigma^2_{dog*day})$

$\gamma\delta_{kl}$: Random interaction between time k and day l, distributed $N(0,\sigma^2_{time*day})$

ε_{ijk} : Random error effect, distributed $N(0,\sigma^2_{\varepsilon})$

i=1 to 3 (group 1) or 4 (group 2), j=1 to 25 (group 1) or 41 (group2), k=1 to 4
l=1 to 2.

When this model was fitted the expected mean squares were decomposed into components of variance for each of the random effects using an extension to the decomposition described in Table 2.1 and coefficients examining the generalizability of the scales over time, observers and day of assessment were calculated (Equation 2.5 to Equation 2.7).

Equation 2.5: Generalizability over observers

$$G_{obs} = \frac{\sigma^2_{dog} + \sigma^2_{dog*day} + \sigma^2_{dog*time} + \sigma^2_{dog*day*time}}{\sigma^2_{dog} + \sigma^2_{obs*dog} + \sigma^2_{dog*time} + \sigma^2_{dog*day} + \sigma^2_{obs*dog*time} + \sigma^2_{obs*dog*day} + \sigma^2_{dog*time*day} + \sigma^2_{\varepsilon}}$$

Equation 2.6: Generalizability over time

$$G_{time} = \frac{\sigma^2_{dog} + \sigma^2_{obs*dog} + \sigma^2_{dog*day} + \sigma^2_{obs*dog*day}}{\sigma^2_{dog} + \sigma^2_{obs*dog} + \sigma^2_{dog*time} + \sigma^2_{dog*day} + \sigma^2_{obs*dog*time} + \sigma^2_{obs*dog*day} + \sigma^2_{dog*time*day} + \sigma^2_{\varepsilon}}$$

Equation 2.7: Generalizability over days

$$G_{day} = \frac{\sigma^2_{dog} + \sigma^2_{obs*dog} + \sigma^2_{dog*time} + \sigma^2_{obs*dog*time}}{\sigma^2_{dog} + \sigma^2_{obs*dog} + \sigma^2_{dog*time} + \sigma^2_{dog*day} + \sigma^2_{obs*dog*time} + \sigma^2_{obs*dog*day} + \sigma^2_{dog*time*day} + \sigma^2_{\varepsilon}}$$

2.2.5.2 Agreement between observers when using the SDS

The pain scores obtained from the SDS were categorical in nature and consequently could not be investigated using the generalizability theory methods described above. For this scale, it was more appropriate to examine the differences between observers using log-linear models and to explore the agreement between observers by calculating Cohen's Kappa coefficient (Cohen 1960).

The loglinear models allowed the degree of association between the SDS score and observer, time, day and interactions between these factors to be explored. The loglinear model can be thought of as and extension to cross tabulations and can be used to explore the association between factors. In particular, whether the level of one factor influences the distribution of observed frequencies across another factor. For example, in a loglinear model an interaction between SDS and observer would indicate whether the distribution of SDS scores was associated with observer, and therefore whether the scores differed between the observers.

Hierarchical loglinear models were fitted; the full, saturated model fitted to data from groups 1 and 2 is shown in Equation 2.8. When fitted, all non-significant terms were removed and the model was refitted. Terms were only removed if they were not included in any significant higher order terms. The model fitted to data collected from group 3

included an additional main effect term and relevant interactions to account for the day on which the assessment took place.

$$\log(\theta_{ijk}) = \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} \qquad \text{Equation 2.8}$$

where θ_{ijk} : probability of score i, being assigned by observer j at time k

α_i : SDS category i

β_j : effect of observer j

γ_k : effect of time k

$\alpha\beta_{ij}$: interaction between observer j and SDS category i

$\alpha\gamma_{ik}$: interaction between time k and SDS category i

$\beta\gamma_{jk}$: interaction between observer j and time k

$\alpha\beta\gamma_{ijk}$:interaction between SDS category i, observer j and time k

i= 0 to 3 (no, mild, moderate or severe pain), j =1 to 3 (group 1) or 4 (group2), and k= 1 to 4

Following this analysis the agreement between any two observers when using the SDS was examined by calculating Cohen's Kappa statistic for each pair of observers. The Kappa statistic provided an indication of the chance-corrected agreement between two observers (Cohen, 1960). The structure of the scores given by the observers is shown in Table 2.2.

The probability of the two observers agreeing on any score (P_o) is the proportion of times they agree out of all the assessments:

$$P_o = \frac{\sum_i n_{ii}}{n_{**}}$$

P_o does not account for agreement by chance. Conditional on the marginal distribution, the probability of agreement by chance (P_c) is as follows:

$$P_c = \sum_i \left(\frac{n_{i*}}{n_{**}}\right)\left(\frac{n_{*i}}{n_{**}}\right)$$

When the two observers are in complete agreement P_o=1 and the maximum possible probability of agreement being better than chance is 1-P_c. The observed probability of agreement being better than chance is P_o-P_c. The Kappa statistic is the ratio of the observed probability of agreement being better than chance to the maximum possible probability of agreement. The Kappa statistic is defined in Equation 2.9, and can be interpreted as a measure of chance-corrected agreement.

Table 2.2: Data structure detailing level of agreement between two observers using the simple descriptive scale to measure pain in a group of dogs. Possible scores are no pain, mild, moderate or severe pain.

		Observer 2				
		No Pain	Mild Pain	Moderate Pain	Severe Pain	Total
Observer 1	No Pain	n_{11}	n_{12}	n_{13}	n_{14}	n_{1*}
	Mild Pain	n_{21}	n_{22}	n_{23}	n_{24}	n_{2*}
	Moderate Pain	n_{31}	n_{32}	n_{33}	n_{34}	n_{3*}
	Severe Pain	n_{41}	n_{42}	n_{43}	n_{44}	n_{4*}
	Total	n_{*1}	n_{*2}	n_{*3}	n_{*4}	n_{**}

where n_{ij}= number of occasions score i was allocated by observer 1 and j by observer 2.

Table 2.3: Interpretation of Cohen's Kappa statistic values, used to explore the level of agreement between two observers when using a categorical scale.

K	Strength of Agreement
0	No better than chance
0.01-0.2	Slight
0.21-0.4	Fair
0.41-0.6	Moderate
0.61-0.8	Substantial
0.81-0.99	Almost perfect
1	Perfect

$$\kappa = \frac{P_o - P_c}{1 - P_c}$$ Equation 2.9

The interpretations of the Kappa coefficient values are given in Table 2.3 (Cohen, 1960).

2.2.5.3 Relationship between the VAS, NRS and SDS

The relationship between the VAS and NRS was investigated by plotting the scores obtained using each scale and fitting a linear regression model to the data. This model allowed the relationship between the VAS and NRS to be compared to a theoretical model. If it can be assumed that the VAS and NRS are used to measure the same aspects of pain, i.e. they measure the same attribute, then it is reasonable that an animal with a 'no pain' score on one scale would be allocated the same score on the other scale. Thus, the relationship between the two scales at this unique point is known. The validity of this assumption was explored by examining whether the intercept of the regression model was significantly different from zero.

To examine whether the relationship between the VAS and NRS scores differed between the observers, the adequacy of a single regression line was investigated. A multiple regression model with individual regression lines for each observer was fitted, and the equality of the slope parameter across observers tested. Non-significant differences would indicate that the relationship between the two scales was not dependent on the observer carrying out the assessment. Thus, the relationship between the scales could be quantified and the scales used interchangeably. This hypothesis was examined using standard F-tests.

The relationship between the SDS and the other two scales was investigated graphically and by calculating confidence intervals for the scores on the VAS and NRS for each level of the SDS.

2.3 Results

Of the 50 animals examined, one from group 1 was thought to be experiencing an unacceptable degree of pain during the examination and was removed from the study and all analyses of the pain scales.

The breeds of animals included in the study and the surgical procedures they underwent are given in Appendix 1. Age and gender distributions of the animals are detailed in Table 2.4. The dogs were aged between 4 months and 15 years old, 27 males were included and 23 females. The demographic details indicated that group 3 had an imbalance in the distribution of the sexes (12 males and 4 females).

2.3.1 Exploratory analysis of scores

The summary statistics for the VAS, NRS and SDS scores allocated by the observers to each group (Table 2.5) indicated that the mean VAS and NRS scores differed between the observers for all three groups. In particular, observer 2 had a higher mean score and larger standard deviation than the other observers, when using the VAS and NRS.

The summary statistics for the VAS, NRS and SDS scores over time indicated that the mean VAS and NRS scores were similar over the 4 time points within each day (Table 2.6 and Table 2.7). This suggested there was less variability over time than between observers.

The VAS and NRS scores for group 3 (Table 2.7) indicated than the pain scores were lower on the day following surgery than in the immediate post-operative period.

These results suggested that the VAS, NRS and SDS scores contained little variability over time within a single day but greater variability between the assessment days, and between observers. This indicated that the reliability of the scales between observers may be questionable.

2.3.2 Investigation of the generalizability of the VAS and NRS

Random effects models were fitted to the scores as described earlier and the resulting components of variance for each factor are shown in Table 2.8.

The results for the VAS and NRS show that the components of variance relating to the 'Dog' effect was large relative to the other factors in the models, for all three groups. Hence, much of the variability occurred between the animals. The results also indicated that, relative to the other factors, the components of variance for the factor 'Observers' was low, but 'Dog*Observer' was high. This suggested that when averaging over all the animals there was little variability between the observers, but when the observers' scores were examined per animal, the variability was higher.

Table 2.4: Demographic details for the three groups of dogs (50 animals) included in a study carried out to compare the performance of the VAS, NRS and SDS when used to measure post-surgical pain.

Study Group	Variable	Statistic	Value
Group 1		N	25
	Age (yr)	Min	0.75
		Mean	7.08
		Max	15
	Sex	Male	15
		Female	10
Group 2		N	41
	Age (yr)	Min	0.33
		Mean	6.25
		Max	15
	Sex	Male	24
		Female	17
Group 3		N	16
	Age (yr)	Min	0.75
		Mean	6.75
		Max	15
	Sex	Male	12
		Female	4

Table 2.5: Summary statistics for pain intensity scores allocated using the VAS, NRS and SDS to three groups of dogs (49 animals). Group 1 was assessed by three observers immediately after surgery, group 2 by four observers on the following day and group 3 by three observers on both of these occasions.

Group	Scale	Statistic	Observer 1	Observer 2	Observer 3	Observer 4
1	VAS	Mean	17.4	21.3	-	15.3
		SD	16.0	17.4	-	14.8
		Min	0	0	-	0
		Max	60	67	-	58
1	NRS	Mean	2.0	2.3	-	1.9
		SD	1.7	1.8	-	1.8
		Min	0	0	-	0
		Max	6	7	-	7
1	SDS	No Pain n (%)	23 (24)	21 (22)	-	29 (30)
		Mild n (%)	49 (51)	57 (59)	-	54 (56)
		Moderate n (%)	21 (22)	18 (19)	-	13 (14)
		Severe n (%)	3 (3)	0 (0)	-	0 (0)
2	VAS	Mean	12.6	27.2	13.5	16.7
		SD	13.5	22.6	15.6	17.0
		Min	0	0	0	0
		Max	57	94	64	61
2	NRS	Mean	1.4	2.9	1.3	2.1
		SD	1.4	2.3	1.4	2.0
		Min	0	0	0	0
		Max	6	9	7	7
2	SDS	No Pain n (%)	63 (38)	35 (21)	58 (35)	52 (33)
		Mild n (%)	82 (50)	66 (40)	78 (48)	79 (48)
		Moderate n (%)	19 (12)	57 (35)	28 (17)	30 (18)
		Severe n (%)	0 (0)	5 (3)	0 (0)	1 (1)
3	VAS	Mean	15.0	23.3	-	13.5
		SD	15.6	17.7	-	14.9
		Min	0	0	-	0
		Max	60	67	-	58
3	NRS	Mean	1.7	2.5	-	1.7
		SD	1.7	1.8	-	1.8
		Min	0	0	-	0
		Max	6	7	-	7
3	SDS	No Pain n (%)	43 (34)	31 (24)	-	47 (37)
		Mild n (%)	62 (48)	60 (47)	-	66 (52)
		Moderate n (%)	23 (18)	37 (29)	-	15 (12)
		Severe n (%)	0 (0)	0 (0)	-	0 (0)

Table 2.6: Summary statistics for pain intensity scores allocated by observers using the VAS, NRS and SDS to two groups of dogs at four time points 20 minutes apart. Group 1 was assessed by three observers immediately after surgery (24 animals), group 2 by four observers on the following day (41 animals).

Group	Scale	Statistic	0 min	20 min	40 min	60 min
1	VAS	Mean	19.1	18.1	16.0	18.7
		SD	17.0	16.7	15.5	15.9
		Min	0	0	0	0
		Max	67	58	51	64
1	NRS	Mean	2.2	2.1	1.9	2.1
		SD	1.8	1.8	1.7	1.8
		Min	0	0	0	0
		Max	7	7	6	7
1	SDS	No Pain n (%)	17 (24)	21 (29)	21 (29)	14 (19)
		Mild n (%)	42 (58)	37 (52)	37 (52)	44 (62)
		Moderate n (%)	13 (18)	13 (18)	13 (18)	13 (18)
		Severe n (%)	0 (0)	1 (1)	1 (1)	1 (1)
2	VAS	Mean	17.7	17.6	17.6	17.0
		SD	19.2	19.0	17.6	17.9
		Min	0	0	0	0
		Max	91	89	94	85
2	NRS	Mean	1.9	2.0	1.9	1.8
		SD	2.0	2.0	1.9	1.8
		Min	0	0	0	0
		Max	9	9	9	8
2	SDS	No Pain n (%)	53 (32)	52 (32)	52 (32)	53 (33)
		Mild n (%)	75 (46)	76 (46)	76 (46)	78 (48)
		Moderate n (%)	34 (21)	35 (21)	35 (21)	30 (18)
		Severe n (%)	2 (1)	1 (1)	1 (1)	2 (1)

Table 2.7: Summary statistics for pain intensity scores allocated by observers using the VAS, NRS and SDS to a group of dogs assessed immediately after surgery and on the following day. Sixteen dogs were assessed, by three observers at four time points each (20 minutes apart) on each day.

Day	Scale	Statistic	0 min	20 min	40 min	60 min
1	VAS	Mean	21.7	18.9	16.7	20.3
		SD	19.2	17.1	16.6	16.7
		Min	0	0	0	0
		Max	67	58	51	64
1	NRS	Mean	2.4	2.2	1.9	2.3
		SD	2.0	1.9	1.9	1.9
		Min	0	0	0	0
		Max	7	7	6	7
1	SDS	No Pain n (%)	14 (29)	15 (31)	15 (31)	10 (21)
		Mild n (%)	22 (46)	22 (46)	22 (46)	27 (56)
		Moderate n (%)	12 (25)	11 (23)	11 (23)	11 (23)
		Severe n (%)	0 (0)	0 (0)	0 (0)	0 (0)
2	VAS	Mean	13.5	14.6	15.9	16.3
		SD	16.8	16.0	14.5	15.3
		Min	0	0	0	0
		Max	57	67	51	49
2	NRS	Mean	1.6	1.8	1.7	1.8
		SD	1.8	1.6	1.5	1.6
		Min	0	0	0	0
		Max	7	7	5	5
2	SDS	No Pain n (%)	21 (44)	15 (31)	15 (31)	16 (33)
		Mild n (%)	19 (40)	26 (54)	26 (54)	24 (50)
		Moderate n (%)	8 (16)	7 (15)	7 (15)	8 (17)
		Severe n (%)	0 (0)	0 (0)	0 (0)	0 (0)

Table 2.8: Mean squares and components of variance derived from fitting random effects model to the VAS and NRS pain intensity scores assigned to 3 groups of dogs. Group 1 was assessed immediately after surgery, group 2 on the following day and group 3 on both of these occasions.

		VAS		NRS	
Group	Factor	Mean Square	Component of Variance	Mean Square	Component of Variance
1	Dog	1584.2	88.0	20.2	1.2
	Observer	876.4	4.3	3.9	0
	Time	128.6	0	1.2	0
	Dog*Observer	443.6	97.2	4.8	1.1
	Dog*Time	139.7	28.2	1.6	0.4
	Observer*Time	9.9	0	0.4	0
	Error	54.8	54.8	0.56	0.56
2	Dog	3068	159.2	32.4	1.66
	Observer	6539	39.2	78.3	0.47
	Time	66	0	1.7	0
	Dog*Observer	363	68.6	3.9	0.75
	Dog*Time	112	8.6	1.3	0.12
	Observer*Time	16	0	0.4	0
	Error	75.6	75.6	0.8	0.81
3	Dog	2710	67.5	32.2	0.81
	Observer	3589	21.2	28.7	0.11
	Time	75	0	0.70	0
	Day	1781	3.4	22.5	0.03
	Dog*Observer	554	41.6	6.19	0.47
	Dog*Time	124	0	1.49	0
	Dog*Day	752	37.3	8.82	0.46
	Observer*Time	6.8	0	0.08	0
	Observer*Day	541	4.51	10.16	0.11
	Time*Day	216	1.05	1.72	0
	Dog*Obs*Time	47	2.45	0.63	0.10
	Dog*Obs*Day	216	43.5	2.49	0.46
	Dog*Time*Day	130	29.3	1.49	0.36
	Obs*Time*Day	78	2.24	1.24	0.05
	Error	42	42	0.42	0.42

Note: Obs = Observer

The results suggested that the generalizability of the VAS and NRS over observers could be expected to be low, since the components of variance for Dog*Observers interaction was high; this was explored formally by calculating generalizability coefficients.

The components of variance also indicated low variability over time and between observers over time, which indicated that, when averaged over the animals, the scores observed did not vary greatly over time. This suggested that the scales could be expected to show good generalizability over time. The component of variance relating to the 'Dog*Day' effect in group 3 was large, as was the 'Dog*Observer*Day' effect, for both the VAS and NRS. This suggested that pain scores changed between the two days over which the study was carried out.

The generalizability coefficients shown in Table 2.9 demonstrated that the generalizability over observers was low for both the VAS and NRS (between 0.27 and 0.49) and that the NRS performed very slightly better than the VAS. These low levels of generalizability across observers indicated that the variability between the observers was contributing significantly to the variability in the VAS and NRS scores. The generalizability of the VAS and NRS scores over time, within each day, was much higher than was seen between observers (between 0.69 and 0.73).

The generalizability of the scores over the two days of the study, calculated using the data from group 3, indicated that the VAS and NRS scores did not show a great deal of generalizability between the assessment days (0.42 and 0.45 for VAS and NRS respectively).

2.3.3 Agreement between observers using the SDS

The summary statistics constructed for the SDS scores showed that the distribution of the scores was similar over observers and time (Table 2.5). This gave an initial indication that the scale was used consistently by the observers. The significance of each effect included in the log-linear models is shown in Table 2.10.

These results indicated that there were significant differences between the SDS scores allocated by the observers for groups 2 and 3 but not for group 1. None of the other factors showed any significant effects. The Cohen's Kappa coefficients calculated are shown in Table 2.11**Error! Reference source not found.**. All the coefficients fell within the category indicating fair agreement between

Table 2.9: Generalizability Coefficients calculated over observers and time for the pain scores collected using the VAS and NRS. Group 1 was assessed by three observers immediately after surgery, group 2 by four observers on the following day and group 3 by three observers on both of these occasions. Generalizability over days was also calculated for animals assessed on both days.

Group	Coefficient	VAS	NRS
1	Generalizability over Observers	0.43	0.49
	Generalizability over Time	0.69	0.71
2	Generalizability over Observers	0.27	0.53
	Generalizability over Time	0.73	0.72
3	Generalizability over Observers	0.51	0.53
	Generalizability over Time	0.72	0.71
	Generalizability over Days	0.42	0.45

Table 2.10: Results of fitting log linear models to the SDS pain intensity scores observed when used to measure post-operative pain in 3 groups of dogs. Group 1 was assessed by four observers immediately after surgery, group 2 by four observers on the following day and group 3 by three observers on both of these occasions. Table shows p-values corresponding to each factor in the model.

Factor	Group 1	Group 2	Group 3
Observer*SDS	0.528 ×	0.000 ✓	0.005 ✓
Time*SDS	0.822 ×	0.979 ×	0.822 ×
Day*SDS	-	-	0.102 ×
Obs*Time*SDS	1.000 ×	1.000 ×	0.989 ×
Obs*Day*SDS	-	-	0.064 ×
Time*Day*SDS	-	-	0.587 ×

× Indicates a non-significant factor
✓ Indicates a significant factor
- Indicates term not included in model
Obs : Observer

Table 2.11: Cohen's Kappa coefficients for agreement between four observers when using the SDS to assess post operative pain in 3 groups of dogs. Group 1 was assessed immediately after surgery, group 2 on the following day and group 3 on both of these occasions. Note that observer 3 did not assess groups 1 or 3.

Observers	Group 1	Group 2	Group 3
1 and 2	0.244	0.313	0.233
1 and 3	-	0.368	-
1 and 4	0.257	0.356	0.319
2 and 3	-	0.211	-
2 and 4	0.299	0.306	0.321
3 and 4	-	0.339	-

the observers as defined in Table 2.3. This suggested that despite its simplicity, the SDS was not used consistently between the observers taking part in this study.

2.3.4 Relationship between VAS and NRS

Scatter plots of the VAS and NRS scores suggested that there was a great deal of variability in the relationship between the scores observed using the two scales (Figure 2.1, Figure 2.2 and Figure 2.3). These plots also suggested that, as pain increased, the observed scores lay slightly below the line of equality.

For each group a linear regression model was fitted to the data observed (Table 2.12). Analysis of these models indicated that the intercept was not significantly different from zero for any of the three groups. Thus, the assumption that no pain on one scale was consistently recorded as no pain on the other scale was valid.

Assuming a zero intercept in the regression model the next step was to explore whether a multiple regression model, with separate regression lines for each observer, provided a better fit to the data and therefore whether the relationship between the two scales was dependent on observer. The fit of the multiple regression model was compared to the simple regression model using standard F-tests. The results of this analysis (Table 2.13) indicated that there were significant differences in the relationship between VAS and NRS scores, over the four observers. Therefore, the relationship between the scales was dependent on the observer and they could not be used interchangeably.

2.3.5 Relationship between the SDS, NRS and VAS

The relationship of the SDS scores with the NRS and VAS scores was initially examined graphically by constructing plots of the observed NRS and VAS scores corresponding to each category in the SDS (Figures 2.4 - 2.9). These plots indicated that, with the exception of the 'no pain' category there was a great deal of variability in the observed VAS and NRS scores corresponding to each SDS category. This suggested that the relationship between these two scales and the SDS was weak and variable.

Following this subjective approach the scores corresponding to each SDS category were examined by calculating summary statistics and 95% confidence intervals for the mean VAS and NRS scores for each SDS category (Table 2.14 and Table 2.15). These results showed that for each level of the SDS, the mean VAS and NRS scores were consistent across the groups but also that wide ranges of scores were observed for each category.

Figure 2.1: Plot of the VAS and NRS pain intensity scores allocated to 25 dogs one hour after the end of surgery. Each dog was assessed by three observers at four time points. Line shows the theoretical line of equality between the NRS and VAS pain intensity scales. Note: NRS scores are jittered.

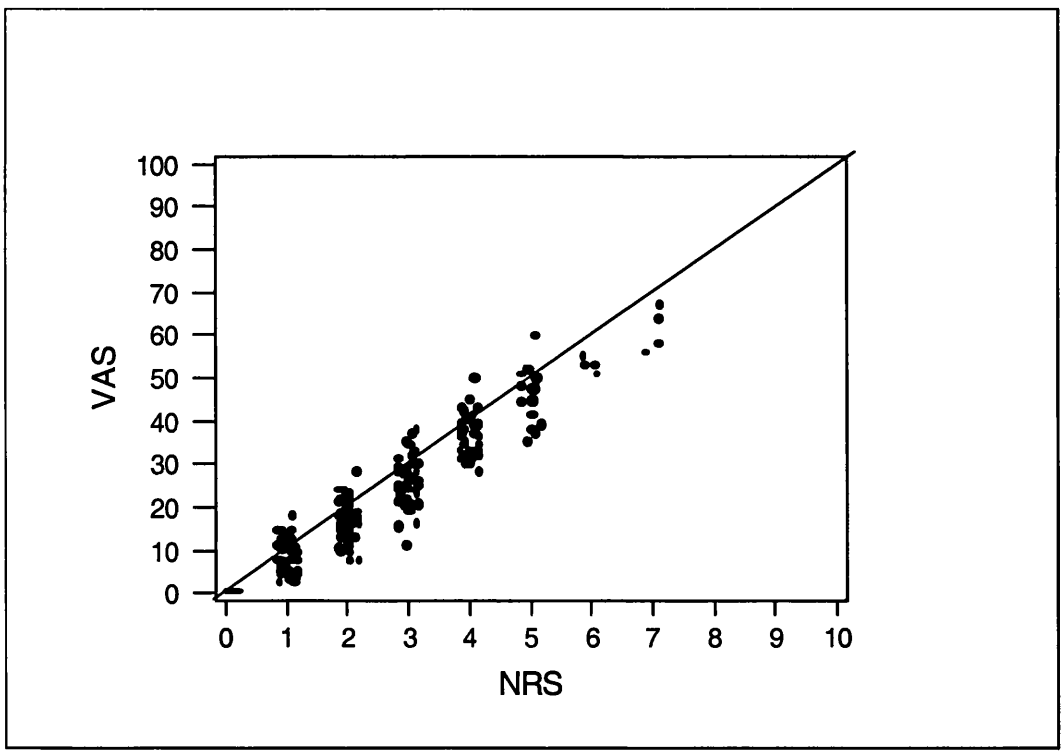


Figure 2.2: Plot of the VAS and NRS pain intensity scores allocated to 41 dogs on the day following of surgery. Each dog was assessed by four observers at four time points. Line shows the theoretical line of equality between the NRS and VAS pain intensity scales. Note: NRS scores are jittered.

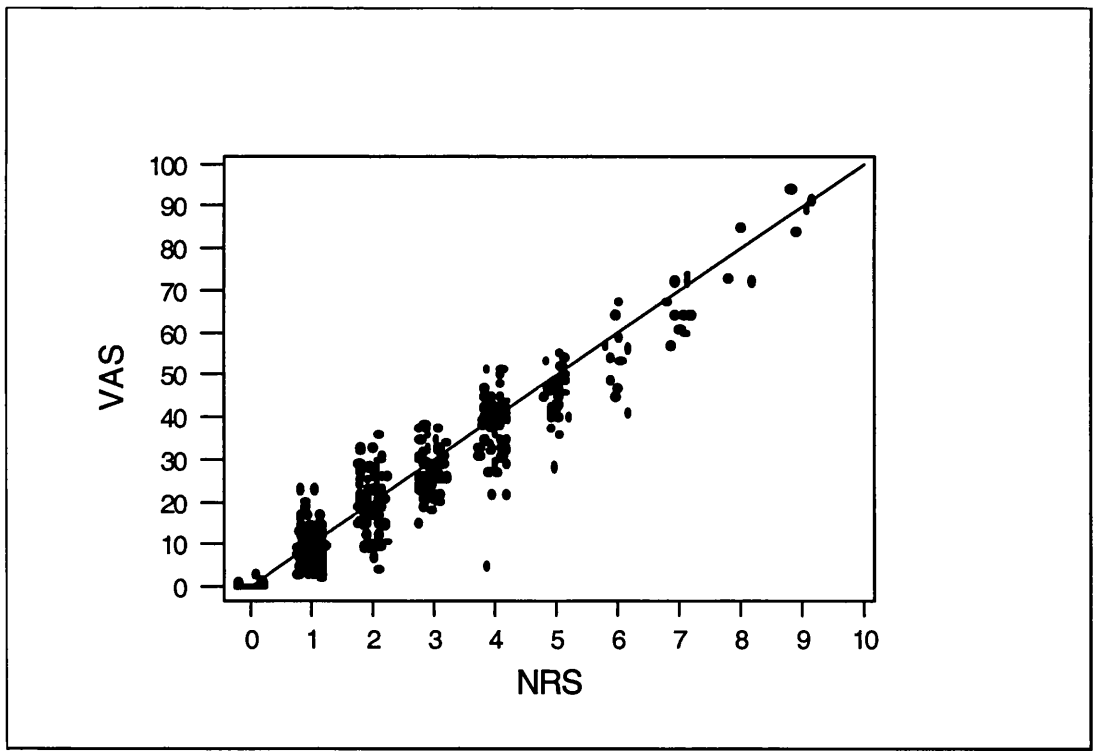


Figure 2.3: Plot of the VAS and NRS pain intensity scores allocated to 16 dogs one hour after the end of surgery and on the day following surgery. Each dog was assessed by three observers at four time points on each day. Line shows the theoretical line of equality between the NRS and VAS pain intensity scales. Note: NRS scores are jittered.

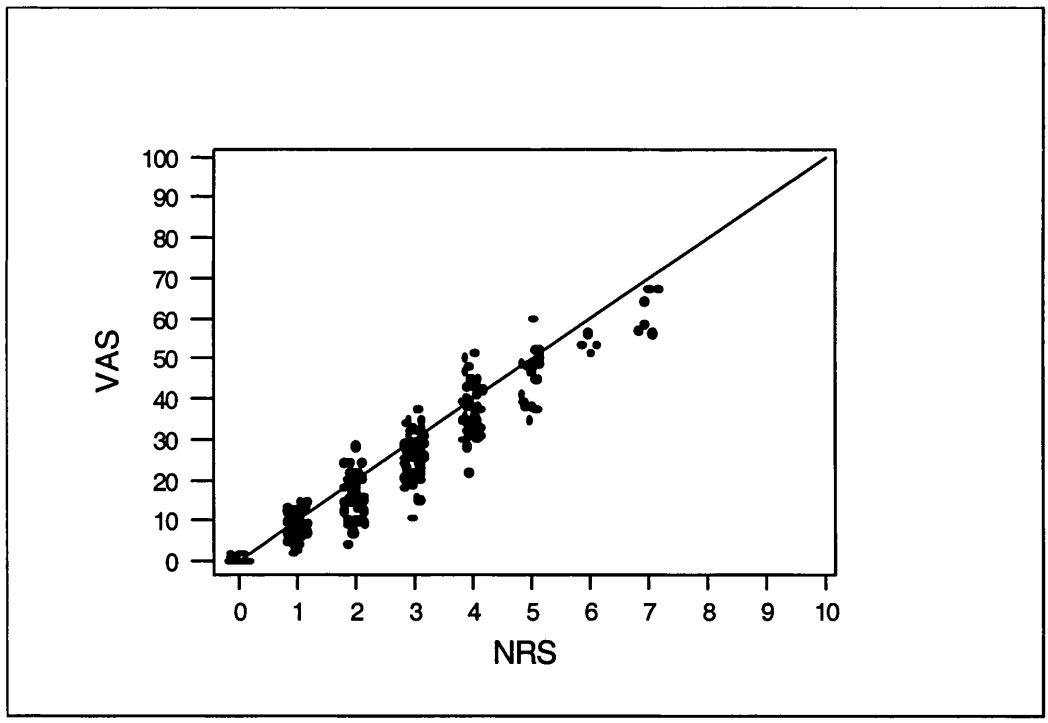


Table 2.12: Results of the linear regression model which was fitted to examine the relationship between the VAS and NRS pain intensity scores allocated to three groups of dogs by a number of observers. Group 1 was assessed immediately after surgery, group 2 on the following day and group 3 on both of these occasions. Table shows parameter estimates and p-value for significance of the intercept.

Group	Intercept Estimate	Slope Estimate	p-value for Intercept
1	-0.40	8.96	0.26
2	-0.09	9.28	0.72
3	-0.34	9.04	0.30

Table 2.13: Results of a multiple regression analysis carried out to examine the consistency of relationship between the VAS and NRS pain intensity scores over a number of different observers. Models were fitted to pain scores allocated to three groups of dogs (Group 1 was assessed immediately after surgery, group 2 on the following day and group 3 on both of these occasions). Table shows estimates of slope parameter for each observer, test statistic to compare the values and corresponding p-value.

Group	Observer 1	Observer 2	Observer 3	Observer 4	F statistic	p-value
1	8.85	9.41	-	8.14	17.8	0.001
2	9.01	9.1	10.48	8.24	46.1	0.001
3	9.06	9.46	-	8.06	28.2	0.001

Figure 2.4: Plot of the VAS and SDS pain intensity scores allocated to 25 dogs one hour after the end of surgery (Group 1). Each dog was assessed by three observers at four time points, 20 minutes apart. Note: SDS scores are jittered.

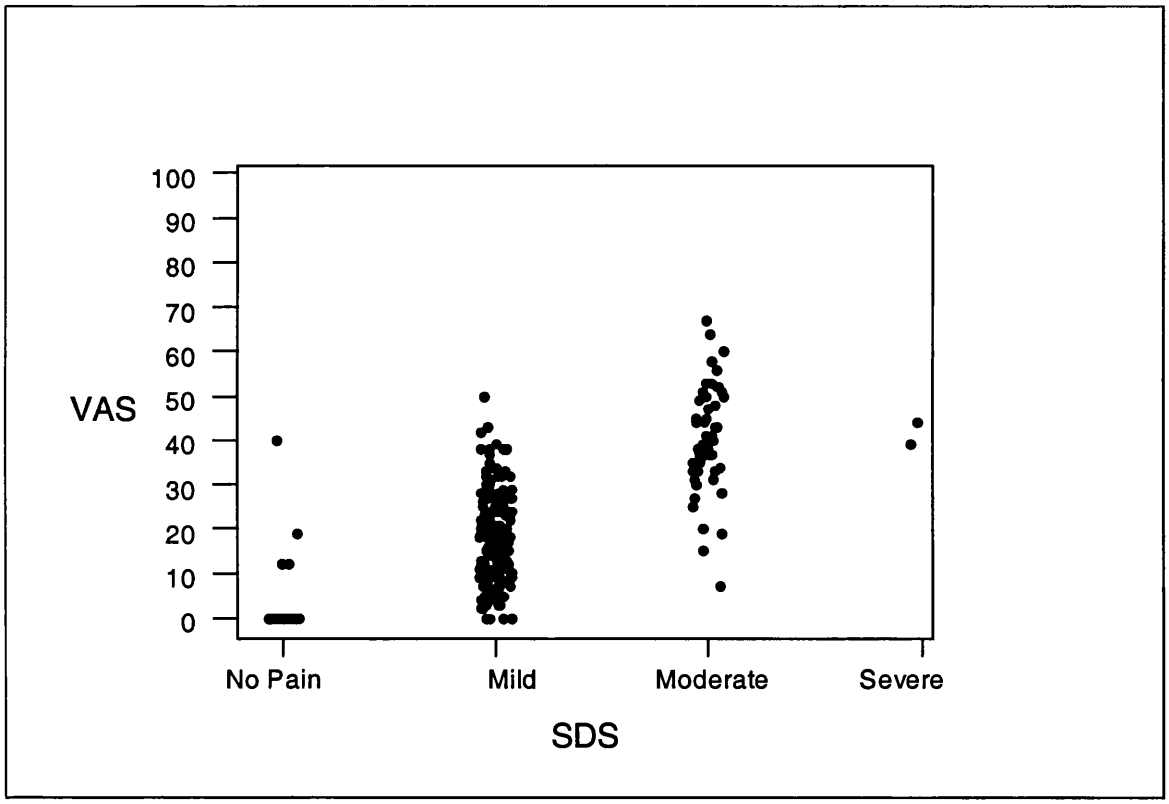


Figure 2.5: Plot of the VAS and SDS pain intensity scores allocated to 41 dogs on the day following surgery (Group 2). Each dog was assessed by four observers at four time points, 20 minutes apart. Note: SDS scores are jittered.

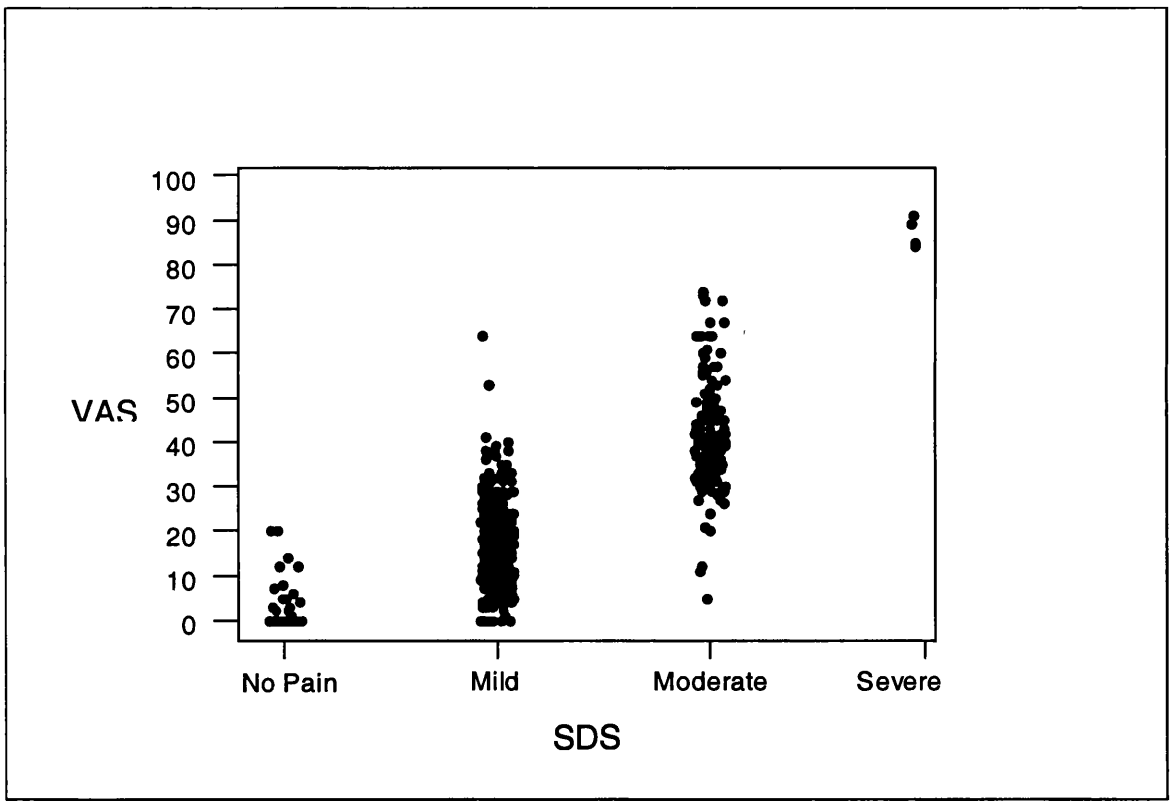


Figure 2.6: Plot of the VAS and SDS pain intensity scores allocated to 16 dogs one hour after surgery and on the day following surgery (Group 3). Each dog was assessed by three observers at four time points, 20 minutes apart. Note: SDS scores are jittered.

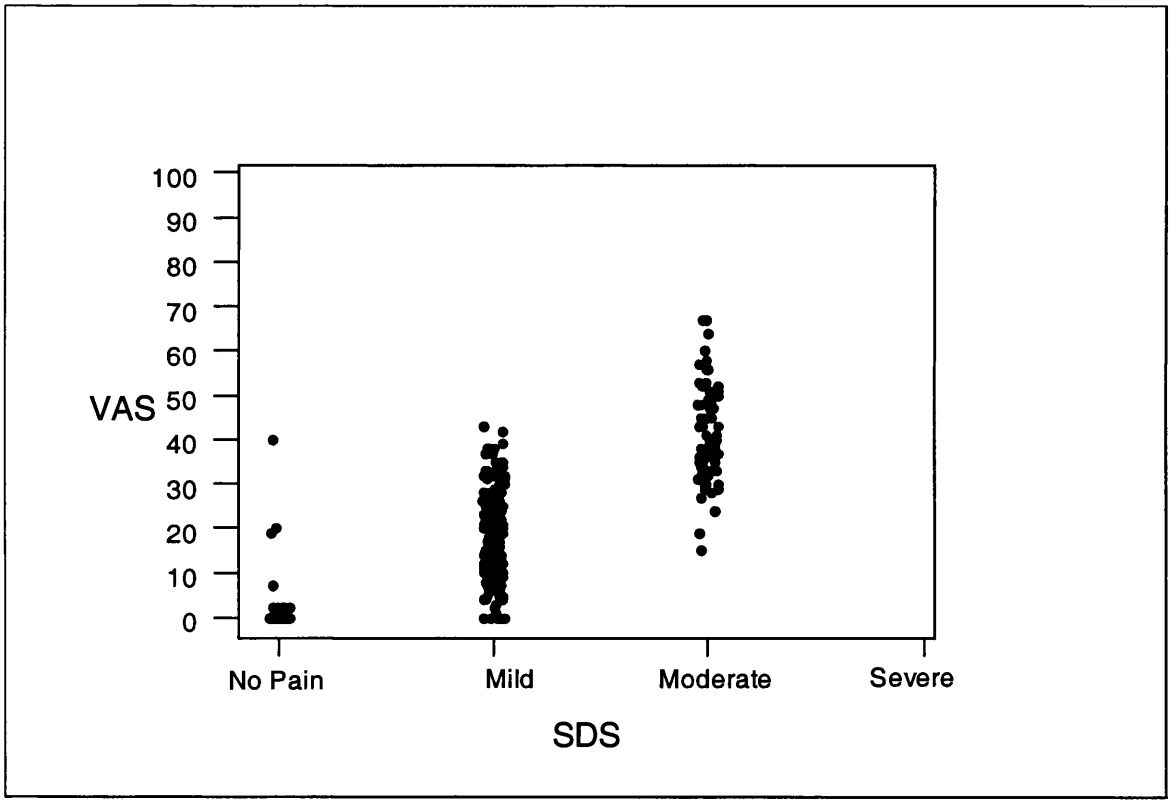


Figure 2.7: Plot of the NRS and SDS pain intensity scores allocated to 25 dogs one hour after the end of surgery (Group 1). Each dog was assessed by three observers at four time points, 20 minutes apart. Note: NRS and SDS scores are jittered.

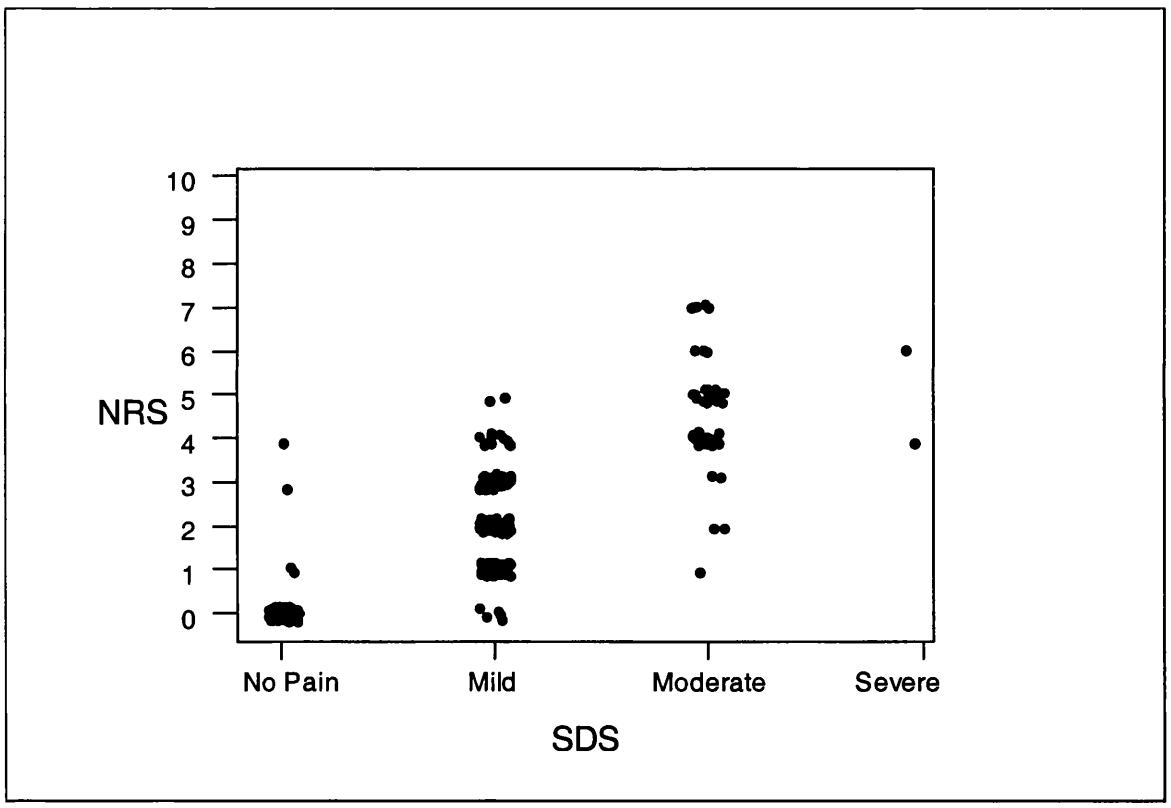


Figure 2.8: Plot of the NRS and SDS pain intensity scores allocated to 41 dogs on the day following surgery (Group 2). Each dog was assessed by four observers at four time points, 20 minutes apart. Note: NRS and SDS scores are jittered

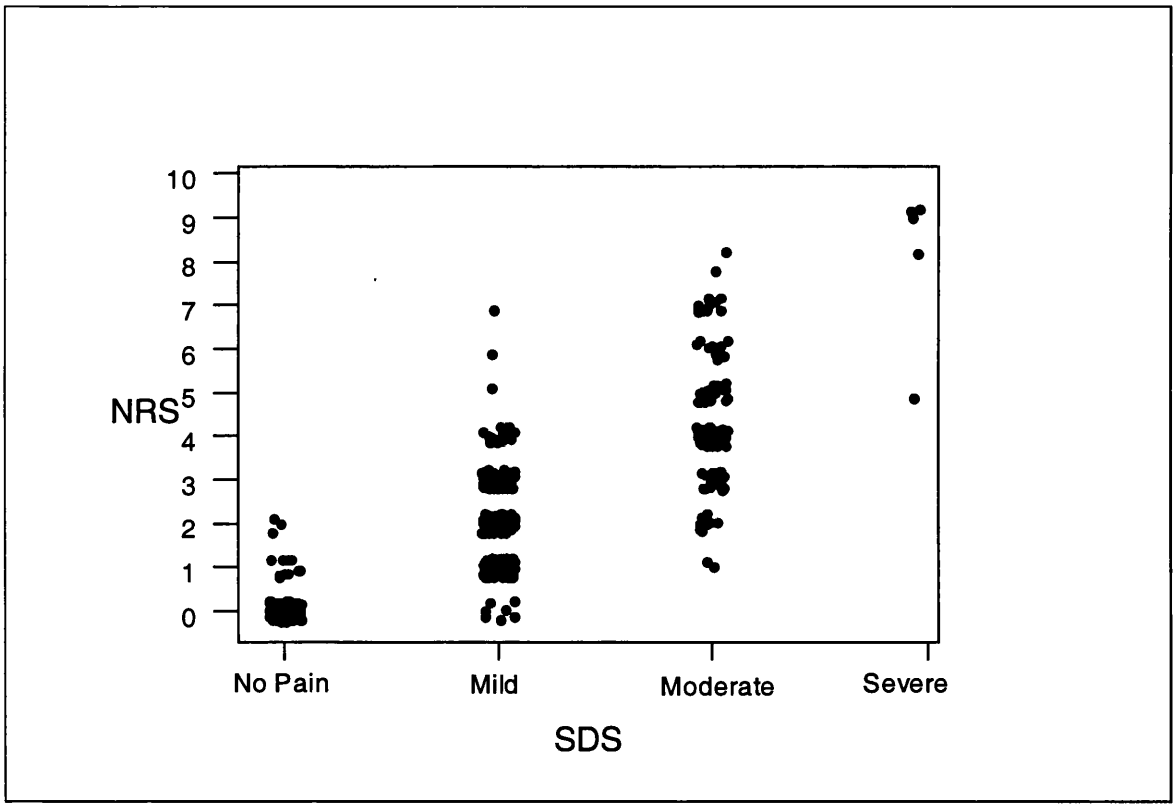


Figure 2.9: Plot of the NRS and SDS pain intensity scores allocated to 16 dogs one hour after the end of surgery and on the day following surgery (Group 3). Each dog was assessed by three observers at four time points, 20 minutes apart. Note: NRS and SDS scores are jittered

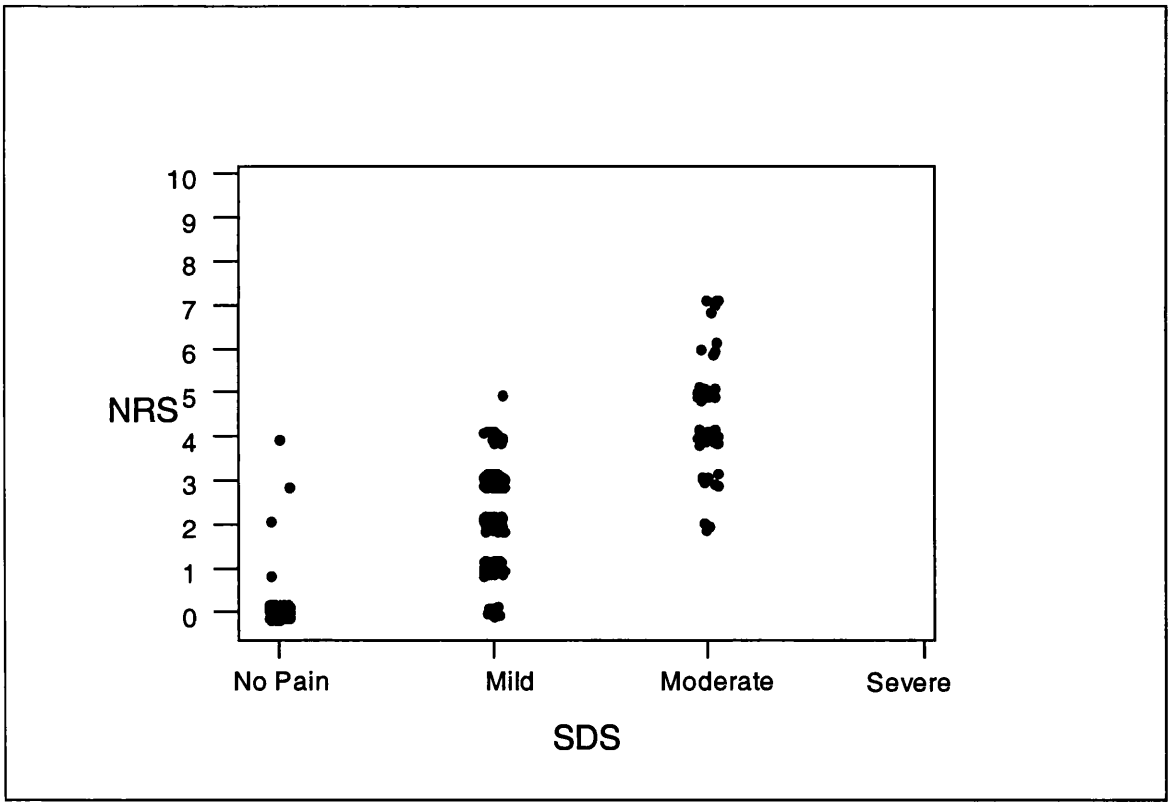


Table 2.14: Summary statistics and 95% confidence intervals for the mean VAS pain intensity scores allocated to three groups of dogs by a number of observers, split by SDS pain intensity category. Group 1 was assessed immediately after surgery, group 2 on the following day and group 3 on both of these occasions.

Group	Statistic	SDS=mild	SDS=moderate	SDS=severe
1	Mean	18.0	40.3	46.0
	SD	10.28	12.09	8.19
	Min	0	7	39
	Max	50	67	55
	C.I.	(16.4, 19.6)	(37.0, 43.7)	(25.7, 66.3)
2	Mean	17.0	42.2	81.3
	SD	10.07	12.61	18.18
	Min	0	5	45
	Max	64	74	94
	C.I.	(15.8, 18.1)	(40.0, 44.4)	(62.3, 100.4)
3	Mean	18.2	41.4	-
	SD	10.06	10.59	-
	Min	0	15	-
	Max	43	67	-
	C.I.	(16.7, 19.6)	(38.9, 43.8)	-

Table 2.15: Summary statistics and 95% confidence intervals for the mean NRS pain intensity scores allocated to three groups of dogs by a number of observers, split by SDS pain intensity category. Group 1 was assessed immediately after surgery, group 2 on the following day and group 3 on both of these occasions.

Group	Statistic	SDS=mild	SDS=moderate	SDS=severe
1	Mean	2.1	4.5	5.0
	SD	1.03	1.20	1.00
	Min	0	1	4
	Max	5	7	6
	C.I.	(1.9, 2.3)	(4.1, 4.8)	(2.5, 7.5)
2	Mean	1.9	4.4	8.2
	SD	1.04	1.41	1.60
	Min	0	1	5
	Max	7	8	9
	C.I.	(1.8, 2.0)	(4.2, 4.6)	(6.5, 9.8)
3	Mean	2.2	4.4	-
	SD	1.04	1.13	-
	Min	0	2	-
	Max	5	7	-
	C.I.	(2.0, 2.3)	(4.2, 4.7)	-

2.4 Discussion

Pain is a personal experience, unique to every individual (Lasanga, 1964; Savage, 1970; Chapman *et al*, 1985), yet to date no effective objective methods of measuring the intensity of clinical pain have been developed (Morton and Griffiths, 1985). Health care professionals must rely solely on subjective assessments, the performance of which must be investigated before they can be accepted as satisfactory. This study was designed to assess the generalizability of the VAS and NRS, agreement between observers when using the SDS, and the relationship between these three scales, when used to measure clinical pain in dogs. This would allow the accuracy with which the scales reflect the patient's pain state to be assessed and would indicate whether the scales could be used interchangeably.

Performance of the VAS, NRS and SDS, in the assessment of pain in humans has been investigated in a number of studies. The majority of the published results support the reliability of the scales when used to measure human pain, which is contrary to the findings of the study reported here. This may be due partly to differences in the use of the VAS, NRS and SDS in human and veterinary medicine. The scales are primarily used as self-reporting tools in human medicine, so the study designs and analysis methods differ from those used in the veterinary setting.

When exploring the reliability of the VAS, Revill *et al* (1976) asked a number of patients to record pain they had experienced one month prior to the assessment. This study demonstrated non-significant differences (p-value=0.112) between scores recorded on three occasions (at first assessment, 5 minutes later, then 24 hours later). Thus, the authors supported the reliability of the VAS in measuring pain. This study examined only *recalled* pain and did not explore the reliability of the VAS when used to record *current* pain. In addition, the statistical methods used did not explore the variability in the scores on each occasion, only the change between assessments.

The reliability of the VAS when used to measure pain due to dyspepsia was also explored by Nyrén *et al* (1987). Each patient's pain was assessed 4 times per day for 7 days. The authors found correlations of 0.42 to 0.6 between scores allocated on consecutive days and so indicated their support for the reliability of the VAS. However, correlation coefficients of this order of magnitude would tend to suggest a large degree of variability in the relationship between the variables. Thus, the evidence provided about the reliability of the VAS in this study may not be as robust as the authors indicated.

The reliability of the VAS and an NRS-type scale when used to record pain following total knee replacement was explored by Boeckstyns and Backer (1989). The authors examined the agreement between successive applications of the scales using Kappa statistics and found good agreement over time (K ranged from 0.35 to 0.88). It is worth noting that the VAS scores were recorded only to the nearest 1cm, so provided scores between 0 and 10 rather than in the range of 0 to 100 that is typically seen. In their analysis, the investigators explored agreement over time and assumed agreement between time points when the second assessment score was within 2 units of the first assessment score. For example, if one patient scored 4 on the first assessment then any score between 2 and 6 at the second assessment was be said to be in agreement. It was possible for the second score to take any value within 50% of the scale and still indicate agreement with the first. It is therefore likely that the results of this study reflect an optimistic view of the reliability of the VAS and NRS.

Although the VAS, NRS and SDS are accepted as reliable when used to measure pain in humans, it is clear that the measurement properties of the scales should not be accepted without question. Many of the results obtained in humans have simply been extrapolated to animals without considering the validity of such a practice. The issues raised regarding the use of these scales in human medicine must also be addressed in the veterinary field if their use in animals is to be ratified. Thus, exploring the measurement properties of the scales in a veterinary setting is critical.

In veterinary medicine, researchers have investigated the agreement between observers when using the VAS (Reid and Nolan, 1991; Welsh *et al*, 1993). Reid and Nolan (1991) investigated the agreement between two observers using the VAS to score pain and sedation in dogs which had been given analgesic drugs post operatively. The study demonstrated no significant disagreement between the two observers and there was little variability in the scores. Based on this, the authors concluded that the VAS was an appropriate scale to use. It is worth noting that the period of assessment was limited to < 6 hours post-anaesthesia, a time when one might expect less variability in a dog's behaviour because of lingering effects of anaesthesia and surgery. It is interesting that, in the study discussed in this chapter, this was also the period where the variability between animals and between observers was smallest. This could be attributed to the fact that the dogs were less responsive because of the prolonged effects of drugs administered during surgery. In retrospect, this may not be the optimum time to make post-operative assessments, and this should be considered in the design of future studies.

Inter-observer reliability of the VAS and NRS has been examined in the context of assessing lameness in sheep (Welsh *et al*, 1993). Although the study was not concerned directly with the assessment of pain, lameness was likely to be closely linked with pain as the lameness was due to footrot, which is thought to be a painful condition. One obvious advantage of assessing lameness rather than pain is that it is easier to define the end-points of the scale (when an animal is sound and when the affected leg is not being used). The intra-observer variability and the relationship between the VAS and NRS when used for scoring lameness were examined. No significant variability between the observers was found, however the investigators concluded that the VAS and NRS were not interchangeable since the relationship between them was not unique.

In contrast to both of the studies discussed above, the study described in this chapter demonstrated poor consistency among observers in the VAS, NRS and SDS scores. The results here suggest that the observers used the scales differently and that the variability between the observers was large relative to the variability between the animals. Reasons for the differences in the conclusions drawn from these studies may be partially due to differences in the design of the study, the statistical analyses used and the attribute under investigation, i.e. pain rather than lameness. Welsh *et al* (1993) and Reid and Nolan (1991) both used two observers, who had been trained specifically to use the scales. In the study reported here up to 4 observers took part and none were given any specific training in the use of the VAS, NRS or SDS, although all were familiar with the scales. It is possible that training the observers improved the agreement in the studies reported by Welsh *et al* (1993) and Reid and Nolan (1991). Universal definitions or instructions for use of the VAS, NRS or SDS are lacking, and so it must be assumed that practitioners who use these scales for assessment of pain would also lack such documentation. Since this study was designed to assess the performance of the scales in a clinical setting, definitions were not assigned to the categories of the SDS, and guidelines were not given for use of the VAS or NRS. Imposition of restrictive definitions or guidelines was not thought to be appropriate as it may have altered the scale properties, and the results would not reflect the performance of these scales when used in practice.

To examine the differences between the observers' scores Welsh *et al* (1993) used paired t-tests. Similarly, Reid and Nolan (1991) used a Mann-Whitney test to compare the median scores allocated by observers. These methods allowed a comparison of the average scores allocated by the two observers, which differs from investigating the generalizability of the scale across observers. Generalizability examines the variability between the observers

relative to the variability between the subjects included in the study, whereas the subject variability is not the focus of the analysis when the mean or median scores are compared. Accordingly, two observers could be shown to have similar average scores but could differ greatly in their assessments of individual animals. This would not be highlighted by simply comparing the average scores but would be demonstrated when calculating generalizability coefficients.

The statistical methods used to calculate the generalizability coefficients for the VAS and NRS scores assume that both of these methods provide continuous interval level measurement for pain. The validity of this assumption may be questioned when exploring the NRS as this scale provides only 11 possible pain scores. However, the observed scores had approximately normal distributions and the random effects model provided an acceptable fit to the data. Therefore, the models used were thought to be appropriate and provided insight into the variability within the scores that would not be possible using categorical methods such as the loglinear modelling techniques used for the SDS. An additional consideration that should be made when examining the results presented is that the data included in the analysis of group 3 were also included in the analyses of groups 1 and 2. This means that the 3 groups are not independent and the results from the analyses should not be regarded as being independent. This complication in the interpretation of the study results may have been avoided if all patients had been examined on the day of surgery and the following day, and therefore the study would have been balanced. This would have allowed the generalizability over time to be examined between the two study days and within each study day using a single model rather than the piecewise analysis discussed. The limitations of carrying out a study in a working veterinary hospital resulted in the imbalance in study design discussed here; it should be considered for the future that the use of balanced design should be used whenever possible as it provides increased efficiency and power.

In the study discussed, no restrictions were placed on the type of surgical procedure, age, sex or breed, hence the dogs included formed a heterogeneous group. This design was used deliberately to investigate the performance of the scales under conditions that were as close as possible to those seen in clinical practice. It is acknowledged that investigation of reliability or generalizability of measurement scales using a very heterogeneous sample can cause an apparent improvement in the performance of the scale in question, compared to the generalizability in a more homogeneous sample (Streiner and Norman, 1995). Despite the heterogeneity of the sample used in this study, both the VAS and NRS demonstrated

poor generalizability over observers. Thus, the scale may perform less well in a research environment where the sample of animals included for study may be more homogeneous.

Investigation of the relationship between the VAS and NRS indicated that there was a linear correspondence between the two scores, but that this differed between the observers. Hence, the VAS and NRS should not be used interchangeably since the relationship between the scales is dependent on the observer. These results agree with those demonstrated by Welsh *et al* (1993) where the VAS and NRS were also said to be non-interchangeable.

Graphical investigation of the relationship between the 3 scales showed that each category of the SDS corresponded to a wide range of VAS and NRS scores, but that the mean scores corresponding to each category were reasonably consistent across groups. Thus, while there did seem to be a somewhat consistent relationship between the SDS and the other two scales, on any single occasion the SDS score could be associated with a very wide range of scores allocated using either the VAS or SDS. Hence, the relationship between the scales appears to be weak and the methods should not be used interchangeably.

Similar investigations into the relationship between the VAS, NRS and SDS have been undertaken when the scales were used to measure pain in humans. In particular, Downie *et al* (1978) explored the relationship between the three scales when used by patients with rheumatic disease. The scales were highly correlated (correlation coefficients between 0.62 and 0.91) when used consecutively. However, the authors noted that each category of the SDS was associated with a wide range of NRS and VAS scores, with the VAS being the more variable. For example, a single category of the SDS was associated with up to 4 points on NRS and 90 points on the VAS. Concurring results were reported by Collins *et al* (1997). This investigation constituted a meta analysis, combining data collected from 11 different studies examining the effects of analgesics. The investigation was concerned with identifying the range of VAS scores corresponding to 'moderate pain' on a SDS scale. The results indicated that no appropriate range of VAS scores existed since the moderate pain category of the SDS was associated with such a wide range of VAS scores (between 0 and 100). Thus, the relationships between the VAS, NRS and SDS demonstrated when measuring human pain reflect the results seen in the measurement of animal pain.

Generally, the VAS, NRS and SDS are used separately. To compare the inter-observer variability, the assessments in the study reported here were performed under identical conditions for each of the 3 scales. To ensure that the pain intensity of the dog being

assessed was constant while the scores were recorded on each scale, the pain assessments were performed concurrently. The design of this study may have allowed the observers to make subconscious comparisons between the three scales when used simultaneously. To minimise this effect the scales were presented to the observers in a random order. However, it is possible that in these circumstances the relationships between the scales could be stronger than if the scales were not used simultaneously. The possibility of such a 'halo effect' was also reported by Downie *et al* (1978) when a similar design was used. Thus, these results should be interpreted with some caution. The generalizability of the scales would be unaffected by this aspect of the design as the analyses examined the variability between observers and over time, not between scales. The implications of these considerations are that the true nature of the relationship between the VAS, NRS and SDS when used separately could be weaker than has been demonstrated here. Thus, any attempts to quantify the relationships between the points on each scale would not be informative and the three scales should be used separately.

The scale that is most appropriate for use in veterinary medicine has not been determined as all three methods explored have been shown to have deficiencies. Results of studies in humans indicate that other factors should be taken into account when making decisions regarding the scale to use. A patient scoring his or her own pain can only differentiate between a maximum of 39 distinct levels of pain (Grossi *et al*, 1983), which suggest that it is safe to assume that observers assessing pain in animals would not improve upon this. In addition, it has been shown that when the VAS is used in humans the visual acuity of subjects can affect the accuracy of VAS scores (as much as ±7 mm on a 100mm line). This was demonstrated when Revill *et al* (1976) asked subjects to place a mark 20% along a VAS line and then to reproduce this mark on a number of occasions. The error over the different occasions was estimated and the results suggested that the large number of possible scores could give a false impression of sensitivity when using the VAS. It has also been indicated that the SDS lacks sensitivity when used in humans (Joyce *et al*, 1975; Seymour, 1982). In an investigation of pain in chronic inflammatory disease, the SDS did not demonstrate any analgesic effect, although this was evident in the VAS scores. Similar results were observed when exploring the analgesic effect of aspirin in dental pain (Seymour, 1982). It is reasonable to assume that the lack of sensitivity of the SDS would be demonstrated, and perhaps even be exacerbated, when the scale is used to assess pain in animals, since the assessments are made by a third party. Hence, it could be argued that, in veterinary as well as human medicine, the NRS provides a suitable compromise between

the over-interpretation, which can be a feature of the VAS, and the lack of sensitivity that has been reported in the SDS.

In conclusion, the results of the study reported here indicated that the three subjective scales (VAS, NRS and SDS) used to assess pain in dogs in a clinical setting demonstrated unacceptable inconsistencies among observers. However, the generalizability of the scales across observations within a short time span, in this case one hour, was satisfactory. Thus, when using any of these scales, careful consideration must be given to study design, such as the timescale over which the study is carried out and the number of observers involved. The use of such subjective pain measurement scales in dogs is limited for both the research worker and the care provider.

3. Development of a Composite Measurement Pain Scale

3.1 Introduction

Traditionally clinicians in human and veterinary medicine have placed great faith in simple tangible entities that can be used to measure single attributes (Wright and Feinstein, 1992). However, according to the literature, pain is a complex, personal experience that cannot be observed directly and, in common with other unobservable attributes such as depression or intelligence, it can only be assessed indirectly (Chapman *et al*, 1985).

The use of subjective measurement scales such as the VAS to measure pain has been questioned since they measure only one aspect of pain, i.e. intensity (Chapman, 1976). Although the problems associated with subjective rating scales when used to measure pain in humans have been demonstrated (Revill *et al*, 1976; Gracely, 1980; Atkinson *et al*, 1982; Carlsson, 1983; Linton, 1983; Kaiko *et al,* 1983), such scales are also commonly used for pain research in veterinary medicine. However, the results presented in Chapter 2 indicate that subjective rating scales such as the VAS, NRS and SDS do not provide reliable measurement of post-operative pain in dogs.

Pain can be thought of as a complex construct in the same way as intelligence or disability (Johnston, 1998). The use of a *composite measurement* tool to tap into such a construct is well recognised in the psychometric literature (Wright and Feinstein, 1992; Streiner and Norman, 1995). The benefits of these methods over subjective tools are widely accepted within the psychometric and medical communities (Guyatt *et al*, 1992; Nunnally and Bernstein, 1994; Streiner and Norman, 1995). Composite measurement scales have been shown to possess greater overall reliability and validity than subjective methods (Wright and Feinstein, 1992; Nunnally and Bernstein, 1994; Johnston, 1998), and one composite measurement scale that is widely used in human medicine is the McGill Pain Questionnaire (MPQ; Melzack, 1975).

The background and properties of the MPQ have been discussed in Chapter 1. The MPQ has been shown to be valid and reliable when used to measure pain in humans and the multidimensional nature of the method has been upheld (Reading, 1982; Prieto and Geisinger, 1983; Reading, 1983; Doctor, 1995).

Of the few pain measurement scales discussed in the human medical or veterinary medical literature, the MPQ is one of the few to have been constructed empirically rather than being

based on the developers' judgement. The initial list of expressions included in the MPQ was derived from a list of pain descriptors discussed by Dallenbach (1939) and from a review of the medical literature. The resulting expressions were then divided into 3 major classes and 16 subclasses. The classification was verified by presenting the list of expressions in their categories to two groups of subjects who were asked to indicate whether they agreed or disagreed with the proposed classification. This investigation demonstrated that the subjects agreed with the classification of the expressions. The intensity of pain associated with each expression was assessed by consulting approximately 180 additional subjects. These subjects comprised groups of students, doctors and patients from diverse social and cultural backgrounds. Each subject indicated the intensity of pain (mild to excruciating) they believed to be associated with each expression. These assessments allowed the pain intensity associated with each expression to be estimated. This study demonstrated excellent agreement in the relative pain intensity for expressions within each subclass. The consistency of the intensity of pain seen in this study prompted the development of the MPQ as a formal tool.

The MPQ is more complex than the subjective rating scales and therefore cannot be used as widely. Nevertheless, it is used throughout human medicine and the composite measurement approach it takes has been accepted in preference to the subjective methods previously used (Coste *et al*, 1995; Streiner and Norman, 1995).

The veterinary community has also become aware of the benefits of taking a composite measurement approach to complex attributes such as pain. This is demonstrated by the recent publication of a number of composite pain measurement scales, as discussed in Chapter 1. The scales published include the numerical rating scale (Conzemius *et al,* 1997), the Colorado State University Veterinary Teaching Hospital Pain Scale (Hellyer and Gaynor, 1998) and the University of Melbourne Pain Scale (UMPS; Firth and Haldane, 1999). Although these scales take a composite measurement scale approach to pain in animals the authors provide very little or no detail regarding how the scales were derived and therefore do not provide any insight into the development of such tools for use in veterinary medicine.

The development of such composite measurement scales is not straightforward and guidelines for scale construction and exploration have been laid down in the psychological and health measurement literature (Coste *et al*, 1995; Streiner and Norman, 1995).

Although the approach taken in constructing a composite measurement scale is at the discretion of the developers, psychometric research has provided guidelines for scale development (Nunnally and Bernstein, 1994; Coste *et al*, 1995; Streiner and Norman, 1995). Despite the apparent acknowledgement and acceptance of these methods in the human medical literature, there has been little recognition of this work in the veterinary literature. The work undertaken in this chapter addresses this issue by using psychometric methods to develop a new composite measurement scale for assessing pain in dogs.

The first step in developing a new composite measurement scale is to identify the attribute of interest and explore the performance of any existing measurement methods. When the investigator has demonstrated that these existing methods are unsatisfactory, items that may be included in the new composite measurement scale should be identified. The process of gathering this information is critical in the development of a scale, as it provides the foundations on which the scale is constructed. A number of techniques can be used to form an initial list of items, such as:

- examination of existing methods and literature,
- discussion with experts in the area of interest,
- discussion with patients, where appropriate and possible,
- clinical observation.

The aim of these methods is to provide a sound base from which a measurement scale can be developed and to ensure the content validity of the resulting scale.

The most appropriate items for inclusion in a measurement scale are identified from the bank of items collected in the first step. Certain items may not be appropriate for practical reasons, some may be difficult to interpret and could lead to confusion, or some items may be similar and their inclusion could result in duplication.

The relevance and ease of interpretation of the items can be judged by consulting a panel of experts, or where appropriate, through discussion with patients. Similarities between the items can be explored in a number of ways, for example, through discussion with experts or using statistical methods. Where items are found to be similar, one item may be removed completely or several may be combined to form a composite item.

Following selection of the items to be included in the measurement scale, a weighting scheme should then be devised. This allows an overall score to be calculated. A number of methods can be used to calculate weights, ranging from very simple subjective estimates to complex scaling models. This area of scale development is discussed in detail in Chapter 4.

The scale developed in this thesis was aimed at providing a valid and reliable pain measurement method for use in dogs. Although a number of composite measurement scales already exist in veterinary medicine, these have not utilised the psychometric principles of scale construction during their development. The construction of the new scale discussed here used methods similar to those described in the development of the MPQ, and so followed the guidelines laid down in the psychological and educational literature (Melzack and Torgerson, 1971; Melzack, 1975).

The primary objectives of the work detailed in this chapter were to:

- collect and collate information on the behaviours and physiological signs thought to be indicative of pain in the dog,
- investigate the degree of pain associated with the behaviours,
- examine the relationships between the items and hence the validity of these for use in a pain measurement scale,
- develop an examination procedure to facilitate the use of the pain measurement scale.

3.2 Materials and Methods, and Results

3.2.1 Collection and collation of behaviours and physiological signs relating to pain

Sixty-nine veterinary surgeons were contacted via Glasgow University Veterinary Hospital and through Continuing Professional Development courses. The veterinary surgeons contacted were thought to be familiar with pain responses and behaviours in dogs. Each was asked to 'list all signs indicative of pain (of any origin) in the dog'. A total of 279 different words and expressions were collected. From this bank of words and expressions, the most consistently cited behaviours and physiological signs were highlighted. An

independent research worker (P. Pawson) rationalised the list of behaviours and signs using the following criteria:

- expressions that described similar pain behaviours but were worded slightly differently, were replaced by a single expression, for example, 'lameness' and 'animal lame when walking' were replaced by 'lame',
- expressions that conveyed causes rather than signs of pain, for example, 'broken leg', were excluded,
- expressions that were very specific to one particular area or cause of pain, such as 'rubbing ear', were replaced by more generic expressions such as 'rubbing painful area',
- expressions that were vague and liable to considerable differences in interpretation were omitted, for example, 'lack of happiness',
- physiological parameters that could not be assessed quickly and easily in a clinical setting, such as blood cortisol levels, were omitted.

Following this rationalisation process, 47 expressions remained (39 behaviours and 8 physiological signs), and these were allocated into categories by the same independent research worker (Table 3.1).

3.2.2 Pain intensity assessment

To investigate the intensity of pain associated with each behaviour and sign, 75 practising veterinary surgeons were contacted via Glasgow University Veterinary School and through Continuing Professional Development courses. These veterinary surgeons were additional to the 69 contacted in the initial collection of expressions and signs. Each veterinary surgeon was provided with the list of behaviours and physiological signs included in each of the 10 categories (Table 3.1).

A 100mm VAS was placed beside each item which had the left and right ends defined as 'no pain' and 'pain could not be worse' respectively. Each veterinary surgeon was asked to 'place a mark on the VAS line at a point that represents the severity of pain implied by each word'. They were also asked to allocate a score to each expression using an NRS, between 0 (no pain) and 10 (pain could not be worse).

Table 3.1: 39 behavioural expressions and 8 physiological parameters regarded as being indicative of pain in dogs, collated from a list of 279 such expressions compiled by 69 practising veterinary surgeons. Categories to which the behaviours were assigned are underlined and shown in bold

Behaviours		
Demeanour	**Posture**	**Response to People**
anxious	curled up	aggressive to people
depressed	hunched up	fearful of people
distressed	rigid	indifferent to people
quiet	tense	sullen
Attention to Wound	**Mobility**	**Response to Touch**
biting wound	lame	crying when touched
chewing wound	slow/reluctant to rise	flinching when touched
licking wound	stiff	growling when touched
looking at wound	stilted	guarding when touched
rubbing/scratching wound	unwilling/unable to rise	snapping when touched
Vocalisation	**Response to food**	**Activity**
crying	disinterested in food	restless
groaning	eating hungrily	sitting/lying still
howling	picking at food	sleeping
screaming	rejecting food	
whimpering		

Physiological Signs		
tachycardia	pyrexia	muscle spasm
panting	salivation	dilated pupils
tachypnoea	trembling	

Of the 75 forms completed and returned, 3 were completed incorrectly and these data were omitted from all subsequent analyses. The pain intensity scores assigned were investigated to assess the pain associated with each item, and to identify the relationships between expressions. The results presented in this chapter relate to investigation of the VAS scores only. Results from the analysis of the NRS, which differ from those for the VAS, will be highlighted and discussed.

3.2.3 Exploratory investigation of pain scores

The behaviours and signs within each category were investigated by constructing dotplots of the VAS scores to illustrate the distribution of scores within the category (Figure 3.1 shows such dotplots for the Posture category). A wide range of VAS scores was observed for each expression, for example, 'hunched' was allocated scores of 20 to 90 and 'curled up' from 0 to 76. The scores for some expressions were similar, however there was a great deal of variability within the data. The distributions of the scores followed normal distributions for many of the expressions, however, since the variability in the scores was large the distributions were flattened. Some behaviours such a 'screaming' had skewed distributions. Dotplots for all of the expressions are given in Appendix 2.

Summary statistics of the VAS scores allocated to each expression (Table 3.2) indicated large variability in the VAS scores, for example, 'salivating' had a standard deviation of 25.0 and a mean of 49.2. In addition, VAS scores of 0-100 and 0-99 were associated with 'aggressive' and 'salivating' respectively. Within the categories some expressions were shown to be very similar in relation to pain intensity, for example, 'lame' and 'slow/reluctant to rise' had the same mean pain score (49.3). However, in the Vocalisation category 'scream' had a much higher mean score (91.4) than the other expressions (55.3 to 66.1). These summary statistics indicated that the relationships between the expressions should be explored further using formal analysis methods and that simplification of the scale through combining behaviours with similar pain intensity was feasible.

The ordering of the expressions within each category was also investigated. The VAS scores allocated within each category, by each veterinary surgeon, were ranked and summary statistics were calculated based on the ranks (Table 3.3). Behaviours with similar mean ranks indicated that the expressions were perceived as implying a similar intensity of pain, relative to the other behaviours in the category. For example, in the Posture category the expressions 'hunched' and 'tense' had similar mean rank scores (2.45 and 2.65 respectively). This indicated that both expressions were considered to be between second

Figure 3.1: Dotplots of VAS scores indicating the pain intensity associated with the behaviours in the Posture category, specifically 'curled up', 'hunched', 'rigid' and 'tense'. Scores allocated by 75 veterinary surgeons, assuming that behaviours were exhibited because of pain.

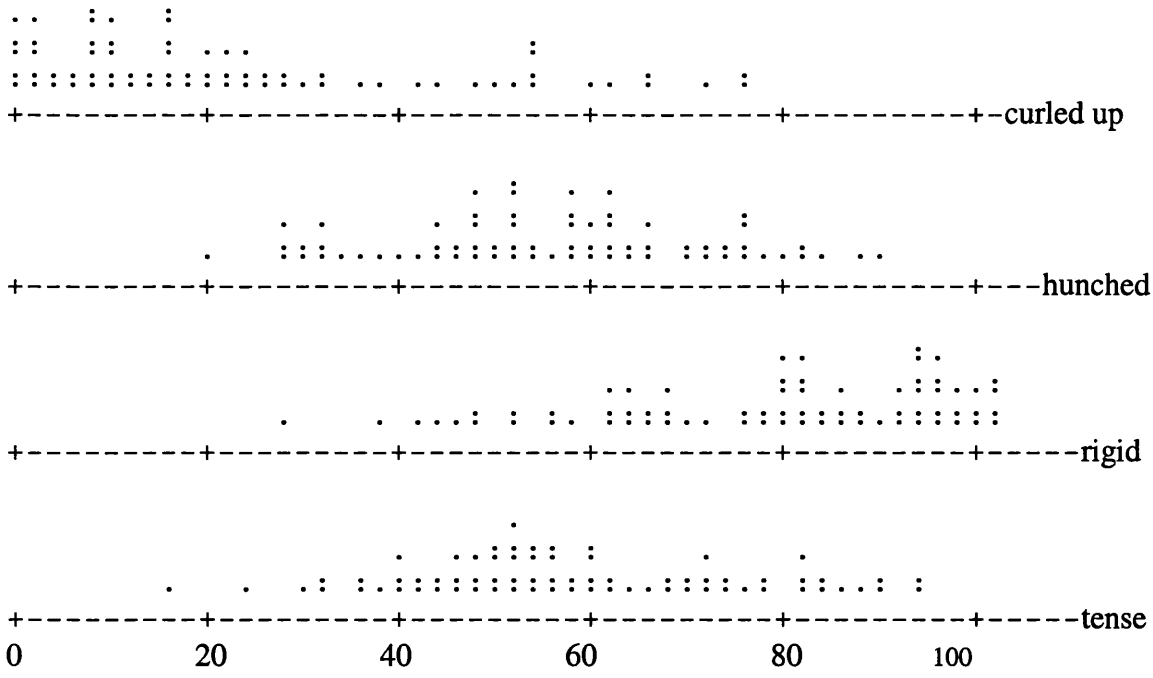


Table 3.2: Summary statistics for VAS scores indicating the pain intensity associated with each behaviour and physiological sign when observed in a dog, as allocated by 72 practising veterinary surgeons using a VAS defined from 0 to 100.

	Mean	S.D.	Minimum	Maximum
Demeanour				
anxious	38.3	20.0	0	88
depressed	43.8	23.4	1	98
distressed	67.2	21.1	1	100
quiet	27.5	23.6	0	89
Response to people				
aggressive	67.6	22.4	0	100
fearful	46.8	19.7	3	87
indifferent	24.3	23.7	0	97
sullen	33.1	19.4	1	77
Posture				
curled	23.7	21.1	0	76
hunched	55.5	16.5	20	90
rigid	75.9	18.4	26	100
tense	58.1	17.6	16	94
Activity				
restless	44.2	20.1	8	87
sit/lie still	35.1	24.0	0	90
sleeping	13.1	15.1	0	86
Vocalisation				
crying	55.3	23.6	2	94
groaning	65.1	21.1	14	100
howling	66.1	26.1	0	100
screaming	91.4	11.8	52	100
whimpering	58.6	21.0	3	100
Mobility				
lame	49.3	20.5	7	97
slow/reluctant to rise	49.3	16.6	20	83
stiff	36.4	16.6	3	67
stilted	38.1	16.0	5	90
unwilling to rise	67.2	19.1	2	100

Table 3.2 cont'd: Summary statistics of VAS scores indicating pain intensity associated with each behaviour and physiological sign when observed in the dog, as allocated by 72 practising veterinary surgeons using a VAS defined from 0 to 100.

	Mean	S.D.	Minimum	Maximum
Attention to painful area				
biting	54.7	24.4	7	97
chewing	48.3	22.7	7	98
licking	37.2	19.2	5	88
looking	33.5	20.9	0	80
rubbing	30.2	17.4	0	75
Response to touch				
crying	61.7	18.6	21	94
flinching	52.9	18.3	16	88
growling	59.8	17.5	15	93
guarding	57.0	20.2	0	94
snapping	71.2	18.0	22	100
Response to food				
disinterested	50.5	19.8	9	90
eating hungrily	9.6	8.5	0	42
picking	34.9	15.8	6	87
rejecting food	57.4	21.8	13	100
Physiological signs				
tachycardia	63.9	17.5	19	98
panting	58.3	20.8	4	92
tachypnoea	67.2	15.7	22	98
pyrexia	58.0	21.3	0	98
salivating	49.2	25.0	0	99
trembling	49.8	21.0	8	89
muscle spasm	65.9	16.4	19	97
dilated pupils	68.4	20.6	11	98

and third lowest in the category, in terms of the intensity of pain they implied. However, 'curled' and 'rigid' were allocated differing scores (1.23 and 3.64 respectively). This indicated that 'curled up' represented least pain in the category and 'rigid' the most pain. The ranks also demonstrated that there was large variability within the data since the standard deviations were high relative to the mean ranks.

3.2.4 Category validation

The validity of the categorisation of expressions was investigated using two statistical techniques. Firstly hierarchical cluster analysis across all of the expressions was used to explore the similarities between the items (Sokal and Sneath, 1963). In this case the categorisation would be supported if the expressions were shown to be most similar, or most highly correlated with expressions from within the same category. The second technique used was to calculate Cronbach's alpha coefficient (Cronbach, 1951). The alpha coefficients provides a measure of the consistency of the expressions within a category and the extent to which the expressions are associated the same attribute.

3.2.4.1 Cluster analysis

Hierarchical cluster analysis is a statistical technique that allows underlying structures in data to be explored by examining the similarities between variables. The hierarchy produced by the cluster analysis can then be used to highlight natural groupings and patterns within the data.

The process of carrying out a hierarchical cluster analysis can be described in a number of steps. Firstly, each variable included in the analysis is defined separately and the similarity between each variable is calculated and entered into a similarity matrix. In the case discussed here, the similarity matrix was defined using the correlation and absolute correlation between the variables. From the similarity matrix, the two variables that are most similar are joined to form a single cluster. Following this combination a new similarity matrix is calculated which includes the similarity between the remaining variables and the newly formed cluster. This process is repeated until all of the variables are combined to form a single cluster. The structures within the variables can be examined graphically using a dendrogram. This illustrates how the clusters are joined together and the distances between them (Chatfield and Collins, 1980).

Table 3.3: Summary statistics for ranked VAS pain intensity scores allocated by 72 practising veterinary surgeons to 39 behaviours and 8 signs thought to be indicative of pain when observed in a dog. VAS pain intensity scores were ranked within each category for each veterinary surgeon then summarised for all behaviours and signs.

	Mean	S.D.	Minimum	Maximum
Demeanour				
anxious	2.22	0.90	1	4
depressed	2.49	0.92	1	4
distressed	3.48	0.76	1	4
quiet	1.65	0.97	1	4
Response to people				
aggressive	3.49	0.85	1	4
fearful	2.70	0.81	1	4
indifferent	1.56	0.92	1	4
sullen	2.20	0.82	1	4
Posture				
curled	1.23	0.52	1	3.5
hunched	2.45	0.73	1	4
rigid	3.64	0.67	1	4
tense	2.65	0.87	1	4
Activity				
restless	2.61	0.54	1	3
sit/lie still	2.20	0.60	1	3
sleeping	1.11	0.28	1	2
Vocalisation				
crying	2.06	0.88	1	4.5
groaning	2.71	1.04	1	5
howling	3.14	1.21	1	5
screaming	4.79	0.56	1	5
whimpering	2.31	1.23	1	5
Mobility				
lame	3.10	1.30	1	5
slow/reluctant	3.32	1.09	1	5
stiff	1.99	0.94	1	5
stilted	2.11	1.08	1	5
unwilling to rise	4.35	1.03	1	5

Table 3.3 cont'd: Summary statistics for ranked VAS pain intensity scores allocated by 72 practising veterinary surgeons to 39 behaviours and 8 signs thought to be indicative of pain when observed in a dog. VAS pain intensity scores were ranked within each category for each veterinary surgeon then summarised for all behaviours and signs.

	Mean	S.D.	Minimum	Maximum
Attention to painful area				
biting	4.01	1.16	1	5
chewing	3.60	1.11	1	5
licking	2.65	1.19	1	5
looking	2.36	1.43	1	5
rubbing	2.22	1.11	1	5
Response to touch				
crying	3.03	1.34	1	5
flinching	2.17	1.19	1	5
growling	3.06	1.12	1	5
guarding	2.74	1.32	1	5
snapping	3.98	1.35	1	5
Response to food				
disinterested	3.09	0.80	1	4
hungry	1.04	0.18	1	2
picking	2.28	0.51	1	4
rejecting food	3.56	0.62	2	4
Physiological signs				
dilated pupils	5.67	2.04	1	8
panting	3.93	2.17	1	8
pyrexia	4.24	2.13	1	8
salivate	3.28	2.20	1	8
muscle spasm	5.29	1.88	2	8
tachycardia	4.94	2.32	1	8
tachypnoea	5.35	2.00	1	8
tremble	2.96	1.88	1	8

A number of methods for calculating the distance between clusters have been defined. The two methods used in this analysis were complete and average linkage. When using complete linkage the distance between two clusters is the maximum distance between a variable in one cluster and any variable in the other cluster (Krzanowski, 1988). This is also known as the further neighbour. When using average linkage the distance between two clusters is the average distance between any variable in one cluster and any variable in another (Krzanowski, 1988).

These analyses indicated ten variable groupings (Table 3.4). With the exception of group 9, all clusters contained expressions from only 1 or 2 different categories, which suggested that the behaviours and signs within each category were more closely related to other items from the same category than items from other categories. Thus, the relationships between expressions and signs within the categories were strong which suggested that the categorisation was sensible. Five expressions did not fall into any obvious clusters and did not appear to be strongly related to any of the other expressions. These were 'eating hungrily', 'screaming', 'howling', 'salivating' and 'quiet'.

3.2.4.2 Internal consistency

The homogeneity, or internal consistency, of each category was examined to determine whether the items addressed the same aspect of pain (Streiner and Norman, 1996). This was investigated by calculating Cronbach's alpha coefficient (Cronbach, 1951).

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_l \sigma_l^2}{\sigma_T^2}\right) \qquad \text{Equation 3.1}$$

Where n : the number of observations

σ_l^2 : the variance of the scores allocated to expression l

σ_T^2 : the variance of the sum of the total over all the expressions in the category.

The interpretation of Cronbach's alpha coefficient is that a value of close to 1 indicates that the expressions within the category are consistent and measure the same attribute, whereas a value of close to 0 indicates that there is little consistency between the expressions and they do not measure the same attribute.

The derivation of the coefficient can be illustrated by examining Cronbach's alpha for a category containing two items (A and B). In this case the variance of the item scores are

Table 3.4: Groupings identified using hierarchical cluster analysis for behaviours and physiological signs thought to be indicative of pain when observed in a dog. Cluster analysis was carried out on VAS scores indicating the pain intensity associated with each behaviour or sign as judged by 72 practising veterinary surgeon. Each expression was previously allocated to a category; this original categorisation is denoted by the letter alongside each expression, and the footnote key*.

Group	Expressions
1	biting (g), chewing (g), rubbing(g) painful area.
2	licking (g), looking (g) at painful area.
3	lame (e), slow/reluctant to rise (e), stiff (e), stilted (e).
4	unwilling/unable to rise (e), panting (j), pyrexia (j), muscle spasm (j).
5	tachypnoea (j), tachycardia (j), trembling (j), dilated pupils (j), disinterested in food (f), picking at food (f), rejecting food (f).
6	growling when touched (h), guarding when touched (h), snapping when touched (h).
7	crying when touched (h), flinching when touched (h), whimpering (c), crying (c), groaning (c).
8	depressed (a), indifferent to people (b), sullen towards people (b)
9	anxious (a), distressed (a), aggressive towards people (b), fearful of people (b), hunched (d), tense (d), rigid (d), restless (i).
10	curled up (d), sitting/lying still (i), sleeping(i).

*a: demeanour, b: response to people, c: vocalisation, d: posture, e: mobility, f: response to food, g: attention to painful area, h: response to touch, i: activity, j: physiological signs.

denoted var(A) and var (B), the variance of the category total is denoted var(A+B). Then Equation 3.1 can be written as follows:

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\text{var}(A) + \text{var}(B)}{\text{var}(A+B)}\right) \qquad \text{Equation 3.2}$$

However, the variance of A plus B can be written as shown in Equation 3.3.

$$\text{var(A+B)=var(A)+var(B)+2cov(A,B)} \qquad \text{Equation 3.3}$$

Therefore, Equation 3.2 can be written as shown (Equation 3.4)

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\text{var}(A) + \text{var}(B)}{\text{var}(A) + \text{var}(b) + 2\text{cov}(A,B)}\right) \qquad \text{Equation 3.4}$$

Hence, when the covariance between the two items is small, i.e. they are unrelated, alpha is close to 0 and when the covariance is large and the items are closely related alpha is close to 1. This illustration extends to categories containing n items (Cronbach, 1951).

With the exception of Demeanour and Response to people, each category had a relatively high value of Cronbach's alpha coefficient (i.e. greater than 0.6) as shown in Table 3.5. Thus, the consistency of expressions within each category was reasonable, which supported the original categorisation of the expressions.

Demeanour and Response to people showed low internal consistency with alpha coefficients of less than 0.55. The two categories were then combined and the alpha coefficient re-calculated. The coefficient increased to 0.63 compared to the two categories investigated separately (0.52 and 0.48 for Demeanour and Response to people respectively), which suggested that the internal consistency of the categories may be improved by combining these two.

3.2.5 Investigation of structure underlying items

The next step in the investigation of the expressions was to examine the similarities between the expressions and thus to explore whether the number of expressions could be reduced or the expressions simplified without loss of information. Summary statistics and exploratory investigation of the data suggested that some items were associated with a similar degree of pain (Table 3.2). However, before any reduction of the items could be

Table 3.5: Cronbach's alpha coefficient calculated to investigate internal consistency of behaviours and signs thought to be indicative of pain in dogs within the allocated categories. Coefficients were calculated on VAS scores indicating the pain intensity associated with each expression as judged by 72 practising veterinary surgeons.

Category	Alpha for Raw Data
Demeanour	0.52
Response to people	0.48
Posture	0.61
Activity	0.66
Vocalisation	0.73
Mobility	0.75
Attention to painful area	0.80
Response to touch	0.77
Response to food	0.63
Physiological signs	0.81
Demeanour and Response to people	0.632

carried out, the relationships between the expressions within each category were examined more closely. No single statistical method could be used to examine all aspects of the relationships between the expressions and so a number of different analyses were carried out. Hierarchical cluster analysis was carried out to examine the correlations between the expressions within each category. The mean pain scores associated with each expression was compared, within categories, using analysis of variance models followed by pairwise comparisons, and finally the distribution of the pain scores were compared using the Kolmogorov Smirnov Test.

3.2.5.1 Cluster analysis

Hierarchical agglomerative cluster analysis was used for a second time to investigate the relationships between the expressions by examining the correlation between items within all categories (Sokal and Sneath, 1963). This analysis differed from the previous cluster analysis in that it was used to investigate the relationships within categories rather than across all of the items. The dendrogram for Posture (Figure 3.2) indicated that, of the four expressions within this category 'hunched' and 'rigid' were the most strongly correlated. This analysis was repeated for all categories and the resulting dendrograms are shown in Appendix 2. Possible combinations of expressions suggested by the cluster analysis are given in Table 3.6.

3.2.5.2 Analysis of variance models

The mean pain intensity scores allocated to each expression within each category were compared using a simple one-way analysis of variance model (ANOVA). Multiple comparison confidence intervals for all pairwise differences in the mean scores were also calculated. Since this produced a large number of confidence intervals, to ensure a global confidence level of 95% multiplicity adjustments were made using Tukey pairwise comparisons (Braun and Tukey, 1983).

The normality of the VAS scores was investigated by examining the residuals resulting from the ANOVA models. Residuals from the categories Demeanour, Response to people, Posture, Mobility and Response to food were found to be normal. Tests of residuals from the categories Activity, Vocalisation, Attention to wound, Response to touch and Physiological signs indicated a lack of normality. Further examination of Q-Q plots for the residual data indicated that for each of these categories the residual data were only slightly

Figure 3.2: Dendrogram resulting from a hierarchical cluster analysis of the VAS scores indicating the intensity of pain associated with behaviours relating to the Posture of a dog. The cluster analysis was carried out using average linkage of the correlation between the 4 behaviours 'curled', 'tense', 'hunched' and 'rigid'.

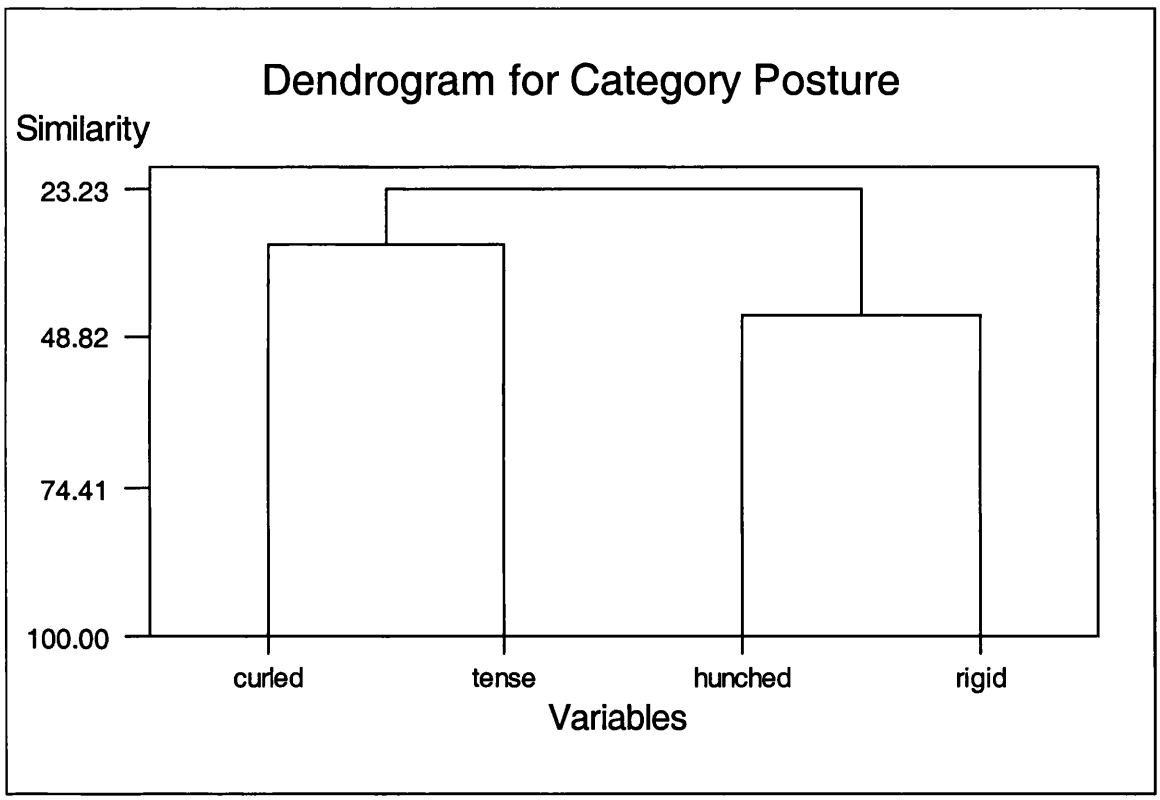


Table 3.6: Behaviours and physiological signs thought to be indicative of pain in a dog, shown to be similar in perceived pain intensity when examined using hierarchical cluster analysis of the VAS pain intensity scores within each category of behaviour.

Category	Behaviours or signs shown to be similar
Demeanour	anxious + distressed
Posture	hunched + rigid
Response to people	indifferent + sullen
Activity	sitting or lying still + sleeping
Vocalisation	crying + groaning + whimpering
Mobility	stiff + stilted; lame + slow or reluctant to rise
Attention to painful area	biting + chewing
Response to touch	crying + flinching; growling + snapping
Response to food	rejecting + picking at food
Physiological parameters	tremble + muscle spasm

skewed and that the deviation from normality was marginal. The ANOVA model was deemed appropriate since the method is robust to small deviations from normality.

The ANOVA models showed significant differences in the mean pain scores between expressions in all 10 categories (p-values<0.001). The Tukey intervals for Posture (Table 3.7) indicated that only 'hunched' and 'tense' did not have significantly different mean pain scores. Complete results for all categories are given in Appendix 2. This analysis indicated that, with the exception of the 'activity' category, at least two expressions within each category could be combined (Table 3.8).

3.2.5.3 Comparison of the empirical cumulative distribution functions

For each expression the distribution of the pain intensity scores allocated was estimated using the empirical cumulative distribution function (ECDF). This takes the form of a step function which indicates the proportion of observations associated with a pain intensity score of less than or equal to the value of interest, for any of the possible pain scores. The form of the ECDF can be expressed as shown in Equation 3.5

$$ECDF(x) = \begin{cases} 0 & \text{if } x < x_1 \\ k/n & \text{if } x_k \le x < x_{k+1} \\ 1 & \text{if } x \ge x_n \end{cases} \qquad \text{Equation 3.5}$$

Where x is the value of interest

x_1 is the smallest observed value

x_n is the largest observed value

n is the number of observations

k is the number of observations with values of less than or equal to x

The ECDFs for any two expressions can be compared using the Kolmogorov-Smirnov test (Hollander and Wolfe, 1973). This test is based on the maximum distance between the ECDFs for variables x and y and can be expressed as shown in Equation 3.6.

$$\text{Test Statistics} = \max |\text{ECDF(x)-ECDF(y)}| \qquad \text{Equation 3.6}$$

When the test statistic is large, i.e. the distance between the two distributions is large, then the distribution of the two variables are said to differ.

Table 3.7: Tukey pairwise confidence intervals, used to compare mean VAS pain intensity score allocated by 72 practising veterinary surgeons, associated with the behaviours 'curled up', 'hunched', 'rigid' and 'tense'.

	Hunched	Rigid	Tense
Curled up	(23.7,39.7)	(44.2,60.1)	(26.4,42.3)
Hunched		(12.4,28.4)	(-5.4,10.7)*
Rigid			(-25.8,-9.8)

* non-significant result

Table 3.8: Behaviours and physiological signs thought to be indicative of pain in a dog, shown to be similar in perceived pain intensity, when examined using Tukey pairwise comparisons of the mean VAS pain intensity scores within each category of behaviour.

Category	Behaviours and signs shown to be similar
Demeanour	anxious + depressed; depressed + quiet
Posture	hunched + tense
Response to people	indifferent + sullen; sullen + fearful
Activity	none
Vocalisation	crying + groaning + whimpering; groaning + howling; howling + whimpering.
Mobility	lame + slow or reluctant to rise; stiff + stilted
Attention to painful area	licking + looking + rubbing/scratching; biting + rubbing/scratching; biting + chewing
Response to touch	crying + growling + flinching + guarding
Response to food	rejecting + disinterested in food
Physiological parameters	panting + pyrexia; tachycardia + tachypnoea + muscle spasm; tachypnoea + dilated pupils; salivate + tremble

Within each category the ECDFs were calculated and pairwise comparisons made between the expressions using the Kolmogorov Smirnov test. Figure 3.3 and Figure 3.4 show the ECDFs for 'curled up' compared to 'rigid' and for 'hunched' compared to 'tense' respectively. The ECDFs for 'curled up' and 'rigid' were quite different since the distance between the two step functions is large (KS test statistic=0.75) indicating that the distributions of the pain intensity scores given to these expressions were not the same. The step functions for 'hunched' and tense' were very close together (KS test statistic=0.08) which indicated that the distributions of the VAS scores were similar for these behaviours.

The Kolmogorov-Smirnov test results were used to identify pairs of items with similar ECDFs. Results for the Posture category suggested that only 'hunched' and 'tense' were similar (Table 3.9). This analysis identified expressions that could be combined in 7 of the 10 categories (Table 3.10). The categories Activity, Response to people and Response to food did not contain any behaviours with similar pain intensity score distributions. In general, this method indicated fewer similarities than the analysis of variance and Tukey intervals, since the method required similarity across the *distribution* of pain intensity scores and not just in the *mean values*. Complete results of the Kolmogorov-Smirnov tests are presented in Appendix 2.

3.2.6 Focus group discussion of results

The results of the statistical analyses highlighted those items in the list that were similar in pain intensity and those which may be unrelated to any of the other items. To maintain the face validity of the pain scale being developed, changes to the items could only be accepted when clinical relevance was maintained. To ensure this was the case a focus group was formed, consisting of 4 qualified veterinary anaesthetists. Each member of the focus group was experienced in the recognition and assessment of pain in animals. The group was presented with and discussed the results of all the statistical analyses and following discussion of the results, it recommended which items could sensibly be combined or removed.

3.2.6.1 Demeanour and Response to people

Both the Demeanour and Response to people categories were thought to reflect an animal's character during the examination procedure and both provided information on similar aspects of the animal's perception of and reaction to its surroundings. Therefore, these two

Figure 3.3: Empirical Cumulative Distribution Functions (ECDF) of the VAS scores indicating the intensity of pain thought to be associated with the behaviours 'curled up' and 'rigid' when exhibited by dogs.

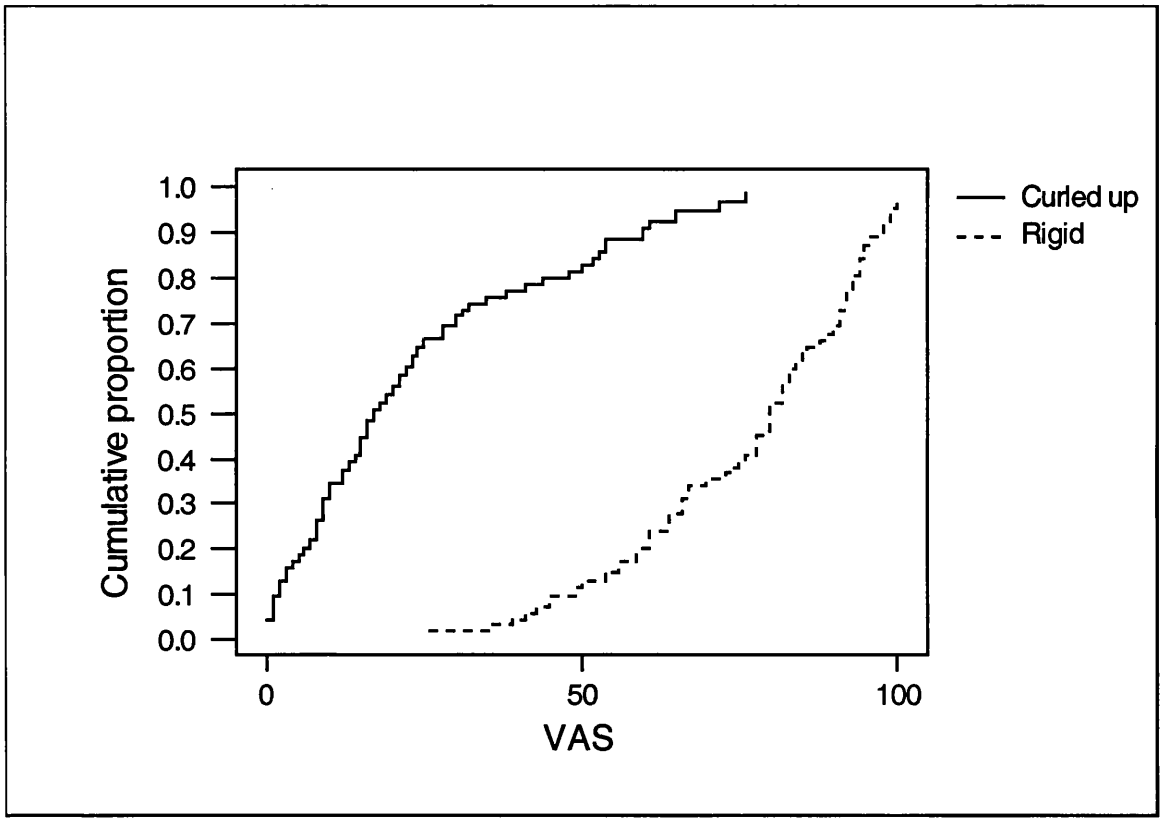


Figure 3.4: Empirical Cumulative Distribution Functions (ECDF) of the VAS scores indicating the intensity of pain thought to be associated with the behaviours 'hunched' and 'tense' when exhibited in dogs.

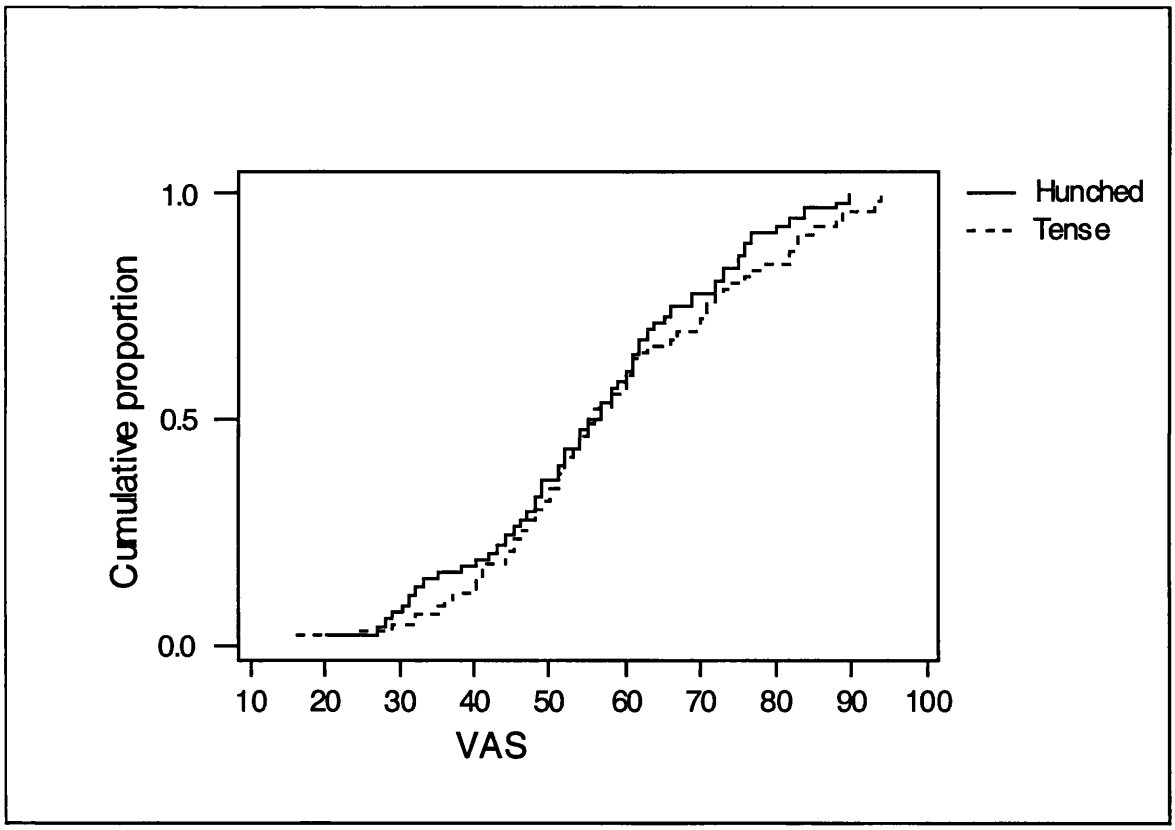


Table 3.9: P-Values from Kolmogorov-Smirnov test used to compare empirical cumulative distribution functions of the VAS pain intensity scores allocated by 72 practising veterinary surgeons to the behaviours 'curled up', 'hunched', 'rigid' and 'tense'.

	Hunched	Rigid	Tense
Curled up	0.66, ×	0.75, ×	0.69, ×
Hunched		0.51, ×	0.08, ✓
Rigid			0.42, ×

× significant difference, ✓ non-significant difference

Table 3.10: Behaviours and physiological signs thought to be indicative of pain in a dog, shown to be similar in perceived pain intensity when examined using the Kolmogorov-Smirnov test to compare the empirical cumulative distribution function of the VAS pain intensity scores within each category of behaviour.

Category	Expressions and signs shown to be similar
Demeanour	anxious + depressed
Posture	hunched + tense
Response to people	none
Activity	none
Vocalisation	crying + groaning + whimpering; groaning + howling
Mobility	slow or reluctant to rise + lame; stiff + stilted
Attention to painful area	chewing + biting; licking + looking; looking + rubbing
Response to touch	growling + guarding + flinching, possibly also crying
Response to food	none
Physiological parameters	tachycardia + tachypnoea + pyrexia + muscle spasm; panting + pyrexia; tachypnoea + dilated pupils; muscles spasm + dilated pupil; salivate + tremble

categories were combined to form an overall Demeanour category. This decision was supported by an increased Cronbach's Alpha coefficient (from 0.521 and 0.484 individually to 0.632 for the combined category).

Once the two categories of Demeanour and Response to people were combined, the focus group felt that some expressions within these categorise could also be combined. The cluster analysis and Tukey intervals indicated that 'indifferent' (mean VAS=24.3) and 'sullen' (mean VAS=33.1) within the Response to people category were similar. 'Quiet' (mean VAS=27.5) and 'indifferent' (mean VAS=24.3), were also thought to convey similar behaviours and were combined to form 'quiet or indifferent', while 'sullen' was removed completely. 'Anxious' (mean VAS=38.3) and 'fearful of people' (mean VAS=46.8) were combined to form 'anxious or fearful'. 'Distressed' was removed, as it was said to be an evocative term that was attributable to an animal's general well being. In order to provide a response for an animal whose pain did not elicit an effect on their demeanour, two expressions, 'happy and content' and 'happy and bouncy' were added. Thus the new category, Demeanour comprised expressions 'aggressive', 'depressed', 'disinterested', 'nervous or anxious or fearful', 'quiet or indifferent', 'happy and content' and 'happy and bouncy'.

3.2.6.2 Posture

Results of the Tukey pairwise comparisons (Table 3.8) and Kolmogorov-Smirnov tests (Table 3.10) indicated that the items 'hunched' (mean VAS=55.5) and 'tense' (mean VAS=58.1) conveyed similar pain intensities. These postures were also thought to be clinically similar and so were combined. 'Curled up' was felt to be inappropriate in this assessment, as an animal could exhibit this behaviour for a number of reasons other than pain. It was not considered an obvious indicator of pain and therefore was removed. The resulting category, Posture, contained the descriptors 'rigid' and 'hunched or tense'. The expression 'neither of these' was added for completeness.

3.2.6.3 Activity

Cluster analysis (Table 3.6) indicated a strong relationship between 'sitting or lying still' (mean VAS=35.1) and 'sleeping' (mean VAS=13.1), even though the mean scores and ECDFs were not similar. The focus group felt that neither expression was particularly informative and so both were removed from the scale. This left only 'restless' in this category and for completeness, 'comfortable' was added.

3.2.6.4 Vocalisation

'Crying' (mean VAS=55.3) and 'whimpering' (mean VAS=58.6) were similar in all the analyses and were combined (see Table 3.6, Table 3.8 and Table 3.10). The cluster analyses (section 0) indicated that the behaviour 'howl' was not closely related to any others in the category. The focus group considered that a dog may howl for a number of reasons unrelated to pain, hence this expression was removed. The resulting behaviours were 'crying or whimpering', 'groaning', 'screaming' and 'not vocalising' the last of these was included for completeness.

3.2.6.5 Attention to painful area

'Licking' (mean VAS=37.2) and 'looking' (mean VAS=33.5) were given comparable VAS scores (Table 3.8 and Table 3.10). This was agreed by the focus group and so the two expressions were combined. The results of the cluster analysis and Tukey intervals also indicated that 'rubbing' (mean VAS=30.2) was similar to both 'licking' and 'looking' (Table 3.8 and Table 3.10). The focus group also considered these items similar and they were combined. The expressions 'chewing' (mean VAS=48.3) and 'biting' (mean VAS=54.7) were shown to be similar (Table 3.6, Table 3.8 and Table 3.10), however the group considered that 'chewing' provided a better description of the behaviour and so 'biting' and 'chewing' were replaced by 'chewing' alone. The expression 'ignoring painful area' was added. The completed category consisted of the behaviours 'chewing painful area', 'licking or looking or rubbing painful area' and 'ignoring painful area'.

3.2.6.6 Response to food

The focus group considered that an animal's response to food would be difficult to determine during the assessment procedure unless food was offered. In addition, nausea could affect this response. Because of the difficulties in assessing this category and the perceived limited information that would be gained from it, this category was removed completely.

3.2.6.7 Mobility

The analyses of the VAS scores indicated that 'stiff' (mean VAS=36.4) and 'stilted' (mean VAS=38.1) conveyed similar intensities of pain (Table 3.6, Table 3.8 and Table 3.10). The focus group felt that 'stilted' was a characteristic of human movement, rather than movement in dogs, therefore 'stiff' and 'stilted' were replaced by 'stiff' alone. The group

agreed that, if an animal was 'unable to rise' then the assessment would not be carried out, thus 'unable to rise' was replaced by 'assessment not carried out'. Also 'unwilling to rise' (mean VAS=67.2) and 'slow or reluctant to rise' (mean VAS=49.3) were considered to be similar types of behaviour, hence 'slow or reluctant to rise ' was included. The expressions 'lame' (mean VAS=49.3) and 'slow or reluctant to rise' (mean VAS=49.3) had very similar VAS scores, however the focus group felt that retaining these as separate items in the scale would provide valuable information for the person carrying out the assessment, hence the expressions were included separately. The expression 'slow or reluctant to rise' was expanded to 'slow or reluctant to rise or sit' to encompass animals with difficulty sitting, as well as standing. Following these changes, the category 'mobility' contained the expressions 'stiff', 'slow or reluctant to rise or sit', 'lame', 'assessment not carried out' and 'none of these', which was added for completeness.

3.2.6.8 Response to touch

The expressions 'growl' (mean VAS=59.8) and 'guard' (mean VAS=57.0) had similar VAS scores (Table 3.8 and Table 3.10) and were combined to form 'growl or guard'. 'Crying' (mean VAS=61.7) was also similar to these two behaviours (Table 3.8), however the three were not combined as the focus group considered that they portrayed different types of behaviour. 'Growl' and 'guard' were thought to be aggressive behaviours, whereas 'crying' was a more submissive behaviour. Thus, the expressions included in this category were 'cry', 'flinch', 'snap', 'growl or guard' and 'none of these'.

3.2.6.9 Physiological signs

The focus group indicated that 'heart rate', 'respiratory rate' and 'pupil dilation' should be included in the scale, as these were thought to provide an insight into an animal's physiological state. The information gathered from assessing 'tachycardia', 'panting' and 'tachypnoea' could be obtained by the assessment of heart and respiratory rates, therefore these original items were replaced. 'Pyrexia' was excluded as it was considered impractical to take an animal's body temperature during a pain assessment. Excessive salivation was judged to be a sign of nausea rather than pain and was excluded. Similarly, 'trembling' was excluded as it was thought to be more likely to be associated with fear, excitement or cold than pain. 'Muscle spasm' was excluded as it was thought too difficult to define and assess during the assessment procedure and could lead to inconsistency in the use of the measurement scale.

3.2.7 Examination Procedure

The focus group also defined how the behaviours and signs should be assessed, and how dogs were to be examined. The examination was to be carried out in and around the animal's kennel, whenever possible. The observer was instructed to remove surgical theatre greens and white lab-coats while carrying out this procedure as the animal may have made some association with these clothes and altered its behaviour in some way.

The assessment procedure required the observer to watch the dog from outside its kennel, which allowed assessment of spontaneous behaviour (i.e. Posture, Activity, Vocalisation and Attention to painful area).

Having assessed spontaneous behaviour, the observer was instructed to approach the kennel and call the animal by name. This allowed assessment of the animal's interaction with the observer, i.e. its Demeanour. At this point, the observer was instructed to enter the kennel and record heart rate, respiratory rate and assess pupil dilation, while the animal was in a resting state. The animal's mobility was assessed by taking it out of the kennel and walking it around the hospital ward area for approximately 10 metres. Then gentle, even pressure was applied to the area around any wound. During this part of the examination, the observer assessed the animal's response to touch. Where the animal had no obvious wound or painful area, the response to touch was assessed by investigating the animal's response to pressure on and around the stifle. The animal was then returned to the kennel. To facilitate the examination procedure and to ensure that it was carried out consistently, a form was devised detailing how the assessments were to be carried out (section 3.3).

3.3 Pain Questionnaire

The questionnaire was made up of a number of sections each of which has several possible answers. The observers were instructed to tick the appropriate answers. The following form details the instructions given to the observers:

Approach the kennel, ensuring you do not have a lab coat or theatre greens on as these may cause a change in the animal's behaviour. While you approach the kennel, look at the dog's behaviour and reactions and answer the following questions.

Look at the dog's posture, does it seem...

Rigid	
Hunched or Tense	
Neither of these	

Does the dog seem to be....

Restless	
Comfortable	

If the dog is vocalising is it...

Crying or Whimpering	
Groaning	
Screaming	
Not vocalising/none of these	

If the dog is paying attention to its wound is it...

Chewing	
Licking or Looking or Rubbing	
Ignoring its wound	

Now approach the kennel door and call the dog by name. Then open the door and encourage the dog to come to you. From the dog's reaction to you when watching him/her assess the animal's character.

Does the dog seem to be...

Aggressive	
Depressed	
Disinterested	
Nervous or Anxious or Fearful	
Quiet or Indifferent	
Happy and Content	
Happy and Bouncy	

The next assessment is the dog's physiological responses. Record the dog's respiratory rate in the space below. Moving the dog as little as possible, record a heart rate (by direct palpation or using a stethoscope) counting for a minimum of 15 seconds. Also, look at the dog's eyes and assess whether the pupils are dilated.

Respiratory Rate	
Heart rate/pulse	
Dilated pupils	

If the dog is mobile open the kennel and put a lead on the dog. If the dog is sitting down get it to stand and then come out of the kennel, then walk slowly up and down the area outside the kennel. On returning to the kennel ask the dog to sit down.

During this procedure did the dog seem to be...

Stiff	
Slow or Reluctant to rise or sit	
Lame	
None of these	
Assessment not carried out	

The next procedure is to assess the dog's response to touch. If a wound is visable apply gentle pressure using two fingers to the wound and an area approx. 4cm around it. If the position of the wound is such that it is impossible to touch then apply the pressure to the closest point to the wound. If there is no wound then apply the same pressure to the stifle and surrounding area.

When touched did the dog...

Cry	
Flinch	
Snap	
Growl or Guard wound	
None of these	

3.4 Discussion

The literature published in veterinary medical research and the results discussed in Chapter 2 show that a valid and reliable pain measurement scale for use in dogs is lacking. The approach taken in this chapter was to develop a composite measurement scale based on pain-related behaviours. The benefits of taking a composite measurement approach are well recognised in psychometric literature and to some extent in the medical literature (Wright and Feinstein, 1992; Nunnally and Bernstein, 1994; Streiner and Norman, 1995). However, to date, the benefits of using a composite measurement approach in the measurement of pain has not been explored in the veterinary literature.

The process of constructing a composite measurement scale and the issues surrounding the development of such a tool for use in health measurement, are discussed within the medical and psychometric literature (Guyatt *et al,* 1992; Nunnally and Bernstein, 1994; Coste *et al*, 1995; Streiner and Norman, 1995). The basic validity of items included (i.e. behaviours and signs in this case) can be investigated and verified during development, only when development is carried out empirically. When a scale is constructed in an ad-hoc manner, there is no scope to examine the relevance of the items included by the developers; the items are in effect the developers' best guess. The number of theoretical assumptions made in constructing the scale is minimised when an empirical approach is taken, since the behaviours included are based on verifiable information which in turn lends weight to the validity of the resulting scale.

The benefits of taking an empirical approach to scale construction were highlighted in a review of the construction of composite measurement scales published by Coste *et al* (1995). The authors noted that the ad-hoc construction of composite measurement scales is scientifically questionable if their measurement properties are not adequately explored. The paper provides a review of scale development in the medical literature and cites a number of areas in which the use of psychometric methods in scale development is crucial. Particular emphasis is placed on the collection and selection of scale items, the development of a scoring system and investigation of the validity, internal consistency and reliability of the scale. Similar criteria for the development and performance of health measurement scales are detailed in the discussion paper by Guyatt *et al* (1992). Again, the collection and selection of items and the validity and reliability of the resulting scales are cited as being fundamental in the development of any health measurement scale. The methods utilised in this chapter fulfil the criteria for scale construction discussed in these

articles since the scale development was based on psychometric principles. The performance of this new CMPS (i.e. its validity and reliability) is explored using psychometric methods in Chapter 5.

In Chapter 1 the review of pain measurement scales developed for use in animals indicated that a number of composite scales have been published. These scales include the numerical rating scale discussed by Conzemius *et al* (1997), the Colorado State University scale (Hellyer and Gaynor, 1998) and the UMPS (Firth and Haldane, 1999). The literature gives little indication of how these scales were constructed. When using the numerical rating scale, Conzemius *et al* (1997) gave no justification for the structure of the scale, items included or weights applied. From the information provided it must be concluded that the numerical rating scale was constructed by the authors in an ad-hoc manner. More information on construction of the Colorado State University scale was provided by Hellyer and Gaynor (1998). The scale items were derived from a review of the literature and the authors' knowledge of the field. The selection criteria, the categorisation and the rationale for the weights applied to the items were not detailed. The most recent composite measure scale reported is the UMPS (Firth and Haldane, 1999). The structure of the UMPS was based on that of the Children's Hospital of Eastern Ontario Pain Scale (CHEOPS; McGrath *et al*, 1985). The items included in the UMPS were derived from a review of the pain measurement literature and from the authors' experience of dogs' physiological and behavioural responses following surgery. The weighting scheme applied to the items was adapted from CHEOPS. However, the article provides no indication of why the authors believed that this approach was appropriate, and no exploration of the psychometric properties of the scale was reported.

This review indicates that, until now, development of composite measurement pain scales in veterinary medicine has been primarily ad-hoc. In itself, this does not invalidate the scales, however their measurement properties must be explored thoroughly before they can be accepted for universal use, and thus far, this appears to have been neglected. On this basis, the CMPS developed here is unique within the veterinary literature, as no other work has taken a similar psychometric approach to the development of a pain measurement scale nor have the measurement properties of such scales been explored (the measurement properties of the CMPS are investigated in Chapter 5).

The initial construction of the CMPS followed similar methods to that used by Melzack and Torgerson (1971) in the development of the MPQ as described briefly in sections 3.1.1.

The MPQ construction consisted of a process of item selection, classification, examination of the validity of the items and investigation of the intensity associated with each. The stages of development of the CMPS are discussed below in relation to the work carried out by Melzack and Torgerson in the construction of the MPQ.

The initial list of behaviours on which the CMPS was based was obtained from a group of practising veterinary surgeons. It was considered that the members of this group represented a cross-section of the veterinary community with broad clinical experience, and would provide a comprehensive list of behaviours relating to pain. The items included in the MPQ were collected from previous work by Dallenbach (1939) and a review of the literature which was also aimed at ensuring a sound basis for further development of the scale (Melzack and Torgerson, 1971). For the CMPS, it was decided to make use of a consultative approach rather than an investigative one since this was novel in the development of pain scales in veterinary medicine and may have provided information that had not been available previously.

The list of 269 behaviours and physiological signs collected from the veterinary surgeons was rationalised and categorised by one person. The criteria used were not formally documented at the time of categorisation and few restrictions were placed on the decisions made by the research worker. This may have caused bias in the resulting list of expressions. However, a cluster analysis indicated that, in general, items from the same category were more closely correlated than items across categories, suggesting that the items within the categories were strongly related and the categorisation was sensible. In addition, the focus group indicated that the face validity of the categories and items was acceptable.

During the initial collation of expressions by the independent researcher, the two behaviours, 'eating hungrily' and 'sleeping', were added. These were included as it was felt that the list was somewhat unbalanced and did not encompass the full range of possible behaviours. In retrospect, the validity of including these items was questionable since they were not said to be related to pain by the 69 veterinary surgeons consulted. In support of this, a cluster analysis indicated that 'eating hungrily' was not closely related to any other expression and had low VAS scores. This suggested that this behaviour should not have been included and following agreement by the focus group, it was removed. Similarly, the focus group did not consider 'Sleeping' to be informative, so it too was removed. Again, the results of the cluster analysis supported this action. No other expressions were added

and no existing expressions altered other than as defined by the criteria detailed in section 3.2.6.

During development of the MPQ, the items were allocated to 3 major classes and 16 subclasses. Melzack (1975) explored the validity of this classification by asking 20 subjects whether they agreed on the categorisation of each item. Items with less than 65% agreement were then presented to an additional 20 subjects, who were asked to assign each item to a category and the number of allocations that agreed with the original categorisation was assessed. This investigation demonstrated that the original categorisation was sensible since only 11 words showed less than 65% agreement. Thus, the categories defined for the MPQ were valid. Although the MPQ and CMPS take differing approaches to exploring the allocation of expressions to categories, both investigations provide information on the validity of the categorisation and support the categorisation adopted for the scale items.

The intensity of pain implied by each item was assessed to allow the relationships between the items to be investigated using formal statistical methods. This information was collected via 72 practising veterinary surgeons, each allocating a pain intensity score to the items using a VAS and NRS. Analysis of both of these scales led to the same conclusion; to avoid unnecessary repetition of results, only the VAS analyses have been presented. When assessing the pain intensity associated with each expression, veterinary surgeons were asked to provide the pain intensity scores assuming that the behaviour or sign was present because the animal was in pain. The wording used indicated that the behaviour was present only because of pain. In reality, this may not always be the case since a behaviour or sign may be exhibited for reasons other than pain. Therefore, the VAS scores may differ slightly from the values that would have been observed if the behaviour was present for any reason. However, it was anticipated that the relative positions and relationships between the behaviours would be unchanged.

In Chapter 2 a number of problems in using the VAS to assess pain were highlighted. Despite these problems, the VAS was used to assess the intensity of pain associated with the scale items. In this situation, the VAS was not used to assess pain directly, but as a cross modality-matching tool to allow the veterinary surgeons consulted to record the pain intensity they believed to be associated with each item (Nunnally and Bernstein, 1994). It has been shown that the VAS is reliable when used in this manner, particularly in the assessment of pain descriptors (Gracely, 1983). Therefore, any problems of

generalizability over observers when using the VAS would not affect the results when used in this context.

When investigating the pain associated with each item in the MPQ Melzack and Torgerson (1971) consulted 180 students, doctors and patients. Each indicated the pain they believed to be associated with the items using a 7-point, SDS-type scale. Melzack and Torgerson's assessment procedure involved a wide variety of subjects, whereas in the development of the CMPS only practising veterinary surgeons were consulted. With hindsight, it may have been of benefit to consult with a wider variety of subjects when assessing pain intensity associated with the items, for example, veterinary nurses and owners. However, the focus group comprised a number of researchers with specialist interest in pain, and could be said to provide an alternative view of pain assessment compared to the practising veterinary surgeons. Thus, it was anticipated that the inclusion of two groups with differing specialist skills would provide a sufficiently broad base of knowledge on which to construct the CMPS.

The internal structure of the CMPS was examined to refine the items by removing any which were deemed redundant or misleading. This was done by examining the relationships between behaviours and signs within each category. No single statistical technique could provide a comprehensive picture of the relationships between the items hence the relationships were examined subjectively and then formally by using three differing statistical methods (cluster analysis, ANOVA followed by multiple comparisons and Kolmogorov-Smirnov tests). These statistical analyses explored the relationships between the items in different ways via the correlation between the items, the mean and standard deviations of the VAS scores and the overall distribution of VAS scores allocated.

The results of the cluster analysis suggested that within every category at least two behaviours could be combined. This method examined the correlation between the expressions, to identify where two behaviours were linearly related even though their mean pain intensity scores may have differed. For example, in the Posture category, both behaviours 'hunched' and 'rigid' indicated tension and immobility, although the mean pain scores were considerably different (mean VAS 55.5 and 75.9 respectively). Therefore, the results of the cluster analysis were informative only when considered in the light of the results of the other analyses.

The Tukey intervals indicated that a large number of items could be combined in some categories, for example, the analysis indicated that 4 behaviours in the Vocalisation

category could be combined. This relatively large number of combinations was due to the large variance in the pain scores assigned to the behaviours. This suggested that the large variability in the VAS scores had caused loss of power in this method and the results may have indicated inappropriate combinations. Nevertheless, the method was useful in identifying global similarities in pain intensity associated with the behaviours, such as the combination of 'crying', 'groaning' and 'whimpering' (mean VAS 55.6, 65.1 and 58.6 respectively). In general, fewer combinations were suggested using the Kolmogorov-Smirnov test than the Tukey intervals. The reason for this was that this method examined the similarity between items across the whole distribution of the scores allocated and not simply on the mean scores. Therefore, this method could be viewed as the most stringent of the three in examining the relationships between the items.

By using these different approaches, it was possible to gain a comprehensive picture of the relationships between the items. Using three different methods ensured that it was unlikely that two items would be combined incorrectly because their perceived pain appeared to be similar when in reality this was not the case. For example, two items may be highly correlated, but have differing mean values, or two items may have similar mean VAS scores but the distribution of the scores may differ. Hence, the use of multiple procedures was a safeguard against inappropriate removal or combining of items that would have resulted in a loss of information in the final CMPS.

The disadvantage of these methods is that a large number of comparisons were made between expressions, which could lead to a significant increase in the probability items being said to be significantly different when in reality they are associated with similar pain intensities. This could have resulted in items not being combined even though it was appropriate to do so. To counter this, the analyses undertaken were used solely as descriptors of the relationships underlying the expressions, to provide information to the focus group. No absolute decisions were made based on the statistical results alone. Therefore, it is anticipated that the possible impact of multiplicity was minimised.

When exploring the pain intensity associated with the items in the MPQ, 180 subjects were consulted. They used a 7-point scale to indicate the pain intensity associated with each item. These scores were converted to scale values for the items using Thurstone's Categorical Judgement model. The model also provided discriminal dispersion values for each item, which indicated the degree of disagreement between the scores assigned to the item (Melzack and Torgerson, 1971). This analysis allowed the scale developers to assess

the level of agreement within the scale items and the relative positioning of the items, in terms of their associated pain intensity. The relationships between the items in the MPQ were not investigated since the objective of the development of the MPQ differed slightly from the development of the CMPS. Melzack and Torgerson were concerned primarily with creating as comprehensive a tool as possible; rationalising the items included was of no concern. However, for any tool that is to be used in veterinary hospitals and clinics its simplicity is paramount. The inclusion of only those items which are valid in the measurement of animal pain was crucial and therefore investigation of the internal structure of the items was necessary.

None of the existing composite measurement pain scales described for use in animals detailed any investigation of the collection or selection of the items included in the scales. Thus, it is difficult to critically review the methods described here in comparison with techniques previously used in the veterinary literature. However, a review of the psychometric literature suggested that the techniques used in the development of the CMPS address all of the relevant principles underlying scale development. In addition, it is possible to compare the items resulting from this investigation with behaviours and signs that are said to be indicative of pain in the veterinary literature.

Many of the behaviours and signs included in the list that resulted from the item rationalisation (Table 3.1) and in the resulting CMPS (Section 3.3) were included in the other behaviour-based pain measurement tools developed for use in dogs (discussed in Chapter 1). The Colorado State University scale includes Vocalisation, Movement, Unprovoked and Interactive behaviour as well as Heart rate and Respiratory rate (Hellyer and Gaynor, 1998). Similarly, Mobility, Vocalisation, Temperament, Locomotion, Heart rate and Respiratory rate are included in the guidelines for pain recognition defined by Morton and Griffiths (1985). Most recently, Firth and Haldane (1999) included Activity, Posture, Vocalisation, Mental status, and Physiological data in their pain measurement tool. Other sources concerned with the investigation of pain also cite behaviours included in the CMPS as being indicative of pain, such as Body movement, Restlessness, Respiratory rate and Whimpering (Potthoff and Carithers, 1989). This similarity between the literature and the items included in the CMPS suggests that the method of data collection and rationalisation resulted in behaviours and signs that are thought to be indicative of pain by the wider veterinary community. In particular, the behaviours and signs highlighted in this investigation and included in the CMPS are in agreement with the behaviours and signs cited by other researchers interested in the measurement of pain.

During its discussions, the focus group identified a number of items that could be combined or removed completely. In the majority of cases, these decisions followed the results of the statistical analyses. Some items were combined or removed following discussions within the focus group alone. For example, the category Response to food was removed completely and 'quiet' and 'indifferent' were combined in the Demeanour category even though the statistical analysis did not directly compare the items in the original Demeanour and Response to people categories. All of these changes were due to one of the following reasons:

- The items were thought to be highly related to attributes other than pain and would reduce the validity of the resulting scale. For example, 'howling' could be related to distress and 'sleeping' or 'sitting still' could reflect a possible degree of sedation resulting from residual anaesthetic drugs. Therefore, inclusion of these items in the assessment could be at best uninformative and at worst somewhat misleading.

- The items were thought to be difficult or inappropriate to assess in the clinical environment. Some physiological parameters were felt to be too time consuming and difficult to measure in a clinical situation. For example, the focus group considered that Response to food would be difficult to assess within the short time period of the assessment procedure, and also that it may be inappropriate to feed a dog at the time of assessment.

The focus group also recommended that the categories Demeanour and Response to people were combined since both relate to the animal's character. This decision was supported by the increased internal consistency of the combined category compared to the separate categories. A number of items were also added to those included in the list, for example 'happy and content', 'ignoring wound' and 'not vocalising'. These items were added to allow a non-pain related response to be recorded when the assessments were made, a *zero response* in effect.

It has been shown that the CMPS was developed on the basis of sound psychometric principles and followed a similar development process to that used in a well known and well respected composite measurement scale used to measure pain in human medicine, the MPQ. The CMPS represents a novel approach to the development of a pain measurement scale for use in veterinary medicine, as existing scales have not been based on psychometric principles.

Having decided upon the expressions to be included in the measurement scale and the examination procedure, the next step was to allocate weights to each item, to allow an overall measurement of pain intensity to be determined. The models investigated and the resulting weights are discussed in Chapter 4.

4. Definitions and scaling model for the Composite Measurement Pain Scale

4.1 Introduction

The development of any composite measurement scale involves a number of discrete stages. The work detailed in Chapter 3 examined the basic construction of a new composite measurement pain scale (the CMPS), but before the CMPS can be used in practice, specific definitions of the items included and an appropriate weighting scheme must be devised (Torgerson, 1958; Streiner and Norman, 1995; Coste *et al*, 1995).

4.1.1 Definition of scale items

The CMPS incorporates a range of behaviours and aims to give a global picture of the pain intensity being experienced. For the developer, the most important issue is to identify the relationships between the items and the attribute (pain in this instance) and so define a valid scale. From the scale user's perspective, the most important aspect is how the scale can be applied in practice. A valid and practical scale can be achieved only when the use of the scale is clearly specified and the items included are well defined (Nunnally and Bernstein, 1994).

It is possible for measurement tools to be constructed and put into use without the provision of such guidance. For example, the VAS, NRS and SDS have been used in human and veterinary medicine for some years with no formal guidance on how they are defined and used. However, as reported in Chapter 2 and in the literature, such scales can be unreliable when used to measure pain (Ohnhaus and Adler, 1975; Revill *et al*, 1976; Nyrén *et al*, 1987). Reliable measurement can only be achieved when the measurement tool is used consistently between observers and over time (Shrout and Fleiss, 1979). For a complex tool such as the CMPS, consistent use relies on clear, concise and specific definitions of how the tool should be used and how the scale items should be identified (Nunnally and Bernstein, 1994).

An example of a well-defined scale is the Glasgow Coma Scale (GCS). The GCS was developed to assess the level of consciousness in hospital patients (Teasdale and Jennett, 1974). When using the GCS, the administrator assesses the patient's response to light, sound and pain. The assessment of these attributes and the possible responses are defined

clearly in the tool to ensure consistent application. The GCS has been in widespread use for some time and has been shown to be both valid and reliable (Starmark *et al*, 1991; Menegazzi *et al*, 1993)

Thus far, the development of the CMPS has defined the procedure by which the animals should be examined (Section 3.3.8). Definitions of each item included in the scale are required before a weighting scheme can be derived.

4.1.2 Level of measurement and scaling models

A major consideration when devising a weighting scheme for a scale is the nature of the attribute of interest and therefore the level of measurement required. For example, if the attribute is thought to be continuous in nature, then measurement based solely on an ordinal scale may result in a loss of information. Pain is assumed to follow an underlying continuous distribution so a pain measurement scale that provides interval or ratio measurement is required.

The level of measurement achieved by any scale is largely dependent on the design of the scale itself and the method for calculating the resulting global score. The weights applied to the items included in composite measurement scales are often devised using a scaling model that allows the attribute to be quantified in a meaningful way. A large number of scaling models have been developed ranging from very simple methods to complex theoretical models (Torgerson, 1958; Nunnally and Bernstein, 1994). Two main families of scaling models exist, direct or subjective estimation techniques and indirect or discriminant techniques (Nunnally and Bernstein, 1994).

Direct or subjective estimation techniques are based on the developer's estimate of the weights that should be assigned to the items included in the scale. These techniques are not based on a formal model; the appropriateness of a particular model can often be assessed by a critical appraisal of the weights and the scale under construction.

Indirect or discriminant techniques are based on information gathered in studies specifically designed to explore item weights. The weight for each item is derived using one of a number of possible scaling models. The design of the studies used to gather such information is dependent on the scaling model used. These types of models were developed mainly for use in measurement scales in Psychology (Thurstone and Chave, 1966; Nunnally and Bernstein, 1994; Streiner and Norman, 1995).

The equally-weighted scaling model is the simplest of the indirect or discriminant scaling models. As the name suggests, this model assumes equal weights for each of the items included in the scale and assigns a score of 1 to each. The total score therefore represents the number of items chosen when the assessment is carried out. This type of scaling model is often used when the scale is in checklist format and requires the patient or administrator to indicate which, of a list of options, is most appropriate. Existing scales that use this model include the McGill Pain Questionnaire and the Speilberger Anxiety State and Trait Scales (Melzack, 1975; Speilberger *et al,* 1972).

The Likert scaling model is commonly used in measurement scales in areas including Marketing, Psychology and Health. The structure of a Likert-type scale is similar to the equally-weighted model. The items form a checklist, but rather than the patient or administrator recording yes/no answers to each item they record how much they agree with the statement or how applicable the statement is. An example from the CES-D depression scale is the item 'I felt depressed' where the possible responses are 'rarely, 'sometimes', 'a moderate amount' or 'most of the time' (Radloff, 1977). When using the Likert model the number of possible responses for each item can vary; between 3 and 15 categories have been used previously (Jacoby and Matell, 1971). Each response category is allocated a number and a patient's overall score is the sum of the item scores over the whole scale (Streiner and Norman, 1995). To date, animal pain measurement scales have not made use of Likert scales.

The Guttman scaling model is used less often than the other models described. The reason for this is that it makes very stringent assumptions about the structure of the items included in the measurement tool. The model assumes that the scale items are answered 'yes' or 'no', that the items are hierarchical in nature and that every item perfectly dichotomises the population (Guttman, 1944; Nunnally and Bernstein, 1994). In such a scale a positive response to item 1 may indicate more of the attribute than a positive response to item 2, which may indicate more of the attribute than a positive response to item 3 and so on. The Guttman model assumes that any person giving a positive response to item 1 *must* also have given a positive response to items 2 and 3, etc. Each item is allocated a score of 0 (negative response) or 1 (positive response); a patient's score is the total of their item scores. Therefore, the score reflects the items to which the patients give a positive response and hence their level of the attribute. To date Guttman scaling has not been used in animal pain scales.

The method of paired comparisons was derived from the classical law of comparative judgement proposed by L. L. Thurstone (1928). To utilise Thurstone's method of paired comparisons the measurement scale must be structured into small groups of items, each relating to the same attribute and associated with differing levels of the attribute. These scales require the respondent to indicate which item from each group is most appropriate. Hence, the level of the attribute associated with each item can be estimated and translated into an appropriate weight. Using this model, the total score is defined by the sum of the weights for the items chosen. The scores produced using this method can be assumed to provide interval level measurement (Nunnally and Bernstein, 1994; Streiner and Norman, 1995). Thurstone's model has not been used previously for any pain measurement scales in animals.

The primary objectives of the work detailed in this chapter are to:

- define the items included in the CMPS clearly and concisely,
- identify an appropriate scaling model for the CMPS and calculate weights for each of the scale items.

4.2 Materials and Methods, and Results

4.2.1 Definition of scale items

Initial definitions for the items included in the CMPS were provided by Professor A. Nolan and Professor J. Reid, from Glasgow University Veterinary School. Both had been involved in the development of the scale and were familiar with its content. In addition, both had extensive experience of pain assessment in animals and were familiar with the behaviours dogs exhibit when experiencing acute pain.

To ensure the item definitions were as clear as possible, 16 external reviewers examined the definitions proposed by Professors Nolan and Reid. All reviewers were respondents to a message circulated on the American College of Veterinary Anaesthetists email list. The aim of this list is to provide information on current developments in veterinary anaesthesia and related areas, including the assessment of pain. The reviewers were all specialist veterinary anaesthetists and were interested and experienced in the assessment of pain in animals. The changes suggested by the reviewer consisted mainly of clarifying ambiguous wording. The definitions are shown in Table 4.1.

Table 4.1: Definitions of acute pain behaviours thought to indicate pain in dogs and included in the Composite Measurement Pain Scale (CMPS).

Posture
Rigid: Animal lying in lateral recumbancy, legs extended or partially extended in a fixed position.
Hunched: When animal is standing, its back forms a convex shape with abdomen tucked up, or back in a concave shape with shoulders and front legs lower than hips.
Tense: Animal appears frightened or reluctant to move, with an overall impression of tight muscles. Animal can be in any body position.
Normal posture: Animal may be in any position, but appears comfortable, with muscles relaxed.

Comfort
Restless: Moving bodily position, circling, pacing, shifting body parts, unsettled.
Comfortable: Animal settled, resting and relaxed, no avoidance or abnormal body position evident. Remains in same body position, at ease.

Vocalisation
Crying: Extension of the whimpering noise, louder and with open mouth.
Whimpering: Often quiet short high pitched sound, frequently closed mouth. Whining.
Groaning: Low moaning or grunting deep sound, intermittent.
Screaming: Animal making a continual high pitched noise, inconsolable, mouth wide open.

Demeanour
Aggressive: Mouth open or lip curled showing teeth, snarling, growling, snapping or barking.
Depressed: Dull demeanour, not responsive, shows reluctance to interact.
Disinterested: Cannot be stimulated to wag tail or interact with observer.
Nervous: Eyes in continual movement, often head and body movement, jumpy.
Anxious: Worried expression, eyes wide with whites showing, wrinkled forehead.
Fearful: Cowering away, guarding body and head.
Quiet: Sitting or lying still, no noise. Will look when spoken to, but not respond.
Indifferent: Not responsive to surroundings or observer.
Happy and Content: Interested in surroundings, has positive interaction with observer, responsive and alert.
Happy and Bouncy: Tail wagging, jumping in kennel often vocalising with a happy and excited noise.

Attention to wound area
Chewing: Using mouth and teeth on wound area, pulling stitches.
Licking: Using tongue to stroke area of wound.
Looking: Turning head in the direction of area of wound.
Rubbing: Using paw or kennel floor etc to stroke wound area.
Ignoring: Paying no attention to the wound area.

Table 4.1 continued: Definitions of acute pain behaviours thought to indicate pain in dogs and included in the Composite Measurement Pain Scale (CMPS).

Mobility
Stiff: Stilted gait, also slow to rise or sit, may be reluctant to move.
Slow to rise or sit: Slow to get up or sit down but not stilted in movement.
Reluctant to rise or sit: Needs encouragement to get up or sit down.
Lame: Irregular gait, uneven weight bearing when walking.
Normal mobility: Gets up and lies down with no alteration from normal.

Response to Touch
Cry: A short vocal response. Looks at area and opens mouth, emits a brief sound.
Flinch: Painful area is quickly moved away from stimulus either before or in response to touch.
Snap: Tries to bite observer before or in response to touch.
Growl: Emits a low prolonged warning sound before or in response to touch.
Guard: Pulls painful area away from stimulus or tenses local muscles in order to protect from stimulus.
No adverse response to touch: Accepts firm pressure on wound with none of the aforementioned reactions.

4.2.2 Selection of a scaling model

The next step was to identify a scaling model that would be appropriate for the CMPS and would provide interval level measurement. Three models were chosen from those discussed in Section 4.1.2. The models explored were the equally-weighted model, a direct estimation technique referred to as the ranked category model and Thurstone's paired comparison model. The rationale behind these choices is discussed further in Section 4.3.

4.2.2.1 Equally-weighted model

When applying the equally-weighted model to the CMPS all pain behaviours were allocated a score of 1. The non-painful behaviour items, such as 'not vocalising' were assigned a score of 0. The overall score was the number of the items chosen, which constitutes a count of the pain behaviours exhibited by the animal during the examination. Total scores could range from 0 to 20, although it was unlikely that any animal would exhibit more than two behaviours from any one category, so a more realistic range may be from 0 to 13.

4.2.2.2 Ranked category model

The second scaling model considered was the ranked category model. This model belonged to the family of direct estimation techniques since it was devised using the judgement of the researchers involved in the development of the CMPS.

Four members of the focus group (described in Chapter 3) were asked to rank the categories to indicate their relative importance when assessing an animal's pain (Table 4.2). From this assessment, each category was allocated a score, and the items within each category were allocated this score as their weight. The scores assigned to each category indicated that the focus group believed that Vocalisation was the most informative, whereas Demeanour, Mobility and Response to Touch were the least informative.

The items indicating no pain behaviour, such as 'normal posture' were assigned a score of 0. The overall pain score was the sum of the weights of the items chosen. Possible scores ranged from 0 to 31, although it was unlikely that any animal would exhibit more than two behaviours from any one category, so a more realistic range may be from 0 to 22.

Table 4.2: Weights assigned to 7 categories of behaviours included in the composite measurement pain scale used assess pain in dogs. Rank weights were assigned based on the perceived importance of each category in the assessment of pain, as agreed by 4 experts in pain measurement in animals.

Category	Rank
Demeanour	1
Posture	2
Comfort	2
Vocalisation	3
Attention to Painful Area	2
Mobility	1
Response to Touch	1

4.2.2.3 Thurstone's method of paired comparisons

Sixteen independent judges were contacted via the American College of Veterinary Anaesthetists email list. These judges were familiar with the items included in the CMPS as they had all previously contributed to the item definitions (Section 4.1.1). The judges were provided with a list of all possible pairs of items within each category (Appendix 4) and the definitions of the items (Table 4.1). They were then asked to indicate which behaviour in each pair implied the highest intensity of pain.

For example, the expressions from the Posture category:

Normal posture vs Rigid

Hunched vs Normal posture

Hunched vs Rigid

The information gathered from these comparisons was collated and probability matrices were estimated for each category (Table 4.3). These matrices indicated the probability that one expression was associated with more pain than another, and hence illustrated the position of each item relative to the others. Thurstone's model assumes that the difference in intensity between any two items follows a Normal distribution, with a constant mean, d and standard deviation, 1. Thus, the distance between any two items in a category can be estimated by transforming the probabilities in the distance matrix into z-scores. Each z-score indicates the number of standard deviations between the corresponding two items. These estimates are likely to be affected by sampling error, which is reduced by averaging the z-scores for each item over the other items in the scale.

The weights calculated for the CMPS ranged from -1.86 to 1.74 (Table 4.4). For ease of interpretation, the weights were transformed to provide continuous scores in the range from 0 to 10 (Equation 4.1).

$$\frac{(z - \text{lowest in category})}{(\text{highest - lowest})} * 10 \qquad \text{Equation 4.1}$$

where z : z-score for each item

lowest in category : lowest z-score of items in category

lowest : lowest possible total score based on z-scores

highest : highest possible total score based on z-scores

Table 4.3: Matrices of estimated probability of the row item being associated with greater pain intensity than the column item. Probabilities are shown for 7 categories of behaviour included in the CMPS.

Category: Posture

	Hunched/Tense	Neither
Rigid	0.58	1.0
Hunched/Tense		1.0

Category: Comfort

	Comfortable
Restless	1.0

Category: Vocalisation

	Groan	Scream	Not Vocalising
Cry/Whimper	0.40	0.01	1.0
Groan		0.01	1.0
Scream			1.0

Category: Attention to painful area

	Lick/Look	Ignore
Chewing	0.92	1.0
Lick/Look		1.0

Table 4.3 continued: Matrices of estimated probability of the row item being associated with greater pain intensity than the column item. Probabilities are shown for 7 categories of behaviour included in the CMPS.

Category: Mobility

	Slow/Reluctant	Lame	None
Stiff	0.83	0.17	1.0
Slow/Reluctant		0.08	1.0
Lame			1.0

Category: Response to touch

	Flinch	Growl/Guard	Snap	None
Cry	0.99	0.60	0.27	1.0
Flinch		0.36	0.18	1.0
Growl/Guard			0.27	1.0
Snap				1.0

Category: Demeanour

	Depressed	Disinterested	Nervous	Quiet	Content	Bouncy
Aggressive	0.36	0.46	0.50	0.73	1.00	1.00
Depressed		0.82	0.91	1.00	1.00	1.00
Disinterested			0.55	0.92	1.00	1.00
Nervous				0.64	1.00	1.00
Quiet					1.00	1.00
Content						0.72

To make this transformation it was assumed that only one item would be chosen from each category. The raw z-scores and transformed weights associated with each of the items are given in Table 4.4. The total pain score was calculated as the sum of the chosen items. If more than one item was chosen from any one category the only the item with the highest weight was used to calculate the total score. The overall pain scores derived using this scaling model can be assumed to provide interval level measurement (Nunnally, 1978; Streiner and Norman, 1995).

The weights defined using this scaling model followed the same ordering as the summary statistics for the VAS pain intensity scores reported in Chapter 3 for all but two of the categories (Mobility and Demeanour). In the Mobility category the mean VAS scores for 'slow' and 'lame' were the same, but using Thurstone's model the weight for 'slow' was smaller than 'lame'. In the 'demeanour' category, 'depressed' was given the largest weight using Thurstone's model, whereas 'aggressive', 'distressed' and 'fearful' are associated with higher VAS pain intensity scores.

4.3 Discussion

The provision of definitions for the items included in the CMPS was aimed at ensuring the scale was used consistently over time and between observers and therefore was intended to improve the reliability of the scale. The benefit of defining items within a pain measurement scale has been acknowledged within the veterinary literature. Of the pain measurement tools developed for use in animals, all provided definitions of how the various behaviours were to be identified and the scores associated with these. For example, in the UMPS proposed by Firth and Haldane (1999) the possible responses to palpation are defined as:

- 'no change from pre-procedural behaviour' – score 0
- 'guards/reacts when touched, including turning head, biting, licking, scratching, snapping at handler or tense muscles and protective posture' – score 2
- 'guards/reacts before touched, including turning head, biting, licking, scratching, snapping at handler or tense muscles and protective posture' – score 3

Table 4.4: Raw and transformed weights for behaviours thought to indicate pain and included in the CMPS to assess pain in dogs. Weights were calculated using Thurstone's method of paired comparisons, and transformed to ensure a maximum total score of 10.

Category	Behaviour	Raw Weight	Transformed Weight
Demeanour	aggressive	0.68	1.22
	depressed	1.37	1.56
	disinterested	0.77	1.26
	nervous	0.51	1.13
	quiet	-0.00	0.87
	content	-1.58	0.08
	bouncy	-1.74	0.00
Posture	rigid	0.85	1.20
	hunched/tense	0.70	1.13
	normal	-1.55	0.00
Comfort	restless	1.16	1.17
	comfortable	-1.16	0.00
Vocalisation	cry/whimper	-0.09	0.83
	groan	0.09	0.92
	scream	1.74	1.75
	not vocalising	-1.74	0.00
Attention to Painful Area	chewing	1.24	1.40
	Lick/look	0.31	0.94
	ignore	-1.55	0.00
Mobility	stiff	0.58	1.17
	slow/ reluctant	-0.01	0.87
	lame	1.17	1.46
	normal	-1.74	0.00
Response to Touch	cry	0.86	1.37
	flinch	-0.25	0.81
	snap	0.89	1.38
	growl/guard	0.36	1.12
	none	-1.86	0.00

Similar definitions are provided for the other items included in the scale, although no definitions were given for the mental status category. The scale states that the observer must assess the dog's dominant/aggressive behaviour before and after surgery. The mental status score is defined as the absolute difference between the pre-surgery and post-surgery scores. The authors gave no indication as to what score should be assigned if the animals mental status score is lower following surgery than before surgery. The authors give no indication of how the definitions were derived for any category.

The pain behaviours included in the Colorado University scale and associated scores were well-defined (Hellyer and Gaynor, 1998). For example, a score of 2 was allocated in the interactive behaviour category if the animal 'vocalised when wound was touched; somewhat restless; reluctant to move but will if coaxed'. The authors did not specify how the definitions were derived or how appropriate they are. The guidelines published by Morton and Griffiths (1985) and Sanford *et al* (1986) were both aimed at ensuring consistent assessment of pain and suffering in laboratory animals. These papers were not aimed at providing formal measurement, but both defined in detail, the behaviours and signs indicative of pain and specified how these should be recognised.

It is evident that the need to define behaviours utilised in the measurement of pain is well recognised within the veterinary literature. However, a review of the scales currently available does not provide any information on the rationale behind the definitions used. In Chapter 3 the development of the veterinary pain scales was explored and found to be ad-hoc in nature. Since the scales themselves were primarily developed on the investigators' judgement, it is reasonable to assume that the item definitions were similarly developed.

Although the review of these by veterinary anaesthetists was aimed at ensuring the item definitions were as clear and concise as possible, the definitions of the items included in the CMPS were also derived on an ad-hoc basis. It is anticipated that consulting with this wider group of veterinary surgeons enhanced the validity of the definitions of the items included in the CMPS.

Of the scaling models used in existing pain measurement scales currently, little information is provided on how the item weights were derived (Conzemius *et al*, 1997; Hellyer and Gaynor, 1998). Only Firth and Haldane (1999), developers of the UMPS, give any indication of the weight scheme used. The weights applied were based on the CHEOPS scale developed for use in paediatric pain, although no details of the rationale or methodology are given. Thus, it must be assumed that the authors used direct or subjective

estimation techniques based on their best estimate of the most appropriate weights. When describing the measurement scales, the authors indicated that the behaviours included were associated with differing degrees of pain and differential weights were assigned to the items. This reflects a consensus in the literature that an animal's behaviour will change with the severity of pain it is experiencing, thus differing behaviours are associated with differing levels of pain intensity. This information should be utilised in the weights assigned, if the scale is to accurately reflect an animal's pain experience. One disadvantage of using direct estimation techniques is that the weights assigned can lack sensitivity. For example, behaviour A may be assigned a score of 1 and behaviour B assigned a score of 2. This implies that behaviour B is twice as important as behaviour A since it is assigned double the weight and contributes twice as much to the final score. The true relative importance of the two behaviours is seldom explored, and so the validity of the weighting scheme is not confirmed.

In the development of the MPQ, Melzack and Torgerson (1971) identified three weighting models for the scale items. The simplest of these is the Number of Words Chosen (NWC). This scheme is equivalent to an equally-weighted model since all items in the questionnaire are allocated a score of 1. The second and third weighting models are known as the Pain Rating Index scores, based on the mean and ranked scale values (PRIS and PRIR respectively). These weights were derived using Thurstone's categorical judgement model, which is similar to the paired comparison model discussed, as it is based on the relative positioning of the items. This model is reported to provide interval level measurement and its use in the MPQ demonstrates an acceptance of the benefits of formal scaling models in the measurement of pain.

The three scaling models explored for application to the CMPS were selected from the possible models reviewed in section 4.1.2.

The Likert scale was thought to be inappropriate for the CMPS as the validity of the three assumptions made when applying this model were questionable. The assumptions made are that all items in the scale address the same attribute (i.e. the scale is unidimensional), the probability of a patient responding positively to the question increases as the level of the attribute increases (i.e. monotonicity) and the relationship between the score and level of attribute for each item is roughly linear (Wright and Feinstein, 1992). The validity of these assumptions when measuring pain in animals must be considered. Although pain in humans has been said to be multidimensional in nature (Melzack, 1975), there has been no

exploration of the dimensionality of pain in animals. Intuitively, it may be questionable to assume that even if a dog could experience the different dimensions of pain, that it would be able to express this through its behaviour. The CMPS is intended to measure the intensity of a dog's pain, and not necessarily to capture the multidimensional aspects of the pain experience. Hence, the first assumption may be valid at least on an intuitive level. The validity of the second assumption is in more doubt. The items included in each category were chosen because they were thought to be associated with distinct levels of pain intensity. Thus, a particular behaviour would be exhibited only if the animal's pain intensity lay within a particular range. This implies that if pain intensity increased beyond the range associated with a behaviour then the probability of the behaviour being exhibited would not increase. Indeed, if the range of pain intensity associated with that behaviour is exceeded then the likelihood of the behaviour being exhibited would fall. For example, if an animal's pain was very severe, it may be more likely to snap when touched than flinch, thus the validity of the monotonicity assumption is questionable. The third assumption of linearity implies that all of the items are associated with the same intensity of pain, and so adding the scores together would not cause bias in the results. When the CMPS was developed, items were chosen specifically because they were related to differing amounts of pain. Therefore, the assumption of linearity is not appropriate. This critical review of the Likert model in relation to the CMPS indicates that two of the three assumptions may not hold, and so the model was deemed inappropriate for this application.

The Guttman scaling model also makes a number of assumptions about the scale items. In particular, it makes very stringent assumptions about the hierarchy of the items included in the measurement tool (Guttman, 1944; Nunnally and Bernstein, 1994). The assumption made is similar to that of monotonicity discussed for the Likert scale, which has been shown to be invalid for the CMPS. The Guttman scaling model is best suited to behaviours and attributes that are developmentally determined where progression to one state guarantees that all previous states have been surpassed (Streiner and Norman, 1995). The fit of a Guttman scaling model to any measurement tool can be assessed, however few clinical or psychological situations have been found to be suitable (Wright and Feinstein, 1992). Even in cases where the conditions for Guttman scaling hold, the model does not provide interval level measurement and its discriminant ability is often poor, hence this method is seldom used and was thought to be inappropriate for the CMPS (Wright and Feinstein, 1992; Streiner and Norman, 1995).

Equally-weighted models have a number of advantages in that they are simple to devise, administer and understand (Streiner and Norman, 1995). In scales where this model has been used, such as in the Speilberger anxiety tests, the scores are often treated as interval level measurements (Melzack, 1975; Speilberger *et al*, 1972). In the case of the CMPS, the assumption of interval level measurement resulting from the equally-weighted model may be invalid for a number of reasons. During the development of the CMPS, the items were combined based on the degree of pain intensity associated with them (Section 3.2). These changes were made to ensure the items included in the scale covered a range of pain intensity and to remove redundancy by combining items that implied the same degree of pain. The equally-weighted model ignores all such differences between the items since each item contributes the same amount to the total score. For example, the items 'screaming' and 'groaning' were associated with different VAS scores (Section 3.2); when using this scaling model both items contribute a score of 1 to the total pain score. Hence the equally-weighted model does not fully exploit the information contained in the items. In addition to this potential loss of information, the equally-weighted model achieves interval level measurement only when the relationship between the scale score and the underlying attribute is constant and linear (Nunnally and Bernstein, 1994). The validity of this is questionable since the items in the CMPS are known to be associated with differing degrees of pain intensity.

Use of the equally-weighted scaling model can also lead to some indirect weighting being introduced to the scale (Streiner and Norman, 1995). This is caused by an unequal number of items being included in the categories. Each category can contribute a different amount to the overall pain score that is dependent only on the number of items included, not on the relative importance of the category (Streiner and Norman, 1995). In the CMPS, the categories contain varying number of items, for example Demeanour includes 7 items whereas Comfort includes only 2, yet Demeanour is not known to be more important than Comfort when measuring pain. Consequently, the use of the equally-weighted model could result in bias in the observed scores.

This review indicates that using the equally-weighted model with the CMPS could have resulted in loss of information and bias in the total scores, and may not have provided interval level measurement. Therefore, this model was not chosen for use with the CMPS.

The ranked category model was proposed to account for the differences in the importance of each category in describing an animal's pain. The categories were weighted according

to their importance in the assessment of pain, as judged by a number of veterinary surgeons (Table 4.2). The perceived benefit of this method over the equally-weighted model was that it would better reflect an animal's pain experience since the importance of each category would be utilised when calculating the total score.

When applying this scaling model the items included in each category were assigned the same weight, and so the model did not account for any differences in the intensity of pain associated with individual items. For example 'lame' and 'stiff' in the Mobility category were assumed to imply the same pain intensity despite having significantly different VAS scores (Section 3.2). Moreover, interval level measurement requires that the relationship between the item scores and the attribute is constant and linear. Therefore, the relationship between pain and the scale items must be constant within each category since each carries the same weight. This assumption does not hold for the CMPS as the items within each category are known to be associated with differing intensities of pain.

It has been noted earlier that an imbalance in the number of items in each category can cause indirect weighting in the scale and result in bias in the total scores. This issue is compounded in the ranked category model by the differential weights between the categories. For example, Comfort was assigned a weight of 2 and Demeanour a weight of 1, thus Comfort is perceived as the more important. However, the combined scores that the two categories could contribute to the total pain score are 2 and 5 for Comfort and Demeanour respectively (assuming the non-pain related behaviours are scored 0). Therefore, despite Comfort being considered more important in pain measurement, Demeanour potentially contributes more to the total score because of the imbalance in the number of items included in the two categories. These issues suggest that the requirement of interval level measurement would not be fulfilled if the ranked category model were applied to the CMPS. This and the questions surrounding the clinical relevance of this scaling model (i.e. behaviours within a category being allocated the same weight, even though they are associated with differing pain intensity), suggested that the ranked category model would not be appropriate for the CMPS.

Thurstone's method of paired comparisons differs from the equal-weighted and ranked category models in that it is an empirical rather than a subjective scaling model. The data used to fit the model were gathered from specialists in pain measurement, and the weights were calculated for the items within each category. The method did not assume any differences between the categories, since the weights were calculated separately within

each category. When fitting this scaling model, the total pain score was the sum of weights for the behaviours observed during the examination. To allow the weights to be transformed to provide positive scores (Section 4.2.2.3) only one item within each of the categories could be used to calculate the total score. To ensure consistency, it was decided that when an animal exhibited more than one behaviour within a category, the item with the highest weight would contribute to the total pain score. The rationale behind this was that if an animal demonstrated more than one behaviour, the behaviour with the highest weight would be most representative of the pain severity.

No existing pain measurement scales developed for animals use Thurstone's paired comparisons model, although it has been used in 'quality of life' scales in human medicine. The use of Thurstone's paired comparison method was criticised by Jenkinson (1991) when it was applied to the Nottingham Health Profile (NHP). The author suggested that this scaling model was not suitable for use in *factual* scales but should be used only in scales concerned with attitude. The author's main criticism of the scaling model was that the weights calculated using Thurstone's model were unstable. One reason that was proposed for this was that the items included in the NHP addressed only the *extreme ends* of the attribute being measured therefore the *distributions* underlying the scale items could have been far apart. This would render the distances between the items difficult to estimate accurately and result in instability in the weights assigned to the items. The author also noted that since the NHP may be multidimensional, the appropriateness of the scaling model was questionable. The criticisms voiced were based solely on the compatibility of the NHP to Thurstone's scaling method, rather than the scaling model itself (Jenkinson, 1991). This highlights the importance of examining the assumptions made when applying this scaling model.

When applying Thurstone's method of paired comparisons two conditions must be fulfilled:

- The pain intensity associated with each item should follow an underlying Normal distribution (Nunnally and Bernstein, 1994).

- There should be transitivity between the items, i.e. if A is thought to be greater than B, and B is greater than C, then A should be greater than C (Nunnally and Bernstein, 1994).

In Chapter 3 the residuals from the analysis of variance models fitted to the VAS scores indicated that in general the VAS scores were Normal (section 3.2). Therefore, it is valid to assume that the pain intensity associated with the items follow Normal distributions. The second assumption of transitivity can be examined using the probability matrices calculated in Thurstone's paired comparisons model. In the Mobility category it can be seen that 'slow/reluctant' is associated with less pain than 'stiff', and that 'stiff' implies less pain than 'lame'. In addition, 'slow/reluctant' is associated with less pain than 'lame', hence in this category the transitivity of the items is maintained. On examination, the condition of transitivity holds for all categories of the CMPS. The validity of the Normality assumption indicates that the weights calculated should accurately reflect the pain intensity associated with each item and transitivity implies the scale measures pain in a unidimensional way (Nunnally and Bernstein, 1994).

As mentioned above, one further point that should be noted when applying Thurstone's paired comparison model is that where the distributions underlying items are far removed from each other the resulting weights can be unstable. In the Demeanour category, a number of items were shown to have large differences between the distributions, for example 'quiet', 'happy and content' and 'happy and bouncy' suggesting that the weights associated with the items in this category are difficult to estimate and may be unstable. This reflected some concerns raised previously regarding the use of the Demeanour category and the suitability of this category in pain assessment will be discussed further in Chapter 6.

Existing work in Psychology has shown that the total scores produced using Thurstone's model possess interval level measurement properties (Nunnally and Bernstein, 1994). It was previously noted that in an interval scale the zero score is arbitrary. Pain is defined as an 'unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage' (Merskey and Bogduk, 1994). When pain is assessed in animals, the patient cannot convey the sensory and emotional experience through any means other than its behaviour. Therefore, the only levels of pain that can be conveyed to the observer are those which cause a patient to alter its behaviour in some way. In humans, some discordance has been shown between a patient's self-report of pain and pain assessed using behavioural observation (Teske *et al*, 1983). This suggests that pain is not fully represented by a patient's behaviour and that it is possible for a patient to experience pain which would be reported as being painful on a self-reporting scale, but may not be sufficiently intense to cause a behavioural change. Hence, a score of zero based

on behaviour cannot be assumed to indicate the complete absence of pain, merely that the patient is not experiencing pain with the intensity that would cause a behavioural change. The assumption that the zero point is arbitrary is valid, thus a scaling model producing interval level rather than ratio level measurement is acceptable.

The assumptions required for the application of Thurstone's paired comparison model do, in the main, hold and the model provides interval level measurement as required in the measurement of pain. In addition, the items included in the CMPS were derived to ensure that they were associated with differing levels of pain intensity; Thurstone's model exploits these differences and provides appropriate weights for each of the items. Thus, Thurstone's paired comparison model was considered the most appropriate of the three models explored.

The rationale behind the use of a formal scaling model in the development of the CMPS has been discussed. The use of such a model constitutes a change in the methodology previously used in the development of any pain scale in veterinary medicine, since the weighting schema used in the pain scales currently published in the veterinary literature take the form of ad-hoc, direct estimation techniques. The numerical rating scale (Cozemius, 1997) and the Colorado University scale (Hellyer and Gaynor, 1998) do not provide any justification for weights derived or any indications of how appropriate these weights are. Thus, the approach taken here is novel in the veterinary literature.

5. Investigation of the performance of the Composite Measurement Pain Scale developed for use in dogs

5.1 Introduction

Having constructed the composite measurement pain scale (CMPS), defined the items included in the scale and allocated appropriate weights to the items, the next step was to examine its performance when used by clinicians to measure pain in dogs.

The two most important psychometric properties of any measurement scale are validity and reliability (Guyatt *et al*, 1992). The investigation of validity and reliability was first developed within Psychology, where research is frequently concerned with the measurement of abstract concepts such as intelligence or anxiety (Cohen *et al*, 1996). In recent years, the general move towards evidence-based medicine has resulted in an increase in the number of scales being developed and used in the measurement of health (Streiner and Norman, 1995). In particular, the use of composite measurement scales for the investigation of complex phenomena such as physical disability, quality of life and pain has grown. Consequently, the need for adequate investigation of the performance of these measurement scales has been acknowledged (Beyer and Wells, 1989; Guyatt *et al*, 1992; Coste *et al,* 1995; Johnston, 1998).

The concepts of validity and reliability are explored in Chapter 1. Recall that the fundamental idea behind the validity of a measurement scale is to examine whether the scale actually does measure the attribute of interest (Kline, 1993; Streiner and Norman, 1995). In the case of the CMPS, validity would be addressed by exploring to what extent the scale actually measures pain in dogs. The reliability or generalizability of a measurement scale reflects the amount of error inherent in the observed scores.

Two pain measurement scales commonly cited for use in adults are the McGill Pain Questionnaire (MPQ) and the visual analogue scale (VAS) (Melzack and Torgerson, 1975; Huskisson, 1983). The validity and reliability of both of these methods have been investigated thoroughly.

The work undertaken to explore the reliability of the VAS and its relationship with other measurement scales when used to measure pain in humans is detailed in Chapter 2. The validity of the VAS, when measuring dental pain, was supported when the scale was shown

to detect the effects of analgesics (Seymour, 1982). Joyce *et al* (1975) also demonstrated the validity of the VAS, as it was shown to change following the administration of analgesics in patients with chronic inflammatory disease. Boeckstyns and Backer (1989) examined the validity of the VAS when measuring pain following total knee replacement surgery. They reported strong relationships between the VAS and a 4-item verbal scale and indicated that this supported the scale's validity. However, the validity of the VAS in the measurement of pain has also been challenged (Langley and Sheppard, 1985). The primary concern was that the multidimensional nature of pain is not addressed when the VAS is used since it addresses only pain intensity. The authors indicated that a multidimensional pain scale would be more appropriate. The VAS has also been reported to be one of the simplest scales used in pain measurement (Huskisson, 1974).

Another method that is commonly used to measure pain in human medicine is the McGill Pain Questionnaire (MPQ). The scale was developed in the 1970s as discussed in Chapter 3. The validity of the MPQ has been examined in a number of studies. The face validity of the MPQ has been well demonstrated as the scale is accepted throughout the medical community and has been translated into a number of languages including Finnish, Norwegian and Italian (Ketovuori and Pontinen, 1981; Reading, 1983; Debenedittis *et al*, 1988; Kim *et al*, 1995). Investigations of the construct validity of the MPQ have focused on exploring the multidimensional nature of the scale. A three-factor structure was originally proposed by Melzack (1975). Studies that have attempted to replicate this using principal components analysis have demonstrated similar, though not identical, structures to that proposed. Principal component analysis was used, in these studies, to investigate structures in the relationships between the items included in the MPQ. This allowed the investigators to explore whether the relationships between the items in the MPQ, when used in practice, actually reflected the dimensionality of the questionnaire that was proposed by the original authors. A five-factor structure was demonstrated when the MPQ was used by a cross-section of volunteers and patients (Crockett *et al*, 1977) and a four-factor structure was seen when the MPQ was used by dysmenorrheic patients (Reading, 1979). The criterion validity of the scale was supported when significant correlations between MPQ scores and a criterion scale were demonstrated (Reading, 1982). The MPQ has also been shown to differentiate between 8 differing clinical pain syndromes, with an accuracy of 77% (Dubisson and Melzack, 1976). The reliability of the MPQ over time has been explored and it has been shown to be reliable when used over time (Melzack, 1975; Hunter *et al,* 1979; Graham *et al*, 1980). Agreement of around 70% over repeated applications of the scale has been demonstrated (Melzack, 1975).

The closest parallel to pain measurement in veterinary medicine is pain measurement in young children, since in both cases the patients cannot communicate effectively with a care provider. The need to investigate the properties of pain measurement scales used in children has been acknowledged (Beyer and Wells, 1989; Johnston, 1998) and studies have been undertaken to investigate the validity and reliability of several pain scales routinely used in children. Two such scales are the Oucher and the Children's Hospital of Eastern Ontario Pain Scale (CHEOPS). The validity of the CHEOPS has been investigated in a number of studies, and supported. The tool has been shown to be responsive to the effects of opioids and the scores observed are highly correlated with behavioural measures of pain (McGrath *et al,* 1985). There has been little investigation of the reliability of the CHEOPS. The validity of the Oucher has also been investigated and supported in a number of studies (Beyer and Aradine, 1988). The scale can detect pain caused by surgery as significant differences between pre- and post-operative scores have been demonstrated, and the scores observed are significantly correlated with other pain measurement scales. The reliability of the Oucher has been explored less thoroughly since it has been claimed that the investigation of reliability in such scales is problematic as the method does not lend itself to standard reliability testing (Beyer and Knapp, 1986; Reading, 1993).

In veterinary medicine, the need to assess pain in an objective and consistent manner so as to minimise inconsistencies caused by personal judgment is accepted (Yoxal, 1978; Morton and Griffiths, 1985; Sanford *et al,* 1986; Chapman, 1989; Potthoff, 1989; Bateson, 1991; Hansen and Hardie, 1993). The ability to recognise and assess pain has been of particular interest to researchers concerned with the welfare of animals undergoing experimentation and with the investigation of analgesic drugs (Morton and Griffiths, 1985; Sanford *et al,* 1986; Hamlin *et al,* 1988; Bateson, 1991; Reid and Nolan, 1991; Nolan and Reid, 1993). However, there is very little evidence within the veterinary literature that the validity and reliability of the scales used to measure pain in animals have been examined formally.

Of the pain measurement scales used in animals that are discussed in this thesis, the simple descriptive scales have undergone the most investigation. The VAS and NRS have been used to measure pain in animals in a number of studies (Reid and Nolan, 1991; Nolan and Reid, 1993; Lascelles, 1994). The generalizability of the VAS, NRS and SDS when used by a number of observers to measure pain was investigated in Chapter 2. These results indicated that there was a great deal of variability between the observers and suggested that these simple descriptive scales were not sufficiently reliable when used to measure pain in dogs. Although the validity of the VAS in pain measurement has not been investigated

explicitly, studies using the VAS have shown that the effect of analgesic drugs can be demonstrated which gives some indication of the validity of the scale (Reid and Nolan, 1991; Lascelles, 1994).

A number of composite measurement scales have recently been constructed to measure pain in dogs (and cats), and these are discussed in Chapter 1. However, very little exploration of the validity and reliability of these scales has been undertaken (Conzemius *et al*, 1997; Hellyer and Gaynor, 1998; Firth and Haldane, 1999). The agreement between observers was investigated in the numerical rating scale developed by Conzemius *et al* (1997), and the scores demonstrated good agreement between the observers. The validity and agreement between two observers was investigated for the UMPS (Firth and Haldane, 1999). The authors reported agreement between the observers and supported the validity of the scale, although the methodology for this investigation requires further discussion which will be presented in Section 0.

It is evident that little formal investigation of the psychometric properties of the pain scales used in veterinary medicine has been undertaken. To ensure that the psychometric properties of the CMPS described in Chapters 3 and 4 were thoroughly investigated, two studies were designed to examine its validity and reliability.

The primary objectives of the first study were:

- to examine the construct validity of the CMPS when used in a clinical setting. Constructs explored were whether the animal had undergone surgery, the grouping to which the animal belonged (defined in section 5.2), and the severity of pain perceived to be associated with the animal's condition,

- to examine, in a clinical setting, the overall reliability of the CMPS and its reliability adjusted for multiple observers.

An additional objective was to explore the relationships between NRS pain scores and three physiological signs (heart rate, respiratory rate and pupil dilation). The rationale behind this investigation was to examine whether these signs were indicative of pain and thus, whether they should be incorporated into the CMPS.

The primary objectives of the second study were:

- to examine the generalizability of the scale over observers,

- to examine the generalizability of the scale over time.

5.2 Materials and Methods

5.2.1 Study 1: Validity and Reliability of the CMPS

This study was carried out over a four month period within Glasgow University Veterinary Hospital. Five observers used the CMPS and associated examination procedures (Section 3.2.7) to assess pain in 80 dogs. The dogs included in the study comprised 4 groups of 20 animals each. The members of 3 groups were patients at Glasgow University Veterinary Hospital, and the fourth group consisted of dogs owned by University staff.

5.2.1.1 Observers

The 5 observers who took part were all qualified veterinary surgeons and were post-graduate students at Glasgow University Veterinary School at the time of the study. All had experience of veterinary practice and none had previously been involved in the development of the CMPS. The objectives of the study were explained to each observer, as were the examination procedure and use of the measurement scale.

5.2.1.2 Animals

Of the 80 dogs included, 20 had undergone orthopaedic surgery (Orthopaedic Group), 20 had undergone soft tissue surgery (Soft Tissue Group) and 20 were hospitalised because of medical conditions (Medical Group). The remaining 20 dogs had no clinical abnormalities (Control Group).

The dogs included in the study were not restricted in terms of their age, breed, surgical procedure or medical condition (see Appendix 5 for details). Only animals considered too aggressive to be handled easily were excluded. The observers were not familiar with the dogs included in the study.

5.2.1.3 Examination Procedure

All examinations were conducted between 12 noon and 4pm in the wards of Glasgow University Veterinary Hospital. Where an animal had undergone surgery, the assessments were performed on the day following surgery, i.e. between 19 and 29 hours after the end of surgery.

The definitions of scale items (Section 4.2) were not given to the observers when making their assessments. The observers were not made aware of the reason for any animal's hospitalisation and they were not aware of any surgical procedures that the animals may have undergone.

The examination procedure was identical to that described in Section 3.2.7. The observers were also asked to assign a subjective global pain intensity score to each animal using an NRS, once the examination procedure was completed. The observers allocated a score between 0 and 10 to each dog, where 0 indicated 'no pain' and 10 indicated 'pain could not be worse'.

Throughout the study, each observer examined each animal on one occasion, independently of the other observers, at any time within the 4-hour window (between 12pm and 4pm). If any observer felt that an animal was experiencing an unacceptable degree of pain, the nursing staff were informed. Such animals were given appropriate pain relief, and any animals receiving opiod analgesia during the assessment period were excluded from the statistical analyses.

5.2.1.4 Perceived Severity of Pain

The validity of the CMPS was examined by studying the relationships between the pain scores allocated and three constructs: presence or absence of surgery, study group and perceived severity of pain.

Information regarding the perceived severity of pain associated with medical conditions and specific surgical procedures was collected through consultation with 25 veterinary surgeons from Glasgow University Veterinary School. Each vet was given a list of surgical procedures and medical conditions and asked to assign a score to each. The scores were 0='Not Painful', 1='Mild Pain', 2='Moderate Pain' and 3='Severe Pain'. Each condition was then allocated an overall score corresponding to the median of the severity scores assigned by the 25 veterinary surgeons.

5.2.1.5 Statistical Methods

All statistical analyses were carried out using SAS version 12.0 for Windows.

The pain scores observed using the CMPS and the NRS were explored graphically and by calculating summary statistics. The relationships between the physiological parameters of

heart rate, respiratory rate and pupil dilation and the NRS were investigated using a number of statistical methods. The relationships between heart rate, respiratory rate and NRS score were examined graphically and the association was quantified by calculating Spearman rank correlation coefficients (Spearman, 1904). Where dogs were assessed as 'panting', they were excluded when calculating the correlation coefficients as their respiratory rate could not be measured accurately. The median NRS score associated with panting and non-panting animals was compared using a Wilcoxon Mann Whitney test. The relationship between pupil dilation and NRS scores was examined by comparing the median pain scores for animals with and without dilated pupils. The distributions of pain scores in these two groups were compared graphically and the median scores were compared using the Wilcoxon Mann Whitney test.

The relationship between CMPS scores and data observed for each of the three constructs was examined graphically. The relationship between whether the animal had undergone surgery and pain score was investigated using Wilcoxon Mann Whitney test. The relationships between study group, pain severity and CMPS scores were investigated by comparing the median scores across the groups using the Kruskall-Wallis test. Pairwise comparisons of the median pain scores between study groups and pain severity groups were carried out using a Wilcoxon Mann Whitney test.

The reliability of the CMPS was examined by calculating reliability coefficients for each group and for all groups combined. The inter-animal and error variability were calculated by fitting a random effects model to the data as described by Glass and Stanley (1970) and Snedecor and Cochran (1980). The random effects model fitted to the pain measurement scale scores for each group and for the groups combined was:

$$X_{ij} = \alpha_i + \beta_j + \varepsilon_{ij} \qquad \text{Equation 5.1}$$

where X_{ij} : Pain score allocated by observer i to dog j

α_i : Random effect of observer i, distributed $N(0, \sigma^2_{obs})$

β_j : Random effect of dog j, distributed $N(0, \sigma^2_{dog})$

ε_{ij} : Random error effect, distributed $N(0, \sigma^2_{\varepsilon})$

i = 1 to 4 and j=1 to 12

The overall reliability of the pain measurement scale is as shown in Equation 5.2 (Streiner and Norman, 1995).

$$R = \frac{\sigma^2_{dog}}{\sigma^2_{dog} + \sigma^2_{\varepsilon}}$$ Equation 5.2

where σ^2_{dog} : variability associated with the dogs
σ^2_{ε} : error variability in the model

The above reliability coefficient makes no adjustment for the multiple observers per animal. The reliability of the scale when averaged over the multiple can be used to gain an insight into the variability in the scale caused by the observers. The reliability coefficient adjusted for multiple observers is calculated as shown in Equation 5.3 (Streiner and Norman, 1995):

$$R_{adj} = \frac{\sigma^2_{dog}}{\sigma^2_{dog} + \frac{\sigma^2_{\varepsilon} + \sigma^2_{obs}}{k}}$$ Equation 5.3

where k : number of observers
σ^2_{dog} : variability associated with the animals
σ^2_{obs} : variability associated with the observers
σ^2_{ε} : error variability in the model

In this case the error and observer variability are divided by the number of observers as this indicates the error variability that would be present in the model if a subject's score was taken as the averaged over the k observers.

5.2.2 Study 2: Generalizability of the CMPS

5.2.2.1 Observers

A total of four observers were involved in this study. All were qualified veterinary surgeons and were post graduate students at Glasgow University Veterinary Hospital. Each observer had experience of veterinary practice and none had been involved in the previous study. The objectives of the study were explained to each observer as were the use of the CMPS and the medium by which the study was to be carried out.

5.2.2.2 Animals

A total of 21 dogs were included in this study, all patients at Glasgow University Veterinary Hospital. The animals had undergone surgery within the veterinary hospital and were housed in the hospital wards, as part of their recovery. Video recordings were made of the CMPS assessment procedure for all 21 dogs. The observers taking part in the study

did not know the dogs and no restrictions were placed on the age, breed or surgical procedure of those included. Only animals considered too aggressive to be handled easily were excluded.

5.2.2.3 Video Recording of Examination Procedure

All examinations were carried out in the wards of the veterinary hospital on the afternoon of the day of surgery. The examinations were carried out by one person (LH) and were recorded by a technician using video recording equipment. Each animal was examined and the video recordings made to allow the CMPS to be completed as defined in section 3.3.

The examination procedure consisted of the following steps:

1. The animal was video recorded within its kennel with no interaction from the person carrying out the examination. This allowed the assessments of Posture, Comfort, Vocalisation and Attention to wound.

2. The kennel was then approached and the animal called by name. The animal's Demeanour could then be assessed.

3. The animal was taken out of the kennel and taken to the clinical examination area where it was walked for approximately 10 metres to allow assessment of Mobility.

4. Following the mobility assessment the animal's Response to touch was examined. The animal was taken into an examination room and gentle even pressure was applied directly to the wound and the area 2cm around it.

The video recordings were edited by Glasgow University Media Services. Short pauses were inserted between each recording and on-screen labels were added to identify each animal. For technical reasons, it was necessary to remove the sound from the video recordings. This meant that it would have been impossible to assess Vocalisation, so the observers were instead provided with written information on each animal's vocalisation.

Prior to the study the video recordings were viewed by two members of the focus group involved in the development of the CMPS, Professor A. Nolan and Professor J. Reid. The suitability of each clip was decided from its length and the clarity with which the procedure and the animal's responses could be seen. The video footage identified as the clearest was re-edited and copied on to a single video tape which could be viewed easily by the

observers making the pain measurements. During this review process, 9 video assessments were thought to be of insufficient quality and were excluded, leaving 12 in total.

5.2.2.4 Pain Measurement using Video Recordings of Examination

The four veterinary surgeons taking part were each given a copy of the pain examination video tape. They were also given the CMPS as detailed in Section 3.3, with definitions of the items included in the scale (Section 4.2). In addition, the observers were provided with the age, sex and breed of the animal, and details of the surgical procedure and the animal's vocalisation.

Each observer was asked to watch the video tape and complete the CMPS for each animal. The observers were also asked to give a NRS score of between 0 and 10 to indicate the global pain intensity they perceived the animal to be experiencing. When using the NRS, 0 was defined as 'No Pain' and 10 as 'Pain could not be worse'.

On completion of the first assessments, all materials were collected from the observers. Between two and four weeks later, the same observers were asked to watch the video recordings for a second time and repeat their assessments of the animal's pain. The second assessments were completed under the same conditions as the first.

5.2.2.5 Statistical Methods

The items within the CMPS were allocated the weights calculated using Thurstone's matched pairs (Section 4.2.2.3) and total scores calculated. The CMPS scores were explored using summary statistics and graphical methods.

To calculate the variance associated with each of the factors included in the generalizability study a random effects model was fitted to the data. The factors fitted in the random effects model were, the observers, the dogs, the times at which the observations were taken and the interactions between all of these. The following model was fitted:

$$X_{ijk} = \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \varepsilon_{ijk} \qquad \text{Equation 5.4}$$

where X_{ijk} : Pain score allocated by observer i to dog j at time t

α_i : Random effect of observer i, distributed $N(0,\sigma^2_{obs})$

β_j : Random effect of dog j, distributed $N(0,\sigma^2_{dog})$

γ_k : Random effect of time k, distributed $N(0,\sigma^2_{time})$

$\alpha\beta_{ij}$: Interaction between observer i and dog j, distributed $N(0,\sigma^2_{obs*dog})$

$\alpha\gamma_{ik}$: Random effect of observer *i*, and time *k*, distributed $N(0,\sigma^2_{obs*time})$

$\beta\gamma_{jk}$: Random effect of dog *j* and time *k*, distributed $N(0,\sigma^2_{dog*time})$

ε_{ijk} : Random error effect, distributed $N(0,\sigma^2_{\varepsilon})$

i =1 to 4, j=1 to 12 and k=1 to 2

The generalizability coefficients were estimated (Streiner and Norman, 1995), as shown in Equations 5.5 and 5.6.

Generalizability over Observers

$$G_{obs} = \frac{\sigma^2_{dog} + \sigma^2_{dog*time}}{\sigma^2_{dog} + \sigma^2_{obs*dog} + \sigma^2_{dog*time} + \sigma^2_{\varepsilon}}$$ Equation 5.5

Generalizability over Time

$$G_{time} = \frac{\sigma^2_{dog} + \sigma^2_{obs*dog}}{\sigma^2_{dog} + \sigma^2_{obs*dog} + \sigma^2_{dog*time} + \sigma^2_{\varepsilon}}$$ Equation 5.6

5.3 Results

5.3.1 Study 1: Validity and Reliability of the CMPS

In total, 80 dogs aged between 5 months and 15 years were included in the study (Table 5.1). During the examination procedure, three animals in the orthopaedic surgery group were thought to be enduring an unacceptable degree of pain. These animals were treated with an opioid analgesic and therefore were excluded from all further statistical analyses.

The CMPS and NRS scores followed a Normal distribution for the orthopaedic group. The other groups showed varying degrees of skewness. As was anticipated, the control group was the most heavily skewed with both scales showing a large number of low scores. The scores combined across groups were skewed and did not follow a Normal distribution, consequently non-parametric statistical methods were used where possible. In a very few cases the CMPS assessments and NRS scores were not completed, this resulted in missing pain scores.

5.3.1.1 Severity of pain associated with medical conditions and surgical procedures

The median pain severity scores associated with the medical conditions and surgical procedures are shown in Table 5.2. The majority (54%) of surgical procedures were

Table 5.1: Summary statistics for the age and sex of 80 dogs included in an investigation of the validity of the composite measurement pain scale (CMPS).

		Orthopaedic n=20	Soft Tissue n=20	Medical n=20	Control n=20
Age(yrs)	min	0.4	0.4	0.4	3.5
	mean	4.25	5.8	7.1	6.6
	max	9	10.6	12.5	15
Sex	Male	10	13	6	12
	Female	10	6	13	5

assigned a severity score of 2, i.e. moderate pain. Of the medical conditions listed, half were thought to cause pain and only two were associated with severe pain. Therefore, it could be expected that the pain scores observed in the medical group would be lower than observed in the two surgical groups. This was the hypothesis used when investigating the construct validity of the CMPS.

5.3.1.2 Investigation of relationship between physiological signs and NRS scores

During the development of the CMPS, the physiological parameters heart rate, respiratory rate and pupil dilation were cited as being indicative of pain (Chapter 3). The validity of this claim was addressed by examining the relationship between these parameters and the pain score allocated using the NRS.

The three physiological signs were recorded for all animals. A number of the animals were said to be breathing very rapidly; this was recorded as 'panting' and these animals were investigated separately in the analysis, as their respiratory rate could not be assessed accurately. The control group had slightly lower median heart rate and respiratory rates than the other groups (heart rate: 92beats per minute (bpm) vs. 108bpm, 96bpm and 118bpm, respiratory rate: 30 breathes per minutes (bpm) vs. 45bpm, 40bpm and 52bpm; Table 5.3). In addition, the medical group had slightly higher heart rates and respiratory rates than the other groups (heart rate: 118bpm vs. 108bpm, 96bpm and 92bpm, respiratory rate: 52bpm vs. 45bpm, 40bpm and 30bpm; Table 5.3).

There was no obvious relationship between heart rate and the severity of pain when examined graphically (Figure 5.1), and the variability in heart rate was constant over all levels of the NRS. Spearman's rank correlation coefficients between heart rate and NRS score were not significantly different from zero for any group other than the medical group (p-value=0.02 for medical group, p-value>0.1 for other groups). When the data from all four groups were combined the correlation between heart rate and NRS scores was significant, although the value of the coefficient was small (correlation=0.12, p-value=0.02). These results suggested that the relationship between heart rate and pain score was weak.

Graphical investigation indicated that there was no obvious relationship between respiratory rate and NRS pain scores (Figure 5.2). The Spearman's rank correlation coefficient between respiratory rate and NRS score was significant only in the soft tissue surgery group (p-value=0.03 for soft tissue group, p-value>0.1 for other groups). It should

Table 5.2: List of surgical procedures and medical conditions presented to 25 veterinary surgeons to allow assessment of the severity of pain associated with each. Pain severity was assessed on a scale of 0 to 3, the median pain severity scores associated with each item are shown

Surgical Procedure	Severity	Surgical Procedure	Severity
Anal gland removal	2	Laminectomy	3
Anal furunculosis therapy	2	Lateral wall resection, ear	3
Arthrotomy, carpus	2	Lung lobectomy	3
Arthrotomy, elbow	2	Major tooth extraction	2
Arthrotomy, shoulder	2	Mandibulectomy	3
Arthrotomy, stifle	2	Minor tooth extraction	1
Biopsy, gut	2	Ovariohysterectomy	2
Biopsy, liver	2	Patella-lateral capsular overlap	2
Biopsy, soft tissue mass	1	Perineal hernia repair	2
Carpal arthrodesis	3	Prostatic cyst removal	2
Castration	1	Remove intestinal foreign body	2
Cataract removal	2	Repair hip dislocation	2
Cruciate repair	2	Soft tissue lump removal, (3cm)	1
Cryosurgery	2	Soft tissue lump removal, (3-10cm)	2
Cystotomy	2	Soft palate resection	2
Diaphragmatic hernia repair	3	Suture pad	1
Entropion repair	1	Total ear canal ablation	3
Exploratory thoracotomy	3	Toe removal	2
Eye removal	2	Tonsillectomy	2
Forelimb amputation	3	Total hip replacement	3
Fracture repair, plate	3	Tibial crest transplant	3
Fracture repair, pin	2	Triple pelvic osteotomy	3
Hindlimb amputation	3	Urethrotomy	2
Implant nasal drain	2	Ventral slot	3
Joint flush	1	Vertebral distraction	3

Medical Condition	Severity	Medical Condition	Severity
Acute moist dermatitis	2	Dilated caridomyopathy	0
Acute otitis externa	2	Endocardiosis	0
Acute pancreatitis	3	Focal erosive gastritis	2
Addison's disease	0	Hepatic failure	0
Chronic nephritis	0	Lymphoma	0
Chronic otitis externa	1	Osteosarcoma	3
Cushings disease	0	Pyrexia, unknown origin	1
Diabetes mellitus	0	Vomiting & Diarrhoea	1

Table 5.3: Median (range) heart rate (beats per minute) and respiratory rate (breaths per minute) observed in 77 dogs. The groups consisted of dogs that had undergone orthopaedic (n=17) or soft tissue (n=20) surgery the previous day, had medical conditions (n=20) or were healthy (n=20). Each dog was assessed by 5 veterinary surgeons.

	Heart Rate	Respiratory Rate
Orthopaedic	108 (40-120)	45 (6-Panting)
Soft tissue	96 (12-160)	40 (15-Panting)
Medical	118 (40-180)	52 (12-Panting)
Control	92 (36-160)	30 (18-Panting)

Figure 5.1: Heart rates (beats per minute) and NRS pain scores for each of 77 dogs, in 4 groups. The groups consisted of dogs that had undergone orthopaedic (n=17) or soft tissue (n=20) surgery the previous day, had medical conditions (n=20) or were healthy dogs (n=20). Each dog was assessed by 5 veterinary surgeons.

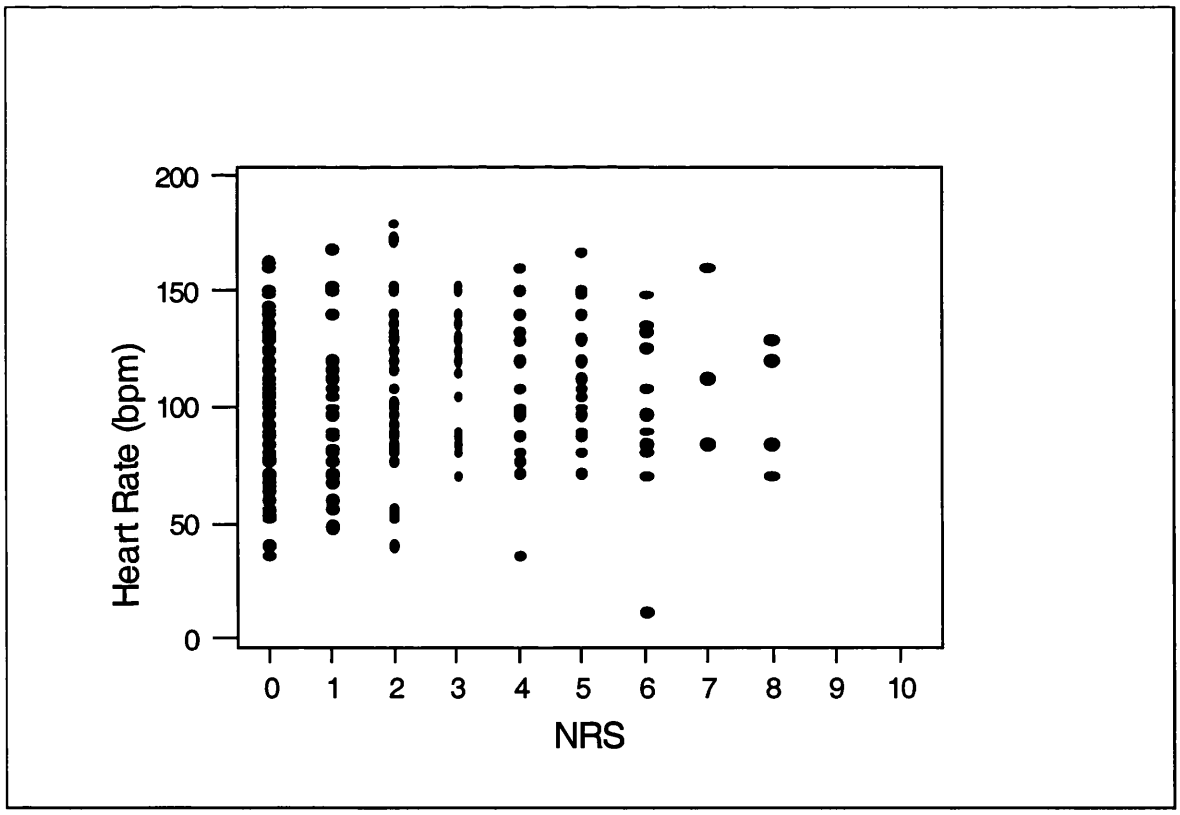
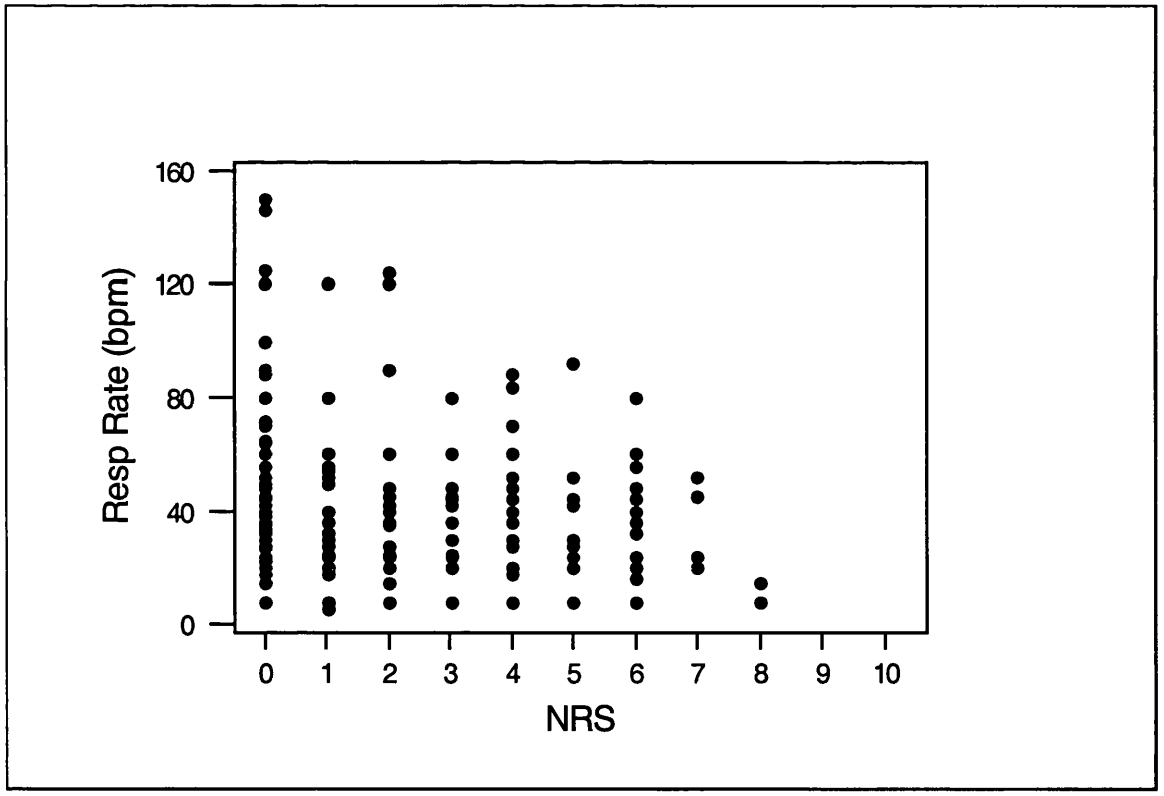


Figure 5.2: Plot of respiratory rates (breaths per minute) and NRS pain scores for a total of 77 dogs in 4 groups. The groups consisted of dogs that had undergone orthopaedic (n=17) or soft tissue (n=20) surgery the previous day, had medical conditions (n=20) or were healthy dogs (n=20). Each dog was assessed by 5 veterinary surgeons.



be noted that the coefficient had a negative value for each group indicating that respiratory rates decreased as pain increased (correlation coefficients were -0.16, -0.26, -0.11, -0.17 for orthopaedic, soft tissue, medical and control groups respectively).

The correlation between respiratory rate and NRS score for all groups combined was small and was not significantly different from zero (correlation coefficient -0.06, p-value=0.30). Hence there was little evidence of a relationship between respiratory rate and NRS score.

The median NRS scores for dogs assessed as panting were not significantly different to those of dogs not assessed as panting for any individual group or for the groups combined (p-values > 0.1 for each group, and all groups combined; Table 5.4).

Graphical investigation indicated differences in the NRS scores assigned to animals with and without dilated pupils in the orthopaedic and soft-tissue groups (Figure 5.3). The Wilcoxon Mann Whitney test (Table 5.5) indicated that the median pain scores differed between dogs with and without dilated pupils in the these groups and for all groups combined. Closer inspection indicated that dilated pupils were associated with higher pain scores in the soft tissue group, but lower pain scores in the orthopaedic group. This suggested that the relationship between pupil dilation and pain scores was not stable.

These results indicated that the relationships between the three physiological signs and pain score were either weak or unstable, therefore these parameters were excluded from the calculations of the pain scores from CMPS.

5.3.1.3 Investigation of the validity of the CMPS

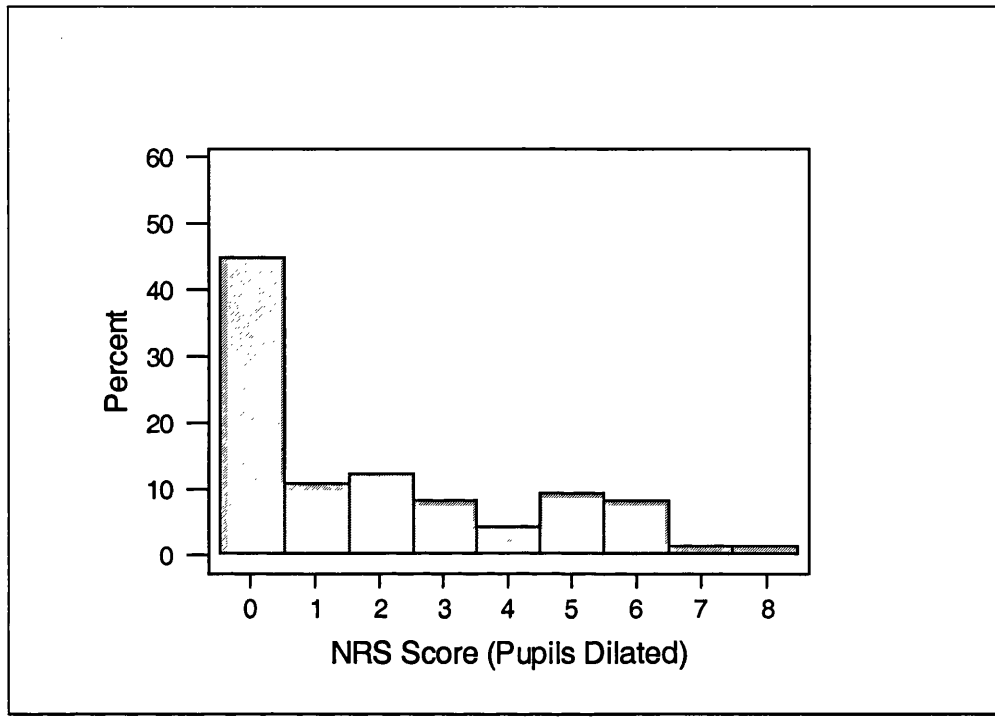
The first construct investigated was the relationship between surgery and pain scores. Graphical comparison of the CMPS scores indicated that the dogs that had undergone surgery had higher pain scores than those that had not (Figure 5.4). Summary statistics supported this since the median CMPS scores were lower in the non-surgical group (median=0.9) than in the surgical group (median=2.4), as shown in Table 5.6. Formal analysis indicated a significant difference between the two groups (p-value<0.01). Therefore, the CMPS scores were sensitive to whether the subjects had undergone surgery.

Summary statistics of the pain scores for each study group calculated are given in Table 5.7 and these suggested that the median CMPS scores may differ between the groups. The orthopaedic and soft tissue surgery groups (median= 3.0 and 2.0 respectively) had higher scores than the medical and control groups (median=1.2 and 0.9 respectively). In addition,

Table 5.4: Mean (median) NRS scores for each of four groups of 77 dogs and all groups combined, split by whether the dog was assessed as panting. Groups consisted of dogs that had undergone orthopaedic (n=17) or soft tissue (n=20) surgery the previous day, had medical conditions (n=20) or were healthy (n=20). Each dog was assessed by 5 veterinary surgeons the p-value shows results of Wilcoxon Mann Whitney test to compare median NRS scores in dogs assessed as panting and those assessed as not panting.

	Orthopaedic	Soft Tissue	Medical	Control	All
Panting	3.3 (3)	1.9 (2)	0.1 (0)	0.2 (0)	1.5 (1)
Not Panting	3.1 (3)	2.4 (2)	0.7 (0)	0.2 (0)	1.5 (0)
p-value	0.66	0.50	0.11	0.72	0.51

Figure 5.3: Histograms of NRS scores for four groups of 77 dogs with and without dilated pupils. Groups consisted of dogs that had undergone orthopaedic (n=17) or soft tissue(n=20) surgery the previous day, had medical conditions (n=20) or were healthy (n=20). Each dog was assessed by 5 veterinary surgeons



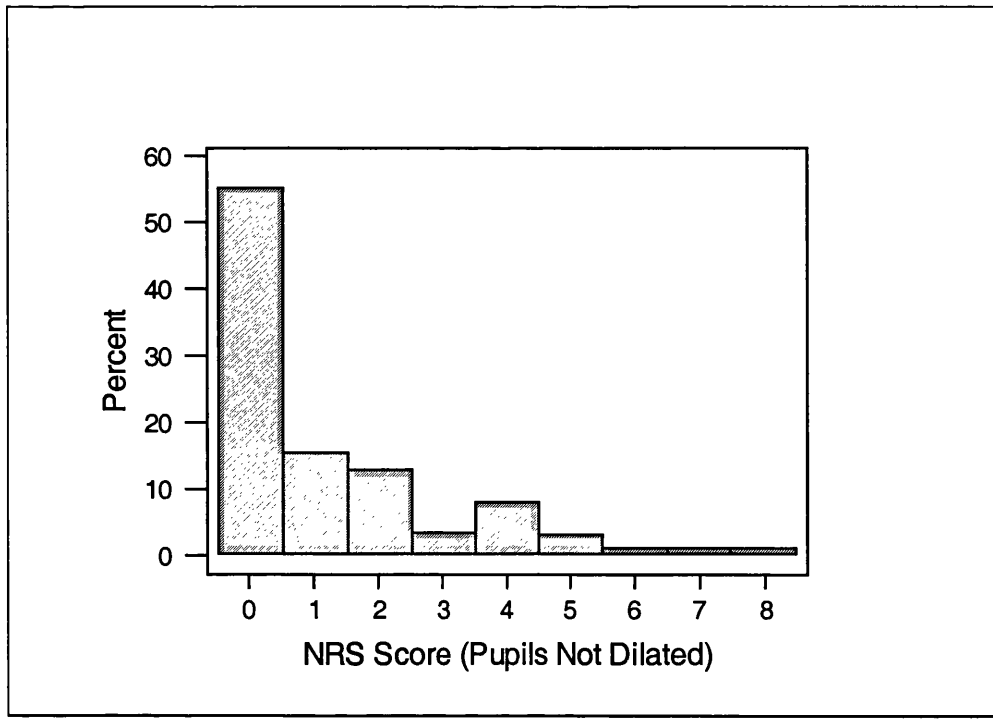


Table 5.5: Mean (median) NRS scores for each of four groups of 77 dogs and all groups combined, split by whether the dog was assessed as having dilated pupils. Groups consisted of dogs that had undergone orthopaedic (n=17) or soft tissue (n=20) surgery the previous day, had medical conditions (n=20) or were healthy (n=20). Each dog was assessed by 5 veterinary surgeons. The p-value shows results of Wilcoxon Mann Whitney test to compare median NRS scores in dogs with and without dilated pupils.

Group	Orthopaedic	Soft Tissue	Medical	Control	All Groups
Dilated	2.73 (2)	3.31 (3)	0.52 (0)	0.18 (0)	1.87 (0)
Not dilated	3.68 (3)	1.53 (1)	0.54 (0)	0.15 (0)	1.17 (1)
p-value	0.04	0.01	0.66	0.79	0.01

Figure 5.4: Boxplots of CMPS scores in a total of 77 dogs, split by whether the dog had undergone surgery. The surgical group consisted of dogs that had undergone orthopaedic (n=17) or soft tissue (n=20) surgery the previous day and the non-surgical group had either medical conditions (n=20) or were healthy (n=20). Each dog was assessed by 5 veterinary surgeons.

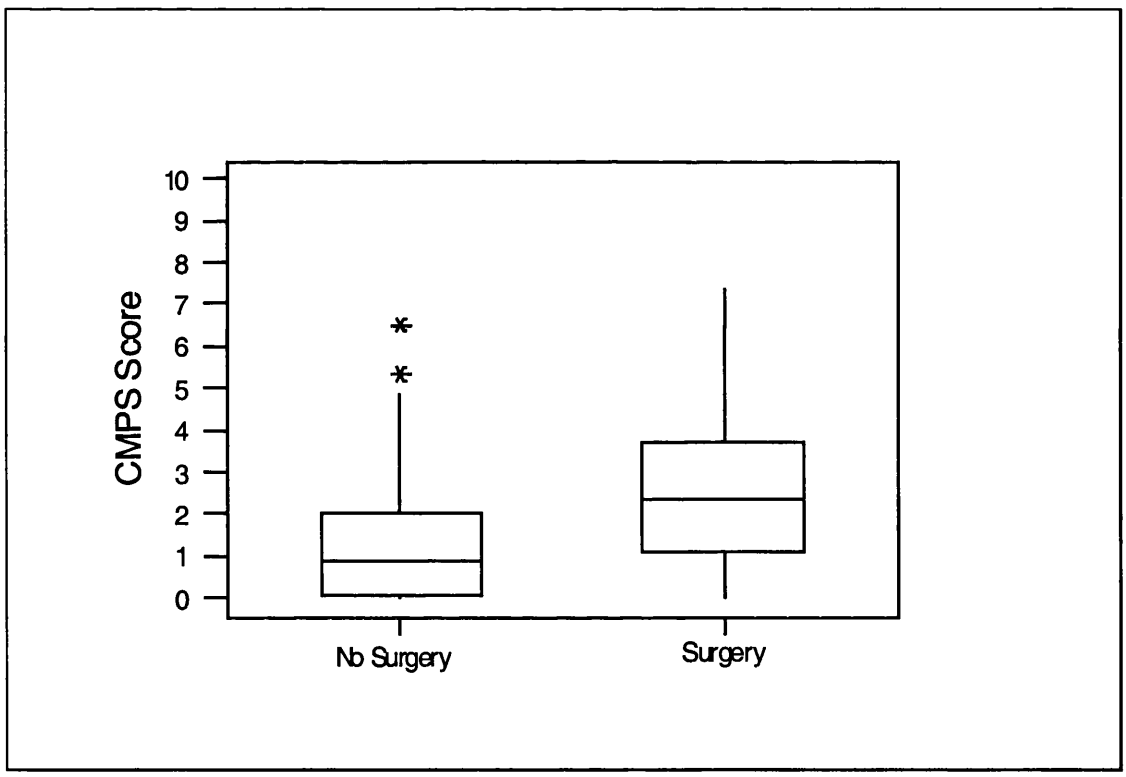


Table 5.6: Summary statistics of CMPS scores allocated to 77 dogs, split by surgical status. The surgical group consisted of dogs that had undergone orthopaedic (n=17) or soft tissue (n=20) surgery the previous day, and the non-surgical group either had medical conditions (n=20) or were healthy (n=20). Each dog was assessed by 5 veterinary surgeons.

	Surgery	Non-Surgery
mean	2.6	1.2
St Dev	1.8	1.2
Min	0	0
Median	2.4	0.9
Max	7.4	6.5

Table 5.7: Summary statistics for CMPS score for 77 dogs split into 4 groups. The groups consisted of dogs that had undergone orthopaedic (n=17) or soft tissue (n=20) surgery the previous day, had medical conditions (n=20) or were healthy (n=20). Each dog was assessed by 5 veterinary surgeons.

	Orthopaedic	Soft Tissue	Medical	Control
Mean	3.2	2.1	1.5	1.0
St Dev	1.9	1.6	1.3	1.1
Min	0.0	0.0	0.0	0.0
Median	3.0	2.0	1.2	0.9
Max	7.4	6.4	6.5	4.9

the maximum observed score was higher for the orthopaedic, soft tissue and medical groups than the control group (max=7.4, 6.4, 6.5 and 4.9 respectively). This gave an initial indication that the scores allocated using the CMPS reflected the differences in the severity of pain between the groups.

Comparison of the median CMPS scores across the groups using a Kruskal-Wallis test showed significant differences between the 4 groups (χ^2=79.98, p-value<0.001). Pairwise comparisons of the median pain scores between the groups indicated that there were significant differences in the median pain scores between all four study groups even at the 1% level (Table 5.8).

The distribution of perceived pain severity associated with surgery and medical conditions is shown in Table 5.9. The orthopaedic and soft tissue surgery groups contain more cases associated with moderate or severe pain than the medical group. The median scores differed among the four severity rating groups with the exception of moderate and severe (median = 0.9, 1.8, 2.4 and 2.4 for no pain, mild, moderate, severe pain; Table 5.10). The mean scores did not demonstrate such strong relationships, and there was a great deal of variability in the scores observed for each severity group (Table 5.10).

Comparison of the median pain scores across the four severity categories indicated that there were significant differences between the groups (χ^2=57.1, p-value<0.001). The pairwise comparison between the groups indicated that the 'no pain' group had a significantly lower median score (median=0.9) than the other groups (medians=1.8 to 2.4; Table 5.11). There were no significant differences between the other severity ratings (p-values>0.3). Thus, the relationship between the CMPS scores and perceived pain severity associated with the condition does not support the hypothesis first proposed, that higher CMPS scores would be seen where the animal's condition was perceived as being more painful.

A plot of the CMPS scores against perceived pain severity illustrated high variability in the scores associated with each severity rating (Figure 5.5). The Spearman's rank correlation coefficient between pain score and severity (correlation=0.37) was significantly different from zero, suggesting that the observed pain score increased as perceived severity increased. Hence, there is a weak relationship between pain severity and the CMPS score.

Table 5.8: P-Values for Wilcoxon Mann Whitney test used to compare median CMPS scores assigned to four groups of dogs. The groups consisted of dogs that had undergone orthopaedic (n=17) or soft tissue (n=20) surgery the previous day, had medical conditions (n=20) or were healthy (n=20). Each dog was assessed by 5 veterinary surgeons.

	Soft Tissue Group	Medical Group	Control Group
Orthopaedic	0.0001	0.0001	0.0001
Soft Tissue		0.0087	0.0001
Medical			0.0033

Table 5.9: Distribution of the perceived severity of pain associated with medical condition or surgical procedure in dogs used to examine the validity of the CMPS.

	Orthopaedic Group	Soft Tissue Group	Medical Group
Missing	1	0	5
No Pain	0	0	13
Mild Pain	2	4	1
Moderate Pain	11	10	1
Severe Pain	3	6	0

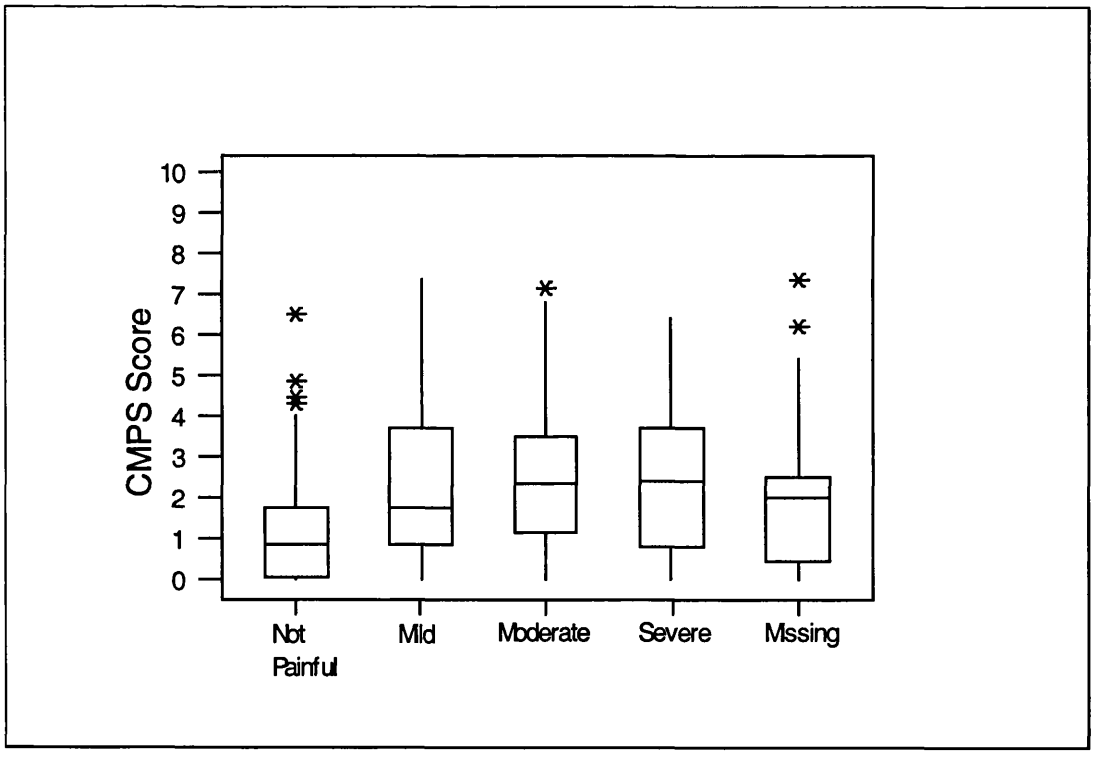
Table 5.10: Summary statistics for the CMPS scores allocated to dogs that had undergone surgical procedures or had medical conditions perceived as causing no pain, mild, moderate or severe pain, split by the perceived severity. Each dog was assessed by 5 veterinary surgeons.

	No Pain	Mild	Moderate	Severe
Mean	1.2	2.3	2.5	2.5
St Dev	1.3	2.0	1.7	1.8
Min	0.0	0.0	0.0	0.0
Median	0.9	1.8	2.4	2.4
Max	6.5	7.3	7.1	6.4

Table 5.11: P-values for Wilcoxon Mann Whitney used to compare median CMPS scores assigned to 4 groups of dogs, split by the severity of pain associated with the condition or surgical procedure undergone by each dog. Each dog was assessed by 5 veterinary surgeons.

	Mild	Moderate	Severe
No Pain	0.0006	0.0001	0.0001
Mild		0.32	0.57
Moderate			0.99

Figure 5.5: Boxplots of CMPS scores assigned to 4 groups of dogs split by the perceived pain severity associated with their medical condition or surgery. Each dog was assessed by 5 veterinary surgeons.



5.3.1.4 Investigation of the reliability of the CMPS.

The reliability of the CMPS when used in the clinical setting was examined by calculating reliability coefficients. The components of variance derived from the random effects model discussed in Section 5.2.2.5 were used to calculate overall reliability and adjusted reliability coefficients (Table 5.12).

The overall reliability coefficients of the pain measurement scale lay between 0.37 and 0.51 when used by a number of observers (Table 5.12). Thus, the proportion of measurement error in the observed scores was high compared to the variability between the animals. The error variability associated with the use of the pain measurement scale is high and therefore reliability is low. When these coefficients were adjusted to account for the multiple observers the reliability coefficients improved to between 0.70 and 0.78 (from the adjusted R values in Table 5.12). These values indicate that when adjusting for multiple observers the reliability of the scale was reasonably high. This implies, the variability between the observers was large; further exploration of this was carried out and is discussed in the second of these two studies.

5.3.2 Study 2: Generalizabiliity of the Composite Measurement Pain Scale

5.3.2.1 Exploratory analysis of variability

Summary statistics presented for both assessments (Table 5.13) indicated that there was a slight drop in the scores allocated at the second assessment for both the NRS and CMPS. The standard deviations observed in the combined NRS scores were slightly greater than in the combined CMPS scores. Standard deviations observed for each observer were greater for the NRS scores than the CMPS scores. This suggests there is less variability in the scores when pain is measured by the CMPS compared to the NRS.

Summary statistics for the pain scores allocated by each observer, over both assessments, indicated differences in the mean and median scores allocated using the CMPS and NRS (Table 5.14). In particular, observer 1 allocated lower scores than the others, and this observer's scores had a smaller standard deviation than the others did. Observer 2 assigned higher scores than the other observers when using the NRS and the difference was more pronounced with the CMPS. The standard deviations of the CMPS scores for each of the two assessments were consistently smaller than those observed for the NRS scores.

Table 5.12: Mean squares, components of variance, reliability coefficients and adjusted reliability coefficients for the CMPS used to measure pain in a total of 77 dogs, split into 4 groups. The groups consisted of dogs that had undergone orthopaedic (n=17) or soft tissue (n=20) surgery the previous day, had medical conditions (n=20) or were healthy (n=20). Each dog was assessed by 5 veterinary surgeons.

	Mean Square	Component of Variance
Orthopaedic Group		
Observer	13.57	0.75
Subject (dog)	8.24	1.41
Error	1.52	1.52
	R=0.48	R_{adj}=0.76
Soft Tissue Group		
Observer	21.61	1.09
Subject (dog)	5.13	0.92
Error	0.89	0.89
	R=0.51	R_{adj}=0.70
Medical Group		
Observer	9.80	0.12
Subject (dog)	2.97	0.50
Error	0.86	0.86
	R=0.37	R_{adj}=0.72
Control Group		
Observer	4.50	0.21
Subject (dog)	3.21	0.56
Error	0.57	0.57
	R=0.50	R_{adj}=0.78

Table 5.13: Summary statistics for CMPS and NRS scores assigned during video assessment of 12 dogs on two occasions. Summary statistics split by assessment. Each dog was assessed by 4 veterinary surgeons on each occasion.

	CMPS		NRS	
	Assessment 1	Assessment 2	Assessment 1	Assessment 2
Mean	2.1	1.9	3.0	2.5
St Dev	1.44	1.48	2.31	2.67
Min	0	0	0	0
Median	1.9	1.7	3.0	2.0
Max	5.6	5.0	8	8

Table 5.14: Summary statistics for CMPS and NRS scores assigned during video assessment of 12 dogs on two occasions. On each occasion, each dog was assessed by 4 observers. Summary statistics split by observer.

Scale		Observer 1	Observer 2	Observer 3	Observer 4
CMPS	Mean	1.5	2.5	2.2	1.8
	St Dev	1.13	1.50	1.55	1.51
	Min	0	0	0	0
	Median	1.4	2.5	1.6	1.9
	Max	4.1	4.6	5.6	4.4
NRS	Mean	1.5	5.3	2.0	2.3
	St Dev	1.69	2.23	2.21	1.98
	Min	0	0	0	0
	Median	1	6	1	2
	Max	5	8	7	6

Summary statistics for the change in scores assigned between the two assessments indicate that, of the four observers, two assigned lower CMPS scores on the second assessment than the first, whereas three assigned lower NRS scores on the second assessment (Table 5.15). In addition, the standard deviation for the change in scores was smaller for the CMPS than the NRS in three of the four observers.

5.3.2.2 Generalizability over observers and time

A random effects model (Equation 5.4) was fitted to the CMPS and NRS pain scores. The expected mean squares for each factor are shown in Table 5.16 and the observed values for the mean square and components of variance are in Table 5.17. For both the CMPS and NRS, the variance component associated with the factor 'Dog' is the largest (0.94 and 2.59 for CMPS and NRS respectively) indicating that most of the variability in the data is due to differences between the animals. The components of variance also indicated that the 'Dog*Observer' interaction term was small relative to the other terms in the model (0.13 and 0.12 for CMPS and NRS respectively) implying that the observers' scores were consistent across the different animals. The variance associated with the NRS 'Dog*Time' interaction was smaller than the corresponding value for the CMPS (0.32 and 0.04 for CMPS and NRS respectively). For both scales the error component of variance was relatively large compared to the other terms (0.67 and 1.25 for CMPS and NRS respectively), which suggested that other sources of variance influenced the pain scores and these had not been examined in this study.

From these components of variance the generalizability coefficients for both pain measurement scales were calculated (Table 5.18). The generalizability coefficients indicated that both the CMPS and NRS exhibited a large amount of variability in the observed pain scores, both between observers (generalizability coefficients = 0.61 and 0.66 respectively) and over time (generalizability coefficients = 0.52 and 0.68 respectively).

5.4 Discussion

5.4.1 Study 1: Validity and reliability of the CMPS

The results explored in Section 5.3.1 fall into three sections: investigating the utility of the physiological parameters, investigating the validity of the CMPS and investigating the reliability of the CMPS.

Table 5.15: Summary statistics for change in CMPS and NRS scores between first and second assessment of pain in 12 dogs. Each dog was assessed by 4 observers. Summary statistics are split by observer.

Scale		Observer 1	Observer 2	Observer 3	Observer 4
CMPS	Mean	-0.6	0.2	-0.7	0.7
	St Dev	1.39	1.18	1.48	1.59
	Min	-2.44	-2.03	-3.65	-2.52
	Median	-0.8	0.4	-0.1	0.8
	Max	1.51	1.73	0.79	3.19
NRS	Mean	-0.2	0.7	-0.3	-2.2
	St Dev	0.58	1.67	2.10	1.64
	Min	-1	-3	-5	-5
	Median	0	0.5	0	-2
	Max	1	3	3	0

Table 5.16: Expected mean squares derived from the random effects model fitted to the CMPS and NRS scores in a video assessment of post-surgical pain in 12 dogs, by 4 observers on 2 occasions.

Source	Expected Mean Square
Observer	$\sigma_\varepsilon^2 + 2\sigma^2_{obs*dog} + 12\sigma^2_{obs*time} + 24\sigma^2_{obs}$
Dog	$\sigma_\varepsilon^2 + 2\sigma^2_{obs*dog} + 4\sigma^2_{dog*time} + 8\sigma^2_{dog}$
Time	$\sigma_\varepsilon^2 + 4\sigma^2_{dog*time} + 12\sigma^2_{obs*time} + 48\sigma^2_{time}$
Observer*Dog	$\sigma_\varepsilon^2 + 2\sigma^2_{obs*dog}$
Observer*Time	$\sigma_\varepsilon^2 + 12\sigma^2_{obs*time}$
Dog*Time	$\sigma_\varepsilon^2 + 4\sigma^2_{dog*time}$
Error	σ_ε^2

Table 5.17: Mean Squares and components of variance derived from the random effects model fitted to CMPS and NRS scores in a video assessment of post-surgical pain in 12 dogs, by 4 observers on 2 occasions.

	CMPS		NRS	
Factor	Mean Square	Component of Variance	Mean Square	Component of Variance
Dog	4.22	0.94	22.34	2.59
Observer	9.69	0.06	69.00	2.51
Time	0.66	0	6.00	0
Dog*Observer	0.92	0.13	1.49	0.12
Dog*Time	2.54	0.32	1.39	0.04
Observer*Time	1.94	0.16	8.56	0.61
Error	0.67	0.67	1.25	1.25

Table 5.18: Generalizability coefficients over time and observers for pain scores assigned using the CMPS and NRS in a video assessment of post-surgical pain in 12 dogs, by 4 observers on 2 occasions.

Scale	Generalizability over Observers	Generalizabity over Time
CMPS	0.61	0.52
NRS	0.66	0.68

When investigating the relationship between the physiological signs and pain, the pain measurement scale used was the NRS. Previous work has shown that such subjective methods lack generalizability when used by multiple observers (Chapter 2), yet the VAS, NRS and SDS are widely accepted within the veterinary literature. The NRS was used in the study discussed here as it was thought to be most acceptable of the three subjective scales currently available (Section 2.4).

Literature concerned with pain recognition, particularly the assessment of pain in experimental animals, cites a number of physiological signs regarded as being indicative of pain. These factors include changes in heart rate, changes in respiratory rate and pupil dilation (Morton and Griffiths, 1985; Bateson, 1991). In their guidelines for assessing the welfare of experimental animals, Morton and Griffiths state that cardiovascular signs such as an alteration in heart rate, pulse quality and peripheral circulation are signs of pain, distress or discomfort in experimental animals. In addition, respiratory signs (abnormal breathing pattern, rate and depth), digestive signs (loss of bodyweight, vomiting and jaundice), locomotory (twitching, fitting, tremors, lameness, unsteady gait and pupil dilation) and a number of other miscellaneous signs were also said to indicate pain, distress or discomfort in experimental animals. The clinical signs that are said to be measurable and recommended for use when assessing welfare include changes in the rate and strength of the pulse, salivation, vomiting, lameness, twitching and convulsions. The authors do not provide any data examining the relationship between these signs and welfare, and so the validity of their inclusion is difficult to judge. It should be considered that this article was concerned with the general welfare of the animals in question and not specifically their pain, therefore it may be reasonable that this broad range of clinical signs are appropriate in this setting.

The physiological signs cited as being indicative of pain in the paper by Bateson (1991) include pupil dilation, blood pressure, respiratory rate, body temperature, sweating and muscle tone. The author does not give any indication of how the relationships between pain and these signs were investigated or the results of that investigation and does not provide any references to such work. The apparent lack of empirical data in support of the relationship between pain and physiological signs raises questions around the validity of these signs as indicators of pain. Nonetheless, physiological signs have been accepted as being indicative of pain and have been incorporated into a number of the recently published pain measurement scales.

The Colorado University scale and the University of Melbourne Pain Scale (UMPS) discussed in Chapter 1, both make use of physiological signs (Hellyer and Gaynor, 1998; Firth and Haldane, 1999). Hellyer and Gaynor (1998) include increased heart rate and respiratory rate, compared to pre-surgical levels, in their assessment of post-surgical pain. The authors also define scores allocated to the patients that are dependent on the size of change from pre-surgery values, for example an increase in heart rate between 16% and 29% is assigned a score of 1, 30% to 45% is assigned a score of 2 and so on, with similar bands defined for increases in respiratory rate. The physiological signs associated with pain and included in the UMPS are dilated pupils, increased heart rate, respiratory rate, temperature and salivation. The authors also define ranges for the increases in heart and respiratory rates and allocate scores accordingly, although these ranges differ from those seen in Hellyer and Gaynor (1998). For example, an increase of greater than 20% is allocated a score of 1, greater than 50% has a score of 2 and so on. The authors' definition of specific ranges of change for the physiological signs and scores associated with these suggest confidence in the relationship between physiological signs and pain intensity. Neither paper provides any investigation into the derivation or validity of these ranges or the scores allocated, nor do they provide any references to such work. Thus, the basis for the acceptance of these factors as indicating pain appears to be anecdotal.

In any clinical study, statistical significance does not automatically equate to clinical relevance. Literature indicates that a high correlation coefficient (approximately 0.6 and greater) is required between two variables before the relationship can be said to be of any benefit in a clinical setting (Nunnally and Bernstein, 1994). Hence, it is often more informative to examine the size of the correlation coefficient observed than the associated p-value, as the correlation coefficient provides more information about the strength of the relationship between variables.

The results showed that the correlation between heart rate and NRS score was significant in the medical group, although the correlation coefficient was small (correlation coefficient=0.38). For this group, NRS scores of 3 or higher were allocated on only 7 occasions and these animals also had high heart rates, which contributed to the correlation between the two (Figure 5.1). Similarly high heart rates were seen in animals assigned low NRS scores, which suggested that, although some relationship does exist between the two variables, heart rate may not be predictive for NRS scores. It is interesting that a wide range of heart rates were recorded in the control group of dogs. These dogs were thought to be pain free, thus observing high heart rates in this group illustrated the weakness in

using heart rate as an indicator of pain. The apparent lack of stability in the relationship and the small correlation between heart rate and NRS score indicated that the utility of heart rate in the evaluation of pain would be limited. Thus, heart rate was not included in the CMPS.

Work that has been undertaken to explore the relationship between heart rate and pain in children has demonstrated a positive relationship. In their review article Sweet and McGrath (1998) cite a number of studies which have demonstrated that following acutely painful procedures, such as heel lancing, the heart rate generally decreased immediately following the procedure then increased slowly and remained elevated for a short period (on average three and a half minutes). However, the work undertaken in paediatric pain in this area focuses primarily on the changes in heart rate over a short period, following acutely painful procedures. No articles exploring the relationship between heart rate and longer term pain, such as post-operative pain, are discussed.

In study one, the correlation coefficients between respiratory rate and NRS score were small and negative for each group, although the value for the soft tissue group was significant. The negative value of the correlation coefficients indicated that as a patient's pain score increased their respiratory rate decreased. This relationship is contrary to common belief and to the relationship advocated in the other pain measurement scales (Hellyer and Gaynor, 1998; Firth and Haldane, 1999). Examination of the data illustrated that a wide range of respiratory rates were seen across the range of NRS scores and that there was a great deal of variability in the relationship between the two variables. This, coupled with the unexpected direction of the relationship between respiratory rate and NRS score, indicated that the inclusion of respiratory rate in the CMPS would not contribute to the validity of the scale.

Investigations of the relationship between respiratory rate and paediatric pain have identified similar problems as conflicting relationships have been found. For example, Craig *et al* (1993) indicated a decrease in respiratory rate following heel lancing, whereas Howard *et al* (1994) demonstrated that respiratory rate increased following circumcision. It appears that the relationship between respiratory rate and pain is unstable, and so the contribution that the inclusion of this measure would make to the overall pain score would be difficult if not impossible to define. For this reason it should not be used in the assessment of pain.

The pain measurement scale proposed by Firth and Haldane (1999) indicated that pupil dilation could be used as an indicator of pain. In the study reported in this chapter, the median NRS scores for dogs with and without dilated pupils were found to differ in only one of the four groups (soft tissue). This suggested that the relationship between pain score and pupil dilation was tenuous. Further examination of the data showed that higher pain scores were associated with dilated pupils in the soft tissue surgery group, while the orthopaedic group showed higher scores in animals without dilated pupils. This suggested that the relationship between pain and pupil dilation was not consistent, and pupil dilation was not included in the CMPS.

When examining the validity of the CMPS the first construct investigated was whether the scores differed between animals that had undergone surgery and those that had not. This provided initial support for the validity of the CMPS, although it could be argued that this level of differentiation in pain scores was not precise enough. To explore the validity of the scale further it was necessary to define a construct that would demonstrate a finer level of differentiation in the scores. This was investigated by exploring the differences in pain scores associated with the four study groups. It was hypothesised that animals in the orthopaedic group may experience the most pain, animals in the soft tissue group may experience the second most intense pain, the medical group third and the control group may experience the least pain. When examining the perceived pain intensities associated with each group (Table 5.9), there was little distinction between the soft tissue and orthopaedic groups. Nevertheless, the CMPS scores differed between the groups, and in particular there were significant differences between the pain scores for the orthopaedic and soft tissue groups, which provided some support to the validity of the CMPS.

The investigation of the validity of the CMPS by examining the perceived severity of pain associated with the various conditions and surgical procedures was not as conclusive as the two previous constructs. The results indicated that there were significant differences in the median pain scores between the no pain group and the others, but the median scores did not significantly differ between the other groups. A plot of pain scores against severity and the associated correlation coefficient indicated that there was a positive, if weak, relationship between these two variables. Therefore, this construct provides some evidence that the CMPS is valid when used in a clinical setting.

The overall results of the investigation of the validity of the CMPS were encouraging in that they support the hypothesis that the scale is valid when used in a clinical setting. The

constructs which were explored were highly inter-related and further work exploring additional constructs would be required before the CMPS could be universally accepted as valid. In particular, the construct of pain severity was dependent on the subjective judgments of a number of veterinary surgeons regarding the pain they associate with certain procedures. Alternative measures that could be explored when investigating construct validity are the duration of hospital stay following surgery, consumption of analgesia or pain as assessed by a veterinary nurse. It could be hypothesised that these would be related to the intensity of pain and exploring them would provide a further insight into the validity of the scale. However, such investigations were not possible within the limitations of the study reported here.

Of the composite measurement scales developed for use in animals, only the UMPS has undergone any investigation of validity (Firth and Haldane, 1999). The authors examined whether the allocated scores differed between four analgesic treatments. In their study, pain was assessed repeatedly over time and the area under the pain/time curve (AUC) was used to summarise the patient's pain score. Use of the AUC effectively removes any variability within an individual's scores over time, hence the variability in the summary score is much reduced compared to the individual assessments. The authors supported the validity of the UMPS as it showed significant differences between some, although not all, of the four analgesic treatment groups. The conclusion that this tool can differentiate between the four treatment regimes may be true, but only when AUC up to 4 hours and AUC up to 12 hours are used as the outcome. From this investigation, it cannot be said that this method can differentiate between treatments when used at a single time point and such a claim requires further investigation.

In the investigation of reliability of the CMPS, the coefficients for each group indicated that the reliability of the scale was low (between 0.31 and 0.51), and hence the error variability within the observed scores was high. From the components of variance, it could be seen that the observer variance was large compared to the subject variance for both surgical groups. This suggested that a great deal of variability in the scores was due to differences between the observers. This was verified by the increase in the reliability coefficients to between 0.70 and 0.78 when adjusted for multiple observers. This coefficient is equivalent to the reliability of the measurement scale when the score for each animal is taken as the average of the scores allocated by the five observers.

Literature discussing the nature of pain indicates that acute pain fluctuates over relatively short periods of time and that the behaviours exhibited because of that pain may also

change (Reading, 1983; Beyer and Knapp, 1986). In study 1 the observers were permitted to visit the wards at any time within a 4-hour time window and each observer examined the dogs independently. The pain an animal was experiencing could have changed over that time period and between the examinations made by the different observers. Consequently, the inter-observer variability may not be due purely to error between the observer but also real changes in the animal's condition. It is probable that the design of this study caused the reliability of the CMPS to be underestimated, thus the time frame over which pain is to be measured must be taken into careful consideration when designing any future studies.

In addition to the possible changes in the underlying pain experienced by the animal, it should also be noted that the observers were not provided with the definitions of the behaviours included in the scale. This was to allow the observers to use their professional judgment when making the assessments. As a result, it is possible that the large variability between observers was partly due to differences in the interpretation of the items included in the scale.

These issues illustrate the difficulties in investigating the reliability of a measurement scale concerned with an attribute, such as pain, that can only be measured indirectly and can change over a short period. Both of these issues suggest that improved study design and observer training could increase the reliability of the CMPS. This aspect of the scale's performance was investigated further in the second study undertaken.

Investigation of the reliability of the measurement scales currently available in veterinary medicine is limited. The numerical rating scale developed by Conzemius *et al* (1997) examined the agreement between two observers when using the scale. Although this shows excellent agreement between two independent observers when assessing post-operative pain (Kappa coefficient greater than 0.90), the authors did not provide any further exploration of the pain scores allocated by the two observers which may have been informative when assessing the performance of the scale. Firth and Haldane (1999) claimed excellent agreement between two observers using the UMPS when examined using the method described by Bland and Altman (1986). However, recall that the authors used AUC as their outcome, which has the effect of removing all intra-animal variability. This reduction in the variability in the patient's outcome could increase the agreement between the two assessors compared to examining individual pain scores. Therefore, these results are not representative of the method's performance at a single time point. Secondly, during Firth and Haldane's study the physiological variables were recorded by the first assessor, only. The second assessor did not record the physiological variables as the

assessments were made via video recordings. The data for the physiological variables were transcribed from the first to the second assessor's scores. These variables accounted for up to 11 out of a possible score of 27, i.e. over 40% of the possible scores were guaranteed to be identical between the two observers. The agreement between the two observers excluding the physiological parameters was not presented. It is therefore impossible to ascertain whether this method would possess high inter-observer reliability when used by a number of observers. The reliability of the measurement scale when used independently by a number of observers is as yet unknown and requires further investigation.

Investigations of the reliability of the MPQ have demonstrated that this scale can be used consistently over time. Repeated administration of the MPQ in cancer patients showed an average of 75% agreement between the first two administrations, however the level of agreement changed over the course of one week (ranging from 66% to 80%) (Hunter *et al*, 1979). Similar results were demonstrated by the developers when the scale was used repeatedly over 3 days (Melzack, 1975). Neither of these investigations made use of the standard methodology of reliability coefficients and instead examined the percentage agreement between repeated administrations of the scale. The authors did not present this data and therefore a comparison of the performance of the MPQ with the CMPS is difficult to ratify.

One consideration in the design of the first study reported in this chapter is that the animals were treated according to standard clinical practice. Any animal thought to be suffering an unacceptable level of pain was treated with analgesics and this was the case on 3 occasions during the assessment period. Consequently, the majority of animals were judged not to be in severe pain. Only on 21 occasions was a pain score higher than 5 on the NRS scale awarded. The results described can only be assumed to hold true for animals with pain scores in a similar range to those recorded in this study. Further investigation would be required to examine the performance of the CMPS at higher intensities of pain.

5.4.2 Study 2: Generalizability of the Composite Measurement Pain Scale

The investigation of the generalizability in a pain measurement scale developed for use in the veterinary field is a novel approach to exploring the performance of any scale in this area. Although classical test theory dictates that two independent studies would have been required to gather the same information (Nunnally and Bernstein, 1994), the methodology of generalizability theory has allowed the performance of the CMPS over observers and over time to be explored in a single study.

One consequence of the novelty of this approach was that no research has been published using this method within veterinary medicine. Therefore, interpretation of the results is not straightforward as comparison of the results with other veterinary studies is not possible. Nevertheless, the approach of generalizability theory has been used within psychological research and to a lesser extent within the medical literature where acceptable levels for generalizability coefficients have been documented. Evans *et al* (1981) indicated that generalizability coefficients greater than 0.80 were required before a measurement scale could be accepted as suitable. However, the generalizability of a scale is not a fixed property and is dependent on the sample of subjects included in the study. Such an absolute cut off for the acceptance of a scale may not be appropriate (Streiner and Norman, 1995)

The design of this second study was comparatively simple as only the generalizability over observers and over time was examined. The two generalizability coefficients calculated from the observed scores were directly equivalent to the inter-observer reliability and test-retest reliability coefficients of classical test theory (Streiner and Norman, 1995). Interpretation of these coefficients gave an insight into the performance of the CMPS when used to measure post-operative pain in dogs. The primary function of these generalizability coefficients was to identify sources of variability within the observed scores. Having identified the sources of variability, possible measures that could be taken to reduce the variability (and thus enhance the performance of the scale) could be explored.

The generalizability of the CMPS over observers was 0.61 indicating that 39% of the observed variability was due to differences between the observers and other measurement error. The same analysis indicated that the generalizability of the NRS over observers was broadly similar (0.66). This study was the first time that any of the observers had used a composite measure type of pain measurement scale, whereas all were familiar with the NRS. It is possible that the observers' lack of familiarity with the CMPS influenced its performance and that the performance of the CMPS could be improved by providing the observers with formal training in its use. Such guidelines for the use of composite measurement scales are common in psychological measurement where a manual is provided describing the use of the method. For example, the Speilberger anxiety state and trait scales and the CES-D depression scale both have such usage guidelines (Speilberger *et al*, 1972; Radloff, 1977). The construction of such a usage manual for the CMPS is an aspect of the scale development that should be explored further in the future.

The generalizability coefficient over time for the NRS was 0.68, and for the CMPS it was 0.52, indicating that the NRS performed better than the CMPS when used over time. This is not as expected, since it was anticipated that the CMPS itself and the provision of item definitions would improve the consistency in the scores across time. However, as noted in Section 5.3.2.1, none of the observers had used this type of measurement scale prior to participation in this study and were therefore unfamiliar with its use. The variability over time suggests that the observers' scores changed between the two assessments, although the video recording guaranteed that the pain behaviours did not change over time. Discussions with the four observers following completion of the study indicated that they felt the assessments were simpler to carry out on the second occasion. This indicated there could be some learning effect in the use of the scale that was causing variability in the scale scores over time.

A number of features of the study design could also have influenced the results, for example the assessment of pain in animals via video recordings has not previously been explored. There is no evidence to indicate whether this method provides an adequate representation of assessments made in the clinical setting or not. From a purely intuitive point of view it seems reasonable to suppose that it is more difficult for the observers to make accurate assessments of an animal's behaviour via a video recording than it would be in a clinical setting. It is anticipated that the use of video recording introduced a source of variability to the assessment procedure that may not have been present had the examinations been carried out in the hospital wards. One reason for this is that the use of a video recording when assessing pain may place an additional barrier between the animal and the observer, thus making the behaviours more difficult to identify. When making an assessment in the clinic the observer would have an opportunity to examine the animal for slightly longer or from an alternative view-point, if they felt this was appropriate, which was not possible when using the video recording.

During the assessments, the observers were provided with definitions of the items included in the scale on a sheet separate from the main measurement scale. The definitions were provided in an effort to improve the observers' interpretation of the items, and hence reduce the variability between the observers and ensure that the observers applied the items in the same way on the two occasions. On discussion with the observers following completion of the study, it became apparent that they seldom referred to the definitions provided and instead used their judgment to identify the behaviours included in the CMPS. This suggested that the anticipated benefit of providing the observers with definitions of

each item in the scale would not have been realised in this study. One possible way that this might be overcome would be to incorporate the definitions in the CMPS itself to ensure that the observers considered the definitions when making their assessments. In addition, during training in the use of the scale, the item definitions could be highlighted and the importance of their use when making assessments emphasised.

The data presented suggested that the generalizability of the CMPS and NRS are comparable over observers, but the NRS has slightly better generalizability over time than the CMPS under these test conditions. However, a number of possible improvements in the study design and use of the CMPS have been identified from the analysis and discussions following the completion of these studies. It is anticipated that the implementation of these improvements would benefit the performance of the CMPS.

6. General Discussion

It is well accepted that the ability to measure pain is of paramount importance in patient care and pain research, yet to date there has been little formal exploration of how well pain is measured in animals. Although pain is said to be a complex and internal experience, the tools currently used in veterinary medicine primarily comprise simple subjective rating scales such as the VAS, NRS and SDS. The issues surrounding the use of subjective scales in pain measurement in humans are recognized and are discussed in Chapter 2 (Chapman, 1976; Gracely, 1980; Carlsson, 1983; Kaiko *et al*, 1983). The use of these methods in veterinary medicine has not been as thoroughly explored. The studies described in this thesis had two main aims, firstly to determine whether the subjective tools currently used to measure pain do so adequately and secondly, to assess whether pain measurement in animals could be improved by developing a composite measurement method.

All the investigations undertaken focused exclusively on the measurement of clinical pain in dogs, primarily following surgery. This work concentrated on one species since the pain behaviours of animals are believed to be species-specific (Morton and Griffiths, 1985; Sanford, 1986). The dog was chosen because a large number of dogs are treated at the Glasgow University Veterinary Hospital where this work was undertaken. Although the focus was on dogs, it is anticipated that the methodology used in the construction and investigation of the composite measurement pain scale (CMPS) could be transferred to other species. The measurement of pain in a clinical setting was investigated, rather than in an experimental one, as it provides an insight into the performance of the scale when used in practice.

The performance of the subjective VAS, NRS and SDS scales when used to measure animal pain was not encouraging. None of the scales possessed acceptable measurement properties when used in a clinical setting. It was shown that the generalizability over observers was low indicating that the scales were used inconsistently among observers. The results of the investigations reported in Chapter 2 were not consistent with the veterinary literature to date, where agreement between observers using the VAS and NRS has been shown to be good (Reid and Nolan, 1991; Welsh *et al*, 1993). This may be due to differences in the design of the studies and in the attributes under investigation. In their study, Reid and Nolan used two observers who were trained in the use of the VAS. It is likely that training the observers before they used the VAS would improve the agreement between them. However, no formal procedures for using the VAS have been described,

and so, unless training guidelines were developed and universally accepted it would be difficult to ensure consistency between observers and institutions. Thus, for an individual research study the training of observers may improve consistency, but it does not guarantee consistency in the use of the scale between studies or when used in patient care. Welsh *et al* (1993) explored the VAS and NRS in the measurement of lameness in sheep. The authors demonstrated that two observers, trained in the use of the VAS and NRS, allocated consistent scores using both scales, but that the scores observed were not interchangeable as there was no unique relationship between the scales. Although there is a relationship between lameness and pain, the performance of the subjective scales when measuring lameness is not necessarily indicative of their performance when measuring pain. It could be argued that lameness is simpler to assess than pain as the extremes of the scale are more easily determined, i.e. the animal is sound and bearing full weight on all legs or the animal does not bear any weight on one leg. Thus, the improved performance of the scales when used to measure lameness compared to pain could be expected. In addition, the results of the study presented in Chapter 2 suggested that pain behaviours may have been less prominent during the assessment in the immediate post-operative period. Future studies investigating the measurement properties of pain scales may benefit from avoiding making assessments during a period when the animal may be under the influence of anaesthetic or sedatives administered during surgery. Comparisons between these results and the performance of similar subjective scales in human medicine are difficult since human patients use the scales as self-rating tools.

The poor performance of the VAS, NRS and SDS indicated that these subjective scales did not provide adequate measurement of post-operative pain in dogs, and so an alternative was sought. One approach that has been advocated in human medicine and has recently become recognised in veterinary medicine is the use of composite measurement scales (Guyatt *et al*, 1992; Conzemius, 1997; Hellyer and Gaynor, 1998; Firth and Haldane, 1999). The McGill Pain Questionnaire (MPQ) is one of the most well known scales of this type used in the measurement of human pain. The construction of the CMPS described in this thesis follows the methods described in the development of the MPQ (Melzack and Torgerson, 1971; Melzack, 1975). The MPQ was developed in the 1970s as a means of allowing patients to express their pain experience by identifying words that best describe their pain. The words form three main factors, which represent the hypothesised dimensions of pain (affective, evaluative and sensory). Within each dimension, words thought to be similar are further grouped into categories. A patient's score is calculated from weights assigned to the words that they choose. The construction of the MPQ

adheres to the steps laid down in the psychological and educational literature (Nunnally and Bernstein, 1994; Coste *et al*, 1995; Streiner and Norman; 1995). The benefit of taking a composite approach to the measurement of an unobservable construct such as pain is that, in psychological research, composite measurement scales have been shown to be more reliable, more objective and more valid than subjective methods (Nunnally and Bernstein, 1994). The CMPS is not the only scale that takes a composite approach to pain measurement in veterinary medicine, however the scale is novel in its construction since no other method developed for use in veterinary medicine has exploited the techniques developed in psychometric and measurement theory.

The construction of the CMPS was described in Chapters 3 and 4. The resulting scale was aimed at providing veterinary surgeons with a well defined examination procedure and list of items that would allow the animal's behaviour to be recorded accurately. The categories of behaviours in the CMPS included posture, comfort, vocalisation, attention the animal was paying to its wound or painful area, demeanour, mobility and response to touch. Each category contained between two (comfort) and seven (demeanour) different items. Within each category, the items described a particular behaviour and were associated with differing intensities of pain. When examining the animal in question, the veterinary surgeon was asked to pick the one item from each category that best described the animal's behaviour. From this, an overall pain score could be calculated.

Despite the similarities between the MPQ and the CMPS, there are a number of fundamental differences underlying the scales. The MPQ was aimed at providing patients with a means of effectively communicating their pain experience (Melzack, 1975). Consequently, a tool encompassing all possible aspects of pain was required. When constructing the MPQ there was no need to simplify the tool or exclude items; all possible pain descriptors were included. In contrast, the CMPS was constructed to provide a method for recording pain-related behaviours in dogs, in a concise way that could be used easily in a clinical setting. To facilitate the use of the scale, expressions thought to be very similar in meaning or pain intensity were combined or removed. This was carried out using statistical analyses to explore the similarities in pain intensity between the items. To ensure that the content validity and clinical relevance of the scale were maintained any possible changes were discussed with a panel of experts. This is the process discussed in Chapter 3.

In addition, in human medicine, pain is said to be multidimensional in nature and the MPQ was designed to explore three different dimensions of pain (Melzack, 1975). However, the multidimensionality of the pain experience in animals is questionable. It is certainly accepted that animals experience pain but whether they can differentiate between the affective, evaluative and sensory dimensions proposed in human pain is an issue for debate. Since pain in animals is potentially unidimensional, providing multidimensional measurement is not a priority. Rather than taking a multidimensional approach the CMPS was aimed at measuring animal pain in a more objective manner than was previously available.

Many of the items included in the CMPS are used in other behaviour-based pain measurement tools, for example, the method developed by Colorado State University includes vocalisation, movement, unprovoked and interactive behaviour, heart rate and respiratory rate (Hellyer and Gaynor, 1998). Similar categories are included in the guidelines for pain recognition defined by Morton and Griffiths (1985). More recently, Firth and Haldane (1999) included activity, posture and vocalisation among others in their pain measurement tool. This commonality between the pain measurement literature and the CMPS indicated that the items included are considered be to indicative of pain by the wider veterinary community. In particular, the CMPS has captured the behaviours and signs thought to indicate pain by those researchers interested in pain measurement in animals, and by veterinary practitioners. This indicated that the methodology used in the scale development was appropriate and supported the content validity of the CMPS.

The validity and reliability of the CMPS were explored by carrying out two studies where veterinary surgeons used the scale to measure pain in a number of dogs. In the first of the two studies reported in Chapter 5, five veterinary surgeons independently examined a total of 80 dogs. The animals comprised four groups; those that had undergone surgery (split into orthopaedic and soft tissue), medical cases and animals that were thought to be sound. The results of this study supported the validity of the CMPS, as there were significant differences in the scores allocated to animals that had undergone surgery and animals that had not. The scores also differed significantly among the four groups included in the study. Slight evidence of a positive relationship between the perceived severity of surgery and the observed scores was seen, although this was non-significant at the 5% level. These preliminary results lend weight to the validity of the scale. Further investigation is required before the CMPS could be accepted as fully valid, which would require the scale to be used within the veterinary community so enabling other groups to appraise its performance.

In that first study, the overall reliability of the CMPS was shown to be relatively poor, although when the reliability coefficient was adjusted for multiple observers a great improvement was seen. This indicated that a large proportion of the variability within the scores was due to perceived differences between the observers. This variability may have been partly due to real changes in the animals' pain intensity, since the observations were carried out at any time within a 4-hour period. The observer generalizability of the CMPS would almost certainly have been improved if the times at which the assessments were made had been more tightly controlled. Taking this into consideration, a second study was designed and carried out, as described in Chapter 5.

In the second study 4 veterinary surgeons used the CMPS and NRS to measure pain in 12 dogs via video recordings of the examination procedure. This method allowed the measurements to be made on two occasions under identical conditions, thus the generalizability of the scales could be explored. The generalizability of the CMPS and NRS over observers and time were investigated. The results indicated that there was considerable variability between the observers and the generalizability of the CMPS over observers was comparable to the NRS.

The only other method that has undergone any similar investigation is the UMPS (Firth and Haldane, 1999). The authors stated that the scale could differentiate between analgesic treatments and was reliable over observers. These conclusions were challenged in Chapter 5, when for example, the authors' use of the area under the pain vs. time curve as a summary of pain was questioned as it distorts the variance structure within the data and thus the reliability of the scale. The issues raised concerning this article demonstrate the need for careful consideration of the study design and investigation of the psychometric properties of pain measurement scales.

Investigations of the CMPS must be thoroughly explored before the performance of the scale can be fully understood. The results of both studies reported in Chapter 5 demonstrated a great deal of variability in the CMPS scores between the observers. In the first study, it was suspected that this could have been due partly to the lack of item definitions. However, the generalizability of the CMPS over observers was still low in the second study, where definitions were available and the anticipated improvement in generalizability by providing item definitions was not demonstrated. Nevertheless, the generalizability of the CMPS was comparable to that of the NRS. Discussions with the participating veterinary surgeons highlighted some problems with the use of the CMPS

definitions. It was considered that the performance of the CMPS could be improved further by finding a mechanism to ensure the observers were aware of and used the definitions when making their assessments. One solution could be to redesign the CMPS form to incorporate the item definitions, and then to reassess the generalizability of the scale. Such changes would not be possible for the NRS as it does not contain individual items that would benefit from clear definition. A final consideration was that the observers who took part in the studies here were familiar with the use of the NRS, but none had previously used the CMPS, which could also have had a detrimental effect on the CMPS results observed.

The second study illustrated that, despite the pain state of the animals being identical over the two assessments, the generalizability over time was low, for both scales. After the study, the observers reported that they found the CMPS easier to use at the second assessment as they became more familiar with the scale. This suggested that there was a learning process involved in the use of the CMPS. Therefore, it is possible that the coefficients of generalizability over time presented in Chapter 5 do not reflect the true performance of the CMPS since any learning effect would have caused variability in the scores. This may have been overcome by carrying out additional assessments. The scores from the first few time-points could have been treated as a training exercise and the generalizability coefficients calculated on the assessments made at later time-points. This would have allowed any learning curve to be identified and the generalizability coefficients may have given a better reflection of the true performance of the scale. When the study was carried out, the observers were familiar with the use of the NRS and made no comments about it being difficult to use. This implies that the performance of the NRS did not suffer from a learning effect, and the observed low generalizability over time was an accurate reflection of the scale's properties. Thus, it is unlikely that the performance of the NRS would benefit from a training programme to the same extent as the CMPS.

Both of the above issues suggest that training in the use of the CMPS would have benefitted the observers and could have improved the performance of the scale. Such training or usage guidelines are commonly developed for composite measurement scales; for example, the Speilberger anxiety scales and the CES-D depression scale have guidelines (Speilberger *et al*, 1972; Radloff, 1977). The CMPS represents a first draft of the final tool, so construction of a usage manual would be premature at this stage of development, although the examination procedure and item definitions available currently could provide the basis of such manual. The production of a training or usage manual for

the CMPS is an aspect of the scale development that could be explored in the future, and could reasonably be expected to improve its generalizability.

The CMPS is the first pain measurement tool constructed for use in animals that has made use of a formal scaling model. The scaling model used was Thurstone's paired comparisons model (Chapter 4). Recall that during the construction of the scale discussed in Chapter 3 a number of veterinary surgeons were asked to use a VAS to indicate the pain intensity they believed to be associated with each expression. These assessments can be thought of as providing an approximate weighting for the scale items, although the items included in the final scale were an adaptation of these original items. To gain an impression of the appropriateness of the Thurstone's weights they were compared to the VAS pain intensity scores. The weights reflected the ordering of the VAS scores associated with the items for all categories except mobility and demeanour. This suggested that, with the exception of those two categories, the weights accurately reflected the relative positioning of the items. The discordance between the VAS scores and Thurstone's weights seen in mobility was not as pronounced as in the demeanour category. The paired comparison probability estimates for demeanour indicated that since the distribution of some items lay far apart, the weights may have been unstable. This difficulty in weighting the demeanour items indicated that the veterinary surgeons consulted had problems in interpreting and scoring the behaviours. Thus, the relationship between demeanour and pain may be more complex than the relationship for the other categories.

Closer investigation of the assessments of the video recordings revealed the demeanour category had by far the highest level of disagreement between the observers and over time. There was disagreement between the observers for 11 of the 12 animals, and when comparing each observer's assessments over time, there was disagreement for between 5 and 8 of the 12 animals. Not only did the observers disagree with each other but their own assessments also changed over time. Hence, demeanour was the most problematic category to assess as well as being the most difficult to interpret when calculating weights. Other scales published in the veterinary literature acknowledge that the relationship between pain and demeanour is complex. An example is the UMPS where the change in demeanour from before to after the assessment is used (Firth and Haldane, 1999). The welfare control guidelines drawn up by Morton and Griffiths (1985) also indicated that to highlight changes in an animal's demeanour the assessment should be made by a person

who is familiar with the animal and who was able to make a comparison with the animal's normal behaviour.

The subjective nature of the VAS, NRS and SDS means that the interpretation of the relationship between demeanour and pain intensity is dependent on the individual observer, and it would be impossible to control this relationship without imposing restrictions on the use of the scale. However, exploration of the relationship between demeanour and pain could be used to greatly improve the performance of the CMPS. For example, the performance of the CMPS in post-operative pain could be improved by assessing demeanour before and after surgery. Alternatively, it may be possible for the animal's owner to be involved in the assessment of demeanour. These issues are dependent on the purpose of the tool and how it is to be used, for example, in a clinical setting it may not always be possible to assess the animal before the incident that causes them pain, as this may be trauma rather than surgery.

One further point to note is that the CMPS takes a global approach to pain measurement in dogs. It does not account for any differences between breeds of dogs. This may be a shortcoming of the scale and may contribute to the difficulties the observers had in assessing demeanour. When using a subjective rating scale the observers can adjust their scores to account for any differences which they believe exist between the breeds. This is not possible when using the CMPS since the contribution of each behaviour is predetermined and cannot be adapted for the breed of the animal being examined. Thus, in the search for a less subjective method, the scale has lost the potential to account for inter-breed differences, whether they are real or perceived. Another possible extension of the work undertaken here would be to investigate the effect of breed on the pain behaviours exhibited and the impact that this may have on the use of the CMPS.

The work undertaken in this thesis has answered some questions surrounding the measurement of pain in dogs, and it has identified areas for future research. The simple subjective rating scales have been shown to provide inadequate generalizability between observers when used in the clinical setting. The solution to this was to construct and implement a composite measurement pain scale using methodology that is acknowledged to provide improved measurement properties (i.e. in validity and reliability) over subjective assessment.

It is only very recently that composite measurement pain scales have been developed for use in veterinary medicine. These scales have all been constructed in an ad-hoc manner

and are based primarily on the personal opinions of their developers. The CMPS developed in this thesis makes use of techniques from statistics and psychometrics to develop a scale that encompasses a body of opinion gathered from the wider veterinary community.

This chapter has highlighted a number of issues that could be explored to improve the performance of the CMPS. Specifically, the areas that demand further investigation are the observers' use of the item definitions, the provision of training for the observers and the relationship between pain and demeanour. Whereas these three areas are unlikely to benefit the subjective rating scales, they provide room to enhance the performance of the CMPS. Hence, these areas offer potential for enhancement of the CMPS, which could then out-perform the subjective rating scales. Investigation of these avenues would provide invaluable information for the ongoing development and appraisal of the Composite Measurement Pain Scale and ultimately for the assessment and management of pain in dogs.

List of references

AMERICAN PSYCHOLOGICAL ASSOCIATION. (1985) Standards for educational and psychological testing. Washington: American Psychological Association Press

ATKINSON, J.H., KREMER, E.F., & IGNELZI, R.J. (1982) Diffusion of pain language with affective disturbance confounds differential diagnosis. *Pain* **12**,375-384.

BANOS, J.E., BOSCH, F., CANELAS, M., BASSOLS, A., ORTEGA, F. & BIGORRA, J. (1989) Acceptability of visual analogues scales in the clinical setting: a comparisons with the verbal rating scale in postoperative pain. *Methods and Finding in Experimental and Clinical Pharmacology*, **11**, 123-127.

BATESON, P.(1991) Assessment of Pain in Animals. *Animal Behaviour* **42,**827-839.

BENSON, G.J., WHEATON, L.G., THURMON, J.C., & TRANQUILLI, W.J. (1989) Postsurgical pain in cats: comparative efficacy of selected analgesics as reflected by plasma catecholamine concentration. *Proceedings of the American College of Veterinary Anaesthetists Annual Meeting,* p247.

BEYER, J.E. (1984) *The Oucher: A User's Manual and Technical report.* Evanston, IL: Judson Press.

BEYER, J.E. & ARADINE, C. (1988) The convergent and discriminant validity of a self-report measure of pain intensity for children. *Children's Health Care* **16,** 274-282.

BEYER, J.E. & KNAPP, T. (1986) Methodological issues in the measurement of children's pain. *Children's Health Care* **14**, 233-241.

BEYER, J.E., MCGRATH, P.J., & BERDE, C.B. (1990) Discordance between self-reported and behavioural pain measures in children aged 3-7years after surgery. *Journal of Pain and Symptom Management* **5,** 350-356.

BEYER, J.E. & WELLS, N. (1989) The Assessment of Pain in Children. *Pediatric Clinics of North America* **36**, 837-854.

BLAND, M.J. & ALTMAN, D.G. (1986) Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet* **i** 307-310.

BOECKSTYNS, M.E.H. & BACKER, M. (1989) Reliability and validity of the evaluation of pain in patients with total knee replacement. *Pain* **38**, 29-33.

BONICA, J.J. (1992) Pain Research and Therapy: History, Current Status, and Future Goals. In *Animal Pain*, ed. VAN POZNACK, A. AND SHORT, C.E., p. 1-29 New York: Churchill Livingston.

BRADSHAW, J.W.S., GOODWIN, D., LEA, A.M., & WHITEHEAD, S.L. (1996) A survey of the behavioural characteristics of purebred dogs in the United Kingdom. *The Veterinary Record* **138**, 465-468.

BRAUN, H.I. & TUKEY, J.W. (1983) Multiple comparions through orderly partitions: The maximum sub-range procedure. In *Principals of Modern Psychological Measurement: A Festschrift for Frederic M. Lord*, ed. WAINER, H. AND MESSICK, S.HILLSDALE, NJ., p. 55-64. Laurence Erlbaum Associates.

CARLSSON, A.M. (1983) Assessment of Chronic Pain, I. Aspects of the Reliability and Validity of the Visual Analogue Scale. *Pain* **16**, 87-101.

CHAMPION, G.D., GOODENOUGH, B., VON BAEYER, C.L., & THOMAN, W. (1998) Measurement of Pain by Self-Report. In *Measurement of Pain in Infants and Children*, ed. MCGRATH, P.J., pp. 123-160. Seattle: International Association for the Study of Pain Press.

CHAPMAN, C.B. (1989) Pain Assessment and Pain Control. *Proceedings of the Eleventh Bain-Falon Memorial Lectures,* p2-11.

CHAPMAN, R.C. (1976) Measurement of Pain: Problems and Issues. *Advances in Pain Research and Therapy* **1**, 345-353.

CHAPMAN, R.C., CASEY, K.L., DUBNER, R., FOLEY, K.M., GRACELY, R.H., & READING, A.E. (1985) Pain measurement: An Overview. *Pain* **22**, 1-31.

CHATFIELD, C. & COLLINS, A.J. (1980) Cluster Analysis. In: *Introduction to Multivariate Analysis*, pp. 82-87. New York: Chapman and Hall.

COHEN, J. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* **20**, 37-46.

COHEN, J. (1968) Weighted Kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin,* **70**, 213-220.

COHEN, R.J., SWEDLIK, M.E., & PHILLIPS, S.M. (1996) Psychological testing and assessment: An introduction to tests and measurement, 3rd edition. California: Mayfield Publishing Company.

COLLINS, S.L., MOORE, R.A., & MCQUAY, H.J. (1997) The visual analogue pain intensity scale: what is moderate pain in millimeters. *Pain* **72**, 95-97.

CONZEMIUS, M.G., HILL, C.M., SAMMARCO, J.L., & PERKOWSKI, S.Z. (1997) Correlation between subjective and objective measures used to determine severity of postoperative pain in dogs. *Journal of American Veterinary Medical Association* **210,**1619-1622.

COSTE, J., FREMANIAN, J., & VENOT, A. (1995) Methodological and statistical problems in the construction of composite measurement scales: A survey of six medical and epidemiological journals. *Statistics in Medicine* **14**, 331-345.

CRAIG, K.D., WHITFILED, M.F., GRUNAU, R.V.E., LINTON, J., HAKJISTAVROPOULOS, H.D. (1993) Pain in the preterm neonate: behavioural and physiological indices. *Pain* **52**: 287-299.

CRANE, S.W. (1987) Perioperative analgesic: A surgeon's perspective; *Journal of the American Veterinary Medical Association.* **191,** 1254-1257.

CROCKETT, D.J., PRKACHIN, K.M., & CRAIG, K.D. (1977) Factors of the language of pain in patient and normal volunteer groups. *Pain* **4**, 175-182.

CRONBACH, L.J. (1951) Coefficient Alpha and the internal structure of tests. *Psychometrica,* **16**, 297-334.

CRONBACH, L.J. (1970) *Essentials of psychological testing*; 3rd edition. New York: Harper and Row.

CRONBACH, L. J., GLESER, G.C., NANDA, H., & RAJARATNAM, N. (1972) The dependability of behavioural measurement: Theory of generalizability scores and profiles. New York: Wiley.

DALLENBACH, K.M. (1939) Somesthesis: Introduction to Psychology. EDS BORING, E.G., LANGFELD, H.S., & WELD, H.P. pp 608-625 New York: Wiley and Son.

DE CONNO, F., CARACENI, A., GAMBA, A., MARIANI, L., ABBATTISTA, A., BRUNELLI, C., LA MURA, A., & VENTAFRIDDA, V. (1994) Pain measurement in cancer patients: a comparison of six methods. *Pain* **57**, 161-166.

DEBENEDITTIS, G., MASSEI, R., NOBILI, R., & PIERI, A. (1988) The Italian Pain Questionnaire. *Pain* **33**, 53-62.

DIXON, J.S. & BIRD, H.A. (1981) Reproducibility along a 10cm vertical visual analogue scale. *Annals of Rheumatic Diseases* **40**, 87-89.

DOCTOR, J.N., SLATER, M.A., & ATKINSON, J.H. (1995) The descriptor differential scale of pain intensity: an evaluation of item and scale properties. *Pain* **61**, 251-260.

DODMAN, N., CLARKE, G.H., COURT, M.H., FIKES, L.L. & BOUDRIEAU, R.K. (1992) Epidural opioid administration for post operative pain relief in the dog. In Animal Pain, pp274-277 eds. SHORT, C.E., VAN POZNA, A. New York: Churchill Livingstone.

DOWNIE, W.W., LEATHAM, P.A., RHIND, V.M., WRIGHT, V., BRANCO, J.A., & ANDERSON, J.A. (1978) Studies with pain rating scales. *Annals of Rheumatic Disease,* **37**, 378-381.

DUBISSON, D. & MELZACK, R. (1976) Classification of clinical pain descriptions by multiple group discriminant analysis. *Experimental Neurology* **51**, 480-487.

DURNAD, M., SANGHA, B., CABAL, L.A., HOPPENBROUWERS, T., & HODGMEN, J.E. (1989) Cardiopulmonary and intracranial pressure changes related to endotracheal suctioning in preterm infants. *Critical Care* **17**, 506-510.

DWORKIN S.F. & WHTNEY, C.W. (1992) Relying on objective and subjective measures of chronic pain: guidelines for use and interpretation. In *Handbook of pain assessment*, pp. 429-446 eds. TURK D. C., MELZACK, R. New York: Guilford Press.

ERICKSON, C.J. (1990) Pain measurement in children: problems and directions. *Development and Behaviour in Paediatrics* **11**, 135-137.

ERICKSSON, H.H. & KITCHELL, R.L. (1984) Pain perception and alleviation in animals. *Proceedings of the 66th Annual Meeting of the Federation of Americal Societies for Experimental Biology,* 307-1312.

EVANS, W.J., CAYTEN, C.G., & GREEN, P.A. (1981) Determining the Generalizability of Rating Scales in Clinical Setting. *Medical Care* **19**, 1211-1220.

FINLEY, G.A. & MCGRATH, P.J. (1998) Introduction: The Roles of Measurement in Pain Management and Research. In: *Measurement of Pain in Infants and Children*, p. 1-4, eds FINLEY, G.A. AND MCGRATH, P.J. Seattle: International Association for the Study of Pain Press.

FIRTH, A.M. & HALDANE, S.L. (1999) Development of a scale to evaluate post-operative pain in dogs. *Journal of the American Veterinary Medical Association* **214**, 651-659.

FORDYCE, W.E. (1983) The Validity of Pain Behaviour Measurement. In: *Pain Measurement and Assessment*, pp.145-153 ed MELZACK, R. New York: Raven Press.

FORS, U., AHLQUIST, M.L., SKAGERWALL, R., EDWALL, L.G.A., & HAEGERSTAM, G.A.T. (1984) Relation between intradental nerve activity and estimated pain in man - a mathematical model. *Pain* **18**, 397-408.

FOX, S.M., MELLOR, D.J., FIRTH, E.C., HODGE, H., & LAWOKO, C.R.O. (1994) Changes in plasma cortisol concentrations before, during and after analgesia and anaesthesia plus ovariohysterectomy in bitches. *Research in Veterinary Science* **57**, 110-118.

FRANK, A.J.M., MOL, J.M.H, & HART, J.F. (1982) A comparison of three ways of measuring pain. *Rheumatology Rehabilitation* **21**, 211-217.

GAITO, J. (1980) Measurement Scales and Statistics: Resurgence of an old misconception. *Psychological Bulletin* **87** (3), 564-567.

GLASS, G.V. & STANLEY, J.C. (1970) *Statistical Methods in Education and Psychology*, New Jersey: Prentice-Hall.

GRACELY, R.H. (1980) Pain measurement in man. In: *Pain, Discomfort and Humanitarian Care*, pp. 111-138 eds NG, L.K.Y. & BONICA, J.J., Amsterdam: Elsevir.

GRACELY, R.H. (1983) Pain language and ideal pain assessment. In: *Pain measurement and assessment*, pP. 71-77, eds by MELZACK, R., New York: Raven Press.

GRAHAM, C., BOND, S.S., GERKOUSCH, M.M., & COOK, M.R. (1980) Use of the McGill Pain Questionnaire in the assessment of cancer pain: reliability and consistency. *Pain* **8**, 377-387.

GRUNAU, R.V.E., & CRAIG, K.D. (1987) Pain expression in neonates: facial and cry responses to invasive and non-invasive procedures. *Pain* **28**, 395-410.

GROSSI, E., BORGHI, C., CERCHIARI, E.L., DELLA PUPPA, T., & FRANCUCCI, B. (1983) Analgoue chromatic continuous scale (ACCS): a new method for pain assessment. *Clinical and Experimental Rheumatology* **1**, 337-340.

GUILFORD, J.P. (1986) Psychometric Methods. New York: McGraw-Hill.

GUION, R.M. (1977) Content validity: Three years of talk-what's the action? *Public Personnel Management* **6**, 407-414.

GUTTMAN, L. (1944) Louis Guttman on Theory and Methodology. Brookfield: Dartmouth

GUYATT, G.H., KIRSHNER, B., & JAESCHKE, R. (1992) Measuring Health Status: What are the necessary measurement properties? *Clinical Epidemiology* **45** (12), 1341-1345.

HALLIN, R.D., & TORBJöRK, H.E. (1974) Methods to differentiate electrically-induced afferent and sympathetic C unit responses in human coetaneous nerves. *Acta Physiology Scandanavia* **92**, 318-331.

HAMLIN, R.L., BEDNARSKI, L.S., SCHULER, C.J., WELDY, P.L., & COHEN, R.B. (1988) Method of objective assessment of analgesia in the dog. *Journal of Veterinary Pharmacology and Therapeutics* **11**, 215-220.

HANSEN, B.D. & HARDIE, E.M. (1993) Prescription and use of analgesics in dogs and cats in a veterinary teaching hospital: 258 cases (1983-1989). *Journal of the American Veterinary Medical Association* **202**, 1484-1494.

HARPIN, V.A. & RUTTER, N. (1990) Development of emotional sweating in the newborn infant. *Archives of Disease in Childhood* **57**, 691-695.

HASKINS, S.C. (1987) Use of analgesics post operatively and in a small animals intensive care setting. *Journal of American Veterinary Medical Association,* **191**, 1266-1268.

HEAVNER, J.E. (1992) Pain recognition during experimentation and tailoring anesthetic and analgesic administration to the experiment. In: *Animal Pain*, p.509-513, ed WALL, D.E., Seattle; International Association for the Study of Pain Press.

HELLYER, P.W. AND GAYNOR, J.S. (1998) Acute Post-surgical Pain in Dogs and Cats. *The Compendium of Continuing Education (Small Animal)* **20** (2), 140-153.

HESTER, N.O. (1993) Assessment of pain in children with cancer. In: *Current and emerging issues in cancer pain: Research and practice*, p. 219-245 eds CHAPMAN, R.C. & FOLEY, K.M. New York: Raven Press.

HOLLANDER, M. & WOLFE, D.A. (1973) *Nonparametric Statistical Methods*, New York: John Wiley & Sons.

HOLTON, L.L., SCOTT, E.M., NOLAN, A.M., REID, J., & WELSH, E.M. (1998a) Investigation of the relationship between physiological factors and clinical pain in dogs scored using a numerical rating scale. *Journal of Small Animal Practice* **39**, 469-474.

HOLTON, L.L., SCOTT, E.M., NOLAN, A.M., REID, J., WELSH, E.M., & FLAHERTY, D. (1998b) Comparison of three methods used for assessment of pain in dogs. *Journal of the American Veterinary Medical Association* **212** (1), 61-66.

HOWARD, C.R., HOWARD, F.M., & WEITZMAN, M.L. (1994) Acteminopahen analgesia in neonatal circumcision: the effect on pain. *Pediatrics* **93**, 641-646.

HUNTER, M., PHILIPS, C. & RACHMAN, S. (1979) Memory of Pain. *Pain,* **6**, 35-46.

HUSKISSON, E.C. (1974) Measurement of Pain. *The Lancet*, **2**, 1127-1131.

HUSKISSON, E.C.(1983) Visual Analogue Scales. In: *Pain measurement and assessment*, p. 33-37, ed MELZACK, R. New York: Raven Press.

IGGO, A. (1984) *Pain in Animals* Potters Bar: Universities Federation for Animal Welfare

JACOBY, J. & MATELL, M.S. (1971) Three-point Likert scales are good enough. *Journal of Marketing Research* **8**, 495-500.

JAY, S.M., OXOLINS, M., ELLIOT, C.H. & CALDWELL, S. (1983) Assessment of children's distress during painful medical procedures. *Health Psychology* **2**, 133-147.

JENKINSON, C. (1991) Why are we weighting? A critical examination of the use of item weights in a health status measure. *Social Science and Medicine* **32**, 1413-1416.

JOHNSON, J.M. (1991) The Veterinarian's Responsibility: Assessing and Managing Acute Pain in Dogs and Cats. Part 1. *The Compendium of Continuing Veterinary Education* **13**, 804-807.

JOHNSON, J.M. (1991) The Veterinarian's responsibility: Assessing and managing acute pain in dogs and cats. Part 1. *The compendium of continuing education* **13**, 320-323.

JOHNSTON, C.C. (1998) Psychometric Issues in the Measurement of Pain. In: *Measurement of Pain in Infants and Children*, p. 5-20. eds FINLEY, G.A. & MCGRATH P.J. Seattle: IASP Press.

JOHNSTON, C.C. & STRADA, M.E. (1986) Acute pain response in infants: a multidimensional description. *Pain* **24**, 373-382.

JOYCE, R.B., ZUTSHI, D.W., HRUBES, V., & MASON, R.M. (1975) Comparison of fiexed interval and visual analogue scales for rating chronic pain. *European Journal of Clinical Pharmacology* **8**, 415-420.

KAIKO, R.F. WALLENSTEIN, S.L., ROGERS, A.G. & HOUDE, R.W. (1983) Sources of variation in analgesic responses in cancer patients with chronic pain receiving morphine, *Pain* **15**, 191-200.

KATZ. E.R., KELLERMAN, J. & SEIGEL, S.E. (1980) Behavioural distress in children with cancer undergoing medical procedure: developmental considerations *Journal of Consultant Clinical Psychology* **48** 356-365.

KEEFE, F.J. & BLOCK, A.R. (1982) Development of an observational method for assessing pain behaviour in chronic low back pain patients. *Behaviour Therapy* **13**, 363-375.

KETOVUORI, H. & PONTINEN, P.J. (1981) A Pain vocabulary in Finnish - The Finnish pain Questionnaire. *Pain* **11**, 247-253.

KIM, H.S., SCHWARTZ-BARCOTT, D., HOLTER, I.M., & LORENSEN, M. (1995) Developing a translation of the McGill Pain Questionnaire for cross-cultural comparison: An example from Norway. *Advanced Nursing* **21**, 421-426.

KITCHELL, R.L. (1987) Problems in defining pain and peipheral mechanisms of pain. *Journal of American Veterinary Medical Association.* **191**, 1195-1199.

KLINE, P. (1993) The handbook of psychological testing, London: Routledge.

KRANE, E.J., JACOBSON, L.E., LYNN, A.M., PARROT, C., & TYLER, D.C. (1987) Caudal morphine for postoperative analgesia in children: a comparison with caudal Bupivacaine and intravenous morphine. *Anesthesia and Analgesia* **66**, 647-653.

KRZANOWSKI, W.J., (1988) Principles of multivariate analysis: a user's perspective, Oxford: Clarendon Press.

KRAVITZ, E., MOORE, M.E., & GLAROS, A. (1981) Paralumbare muscle activity in chronic low back pain. *Archives of physiology and medical rehabilitation* **62**, 172-176.

LANDY, F.J. (1986) Stamp collecting versus science. *American Psychologist* **41**, 1183-1192.

LANGLEY, G.B. & SHEPPARD, H. (1985) The visual analogue scale: Its use in pain measurement. *Rheumatology International* **5**, 145-148.

LASANGA, L. (1964) The Clinical Measurement of Pain. *Annals of the New York Academy of Science* **86**, 28-37.

LASCELLES, B.D.X., BUTTERWORTH, S.J., & WATERMAN, A.E. (1994) Postoperative analgesic and sedative effects of carprofen and pethidine in dogs. *The Veterinary Record* **134**, 187-191.

LAWRIE, S.C., FORBES, D.W., AKHTAR, T.M., & NORTON, N.S. (1990) patient-controlled analgesia in children. *Anaestheisa* **46**, 1074-1076.

LEAVITT, F., GARRON, D.C., WHISLER, W.W., & SHEINKOP, M.B. (1978) Affective and sensory dimensions of pain. *Pain* **4**, 273-281.

LEHMANN, D.R. & HULBERT, J. (1972) Are three-point scales always good enough. *Journal of Marketing Research* **9**, 444-446.

LE RESCHE, L. (1982) Facial expression of pain: a study of candid photographs. *Journal of Non-verbal Behaviour* **7**, 46-56.

LEWIS, T. (1942) Pain, New York: MacMillan Press.

LIGHT, G.,S., HARDIE, E.,M., YOUNG, M.,S., HELLYER, P.,W., BROWNIE, C., & HANSEN, B.,D. (1993) Pain and anxiety behaviors of dogs during intravenous catheterization after premedication with placebo, acepromaxine or oxymorphone. *Applied Animal Behaviour Science* **37**, 331-343.

LINTON, S.J. & GOTESTAM, G. (1983) A clinical comparison of two pain scales: correlation, remembering chronic pain and a measure of compliance. *Pain* **1**, 57-65.

LIVINGSTON, A., WATERMAN, A.E., LASCELLES, B.D.X., JONES, A., & HENDERSON, G. (1994) Correlation of blood levels of carprofen given either pre or post-operatively with mechanical thresholds and humoral factors as indicators of post-operative pain in the dog. Proceedings of 6th EAVPT Congress:191-192.

LOEWENTHAL, K. M. (1996) An introduction to psychological tests and scales; ed K LOEWENTHAL, London: UCL Press.

MAMGIONE, C.M., MARCANTONIO, E.R., GOLDMAN, L., COOK, E.F., DONALDSON, M.C., SUGARBAKER, D.J., POSS, R., & LEE, T.H. (1993) Influence of Age on Measurement of health status in patients undergoing elective surgery. *Journal of the American Geriatric Society* **41**, 377-383.

MASON, J.W. (1968) Overall hormonal balance as a key to endocrine organisation. *Psychosomatic Medicine* **30**, 791-808.

MAXWELL, L.G., YASTER, M., WETZEL, R.C., & NEIBYL, J.R. (1987) Penile nerve block for newborn circumcision. *Obstetrics and Gynaecology* **7,** 415-419.

MCCULLAGH, P. & NELDER, J.A. (1989) *Generalized Linear Models*, London: Chapman and Hall.

MCGRATH, P.A. (1987) An assessment of children's pain: a review of behavioural, physiological and direct scaling techniques. *Pain* **3,** 147-176.

MCGRATH, P.A. & BRIGHAM, M.C. (1992) The Assessment of Pain in Children and Adolescents. In: *Handbook of Pain Assessment*, p. 295-314, eds TURK, D.C. & MELZACK, R. New York: The Guilford Press.

MCGRATH, P.A. & HILLIER, L.M. (1989) The Enigma of Pain in Children: An overview. *Pediatrician* **16,** 6-15.

MCGRATH, P.J. (1990) Paediatric pain: a good start. *Pain* **4,** 253-254.

MCGRATH, P.J. (1996) There is more to pain measurement in children than "ouch". *Canadian Psychology* **37**, 63-75.

MCGRATH, P.J., JOHNSON, G., GOODMAN, J.T., SCHILLINGER, J., DUNN, J., & CHAPMAN, J.A. (1985) CHEOPS: a behavioural scale for rating post-operative pain in children. In: *Advances in Pain Research and Therapy,* p 395-402, eds FIELDS, H.L., DUBNER, R., & CERVERO, F. New York: Raven Press.

MCGRAW, M. B., (1945) The Neuromuscular Matruation of the Human Infant, New York; Hafner.

MELZACK, R. (1975) The McGill Pain Questionnaire: Major properties and scoring methods. *Pain* **1**, 277-299.

MELZACK, R. (1983) The McGill Pain Questionnaire. In: *Pain measurement and assessment*, p. 41-47, ed MELZACK, R. New York: Raven Press.

MELZACK, R. & TORGERSON, W.S. (1971) On the language of Pain. *Aneasthesiology* **34**, 50-59.

MENEGAZZI, J.J., DAVIS, E.A., SUCOV, A.N., & PARIS, P.M. (1993) Reliability of the Glasgow Coma Scale when used by emergency physicians and paramedics. *The journal of trauma* **34**, 46-48.

MERSKEY, H. & BOGDUK (1994) Classification of chronic pain. Descriptions of chronic pain syndromes and definition of pain terms (2nd Edition) Seattle: IASP Press.

MESSICK, S. (1980) Test validity and the ethics of assessment. *American Psychologist* **35**, 1012-1027.

MORTON, D.B. AND GRIFFITHS, P.H.M. (1985) Guidelines on the recognition of pain, distress and discomfort in experimental animals and a hypothesis of assessment. *The Veterinary Record* **116**, 431-436.

NOLAN, A.M. (1994) Analgesics in veterinary medicine. *Proceedings of the 6th European Association of Veterinary Pharmacology and Therapeutics,* p143-145.

NOLAN, A.M. & REID, J. (1993) Comparison of the postoperative analgesic and sedative effects of carprofen and papaveretum in the dog. *The Veterinary Record* **133,** 240-242.

NUNNALLY, J.C. (1978) *Psychometric Theory*, New York: McCraw-Hill.

NUNNALLY, J.C., BERNSTEIN, I.H. (1994) Psychometric theory, 3rd edition, New York: McGraw-Hill.

NYREN, O., ADAMI, H.O., BATES, S., BERGSTROM, R., GUSTAVSSON, S., & LOOF, L. (1987) Self-rating of pain in nonulcer dyspepsia: A methodological study comparing a new fixed pain scale and the visual analogue scale. *Journal of Clinical Gastroenterology* **9**, 408-414.

OHNHAUS, E.E. & ADLER, R. (1975) Methodological problems in the measurement of pain: a comparison between the verbal rating scale and the visual analogue scale. *Pain* **1**, 379-384.

PAIN, R. MAX, M., & INTURRISI, C. (1986) Principles of analgecis drugs in the treament of acute pain and chronic cancer pain *Syllabus Committee Meeting, American Pain Society Board of Directors,* p. 1-9

POTTHOFF, A. & CARITHERS, R.W. (1989) Pain and analgesia in dogs and cats. *The Compendium of Continuing Veterinary Education* **11**, 887-896.

PRICE, D.D., BUSH, F.M., LONG, S., & HARKINS, S.W. (1994) A comparison of pain measurement characteristics of mechanical visual analogue and simple numerical rating scales. *Pain* **56,** 217-226.

PRICE, D.D. & HARKINS, S.W. (1992) Psychophysical approaches to pain measurement and assessment. In: *Handbook of Pain Assessment*, p. 111-134 eds TURK, D.C. & MELZACK, R., New York: Guilford Press.

PRIETO, E.J. & GEISINGER, K.F. (1983) Factor-analytic studies for the McGill Pain Questionnaire. In: *Pain measurement and assessment*, p 63-70, ed MELZACK, R. New York: Raven Press.

PRIETO, E.J., HOPSON, L., BRADLEY, L.A., GEISINGER, K.F., MIDAX, D., & MARCHISELLO, P.J. (1980) The language of low back pain: Factor structure of the McGill Pain Questionnaire. *Pain* **8**, 11-20.

RADLOFF, L.S. (1977) The CES-D scale: A self-reporting depression scale for research in the general population. *Applied Psychological Measurement* **1**, 385-401.

RAWLINGS, C.A., TACKETT, R.L., & BJORLING, D.E. (1989) Cardiovascular function and serum catecholamine concentration after anesthesia and surgery in the dog. *Veterinary Surgery* **18**, 225-260.

READING, A.E. (1979) The internal structure of the McGill Pain Questionnaire in dysmenorrhea patients. *Pain* **7**, 353-358.

READING, A.E. (1980) A comparison of pain rating scales. *Journal of psychosomatic research* **24**, 119-124.

READING, A.E. (1982) A comparison of the McGill pain questionnaire in chronic and acute pain. *Pain* **13**, 185-192.

READING, A.E. (1983) The McGill Pain Questionnaire: An appraisal. In: *Pain measurement and assessment*, p. 55-61, ed MELZACK, R. New York: Raven Press.

REID, J. & NOLAN, A.M. (1991) A comparison of the post-operative analgesic and sedative effects of flunixin and papaveretum in the dog. *Journal of Small Animal Practice* **32**, 603-608.

REVILL, S.I., ROBINSON, J.O., ROSEN, M., & HOGG, M.I.J. (1976) The reliability of a linear analogue for evaluating pain. *Anaesthesia* **31**, 1191-1198.

ROBERTSON, J. (1993) Pediatric pain assessment: Validation of a multidimensional tool. *Pediatric nursing* **19** (3), 209-213.

SACKMAN, J.E. (1991) Pain: It's perception and alleviation in dogs and cats. Part 1: the physilogy of pain. *The compendium of continuing education* **1** (1), 35-40.

SANFORD, J., EWBANK, R., MOLONY, V., TAVERNOR, W.D., & UVAROV, O. (1986) Guidelines for the recognition and assessment of pain in animals. *Veterinary Record* **118,** 334-338.

SAVAGE, C.W. (1970) In: *The Measurement of Sensation: A Critique of Perceptual Psychophysics* Berkley: University of California Press.

SCOTT, J. & HUSKISSON, E.C. (1979) Accuracy of subjective measurements made with or without previous scores: an important source of error in serial measurement of subjective states. *Annals of Rheumatic Diseases* **38**, 558-559.

SEYMOUR, R.A. (1982) The use of pain scales in assessing the efficacy of analgesics in post-operative dental pain. *European Journal of Clinical Pharmacology* **23**, 441-444.

SHAVELSON, R.J., WEBB, N.M., & ROWLEY, G.L. (1989) Generalizability Theory. *American Psychologist* **44**, 922-932.

SHROUT, P.E. & FLEISS, J.L. (1979) Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin* **86,** 420-428.

SKEVINGTON, S.M. (1983) Activities as indices of illness behaviour in chronic pain. *Pain* **3**, 295-307.

SMIRNOV, N.V. (1939) One the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bulletin of Moscow University* **2**, 3-16.

SNEDECOR, G.W. & COCHRAN, W.G. (1980) *Statistical Methods*, Ames: The Iowa State University Press.

SOKAL, R.R. & SNEATH, P.H.A. (1963) Principles of numerical taxonomy; San Francisco: W.H. Freeman.

SPEARMAN, C. (1904) The proof and measurement of association between two things. *American Journal of Psychology* **15**, 72-101.

SPECTOR, P.E. (1997) Choosing Response Categories for Summated Rating Scales. *Journal of Applied Psychology* **61**, 374-375.

SPEILBERGER, C.D., GORSUCH, R.L., & LUCHENE, R.E. (1972) *Manual for the State-Trait Anxiety Inventory, revised edition*, Palo Alto: Consulting Psychological Press.

SPINELLI, J.S. (1987) Clinical recognition and anticipation of situations likely to induce suffering in animals. *Journal of the American Veterinary Medical Association* **191** (10), 1216-1218.

SPINELLI, J.S., & MARKOWITZ, H. (1987) Clinical recognition and anticipation of situations likely to induce suffering in animals. *Journal of American Veterinary Medical Association.***191**, 1216-1218.

STARMARK, J., HOLMGREN, E., STALHAMMAR, D., LINDGRE, S., & OLANDERS, S. (1991) Reliability and Accuracy of the Glasgow Coma scale. *The Lancet* **337,** 1042-1043.

STERNBACH, R.A. (1983) The tourniquet pain test. In: *Pain measurement and assessment*, p. 27-31, ed MELZACK, R. New York: Raven Press.

STEVENS, B.J., JOHNSTON, C.C., & HORTON, L. (1993) Multidimensional pain assessment in prematrue neonates: a pilot study. *Journal of gynaecology and neonatal nursing* **22**, 531-541.

STRAND, L.I. & WISNES, A.R. (1991) The development of a Norwegian pain questionnaire. *Pain* **46**, 61-66.

STREINER, D.L. & NORMAN, G.R. (1995) *Health Measurement Scales: A Practical Guide to their Development and Use*, Oxford: Oxford Medical Publications.

SWEET, S.D. & MCGRATH, P.J. (1998) Physiological measure of pain. In: *Measures of pin in infants and children*, p59-81, eds by FINLEY, G.A. & MCGRATH, P.J. Seattle: IASP Press.

TAENZER, P. (1983) Postoperative pain: relationships among measures of pain, mood and narcotic requirements. In: *Pain measurement and assessment*, p. 111-118, ed MELZACK, R. New York: Raven Press.

TARBELL, S.E., COHEN, T., & MARSH, J.L. (1992) The toddler-preschooler postoperative pain scale: an observational scale for measuring postoperative pain in children aged 1-5. Preliminary report. *Pain* **50**, 273-280.

TAYLOR, P. (1985) Analgesia in the dog and cat *Journal of Small Animal Practice* **7** (1), 5-13.

TAYLOR, P.M & HERRTAGE, M.E. (1986) Evaluation of some drug combinations for sedation in the dog. *Journal of Small Animal Practice* **27**, 325-333.

TAYLOR, P.M. & HOULTON, J.F. (1984) Post-operative analgesia in the dog: a comparison of morphine, buprenorphine and pentazocine. *Journal of Small Animal Practice* **25,** 437-451.

TEASDALE, G. & JENNETT, B. (1974) Assessment of Coma and Impaired Consciounsess: A practical scale. *The Lancet.* **2** 81-84.

TESKE, K., DAUT, R.L., & CLEELAND, C.S. (1983) Relationships between nurses' observation and paients' self-report of pain. *Pain.* **16,** 289-296.

THURSTONE, L.L. (1928) Law of comparative judgment *American Journal Sociology* **33**, 529-554.

THURSTONE, L.L. & CHAVE, E.J. (1966) *The Measurement of Attitude*, Chicago: The University of Chicago Press.

TORGERSON, W.S. (1958) Theory and methods of scaling, New York: Wiley.

TORGERSON, W.S. & BENDEBBA, M. (1983) The sturcture of pain descriptors. In: *Pain measurement and assessment*, p. 49-54, ed MELZACK, R. New York: Raven press.

TOWNSEND, J.T. & ASHBY, F.G. (1984) Measurement Scales and Statistics: The Misconception Misconceived. *Psychological Bulletin* **96**, 394-401.

TYLER, D.C., DOUTHIT, A.T., & CHAPMAN, R.C. (1993) Toward validation of pain measurement tools for children: a pilot study. *Pain* **52**, 301-309.

WALL, P.D. (1992) Defining pain in animals. In Animal pain, p 63-79, ed SHORT, C.E. & VAN POZNAK, A. Edinburgh: Churchill Livingston.

WATERMAN, A.E. & KALTHUM, W. (1988) Pharmacokinetics of intramuscularly administered pethidine in dogs and the influence of anaesthesia and surgery. *The Veterinary Record* **124**, 293-296.

WECHSLER, D. (1981) *WAIS-R manual: Wechsler Adult Intelligence Scale-Revised.* New York: Psychological Corporation.

WELSH, E.M., GETTINBY, G., & NOLAN, A.M. (1993) Comparison of a visual analgoue scale and a numerical rating scale for assessment of lameness, using sheep as a model. *American Journal of Veterinary Research* **54**, 976-983.

WILCOXON, F. (1945) Individual comparisons by ranking methods. *Biometrics* **1**, 80-83.

WILLIS, W.D., CHUNG, J.M. (1987) Central mechanisms of pain, *Journal of the American Veterinary Medical Association,* **191**, 1200-1202.

WOLF, S.L., NACHT, M., & KELLY, J.L. (1982) EMG feedback training during dynamic movement for low back pain patients. *Behaviour Therapy* **13**, 395-406.

WONG, D.L. & BAKER, C.M. (1988) Pain in Children: Comparison of assessment scales. *Pediatric nursing* **14** (1), 9-17.

WRIGHT, J.G. & FEINSTEIN, A.R. (1992) A Comparative contrast of clinimetric and psychometric methods for constructing indexes and rating scales. *Clinical Epidemiology* **45** (11), 1201-1218.

YAKSH, T.L. & HAMMOND, D.L. (1982) Peripheral and central substrates involved in the rostrad transmission of nociceptive information, *Pain* **13**, 1-85.

YOXALL, A.T. (1978) Pain in small animals - its recognition and control. *Journal of Small Animal Practice* **19**, 423-438

Appendices

Appendix 1: Details of animals included in study to examine performance of three pain scales, VAS, NRS and SDS

Table 1: Listing of breeds of dogs and procedures undergone included in a study to compare the performance of the VAS, NRS and SDS when assessing pain. Animals were examined one hour after surgery.

Breed	Procedure	Age (yrs)	Sex
Cross-breed	Laminectomy	10	FN
Dobermann Pincher	Culposuspension	7.5	FN
Golden Retreiver	Ovariohysterectomy	6	F
English Springer Spaniel	Mastectomy	10	FN
Border Collie	Perineal hernia repair	9	M
Labrador/Retriever	Fracture repair	10	FN
Tibetan Spaniel	Gastroduodenostomy	5	M
West Highland White Terrier	Parotid duct transplant	4	FN
Labrador	Castration	12.5	M
Golden Retreiver	Biopsy, soft tissue	4	FN
Labrador Cross	Cruciate repair	8	MN
Bullmastiff	# Anconeus	0.75	M
Cavalier King Charles Spaniel	Castration	1	M
Cross-breed	Eye removal	11	FN
Dachshund	Fenestration	8	M
West Highland White Terrier	Cruciate repair	7	F
Cross Breed	Episiotomy	15	F
Cocker Spaniel	Entropion repair	11.5	MN
Greyhound	Castration	*	M
Chow Chow	Entropion repair	3	M
Greyhound	Fracture repair and dental	8	M
Border Collie	Fracture repair (radius/ulna)	4	M
Border Collie Cross-breed	Fracture repair (femur)	8	MN
Poodle	Castration	*	M
Golden Retriever	Arthrotomy, shoulder	0.67	M

Table 2: Listing of breeds of dogs and procedures undergone included in a study to compare the performance of the VAS, NRS and SDS when assessing pain. Animals were examined on the day following surgery.

Breed	Procedure	Age (yrs)	Sex
Labrador/Retriever	Insulinoma	11	FN
Cocker Spaniel	Ear cleaning	4	F
Dalmation	Eye removal	9.5	M
Cavalier King Charles Spaniel	*	1	M
Flat Coat Retriever	Cruciate repair	5.6	M
Golden Retriever	Cruciate repair	4.5	M
German Shepherd cross	Fracture repair	7.5	F
Ridgeback	Fracture repair	3	F
German Shepherd cross	Mammary strip	8	FN
West Highland White Terrier	Hemi-mandibulectomy	8	F
Border Collie	Femoral Osteotomy	0.33	F
German Shepherd Dog	Anal furunculosis therapy	6	M
German Shepherd Dog	Total ear canal ablation	8.5	M
Labrador	Exploratory laparotomy	10	M
Cavalier King Charles Spaniel	Corneal transplant	8	FN
Cross-breed	Fracture repair (Femur)	2	MN
Cross breed	Carpal Arthrodesis	1.33	MN
Labrador	Laminectomy	7	M
Golden Retriever	Lumpectomy	10	FN
Jack Russell	Para-aural absess removal	6.33	M
German Shepherd Dog	Anal furunculosis therapy	8	M
Cross-breed	Exploratory laparotomy	7.67	FN
Cavalier -King Charles Spaniel	Pedicle flap	1.16	FN
Border collie	Cruciate repair	5	M
German Shepherd Dog	Exploratory laparotomy	2	F

* Missing value

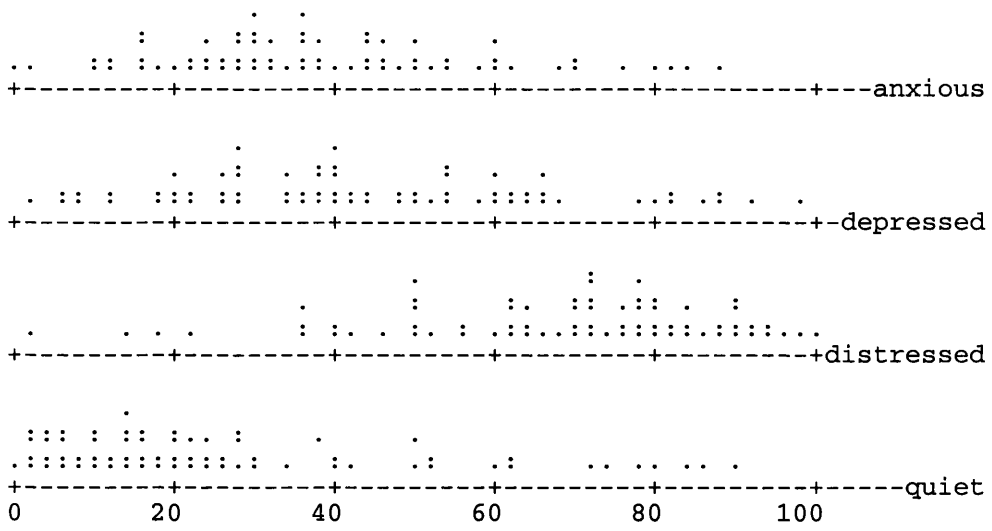
Table 3: Listing of breeds of dogs and procedures undergone included in a study to compare the performance of the VAS, NRS and SDS when assessing pain. Animals were examined one hour after surgery and on the day following surgery.

Breed	Procedure	Age (yrs)	Sex
Border Collie	Perineal hernia repair	9	M
Labrador/Retriever	Fracture repair	10	FN
Tibetan Spaniel	Gastroduedenostomy	5	M
Golden Retriever	Biopsy, soft tissue	4	FN
Labrador Cross	Cruciate repair	8	MN
Bullmastiff	# Anconeus	0.75	M
Daschund	Fenestration	8	M
Chow Chow	Enropion repair	3	M
Border Collie Cross	Fracture repair (femur)	8	MN
Greyhound	Fracture repair and dental	8	M
Golden Retriever	Arthrotomy, shoulder	0.67	M
Border Collie	Fracture repair (radius/ulna)	4	M
Greyhound	Castration	*	M
Cross-breed	Episiotomy	15	F
West Highland White Terrier	Cruciate Rupture	7	F
Labrador	Castration	12.5	M

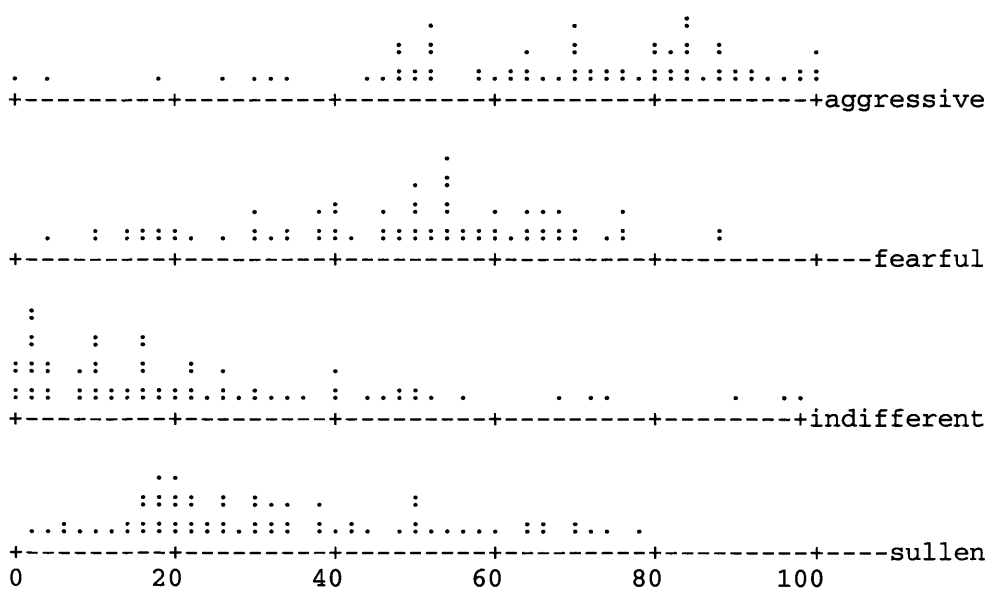
Appendix 2: Output of investigation of scale structure

Section 1: Dotplots of VAS pain intensity scores allocated to behaviours and signs of pain by 72 practicing veterinary surgeons.

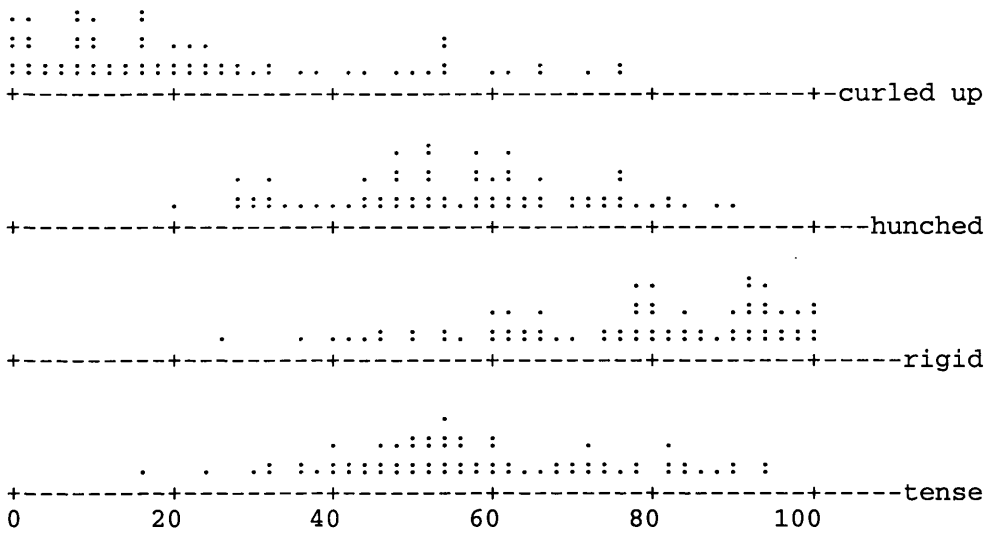
Demeanour



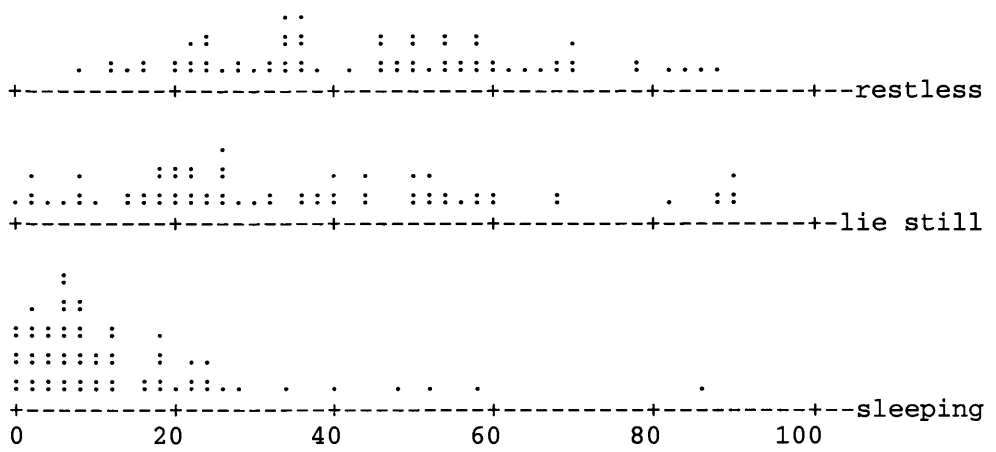
Response to people



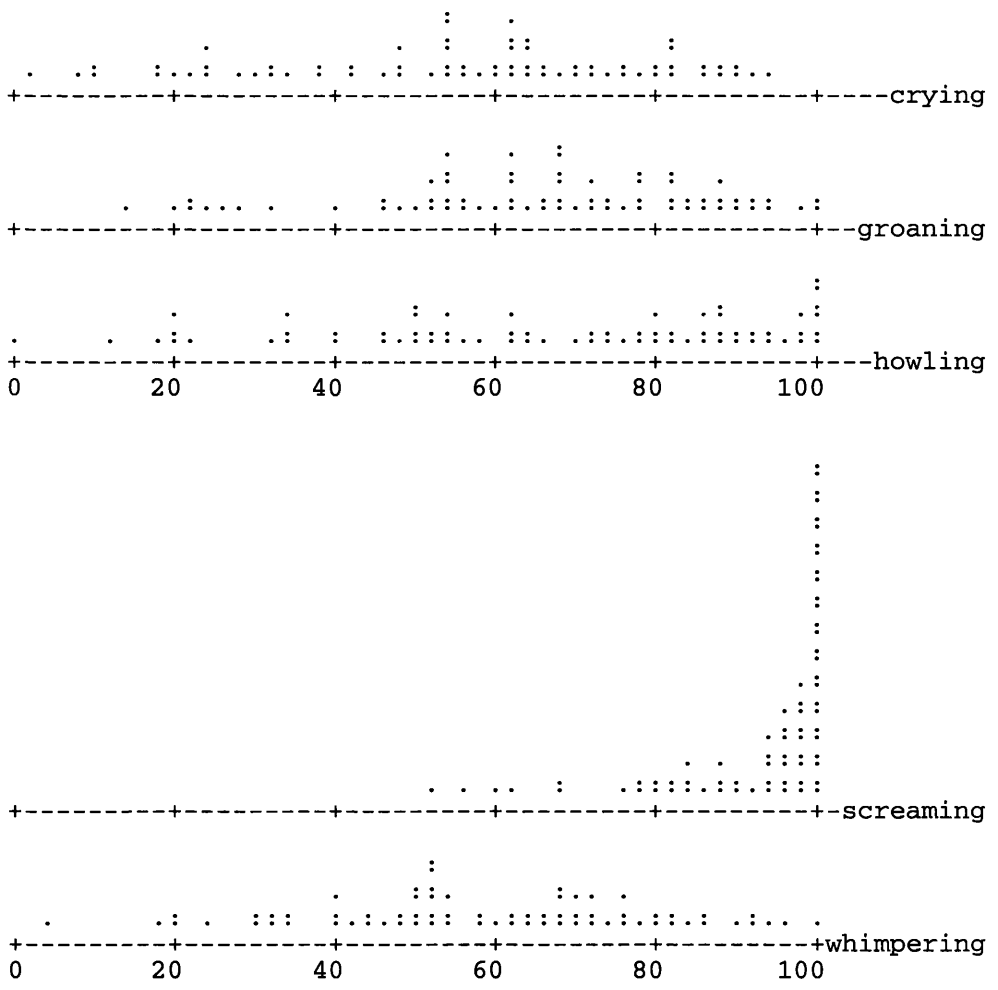
Posture



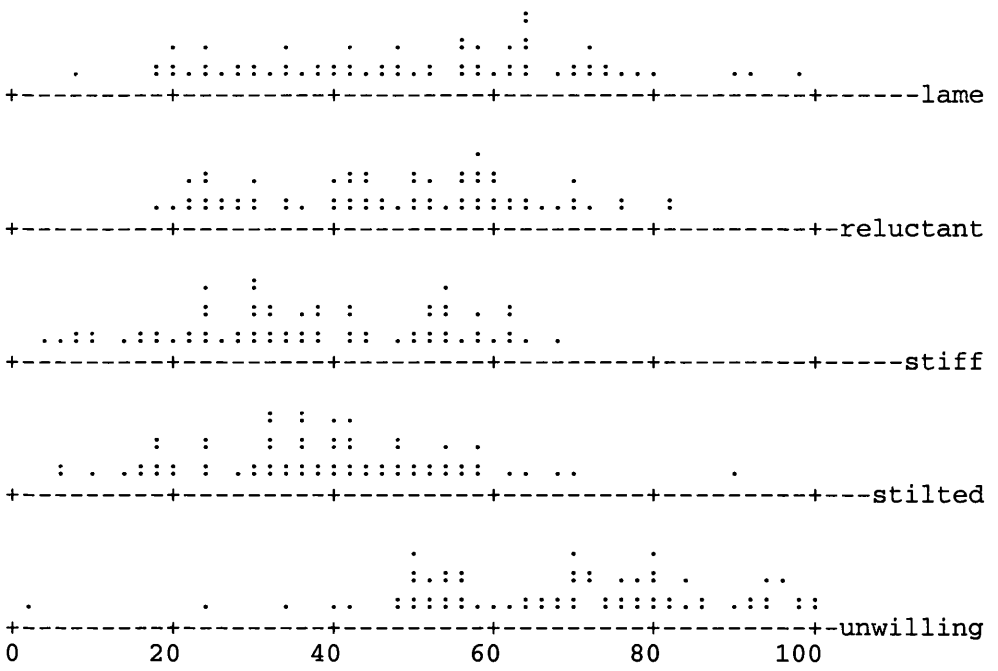
Activity



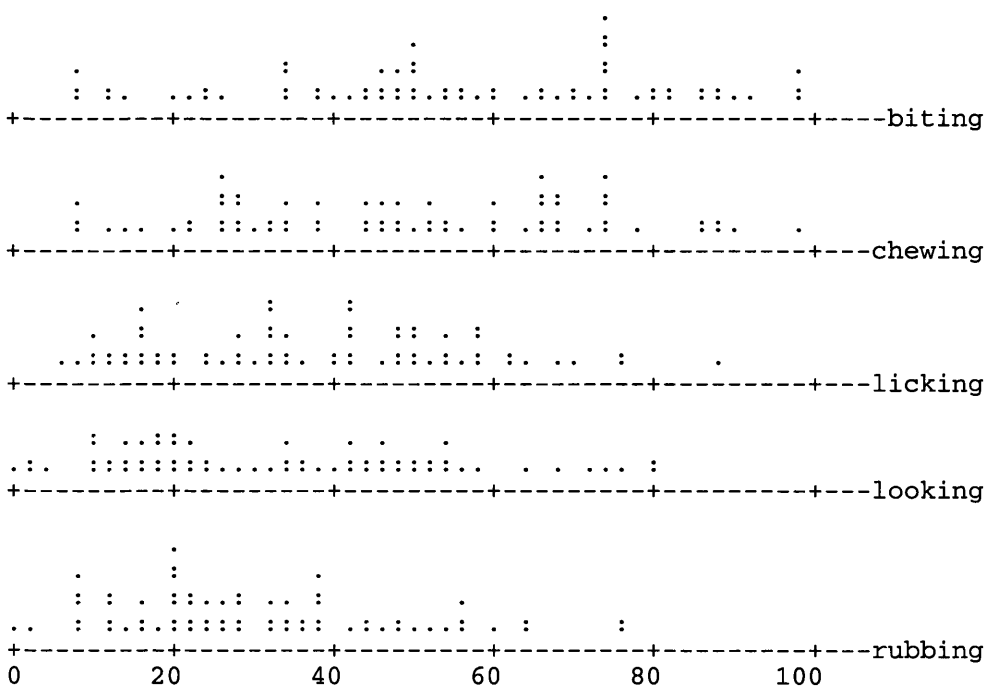
Vocalisation



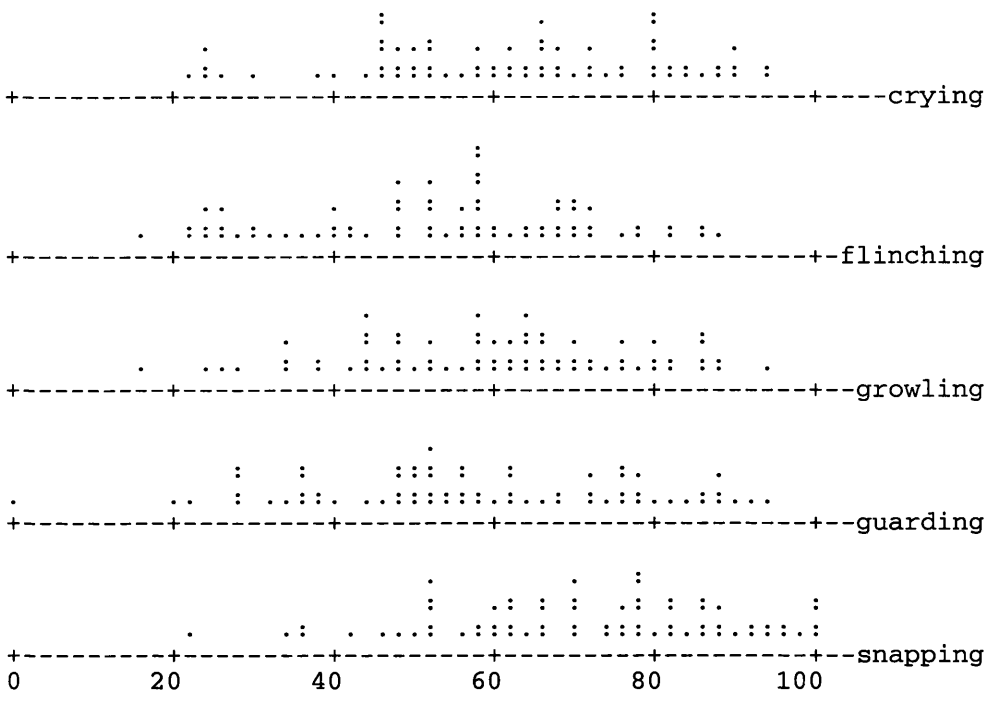
Mobility



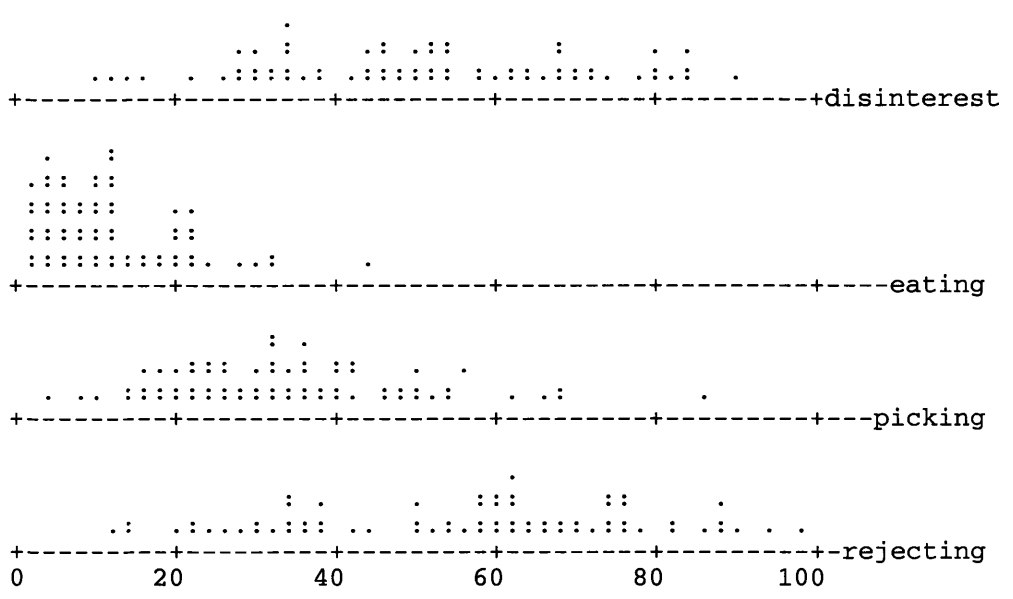
Attention to painful area



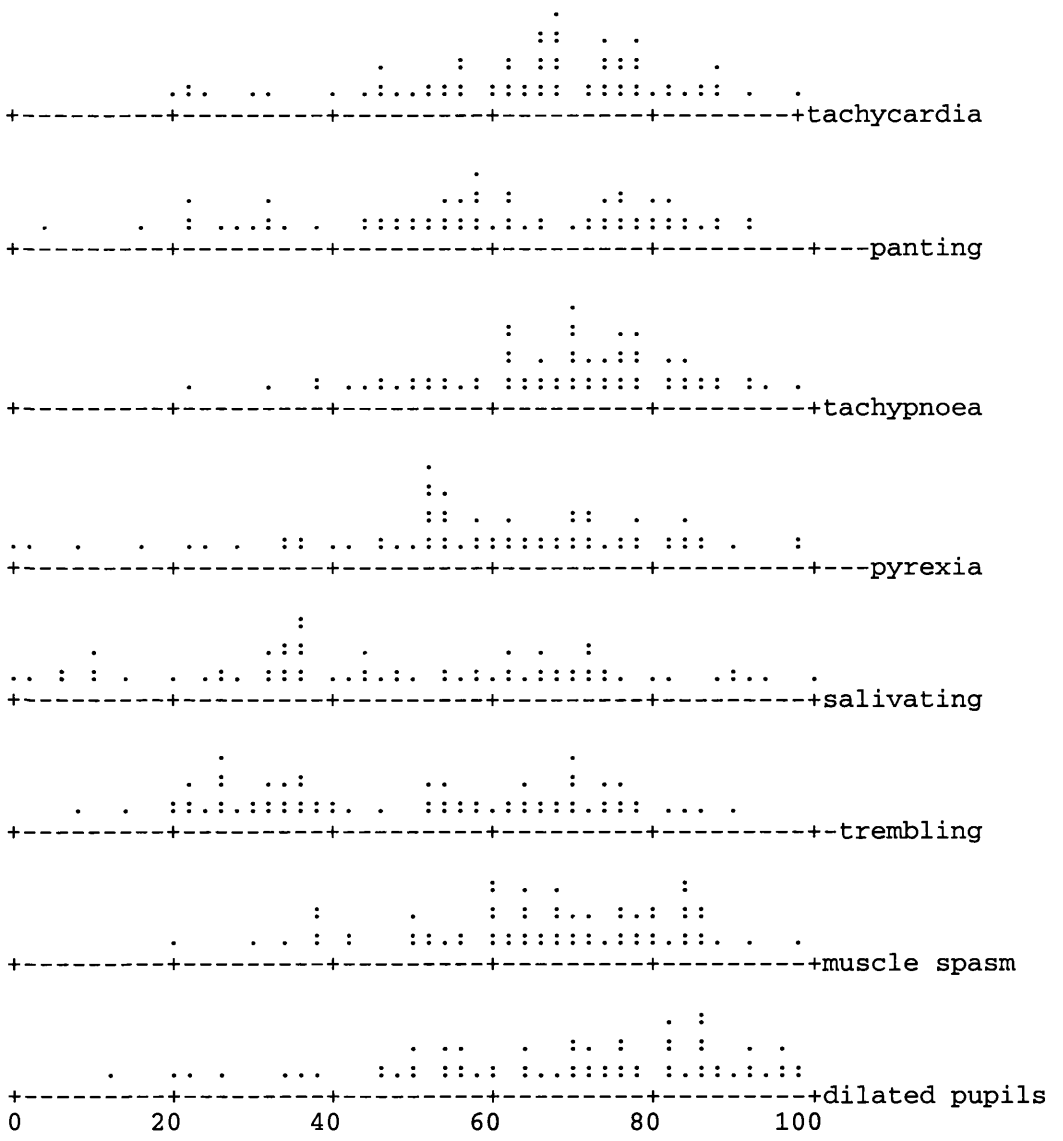
Response to touch



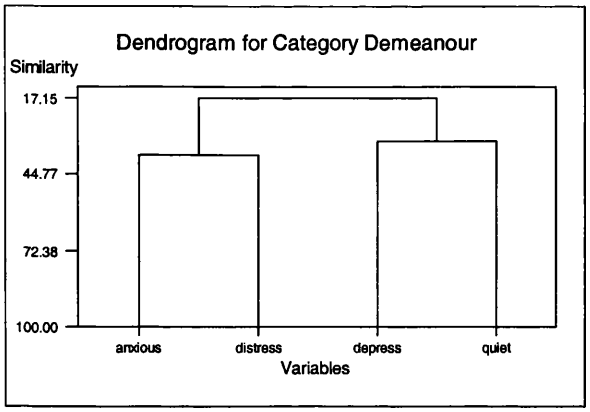
Response to food

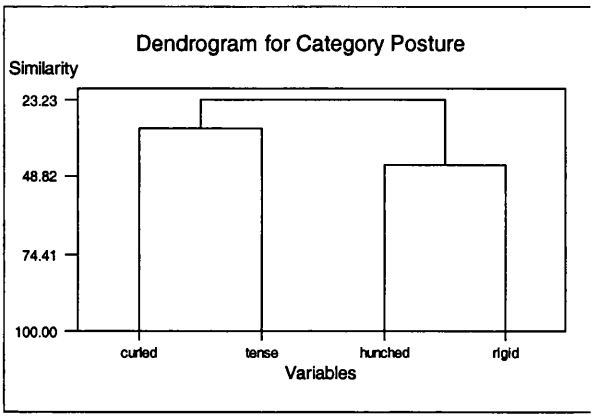


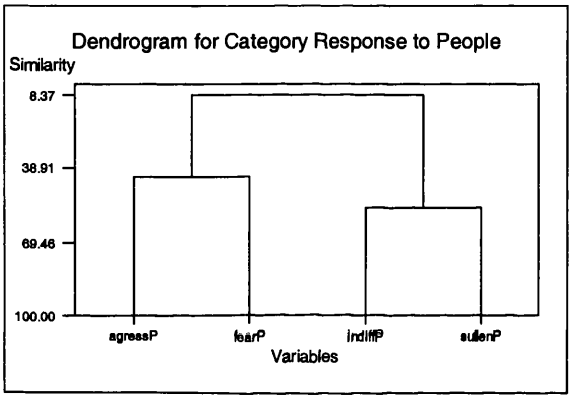
Physiological parameters

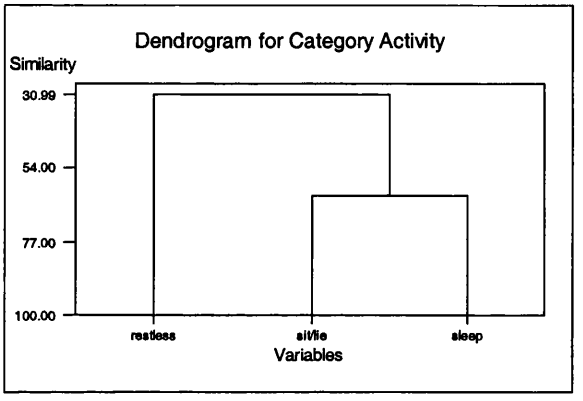


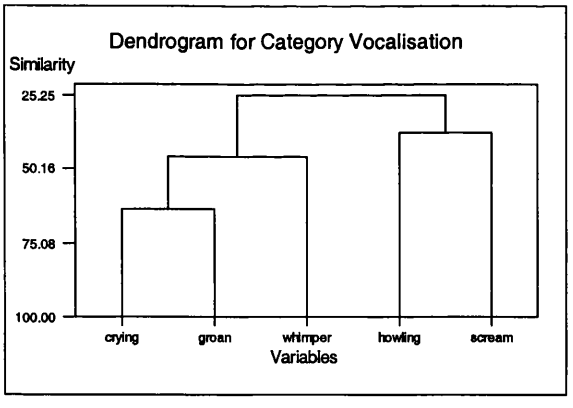
Section 2: Dendrograms for cluster analysis of behaviours and physiological signs

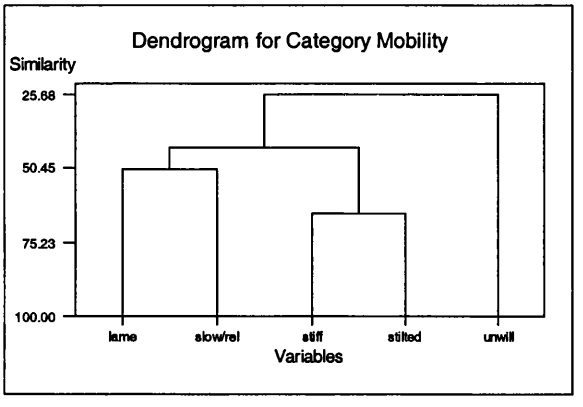




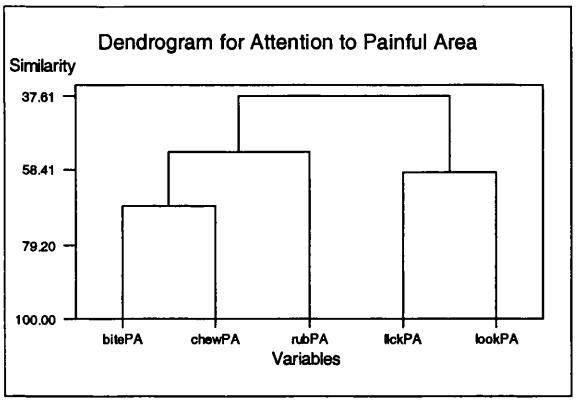


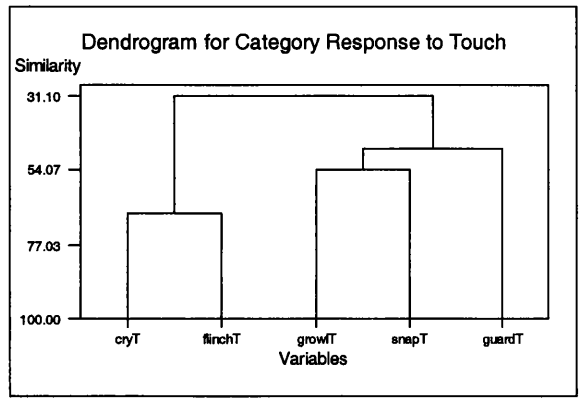


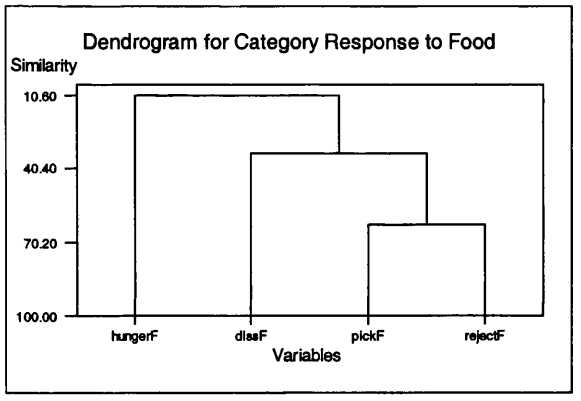


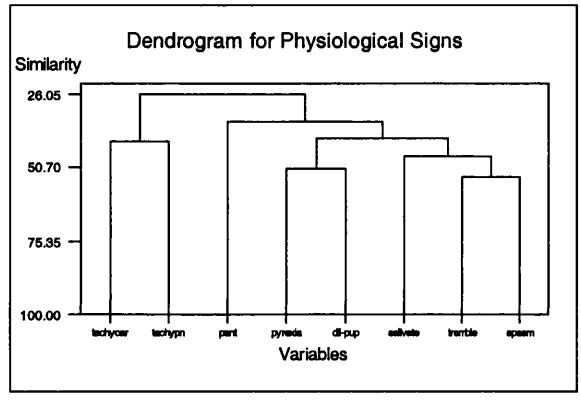


Section 2 continued: Dendrograms for cluster analysis of behaviours and physiological signs









Section 3: Tukey intervals for pairwise comparisons used to compare the mean VAS pain intensity scores allocated to behaviours and physiological signs by 72 practicing veterinary surgeons.

Category: Demeanour

	Depressed	Distressed	Quiet
Anxious	(-4.1,15.1)	(19.3,38.5)	(-20.5,-1.1)
Depressed		(13.9,32.)	(-25.9,6.7)
Distressed			(-49.3,-30.1)

Category: Posture

	Hunched	Rigid	Tense
Curled up	(23.7,39.7)	(44.2,60.1)	(26.4,42.3)
Hunched		(12.4,28.4)	(-5.4,10.7)
Rigid			(-25.8,-9.8)

Category: Response to People

	Fearful	Indifferent	Sullen
Aggressive	(-30.0,-11.6)	(-52.6,-34.1)	(-43.8,-25.3)
Fearful		(-31.7,-13.3)	(-22.9,4.5)
Indifferent			(-0.5,18.1)

Category: Activity

	Sitting/lying still	Sleeping
Restless	(-17.1,-1.0)	(-39.2,-23.1)
Sitting/lying still		(-30.1,-13.9)

Category: Vocalisation

	Groaning	Howling	Screaming	Whimpering
Crying	(-0.1,19.6)	(1.0,20.6)	(26.3,45.9)	(-6.6,13.0)
Groaning		(-8.8,10.8)	(16.5,36.1)	(-16.3,3.3)
Howling			(15.5,35.1)	(-17.4,2.3)
Screaming				(-42.7,-23.1)

Category: Mobility

	Slow/Reluctant	Stiff	Stilted	Unwilling/Unable
Lame	(-8.4,8.4)	(-21.2,-4.6)	(-19.5,2.9)	(9.6,26.3)
Slow/Reluctant		(-21.2,-4.6)	(-19.45,-2.9)	(9.6,26.2)
Stiff			(-6.5,9.9)	(22.6,39.1)
Stilted				(20.8,37.3)

Category: Attention to Painful Area

	Chewing	Licking	Looking	Rub/Scratch
Biting	(-16.3,3.4)	(-27.3,-7.7)	(-31.2,-11.3)	(-34.5,14.7)
Chewing		(-20.9,-1.3)	(-24.8,-5.0)	(-28.0,-8.3)
Licking			(-13.7,6.1)	(-16.9,2.8)
Looking				(-13.3,6.7)

Category: Response to Touch

	Flinching	Growling	Guarding	Snapping
Crying	(-17.3,0.2)	(-10.4,6.7)	(-13.2,3.9)	(1.0,18.2)
Flinching		(-1.6,15.5)	(-4.4,12.7)	(9.7,26.9)
Growling			(-11.3,5.8)	(2.8,20.0)
Guarding				(5.6,22.8)

Category: Response to Food

	Eating hungrily	Picking at food	Rejecting food
Disinterested	(-48.5, -33.5)	(-23.1, -8.1)	(-0.7,14.3)
Eating hungrily		(17.9,32.9)	(40.3,55.3)
Picking at food			(15.0,29.9)

Category: Physiological signs

	Panting	Tachy pnoea	Pyrexia	Salivate	Tremble	Muscle Spasm	Dilated Pupils
Tachy cardia	(-15, 5)	(-7, 14)	(-16, 4)	(-25, -4)	(-24, -4)	(-8, 12)	(-6, 15)
Panting		(-1, 19)	(-11, 10)	(-20, 1)	(-19, 2)	(-3, 18)	(0, 21)
Tachy pnoea			(-19, 1)	(-28, 7)	(-28, -7)	(-11, 9)	(-9, 12)
Pyrexia				(-19, 2)	(-18, 2)	(-2, 18)	(0, 21)
Salivate					(-10, 11)	(6, 27)	(9, 30)
Tremble						(6, 26)	(8, 29)
Spasm							(-13, 8)

Section 4: Test statistics for Kolmogorov Smirnov tests for each pairwise comparison of the mean VAS pain intensity scores allocated to behaviours and physiological signs by 72 practicing veterinary surgeons, within each category .

Where ✓: non-significant result, × : significant result, ? = borderline result.

Category: Demeanour

	Depressed	Distressed	Quiet
Anxious	0.16, ✓	0.58, ×	0.39, ×
Depressed		0.47, ×	0.40, ×
Distressed			0.68, ×

Category: Posture

	Hunched	Rigid	Tense
Curled up	0.66, ×	0.75, ×	0.69, ×
Hunched		0.51, ×	0.08, ✓
Rigid			0.42, ×

Category: Response to People

	Fearful	Indifferent	Sullen
Aggressive	0.47, ×	0.70, ×	0.62, ×
Fearful		0.51, ×	0.37, ×
Indifferent			0.31, ×

Category: Activity

	Sitting/lying still	Sleeping
Restless	0.27, ×	0.69, ×
Sitting/lying still		0.53, ×

Category: Vocalisation

	Groaning	Howling	Screaming	Whimpering
Crying	0.20, ✓	0.24, ?	0.67, ×	0.11, ✓
Groaning		0.14, ✓	0.59, ×	0.19, ✓
Howling			0.48, ×	0.26, ×
Screaming				0.69, ×

Category: Mobility

	Slow/Reluctant	Stiff	Stilted	Unwilling/Unable
Lame	0.09, ✓	0.29, ×	0.31, ×	0.38, ×
Slow/Reluctant		0.33, ×	0.33, ×	0.42, ×
Stiff			0.14, ✓	0.60, ×
Stilted				0.63, ×

Category: Attention to Painful Area

	Chewing	Licking	Looking	Rub/Scratch
Biting	0.16, ✓	0.33, ×	0.37, ×	0.49, ×
Chewing		0.25, ?	0.31, ×	0.39, ×
Licking			0.16, ✓	0.23, ?
Looking				0.15, ✓

Category: Response to Touch

	Flinching	Growling	Guarding	Snapping
Crying	0.21, ?	0.09, ✓	0.15, ✓	0.25, ?
Flinching		0.20, ✓	0.16, ✓	0.42, ×
Growling			0.15, ✓	0.30, ×
Guarding				0.33, ×

Category: Response to Food

	Eating hungrily	Picking at food	Rejecting food
Disinterested	0.86, ×	0.42, ×	0.22, ?
Eating hungrily		0.74, ×	0.87, ×
Picking at food			0.52, ×

Category: Physiological Signs

	Panting	Tachy pnoea	Pyrexia	Salivate	Tremble	Muscle Spasm	Dilated Pupils
Tachy cardia	0.24, ?	0.12, ✓	0.19, ✓	0.35, ✗	0.34, ✗	0.08, ✓	0.23, ?
Panting		0.26, ✗	0.09, ✓	0.24, ?	0.25, ?	0.26, ✗	0.25, ?
Tachy pnoea			0.24, ?	0.37, ✗	0.37, ✗	0.07, ✓	0.17, ✓
Pyrexia				0.27, ✗	0.26, ✗	0.23, ✓	0.24, ✗
Salivate					0.11, ✓	0.37, ✗	0.36, ✗
Tremble						0.35, ✗	0.34, ✗
Spasm							0.18, ✓

Appendix 3: Paired comparisons of items included in the Compoiste Measurement Pain Scales

When a dog is in pain it will exhibit a number of spontaneous and evoked behaviours which give an indication of the animal's pain intensity. A number of these behaviours are listed in pairs below. For each of the pairs please indicate which would imply a higher intensity of pain if observed in a dog. To do this please circle one of the expressions. Please complete this for all pairs of behaviours paying particular attention to the context in which the behaviours are observed.

When assessing the dog's posture, for each of the following pairs, which behaviour would indicate the highest pain intensity? Please circle the appropriate expression.

Normal Posture	vs	Rigid
Hunched	vs	Normal Posture
Hunched	vs	Rigid

When assessing the animals level of comfort which of the following expressions indicate the highest intensity of pain? Please circle.

Restless	vs	Comfortable

If the dog was vocalising, which behaviour from the following pairs would indicate the highest pain intensity? Please circle.

Screaming	vs	Crying or Wimpering
Crying or Wimpering	vs	Not Vocalising
Groaning	vs	Not Vocalising
Not Vocalising	vs	Screaming
Groaning	vs	Crying or Wimpering
Screaming	vs	Groaning

If the dog was paying particular attention to its' wound, which behaviour from the following pairs would indicate the highest pain intensity? Please circle.

Chew	vs	Ignoring wound
Lick/Look/Rub	vs	Ignoring wound
Chew	vs	Lick/Look/Rubbing wound

When assessing the dog's demeanour, for each of the following pairs, which characteristic would indicate the highest pain intensity? Please circle.

Nervous	vs	Depressed
Content	vs	Disinterested
Nervous	vs	Bouncy
Aggressive	vs	Depressed
Quiet	vs	Depressed
Disinterested	vs	Depressed
Depressed	vs	Content
Nervous	vs	Disinterested
Disinterested	vs	Quiet
Aggressive	vs	Disinterested
Bouncy	vs	Depressed
Content	vs	Nervous
Bouncy	vs	Disinterested
Content	vs	Nervous
Quiet	vs	Aggressive
Content	vs	Aggressive
Nervous	vs	Quiet
Content	vs	Quiet
Aggressive	vs	Nervous
Quiet	vs	Bouncy
Bouncy	vs	Content

When assessing the dog's mobility, which behaviour from the following pairs would indicate the highest pain intensity? Please circle.

Slow	vs	Stiff
Slow	vs	Mobile
Lame	vs	Mobile
Stiff	vs	Mobile
Stiff	vs	Lame
Slow	vs	Lame

When assessing the animal gentle even pressure is applied to the area immediately surrounding any wound that is apparent, either from surgery or trauma. Which reaction to this touch, from the following pairs, would indicate the highest pain intensity? Please circle.

Flinch	vs	Cry
Growl	vs	Cry
Snap	vs	Flinch
Snap	vs	Cry
Growl	vs	Flinch
No reaction	vs	Snap
No reaction	vs	Flinch
Snap	vs	Growl
No reaction	vs	Cry
Growl	vs	No reaction

Appendix 4: Definitions of items included in the Composite Measurement Pain Scales.

When recording a dogs behaviours using the pain assessment scale each of the behaviours described in the scale are defined as follows.

Posture

Rigid: Animal lying lateral recumbancy, legs extended or partially extended in a fixed position.
Hunched: When animal is standing, its back forms a convex shape with abdomen tucked up, or its back in a concave shape with shoulders and front legs lower than hips.
Tense: Animal appears frightened or reluctant to move, with an overall impression of tight muscles. Animal can be in any body position.
Normal posture: Animal may be in any position but appears comfortable, with muscles relaxed.

Comfort

Restless: Moving bodily position, circling, pacing, shifting body parts, unsettled.
Comfortable: Animal settled, resting and relaxed no avoidance or abnormal body position evident. Remains in same body position, at ease.

Vocalisation

Whimpering: Often quiet short high pitched sound, frequently closed mouth. Whining
Crying: Extension of the whimpering noise, louder and with open mouth.
Groaning: Low moaning or grunting deep sound, intermittent.
Screaming: Animal making a continual high pitched noise, inconsolable, mouth wide open.

Attention to wound area

Chewing: Using mouth and teeth on wound area, pulling stitches

Licking: Using tongue to stoke area of wound.

Looking: Turning head in direction of area of wound.

Rubbing: Using paw or kennel floor etc to stroke wound area.

Ignoring: Paying no attention to the wound area.

Demeanour

Aggressive: Mouth open or lip curled showing teeth, snarling, growling, snapping or barking.

Depressed: Dull demeanour, not responsive, shows reluctance to interact.

Disinterested: Cannot be stimulated to wag tail or interact with observer.

Nervous: Eyes in continual movement, often head and body movement, jumpy.

Anxious: Worried expression, eyes wide with whites showing, wrinkled forehead.

Fearful: Cowering away, guarding body and head.

Quiet: Sitting or lying still, no noise. Will look when spoken to, but not respond.

Indifferent: Not responsive to surroundings or observer.

Happy and Content: Interested in surroundings, has positive interaction with observer, responsive and alert.

Happy and Bouncy: Tail wagging, jumping in kennel often vocalising with a happy and excited noise.

Mobility

Stiff: Stilted gait, also slow to rise or sit, may be reluctant to move.

Slow to rise or sit: Slow to get up or sit down but not stilted in movement

Reluctant to rise or sit: Needs encouragement to get up or sit down.

Lame: Irregular gait, uneven weight bearing when walking

Normal mobility: Gets up and lies down with no alteration from normal.

Response to Touch

Cry: A short vocal response. Looks at area and opens mouth, emits a brief sound.

Flinch: Painful area is quickly moved away from stimulus either before or in response to touch

Snap: Tries to bite observer before or in response to touch.

Growl: Emits a low prolonged warning sound before or in response to touch.

Guard: Pulls painful area away from stimulus or tenses local muscles in order to protect from stimulus.

No adverse response to touch: Accepts firm pressure on wound with none of the aforementioned reactions.

Appendix 5: Details of animals included in study to examine the performance of the Composite Measurement Pain Scale

Table 1: Listing of breeds of dogs and procedures undergone included in a study to compare the performance composite measurement pain scale (CMPS). Animals had undergone orthopaedic surgery

Breed	Procedure	Age (yrs)	Sex
Golden Labrador	Cruciate repair	2.5	F
Golden Labrador	Joint flush	8	M
Old English Sheep Dog	Aspergillosis	6	F
King Charles Spaniel	Tibial crest transplant	1	F
King Charles Spaniel	Carpal arthrodesis	1.75	M
Staffordshire Terrier	Biopsy, soft tissue mass	1.5	M
Cross-breed	Repair hip dislocation	*	M
Yorkshire Terrier	Cruciate repair	6.16	M
Collie Cross-breed	Cruciate repair	9	FN
Cross-breed	Cruciate repair	8	F
Rottweiller	Triple pelvic osteotomy	0.42	M
Rottweiller	Cruciate repair	2.5	M
Beagle	Ventral slot	7	M
Labrador	Ventral slot	9	F
Cross-breed	Arthrotomy, shoulder	1	F
Collie Cross-breed	Mandibulectomy	5	FN
Cross-breed	Arthrotomy, shoulder	0.58	F
Great Dane	Cruciate repair	4	MN
Airedale Terrier	Cruciate repair	0.58	M
Cross-breed	Missing	6.5	FN

* Missing value

Table 2: Listing of breeds of dogs and procedures undergone included in a study to compare the performance composite measurement pain scale (CMPS). Animals had undergone soft tissue surgery

Breed	Procedure	Age (yrs)	Sex
Lurcher Cross	Perineal hernia repair	9	MN
German Shepherd Dog	Implant nasal drain	1.5	M
Rottweiller	Total ear canal ablation	8	*
Golden Retriever	Repair macerated foot	0.42	F
Bernese Mountain Dog	Exploratory thoracotomy	4.5	M
German Shepherd Dog	Exploratory laparotomy	9.5	FN
Staffordshire Terrier	Total ear canal ablation	10	FN
English Springer Spaniel	Foreign body removal	4.5	M
French Bull Dog	Lateral wall resection, ear	4	M
Boxer	Prostatic cyst removal	7	M
Cross-breed	Anal furunculosis	6.5	M
German Shepherd Dog	Total ear canal ablation	5	M
Labrador	Cataract removal	8	FN
Greyhound	Castration	*	M
Yorkshire Terrier	Cataract removal	8.67	M
Doberman Pinscher	Lung lobectomy	7	F
Chow Chow	Biopsy, soft tissue mass	3	M
German Shepherd Dog	Biopsy, soft tissue mass	10.58	M
Cross-breed	Exploratory laparotomy	3.5	FN
Golden Retriever	Biopsy, soft tissue mass	4.75	M

* Missing value

Table 3: Listing of breeds of dogs and medical conditions included in a study to compare the performance of the composite measurement pain scale (CMPS). Animals were hospitalised because of medical condition.

Breed	Condition	Age (yrs)	Sex
Cocker Spaniel	Hepatic failure	3.5	FN
Airedale	Immunological investigation	8	M
Sheltie	Dermatological investigation	7.1	FN
English Springer Spaniel	Cardiomyopathy	11	FN
Yorkshire Terrier	Diabetes mellitus	*	*
Labrador	Diabetes mellitus	9	M
Doberman Pinscher	Diabetes mellitus	8.5	M
West Highland White Terrier	VSD	4.5	FN
Rottweiller	Diabetes mellitus	6	FN
German Shepherd Dog	CDRM	11.3	FN
Labrador	Hypothyroid	5	F
Cocker Spaniel	Lymphoma investigation	8.5	F
Flat Coat Retriever	Missing	2	FN
West Highland White Terrier	Heart condition	5	F
Corgie	Cushings disease	10	FN
Dalmation	Peripheral shunt	0.42	F
Boxer	Dilated cardiomyopathy	9	M
English Springer Spaniel	Pyrexia, unknown origin	5.5	FN
Cocker Spaniel	Cushings disease	12.5	M
West Highland Terrier	Diabetes mellitus	8	M

* Missing value

Appendix 6: Details of animals included in a video study to examine generalizability of the composite measurement pain scale (CMPS)

Table 1: Listing of breeds of dogs and procedures undergone included in a video study to examine the generalizability composite measurement pain scale (CMPS). Animals had undergone surgery in the morning, prior to the video assessment being carried out.

Breed	Procedure	Age (yrs)	Sex	Video Study?
Labrador	Pelvic split	8	FN	No
Flat Coat Retriever	Bone marrow transplant	10.5	FN	No
King Charles Spaniel	Repair luxated patella	0.83	M	No
West Highland White Terrier	Bone marrow biopsy	6	M	No
Border Terrier	Castration	11	M	No
Border Collier	Removal of anal mass	3.5	M	No
English Springer Spaniel	Cruciate repair	8	F	No
Cross breed	Biopsy mouth tumour	11	M	Yes
Golden Retriever	Laryngeal tieback	6.75	FN	Yes
Great Dane	Arthrotomy, shoulder	0.75	M	Yes
Labrador	Cruciate repair	6	M	Yes
Sheltie	Hernia Repair	1	M	Yes
Weimeranea	Colopexy	5	M	Yes
Collie Cross breed	Fracture repair	1	M	Yes
Labrador	Remove rectal polyps	4.5	M	No
Corgie	Spey	3.67	F	No
Retriever	Bone marrow biopsy	12	FN	Yes
Weimeranea	Bone marrow biopsy	8.92	FN	Yes
Scottie	Remove rectal polyps	7	M	Yes
King Charles Spaniel	Cruciate repair	8	M	Yes
Border Terrier	Total ear canal ablation	10	F	Yes

Appendix 7: The Composite Measurement Pain Questionnaire

Approach the kennel, ensuring you do not have a lab coat or theatre greens on as these may cause a change in the animal's behaviour. While you approach the kennel, look at the dog's behaviour and reactions and answer the following questions.

Look at the dog's posture, does it seem...

Rigid	
Hunched or Tense	
Neither of these	

Does the dog seem to be....

Restless	
Comfortable	

If the dog is vocalising is it...

Crying or Whimpering	
Groaning	
Screaming	
Not vocalising/none of these	

If the dog is paying attention to its wound is it...

Chewing	
Licking or Looking or Rubbing	
Ignoring its wound	

Now approach the kennel door and call the dog by name. Then open the door and encourage the dog to come to you. From the dog's reaction to you when watching him/her assess the animal's character.

Does the dog seem to be...

Aggressive	
Depressed	
Disinterested	
Nervous or Anxious or Fearful	
Quiet or Indifferent	
Happy and Content	
Happy and Bouncy	

The next assessment is the dog's physiological responses. Record the dog's respiratory rate in the space below. Moving the dog as little as possible, record a heart rate (by direct palpation or using a stethoscope) counting for a minimum of 15 seconds. Also, look at the dog's eyes and assess whether the pupils are dilated.

Respiratory Rate	
Heart rate/pulse	
Dilated pupils	

If the dog is mobile open the kennel and put a lead on the dog. If the dog is sitting down get it to stand and then come out of the kennel, then walk slowly up and down the area outside the kennel. On returning to the kennel ask the dog to sit down.

During this procedure did the dog seem to be...

Stiff	
Slow or Reluctant to rise or sit	
Lame	
None of these	
Assessment not carried out	

The next procedure is to assess the dog's response to touch. If a wound is visable apply gentle pressure using two fingers to the wound and an area approx. 4cm around it. If the position of the wound is such that it is impossible to touch then apply the pressure to the closest point to the wound. If there is no wound then apply the same pressure to the stifle and surrounding area.

When touched did the dog...

Cry	
Flinch	
Snap	
Growl or Guard wound	
None of these	