ASSESSING THE PSYCHOMETRIC PROPERTIES OF THE BRITISH COLUMBIA
PAIN BEHAVIOUR TAXONOMY (BCPBT)

by

Elizabeth A. Hughes

B.A. (Hon.) Simon Fraser University, 1995

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
in
PSYCHOLOGY

THE UNIVERSITY OF NORTHERN BRITISH COLUMBIA

November, 2003

© Elizabeth A. Hughes, 2003
NOTICE:
The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.
ABSTRACT

A new observational pain behaviour taxonomy, the British Columbia Pain Behaviour Taxonomy (BCPBT), was developed by modifying a previously existing taxonomy created by Keefe and Block (1982). The BCPBT was designed to be consonant with recent research findings and the pragmatic concerns of the Workers' Compensation Board of British Columbia (WCB-BC), for whom the modifications were made. A series of six studies was conducted to analyze the psychometric properties of this new measure. In Study 1, the BCPBT was assessed for accuracy of application over time. A multiple-baseline design was used to measure the extent to which a select group of judges could master the taxonomy. Training in the BCPBT was staggered, with the Immediate Training group receiving the first 5-hour training seminar. Judges had moderately high agreement scores on the pretest; the training seminar did improve coding agreement, but these improvements failed to reach statistical significance. However, all judges did eventually reach an adequate level of coding proficiency (mean agreement = 83%, $\kappa = .66$). Study 2 involved the demonstration of test-retest reliability across judges trained in the BCPBT over a maximum delay of two weeks; a mean Pearson product-moment correlation was computed incorporating all post-training percent agreement scores (mean $r = .89$), indicating a high level of reliability over time. In Study 3, judges were tested on their retention of the taxonomy five months after their first exposure, and feedback of their performances was given in an attempt to increase mean percent agreement. It was determined that trained judges could surpass the minimum agreement criterion of 80% (mean agreement = 83.8%, $\kappa = .67$), but were unable to reach a mean agreement of 90%, a criterion that has been used by past researchers. In Studies 4A and 4B, WCB-BC
judges trained in the BCPBT were shown to achieve adequate agreement levels after one training session, both in an experimental setting (mean agreement$_{4A} = 80.9\%$, $\kappa = .62$), and in a clinical setting (mean agreement$_{4B} = 87.7\%$, $\kappa = .69$). For Study 5, the scores of 67 patients who had been administered the BCPBT twice within a three day period were used to assess the test-retest reliabilities of the measure and each behavioural category. The results indicated that facial expression, guarding, and total pain behaviour score had strong relationships over time, while the remaining behavioural categories had moderately sized correlations. Study 6 focused on refining the BCPBT's psychometric properties by assessing its internal consistency and component structure using the data collected from 120 WCB-BC claimants. Using Cronbach's alpha, the BCPBT was streamlined by removing all items that had consistently low item-total correlations. Principal components analysis revealed a strong underlying component structure for four of the five behavioural types. On the strength of all the combined information gathered over the six separate studies, it appears that the BCPBT is easy to learn, and demonstrates satisfactory inter-rater and test-retest reliability. The psychometric changes made to the BCPBT in response to the internal consistency and component structure information generated by Study 6 must be examined in subsequent research.
# TABLE OF CONTENTS

Abstract iii

List of Tables viii

List of Figures ix

Acknowledgement x

Introduction 1

The Low Back Disability Project 6
Previous Pain Behaviour Observational Systems 12
The Behavioural Assessment Protocol 23
The British Columbia Pain Behaviour Taxonomy (BCPBT) 25
The Present Studies 30

Study 1 Validation of the Training Protocol for the BCPBT 31
Method 32
Judges 32
Materials 32
Training 32
Testing 33
Equipment 34
Design 35
Procedure 35
Training 36
Data scoring 38
Data analysis 39
Results 39

Study 2 Test-retest Reliability of the UNBC Judges 45
Method 45
Judges 45
Materials 45
Design and procedure 46
Data scoring 46
Results 46

Study 3 The Effect of Feedback on BCPBT Accuracy 48
Method 49
Judges 49
Materials 49
Design and procedure 49
Properties of the BCPBT

Data scoring 50
Results 50

Study 4
Part I: Effects of Training on WCB Judges 52
Method 52
Judges 52
Materials 52
Design and procedure 53
Data scoring 53
Results 53

Part II: Using the BCPBT in a Clinical Setting 54
Method 54
Judges 54
Materials 54
Design and procedure 55
Data scoring 57
Results 57

Study 5
Clinical Test-Retest Reliability of the BCPBT 58
Method 58
Judges 58
Design and procedure 59
Data scoring 59
Results 60

Study 6
Part I: The Internal Consistency of the BCPBT 63
Method 63
Judges 63
Design and procedure 64
Data scoring 64
Results 64

Part II: The Factor Structure of the BCPBT 71
Method 71
Judges 71
Design and procedure 71
Data scoring 72
Data screening 72
Results 73

Discussion
Reliability 83
Internal Consistency 83
Principal Components Analyses 89
List of Tables

<table>
<thead>
<tr>
<th>Table 1.</th>
<th>Comparison of Pre-Training and Post-Training Agreement Scores for Immediate and Delayed Training Groups (N = 8)</th>
<th>41</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 2.</td>
<td>Epochs For the WCB-BC Standardized Physical Examination</td>
<td>56</td>
</tr>
<tr>
<td>Table 3.</td>
<td>The Mean Frequency of Pain Behaviours in the Test-Retest Group of WCB Claimants (N=67)</td>
<td>61</td>
</tr>
<tr>
<td>Table 4.</td>
<td>The Test-Retest Correlations of the Pain Behaviours of the BCPBT Across Claimants Over a Three-Day Period (N=67)</td>
<td>62</td>
</tr>
<tr>
<td>Table 5.</td>
<td>The Mean Frequency of Pain Behaviours in the One-Time Examination of WCB Claimants (N=120)</td>
<td>66</td>
</tr>
<tr>
<td>Table 6.</td>
<td>The Change in Cronbach’s Alpha Statistics for the Five Behavioural Categories, Before and After the Removal of Low-Loading Coding Epochs (N = 120)</td>
<td>67</td>
</tr>
<tr>
<td>Table 7.</td>
<td>Item-Total Correlations For The Nine Excluded Epochs (N=120)</td>
<td>70</td>
</tr>
<tr>
<td>Table 8.</td>
<td>Principal Components Analysis: Factor Loadings For Each Pain Behaviour Category in the First Examination of WCB Claimants, Summing Across Epochs (N=120)</td>
<td>74</td>
</tr>
<tr>
<td>Table 9.</td>
<td>Principal Components Analysis: Primary Eigenvalue Extracted For Each Pain Behaviour Category in the First Examination of WCB Claimants, Across Epochs (N=120)</td>
<td>78</td>
</tr>
<tr>
<td>Table 10.</td>
<td>Principal Components Analysis: Factor Loadings For Individual Behavioural Category-Epochs For the First Examination of WCB Claimants (N=120, Variables=106)</td>
<td>79</td>
</tr>
<tr>
<td>Table 11.</td>
<td>Principal Components Analysis, Split-Half 1: Factor Loadings For Individual Behavioural Category-Epochs For the First Examination of WCB Claimants (N=120, Variables=53)</td>
<td>80</td>
</tr>
<tr>
<td>Table 12.</td>
<td>Principal Components Analysis, Split-Half 2: Factor Loadings For Individual Behavioural Category-Epochs For the First Examination of WCB Claimants (N=120, Variables=53)</td>
<td>81</td>
</tr>
</tbody>
</table>
List of Figures

Figure 1. Mean Percent Agreement Over Test by Group 40
The author would like to acknowledge the gracious assistance of the Workers’ Compensation Board of British Columbia for making this research possible, and thank them for their funding efforts, the use of their materials and facilities, and their general spirit of co-operation. Additional thanks must go to Dr. Kenneth Prkachin for his unflagging patience and support in this endeavour, and to Dr. Peter MacMillan and Dr. Cindy Hardy for their invaluable input into the project.
Assessing the Psychometric Properties of the British Columbia Pain Behaviour Taxonomy (BCPBT)

Disability due to chronic low back pain is ubiquitous and expensive (Osterweis, Kleinman, & Mechanic, 1987). In both human and financial terms, it exacts a terrible cost from those it affects. Its incidence and prevalence have increased so dramatically in the last fifty years that at least one researcher has likened it to an epidemic (Waddell, 1987). Fordyce (1985) estimated that the rate of low-back-pain-induced disability in the United States increased fourteen times faster than the population growth between 1957 and 1976. It is estimated that 80% of all adults will report experiencing at least one disabling bout of low back pain at some time in their lives (Waddell, 1987). In the 1994-95 census, 3.3 million adult Canadians, 13% of the total population, reported experiencing chronic low back pain (Statistics Canada, 1996). Annually, a small but significant number of the North American working population will qualify for compensation due to work-related low back pain. Between 1% and 5% of those in the workforce will suffer from acute low back pain resulting in compensable absences from

---

1 The definition of “chronic” pain is not a simple one. “Chronic” pain may be both qualitatively and quantitatively different from “acute” pain; while acute pain is often a direct reflection of tissue damage to the body, and thus is driven by external sensory input, many people believe that chronic pain is more centrally driven and reactive to operant factors (Crue, Kenton, Carregal & Pinsky, 1980). It would appear that chronic pain is a more complex phenomenon than acute pain, in that previously adaptive pain-relieving functions of the organism appear to become maladaptive as the condition progresses. A vicious cycle, chronic pain often perpetuates itself through the pain-relieving behaviours of those afflicted with it (Schwartz, Tapp & Brucker, 1985). However, most researchers’ operational definitions merely focus on the differential duration of each type of pain as their definitive attributes. Vesudevan (1992) stated that many pain researchers arbitrarily consider six months to be delineation between “acute” and “chronic” pain; the Subcommittee on Taxonomy of the International Association for the Study of Pain (1986) reduced this period to three months. Others have taken a more relativistic view. Bonica (1990) wrote that acute pain becomes chronic after the normal amount of time it usually takes the specific type of injury to heal; other researchers have said that the critical determination point occurs one month after Bonica’s definition. Regardless of exactly how “chronic pain” is defined, those who have it had acute pain at one time, and it persisted.
work (Frymoyer & Cats-Baril, 1987), although later studies have estimated this figure to be as high as 7.6% (Carey, Evans, Hadler, Lieberman, Kalsbeek, Jackman, Fryer, & McNutt, 1996). A smaller number of these people will retain their affliction and acquire disabled status with disastrous ramifications for both the person and the system that must then provide financial support. In British Columbia, the annual incidence of low back pain-related disability is small, approximately 0.6% of the total population. However, the prevalence of disabled workers is cumulative, and each of these workers depends upon the Workers' Compensation Board of British Columbia (WCB-BC) for the economic survival of themselves and their families. In total, it is estimated that the annual cost of supporting workers disabled due to low back pain is over $50 million (WCB-BC, 1996). Across Canada, the annual cost of chronic low back injury is believed to be between $1.875 and $2.25 billion (WCB-BC, 1996).

One of the biggest hindrances in addressing the expensive problem of disability caused by chronic low back pain is that it is a little understood condition, and the antecedent factors instrumental to its development remain unclear (Turk & Melzack, 1992). The relationships among tissue injury, chronic low back pain, and disability are not isotonic (Osterweis et al., 1987; WHO, 1980). The extent to which tissue is damaged predicts neither the intensity of the pain experienced (Melzack & Wall, 1983) nor the development of a permanent disability² (Waddell, 1987; WHO, 1980). A direct, positive relationship between tissue damage and chronic pain intensity has been observed in a few studies (Hunter, 2001), and some research has noted a strong correlation between chronic pain levels and disability (Hazard, Haugh, Reid, Preble, & MacDonald, 1996). However,
these relationships have not been consistently found by other researchers (Astrand & Isaacsson, 1988; Fordyce, Lansky, Calsyn, Shelton, Stolov, & Rock, 1984; Lehmann, Spratt, & Lehmann, 1993; Waddell, 1987; Waddell, 1991). The connections among the three phenomena remain inconclusive, which suggests that other factors are mediating the relationships among them. For over 40 years, and in a multiplicity of disciplines, dozens of possible factors have been examined, such as age and sex (WCB-BC, 1996; Williams, 1988), depression (Keefe, Wilkins, Cook, Crisson, & Muhlbaier, 1986), social skills (Williams, 1988), coping strategies (Devine & Spanos, 1990; Keefe, Crisson, Urban, & Williams, 1990), beliefs of the etiology of low back pain (Walsh & Radcliffe, 2002), fear of pain (Crombez, Vlaeyen, Heuts, & Lysens, 1999; Waddell, Newton, Henderson, Somerville, & Main, 1993), and functional assessments (Hazard et al, 1996; WCB-BC, 1996), but consistent results in this area remain elusive.

In fact, the results of previous research have not only been elusive, the results of different studies have often been completely contradictory. For example, the WCB-BC has estimated that, using the information gathered as a standard part of its evaluative process in work-related low back injury cases, only 21.7% of the variance of disability from chronic low back pain can be explained (WCB-BC, 1996). Yet Frymoyer and Cats-Baril (1987), after using a team of experts to determine the possible predictive factors of chronic pain and disability and examining these factors in a cohort study of a mixed sample of acute and chronic pain patients, noted that a combination of demographic, personal and cognitive factors explained 89% of the variance of long-term disability, far more than could be explained by physical and psychological measures alone. Amongst

2 "Permanent disability" has been defined as the inability to fulfill role or employment specific expectations
the primary predictors was self-rated pain intensity, self-prediction of disability, perceptions of responsibility for the injury, income, education, length of pain problem, and job-related factors.

Conversely, Astrand and Isaacsson (1988), when examining the results of a rigorous longitudinal study more than two decades in duration, found that self-report of general health was the greatest predictor of long-term disability and job turnover due to back pain, and that neither demographic factors, such as age, nor physical factors, such as pathoanatomical findings and the self-rated severity of back pain, were strongly associated with these outcomes.

Lehmann, Spratt and Lehmann (1993) also observed only one significant correlation with disability due to chronic low back pain in their investigations, but it was not the same factor noted by Astrand and Isaacsson. To the exclusion of all other demographic variables, work and health indices, physical factors and pain level, only marital status was predictive of disability. Unmarried patients were much more likely to develop disabling low back pain than their married counterparts. This finding can be contrasted with that of other researchers, such as Block, Kremer and Gaylor (1980), Fordyce (1976) and Romano and Turner (1995) who noted that simply having a spouse does not prevent low back pain based disability. Being married to a solicitous, or particularly sensitive and assisting, spouse had a positive linear association with the severity of disability in depressed patients and with the frequency of pain behaviour in nondepressed patients. Therefore, not only is living alone predictive of eventual...
disability, apparently so is the presence of family dynamics that encourage dependent behaviour.

Other research attempting to tie neuropathy, pain intensity and disability together has been equally inconsistent. Council, Ahern, Follick and Kline (1988) reported that only the chronic low back pain patients’ own beliefs of their functional abilities predicted their performance on a physical examination, to the exclusion of all other factors including pain intensity. Similarly, Waddell et al (1993) noted that it was patients’ fears of returning to work and of injuring themselves further that predicted approximately 44% of the variance in return to work after developing a chronic low back pain condition. Once the severity of the patient’s pain was taken into account, the relationship dropped somewhat, but these pain-related fears still remained potent predictors of the patients’ re-entry into the workforce.

Hazard et al (1996) constructed an 11-item questionnaire to predict disability from chronic low back pain. The questionnaire incorporated questions regarding pain history and self-reported current intensity, self-predicted disability, perceptions of blame for the injury, marital status and job characteristics. The questionnaire proved to be both sensitive and specific in the prediction of disability three months later in an acute low back pain sample. A significant association was observed between self-reported pain intensity and disability, a finding unlike many that have come before it (Waddell, 1987). While this may be a veridical association for their sample, Hazard et al experienced substantial self-selection in the composition of this sample. Of the 699 workers with low back pain first approached, only 166 of them (23.7%) consented or were able to participate in the study, which could possibly explain the unusual finding.
Plainly, the relationship between chronic pain and disability is unclear, and more rigorous and inclusive research is needed if the antecedents of low-back-pain-induced disability are to be uncovered. As can be seen from the brief overview, certain factors may predict the development of a disability due to chronic low-back-pain, but the clarification that only rigorous and thorough research (that includes all the aforementioned factors in the equation) can provide is needed to say definitively (Osterweis et al, 1987).

The Low Back Disability Project

In 1997, the WCB-BC undertook to determine the factors predictive of eventual disability due to chronic low back pain by conducting a prospective study of new claims for work-related back pain. The Multivariate Prediction of Disability, Low Back project was based on an extensive new data collection process framed within a biopsychosocial approach to human behaviour. There were five domains of study: the standard information usually collected by the WCB-BC (which consists of demographic factors, conditions causing the injury, the severity and duration of injury, and other information immediately pertinent to the injury itself), in-depth information on the workplace, psychological and social data, a new standardized physical examination to provide consistent physical evaluations on each new claimant, and behavioural observation information. From the information gathered in these areas, the goal was to construct a predictive model to clarify the causal factors of disability in order to develop effective prevention resources, and to focus present resources in the most parsimonious manner.
It is the last domain, that of behavioural observation, that is of most interest to behavioural psychologists. The currently *zeitgeist* in pain research maintains that pain is multidimensional (Turk & Rudy, 1992); it is not a sensation, but an aggregate perception (Cailliet, 1993), and as such is composed of physiological, affective, cognitive and behavioural components (Turk & Melzack, 1992).

Pain is an inherently subjective experience that must be communicated if it is to be perceived and acted upon by others (Prkachin & Craig, 1994). The term coined for this communication, in whatever form it might take, was “pain behaviour” (Fordyce, 1976, p. 152). Fordyce (1966) first conceptualized pain behaviour in chronic pain patients as maladaptive perseverant behaviour that could be eliminated through the judicious application of operant conditioning techniques. Pain behaviour has since come to be recognized as the result of a complex interaction between the biological response to pain and learning factors (Bonica, 1990), and for the current research, is now simply defined as the behaviour demonstrated by someone experiencing pain, regardless of the function it serves.

Much like pain itself, pain behaviour can be viewed as multidimensional (Turk & Rudy, 1992), and its different subtypes vary in both nature and function (Prkachin, 1986). As has been discussed, some types of behaviour associated with pain are essentially communicative. For example, the characteristic facial expressions of pain (Prkachin, 1992) appear to have little to do with the relief or attenuation of pain, and probably have evolved as a system to alert others of one’s internal state (Buck, 1984; Williams, 1988). Communicative pain behaviour may also be self-preservational, in that a display of pain behaviour can serve to elicit assistance from others in escaping or
circumventing the painful stimulus (Fordyce, 1976; Fordyce, Roberts, & Sternbach, 1985). Alternately, some types of behaviour associated with pain are both communicative and amelioratory; for example, rubbing a painful area or taking analgesic medication are actions that not only potentially convey to others that one is hurt, but relieve some of the discomfort as well. Finally, certain behaviours can serve a compensatory function, allowing the afflicted to avoid further pain by avoiding action or using other muscle groups or external supports to compensate for the injured area. Examples of this type of behaviour are resting, limping, supporting the body’s weight on a walking aid, and using a wheelchair. Researchers tend to agree on the existence and many of the characteristics of pain behaviour, but their operational definitions of these behaviours and their interpretations of the structural interrelationships of these behaviours often varies from study to study (Feuerstein, Greenwald, Gamache, Papciak, & Cook, 1985; Keefe & Block, 1982; Richards, Nepomuceno, Riles, & Suer, 1982; Turk & Rudy, 1992).

The measurement of pain behaviour as an indicator of overall level of pain experienced is becoming more frequent in pain research, particularly for chronic pain. Previous research on the relationship between pain and disability often has focused on the self-reported intensity of pain to the exclusion of indices of other forms of pain behaviour (Craig & Prkachin, 1983). While self-report of pain intensity can be informative about the patient’s subjective state (Jensen & Karoly, 1992), it assesses only the cognitive aspect of the conscious awareness of pain, which is only one dimension of the complex pain experience.
Although the frequency of pain behaviour correlates well with self-rated pain intensity in some studies (Keefe & Block, 1982; Keefe, Crisson, & Snipes, 1987) and has been known to predict the chronicity of low back pain in others (Hasenbring, Marienfeld, Kuhlendahl, & Suyka, 1994), there is a substantial portion of unique variance between the two phenomena (Craig & Prkachin, 1983), which suggests that it is possible that pain behaviour may be the visible manifestation of processes that are not necessarily consciously available. Pain behaviour may be a phylogenetically earlier form of pain expression than is self-report, which is reliant upon the higher brain functions of self-awareness and introspection. Evidence of this can be seen in experimental pain situations. For example, while it is always possible to verbally dissimulate the absence of pain, certain forms of pain behaviour are difficult to consciously control; for example, contractions of the orbicularis oculi, the muscles that surround the eyes, are near impossible to suppress when moderately high to high nociceptive stimulation is applied, regardless of conscious effort (Craig, Hyde, & Patrick, 1997). Therefore, by studying pain behaviour independently from self report, it may be possible to gain insight into the characteristics of chronic low back pain, and perhaps even eventual disability.

Moreover, the self-report of pain has been observed to be potentially influenced by a number of factors independent of nociception. Demand characteristics, state-specific memory, social modeling (Prkachin, 1997), the cognitive reinterpretation of the nociceptive stimulation (Melzack & Wall, 1983) and even outright dissimulation can all alter the self-report of pain in both experimental and clinical situations.

Fordyce (1976; Fordyce, Roberts, & Sternbach, 1985), a pioneer in the field of pain behaviour and management, believes that pain behaviour in chronic pain patients
Properties of the BCPBT

gives the clinician a glimpse of the learning history of each patient, and perhaps of the
information needed to predict long-term disability. Pain behaviour is susceptible to
learning factors. The cessation or amelioration of pain that occurs after employing certain
pain behaviours, such as limping, can reinforce the behaviour and increase the probability
that the behaviour will occur again. Also, the operant conditioning that results from the
secondary gain garnered from the adoption of the "sick role" (Dworkin & Whitney, 1992)
can also increase the frequency of pain behaviours. This learning can occur without the
conscious awareness of the patients themselves, and, due to the behaviour-perpetuated
nature of chronic pain (Schwartz, Tapp, & Brucker, 1985), may be associated with
eventual chronic low back pain induced disability.

From a pragmatic perspective, it is well known that pain behaviour is clinically
informative to health care professionals (Craig & Prkachin, 1983). Past studies have
shown that in clinical settings, health care professionals often use the information
communicated by pain behaviours to make decisions about a patient's condition (Craig &
Prkachin, 1983; Hadjistavropoulos & Craig, 1994). However, research into clinical
practice suggests that pain behaviour is neither observed systematically (Keefe, Crisson,
& Snipes, 1987), nor is the received information used consistently.

Assessing pain behaviour on an intuitive level creates uncertainty and the
possibility of bias in judgments based on these assessments (cf. Tversky & Kahneman,
1974). Both patient and observer characteristics can influence pain perception in
observers. Bond and Pilowsky (1966) observed that nurses tending to terminal cancer
patients were much more likely to administer opiate medication to females than to males,
to the extent that some female patients who neither needed nor asked for narcotic
medication received it nevertheless. Males who complained of pain were routinely ignored, or given low doses of aspirin or other NSAID medications. Hadjistavropoulos, Ross and von Baeyer (1990) found that observers tended to assign lower ratings of pain to attractive patients than they did to unattractive patients, a phenomenon the researchers attribute to the 'halo effect'. The findings of Prkachin, Solomon, Hwang and Mercer (2001) suggest that past experience with chronic pain patients can influence current pain estimations; observers who had a family history of chronic pain rated test patients as experiencing more pain than did observers with no such family history, while health care professionals exhibited the opposite pattern, assigning lower ratings of the pain of test patients than did a group of non-health care professionals. In a study of the ability to discriminate between real and dissembled facial expressions of pain on videotaped lower back pain patients, Poole and Craig (1992) observed that the participants were unable to distinguish between real and dissembled facial expressions of pain, even when warned of the possibility that faked pain expressions were present. Instead, participants so warned had significantly lower judgments of pain for every patient shown, indicating that, in this population, the suggestion of possible malingering reduced the judged severity of the patients’ sufferings indiscriminately. These and many other biases in the pain perception of others exist and operate subliminally, subtly and not-so-subtly influencing one’s perceptions without conscious awareness (Tversky & Kahneman, 1974). By assessing pain behaviour through retrospective techniques (Richards, Nepomuceno, Riles, & Suer, 1982), global ratings (Feuerstein, Greenwald, Gamache, Papciak, & Cook, 1985; Richards, Nepomuceno, Riles, & Suer, 1982) and other unsystematic methods, the probability that observations will be tainted by bias is increased.
It is logical to assume, then, that techniques that standardize the observation of pain behaviour could reduce the effects of patient and observer-related biases, and furthermore could foster more objective perceptions in the health care field. In line with this reasoning, several pain behaviour coding systems have already been developed in attempts to achieve objectivity in pain behaviour assessment. Many of the pain behavior scales developed have similar factor structures (Turk, Wack, & Kerns, 1985), but minor differences in the definition of the behaviours and disagreements regarding the level at which certain behaviours should be combined into definitive factors has made each coding system seemingly unique and, as a whole, difficult to compare with others.

*Previous Pain Behaviour Observational Systems*

One of the first pain behaviour scales was formulated by Richards, Nepomuceno, Riles, and Suer (1982). Named the University of Alabama Pain Behavior Scale (UAB), it is a summary of the general frequency with which patients perform ten types of pain behaviours: (1) vocal complaints (verbal), (2) vocal complaints (non-verbal), (3) downtime, (4) facial grimaces, (5) standing posture, (6) mobility, (7) body language (clutching/rubbing), (8) use of visible support equipment, (9) movement while stationary (fidgeting) and (10) medication use. When used in a clinical setting, each type of behaviour is rated retrospectively by the attending health care professional, usually a nurse, on a three-point scale. The possible scores are 'little', 'moderate', or 'frequent' for each type of behaviour. It can be administered in under five minutes and quickly scored, making it practical to use in such settings. Furthermore, it can be used accurately and consistently by health care professionals, and preliminary validity tests were promising.
However, as delineated above, retrospective and global pain behaviour assessment techniques such as this may increase the likelihood of cognitive biases (e.g. the availability heuristic, see Tversky & Kahneman, 1974) influencing observers’ ratings.

The UAB was later modified for use with outpatient populations, and renamed the “Pain Behavior Scale” or PBS (Feuerstein et al, 1985). The measure was reduced to eight of the original ten types of pain behaviours. Medication use and downtime were removed because they were not immediately observable in an outpatient context. With this variation, coders observe outpatients as they perform a very brief, standardized motor task which consists of standing from sitting in a chair, walking 20 paces in total, standing momentarily, and returning to the initial sitting position. The outpatients are then rated on the eight pain remaining behaviours discussed above, each of which is rated on a three-point scale.

Although one of the main problems of the original UAB, the retrospective manner of coding, was solved in this modification, the globality of the assessments of each of the pain behaviours remained. In addition, new problems were introduced. The primary problem concerns the motor task upon which the behavioural assessment is based. Not only is it far too brief to provide an adequate sampling of pain behaviour, the motor task is relatively insensitive to pain other than intense back pain (one third of the sample in Feuerstein et al’s study had no back pain at all, instead suffering from facial, neck, upper body extremity and headache pain). As the characteristics of back pain and subsequent impairment can vary widely between patients, a rigorous pain assessment comprised of a set of range-of-motion tasks should be employed to fully determine a patient’s unique set
Properties of the BCPBT

of limitations. Simply observing one instance of a patient standing and walking cannot convey the patient’s true score on the underlying construct of pain behaviour.

The second point of concern is the lack of reliability testing performed on the PBS. When a measure is altered in any way, the reliability of its application is altered as well. The authors correlated the obtained scores on the PBS with the hypothesized outcome on the UAB, which was obtained by combining the PBS with estimated scores for the removed behaviours medication use and downtime extrapolated from the interview that accompanied the behavioural assessment. As it was impossible for the PBS scores to decrease when converted to the hypothesized UAB scores, and the range of possible increase was small (2 points), it was not surprising that the correlation between the two sets of scores was high (r = .98). Additionally, conventional forms of reliability assessment, such as inter-rater and test-retest reliabilities (Dworkin & Whitney, 1992; Keefe & Williams, 1992), were not performed, thus leaving the extent to which the measure is reliable unknown.

Cincirpini and Floreen (1983) developed a behavioural observational system for pain that was used in conjunction with a structured interview to assess the pain behaviour of a sample of chronic pain patients, consisting of approximately 60% low back pain patients and 40% upper body and extremity pain patients. Four target behaviours were coded during a simple sequence of motor tasks - rising from a seated position, walking around the room, bending to pick up a small object from the floor, and carrying a chair across the room. These target pain behaviours were (1) touching, (2) grimacing, (3) gesturing and, uniquely, (4) smiles. Gesturing was defined as “gross body movements used to express discomfort” (p. 119), such as limping, wringing of the hands and bracing.
The motor task was videotaped for later coding. Behaviours were coded by counting the total frequency of the occurrence of each type.

For all behaviours, the mean inter-rater reliability was calculated to be over .8, which, on the surface, appears to be adequate (Dworkin & Whitney, 1992). However, the method of calculating this figure seems questionable. The coding dyads compare their coding, and, for each minute of videotape, the smaller number of codes are divided into the larger number of codes to arrive at a percentage. These percentages are averaged across each videotape, and this is the index of inter-rater reliability. Cincirpini and Floreen admitted that there are shortcomings to this method (presumably, the possibility of attaining a perfect inter-rater reliability score of 1 without ever actually achieving agreement on the occurrence of a behaviour), but insisted that the relative infrequency with which the behaviours were displayed and the short coding intervals made it necessary. Unfortunately, this unique method of calculating inter-rater reliability hampers the comparison of this study with others like it that employed more conventional reliability equations, which makes it a less desirable choice when searching for a pain behaviour observational coding system.

Another pain behaviour measure, the Checklist for Interpersonal Pain (CHIP) was created by Vlaeyen, Pernot, Kole-Snijers, Schuerman, Van Eek, and Groenman (1990). The CHIP has six factors: distorted mobility, verbal complaints, nonverbal complaints, nervousness, depression, and day sleeping. While the names of the other categories are basically self-explanatory, it must be noted that nonverbal complaints is a vast category that includes nonverbal utterances such as sighing and groaning, rubbing the painful body part, and shifting position when sitting. The CHIP was designed to be administered by a
health care professional, usually a nurse, over the course of a week. At the end of the observation period, frequency estimations of the six behavioural categories were made on a 5-point Likert scale.

The primary concern with the CHIP is that the method of coding used in it requires the coder to make broad judgments about the frequency of pain behaviour on a Likert scale. In the CHIP validation study conducted by Vlaeyen et al (1990), it is likely that, during the week of observation, the nurses were also performing other tasks in the rehabilitation program, dividing their attention between their nursing duties and observing the patients’ pain behaviours. In this way, the CHIP fails to employ a specific behavioural sampling technique, instead relying on convenience sampling to determine the frequency of pain behaviour. Thus, because of the lack of raw frequency counts of pain behaviour, and the haphazard method used to sample behaviours, coders using the CHIP scale are at a greater risk of producing biased observations than if they used a more rigorous, systematic measure.

Another problem with the CHIP is the level of inference regarding the patient’s state of mind necessary to complete it. The behaviour categories of both nervousness and depression require coders to rate the patients’ cognitive states, which are difficult to define even when using self-report rather than behavioural observation. Vlaeyen et al (1990) fail to use concrete examples of behaviours as they did for their other four factors, instead incorporating such vague terms as “restless and nervous” and “tense” (Vlaeyen et al, 1990, p. 339) into the scale as indicators of “nervousness”, and “appears blue, down” and “appears quiet/withdrawn” (p. 339) as indicators of depression.
Another pain behaviour classification tool, the Audiovisual Taxonomy, was developed by Follick, Ahearn and Aberger (1985). Unlike most other pain behaviour scales, the behavioural categories were empirically derived from 2105 examples of pain behaviour observed in a clinical setting. Using the rational-intuitive approach to classification, these examples were categorized into 80 different types of pain behaviour. From these 80 categories, 16 relatively frequent categories were chosen to serve as the constituents of the Audiovisual Taxonomy. However, the still-large number of categories coupled with the variability of the frequency of certain categories and the wide range in inter-rater reliability coefficients (biserial correlations $r_{bi}$ from .00 to .88) prompted Follick, Ahearn and Aberger to reduce the number of behavioural categories to seven. The final taxa were: guarded movement, which was defined as “slow, cautious movement relative to baseline; nonmethodical or jerky movement” (p. 560); bracing, or using an extremity on the body or another object for support; position shifts; partial movement; grimacing, which consists of biting the lips, gritting the teeth, and pulling back the corners of the mouth; limitation statements, relating to the patient’s inability to perform a task; and sounds.

Discriminant function analysis performed on the data of 57 participants revealed that only four of the categories were useful in discriminating participants with pain from non-pain control participants. These four factors were: partial movement, limitation statements, sounds and position shifts. However, the inter-rater reliability of these categories was only moderate (Dworkin & Whitney, 1992); the mean biserial correlation was approximately .76, which is lower than Dworkin and Whitney’s recommended minimum level of inter-rater reliability of .80. Also, as will be discussed in detail
presently, the operational definition of one of their categories, grimacing, was not consonant with what has since been discovered about the facial expression of pain (Prkachin, 1992). This may have contributed to the lack of discriminant ability of this category in experimental testing.

One of the most frequently used pain behaviour coding systems was developed by Keefe two decades ago (Keefe & Block, 1982; Keefe, Crisson, & Snipes, 1987; Keefe & Hill, 1985). It has been utilized in both experimental and clinical settings, and has produced some impressive psychometric data. Keefe’s taxonomy consists of three types of mutually exclusive behaviour (standing, sitting, and reclining), and five types of concomitant behaviour (guarding, bracing, rubbing, sighing, and grimacing).

The mutually exclusive behaviours are fairly self-explanatory, and Keefe noted that they could be coded extremely reliably, but there are limits to their clinical utility. To observe them, one must observe the patient over a long period of time, or rely on the patient’s own retrospective self report, which defeats the underlying reason for employing a behavioural observation system. However, Keefe found the concomitant behaviours easy to observe and record, and full of psychometric promise. The frequency of the concomitant behaviours showed a strong positive association with the patients’ self-reports of pain and with naïve observers’ ratings of the patients’ pain levels. They also showed evidence of discriminant validity in a mixed population of pain patients and a control group, which consisted of normal and depressed subjects.

There are five categories of pain behaviour in Keefe’s system: guarding, bracing, rubbing, sighing and grimacing. For Keefe, guarding occurs when, during movement, patients act to avoid pain by shifting their weight, interrupting movement, moving in a
stiff or rigid manner, or using support equipment to move. Bracing was defined as using an object for support while stationary, and can be determined by observing whether parts of a patient's body other than the injured area are bearing abnormal amounts of weight. Rubbing occurs if either of a patient's hands makes contact with the afflicted area, or the area immediately surrounding it (Keefe & Block, 1982; Keefe, Crisson, & Snipes, 1987). Sighing is an "obvious exaggerated exhalation of air, usually accompanied by the shoulders first rising and then falling" (Keefe, Crisson, & Snipes, 1987, p. 28). The predominantly visual cues to the occurrence of sighing are owed to the lack of sound on the videotapes first used to develop the scale (Ahles, Coombs, Jensen, Stukel, Maurer, & Keefe, 1990). This may also explain the reason that other forms of sound, such as grunting, groaning and screaming, are not included in Keefe's observational system. Finally, the behavioural category grimacing incorporates many different facial events, including brow furrowing, narrowing of the eye apertures, and a tightening and pulling back of the lips over clenched teeth. Three of the types of behaviour, bracing, rubbing and grimacing, are subject to a duration criterion; to be codable, any occurrence of these behaviours must be held for at least three seconds.

Keefe's observational system was designed to use a time-based behavioural sampling technique. Coders using the system sample for occurrences of the aforementioned behaviours by observing patients for twenty seconds, and subsequently recording their observations for ten seconds. This thirty-second cycle is continuous throughout the entire examination period, which is usually ten minutes long, producing twenty coding opportunities per patient per session.
Research on Keefe’s measure suggests that it is reliable between coders and over time, with inter-rater reliability estimates between 93% and 99% (Keefe & Block, 1982; Keefe, Wilkins, & Cook, 1984). Moreover, it has been demonstrated to be it is valid: it correlates moderately with self-reported pain intensity, and frequency of pain behaviours decreases after treatment for chronic low back pain. Keefe and Block (1982) reported that naïve observers consistently rated those patients who displayed pain behaviour more frequently as experiencing more pain. Keefe, Crisson, Maltbie, Bradley, and Gil (1986) found a strong relationship between Keefe’s measure and another behaviour scale, the Illness Behaviour Questionnaire. Finally, Keefe and Block (1982) demonstrated the pain behaviour measure discriminated between samples of participants: chronic pain patients and pain-free patients with depression, and chronic pain patients and a control sample of normals. The frequency of pain behaviours was greater for the chronic pain patients in both comparisons, illustrating the measure’s discriminant validity.

Criticisms of Keefe’s measure, while few, are important to note. The definitions of the categories do not appear to be conceptually mutually exclusive. Guarding and bracing are different on only the dimension of movement. The same behaviour would be classified as guarding if the patient is moving, and bracing if the patient is stationary (Keefe, Crisson, & Snipes, 1987). Both guarding and bracing pain behaviours serve similar purposes, and their separation into discrete, nominally mutually exclusive categories is arbitrary. In clinical settings, pain patients often engage in jerky, stop-start type movements, particularly during tasks that involve walking, which can reduce discriminability amongst observers. Thus bracing can interrupt examples of guarding
without the occurrence of a qualitative change in pain behaviour other than the cessation of movement.

Also, Keefe's definition of the typical pain facial expression, while consonant with the popular conception of a 'pain face', and widely used (e.g. Cincirpini & Floreen, 1983; Richards, Nepomuceno, Riles, & Suer, 1982), has not been supported by subsequent research. Recent research has isolated the facial events that typify the experience of pain (Craig, Hyde, & Patrick, 1997; Craig, Prkachin, & Grunau, 2001). Unlike Keefe's definition of a grimace, the definitive facial features of pain do not include horizontally stretched lips; in fact, the above mentioned researchers have found that the mouth is only peripherally involved in communicating pain. The main communicating actions appear to be narrowed eyes, a furrowed brow, a wrinkled nose, and a raised upper lip. These actions do not necessarily occur together, but both individually and in combination, they are the facial actions that are most associated with the internal experience of pain (Craig, Prkachin, & Grunau, 2001). This is not to say that horizontally stretched lips are never seen in patients facially expressing pain; a few studies that specifically focused on the individual movements of the face have noted such an association (LeResche, 1982; LeResche, Ehrlich, & Dworkin, 1990). It has also been found to be indicative of pain in neonates (Grunau & Craig, 1987). However, many more studies have observed either a significant relationship between the horizontal lip stretch and exaggerated or faked pain facial expressions (Craig, Prkachin, & Grunau, 2001; Galin & Thorn, 1993; Hadjistavropoulos & Craig, 1994), or a nonsignificant relationship between these two factors in adults (Craig, Hyde, & Patrick, 1997; Patrick, Craig, & Prkachin, 1986; Prkachin, 1992).
Another concern is that Keefe’s measure neglects both verbal and paralinguistic pain behaviours. Turk, Wack and Kerns (1985) identified that a major dimension upon which pain behaviour tend to be grouped is that of “audio-visual”. Pain behaviours fall in varying places along a continuum, of which the anchors are whether the behaviour is heard or seen. Keefe’s taxonomy does not include any behaviours at the auditory end of the continuum. The auditory behavioural category, sighing, is only nominally so; occurrences of sighing are identified by visual means, primarily the rising and falling movement of the shoulders. All other sound-based behaviours were excluded from consideration, but only because the audiovisual equipment used in the studies were unable to reproduce sounds reliably (Ahles et al, 1990). Other researchers have used broad based sound and verbal complaint-type categories to satisfactory effect (Ahles et al, 1990; Follick, Ahearn & Aberger, 1985; Hasenbring, Marienfeld, Kuhlendahl, & Soyka, 1994; Waddell, McCulloch, Kummel, & Venner, 1980).

Finally, although the three-second rule regarding the duration of bracing, rubbing and grimacing behaviours maximizes reliability between observers and over time by creating a situation in which only the broadest behaviours are coded, it ignores valid communications of pain merely because they are short in length. As Ekman (1992) has discussed, fleeting facial behaviours are still behaviours, regardless of their brief appearance. In fact, the “leakage” (Ekman, 1992, p. 25) of facial behaviour can be very informative, as it can indicate the presence of suppressed or masked emotional states. Past research focusing on the facial expression of emotion (Ekman, 1984; Ekman, 1992) and of pain (Prkachin, 1992; Prkachin & Craig, 1994) suggest that these phenomena rarely last longer than a few seconds, and most instances are considerably shorter. These
findings may explain the infrequency of the observed occurrence of Keefe’s grimacing category in clinical situations (Ahles et al., 1990; Keefe & Block, 1982; Keefe, Crisson, & Snipes, 1987), and the low correlations noted between grimacing and the other pain behaviours in studies seeking to establish the validity of Keefe’s system (Keefe & Block, 1982; Keefe, Crisson, Maltbie, Bradley, & Gil, 1986).

The concept of behavioural leakage is not restricted to facial behaviour. It can be generalized to apply to the other two behaviours constrained by the three-second rule as well. Within a comprehensive observational system, the short duration of behavioural displays of bracing and rubbing should not negate their inclusion as instances of codable pain behaviours. To do so would penalize those patients with rapid styles of behavioural display, whether it be idiosyncratic or culturally determined, and would ignore potentially useful sources of information.

The Behavioural Assessment Protocol

Although these observational pain behaviour coding systems are, on the whole, useful, each has its own subset of shortcomings that might be improved. The present study reports on the preparation of a new pain behaviour assessment protocol devised to avoid the pitfalls of other pain behaviour observational systems. The new protocol was designed to meet several theoretical, psychometric and practical criteria.

From a theoretical standpoint, the proposed protocol would have to sample meaningful types of pain behaviour; these types of behaviour would have to be associated with the experience of pain, or at least such an association should be suspected. Moreover, the definitions of these behaviours must be current with the latest research
findings in this area. Psychometrically, the protocol would have to be reliable in its application both between coders and over time. It should also be sensitive to individual and population differences, and valid, in that it truly measures what it purports to measure - the frequency of certain pain behaviours (Dworkin & Whitney, 1992; Traub, 1994; Trochim, 1999).

The practical criteria are both those generally advisable and those imposed by the circumstances particular to the intended use of the protocol in the WCB-BC's Multivariate Prediction of Disability, Low Back study. In general, the protocol would have to be easy to learn, with clear standards to assist in making borderline judgements. Also, administering the protocol should not be cognitively taxing; it must be simple to use, with little or no special equipment needed.

Because of practical concerns about developing an instrument that can be adapted to realistic clinical circumstances, it was considered important that the coders who use it be able to implement the protocol clinically, and in real-time. Although most other observational behaviour protocols use videotape to capture a record of the behaviour for later coding (Cincirpini & Floreen, 1983; Follick, Ahearn, & Aberger, 1985; Keefe & Block, 1982; Keefe & Williams, 1992), the legal issues surrounding the video recording of compensation claimants are too dense and impermeable to allow it. Observers must be able to implement the protocol in situ with ease, which means that it could not require the coder to be active for a sustained period of time. Furthermore, the coding processes had to be noninvasive and inconspicuous as possible, to keep the patients' awareness of the coders' presence to a minimum. Finally, the protocol also had to be structurally parallel to the WCB-BC's standardized physical examination protocol, as the pain behaviour
observations were to be made during these examinations. Adherence to these parameters assured the theoretical, psychometric and practice usefulness of such a measurement entity.

_The British Columbia Pain Behaviour Taxonomy (BCPBT)_

Because of the relatively good psychometric properties of Keefe's behavioural coding system, it provided the foundation for the new pain behaviour observation protocol, called the British Columbia Pain Behaviour Taxonomy that was developed for WCB-BC. Keefe's system was modified both to reflect the criticisms specific to his protocol, and to accord with criteria that assure a sound measurement entity. It was altered in a number of ways.

The behaviour sampling method was fundamentally changed. The administration of the BCPBT was yoked to the WCB-BC standardized physical examination in which the range of motion and flexibility of each claimant is ascertained. The physical examination was structured such that each examination session takes approximately 45 minutes, which is significantly more time than the 10-minute examinations used by Keefe. Forty-five minutes is excessively long to maintain the sustained concentration necessary to code pain behaviours, particularly if the coder must code multiple sessions sequentially. Therefore, the time-based system of sampling behaviour employed by Keefe was replaced by a hybrid event/interval based coding system. The physical examination was divided into 56 discrete, conceptually distinct epochs of varying length (see Appendix F), and observers coded pain behaviours during 36 of these epochs. Also, as Keefe noted (Keefe, Wilkins, & Cook, 1984; Keefe, Crisson, & Snipes, 1987), almost
Properties of the BCPBT

all pain behaviour occurred during times of patient movement. The epochs of the BCPBT were centered on the standardized movement of the patients in order to maximize the possibility of observing pain behaviour. The foremost concern was to minimize the cognitive effort of the coders while simultaneously optimizing the amount of information gathered.

This new kind of hybrid event/interval based coding system is very different from those used in other coding systems. Most other systems utilize coding units that are uniform in duration to control for the differential probability of observing pain behaviour over varying time periods. The heterogeneous duration of the coding unit in the BCPBT does present a possible problem, however, this increased likelihood of a pain behaviour occurrence over time is partially controlled by the fact that BCPBT samples each of the five behaviours only once per epoch. Regardless of whether the epoch lasts twenty seconds or five minutes, the coder will score a maximum of one occurrence of each behaviour per epoch.

This method is conceptually similar to Waddell's (Waddell et al, 1980) approach of assessing the inorganicity of pain behaviour by the presence of at least three out of five possible inorganic signs. The BCPBT merely extends this technique, and makes the results additive and interpretable using interval-scale measurement techniques.

The next modification targeted Keefe's behavioural categories. Four of the behavioural categories were redefined, three were renamed, and one was created. Keefe’s categories of (1) guarding, (2) bracing, (3) rubbing, (4) sighing, and (5) grimacing were revised to form guarding, touching, words, sounds and facial expression. Keefe’s category of guarding has been expanded to include bracing, which was initially conceived
as a stationary form of guarding (Keefe, Crisson, & Snipes, 1987) but analysis of WCB-BC pain patient videotapes revealed that it could occur during movement as well, and flinching, which was also observed in select WCB-BC pain patients early in the protocol development process. A new definition was developed, as follows: "(g)uarding is behaviour that prevents or alleviates pain. It is the most encompassing behavioural category, and its subtypes include stiffness, hesitation, limping, bracing, and flinching. Stiffness is a marked lack of normal flexibility in movement or maintaining a rigid posture. Hesitation is the apparent reluctance to move or an interruption in movement. Limping occurs when the patient fails to apply weight to or favours one leg while walking or walks with an abnormal gait. Bracing is behaviour in which the patient places an abnormal amount of weight upon a part of the body. Bracing cannot be described as limping, but it can occur both when moving or when stationary. It includes using objects or the self as an aid (e.g. leaning on a table for support, using a cane, pushing against oneself to rise from a seated position). Flinching is a sudden withdrawal or spasm of a part of the body." Coders did not have to categorize every guarding pain behaviour they observed into one of the five subtypes, but they did have to be able to identify the components of this pain behaviour.

Keefe’s category of sighing was expanded to include a variety of sounds, including grunts, moans, screams and cries. The new definition for sounds was "voluntary or involuntary production of one of the following: (a) sighing - a puffing or slow exhalation of breath; it can be long or short, and may be accompanied by a shrugging movement of the shoulders or a deflation of the chest; (b) moaning; (c)
grunting; (d) screaming; (e) crying, but it must be an auditory occurrence; do not code sounds if the patient only tears up or sheds tears silently."

Spontaneous verbal behaviour was added to the taxonomy as a new category of pain behaviour. The common conception is that pain behaviour does not include verbal phenomena; however, Fordyce (1976), a strict behaviourist, considered speaking a behaviour rather than an indication of cognition, and thus included verbal complaints in his categorization of pain behaviour, as did Ahles et al (1990). Waddell (Waddell, McCulloch, Kummel, & Venner, 1980) also used verbal complaints in his system to identify nonorganic signs of low back pain; he included "disproportionate verbalizations" (p. 119) as a criteria for identifying patient “overreaction” (p. 119) to nonpainful or minimally painful stimulation. For the BCPBT, words is defined as “as any spontaneous (i.e. not elicited from the examiner’s questions) verbal complaint that relates to the patient’s pain. Examples include: (a) commands - “Stop it.”; (b) information: “That hurts.”, “It aches when I move it.”, “That’s as far as I can go.”; and (c) interjections / expletives: “Ouch.”, “Yikes!”.” It is important to note that the category was clearly defined to include only “spontaneous verbal complaint” because clinical pain assessments commonly require the examiner to inquire directly about pain. Such solicited responses are, by their nature, subject to different determinants than spontaneous behaviour.

The words category of the BCPBT cannot be considered the equivalent to self-report, a measure that historically has an inconsistent relationship with observed pain behaviour (Jensen & Karoly, 1992), and with disability itself (White & Gordon, 1982). The words category only includes spontaneous utterances of pain complaint from the
patient, and the frequency of such behaviour is summed. No content evaluation is performed on the utterances.

Finally, the category of grimacing was completely reformulated according to the findings of Prkachin (1992; Prkachin, 1997; Prkachin & Craig, 1994). The “tightened lips, corners of mouth pulled back and clenched teeth” (Keefe, Crisson, & Snipes, 1987, p.11) aspects of the definition were discarded, and the four characteristic components of the pain facial expression - the sneer, the nose wrinkle, the furrowing of the brow and the narrowing of the eye aperture (the orbital squeeze) - defined the category. The definition of facial expression is much longer and more complex than the definitions for the other four behavioural categories. For a comprehensive definition of facial expression, please refer to Appendix D.

The modifications made to the categories of sighing and grimacing have made the BCPBT extremely sensitive to the occurrence of sounds and facial expressions, and these changes are expected to improve the taxonomy's construct validity. Keefe consistently found that his versions of these two behavioural categories were observed infrequently, and were usually weakly related to overall pain behaviour scores (Keefe & Block, 1982; Keefe, Crisson, & Snipes, 1987; Keefe, Wilkins, & Cook, 1984). These findings may have been due to the relative insensitivity of Keefe's system to the types of behaviour in question.

The next modification made to Keefe's taxonomy was that the 'three-second rule' was eliminated. Any and all observable pain behaviour was coded, regardless of its duration.
The final change made to Keefe's taxonomy was that coding epochs were specifically chosen to be movement-oriented, as Keefe, Wilkins and Cook (1984) found that most of the pain behaviours they observed occurred during physical assessment tasks and periods of transition from one task to another. Thus for the BCPBT, coding was limited to pre-selected transition epochs (moving from one task to another), and to the physical examination tasks themselves. Focusing observers' attention on actions occurring during movement was designed to reduce the total attention time expended per examination while maximizing the possibility of observing pain behaviours.

The Present Studies

The goals of the present studies were fivefold: (1) to evaluate a program designed to train observers in the use of the BCPBT; specifically, to investigate if accuracy in detecting and classifying pain behaviour increases due to exposure to the five-hour BCPBT training procedure; (2) to determine whether persons trained in the BCPBT can maintain an acceptably high standard of accuracy in coding over time; (3) to determine if customized feedback can improve coding accuracy; (4) to assess the inter-rater and test-retest reliabilities of the BCPBT in a clinical setting; and (5) to evaluate the psychometric properties of the BCPBT so that the overall internal consistency and the factor structure of the test can be maximized while maintaining its brevity and ease of administration. These goals were accomplished over a series of six separate studies.
Study 1

*Validation of the Training Protocol for the BCPBT*

A one-day training program to instruct coders (hereafter referred to as "judges") in the administration of the BCPBT was designed and experimentally implemented. This was considered to be a reasonable method of training and timeframe, as the administration of other pain observational scales had been taught in this manner (Keefe & Block, 1982; Keefe, Wilkins, Cook, Crisson, & Muhlbaier, 1986), and the judges achieved adequate levels of mastery.

It was predicted that the training session would increase the trained judges’ abilities to detect and classify pain behaviours significantly over the abilities of judges not so trained. The judges’ sensitivity to pain behaviour was measured by their agreement with the coding of an expert (percent correct), and not by their agreement amongst themselves (inter-rater reliability) for two reasons. Firstly, if there was a systematic error in the training program, or a systematic lack of understanding in all judges about one or more aspects of the BCPBT, then the inter-rater reliabilities would reflect only the reliability of the measure, but not its construct validity. Secondly, the researcher, a co-developer of the coding system, had constructed the training materials to specifically test all types of pain behaviours and thus all behaviours showcased in the training materials were exhaustively coded. This type of training outcome measure was modeled after the final test of the Facial Action Coding System (Ekman & Friesen, 1976; Ekman & Friesen, 1978).
Properties of the BCPBT

Method

Judges

Ten University of Northern British Columbia students, eight females and two males, were recruited by word of mouth and by an announcement posted on an electronic university mailing list. All judges received a $100 honourarium for completing the study. Two females withdrew when they realized that their role was to observe and rate the pain behaviour of others. The mean age of the eight remaining judges was 27.63 years. Judges who were relatively older than the average undergraduate university student were recruited intentionally to make the sample comparable to the population to whom the BCPBT would be taught later, specifically the employees of the WCB-BC, all of whom have some level of university education.

Materials

Training Two written training manuals were produced for use in this research. The first, a brief description of the pain behaviours to be observed and instructions on the method of coding these behaviours, was created to be given to judges before the pretest (Test 1) was administered. The second was more comprehensive and was written to accompany the two training videotapes. These manuals appear in Appendix C and D, respectively.

Two instructional videotapes were created to assist in the training of judges. The main videotape contained material pertaining to four of the behavioural categories of the scale: (1) guarding, (2) touching, (3) words, and (4) sounds. Positive and negative examples of each of the behavioural categories taken from actual video footage of WCB-
BC low back pain examinations were inserted between written descriptions of the behaviours. The WCB-BC supplied this source video footage. Every patient depicted on the tapes had consented to the use of their image for research purposes. After each behavioural category had been defined and demonstrated, audiovisual practice items, in which a patient demonstrated that type of behaviour, were shown. At the end of the main training tape, 13 general practice items were presented. The number and type of pain behaviours depicted in these final items varied, testing the judges on their abilities to detect the full range of pain behaviours.

The content of the second videotape consisted of definitional information on the four facial action units that are pertinent to the coding of facial expression. Example and practice video segments of people making facial expressions that incorporated one or more of the four critical action units needed to judge the presence of pain (Prkachin, 1992). Only the head and shoulders of these people were visible in the video. The facial expressions were either posed by an expert FACS coder (Ekman & Friesen, 1976), or were taken from an earlier study on the facial effects of nociceptive electrical stimulation (Solomon, 1995; Solomon, Prkachin & Farewell, 1997). There were two example video segments of each behaviour, and twelve practice video segments at the end of the videotape.

**Testing**

Four videotaped tests were used to measure the judges’ mastery of BCPBT. These videotapes consisted of 40 video segments, ranging from approximately 20 to 40 seconds in duration, and these segments were randomized for order. The video segments, like the ones from the main training video, were taken from actual video
footage of low back pain examinations at the WCB-BC. None of the test video segments were of patients shown in the training videotapes.

Video segments were used for testing purposes rather than a whole videotaped examination for a number of reasons. Although the use of a whole examination as an accuracy test would have resulted in a greater degree of external validity, employing video segments instead allowed for greater variance in the types of behaviour shown. Patients in the three physical examination videotapes that were used to construct the test videotapes had a tendency to display idiosyncratic constellations of pain behaviours throughout their examinations. Samples from all three examinations were taken to give breadth to the tests. Also, because the standardized physical examination protocol was being developed at the same time as the BCPBT, the protocol was in a constant state of flux. Consequently, it was not possible to get more than one videotaped physical examination that used the protocol. Without at least two videotaped physical examinations that used the same protocol, one for training and one for testing, teaching the judges about interval-specific coding was not feasible. Also, the ideal testing materials, four parallel test videotapes, could not be created due to limited video footage. Therefore, the 40 test video segments were randomized for order to reduce the effect of memory on accuracy. This measure to ensure internal validity seems to have been effective. Several judges reported that they were unaware that any of the 40 video segments were the same across test videotapes until the fourth administration of the test.

**Equipment** All videotapes created for use in this research were in SVHS format, and the information contained within these videotapes was recorded on to them in SVHS mode using VM Studio equipment and software. In addition to the training and
testing materials, a 50 centimeter-screen television and a Sanyo SVHS videotape player were used. They were mounted on a trolley that was approximately 1 meter in height, and positioned not more than 3 meters from any one judge. Judges self-selected their own positions relative to the screen. Following the example of the FACS training system (Ekman & Friesen, 1976), the experimenters supplied the judges with small mirrors to assist them in the identification of the components of a pain facial expression on their own faces.

*Design*

A multiple baseline design was used (Watson & Workman, 1981). Research has shown this type of design to be useful in partialling out the effects of learning from mere exposure in experiments that use repeated measures (Kazdin, 1994; Pancsofar & Bates, 1984). Specifically, a 2x4 mixed design was employed. The between-judges variable (Group) had two levels: Immediate Training and Delayed Training. The test itself was a repeated measures, within-judges variable; it was administered four times over an eight-day period.

*Procedure*

After introductions were made and informed consent obtained (Appendix A), judges were given fifteen minutes to review the brief manual, and allowed to ask any questions they may have had. Then the pretest (Test 1) was administered. In a dimly lit room, judges were shown Test Tape 1, coding their observations on the provided coding forms (see Appendix D).
Following Test 1, judges were quasi-randomly assigned to either the Immediate or the Delayed Training groups. As research has historically found sex differences in emotional decoding ability (Hall, 1979), groups were matched for sex; one male was assigned to each group. The judges in the Delayed Training group were allowed to leave for the bulk of the day. They returned for Test 2 that afternoon. The Immediate Training judges were then given five hours of training in the administration of the BCPBT scale (see Training below).

When all judges were assembled at the end of the first day of the study, Test 2 was administered in precisely the same conditions as was Test 1. All judges were then allowed to leave.

One week later, all judges were administered Test 3 under the same conditions as the first two tests. Then the Immediate Training group was allowed to leave for the day, while the Delayed Training group received training. At 5 p.m., all judges were given Test 4 and then departed.

Training Although the two training sessions followed a curriculum and were identical in the material discussed therein, they were also interactive in that the instructor geared the session to the learning needs of the judges. Thus, the two training sessions were not strictly identical. However, both sessions proceeded as follows. Judges were first given the comprehensive manual so they could read the descriptions of the pain behaviours of note and the conditions under which these behaviours should be coded before they saw the corresponding videotaped information. The main training videotape was shown first, followed by the facial expression training videotape. Every word that appeared on the television was read aloud by the instructor. The videotapes were not
shown uninterrupted; they were stopped, rewound, and even shown in slow motion until the judges felt that they understood the concepts inherent in the taxonomy. During the facial expression training videotape, judges were encouraged to produce the critical expressions themselves. Then the videotape was stopped, and the judges practiced the facial expressions both into a mirror and with a partner who provided feedback. After the informational content of each videotape was finished, judges engaged in practice coding on which they received immediate, specific feedback.

After training in the content of the scale, judges were instructed in the method of coding. After becoming more familiar with the coding forms and the units of coding, a mnemonic strategy was taught to enable the judges to retain their observations of codable pain behaviours over the length of each test clip without the loss of any information. Judges were encouraged to label each finger of their non-dominant hands as one pain behaviour. When a pain behaviour occurred, the finger corresponding to that behaviour was moved in a characteristic manner to indicate a positive instance. At the end of each test clip, judges merely had to refer to their hands to determine the pain behaviours that had occurred. This method was recommended to avoid the problem of the retention of positive instances of pain behaviours in short-term memory while the remainder of the test clip played out. Without rehearsal, and with such potent distracters as the occurrence of other behaviours, the information stored in short-term memory could have been fleeting if it were not for the use of such a mnemonic device.

Finally, judges practiced their newly acquired behavioural observation skills by coding pain behaviours from the physical examination videotape A, verbally identifying the type of behaviour displayed, and its onset and offset.
Data scoring

For four of the five behavioural categories (1. guarding, 2. touching, 3. words, and 4. sounds), judges made binary, “present/absent” decisions, resulting in scores of either “1” or “0”, respectively. For the behavioural category of facial expression, however, the BCPBT scale required that judges score behaviour on a three-point intensity scale. Judges rated facial expression of pain as absent, as low/medium in intensity, or as high in intensity, which would correspond to scores of “0”, “1” or “2”, respectively.

Errors were calculated by comparing the coding of the judges against a master coding protocol. Every deviation from this master protocol resulted in a decrement in the overall accuracy score of each judge. All errors were of equivalent weight, in that any error of a magnitude of more than one point (i.e., a two-point error in the behavioural category of facial expression) was the same as an error in any other behavioural category.

The index of accuracy upon which all subsequent data analysis was based was the “percent effective agreement” statistic as outlined by Hartmann (1977) and Bakeman and Gottman (1997). Percent agreement was calculated in the following manner:

\[
PA = \frac{AGREEMENTS}{AGREEMENTS + DISAGREEMENTS} \times 100
\]

This statistic is consistent with the methodology of earlier research (Keefe & Block, 1982; Keefe & Hill, 1985; Keefe, Crisson, & Snipes, 1987), thus facilitating the direct comparison of agreement levels, and ultimately, of the psychometric properties of the measure.
However, Cohen (1960; Cohen, 1968; Cohen, 1988) has stated that percent agreement, by itself, is inadequate to reflect the actual level of accuracy in categorical coding systems, which is a position supported by Rosenthal (1982). As with any observational measure of behaviour, there will always be a level of chance agreement between judges. To control for chance agreement, kappa (κ) statistics (Bakeman & Gottman, 1997; Cohen, 1960) were computed, and are reported in conjunction with every percent agreement calculated.

It was determined that the baseline frequencies of positive and negative instances of pain behaviours in the test bank of video segments were such that the kappa statistics were not unreasonably limited (Turk & Rudy, 1992), and the calculation of Yule statistics (Turk & Rudy, 1992) in their stead was deemed unnecessary.

Data analysis

The data were analyzed using the SPSS 6.1.1 for Students software package for the Macintosh.

Results

Figure 1 and Table 1 display the mean percent agreement over each of the four experimental phases by group. The steepest slopes of the lines representing the data occur for both groups between the test immediately before training in the BCPBT, and the test given immediately after. Training was given to the Immediate Training group between Tests 1 and 2, and to the Delayed Training group between Tests 3 and 4. The solid vertical line at Test 2 indicates the first test given after the Immediate Training
Figure 1. Mean Percent Agreement Over Test by Group
Table 1. Comparison of Pre-Training and Post-Training Agreement Scores for Immediate and Delayed Training Groups (N = 8)

<table>
<thead>
<tr>
<th>Test</th>
<th>Immediate Training</th>
<th>Delayed Training</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test 1</td>
<td>74.75%</td>
<td>76.38%</td>
</tr>
<tr>
<td>Test 2</td>
<td>82.00%</td>
<td>76.25%</td>
</tr>
<tr>
<td>Difference</td>
<td>7.25%</td>
<td>-1.25% (1)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test</th>
<th>Immediate Training</th>
<th>Delayed Training</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test 3</td>
<td>84.00%</td>
<td>77.63%</td>
</tr>
<tr>
<td>Test 4</td>
<td>85.63%</td>
<td>82.25%</td>
</tr>
<tr>
<td>Difference</td>
<td>1.63%</td>
<td>4.63% (2)</td>
</tr>
</tbody>
</table>

(1) - p = .038 significance
(2) - p = .161 failed to reach significance
Properties of the BCPBT

The omnibus F-test for the 2x4 repeated measures ANOVA showed that there was a main effect for Test, $F(3, 18) = 21.78$, $p < .001$, partial $\eta^2 = .784$. This effect was large (Cohen, 1992). The mean percent agreement scores for Test differed significantly over administration, across Group.

A main effect was not hypothesized for Group. It was expected that both groups would reach equivalent levels of coding ability after both had received training (Test 4). As predicted, there was no main effect for Group, $F(1, 6) = 1.41$, which was not statistically significant.

The hypothesized interaction between Test and Group was observed, $F(3, 18) = 5.90$, $p < .005$, partial $\eta^2 = .496$. This interaction also had a large effect magnitude (Tabachnick & Fidell, 1996). To elucidate if the significant differences were between the hypothesized scores, two pairwise comparisons were employed.

Two \textit{a priori} independent samples t-tests comparing the Immediate Training and the Delayed Training groups were computed using the difference between scores on Test 1 and 2, and scores on Test 3 and 4, which can be seen in Table 1. Although the use of difference scores has been criticized in the past on the basis of its inflated error component (Spector, 1981), both Maxwell and Howard (1981) and Kenny (1975) support the use of such scores in research. A Bonferroni correction was applied to the alpha
levels of the t-tests to control for familywise error (Shutz & Gessaroli, 1987). Because the predicted outcome of the t-tests was directional, the entire alpha was applied to one tail of the distribution, and so the adjusted critical value of alpha for the a priori pairwise comparisons was $\alpha \leq .05$.

The main hypothesis, that BCPBT training would produce greater test accuracy, predicted that the difference between Test 1 and Test 2 would be significantly larger for the Immediate Training group than for the Delayed Training group, due to the intervening training session given to the former. This hypothesis was, in direction, supported by the data ($M_{\text{IMM}} = 7.25, M_{\text{DEL}} = -.125, t(6) = 2.64, p = .038$), which met statistical significance. The corresponding effect size was large ($r_{pb} = .733$) (Cohen, 1992).

The second hypothesis, that the Delayed Training group would have a larger difference between their scores on Test 3 and Test 4 than would the Immediate Training group, due to the intervening training session given to the former, was not confirmed by the data ($M_{\text{IMM}} = 1.63, M_{\text{DEL}} = 4.63, t(6) = -1.6, p = .161$). Even without the Bonferroni correction, this statistic would not have reached significance. This may be due to practice effects, as the Delayed Training group did show small but steady increments in agreement over the three tests that preceded their training. However, the calculation of the effect size ($r_{pb} = .713$), large by Cohen’s criteria, indicates that, again, given a slightly larger sample size producing similar data, this comparison would have been statistically significant.

Although only one of the a priori pairwise comparisons was statistically significant, the judges did reach an acceptable level of coding agreement. After all judges had been trained (Test 4), the mean level of coding agreement was 83.9%, above the
minimum criterion of 80% set out by Dworkin and Whitney (1992). Only two judges failed to score above 80% agreement on Test 4; both of these judges were 78% accurate.
Study 2

Test-retest Reliability of the UNBC Judges

Once an adequate level of coding accuracy was achieved for the BCPBT, its test-retest reliability needed to be assessed. Using the judges from UNBC, one additional accuracy test was administered to aid in the determination of the reliability of the judges' coding over time. It was hypothesized that all judges would be able to retain their level of accuracy over the seven day period between tests, and that the mean of all post-training tests would meet or exceed Keefe's standard of 90% agreement.

Method

Judges

The judges from Study 1 were also the judges for Study 2.

Materials

All the materials used in Study 1, except for the brief training manual, were used in Study 2. An additional testing videotape, Test Videotape 5, was used as well. This videotape contained all of the same video segments as Tests 1 through 4, but the order of presentation of these segments was unlike those of earlier tests. In this way, the equivalence of the testing materials across testing conditions was assured, and there was no measurement error added due to a difference in materials (Carmines & Zeller, 1979).
Design and procedure

Study 2 was simply an extension of Study 1. One week after the Delayed Training group were trained and Tests 3 and 4 were administered, an additional test was given. This test, Test 5, was solely for the purpose of assessing the test-retest reliability of the BCPBT Scale for both groups. As all post-training test scores reflect the stability of the measure over multiple administrations, Test 5 was administered specifically to tap the measure's stability in the Delayed Training group, and to provide an estimation of the test-retest reliability over a longer period of time, as the Immediate Training group had been trained two weeks earlier. For the Immediate Training group, a wider range of scores, those from Tests 2, 3, 4, and 5, were used to calculate a more reliable test-retest reliability coefficient.

Data scoring

The data were scored as they were in Study 1.

Results

The mean agreement for Test 5 was 82.6% (with a range from 72% to 87% agreement), which was slightly lower than the 84% mean of Test 4, and substantially lower than Keefe's criterion of 90%.

All post-training percent agreement scores were used to calculate the mean Pearson's product-moment correlation, which can serve as the coefficient of reliability in test-retest conditions (Carmines & Zeller, 1979; Ghiselli, Campbell, & Zedeck, 1981). The correlations between Tests 2, 3, and 4 for the Immediate Training group, and the
correlation between Tests 4 and 5 for all judges were standardized through the use of the Fisher Z transformation (Glass & Hopkins, 1996). These standardized amounts were then averaged, and transformed back into the mean Pearson's product moment correlation. This test-retest reliability coefficient was equal to .887 ($\kappa = .66$, $p < .001$), indicating that the BCPBT was sufficiently reliable in its application (Dworkin & Whitney, 1992). This correlation could be considered "high" (Williams, 1988, p. 240).
Study 3

The Effect of Feedback on BCPBT Accuracy

Other researchers of pain behaviour have reported that their judges were able to reach 90% coding agreement levels (Keefe & Block, 1982; Keefe, Crisson, & Snipes, 1987). However, upon closer inspection of these two studies, it is clear that these high levels of agreement were not achieved in one training session. Judges in Keefe’s studies needed between 4 to 8 hours of training that was self-administered, and also received regular feedback on their coding performance. This approach has been used successfully in other coding methodologies, most notably Ekman and Friesen’s Facial Action Coding System training protocol (1978). Therefore, it is possible that the judges’ performance in Studies 1 and 2, which was below the level of agreement expected, may have been due to insufficient feedback on their coding strategies. Thus the judges who learned the BCPBT in the two previous studies were recruited again to see if a review of the administration of the scale coupled with feedback on their general coding performance would result in agreement scores that matched or surpassed the 90% agreement benchmark. It was believed that feedback, in a general form, would increase the judges’ awareness of their erroneous coding strategies and assist the judges in recalibrating their knowledge and application of the BCPBT. General feedback was necessary because specific feedback on each testing item would invariably result in overwhelming carry-over effects for the following test. Thus the main hypothesis of this experiment was that an intensive review of the BCPBT coupled with general feedback about the judges’ personal coding strategies would increase their sensitivity to pain behaviour, increasing their mean agreement to a criterion of 90% agreement (Keefe, Crisson, & Snipes, 1987).
Method

Judges

Five of the eight original judges were available to take part in Study 3, which occurred approximately five months after the end of Study 1. Four judges were female, and one was male. The mean age of the returning judges was 27.4. They each received a $50 honourarium for completing the study.

Materials

All the training materials from Study 1, save for the brief training manual, were used again in Study 3. Two of the test videotapes from Study 1 served as the test materials for Study 3. Test videotape 4 was administered as the pretest, and Test videotape 5 was administered as the post-retraining measure.

Design and procedure

After completing an informed consent document (see Appendix B), judges were administered an accuracy test immediately after they had completed their informed consent paperwork, without any review of the BCPBT scale. Their tests then were scored, and general feedback statistics were composed for each judge. General feedback consisted of a summary of the number and type of errors made for each behavioural category by each judge. The types of errors were “misses”, in which judges did not score behaviours that occurred, and “false alarms”, in which judges scored behaviours that had not occurred. Changes in personal coding strategies were recommended depending upon the pattern of errors each judge made. If judges had a preponderance of omissions or
"misses", they were instructed to be more "liberal" in their coding; if they tended to code more "false alarms", they were told to be more "conservative" in their judgments. The terms "liberal" and "conservative" were defined and described to the judges, and a personalized breakdown of the types of errors each judge made was presented to that judge during the private feedback session with the trainer.

Due to previous commitments, two judges had to be trained and tested in individual sessions. The remaining three judges were exposed to the training information and general feedback in a group setting. The same trainer was used for all three session, and there were no differences in scores between judges trained in a group and those trained individually.

Data scoring

Data scoring was identical to that of Study 1. Percent agreement figures were calculated between the judges' coding and a master coding protocol, and these figures were used when calculating the inferential statistic, a paired-samples t-test. Kappa statistics were also calculated to describe the level of chance agreement between the master protocol and the judges.

Results

Although the judges' scores on the first follow-up test were slightly higher than their scores on the very first test they ever received, before anything but a brief 15-minute training procedure had been administered, this difference was not significant statistically nor practically ($M_{test1} = 75.6\%, M_{followup1} = 76.1\%, t(4) = -0.43, p = .69, r_{pb} = .23$),
suggesting that the judges had lost all their previously gained coding abilities over the intervening five month period.

The difference between the first and the second follow-up tests was significant ($M_{\text{followup1}} = 76.1$, $M_{\text{followup2}} = 83.6$, $t(4) = -3.89$, $p = .018$, $r_{pb} = .69$), which indicated that the agreement of the judges was improved by the instructional review and the general coding feedback. However, the comparison of interest, whether the judges could meet the 90% accuracy criterion, was not significant. Although the judges’ mean agreement of 83.6% ($\kappa = .67$) was still significantly larger than Dworkin and Whitney’s recommended minimum agreement level of 80% (1992), ($t(4) = 4.06$, $p = .015$), the mean scoring agreement for the second follow-up test was substantially below the 90% benchmark of agreement reported by Keefe (Keefe, Crisson, & Snipes, 1987).
Study 4

Part I: Effects of Training on WCB Judges

Although the administration of the BCPBT in an experimental situation resulted in adequate coding accuracy, it was necessary to examine if the BCPBT could be successfully utilized in a clinical context. The clinical judges were first trained in a similar manner to those in Study 1; they learned and practiced the BCPBT using videotapes before moving on to administer the BCPBT using real WCB-BC patients.

Method

Judges

Five female employees of the WCB-BC completed the training. They were all between 27 and 35 years of age. All had at least some post-secondary education and were proficient in the English language.

Materials

All the videotape materials from Study 1 were used in Study 4. The written manuals, however, were altered. The brief manual was not used. The comprehensive manual was used in its entirety, but passages describing the nature of coding epochs, and instructions on how to parse the standardized physical examination into these epochs were added. These added entries are included in Appendix E.

A 24-inch screen television and a SVHS videotape player unit were used to show the training and test videotapes. The television was mounted upon a 4 foot trolley; all
judges were no further than ten feet away from the television screen, and self-selected their seating. All judges had a full frontal view of the monitor.

*Design and procedure*

The WCB-BC judges were trained in the same way as the UNBC judges in Study 1 were. The only difference was that the WCB-BC judges were all trained together simultaneously, and additional information on the parsing of the standardized physical examination into intervals was given (see Appendix E). Therefore the training seminar lasted approximately 1.5 hours longer than the one administered to the UNBC judges, for a total of 6.5 hours.

The judges were given an accuracy test (Test video 4) immediately after training, and were then tested again two days later using an alternate form of the test (Test video 5).

*Data scoring*

All data were scored in the manner described in Study 1.

*Results*

The mean agreement of the judges over both tests was 80.9% ($\kappa = .62$) which was statistically equivalent to the minimum criterion of 80% suggested by Dworkin and Whitney (1992) ($t(4) = .73$, $p = .508$). However, the judges’ mean agreement was lower than the 90% agreement criterion advocated by Keefe (Keefe & Block, 1982; Keefe, Crisson, & Snipes, 1987).
Part II: Using the BCPBT in a Clinical Setting

Judges working at the WCB-BC used the BCPBT Scale to rate the pain behaviours of a small practice sample of WCB-BC claimants undergoing a standardized physical examination, thus demonstrating the generalizability of their knowledge set from the highly controlled training setting to a less controlled, real life clinical setting. If the BCPBT knowledge set can be thus transferred, which will be evident if the judges’ agree with each other to an acceptable level, then the BCPBT can be utilized in the WCB-BC Multivariate Prediction of Disability, Low Back project.

Method

Judges

All but one of the same judges that completed Part I also participated for Part II. Judge 3 left the employ of the WCB-BC before this phase of the study could be completed. The judges worked in pairs, both simultaneously coding the examination of the claimant, which provided agreement data between each pair.

Materials

Judges used a coding form reflective of the standardized physical examination protocol (see Appendix F for the form).
Design and procedure

Each pair of judges observed a live standardized physical examination from a vantage point inside the examination room, rating the observed claimant on the number and type of pain behaviours displayed. Although the examiner was not a constant throughout all sessions, the six examiners who took turns administering the examinations were experienced physicians or physiotherapists who were well trained in the administration of the standardized physical examinations, and the physical examinations all followed the standardized protocol to the best of the examiners’ abilities.

Six WCB-BC claimants were observed so that each judge was paired with every other judge. The judges remained as unobtrusive as possible during the sessions. They were instructed to avoid eye contact with the claimants, and to remain silent throughout the examination. They were also instructed to restrain their facial expressions, and any other sign of their internal states so as not to influence the other judge, the claimant, or the examiner in their evaluation of the exam.

The judges coded the claimants' pain behaviours only during the 36 epochs as outlined in Table 2. The beginnings and endings of each epoch are described in Appendix D. The other 21 epochs listed in Table 2 were not coded for theoretical or pragmatic reasons, and, due to the instructions of the examination protocol, two of the codable epochs (33. axial rotation and 34. simulated rotation) could only be partially coded.
Table 2. Epochs For the WCB-BC Standardized Physical Examination

<table>
<thead>
<tr>
<th>epoch designation</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. to sitting 1 (*)</td>
<td>30. lateral extension right 4</td>
</tr>
<tr>
<td>2. introduction</td>
<td>31. lateral extension left 5 (*)</td>
</tr>
<tr>
<td>3. to scale (*)</td>
<td>32. lateral extension right 5 (*)</td>
</tr>
<tr>
<td>4. weight/height</td>
<td>33. axial compression</td>
</tr>
<tr>
<td>5. to landmarks (*)</td>
<td>34. simulated trunk rotation</td>
</tr>
<tr>
<td>6. landmarks</td>
<td>35. to kneeling</td>
</tr>
<tr>
<td>7. lordosis</td>
<td>36. ankle reflexes</td>
</tr>
<tr>
<td>8. heel raise</td>
<td>37. to sitting 2</td>
</tr>
<tr>
<td>9. forward/back</td>
<td>38. knee reflexes (*)</td>
</tr>
<tr>
<td>10. rotation</td>
<td>39. knee extension 1</td>
</tr>
<tr>
<td>11. side-to-side</td>
<td>40. knee extension 2</td>
</tr>
<tr>
<td>12. stand for measurement</td>
<td>41. muscle strength 1</td>
</tr>
<tr>
<td>13. lumbar extension 1</td>
<td>42. muscle strength 2</td>
</tr>
<tr>
<td>14. lumbar flexion 1</td>
<td>43. to supine</td>
</tr>
<tr>
<td>15. lumbar extension 2</td>
<td>44. ankle dorsiflexion</td>
</tr>
<tr>
<td>16. lumbar flexion 2</td>
<td>45. toe extensor</td>
</tr>
<tr>
<td>17. lumbar extension 3</td>
<td>46. thigh &amp; calf muscle bulk</td>
</tr>
<tr>
<td>18. lumbar flexion 3</td>
<td>47. sensation</td>
</tr>
<tr>
<td>19. lumbar extension 4</td>
<td>48. passive straight leg raise right</td>
</tr>
<tr>
<td>20. lumbar flexion 4</td>
<td>49. passive straight leg raise left</td>
</tr>
<tr>
<td>21. lumbar extension 5 (*)</td>
<td>50. to prone</td>
</tr>
<tr>
<td>22. lumbar flexion 5 (*)</td>
<td>51. palpation</td>
</tr>
<tr>
<td>23. lateral extension left 1</td>
<td>52. McKenzie push-up</td>
</tr>
<tr>
<td>24. lateral extension right 1</td>
<td>53. prone active extension (*)</td>
</tr>
<tr>
<td>25. lateral extension left 2</td>
<td>54. to supine</td>
</tr>
<tr>
<td>26. lateral extension right 2</td>
<td>55. active situp</td>
</tr>
<tr>
<td>27. lateral extension left 3</td>
<td>56. bilateral active straight leg raise</td>
</tr>
<tr>
<td>28. lateral extension right 3</td>
<td>57. to standing</td>
</tr>
<tr>
<td>29. lateral extension left 4</td>
<td></td>
</tr>
</tbody>
</table>

**bold** - coded epochs

(*) - psychometrically poor epochs
Data scoring

As real claimants were used in this study, and there could be no master coding protocol with which to compare the judges’ ratings, percent agreement scores were calculated between each pair of judges. This percent agreement was a statistical summary of the confluence of the two judges’ scoring rather than the accuracy of any one judge. Although percent agreement is not a recommended statistical method for summarizing reliability between judges (Carmines & Zeller, 1979; Rosenthal, 1982), percent agreement was utilized to make this research comparable to previous studies. Mean inter-rater correlations and kappa statistics were also computed for a more conservative estimate of inter-rater reliability.

Results

The mean agreement of the judges was 87.7%. This level of agreement was well over the 80% minimum criterion for reliability recommended by Dworkin and Whitney (1992). It was not significantly different from a 90% level of agreement (t(5) = -.68, p = .525). The mean inter-rater reliability coefficient between the judges, calculated using a Fisher Z transformation, was r = .774. The mean kappa of the judges was calculated to be .693, which numerically exceeds the minimum kappa criterion outlined by Dworkin and Whitney (1992) of .60, but is not significantly different from it (t(5) = 1.38, p = .226).
Study 5

Clinical Test-Retest Reliability of the BCPBT

The test-retest reliability of the BCPBT and its components was determined when judges observed each of 67 WCB-BC claimants twice over a three-day period. Although pain behaviour is subject to variation in frequency from day to day due to the cyclical nature of chronic low-back-pain, for a measure of this phenomenon to be useful, it must demonstrate some consistency from one administration to the next (Dworkin & Whitney, 1992). The goal of this study was to examine the test-retest reliability of the BCPBT on a clinical population of low-back-pain patients over time.

Method

Participants

Sixty-seven WCB-BC claimants were recruited to be observed by BCPBT judges during a standardized physical examination as part of the larger Multivariate Prediction of Disability, Low Back project for the WCB-BC. All participants were between the ages of 18 and 60 years old (mean age = 41), and had an open claim with the WCB (i.e. were receiving or had recently applied to receive compensation for their injuries). All participants could read and speak English. There were 50 male and 17 female participants in the sample. Pregnant women were excluded from the sample.

Participants were contacted by a WCB-BC employee if they met the above criteria, and were asked to participate over the telephone. If they agreed to participate, two appointments at the WCB-BC clinic were arranged. Participants were required to give informed consent before the examination took place.
**Design and procedure**

The participating claimants were observed during two WCB-BC standardized physical examinations for low back pain. These examinations were administered by a licensed physician or physiotherapist. All participants were observed twice within a maximum three day period by one of the four judges trained in Study 4. The attendant judge sat in the corner of the examination room as unobtrusively as possible, making no eye contact with the participants. The judges were further instructed to keep silent and their faces blank from emotional expression.

The judges coded the participants' pain behaviours only during the intervals as outlined in Table 2. The beginnings and endings of each interval are described in Appendix D. Table 2 notes the epochs that were not coded for pragmatic reasons, either to reduce the cognitive load on judges, or because the examiners' actions hamper the judges' ability to observe fully. As in Study 4B, in two epochs (33. axial rotation and 34. simulated rotation), coding was curtailed by the standardized examination instructions; the examiner explicitly stated that the patients should touch their own hip area for simulated rotation (which precludes the coding of touching), and in both simulated rotation and axial rotation, in which the examiner asked the participants to verbally indicate their pain experiences, clearly precluding the coding of words.

**Data scoring**

All data were left in their original form, coded either “0”, “1”, or “2”. Totals for each type of pain behaviour, as well as for total pain behaviour, were calculated for each examination. All scores of "cannot code", which indicated that the judge was prevented
from coding that epoch due to some situational factor, and "did not code", which indicated that the scheduled action for that epoch did not occur or the epoch was not administered, were removed from the data set.

Results

Initially, the mean and standard deviation of each of the pain behaviour categories were calculated (see Table 3). The pain behaviours varied considerably in their frequencies. Facial expression, sounds and guarding all had relatively high means; words and touching both had relatively low means, and standard deviations that exceeded those means.

The test-retest reliabilities between the observed behaviours were calculated using a Pearson product-moment correlation between the first score (Time 1) and the second score (Time 2) for each behavioural category, and for the total pain behaviour score. All the correlations were statistically significant, and are listed in Table 4.

The three highest correlations were total behaviour ($r = .595$), guarding ($r = .618$) and facial expression ($r = .713$), and although Cohen (1988) indicates that correlations of this magnitude are very large, pragmatically speaking, for a system of behavioural assessment, they are really only moderate in size (Dworkin & Whitney, 1992; Williams, 1988). Touching and words displayed the lowest correlations ($r = .358$ and .339, respectively).
Table 3. The Mean Frequency of Pain Behaviours in the Test-Retest Group of WCB Claimants (N=67)

<table>
<thead>
<tr>
<th>behavioural category</th>
<th>Time 1</th>
<th></th>
<th>Time 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>SD</td>
<td>mean</td>
<td>SD</td>
</tr>
<tr>
<td>guarding</td>
<td>11.06</td>
<td>7.73</td>
<td>10.42</td>
<td>7.20</td>
</tr>
<tr>
<td>touching</td>
<td>1.40</td>
<td>1.78</td>
<td>1.77</td>
<td>2.33</td>
</tr>
<tr>
<td>words</td>
<td>4.93</td>
<td>4.31</td>
<td>3.06</td>
<td>3.24</td>
</tr>
<tr>
<td>sounds</td>
<td>12.12</td>
<td>7.78</td>
<td>10.28</td>
<td>7.62</td>
</tr>
<tr>
<td>facial expression</td>
<td>18.42</td>
<td>10.77</td>
<td>16.42</td>
<td>10.28</td>
</tr>
<tr>
<td>total behaviour</td>
<td>41.95</td>
<td>24.86</td>
<td>41.96</td>
<td>22.46</td>
</tr>
</tbody>
</table>
Table 4. The Test-Retest Correlations of the Pain Behaviours of the BCPBT Across Claimants Over a Three-Day Period (N=67)

<table>
<thead>
<tr>
<th>behavioural category</th>
<th>r</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>guarding</td>
<td>.618</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>touching</td>
<td>.358</td>
<td>.003</td>
</tr>
<tr>
<td>words</td>
<td>.339</td>
<td>.005</td>
</tr>
<tr>
<td>sounds</td>
<td>.436</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>facial expression</td>
<td>.713</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>total taxonomy</td>
<td>.595</td>
<td>&lt; .001</td>
</tr>
</tbody>
</table>
Study 6

Part I: The Internal Consistency of the BCPBT

The collection of BCPBT scores for the moderately sized sample of 120 WCB-BC claimants was an opportunity to examine and to refine the psychometric characteristics of the measure. Internal consistency is an essential aspect of any good observational system (Dworkin & Whitney, 1992), and the reliability gained from the length of a measure should not supercede the reliability gained by the quality of the items contained therein. A good measure should be brief as possible, but without sacrificing the integrity of the interplay of its items. This study focused on the contribution of each item to the overall reliability of the BCPBT, and was used to winnow the functional items from the nonfunctional items to streamline the measure.

Method

Participants

The participants in this study were also the judges in Study 5. One hundred and twenty WCB-BC claimants, including the sixty-seven participants from Study 5, were recruited to be observed by BCPBT judges during a standardized physical examination as part of the larger Multivariate Prediction of Disability, Low Back project for the WCB-BC. All participants were between the ages of 18 and 60 years old (mean age = 40), and had an open claim with the WCB (i.e. were receiving or had recently applied to receive compensation for their injuries). All participants could read and speak English. There were 88 male and 32 female participants in the sample. Pregnant women were excluded from the sample.
Design and procedure

The data produced for Study 5 was analyzed for internal consistency using Cronbach’s alpha statistics produced by SPSS 6.1.1 for Students software package for the Macintosh.

Data scoring

All data were left in their original form, coded either “0”, “1”, or “2”. Data were not transformed. Every behavioural category in every epoch was considered to be an item, resulting in a total of 177 items. For behavioural category-specific analyses, guarding, sounds and facial expression had 36 items per participants; due to aforementioned administration issues, touching and words had fewer items, 35 and 34, respectively.

Results

The data were first analyzed to indicate the mean frequencies of each category of behaviour, the outcome of which is displayed in Table 5. As in Study 5, the most frequent behaviours were facial expression, sounds and guarding, while the least frequent were words and touching.

The data were then entered into a total Cronbach’s alpha calculation using each datapoint as an item on a test. As there were 36 codable epochs with five behavioural categories in each and three coding exceptions, each participant’s score contained 177 possible “items” or variables. (It was possible to have fewer items per participant if the administrator of the physical exam omitted any epoch, or if the observing judge could not
fully view the participant during a given epoch.) The data were not summarized or
transformed; each behavioural category of each coding epoch was a separate variable.

As is shown in Table 8, the Cronbach’s alpha for the total score of all the epochs
was .955, which is very high according to Dworkin and Whitney (1992). However, further
inspection of the individual item-test correlations showed that 43 item-total correlations
were below .30. The majority of these low-loading items came from the behavioural
category of touching.

To investigate this result, Cronbach’s alpha statistics were calculated for each
behavioural category individually, across epochs. These alphas can be seen in Table 6.
Overall, the alpha statistics were well in the acceptable range, with a Fisher Z transformed
mean alpha of .873 across all the categories (Dworkin & Whitney, 1992). However, as was
suggested by the item-total correlations of the alpha of the total scores, touching had an
alpha that was substantially lower than the other categories’ (alpha = .68). Without
touching, the Fisher Z transformed mean alpha of the remainder of the behavioural
categories increased to .901.
Table 5. The Mean Frequency of Pain Behaviours in the One-Time Examination of WCB Claimants (N=120)

<table>
<thead>
<tr>
<th>behavioural category</th>
<th>mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>guarding</td>
<td>8.93</td>
<td>7.43</td>
</tr>
<tr>
<td>touching</td>
<td>1.33</td>
<td>1.80</td>
</tr>
<tr>
<td>words</td>
<td>4.57</td>
<td>4.61</td>
</tr>
<tr>
<td>sounds</td>
<td>11.02</td>
<td>7.93</td>
</tr>
<tr>
<td>facial expression</td>
<td>16.38</td>
<td>11.00</td>
</tr>
<tr>
<td>total behaviour</td>
<td>42.22</td>
<td>24.86</td>
</tr>
</tbody>
</table>
Table 6. The Change in Cronbach’s Alpha Statistics for the Five Behavioural Categories, Before and After the Removal of Low-Loading Coding Epochs (N = 120)

<table>
<thead>
<tr>
<th>behavioural category</th>
<th>Old α (n)</th>
<th>New α (n)</th>
<th>change</th>
</tr>
</thead>
<tbody>
<tr>
<td>guarding</td>
<td>.921 (36)</td>
<td>.925 (27)</td>
<td>.004</td>
</tr>
<tr>
<td>touching</td>
<td>.684 (35)</td>
<td>.616 (26)</td>
<td>-.062</td>
</tr>
<tr>
<td>words</td>
<td>.837 (34)</td>
<td>.858 (25)</td>
<td>.021</td>
</tr>
<tr>
<td>sounds</td>
<td>.925 (36)</td>
<td>.930 (27)</td>
<td>.005</td>
</tr>
<tr>
<td>facial expression</td>
<td>.899 (36)</td>
<td>.913 (27)</td>
<td>.014</td>
</tr>
<tr>
<td>total taxonomy</td>
<td>.955 (177)</td>
<td>.954 (132)</td>
<td>-.001</td>
</tr>
</tbody>
</table>
It was noted that several of the epochs had consistently low item-total correlations (below or equal to an alpha of .30) across the different behavioural categories (see Table 7), which indicates the possibility of problems at both the practical and the psychometric levels. For example, epoch 38, Knee Reflexes, loaded poorly on four of the five behavioural categories (touching, words, sounds and facial expression), suggesting that this epoch does not significantly add to the internal consistency of the BCPBT, and may not be a productive use of the judges' time. In all, nine of the epochs loaded poorly on three or more of the behavioural categories. These epochs are as follows: 1. To Sitting, 3. To Scale, 5. To Landmarks, 21. Lumbar Extension 5, 22. Lumbar Flexion 5, 31. Lateral Flexion Left 5, 32. Lateral Flexion Right 5, 38. Knee reflexes, and 53. Prone Active Extension. Epochs 21, 22, 31 and 32 had low item-total correlations because these epochs were not administered in any of the patients' examinations. Other epochs were merely infrequently administered, resulting in reduced means and high variability in frequency. Epoch 3 was a borderline case; its item-total correlations were acceptable for words and facial expression, and was borderline on touching. It was deemed to be minimally informative, and was included in the group of nine poorly loading epochs.

Cronbach’s alpha statistics were calculated again after the nine poorly loading epochs were removed from the dataset. The results are also listed in Table 6. The alpha of the total BCPBT score decreased almost unnoticeably, from .955 to .954. Such a miniscule decrease after omitting 45 items (approximately 25% of the total scale) from the calculation indicates that these omitted epochs were not contributing to the internal consistency of the BCPBT. In fact, once a Fisher’s Z transformation was applied to the alphas of each of the behavioural categories and the mean was calculated, a small increase in the mean alpha was observed,
from .873 to .878, even though the alpha for touching dropped significantly from .684 to .616. The alphas for all the other pain behaviour categories were very strong; all the other behavioural categories had individual Cronbach’s alpha scores of at least .858, and three of the categories (guarding, sounds and facial expression) had alphas that exceeded .91.

Because touching had such a drastically lower alpha than the other behavioural categories, and because even after weeding out the lowest loading epochs in the scale, twenty-three out of the twenty-seven epochs in touching had item-total correlations of less than or equal to .30, it was decided that touching was a candidate for exclusion from the taxonomy. The Cronbach’s alpha for the total BCPBT excluding touching was .954, which was only a slight increase despite the loss of 26 items.
Table 7. Item-Total Correlations For The Nine Excluded Epochs (N=120)

<table>
<thead>
<tr>
<th>Behavioural Category</th>
<th>guarding</th>
<th>touching</th>
<th>words</th>
<th>sounds</th>
<th>facial expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epoch</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. to sitting</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>3. to scale</td>
<td>.43</td>
<td>X</td>
<td>X</td>
<td>.49</td>
<td>X</td>
</tr>
<tr>
<td>5. to landmarks</td>
<td>.45</td>
<td>X</td>
<td>X</td>
<td>.41</td>
<td>X</td>
</tr>
<tr>
<td>21. lumbar extension 5</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>22. lumbar flexion 5</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>31. lateral flexion left 5</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>32. lateral flexion right 5</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>38. knee reflexes</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>53. prone active extension</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

X - the item-total correlation less than or equal to .30
Part II: The Component Structure of the BCPBT

The dataset obtained from the WCB-BC is also an opportunity to ascertain the latent component structure of the BCPBT. Prevailing theory states that, like pain itself, pain behaviour may not be unitary in nature. Each type of pain behaviour appears to serve a different function, and each may have different physiological antecedents (Prkachin, 1986).

Principal components analysis was chosen over factor analysis to explore the latent structure of the variables, as factor analysis only utilizes the common variances between the variables, and not the total variance present (Dunteman, 1989). The component structure produced will reveal if pain behaviour is indeed a comprised of a number of distinct subtypes of behaviour, or if it is a unified construct.

Method

Participants

The participants in this study were also the participants in Study 5.

Design and procedure

The data produced for Study 5 was used for the exploratory principal components analyses, conducted to illuminate the underlying component composition of the BCPBT for possible future psychometric improvements. To fully analyze the component structure of the BCPBT required three separate sets of principal components analyses. The first set of analyses was on the overarching component structure of the summed totals of the behavioural categories. The second set examined the latent structure of each behavioural
category. The third, and most exploratory, set of analyses utilized every useable item of the BCPBT individually to see if a pattern of component loadings would emerge.

**Data scoring**

All data were left in their original form, coded either "0", "1", or "2". Data were not transformed. Every behavioural category in every epoch was considered to be an item, or variable, resulting in a total of 177 variables. For behavioural category-specific analyses, guarding, sounds and facial expression had 36 variables per patient; due to aforementioned administration issues, touching and words had fewer variables, 35 and 34, respectively. Initially, in total, 177 variables were entered into the analyses.

**Data screening**

Preparatory data screening was performed. The first principal components analysis utilized total scores across each behavioural category, so no data needed to be screened. For the second principal components analyses, all variables that had a mean of 0 were removed before the analyses were run. In total, 16 variables were removed; 10 of these were from the behavioural category of touching, 2 were from the category guarding, 2 were from the category facial expression and 2 were from the category words. The remaining 161 variables were entered into the principal components analyses. For the third and final set of principal components analyses, the total number of variables exceeded the number of observations per participant. To reduce the number of variables, the behavioural category of touching was removed in accordance with Study 6A, which found that touching was not as internally consistent as the other four behavioural
categories. Also, all epochs that had item-total correlations of less than or equal to .30 in three or more of the behavioural categories (see Table 7) were removed from the dataset as well for the third set of analyses. These exclusions left 106 variables to enter into the analyses.

Results

The first analysis, that which included only the summed totals of each behavioural category (and thus included only five variables), yielded only one component larger than 1 (eigenvalue = 2.65, variance explained = 52.9%), and as such, no component rotation was necessary. All the behavioural categories’ loadings on this lone component were .60 or above (please see Table 8 for the specific size of the loadings).
Table 8. Principal Components Analysis: Factor Loadings For Each Pain Behaviour Category in the First Examination of WCB Claimants, Summing Across Epochs (N=120)

<table>
<thead>
<tr>
<th>Behavioural Category</th>
<th>Total Dataset</th>
<th>Without Touch and Excluded Epochs</th>
</tr>
</thead>
<tbody>
<tr>
<td>eigenvalue</td>
<td>2.65</td>
<td>2.38</td>
</tr>
<tr>
<td>variance</td>
<td>52.9%</td>
<td>59.4%</td>
</tr>
</tbody>
</table>

**Component Loadings**

- guarding: .779, .780
- touching: .613, -
- words: .749, .748
- sounds: .796, .835
- facial expression: .685, .714
Because of the findings of Study 6a, that nine epochs and the entire behavioural category of touching did not significantly contribute to the overall internal consistency of the BCPBT, the first analysis was repeated, but excluding these variables. The results are also displayed in Table 8. After the deletion of the chosen epochs, the analysis showed that the primary (and only) component accounted for approximately 6% more variance than the primary component of the five component model (52.9% and 58.6%, respectively), and that the four remaining behavioural categories loaded even more strongly onto that first component than in the initial analysis.

The second set of analyses did nothing to contradict this unitary component hypothesis. Each behavioural category was analyzed separately to detect the latent component structure contained within. Five behavioural category-specific principal components analyses were conducted. A quartimax rotational strategy was employed for each analysis to rotate the component matrices into more interpretable configurations. Quartimax rotation was used rather than the more popular varimax rotation (Dunteman, 1989) because quartimax rotation tends to maximize the interpretability of the grouped variables (Tabachnick & Fidell, 1996).

Scree plots, visual analyses of the magnitude of the eigenvalues plotted against the components (Cattell, 1966), were produced for all five behavioural category results. Inspection of these scree plots indicated that all categories but touching extracted one very large primary component upon which most, if not all, epochs loaded strongly, but each behavioural category differed as to the magnitude of this component (see Table 9 for a listing of the behavioural categories’ individual eigenvalues). Touching had one primary component (with an eigenvalue = 3.19), but also had another possibly viable
component with an eigenvalue of 2.49. From the scree plot for touching, a break in the slope of the graph between Factor 2 and Factor 3 was observed, indicating that touching probably has two smallish latent components rather than the single components demonstrated by the other behavioural categories.

The third and final set of principal components analyses performed examined the component structure of the whole dataset; 106 variables (one for each behavioural category of each epoch, not including touching or any of the excluded variables mentioned in Data Screening) were entered into a principal components analysis. Again, a quartimax rotational strategy was utilized. Although 28 components greater than 1 were extracted, only the first five of these components were both interpretable and psychometrically relevant.

The epochs of each behavioural category tended to group with others from the same behavioural categories. Table 10 contains the component structure, and a breakdown of the larger (i.e. .30 or above) behavioural category epoch component loadings. (It must be noted that, although a cut-off of .30 was used to describe the component loadings, most of the loadings noted were quite high, between .50 and .80). As can be seen from the table, sounds epochs tended to load most heavily on the very large first component, while the guarding epochs loaded most heavily on Factor 2. A variety of behavioural category epochs loaded onto Factor 3, but these behavioural category epochs tended to be from the same total epoch, suggesting that this component reflects a natural grouping of variance caused by the behavioural tasks of the standardized physical examination. The behavioural categories of five total epochs loaded heavily on this component: 37. To Sitting, 43. To Supine, 50. To Prone, 54. To Supine 2, and 57. To
Standing. (It is interesting, in that these are all transition epochs, in that they are coded periods of the examination in which the patients move from one position to another, and, as such, are periods during which patients do not usually expect to be observed.) Facial expression epochs loaded primarily onto Factor 4, and finally, the epochs of the behavioural category of words loaded primarily onto Factor 5.

As the number of variables approached the number of participants in the dataset in this analysis, it was decided that a set of split-halves principal components analyses would be performed to test if the component structure observed in the total analysis would be maintained. The behavioural category epochs were randomly assigned into one of two groups, each with 53 variables, and two principal components analyses were performed on them.
Table 9. Principal Components Analysis: Primary Eigenvalue Extracted For Each Pain Behaviour Category in the First Examination of WCB Claimants, Across Epochs (N=120)

<table>
<thead>
<tr>
<th>Behavioural Category</th>
<th>Primary Extracted Component Eigenvalue</th>
<th>Variance Explained</th>
</tr>
</thead>
<tbody>
<tr>
<td>guarding</td>
<td>9.72</td>
<td>30.38%</td>
</tr>
<tr>
<td>touching</td>
<td>3.19</td>
<td>11.38%</td>
</tr>
<tr>
<td>words</td>
<td>5.76</td>
<td>19.20%</td>
</tr>
<tr>
<td>sounds</td>
<td>10.10</td>
<td>31.57%</td>
</tr>
<tr>
<td>facial expression</td>
<td>8.17</td>
<td>25.54%</td>
</tr>
</tbody>
</table>
Table 10. Principal Components Analysis: Factor Loadings For Individual Behavioural Category-Epochs For the First Examination of WCB Claimants (N=120, Variables=106)

<table>
<thead>
<tr>
<th>Factor</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eigenvalue</td>
<td>20.42</td>
<td>6.80</td>
<td>6.07</td>
<td>5.07</td>
<td>3.80</td>
</tr>
<tr>
<td>Variance</td>
<td>19.3%</td>
<td>6.4%</td>
<td>5.7%</td>
<td>4.8%</td>
<td>3.6%</td>
</tr>
</tbody>
</table>

**Items Loading .30 or Above**
- sounds 08 guarding 09 Face 37 face 08 words 09
- sounds 09 guarding 11 guarding 37 face 09 words 13
- words 09 guarding 13 sounds 37 face 11 words 14
- sounds 11 guarding 14 sounds 40 face 13 words 17
- sounds 13 guarding 17 Face 43 face 14 words 18
- sounds 14 guarding 18 guarding 43 face 17 words 23
- sounds 17 guarding 23 sounds 43 face 18 words 24
- sounds 18 guarding 24 Face 50 face 23 words 28
- sounds 23 guarding 27 guarding 50 face 24 sounds 39
- sounds 24 guarding 28 sounds 50 face 27 sounds 42
- sounds 27 guarding 37 Face 54 face 28 words 42
- sounds 28 guarding 42 guarding 54 face 34 sounds 48
- sounds 43 guarding 43 sounds 54 face 48 words 48
- sounds 48 guarding 50 words 54 face 52 words 49
- sounds 49 guarding 52 Face 57 face 54 words 50
- sounds 50 guarding 54 guarding 57 face 55 words 57
- sounds 52 guarding 55 sounds 57 face 56
- sounds 54 guarding 57
- sounds 55
- sounds 56
Table 11. Principal Components Analysis, Split-Half 1: Factor Loadings For Individual Behavioural Category-Epochs For the First Examination of WCB Claimants (N=120, Variables=53)

<table>
<thead>
<tr>
<th>Factor</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eigenvalue</td>
<td>10.11</td>
<td>3.53</td>
<td>3.44</td>
<td>2.65</td>
<td>2.12</td>
</tr>
<tr>
<td>Variance</td>
<td>19.1%</td>
<td>6.7%</td>
<td>6.5%</td>
<td>5%</td>
<td>4%</td>
</tr>
</tbody>
</table>

Items Loading .30 or Above

- sounds 08 guarding 11 Face 08 sounds 37 sounds 08
- sounds 11 guarding 13 Face 09 face 43 words 08
- sounds 18 guarding 17 Face 14 guarding 43 face 56
- sounds 23 guarding 23 Face 17 face 54 sounds 56
- sounds 27 guarding 42 Face 18 guarding 54 sounds 56
- sounds 37 guarding 43 Face 28 sounds 54
- sounds 48 words 43 Face 54 guarding 57
- sounds 49 guarding 48 Face 55
- words 49 guarding 54
- sounds 52 guarding 57
- sounds 54
- sounds 56
Table 12. Principal Components Analysis, Split-Half 2: Factor Loadings For Individual Behavioural Category-Epochs For the First Examination of WCB Claimants (N=120, Variables=53)

<table>
<thead>
<tr>
<th>Factor</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eigenvalue</td>
<td>11.18</td>
<td>4.12</td>
<td>3.16</td>
<td>2.98</td>
<td>2.23</td>
</tr>
<tr>
<td>Variance</td>
<td>21.1%</td>
<td>7.8%</td>
<td>6%</td>
<td>5.6%</td>
<td>4.2%</td>
</tr>
</tbody>
</table>

Items Loading .30 Or Above

- sounds 09
- words 09
- guarding 08
- guarding 09
- face 11
- words 09
- words 13
- guarding 09
- words 09
- face 13
- sounds 13
- words 14
- guarding 14
- face 37
- face 23
- sounds 14
- words 17
- guarding 18
- guarding 37
- face 24
- sounds 17
- words 18
- guarding 24
- sounds 39
- face 27
- sounds 24
- words 23
- guarding 27
- sounds 43
- face 33
- sounds 28
- words 24
- guarding 28
- face 50
- face 34
- sounds 43
- words 27
- guarding 37
- guarding 50
- face 52
- sounds 50
- sounds 39
- guarding 50
- sounds 50
- face 57
- sounds 55
- sounds 42
- guarding 52
- words 54
- face 57
- sounds 57
- words 42
- face 57
- sounds 57
- words 57
The results of these two analyses can be seen in Tables 13 and 14. The behavioural categories still tended to group together, but the relative order of the components they loaded onto shifted. On both, the largest, first components were approximately the same size (eigenvalues of 10.11 and 11.18), and both first components were the primary loading component for sounds epochs. In split-half analysis number 1 (Table 11), guarding epochs loaded mostly strongly on component 2 (eigenvalue = 3.53), facial expression epochs loaded on component 3 (eigenvalue = 3.44), and the transitional epochs loaded best on component 4 (eigenvalue = 2.65). Factor 5 (eigenvalue = 2.12) seems to be the result of variance in BCPBT scores caused by similar movements, as the five behavioural category epochs that loaded strongly on it were from the epochs 8. Heel Raise and 56. Bilateral Straight Leg Raise.

The second split-half principal components analysis (seen in Table 12) also exhibited components upon which the epochs of guarding and of facial expression loaded heavily, component 3 (eigenvalue = 3.16) and component 5 (eigenvalue = 2.23) respectively, but in this second analysis, words epochs loaded almost exclusively on component 2. Finally, the transitional epochs loaded exclusively onto component 4 (eigenvalue = 2.98).

Although the order of the components varied between the two split-half principal components analyses, the epochs from each behavioural category clearly clustered together, loading strongly on the same components. Moreover, the transitional epochs loaded exclusively on component 4 of each analysis, showing a surprisingly consistency between the two analyses.
Discussion

There were five main objectives to this body of research: to ascertain whether the BCPBT could be taught to be administered with a minimum of error, to determine if coding accuracy could be maintained over time, to learn if personalized feedback could boost coding accuracy, to observe the reliability of the measure in a clinical setting, and to evaluate the psychometric properties of the measure for possible test refinement. Each of these goals was met to varying degrees in the six studies outlined above.

Reliability

From the results of Study 1 though Study 4, it is evident that the BCPBT can be learned easily and later administered with a high level of precision. Specifically, in Study 1, the mean scores on the behaviour coding tests increased subsequent to training, and the improvement was linked temporally to the occurrence of training through the use of the multiple baseline design. The lack of significant statistics of the difference scores between the Immediate and the Delayed training groups was primarily due to small sample size, as there were large effect sizes associated with each test. Study 2 illustrated that once training had occurred, the judges were able to retain their knowledge of the BCPBT and to administer it consistently over a period of either eight or fifteen days (test-retest reliability $r = .89$).

However, the UNBC judges were unable to retain this information over the intervening five months, and when they were tested again at follow-up, they had reverted to their pre-training accuracy levels. This decrement in their performance was easily rectified by the brief follow-up training session and feedback. This finding makes a
potent argument for constant practice of the administration of the BCPBT amongst judges, as well as the need for periodic recalibration session to ensure consistency and to curtail coder drift (Keefe & Williams, 1992).

The agreement scores of the post-training groups were consistently over 80% both in the training situations and in the clinical setting. The mean scores of all judges after training ($M_{\text{study1, 2 & 3}} = 82.24\%, M_{\text{study 4A}} = 80.70\%, M_{\text{study 4B}} = 87.71\%$) consistently exceeded the acceptability levels set out by Dworkin and Whitney (1992). However, in the same studies, the judges showed just as consistently that they were unable to meet the 90% coding criterion set by Keefe for the users of his taxonomy (Keefe & Block, 1982; Keefe, Crisson, Maltbie, Bradley, & Gil, 1986; Keefe, Crisson, & Snipes, 1987). As the BCPBT is based on Keefe's taxonomy, this inability to achieve agreement scores of 90% or greater is a conundrum.

Given that in the past, other researchers have had no difficulty in achieving the 90% criterion with their judges, even when they altered Keefe’s taxonomy slightly to suit the population under observation (most notably, the work done with cancer patients by Ahles et al, 1990), the cause of the current inability to surpass the 90% criterion may lie with the drastic modifications made to the taxonomy to create the BCPBT. One possible source of additional judgement error in the BCPBT was the elimination of Keefe’s three-second rule, which defined pain behaviours as codable only if they endured for three seconds or longer. While the decision to eliminate this rule was theoretically sound (Ekman, 1984; Ekman, 1992; Prkachin, 1997), it may have had unforeseen practical drawbacks in the administration of the measure.
By removing the time duration criteria for the pain behaviour categories, all perceptible pain behaviours are now deemed codable, regardless of how close to the judges’ perception thresholds they may be. Even those behaviours that Ekman termed “microbehaviors” (Ekman, 1992) should be coded by observant judges. While this can be seen as a beneficial alteration in that it increases the measure’s sensitivity to pain behaviours, it may have also lowered the judges’ accuracy scores by making codable behaviours more difficult to perceive.

The quality of the videotape training and testing materials may have also exacerbated this problem. The source tapes for all the audiovisual training and testing materials were two generations (or “steps”) away from the original copy, making the information they contained less clear than had the original tapes been used. Moreover, the video recording sessions were plagued with poor lighting and auditory interference from the video recorder itself and other ambient noise in the examination room. However, it was necessary to use the videotapes as they were received. This factor may explain the dramatic increase of percent agreement scores between Study 4 Part 1 and Part 2; WCB-BC judges’ experienced a leap in percent agreement (from 80.90% to 87.71%) in the shift from training to a clinical setting. It is probable that judges found the codable pain behaviours to be more salient (and more easily codable) once they could observe them in person at close range.

It should be noted that the judges who took part in Study 1 learned the BCPBT coding technique remarkably quickly. After only 15 minutes of training with the brief manual (see Appendix C) and no visual examples, the judges were able to achieve a mean accuracy score of 75.6% on Test 1, the pre-test. This indicates that the BCPBT is easy to
learn, and intuitive to administer. For future training groups, it may be possible to reduce the duration of the existing training protocol from its current five hour timeframe.

Study 3 showed that personalizing feedback to the judges regarding their own coding strategy mistakes was overall beneficial to their percent agreement scores, but not as much as was anticipated. UNBC judges only increased their coding accuracy by 7.5% between the follow-up pre-test and post-test, from 76.1% to 83.6%. As has been mentioned earlier, Keefe’s coders all reached 90% agreement or more after training (Keefe & Block, 1982; Keefe, Crisson, & Snipes, 1987; Keefe & Williams, 1992). It is possible that the UNBC judges hit a ceiling to their ability to perceive the pain behaviours on the test videotapes, because even after the initial intensive training sessions, the judges were not able to exceed a mean percent agreement of approximately 84%. As the WCB-BC judges displayed a dramatic jump in percent agreement between their training test and the clinical observation pilot (from 80.9% to 87.7%), it can be hypothesized that the training and testing videotapes may have inherent accuracy-limiting drawbacks that prevented the judges from meeting the 90% criterion observed by other researchers (Ahles et al, 1990; Keefe & Block, 1982; Keefe, Crisson, & Snipes, 1987).

It is also possible that the self-directed training program that Keefe employs to train his judges is somehow superior to the one outlined here. Keefe’s judges can spend up to eight hours studying the materials and taking progress tests at their own speed before they are considered to be a acceptable judge in a research setting (Keefe, Crisson, & Snipes, 1987). Moreover, not all judges who begin training finish it; judges who do not reach criterion coding levels after their training do not continue coding. In the present studies (specifically, Studies 1 to 5), two distinct subgroups of judge can be seen from the
raw percent agreement scores in these studies – those with scores in the mid to high 80s range, and those with scores consistently under 80 - there may be a personal suitability factor at work here as well.

The results of Study 5 indicate that, although the test-retest correlations between the scores obtained on the BCPBT after two different administrations are not uniformly strong, the correlations obtained for guarding, facial expression and the overall total score were moderate to good by Dworkin and Whitney’s standards (1992). Touching and words displayed rather low correlations, in the .30 range. However, it is imperative to remember that chronic low-back-pain can be cyclical in nature; patients report (and demonstrate) daily, even hourly fluctuations in the severity of their pain. In this way, the psychometric characteristics of chronic low-back-pain may be unlike that of other, presumably more stable characteristics, like the presence of allergies, or a measurement of intelligence. Acceptable test-retest correlation levels for these, less mutable phenomena are not necessarily good benchmarks for the test-retest reliability of a measure of chronic pain.

To estimate the true test-retest correlation of the BCPBT, the random fluctuations in score due to a change in the severity of the patients' pain experiences would have to be factored out prior to the correlation calculation. Unfortunately, not enough information was collected in this study to allow for such an estimate to be computed. Further research on this area is clearly needed.
Internal Consistency

Based on the findings of Study 6A, the behavioural category of touching was removed from the scale. Keefe and Williams (1992) have stated that for behavioural observation purposes, a target behaviour needs to occur with relative frequency to achieve a degree of variance needed to make discriminations. After the item-total correlations analysis, it became evident that touching was not occurring with the frequency necessary to make it a potent component of the BCPBT. To lend support to this argument, without touching, the internal consistency indicator, Cronbach's alpha, actually increased. The overall alpha did not significantly change, dropping almost imperceptibly from .954 for 177 items (including touching) to .955 for 144 items (excluding touching). Although they are both reflect excellent levels of reliability (Dworkin & Whitney, 1992), the fact that the alpha increased by .001 despite a considerable reduction in the number of coded items suggests that the removal of the category touching was not counterproductive.

Also deleted from the scale were the following epochs: 1. To Sitting, 3. To Scale, 5. To Landmarks, 21. Lumbar Extension 5, 22. Lumbar Flexion 5, 31. Lateral Flexion Left 5, 32. Lateral Flexion Right 5, 3.8 Knee Reflexes, and 53. Prone Active Extension. These were deleted because fewer than three of the remaining four behavioural categories were significantly associated with them. Specifically, three or more behavioural categories had item-total correlations of .30 or less on that epoch. Two epochs that fit this criterion were not removed because the epochs contain the Waddell signs, 33. axial loading and 34. simulated rotation (Waddell, McCulloch, Kummel, & Venner, 1980, p.118), and they were of particular research interest to the WCB-BC. The removal of the nine epochs was warranted because BCPBT, while relatively simple to
use, was still somewhat cognitively taxing to observers; therefore, lowering the number of coding epochs was a pragmatic necessity. Also, these epochs were serving little purpose within the scale. After the removal of the poorly functioning epochs, the Cronbach’s alphas of the behavioural categories actually increased, albeit in a small way (see Table 6). As decreasing the number of items used to compute Cronbach’s alpha usually decreases the resultant alpha, these small increases were hard won, and of practical significance.

The alpha for the total measure, however, did not increase; its magnitude was reduced slightly. After the removal of 74 items, consisting of nine complete epochs, two partial epochs (each in the category words due to aforementioned administration problems), and the entire behavioural category of touching, the overall Cronbach’s alpha decreased from .955 to .954. This decrement (which was actually a .0006 decrease) in the overall alpha suggests that the removal of these items did not alter the reliability of the scale substantially, and that the internal consistency for the modified measure is excellent.

**Principal Components Analyses**

The principal components analyses performed on the datasets gleaned from the WCB-BC’s use of the BCPBT during 120 standardized physical examinations yielded results that may have long term effects on the way pain behaviour is perceived. The global principal components analysis conducted on the summed totals of each behavioural category resulted in a strong unidimensional solution. Only one component emerged with an eigenvalue greater than one, and it accounted for up to 57.6% of the total variance in the dataset (when touching and the nine poorly performing epochs were
excluded from the analysis). This finding seems to provide strong evidence that the underlying concept associated with this component is a general construct of pain behaviour. More evidence to support this view was found when the second series of principal components analyses were performed on the individual behavioural categories. All but one of the pain behaviours emerged with an overwhelmingly large first component that accounted for the bulk of the variance. However, it wasn’t until the third set of principal components analyses were conducted that the results of the first two sets were reinterpreted to perhaps reflect a multidimensional view of pain behaviour.

The third set of PCA uncovered a series of component loadings that shows that each category of pain behaviour has its own unique variance, which indicates that pain behaviour, as a construct, may have distinct subtypes of behaviour, and as such, may not be unitary in nature. The results from third set of analyses indicated that there was primary latent component that explained a bulk of the variance, and that the behavioural category of sounds loaded strongly on that component. Other components followed which were, of course, smaller, but no less interpretable; the second latent component was loaded exclusively with guarding epochs, and the fourth and fifth components loaded primarily with facial expression epochs and words epochs, respectively. Each of these components had their own unique variance, completely orthogonal from the components that preceded them. This emergent component structure is extremely suggestive of a concept of pain behaviour as a construct that has a multiplicity of forms that are related, but not identical.

However, there is another possible explanation for these findings. Artificially constructed perceptions of types of pain behaviours instilled by the training process may
have created the data pattern seen in the analyses. In other words, systematic differences between types of pain behaviours may have been observed because the judges were trained to perceive them in that manner. The training may have created synthetic boundaries in the judges' perceptions of pain behaviour when in reality, there are no such boundaries. Unfortunately, this study was designed to neither refute nor support this hypothesis, and as such, it will have to be addressed in future research.

Another fascinating finding of the final set of the principal components analyses was the relatively high component loading of the transitional epochs, the substantial amount of variance they accounted for, which exceeded even that of facial expression and of words. It is unlikely that this high component loading was an unintentional by-product of the coding process, as the transitional epochs were not singled out from the other task-related epochs in any way during training. The definitions of all epochs were presented in a similar manner, beginning with the movement of the participant, and ending with the cessation of movement; no special attention was paid to the definitions or the coding of the transitional epochs. Therefore, because it was found that the transitional epochs loaded significantly on the third component in the solution matrix and accounted for approximately 6% of the variance of the dataset, meaning that the transitional epochs represent a unique portion of variance in the dataset, this suggests the existence of a latent construct tied to the transitional epochs, although it is doubtful that the nature of the latent construct is that of transition. As the transitional epochs occur between standardized tasks in the physical examination, and people rarely expect their behaviour between tests to be observed and recorded, it is possible that transitional epochs contain pain behaviour that is more candid than that in the other epochs. This effect may be
heightened because of the evaluative (and potentially adversarial) nature of the setting in which the behaviours occurred (the WCB-BC). The latent construct may be of candid pain behaviour, or pain behaviour that occurs when the participants are not monitoring their own reactions. If this is indeed the case, the transitional epochs may be clinically more informative than the other epochs, and it may be wise to concentrate on the development of these epochs if the BCPBT is to be refined further.

**Future Directions**

While this research’s results are promising for the future use of the BCPBT, if the measure is to be used beyond its intended purpose for the WCB-BC, its psychometric properties must be further examined. Several epochs were removed, as was an entire behavioural category; sweeping structural changes like these can radically alter the psychometric properties of even a measure that has a proven track record of reliability and validity over time.

Moreover, the BCPBT should be compared with a normal population for discriminant purposes. It is important to establish that a measure specifically designed to assess abnormal populations be administered to those who are ostensibly not part of that population, to determine if the measure can discriminate group membership with any accuracy, as was done in Keefe’s taxonomy (Keefe & Block, 1982; Keefe, Crisson, & Snipes, 1987). This would provide a firm foundation on which to build the evidence of the BCPBT’s validity.

After the requisite test-building has been done, and if the BCPBT appears to be stable, reliable and valid in its current incarnation, future researchers may want to
examine the possibility of adapting the BCPBT for use in conjunction with other
standardized examinations for other pain conditions. Pain behaviour caused by other
types of pain, such as cancer pain, postoperative pain, and pain from arthritis (just to
name a few) needs to be measured accurately in health care research with surprising
frequency. The form of the BCPBT may be flexible enough to be adapted to any pain-
causing disorder, as long as there is an appropriately standardized physical examination
protocol to which to yoke the measure. At that point, of course, the epoch definitions will
have to be redefined to reflect the new standardized physical examination, but the general
structure of the BCPBT may prove to be sufficiently plastic survive the change from the
low-back-pain context.
References


Appendix A. Informed Consent Form for Study 1
Properties of the BCPBT

Consent Form

The goal of this study, “Assessing the Reliability of the British Columbia Pain Behaviour Taxonomy (BCPBT)”, is to determine if individuals trained in the BCPBT can identify pain behaviours more accurately than can individuals not trained in the BCPBT. There are several things you will need to know before agreeing to participate in this research. To begin, within the training protocol, you will see videotaped clips of people undergoing a standardized physical exam who may or may not be experiencing pain during the examination. Some people may find these images disturbing. If you agree to participate in this research, please be aware that these images will be shown and studied.

Also, this research will take place over three consecutive Saturdays. On both the first and the second Saturday, you will be required to take two 30-minute tests; one in the morning (9 a.m.) and one in the evening (5 p.m.). On one of these Saturdays, you will be required to take part in a 6-hour training seminar to learn the BCPBT. On the third Saturday, you will be required only to complete one 30-minute test in the morning, after which you are free to go. If you do not attend all three Saturdays, you will not be eligible for the $100 honorarium.

Finally, the Workers’ Compensation Board of British Columbia is funding this research, and will have access to the work that you do. You will not be required to do anything for them but learn and be tested on the BCPBT. However, they may use your test scores in further research.
If you agree to become a participant in this study, you have certain rights and responsibilities. You have the right to:

- withdraw at any time
- ask questions about the experiment at appropriate times
- have the experiment explained to you fully
- be paid an honourarium of $100 if all three sections of the experiment is completed

You also have a number of responsibilities. By agreeing to participate, you are agreeing to:

- show up for all three sections of the experiment
- learn the material to which you will be exposed to the best of your ability
- keep confidential all aspects of the materials to which you will be exposed

By signing below, you are indicating that you understand these rights and responsibilities as outlined above, and agree to participate in this research.

name: __________________________________________  date: __________________

signature: ______________________________________________________________________

witness: ________________________________________________________________________
If you have any problems with your participation in the research after you have signed this agreement, you should contact E. A. Hughes at 562-6687 or Dr. Prkachin at 960-6633.
Appendix B. Informed Consent Form for Study 3
The goal of this study, "Assessing the Reliability of the British Columbia Pain Behaviour Taxonomy (BCPBT)", is to determine if individuals trained in the BCPBT can identify pain behaviours more accurately than can individuals not trained in the BCPBT. There are several things you will need to know before agreeing to participate in this research. To begin, within the training protocol, you will see videotaped clips of people undergoing a standardized physical exam who may or may not be experiencing pain during the examination. Some people may find these images disturbing. If you agree to participate in this research, please be aware that these images will be shown and studied.

Additionally, the Workers’ Compensation Board of British Columbia is funding this research, and will have access to the work that you do. You will not be required to do anything for them but learn and be tested on the BCPBT. However, they may use your test scores in further research.

If you agree to become a participant in this study, you have certain rights and responsibilities. You have the right to:

- withdraw at any time
- ask questions about the experiment at appropriate times
- have to the experiment explained to you fully
- be paid an honourarium of $50 if the experiment is completed
You also have a number of responsibilities. By agreeing to participate, you are agreeing to:

- learn the material to which you will be exposed to the best of your ability
- keep confidential all aspects of the materials to which you will be exposed

By signing below, you are indicating that you understand these rights and responsibilities as outlined above, and agree to participate in this research.

name: ___________________________ date: _______________________

signature: ______________________________

witness: ______________________________

If you have any problems with your participation in the research after you have signed this agreement, you should contact E. A. Hughes at 562-6687 or Dr. Prkachin at 960-6633.
Appendix C. Brief Coding Manual
BRIEF TRAINING

TYPES OF BEHAVIOUR

You can usually tell when someone is in pain by the way that they act. Sometimes they rub the affected part of their body like this (demonstrate), or limp like this (demonstrate).

Past research has isolated five categories of pain behaviour, behaviour that lets other people know that someone is experiencing pain. They are as follows:

1. “guarding” is behaviour that prevents or alleviates the experience of pain. It is the most encompassing behavioural category, and its subtypes include stiffness, hesitation, limping, bracing, and flinching. (Take note: in coding situations, you do not have to identify the sub-type of behaviour that occurred. You just have to indicate that “guarding” was observed.

- stiffness - a marked lack of normal flexibility in movement; maintaining a rigid posture
- hesitation - a reluctance to move or an interruption in movement
- limping - the patient fails to apply weight to or favours one leg while walking or walks with an abnormal gait
- bracing - behaviour that cannot be described as limping, but in which the patient places an abnormal amount of weight upon a part of the body. Bracing can occur both when moving or stationary, and includes using objects or the self as an aid
Properties of the BCPBT

(e.g. leaning on a table for support, using a cane, pushing against oneself to rise from a seated position).

- flinching - a sudden withdrawal or spasm of a part of the body.

2. "**Touching**" is defined as any contact between the patient's hands and the lower back, the hips, buttocks and the outer aspect of the thighs (i.e. the area around the painful site). The behaviour exhibited may be passive touching, active rubbing or massaging (which would entail noticeable movement of the hand(s) over the painful area of the body), grabbing, squeezing, holding or pushing/supporting the aforementioned area.

3. "**Words**" is defined as any *spontaneous* (i.e. not elicited from the examiner's questions) verbal complaint that relates to the patient's pain.
   - commands: "Stop it."
   - information: "That hurts.", "It aches when I move it.", "That's as far as I can go."
   - interjections / expletives: "Ouch.", "Yikes!"

4. "**Sounds**" - voluntary or involuntary production of one of the following:
   - sighing: a puffing or slow exhalation of breath. It can be long or short, and may be accompanied by a shrugging movement of the shoulders or a deflation of the chest.
   - moaning
   - grunting
   - screaming
• crying: must be an auditory occurrence; do not code ‘Sounds’ if the patient only tears up or sheds tears silently.

5. “Facial Expression”, as a behavioural category, is more complex than the preceding four. Not only can it be divided into four subtypes of pain behaviour, facial expression can also be classified into degrees of intensity. First, however, we will discuss the subtypes of this behaviour. Research has shown that pain is most often expressed by the human face using one or more of the following four behaviours:

• furrowing your forehead
• squinting your eyes
• wrinkling your nose
• sneering with your upper lip.

Moreover, you can have variable intensities of facial expression. Dr. Prkachin and I have decided, for ease of application, that you should use the categories “none”, “some”, and “lots” when classifying facial expression. If you don’t see any facial expression, that would mean you should indicate that there was “none”; if you only saw a little facial expression, then that would be “some”; and if you saw an extreme facial expression, then that would be “lots”.

Now that you know what behaviours you should be looking for to indicate that someone is in pain, you need to know how to mark it down when you see it. The test you will be getting in a couple of minutes will show you 40 videotaped clips of people undergoing
standardized physical examinations. Some of them will be showing pain behaviours during the video clip, and some of them won’t be showing any pain behaviours. Your job will be to identify the pain behaviours they do show, and mark them down in a table that looks like this:

<table>
<thead>
<tr>
<th>clip #</th>
<th>guarding</th>
<th>touching</th>
<th>words</th>
<th>sounds</th>
<th>facial expres</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1 2</td>
</tr>
<tr>
<td>2.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1 2</td>
</tr>
</tbody>
</table>

The coding system is only yes or no (except with “facial expression”, but this will be discussed presently). You only need to worry about identifying whether each type of pain behaviour occurred during the video clip. If the person in the clip demonstrates “guarding” twice, you only have to put a checkmark under “guarding” once to indicate that it happened. The frequency with which the behaviour occurred is not needed.

For “facial expression”, you have to judge not only whether you saw a pain facial expression, but you also have to judge how intense the facial expression was. If you didn’t see a facial expression of pain in the clip, leave the square blank. If you saw a mild one, circle “1” (for “some” facial expression); if you saw an intense pain facial expression, circle “2” (for “lots” of facial expression).
Always keep in mind that people use these pain behaviours in different ways to express pain. Pain expression is not exactly the same between people, and it is not exactly the same within the same person over time. So you must pay careful attention to each video clip to be able to identify the unique pattern of pain behaviour expressed there. Do not rely on what you have seen before, or what you expect to happen.
Appendix D. Extended Coding Manual
Introduction

The characteristic behaviours displayed by persons experiencing pain can be very informative in clinical settings. However, to get an accurate assessment of a patient’s pain behaviour, it must be observed systematically. This manual will teach you a system of coding designed to detect pain behaviour in a specific clinical setting, that of the Worker’s Compensation Board of BC.

The system may seem complex at first, but with practice, it will become easier. Your goal is to become so comfortable with the system that it becomes automatic. This will come with time and experience.

There are a number of aspects to the present system that are not similar to other systems of its kind: the system is specifically structured around the standardized physical examination of the WCB, ignores duration of behaviour, and is both patient- and movement-driven. The standard examination of the WCB was divided up into discrete sections, or epochs, of activity. Pain behaviours are coded only during these epochs, and only as present or absent (except in the case of facial expression, but this will be discussed separately). There is no cumulative count of the behaviours within an epoch, only over the total number of epochs. Thus, it is not necessary to attend to the frequency or the duration of each type of pain behaviour within an epoch. Also, there is no time
restriction for the length of an epoch. The beginning and ending of each epoch is determined by the movement of the patient (there are a few exceptions to this rule, but they will be discussed separately).

Section 1: Operational Definitions of the Behaviours

There are five general types of behaviours relevant to this behavioural observation system. These are: guarding, touching, words, sounds, and facial expression. Each are overarching categories of behaviour that contain much variation. You will be expected to code only for these broad categories, and not for their numerous subtypes.

I. guarding is behaviour that prevents or alleviates the experience of pain. It is the most encompassing behavioural category, and its subtypes include stiffness, hesitation, limping, bracing, and flinching.

- stiffness - a marked lack of normal flexibility in movement; maintaining a rigid posture
- hesitation - a reluctance to move or an interruption in movement
- limping - the patient fails to apply weight to or favours one leg while walking or walks with an abnormal gait
- bracing - behaviour that cannot be described as limping, but in which the patient places an abnormal amount of weight upon a part of the body. Bracing can occur both when moving or stationary, and includes using objects or the self as an aid
(e.g. leaning on a table for support, using a cane, pushing against oneself to rise from a seated position).

- flinching - a sudden withdrawal or spasm of a part of the body.

General points:

Do not code ‘Guarding (bracing)’ if you suspect that the patient is only touching the object/other person/self for balance. An abnormal amount of body weight must be borne by the limb doing the bracing.

Also, when judging patients as they raise or lower themselves into prone or supine positions: do not code ‘Guarding (bracing)’ for their use of their arms for assistance unless the amount of weight being borne by the arms is abnormal.

II. Touching is defined as any contact between the patients hands and the lower back, the hips, buttocks and the outer aspect of the thighs (i.e. the area around the painful site). The behaviour exhibited may be passive touching, active rubbing or massaging (which would entail noticeable movement of the hand(s) over the painful area of the body), grabbing, squeezing, holding or pushing/supporting the aforementioned area. Please note that contact between the patient’s hands and the above mentioned areas of the body must occur to code ‘Touching’

General points:
Do not code ‘Touching’ if you cannot see the patient’s hands and cannot infer with certainty that the behaviour is indeed occurring.

‘Touching’ should not be coded when the patient’s hands are folded in his or her lap, or if the patient is following the directions of the examiner (e.g. “Reach down the outside your leg and stretch.”).

III. Words is defined as any spontaneous (i.e. not elicited from the examiner’s questions) verbal complaint that relates to the patient’s pain.

Examples:

commands: “Stop it.”

- information: “That hurts.”, “It aches when I move it.”, “That’s as far as I can go.”
- interjections / expletives: “Ouch.”, “Yikes!”

General Points:

‘Words’ must be spontaneous. You should not code words that refer to the patient’s pain in the following epochs because the examiner issues the general instruction, “Tell me when it hurts.” at the beginning of each: palpation (no coding anyway), axial compression, simulated rotation, and passive straight leg raise 1 and 2.
Do not code for “**Words**” during any other epoch in which the examiner breaks from the script and explicitly asks the patient to verbalize if/when it hurts.

IV. **Sounds** - voluntary or involuntary production of one of the following:

- **sighing**: a puffing or slow exhalation of breath. It can be long or short, and may be accompanied by a shrugging movement of the shoulders or a deflation of the chest.
- **moaning**
- **grunting**
- **screaming**
- **crying**: must be an auditory occurrence; do not code ‘**Sounds**’ if the patient only tears up or sheds tears silently.

General points:

Code ‘**Sounds**’ if you observe an audible behaviour, but cannot understand it or place it as a word.

V. **Facial Expression**, as a behavioural category, is more complex than the preceding four. Not only can it be divided into four subtypes of pain behaviour, facial expression can also be classified into degrees of intensity. First, however, we will discuss the subtypes of this behaviour. Research has shown that pain is most often expressed by the
human face using one or more of the following four behaviours: the forehead furrow, the orbital squeeze, the nose wrinkle, and the sneer.

i. forehead furrow (FACS 4)

This movement is located in the forehead region of the face, and it entails a lowering of the brow, so that the eyebrows move down and possibly closer together. Wrinkling, puckering and/or bulging of the forehead skin is common. Vertical and/or diagonal lines or ridges may also form in this region, and horizontal lines may appear at the bridge of the nose. Finally, the under-eyebrow folds of skin may descend to cover more of the visible eyelids than is usual.

Note: for all the ‘Practice’ sections, use a mirror to watch yourself practicing the facial expressions.

Practice: For the forehead furrow, pull the muscles covering your brow down. If you are having difficulties in performing this movement, pretend to be angry. For many people, it is an intrinsic part of the expression of anger. If this does not work, use your fingers to gently push the brows down towards the eyes, then engage your muscles to keep your brows lowered, and take your fingers away. Repeat the expression several times. Hold it. Look at how the skin bunches up between the brows, and how the brows lower and, for many people, come closer together. Try to do it as strenuously as you can; then do it as lightly as you can. What are the differences? What are the similarities?
ii. orbital squeeze (FACS 6, 7)

The **orbital squeeze** occurs only in the upper part of the face, specifically around the eyes. It entails a narrowing of the normally visible portion of the eyes through a movement of the top lid, the bottom lid, or both (i.e. the top lid descends, the bottom lid ascends, or they both come together). This movement or movements create visible effects around the eyes: the skin around the eye contracts, possibly producing crow’s feet at the sides and bagging or puckering underneath the eyes, and the tops of the cheeks may be pulled up as well.

Practice: Narrow your eyes to slits. Notice how the skin wrinkles around your eyes? Also notice how the amount of white and pupil are visibly reduced. Make your eyes as ‘slit-like’ as possible. Now only narrow them slightly. See how the difference in intensity has an effect on the amount of wrinkling in the skin around the eyes. Next, narrow your eyes, but only use your bottom lids. See how the bottom lids bulge as they rise to cover more of the sclera? Do this move as intensely as you can, and then as lightly as you can.

All of these movements are variations of the **orbital squeeze**, and should one or all occur, it is classified as a pain behaviour.

iii. nose wrinkle (FACS 9)
The **nose wrinkle** involves several areas of the face. When it occurs, there is a characteristic movement of the nose, but also of the forehead, the apples of the cheeks, and the upper lip. As its name implies, the **nose wrinkle** causes the nose to wrinkle; it also causes a **forehead furrow**, the apples of the cheeks to be pulled straight up towards the inside corners of the eyes, and the upper lip to rise vertically as well. The lines that connect the sides of the nose with the corners of the mouth (the nasolabial furrow - think Fred Flintstone) deepen and rise straight up. The nostrils usually flare (get bigger), and in some cases, the eyes may narrow into a squint.

**Practice:** Pull the muscles around your nose upwards. If you have difficulty performing this movement, another way to approach this facial expression is as the “Eeuw!” face (it is typically indicative of disgust), alternately known as the “I smell something horrible” face. Notice how the wrinkles on your nose aren’t all the same shape? Observe how your browline drops and develops ridges, exactly like in the **forehead furrow**. See if your eyes narrow. Look at the shape of your nasolabial furrow; it should look like this /\.

Perform the expression as strenuously as you can; repeat as lightly as you can. What are the differences and similarities?

**iv. sneer (FACS 10)**

The **sneer** is fairly self-explanatory. The movement is centered on the upper lip; it rises upwards, but it also spreads diagonally as well, flattening itself slightly. The nasolabial furrow takes on a more curved disposition, appearing arch-like, as opposed to the straight
up-and-down motion inherent to the nose wrinkle. The nostrils also flare, but the nose does not wrinkle.

Practice: When performing the sneer, be careful about wrinkling your nose. Although the sneer and the nose wrinkle are similar, they are distinct movements. However, they can also occur together.

Intensity rating of Facial Expression

The facial expression type of pain behaviour is the only one of the five to have an intensity rating. This means that you are not merely coding for the presence of the behaviour; you are also coding for its strength as well. This is done on a three-point scale that ranges from zero to two. A handy mnemonic to remember the scale is “none, some or lots”, in which “none” is represented by zero, “some” by one, and “lots” by two. Intensity decisions are a gestalt process, one for which it is very difficult to design hard and fast decision rules. Defining “none” and “lots” may be straightforward; however, delineating the boundaries between “some” and “lots” is quite problematic. Intensity decisions cannot be made based solely on the number of the expressions, nor should you use duration information, as there is no time criterion for accepting a facial expression as pain behaviour (thus even fleeting facial expression should be coded). This topic will be pursued during the seminar, during which time examples will be presented, and the relevant aspects dissected.
**Combinations of Facial Expression**

All of the four facial expressions presented above will not occur in isolation with any frequency. They will usually be embedded into more complex facial expressions from which you will have to discern the relevant movements. It is important that you zero in on the expressions listed above, and do not become distracted by the presence or intensity of other, irrelevant, expressions. For example, avoid taking the whole mouth into consideration when making coding decisions; the only part of the mouth that is relevant for any of the expressions is the upper lip. Whether the mouth is open or shut, or stretched or slack, is of no consequence in this coding system, but such expressions can be highly distracting if you are not careful in your observations.

**How to Code Facial Expression When Part of the Face is Not Visible**

Coding 'Facial Expression' can be done even if part of the patient’s face is covered. Many of the facial expressions that you will be witnessing will be symmetrical, meaning that, barring some sort of facial deformity or enervation problem, the expressions will happen on both sides of the face at the same time to the same degree. Thus it is possible to infer that a patient is making a certain facial expression by examining as little as half the patient’s face. This holds true when the face is lowered, turned, or occluded (such as by the patient’s hair or hands, the examiner, etc...)

Moreover, with practice, you will be able to infer a great deal from isolated glimpses of a patient’s face. For example, if less than half of the patient’s face is visible but you still
have a clear view of the outline of the forehead, forehead furrowing can still be inferred from the characteristic lowering and bulging of the brow area.

Section 2. How to Code

The Coding Sheet

Once you have noticed that a behaviour has occurred, you should transcribe the occurrence for later study. A coding sheet has been developed for the purpose. As you are primarily making “present/absent” decisions about the behaviours (the exception is facial expression, which has an intensity decision as well), the coding sheet relies on a checkmark system. After an epoch is over, you must mark off the behaviours observed during that epoch, each behaviour within its own column. If an epoch is skipped within the examination, you would place a check in the “did not do” box on the Checklist for that epoch; if something unforeseen occurs during an epoch that prevents the accurate observation of the pain behaviours, you would mark off the “cannot code” box on the coding sheet for that epoch.

Facial expression is coded slightly differently, in that you are required to circle the perceived intensity of the observed facial expression exhibited during the epoch. If several facial expressions are exhibited over the same epoch, the most severe is chosen for transcription.

As the physical requirements of the examination can sometimes interfere with the observation of the aforementioned pain behaviours, several such epochs have been omitted from the coding system. However, these epochs are included on the coding form to keep you aware of the sequencing of the examination, but are darkened to prevent
accidental coding. If an entire row is darkened, this indicates that you should not code at all during this part of the examination. If less than the entire row of the coding sheet has been darkened, this means that the examination somehow limits or blocks the performance of certain behaviour(s), and that you should not code for that behaviour(s) during that epoch.

The coding sheet may seem long and confusing now, but after you are acquainted with the examination epochs, and learn how to identify and code the behaviours quickly, it will become easier to read and use.

The Marker System

As was mentioned earlier, the epochs have no time limit. Because the boundary of short-term memory is approximately thirty seconds without a distracter task, and considerably less than that with a distracter task, and as continued observation of a patient after a pain behaviour has been exhibited is a daunting distracter, a physical reminder system was deemed necessary to ensure the accuracy of the observations. Based on the concept of “Finger Math” in which children are taught to represent numbers using their fingers, the marker system developed for this project uses your nondominant hand to keep track of the pain behaviours exhibited during an epoch. Each digit of the nondominant hand represents a category of pain behaviour; when a pain behaviour is observed during an epoch, the finger that represents that pain behaviour is moved in a characteristic way, depending upon your preference. The thumb is reserved for the pain behaviour facial expression because the face is coded on a three-point scale, and, for most people, the thumb is the most flexible digit. So, depending upon the severity of the facial expression
observed, the thumb can be moved incrementally (no movement = 0, half-movement = 1, and full movement = 2).

Tip: When you first begin to learn this coding system, it may be beneficial to write the letters of each type of behaviour on to sticky dots and place them on your fingers.
Appendix E. Experimental Coding Form
<table>
<thead>
<tr>
<th></th>
<th>guarding</th>
<th>touching</th>
<th>words</th>
<th>sounds</th>
<th>facial expres.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td></td>
<td></td>
<td></td>
<td>1 2</td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td></td>
<td></td>
<td></td>
<td>1 2</td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td></td>
<td></td>
<td></td>
<td>1 2</td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td></td>
<td></td>
<td></td>
<td>1 2</td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td></td>
<td></td>
<td></td>
<td>1 2</td>
<td></td>
</tr>
<tr>
<td>6.</td>
<td></td>
<td></td>
<td></td>
<td>1 2</td>
<td></td>
</tr>
<tr>
<td>7.</td>
<td></td>
<td></td>
<td></td>
<td>1 2</td>
<td></td>
</tr>
<tr>
<td>8.</td>
<td></td>
<td></td>
<td></td>
<td>1 2</td>
<td></td>
</tr>
<tr>
<td>9.</td>
<td></td>
<td></td>
<td></td>
<td>1 2</td>
<td></td>
</tr>
<tr>
<td>10.</td>
<td></td>
<td></td>
<td></td>
<td>1 2</td>
<td></td>
</tr>
<tr>
<td>11.</td>
<td></td>
<td></td>
<td></td>
<td>1 2</td>
<td></td>
</tr>
<tr>
<td>12.</td>
<td></td>
<td></td>
<td></td>
<td>1 2</td>
<td></td>
</tr>
<tr>
<td>13.</td>
<td></td>
<td></td>
<td></td>
<td>1 2</td>
<td></td>
</tr>
<tr>
<td>14.</td>
<td></td>
<td></td>
<td></td>
<td>1 2</td>
<td></td>
</tr>
<tr>
<td>15.</td>
<td></td>
<td></td>
<td></td>
<td>1 2</td>
<td></td>
</tr>
<tr>
<td>16.</td>
<td></td>
<td></td>
<td></td>
<td>1 2</td>
<td></td>
</tr>
<tr>
<td>17.</td>
<td></td>
<td></td>
<td></td>
<td>1 2</td>
<td></td>
</tr>
<tr>
<td>18.</td>
<td></td>
<td></td>
<td></td>
<td>1 2</td>
<td></td>
</tr>
<tr>
<td>19.</td>
<td></td>
<td></td>
<td></td>
<td>1 2</td>
<td></td>
</tr>
<tr>
<td>20.</td>
<td></td>
<td></td>
<td></td>
<td>1 2</td>
<td></td>
</tr>
<tr>
<td>21.</td>
<td></td>
<td></td>
<td></td>
<td>1 2</td>
<td></td>
</tr>
<tr>
<td>22.</td>
<td></td>
<td></td>
<td></td>
<td>1 2</td>
<td></td>
</tr>
<tr>
<td>23.</td>
<td></td>
<td></td>
<td></td>
<td>1 2</td>
<td></td>
</tr>
<tr>
<td>24.</td>
<td></td>
<td></td>
<td></td>
<td>1 2</td>
<td></td>
</tr>
<tr>
<td>25.</td>
<td></td>
<td></td>
<td></td>
<td>1 2</td>
<td></td>
</tr>
<tr>
<td>26.</td>
<td></td>
<td></td>
<td></td>
<td>1 2</td>
<td></td>
</tr>
<tr>
<td>27.</td>
<td></td>
<td></td>
<td></td>
<td>1 2</td>
<td></td>
</tr>
<tr>
<td>28.</td>
<td></td>
<td></td>
<td></td>
<td>1 2</td>
<td></td>
</tr>
<tr>
<td>29.</td>
<td></td>
<td></td>
<td></td>
<td>1 2</td>
<td></td>
</tr>
<tr>
<td>30.</td>
<td></td>
<td></td>
<td></td>
<td>1 2</td>
<td></td>
</tr>
<tr>
<td>31.</td>
<td></td>
<td></td>
<td></td>
<td>1 2</td>
<td></td>
</tr>
<tr>
<td>32.</td>
<td></td>
<td></td>
<td></td>
<td>1 2</td>
<td></td>
</tr>
<tr>
<td>33.</td>
<td></td>
<td></td>
<td></td>
<td>1 2</td>
<td></td>
</tr>
<tr>
<td>34.</td>
<td></td>
<td></td>
<td></td>
<td>1 2</td>
<td></td>
</tr>
<tr>
<td>35.</td>
<td></td>
<td></td>
<td></td>
<td>1 2</td>
<td></td>
</tr>
<tr>
<td>36.</td>
<td></td>
<td></td>
<td></td>
<td>1 2</td>
<td></td>
</tr>
<tr>
<td>37.</td>
<td></td>
<td></td>
<td></td>
<td>1 2</td>
<td></td>
</tr>
<tr>
<td>38.</td>
<td></td>
<td></td>
<td></td>
<td>1 2</td>
<td></td>
</tr>
<tr>
<td>39.</td>
<td></td>
<td></td>
<td></td>
<td>1 2</td>
<td></td>
</tr>
<tr>
<td>40.</td>
<td></td>
<td></td>
<td></td>
<td>1 2</td>
<td></td>
</tr>
</tbody>
</table>
Appendix F. WCB Coding Manual Additions
Section 2. How to Code

If an epoch is skipped within the examination, the coder would place a check in the “did not do” box on the checklist for that epoch. If something unforeseen occurs during an epoch that prevents the accurate observation of the pain behaviours, the coder would mark off the “cannot code” box on the coding sheet for that epoch.

Also, there is no time restriction for the length of an epoch. The beginning and ending of each epoch is determined by the movement of the patient (there are a few exceptions to this rule, but they will be discussed separately).

Section 3. The Epochs: When to Code

Epochs are usually followed by a period of nonobservation to allow you to transcribe the observations for the preceding epoch. Moreover, occasional brief coding breaks have been built in to the protocol to allow some recuperative “downtime” as the total examination time of approximately thirty minutes is too long a period to maintain uninterrupted concentration.

The epochs, and the considerations one must take into account during each separate epoch, are listed below.

1. to sitting 1: This epoch begins as the patient enters the room, and ends when the patient is settled into a seated position. A patient is ‘settled’ if his or her buttocks have made full contact with the seat, and are bearing most, if not all, of the weight of the body. A clue to settling is the cessation of both downward movement and “getting comfortable” fidgeting.
2. introduction: This epoch is not coded.

3. to scale: This epoch is coded from the time that the patient begins to rise from the chair until the patient touches the scale with one foot.

4. weight/height: This epoch is not coded.

5. to landmarks: This epoch is coded from the time that the patient breaks physical contact with the scale completely (i.e. is no longer touching the scale) until the patient adopts a stationary standing position in preparation for the drawing of landmarks.

6. landmarks: This epoch is not coded.

7. lordosis: This epoch is not coded.

8. heel raise 1: This epoch is not coded.

9. heel raise 2: This epoch is not coded.

10. forward/back: Coding begins as the patient begins to perform the requisite movement, and ends when the patient has returned to the original, upright position. If
this position cannot be achieved, then the coding ends when the patient ceases to perform the movement.

11. rotation: This epoch is not coded

12. side-to-side: Coding begins as the patient begins to perform the requisite movement, and ends when the patient has returned to the original, upright position. If this position cannot be achieved, then the coding ends when the patient ceases to perform the movement.

13. SFM: This acronym means “stand for measure”, and is used to indicate times of stationary standing during which coding is not done.

14. lumbar extension 1: Coding begins as the patient begins to perform the requisite movement, and ends when the patient has returned to the original, upright position. If this position cannot be achieved, then the coding ends when the patient ceases to perform the movement.

15. SFM

16. lumbar flexion 1: Coding begins as the patient begins to perform the requisite movement, and ends when the patient has returned to the original, upright position. If
this position cannot be achieved, then the coding ends when the patient ceases to perform
the movement.

17. SFM

18. lumbar extension 2: This epoch is not coded.

19. SFM

20. lumbar flexion 2: This epoch is not coded.

21. SFM

22. lumbar extension 3: Coding begins as the patient begins to perform the requisite
movement, and ends when the patient has returned to the original, upright position. If
this position cannot be achieved, then the coding ends when the patient ceases to perform
the movement.

23. SFM

24. lumbar flexion 3: Coding begins as the patient begins to perform the requisite
movement, and ends when the patient has returned to the original, upright position. If
this position cannot be achieved, then the coding ends when the patient ceases to perform the movement.

25. SFM

26. lateral extension left 1: Coding begins as the patient begins to perform the requisite movement, and ends when the patient has returned to the original, upright position. If this position cannot be achieved, then the coding ends when the patient ceases to perform the movement.

27. SFM

28. lateral extension right 1: Coding begins as the patient begins to perform the requisite movement, and ends when the patient has returned to the original, upright position. If this position cannot be achieved, then the coding ends when the patient ceases to perform the movement.

29. SFM

30. lateral extension left 2: This epoch is not coded.

31. SFM
32. lateral extension right 2: This epoch is not coded.

33. SFM

34. lateral extension left 3: Coding begins as the patient begins to perform the requisite movement, and ends when the patient has returned to the original, upright position. If this position cannot be achieved, then the coding ends when the patient ceases to perform the movement.

35. SFM

36. lateral extension right 3: Coding begins as the patient begins to perform the requisite movement, and ends when the patient has returned to the original, upright position. If this position cannot be achieved, then the coding ends when the patient ceases to perform the movement.

37. SFM

38. axial compression: This epoch is coded differently than most of the other epochs. It begins and ends based on the actions of the examiner. As soon as the examiner places both hands upon the patient's head, you should begin coding. When the examiner's hands are removed, coding should cease. Please note that due to the nature of the examiner's questions during this epoch, the production of any example of the behaviour
‘Words’ must be disregarded. The appropriate box on the coding sheet has been darkened as a reminder.

39. simulated rotation : As with ‘axial compression’, this epoch must be coded differently. It begins when the examiner begins to rotate the patient’s body, and ends when the examiner removes his or her hands from the patient’s arms. Please note that due to the nature of the examiner’s questions and the physical constraints present during this epoch, the production of any example of the behaviours ‘words’ or ‘touching’ must be disregarded. The appropriate boxes on the coding sheet are darkened as a reminder.

40. to kneeling : This epoch is not coded.

41. ankle reflexes : This epoch is not coded.

42. to sitting 2 : This epoch begins when the patient is clear of the chair (i.e. is no longer touching the chair), and ends when he or she is settled on the examination bed.

43. knee reflexes : This epoch begins when the patient makes the first move to clasp their hands together

44. knee extension : The beginning of this epoch is not dependent upon the patient’s movement, but on the actions of the examiner. It begins when the examiner raises the patient’s leg, and ends when the leg is returned to the original, lowered
position. Please watch for any overt verbal pain probes from the examiner during this and the next epoch. As in all epochs, if one occurs, and then the patient gives some verbal indication of pain, then ‘Words’ cannot be coded.

45. muscle strength 1: This epoch begins when the patient raises his or her leg and does not end until the leg is lowered.

46. muscle strength 2: This epoch begins when the patient raises his or her leg and does not end until the leg is lowered.

47. to supine: This epoch begins when the patient first shifts position in order to lie down (it does not have to be a large movement), and does not end until the patient is fully settled on his or her back.

48. ankle dorsiflexion 1: This epoch is not coded

50. toe extensor 1: This epoch is not coded

51. ankle dorsiflexion 2: This epoch is not coded

52. toe extensor 2: This epoch is not coded

53. thigh & calf muscle bulk: This epoch is not coded.
54. sensation : This epoch is not coded.

55. passive SLR 1 : The beginning of this epoch is not dependant upon the patient’s movement, but on the actions of the examiner. It begins when the examiner raises the patient’s leg, and ends when the leg is returned to the original, lowered position.

56. passive SLR 2 : The beginning of this epoch is not dependant upon the patient’s movement, but on the actions of the examiner. It begins when the examiner raises the patient’s leg, and ends when the leg is returned to the original, lowered position.

57. to prone : This epoch begins as the patient first moves to turn over, and ends when the patient is fully settled.

58. palpation : This epoch is not coded.

59. MacKenzie push-up : This epoch begins as the patient begins the upwards motion necessary to complete the movement, and ends when the patient has returned to the resting position (one possible clue is that the patient’s head has made contact with the table, and all stress from the neck is gone). Please note that you cannot use space under the patient’s upper body as an indicator of this movement, as many people will not necessarily clear the table.
60. prone active extension: This epoch begins as the patient begins the upwards motion necessary to complete the extension, and ends when the patient has returned to the resting position (one possible clue is that the patient's head has made contact with the table, and all stress from the neck is gone). Please note that you cannot use space under the patient’s upper body as an indicator of this movement, as many people will not necessarily clear the table.

61. to supine: This epoch begins as the patient first moves to turn over, and ends when the patient is fully settled.

62. active sit-up: This epoch begins as the patient begins the upward curl of the upper body, and ends as the patient returns to the original position (one possible clue is that the patient’s head has made contact with the table, and all stress from the neck is gone).

63. bilateral active SLR: This epoch begins as the patient’s legs are lifted upwards, and ends as the legs are returned to their original, resting position. Please note: gripping the examination table at the suggestion of the examiner should not be counted as ‘Guarding (bracing)’.

64. to standing: This epoch begins as the patient first moves to rise from the supine position, and ends as the patient is standing on both feet, and is clear of the table.
Appendix G. WCB Coding Form
<table>
<thead>
<tr>
<th>Epoch</th>
<th>didn't do</th>
<th>cannot code</th>
<th>guard</th>
<th>touch</th>
<th>words</th>
<th>sounds</th>
<th>face</th>
</tr>
</thead>
<tbody>
<tr>
<td>to sitting 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>introduction</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>to scale</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>weight/height</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>to landmarks</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>landmarks</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lordosis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>heel raise</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>forward/back</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rotation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>side-to-side</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SFM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lumbar extension 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lumbar flexion 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lumbar extension 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lumbar flexion 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lumbar extension 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lumbar flexion 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lumbar extension 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lumbar flexion 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lumbar extension 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lumbar flexion 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lateral extension left 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lateral extension right 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lateral extension left 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lateral extension right 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lateral extension left 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lateral extension right 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lateral extension left 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lateral extension right 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lateral extension left 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lateral extension right 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>axial compression</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>simulated rotation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>to kneeling</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ankle reflexes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>to sitting 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>knee reflexes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>knee extension 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>knee extension 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------------------------------</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>muscle strength 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>muscle strength 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>to supine</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ankle/toe testing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>thigh &amp; calf muscle bulk</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sensation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>passive SLR 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>passive SLR 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>to prone</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>palpation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>McKenzie push-up</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>prone active extension</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>to supine</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>active situp</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bilateral active SLR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>to standing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: