

**Assessment of Perceived Functional Capacity:
Using Rasch Analysis to Evaluate the Measurement Properties
of Four Perceived Pain & Disability Scales**

Lois Lochhead

B.S.R. University of British Columbia, 1984

Thesis Submitted in Partial Fulfillment of

The Requirements for the Degree of

Master of Science

In

Community Health Sciences

The University of Northern British Columbia

August 2009

© Lois Lochhead, 2009



Library and Archives
Canada

Published Heritage
Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file *Votre référence*
ISBN: 978-0-494-60828-9
Our file *Notre référence*
ISBN: 978-0-494-60828-9

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.

AVIS:

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protègent cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

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

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

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

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

■❖■
Canada

Abstract

Functional Capacity Evaluations (FCEs) include comparisons of self-report and performance-based measures. A difference in the two scores can be interpreted as symptom magnification which can impact eligibility for benefits. FCEs typically include scales such as the Oswestry Disability Index (ODI), the Dallas Pain Questionnaire (DPQ), the Spinal Function Sort (SFS) and the Neck Disability Index (NDI). Rasch Modeling was used to evaluate their original classification categories. Examination included fit of data to model expectations, threshold ordering of items, differential item functioning and item difficulty. None of the scales demonstrated unidimensionality. For the ODI and DPQ, rescaling and/or eliminating items improved the scales. The SFS is not a unidimensional scale and demonstrates differential item functioning. The NDI demonstrates unidimensionality when two of the items are eliminated but disordered thresholds could not be fixed. Health Professionals using these measures should be aware that these scales do not perform as well as expected.

TABLE OF CONTENTS

ABSTRACT.....	ii
TABLE OF CONTENTS.....	iii
Index of Tables	vi
Table of Figures.....	vii
Glossary	viii
Acknowledgement	x
CHAPTER ONE: INTRODUCTION.....	1
CHAPTER TWO: LITERATURE REVIEW	8
Instruments	10
Oswestry Disability Index (ODI)	10
Dallas Pain Questionnaire (DPQ).....	15
Spinal Function Sort (SFS).....	19
Neck Disability Index (NDI)	23
Measurement Theory.....	27
Item Response Theory (IRT).....	28
Rasch Analysis	31
Dichotomous Model.....	32
Rating Scale Model.....	33
Partial Credit Model.....	35
Research Questions	36
Significance of Proposed Study	37

CHAPTER THREE: DESIGN AND METHODOLOGY	38
Subjects	38
Instrumentation.....	39
Procedures	40
Data Analysis	40
CHAPTER FOUR: RESULTS	42
The Oswestry Disability Index	42
Diagnostic Measures	43
Item Difficulty	50
Differential Item Functioning (DIF)	52
Dallas Pain Questionnaire (DPQ).....	53
Diagnostic Measures for the DPQ	54
Item Difficulty	58
Differential Item Functioning (DIF)	59
PACT Spinal Function Sort.....	60
Diagnostic Measures for the SFS.....	61
Item Difficulty	62
Differential Item Function (DIF) for the SFS	63
Neck Disability Index.....	66
Diagnostic Measures for the NDI	67
Item Difficulty	70
Differential Item Functioning	71
CHAPTER FIVE: DISCUSSION.....	72
The Oswestry Disability Index (ODI)	72

Dallas Pain Questionnaire (DPQ).....	75
Spinal Function Sort (SFS).....	76
Neck Disability Index (NDI)	77
Conclusion.....	78
Limitations of Design.....	78
Recommendations for Practitioners	78
Recommendations for Future Research	79
Oswestry Disability Index.....	79
Dallas Pain Questionnaire.....	79
Spinal Function Sort	79
Neck Disability Index	79
Appendix 1 – Oswestry Disability Index – Version 1.0	86
Appendix 2 – Dallas Pain Questionnaire	88
Appendix 3 – Spinal Function Sort Instructions, Sample & Score Sheet.....	90
Appendix 5 – Letter of Consent.....	95
Appendix 6 – Consent to Evaluate.....	96
Appendix 7 – DOT Physical Demand Characteristics of Work	97
Appendix 8 – Oswestry Disability Index – Rescaled	98

Index of Tables

Table 1	<i>Physical Occupational Demands</i>	39
Table 2	<i>Diagnostic Measures for the Oswestry Disability Index</i>	44
Table 3	<i>Oswestry Categories Collapsed</i>	48
Table 4	<i>Diagnostic Measures for Rescaled Oswestry Disability Index</i>	49
Table 5	<i>Diagnostic Measures for the Dallas Pain Questionnaire</i>	54
Table 6	<i>Spinal Function Sort Misfitting Items</i>	61
Table 7	<i>Diagnostic Measures for NDI</i>	67

Table of Figures

<i>Figure 1</i>	Age distribution of study sample	38
<i>Figure 2</i>	Standing item #6 Oswestry Disability Index.	45
<i>Figure 3</i>	Pain item #1 Oswestry Disability Index	46
<i>Figure 4</i>	Rescaled pain item #1	50
<i>Figure 5</i>	Item Characteristic curves depicting item difficulty on the rescaled ODI.....	51
<i>Figure 6</i>	DIF by Gender for the ODI.....	52
<i>Figure 7</i>	DPQ Item 3 Lifting Category Probability Curve	55
<i>Figure 8</i>	Category probability curve of DPQ item 10 “Vocational”	56
<i>Figure 9</i>	Category probability curve for the 4-point DPQ rating scale	58
<i>Figure 10</i>	Item characteristic curves depicting item endorsement for the 4-point DPQ.....	59
<i>Figure 11</i>	Item Person Map for Spinal Function Sort	62
<i>Figure 12</i>	Spinal Function Sort Item DIF size by gender.....	64
<i>Figure 13</i>	Category Probability Curves NDI “Lifting” item #3	68
<i>Figure 14</i>	Category probability curves for the NDI “Headache” item #5	69
<i>Figure 15</i>	Item Difficulty for the NDI.....	70
<i>Figure 16</i>	DIF by Gender for the NDI.....	71

Glossary

Ability Estimate	The location of a person on a variable, inferred by using the collected observations.
Calibration	The process of estimating item difficulty/person ability by converting raw scores to logits on a measurement scale.
DIF	Differential Item Functioning is the loss of invariance of item estimates across testing situations such as when an item functions differently with men and women. DIF is evidence of item bias.
Infit Mean Square	Indicates degree of fit of an item or person to the Rasch Model and is a transformation of the residuals, the difference between predicted and observed. Expected value of 1 with ranges from .6 to 1.4 deemed acceptable for rating scale survey items. Infit statistic is more sensitive to inlier patterns i.e. unexpected response patterns by persons on items that are targeted on them.
Item Separation Index	An estimate of the spread of items on a measure variable expressed in standard error units i.e. the adjusted item standard deviation divided by the average measurement error.
Latent Trait	Attribute of an individual that can be inferred from observation of behavior.
Logit	The unit of measurement resulting from the transformation of raw scores from ordinal data to log-odds ratios on a common interval scale. The log-odds of an event is the logit of the probability of the event.
Outfit Mean Square	Unstandardized estimates of degree of fit that are more sensitive to outliers – unexpected responses by persons on items that are distant to the subject's ability. Values of .6 to 1.4 are acceptable for rating scale items.

Partial Credit Model	Masters' Rasch Model for polytomous data which allows the item categories and/or threshold values to vary from item to item.
Person Separation Index	Estimate of the spread of persons on the measured variable expressed in standard error units.
Rating Scale Model	Andrich's Rasch Model for polytomous data generated from Likert scales. It applies one set of threshold values to all items on the test.
Threshold	The point of equal probability of adjacent categories where the likelihood of not endorsing the item turns to the likelihood of endorsing the item.
ZSTD	Tests the significance of a particular mean square value. Values from -2.0 to +2.0 are acceptable.

Acknowledgement

It is not often that we are given a second chance to fulfill a dream. I am truly grateful for this opportunity and it is a pleasure to thank those who made this possible.

Dr. Peter MacMillan, with his undying enthusiasm for all things Rasch, has been the ideal supervisor. His sage advice, insightful criticisms and patient encouragement aided the writing of this thesis from the formative stages to the final draft. I could not have done it without him.

My committee members, Dr. Henry Harder and Dr. Ken Prkachin, kindly made time in their busy schedules to review the thesis and to prepare for the defense on short notice. Dr. Saif Zahir, generously offered his services as an external examiner so the defense could go forward. I am indebted to all of you.

My husband, Chuck Attwater, patiently allowed me to study at the expense of household chores, social and family obligations as well as companionship; all of which only love could endure. In addition, he proof-read the final text, eliminated numerous errors and made valuable suggestions to improve the clarity of the work.

My sons, Anthony and Patrick Daniele, have believed and trusted in me and have given their love, understanding and companionship throughout the years as we navigated some difficult waters. To them, I dedicate this thesis.

“Lately it occurs to me...What a long strange trip it’s been.” (Grateful Dead, 1970)

CHAPTER ONE: INTRODUCTION

Determination of an individual's readiness to return to work following injury or illness often involves having the individual participate in a Functional Capacity Evaluation (FCE) or Work Capacity Evaluation (WCE). These two terms are used interchangeably within the rehabilitation literature. Isernhagen (1995) defines Functional Capacity Evaluation as follows:

FCE is a standardized battery of clinical tests that purport to measure a patient's safe physical ability for work-related activity. Physical capacity as found in the FCE testing is compared to required physical job demands of the patient's occupation. Critical job demands are assessed by a job analysis involving collecting relevant information by either direct observation, an interview with employer or employee, or existing job descriptions. (p.410)

FCEs are typically performed by Physical Therapists or Occupational Therapists who have specialized training and certification in the administration of the test batteries that make up the evaluation. One such certification is the Certified Work Capacity Evaluator (CWCE) designation available through Roy Matheson and Associates. The training consists of a five day education program where therapists are taught the protocols for administration and scoring of the individual tests. Evaluation of perception of ability, level of pain, physical effort, musculoskeletal evaluation, mobility, positional tolerances, dexterity, cardiovascular fitness and material handling are all included in the five days of training. There is little or no discussion about the psychometric properties of the tests; the focus is on proper administration. The participants in the training are assured that each test has undergone extensive reliability and validity testing.

The referring agencies (Worker's Compensation Board (WCB), Insurance Corporation of British Columbia (ICBC), Long Term Disability carriers, lawyers, etc.) usually provide written questions to the therapist who will be performing the FCE. Matheson (2006) gives the following examples of standardized referral questions:

- A. Did the client demonstrate full physical effort during the evaluation?
 - B. Are the client's subjective reports reliable?
 - C. Is the client able to return to work at this time?
 - D. If unable to return to his/her usual and customary job: What physical deficits hinder the worker's ability to return to work at this time? What modifications are needed to return to modified work? Which rehabilitation options exist at this time?
 - E. What are the client's current functional abilities?
 - F. What is the client's loss of function as compared to pre-injury ability?
 - G. Would the client benefit from additional rehabilitation services at this time?
- (p. 6)

To answer question B above, one or more pen and paper tests such as the Oswestry Disability Index (ODI), the Dallas Pain Questionnaire (DPQ), the Performance Assessment and Capacity Testing (PACT) Spinal Function Sort (SFS) and the Neck Disability Index (NDI) are completed by the client. The scores on these questionnaires are compared with the client's actual performance during the evaluation.

All of these measures use a Likert-type or rating scale to score each item. The Dallas Pain Questionnaire uses a Visual Analogue Scale of varying lengths for each item whereas the other three instruments (ODI, SFS, and NDI) use scales of a consistent length for each item. For instance, in the PACT Spinal Function Sort (SFS) the items are scored on a "1-5"

Likert-type scale. The range is from “1” (Able) to “5” (Unable) with “2”, “3” and “4” being used to depict abilities in the range between unable and able. The rating of “2” relates to a mild degree of restriction of ability, “3” relates to moderate restriction of ability and “4” relates to significant restriction of ability on the task. Scoring is done by adding the responses in each column and multiplying the number of “1” responses by four, “2” responses by three, “3” responses by two and “4” responses by one. No points are awarded for “5” or for items where the individual selected “?”. This presumes that selection of “4” - significant restriction indicates twice as much restriction as a selection of “2” - mild restriction. It also means that an answer of “don’t know” is equivalent to an answer of “unable”. The ODI and NDI offer a set of six statements for each response. The respondent checks off the statement that most represents his/her feelings for each item. Each of the statements was designed to sequentially represent more disability than the previous statement. These four measures have been developed using classical test theory (CTT).

Measurement Theory

The definition of measurement as the “assignment of numerals to objects or events according to some rule” as proposed by Stevens (1946) has been widely adopted in the human sciences and forms the basis of questionnaire development and analysis. Responses are summed to form a score that represents the individual’s level of ability or agreement using classical test theory (CTT). CTT is based on the classical true score model as outlined by Crocker & Algina (1986) who say that “any observed test score could be envisioned as the composite of two hypothetical components – a true score and a random error component” (p. 66).

This is expressed as

$$X = T + E$$

where X is the observed test score and T is the individual's true score and E is the error that occurs between the true score and the observed score for that individual on the given test. CTT examines the success rate of a group of examinees on an item. The success rate of this group on any given item is known as the p value of the item and is used as the measurement of item difficulty. The higher the p value, the easier the item is to endorse. Item discrimination refers to the ability of an item to discriminate between higher and lower levels of the ability or trait we are attempting to measure. This is often expressed as the Pearson product-moment correlation coefficient (r) between the scores on the item and the total scores on the test. Alternatively a discrimination index, D , is defined as the difference between endorsement/difficulty values of a high trait group and a low trait group. Fan (1998) summarized the limitations of CTT as follows:

The major limitation of CTT can be summarized as circular dependency: (a) The person statistic (i.e., observed score) is (item) sample dependent, and (b) the item statistics (i.e., item difficulty and item discrimination) are (examinee) sample dependent. This circular dependency poses some theoretical difficulties in CTT's application in some measurement situations such as test equating, computerized adaptive testing (p.357).

This means that the item parameters change even in their order of difficulty/ease of endorsement depending on the population to which it is administered; the standard error of measurement (SEM) is the same for all scores and reliability changes with population. This limits the ability to generalize from one population to another. Each of the four measures, the DPQ, ODI, SFS, and NDI has been extensively researched in terms of reliability and validity using CTT statistical analyses with the limitations as described above.

Inherent in these four measures is a scoring scale that assumes an equal interval between each score and that each item contributes equally to the total score. While these tools are built from ordinal items, the scoring is done in a manner that would assume the data is interval in nature. Bond and Fox (2007) point out that:

while classification and seriation are necessary precursors to the development of measurement systems, they are not sufficient for measurement. The distinctive attribute of a measurement system is the requirement for an arbitrary unit of difference that can be iterated between successive lengths. (p. 4)

In contrast to CTT, Item Response Theory (IRT), also known as latent trait theory, focuses on the item level information as opposed to test level information. IRT is applied in the development and refining of measurement instruments as well as for equating tests. The probability of a correct response to an item is expressed in terms of person and item parameters. Person parameters may be the ability of the individual or the strength of his convictions. Item parameters include difficulty, and in some models, discrimination and pseudo-guessing. Items may be in the form of right/wrong responses, statements that relate to level of agreement or presence/absence/degree of symptoms. IRT models scale the ability of examinees and item difficulty on the same metric allowing meaningful comparison of an item and the ability of the person.

Within IRT there are one, two and three parameter logistic models. The three-parameter (3 PL) model is so named because it employs three item parameters – item difficulty, item discrimination and pseudo-guessing. It is a logistic model as it converts the raw score summary into its natural logarithmic odds ratio to produce a linear (interval) measure. The two-parameter (2 PL) model assumes minimal guessing but items can vary in terms of difficulty and discrimination. The one-parameter model assumes that there is

minimal guessing and equivalent item discrimination so that items are only described by a single parameter – item difficulty. (Crocker & Algina, 1986)

Karabatsos (1999) reported to the 32nd annual conference of the Society for Mathematical Psychology that:

There is strong support that almost 100% of the time, the parameters of the 2 PL and 3 PL violate interval scaling. On the other hand, the theoretical probabilities of the Rasch models will always support a stable, interval scale structure. (p. 18)

There are two approaches to evaluate the single parameter of item difficulty; the One-parameter Logistic model (Birnbaum, 1968) and the Rasch Model (Rasch, 1960). For practical purposes, when the person sample is parameterized by a mean and standard deviation for item estimation, it is a 1 PL IRT model. When each individual in the person sample is parameterized for item estimation, it is Rasch. While the 1 PL model is primarily a descriptive, computationally simpler approximation to the Normal Ogive Model of L.L. Thurstone (1927), as developed by Lord (1952), the Rasch Model is prescriptive offering distribution-free person ability and item difficulty estimates on a linear latent variable.

The Rasch model was developed by Georg Rasch (1901-1980), a Danish mathematician whose initial work in the field was done in the field of educational measurement. For several decades these methods have been applied in the health care field to improve the psychometric reliability and validity of self-report test batteries. Rasch analysis facilitates the calibration of ordinal measures to interval measures and therefore can improve confidence in scores obtained on these self-report tests.

Rasch analysis is also useful for equating two or more instruments that purport to measure the same construct. By equating instruments, it can be determined if the instruments

measure the same construct and if so, do they measure it at the same level. Analysis with the Rasch model allows the researcher to order persons according to their perceived level of the latent trait and items according to their perceived difficulty/ease of endorsement. From this analysis, a method of scoring may be developed that improves the sensitivity and specificity of the tests. If all tests measure the same construct equally, a case can be made for administering only one of the tests to save time in the evaluation. It may also be determined that certain of the tests are more effective at measuring certain populations thus assisting the examiner with test selection.

The purpose of this study is to evaluate the psychometric properties of the four instruments – the Oswestry Disability Index (ODI), The Dallas Pain Questionnaire (DPQ), the PACT Spinal Function Sort (SPS) and the Neck Disability Index (NDI) using Rasch Analysis to establish evidence for reliability and validity of each instrument as well as to attempt to equate the four instruments to evaluate the abilities of the persons and difficulty of the items. This is important as these measurement instruments are currently being used by Work Capacity Evaluators to determine reliability of the client's subjective reports as compared to demonstrated abilities. This practice can potentially affect the continuation of disability benefits and allotment of funds for rehabilitation as well as monetary awards in litigation. Since these instruments are frequently used as pre and post treatment measurements, efficacy of treatment is often established using a change in score on the questionnaire as the indicator. Improving the sensitivity and specificity of these instruments will, in turn, improve the accuracy of measurement of the trait in question.

CHAPTER TWO: LITERATURE REVIEW

Thew (2007) identified:

a multitude of purposes for which Functional Capacity Evaluations are completed in Prince George. These include: (a) determining safe return to work to current employment; (b) pre-employment physical ability assessments; (c) determining current functional abilities or level of functioning or physical abilities; (d) comparing functional abilities to job demands; (e) developing a rehabilitation treatment plan; (f) assisting with return to work planning either to current occupation or alternative occupation; and (g) determining if evaluatees are accurately reporting their abilities. (p. 20).

Self report measures are used as part of Functional Capacity Evaluations to measure the individual's perception of pain and disability and the results are compared to clinical examination and functional testing scores. This comparison gives the evaluator insight into the accuracy of the individual's perception i.e. if they are magnifying or minimizing their abilities/symptoms. Clinicians have come to trust the reliability and validity of these self report measures. In the study by Thew (2007) in the transcript of the interview, the Clinician was quoted as saying:

We try to use the forms as much as possible. The research that was done to validate these questionnaires can back up that they're valid and reliable. If the person says they can't do this and then they demonstrate it, you know that the questionnaire is a valid questionnaire. (p. 52)

In a case such as the one described above by the Clinician, the individual might be considered to be magnifying his symptoms. Symptom Magnification Syndrome is a term which was coined by Leonard Matheson (1988) and it refers "to the conscious or sub-conscious tendency of an individual to under-rate his/her abilities and/or over-state his/her limitations" (p.11). An opinion rendered by a professional indicating symptom

magnification can have tremendous negative consequences for a disabled person including loss of access to rehabilitation and loss of financial support. Conversely, if the individual is minimizing his symptoms and he overestimates his abilities compared to his true abilities, he can be returned to work and injure himself or others. For these reason, perceived functional capacity is an important part of every Functional Capacity Evaluation.

The Matheson System for Functional Capacity Evaluation (2006):

embodies the professional value system endorsed by the American Psychological Association, American Physical Therapy Association, and the National Institute of Occupational Safety and Health (NIOSH), which states that each evaluation must address these five hierarchical components: Safety, Reliability, Validity, Practicality and Utility.” (Chapter 2 – page 3).

Since the FCE is a battery of tests including the self report measures such as the Oswestry Disability Index (ODI), the Dallas Pain Questionnaire (DPQ), the PACT Spinal Function Sort (SFS) and the Neck Disability Index (NDI) each of these measures needs to satisfy the components as outlined above.

Instruments

Oswestry Disability Index (ODI)

The Oswestry Disability Index (ODI), see Appendix 1, is one of the most commonly used condition-specific assessments. The ODI is a ten item scale that measures pain intensity, personal care, standing, sleeping, lifting, walking, sitting, sex life, social life and ability to travel. Each item includes six potential responses such that each response is presumed to describe a greater degree of disability ranging from no disability to total disability. Individual items are scored from “0” to “5” and then summed and doubled to obtain a percentage score. Fairbank, Couper, Davies and O’Brien, (1980) designated five categories to interpret the Oswestry score. Low percentage scores represented less disability, whereas higher percentage scores indicated more disability. These five categories of disability are 0% to 20%, minimal; 20% to 40%, moderate; 40% to 60%, severe; 60% to 80%, crippled; and 80% to 100%, bed bound or exaggerating. The overlap at the transition points of each category are concerning; a score of 20% could relate to minimal or moderate disability. Despite the wide use of this scale and the many validation studies that have been done, this issue has not been addressed. Categories of disability provide the context from which to interpret a score and this can be important in the awarding of disability benefits. The scores are frequently used in case studies and clinical research as a reference point to interpret outcomes.

The Oswestry Disability Index is also cited in the literature as the Oswestry Pain Questionnaire and the Oswestry Low Back Pain and Disability Index, the latter being its original name (Fairbank et al., 1980). The name has now been shortened to the Oswestry

Disability Index (ODI). It has become one of the most frequently used outcome measures for spinal disorders. A Medline review done on October 25, 2008 using a simple name search revealed 402 hits in the database compared to 2 hits for the Spinal Function Sort, 33 hits for the Dallas Pain Questionnaire and 175 hits for the Neck Disability Index. There are several versions (1.0, 1.1 and 2.0) of the ODI with small wording differences. There is also a revised version developed by a chiropractic study group in the United Kingdom with the intent to improve the sensitivity of the scale for less disabled persons (Hudson-Cook, Tomes-Nicholson & Breen, 1989). In the revised version, Section 8, the sex life question was omitted and a new section was added that was called “changing degree of pain” – this related to whether the pain was improving or getting worse overall. The original version focused on current pain levels not changes in pain. Several other sections were reworded. For instance in Section 4 – *Walking* - the ODI version 2.0 offered the following choices:

1. Pain does not prevent me walking any distance.
2. Pain prevents me walking more than 1 mile
3. Pain prevents me walking more than ½ of a mile.
4. Pain prevents me walking more than 100 yards.
5. I can only walk using a stick or crutches
6. I am in bed most of the time and have to crawl to the toilet.

The Revised Oswestry Disability Index offered these choices instead:

1. I have no pain on walking.
2. I have some pain with walking but it does not increase with distance.
3. I cannot walk more than 1 Mile without increasing pain.
4. I cannot walk more than ½ Mile without increasing pain.
5. I cannot walk more than ¼ Mile without increasing pain.

6. I cannot walk at all without increasing pain.

Fairbank and Pynsent (2000) had this to say about the revised version:

Its (the Revised Oswestry Disability Index) objective was to increase the sensitivity of the scale for less disabled patients, but it confuses impairment with disability...In the authors' view, this version is not acceptable, because it confuses impairment questions with disability questions. Its wording is often complex, and some sections do not allow for no symptoms. It allows a measurement of changing symptoms, however.

The statements "Pain does not prevent me walking any distance" from the ODI 2.0 and "I have no pain on walking" as used in the Revised ODI are not equivalent in my view. One statement refers to how pain affects the individual and the other refers to the amount of pain the person is experiencing. An individual could have pain with walking but not let that pain prevent him or her from walking any distance. Also the increments between distances walked are different. With the utilization of the metric system in Canada, younger respondents may have difficulty with distances in Imperial units.

All versions are scored in the same manner for a score out of 50 possible points. The score is then doubled to provide a percentage which relates to the amount of perceived disability. An individual taking the Revised ODI might have a score that indicates a higher or lower level of disability than a similar individual would score on the ODI. The Matheson Functional Capacity Evaluation Certification Program uses only the ODI 1.0 and therefore only this version will be discussed with regard to development, testing, reliability and validity.

Development of the Oswestry Disability Index

In 1976, the development of this condition specific outcome measure was initiated by John O'Brien. An orthopedic surgeon, Stephen Einstein and an occupational therapist, Judith Cooper interviewed patients with back pain and from responses obtained regarding limitations to activities of daily living, several drafts of the questionnaire were tried. The final version was tested on 22 patients with a one day test-retest reliability of $r = .99$, $p < .001$ (Fairbank et al., 1980). This high correlation may contain a memory effect and in fact subsequent studies showed lower correlations when the test-retest interval was increased. With a four day interval, Kopec et al. (1996) reported a reliability of $r = .91$ and with a 1 week interval, Grönblad et al. (1993) found a reliability of $r = .83$. With longer intervals between tests, natural symptom fluctuation may influence the results.

Within CTT internal consistency is commonly calculated using Cronbach's α which is defined as:

$$\alpha = \frac{N}{N-1} \left(1 - \frac{\sum_{i=1}^N \sigma_{Y_i}^2}{\sigma_X^2} \right)$$

where N is the number of items, σ_X^2 is the variance of the observed total test scores, and $\sigma_{Y_i}^2$ is the variances for the N individual items (Cronbach, 1951). When α is large it can be assumed that the total score is a reasonable representation of the individual item scores. Strong, Ashton and Large (1994) reported an internal consistency for the ODI of $\alpha = .71$ on a sample of 100. The sample size of the original study was 22 so this sample size is better. Given that the formula for standard error is σ/\sqrt{n} , where sigma is the standard deviation

and n is the number of participants, it is easy to understand that the larger the sample the greater precision there is in the results.

One validation study of the ODI was reported by Fairbank et al. in 1980. They evaluated a group of 25 individuals with a first episode of low back pain over a three week period. Since this was a first episode of low back pain, there was a reasonable expectation of improvement over the three week period. They reported that the results of a t-test using the scores at the beginning and end of the three week period were significant at a level of $p < .05$. Overall the percentage of disability had dropped by 28%. Again the sample size is small allowing a greater chance of error. The sample is not representative of all back pain patients which is the population it is currently used to assess.

Beurskens, de Vet, & Köke (1996) performed a study with 81 subjects who had suffered with non-specific back pain for at least 6 weeks. These subjects were tested with the ODI before and after treatment. Using external criterion for improvement, it was determined that 38 of the subjects had improved. The effect size was calculated to be 0.8. Grönblad et al. (1993) showed a moderate correlation of $r = .62$ with the Huskisson's (1974) Visual Analogue Scale, an established pain measure, on a sample of 94 patients.

Following these studies, the ODI became the "gold standard" in the rehabilitation field. It was then used to validate other instruments such as the Pain Disability Index (Pollard, 1984), the Low Back Outcome Score (Greenough & Fraser, 1992) and many others. The issues with small sample size in the original studies seem to have been overlooked. Additionally, the use of differing populations to validate the instrument i.e. patients presenting with a first episode of back pain vs. non-specific back pain of more than 6 weeks duration limit the generalizability of the results due to the circular dependency inherent in

CTT. Fairbank and Pynsent (2000) proposed that “The wide use of the ODI is part of the validation process.” Thom Walsh (2000) responded to this claim as follows:

The thought that wide use and reasonable performance as expected on a small sample are synonymous with validation and a rigorous review is one that falls short of current capabilities in the field. It should no longer be enough to simply report findings that turned out as expected, or that a gold-standard measure is crowned as a result of widespread use. Good validation studies should state a clear hypothesis and test it using a rigorous design and statistical analysis. This review article nicely compiles a wide range of work utilizing the ODI over the past 20 years. While the breadth of this compilation is notable, and the validation steps taken at various times have raised interesting questions, it has not, in my opinion, established a gold-standard measure. (p. 2953)

The current researcher’s enthusiastic endorsement of Walsh’s (2000) views should be recognized as the *raison d’être* for this study. This is not simply for investigation of the ODI but for other instruments as well.

In addition to the reliability and validity issues, the ODI consists of ordinal items totaled as a sum of equal-valued items to produce a disability rating. However, item values have been assigned rationally rather than empirically resulting in total scores that do not proportionally indicate the trait. There have been many changes in the field of measurement since the Oswestry Disability Index was first developed. Applying these new techniques such as Rasch analysis can improve the ability of the ODI to assess self-reported disability levels for this instrument as well as for the DPQ, SFS and NDI.

Dallas Pain Questionnaire (DPQ)

Lawlis, Cuenas, Selby and McCoy (1989) developed the Dallas Pain Questionnaire (DPQ) to measure the impact of chronic spinal pain on four aspects (daily and work-leisure activities, anxiety-depression, and social interest) of a respondents’ life; see Appendix 2 for the full questionnaire. The DPQ is considered to measure two factors: Functional Activities

and Emotional Capacities. For this study only the items that report on the Functional Activities Factor are included. This relates to sections I to X which includes Pain Intensity, Personal Care, Lifting, Walking, Sitting, Standing, Sleeping, Social Life, Travelling and Vocational. It should be noted that a factor analysis conducted by Lawlis et al., (1989) showed loadings of .495 and .202 for the pain item on the Functional Activities Factor and the Emotional Capacities Factor respectively. Other item factor loadings were .615 and above. Correlation studies of the item scores with total scores showed that the pain item correlated .65 with daily activities and .52 with work/leisure activities ($p < .0001$). Correlation coefficients for all the other items were .78 and higher.

Each item is scored on a visual analog scale (VAS). The standard VAS is a 10 cm continuous line between two points where the respondents indicate their level of agreement to a statement by indicating a position on the line. The score is then determined by measuring the distance from the end point to the indicated position. In the DPQ discrete values are created for each item by adding breaks delineated by “:” to the scale. The scales are anchored at the beginning with words such as “I can lift as much as I did” and at the end with words such as “I cannot lift at all”. The respondents indicate their level of agreement by placing an “x” in one of the delineated segments. The length and number of segments of each scale varies – Sections I –VI are six units in length, VII is five units in length. The score is added for Sections 1-VII and multiplied by three to obtain a percentage score for that aspect (daily activities) of the Functional Activities Factor. Similarly for the second aspect, work-leisure activities, the scores on the VIII to X items are added. These items are eight, seven and eight units in length, scored “0-7”, “0-6” and “0-7” respectively. The score is multiplied by five to gain a percentage. Lawlis et al. (1989), who developed this instrument

with varying scale lengths, explain the rationale for the lack of uniformity in scale length as follows:

Using previous pilot studies, differential weighting of each segment accounted for variances of total scores; therefore, by applying different numbers of segments with respect to high predicting variables, the scoring could be done without complicating the process by multiplying each segment before summing. For example, “lifting interference” was weighted slightly more than “sleeping interference” and hence was segmented into six rather than five scoring weights[U1]. (p. 512)

Speed of administration and scoring is identified by the authors as positive features of the measurement tool and they comment that the test can be scored “in 60 seconds or less.” (Lawlis et al., 1989).

Lawlis et al. (1989) reported a “stability reliability coefficient of 0.970 using the method described by Anastasi and Cronbach in 1961.” Their analysis included a total of 143 subjects divided into “pain” and “non-pain” groups. The pain group consisted of 104 chronic back pain patients, 48 women and 56 men, undergoing pain management training and treatment in an inpatient program who had been medically diagnosed and referred as well as 15 patients, five women and ten men, who had been discharged from the inpatient program to work and who were working. (Many people are discharged from chronic pain programs to work but do not return to work.) The comparison group consisted of 24 controls recruited from clinic staff and local airline employees. The problems with small samples (the working chronic pain group and the control group) and short time frames such as this was discussed in the Oswestry Disability Index section. Also no mention is made in the published results regarding any attempt to match the controls with the cases for demographics such as race, age or gender, nor to screen for prior back injury or current back pain. They report on a t-test that demonstrated that chronic pain patients have significantly higher DPQ scores than

“normals”. Concurrent functional validity was tested by comparing the scores of a small sample of 15 patients who were returning to work with scores obtained from functional capacity evaluation on tasks such as manual material handling. There was a negative correlation between this measured ability and scores on the DPQ. Citing this, Lawlis reported that “because these findings support its statistical properties, the DPQ appears to have utility for clinical and research purposes.” There is a large population of back pain patients who have not undergone extensive inpatient rehabilitation who were not represented in the sample. In fact, it could be argued that the largest population of individuals with chronic back pain remains in the workforce.

Despite the fact that few studies have been done to further establish the psychometric properties of this instrument, the DPQ is now widely used to assess the consequences of chronic low back pain (LBP) and outcomes of treatment. Christensen, Laursen, Gelineck, Hansen and Bünger (2001) used it to assess the functional outcomes of posterolateral spinal fusion at unintended levels due to bone-graft migration. They reported that there was no significant difference in functional ability following this complication based on the results from the DPQ.

Roche et al. (2007) used the DPQ as one of the measures for comparison of a functional restoration program with active individual physical therapy for patients with chronic low back pain. This population of working individuals with long-standing back pain was not included in the instrument development but the DPQ is being used to evaluate them. In France, Marty, Blotman, Avouac, Rozenberg and Valat (1998) developed a French language version which they reported to be reproducible, valid and sensitive. Ozguler, et al., (2002) used the French DPQ to classify individuals with low back pain in a working

population into 4 groups ranging from slightly disabled to disabled with emotional consequences. They first changed the scoring from the visual analogue scale to a numerical rating scale of 1-10 for all items. They felt that this made it more homogeneous but did no validation studies to support their conclusion. There are numerous other examples of use (and misuse) of this measurement tool within the literature.

Spinal Function Sort (SFS)

The PACT Spinal Function Sort (SFS) is a self-report measurement of physical work capacity that employs the use of pictorial activity and task sorts (PATs). See Appendix 3 for a copy of this instrument. According to Matheson (2004):

the PATs approach is an efficient means of gathering information about ability to perform a wide variety of work activities and tasks in a brief period of time. In addition to providing information about abilities, these measures can provide information about the evaluatee's psychological status that may be valuable for rehabilitation planning (p. 175).

The Spinal Function Sort (SFS) was developed by Leonard Matheson and Mary Matheson (1989). The items are a set of 50 pen and ink drawings presented in booklet format, 2 drawings to a page. Each drawing depicts a person involved in a work task with a brief description of the activity below the drawing. Standard instructions are read to the individual regarding how they should score each item. The examinee indicates his/her ability to perform each task as "Able" (scored as "1") to "Unable" (scored as "5") with "2", "3" and "4" indicating slightly restricted (scored "2") to very restricted (scored "4"). There is also a "Don't Know" category. Two pairs of items within the 50 items are identical to gauge response consistency.

To obtain a score on the instrument, the assessor counts the number of responses in each category "1" through "5". All of the "1" responses are multiplied by a factor of four,

the “2” responses are multiplied by a factor of three, the “3” responses are multiplied by a factor of two, and the “4” responses are multiplied by a factor of one. These products are then added to determine an overall rating of perceived capacity (RPC) score which is then related to the Physical Demand Characteristics chart found in Appendix 5.

The Spinal Function Sort was developed by the Mathesons in 1989 in response to a perceived need for an assessment that emphasized material handling tasks or activities of daily living tasks that involved spinal movements or loading. To develop the instrument, 500 pictures of men and women performing such tasks were collected. Line drawings were made of the photographs for each task. The 208 resulting tasks were made into a card sort deck and given to 5 evaluators who grouped tasks according to biomechanical demands of the task. During the sorting process 43 groups were determined and a representative task was selected for each group by the test developers. Five tasks were added based on suggestions by the evaluators to give a final count of 48 tasks, two of which would be replicated within the set to measure consistency.

Test-Retest reliability was established with a two day test-retest Pearson product moment correction of $r = .85$. (Matheson et al., 1989). The developers report a variety of test-retest reliabilities ranging from $r = .85$ for the two day test-retest to $r = .77$ for the eight day test-retest. Matheson, Matheson, and Grant (1993) reported further reliability studies in the Journal of Occupational Rehabilitation and suggested that additional research was needed.

The validity of this instrument in terms of its relationship to functional performance status and to other measurable changes in status that occur with treatment is also an appropriate focus of investigation. Finally, research is needed to analyze the factor structure of the SFS. This will be useful to better

understand the underlying dynamics of the components of perceived functional ability that are sampled by the SFS. (p. 27)

Gibson and Strong (1996) published a study evaluating the reliability and validity of the SFS. The sample consisted of 34 men and eight women who had diagnoses including lumbar, thoracic, neck and/or shoulder sprain, along with chronic illnesses such as systemic lupus erythematosus presenting for functional capacity evaluation. A sub sample of 14 of the 42 subjects (ten men and four women) in the study attended for a second administration of the SFS four to fourteen days later. No indication of why this sample size was used or how participants were selected is given in the article. Test-retest validity was established using intraclass correlation coefficient (ICC) $ICC = .89$. Internal consistency was measured using Cronbach's alpha ($\alpha = .97$). Again with this study, I have issues with sample size and the high internal consistency which can indicate item redundancy.

To further examine the validity of the Spinal Function Sort as a measure of perceived capacity for work-related tasks in persons with chronic back pain, Gibson and Strong (1996) used correlational methods to determine the relationship between scores on the Spinal Function Sort and scores on other scales with established validity for measuring similar constructs in persons with chronic pain. Multiple regression was used to examine the prediction of scores on the Spinal Function Sort by the other measures used in the study. The Spinal Function Sort correlated significantly ($p < .001$), adjusted $R^2 = .64$, $df = 5$ with the Pain Self-Efficacy Questionnaire, Self-Efficacy Scale, Pain Disability Index, and Work Reentry Questionnaire 24. With Bonferroni adjustment for multiple comparisons requiring probability of less than .003, the Spinal Function Sort still correlated significantly with each of these measures and, of course, in the anticipated direction.

Gibson and Strong (1996) reported that the study results supported the test-retest reliability of this instrument as well as its internal consistency. They expressed that some support for construct validity of the SFS as a measure of perceived capacity for work had been obtained. They comment that “the SFS depicts tasks that are compatible with those assessed in a functional capacity evaluation (FCE), thus allowing comparison of perceived capacity with the capacity observed in a functional capacity evaluation.” They administered the SFS to the 42 clients presenting for FCE but did not report on any relationship between perceived capacity and actual performance. In the clinical setting this difference is often used as an indicator of symptom magnification.

Robinson et al., (2003) evaluated the clinical utility of the Spinal Function Sort with a group of postoperative and non-operative back patients who had completed a functional restoration program. The SFS was administered both before and after the functional restoration program and was found to measure change effectively. They reported that “Overall, the SFS was found in the present study to be sensitive enough to detect improvement in the functioning capacity of a postoperative spinal group as a result of a functional restoration program”.

Certified Work Capacity Evaluators (CWCEs) are trained to use this instrument and it is part of the standard test battery in the Matheson Functional Capacity Evaluation yet the method for scoring this instrument seems to have been arbitrarily developed. The application of Rasch Analysis can determine if the current method of scoring is effective or if items should be scaled differently. Test equating could confirm the assumption that perceived physical work capacity as measured by the SFS is a similar construct to perceived disability

as measured by the Oswestry Disability Index, Dallas Pain Questionnaire and the Neck Disability Index.

Neck Disability Index (NDI)

Vernon and Mior (1991) developed the NDI to assess how neck pain in individuals affects their activities of daily living. See Appendix 4 for a copy of the NDI. It was adapted from the Oswestry Disability Index for use with populations who have neck pain rather than low back pain. The ten items measure various levels of neck pain, headache, personal care, work, driving, lifting, recreational activities, reading, sleeping and concentration. Each item includes six potential responses, each describing a greater degree of disability, ranging from no disability to total disability. The NDI's total percentage score is calculated by adding the individual item scores (which range from 0 to 5), doubling the total and expressing the result as a percentage. A higher score is indicative of greater perceived disability associated with the neck disorder.

Vernon and Mior (1991) reported on the reliability and validity of the NDI after it had been tested on a small cohort of 17 patients. Small sample size seems to be a recurring issue with these measurement instruments. They asserted that the test-retest reliability of $r = .89$ (over 48 hours) showed response stability, that it was responsive to change in condition as assessed by comparing the percentage of change on a subset of 10 patients before and after treatment and found that it correlated significantly with the Pain Visual Analogue Scale (Huskisson, 1974) and the McGill Pain Questionnaire (Melzack, 1975). They also postulated that the NDI might be assessing two different factors which were represented by tasks that were voluntary vs. obligatory. They reported an internal consistency of $\alpha = .80$ but did not address the possibility of item redundancy as a possible contributor to the high correlation

and did not divulge the inter-item correlations. Inter-item correlation of .9 or more indicates item redundancy (Streiner & Norman, 1989).

In 1998, Hains, Whalen and Mior, postulated that the NDI may contain response set bias as all of the items start with the lowest degree of difficulty and progress to the highest. The patient could be responding consistently by selecting the same level on each question regardless of the question. They developed seven variations of the NDI to determine if this was an issue. However, they determined that the responses obtained from the 237 subjects were related to content rather than response set bias. Hains et al. (1998) concluded that, “This study supports the use of the NDI as a homogeneous instrument possessing stable psychometric characteristics that could provide a means of assessing the disability and the response to treatment over time for individual patients suffering from neck pain” (p. 77)

Ackelman and Lindgren (2002) when validating a Swedish translation of the NDI reported that “The NDI for the neck pain subjects was well distributed and neither ceiling nor floor-effects could be seen” (p. 286). Cleland, Childs and Whitman (2008) developed a study with 137 mechanical neck pain participants “to examine the psychometric properties including test-retest reliability, construct validity, and minimum levels of detectable and clinically important change for the Neck Disability Index (NDI).”

They found that:

the NDI and NRS (Numeric Pain Rating Scale) exhibit fair to moderate test-retest reliability in patients with mechanical neck pain. Both instruments also showed adequate responsiveness in this patient population. However, the MCID (minimal clinically important difference) required to be certain that the change in scores has surpassed a level that could be contributed to measurement error for the NDI was twice that which has previously been reported. Therefore the ongoing analyses of the properties of the NDI in a patient population with neck pain are warranted. (p. 73).

This publication appeared to upset one developer of the instrument (Vernon) and he responded in the July 2008 Letters to the Editor of the Archives of Physical Medicine and Rehabilitation citing numerous other studies which reported better test-retest reliability than the .50 reported by Cleland et al. He felt that the treatment interval of 2-4 days was too short to show change in the patient and that an interval of 2 weeks would be more appropriate.

In a rebuttal to Vernon's defense of the NDI, Cleland et al. (2008) defended their experimental design in the same issue of the Archives of Physical Medicine & Rehabilitation and stated that "Examination of the psychometric properties is an ongoing process and we urge more investigation into the NDI, as well as continued work in the research community on some sort of standardized approach to examining the psychometric properties of self-report questionnaires in general. (p. 1416)"

Recently, a study by van der Velde et al. (2009) evaluated the measurement properties of the Neck Disability Index using Rasch Modeling for a sample of 521 subjects with neck pain. They reported that the NDI in its original form was not a unidimensional interval-level scale but that this could be, and was, accomplished with the removal of two misfitting items; headaches and lifting. They found disordered thresholds in five of the items (personal care, lifting, headaches, work and recreation) but chose not to correct the disordered thresholds because doing so would preclude the possibility of providing a straightforward exchange between the everyday summed ordinal score and its corresponding interval score. They also felt that collapsing the categories would result in a varying number of categories across items which was a significant departure from the original design. They recommended further examination of this instrument with consideration being given to collapsing the categories in a systematic and clinically relevant way.

I agree with Cleland et al. (2008) regarding the need for a standardized approach to examining the psychometric properties of self-report questionnaires and I am heartened to see the NDI which was developed and validated with small samples, rigorously examined using a sample of 521 patients. I am concerned that researchers can, and do, alter the questionnaires to suit their needs and apply the values for reliability and validity that were obtained using another form of the instrument as was done by Ozguler et al. (2002) with the Dallas Pain Questionnaire. Measurement in the social sciences needs to meet the standards of the hard sciences when it comes to development and use of measurement tools. We do not change the length or scale of a ruler because it fits better in our pocket, so neither should we change the scale of a pen and paper measurement tool because it suits us at that time. We know that with a twelve inch ruler, each inch contributes equally to the one foot measurement; we need to endeavor to develop instruments that can be relied upon in this fashion in the social sciences.

Measurement Theory

The four measures (ODI, DPQ, NDI and SFS) to be studied were all developed using classical test theory (CTT). Using this model, item values are assigned rationally rather than empirically. This results in scores that do not proportionally indicate the trait, cannot be compared proportionally over time and cannot be linked to an external standard. Matheson et al., (2008) clarifies the problems with items constructed using classical test theory:

While scientific measures rely on proportional values, many self-report instruments used in healthcare do not. Many count each item selected by the patient without assurance that all items have the same unit value. Others add number values assigned to ordinal scale items without assurance that the proportionality indicated by the numbers reflects the item's true value. Both types of measures also sum the item scores without the assurance that the measure's total scores have proportional value. Addition of item scores is used to derive a total score or division of the total by the number possible is used to derive a percent score, suggesting that these instruments have mathematical qualities that they do not have. The absence of proportional value calibration in items limits the ability of such instruments to quantify a patient's status dependably across the range of reported scores. (p. 46)

Item Response Theory (IRT) addresses the issues raised by Matheson and some researchers have begun to develop and refine measurement instruments for the Social Sciences using these models. Using Rasch analysis, Davidson (2008) compared three versions of the Oswestry Disability Questionnaire given to 100 patients at their first admission to one of seven outpatient hospital clinics or one of nine private practice physiotherapy practices in Australia. The initial questionnaire completion was at admission in person and the follow up was done by mail four weeks later. Her results showed unidimensionality on all items except for the "changing degree of pain" that had been added on the chiropractic form.

Page, Shawaryn, Cernich, & Linacre (2002) applied Rasch modeling to the Revised Oswestry Disability Questionnaire (RODQ) which was the version developed by the

chiropractors. Their findings were as follows: “Several Rasch analyses were performed, with Item 1 Pain deleted and 2 response categories collapsed, creating a better test without increased error. A schema for item administration and evaluation was also developed” (p. 1579). Page et al. (2002) suggest that the revised instrument “boasts good psychometric characteristics, although future researchers may want to subject it to further analysis” (p. 1583).

White and Velozo (2002) applied the Rasch Model to original Oswestry Disability Questionnaire responses from 942 patients with the following results:

All items from the Oswestry except the pain item fit the Rasch model. Construct validity of the scale using the Rasch model required the structure of the rating scale to be modified from 6 response levels to 4. A hierarchical representation of LBP disability was supported. A comparison of the disability categories based on Likert and Rasch scaling revealed them to be non-equivalent. The new scaling changed the disability categories for 44% of patients. (p. 822)

It should be noted that the White and Velozo (2002) results are similar to the Page et al. (2002) results even with a sample size that was ten times that of the latter sample. No further analysis has been done to validate the resulting ODI revision. This is necessary before it is used as a clinical or research measurement tool.

A literature search for Item Response Theory or Rasch Modeling with the DPQ, and SPS rendered no results.

Item Response Theory (IRT)

Item response theory (IRT) is a psychometric theory that consists of a series of mathematical models which relate person and item parameters to the probability of the responses on a discrete outcome, such as a correct response to an item or an endorsement of a category of a trait. The attraction of Item Response models lies in their promise of invariant

item and person parameters provided there is data-model fit. That is, estimates of ability are not dependent on the difficulty of items. IRT also provides a basis for estimating several item parameters (i.e., difficulty, threshold, guessing, and category intersection), ascertaining how well data fits a model, and investigating the psychometric properties of assessments. As there is a single person characteristic assumed to account for the responses, the model is described as “unidimensional”. “Compared with classical test theory, IRT generally provides more sophisticated information regarding the psychometric properties of individual assessment items. The application of IRT has been wide, including the measure of personality traits, moods, behavioral dispositions and attitudes, as well as cognitive traits. Moreover, IRT is frequently applied to many health measurements” (Tsutsumi et al., 2008, p.110).

Within IRT there are three probabilistic measurement models: the 1-parameter (1PL), 2-parameter (2PL) and 3-parameter (3PL), named by the number of item parameters estimated in each model. All three models can be derived from the equation below for the 3-parameter model:

$$P(\theta) = c_i + \frac{1 - c_i}{1 + \exp[-1.7a_i(\theta - b_i)]}$$

where the three item parameters are

c_i = low asymptote of ogive (guessing)

b_i = median intercept of ogive (difficulty)

a_i = slope of ogive at inflection (discrimination), and the one person parameter is

θ = ability of a person on the variable

To derive the 2-parameter model, “ c_i ” is held constant eliminating “guessing”. For the 1-parameter model, “ a_i ” is held constant for all items and is often scaled to equal one. When “ a_i ” is held constant this implies that all items on a test are equally discriminating. This leaves item difficulty as the sole parameter being estimated. The 1PL is expressed as follows:

$$P(\theta) = \frac{1}{1 + \exp[-1.7a_i(\theta - b_i)]}$$

Mathematically, the Rasch Dichotomous Model is identical to the 1-parameter IRT model with a formula of

$$P(\theta) = \frac{e^{(\beta_n - \delta_i)}}{1 + e^{(\beta_n - \delta_i)}}$$

where

e = base of natural logarithm or Euler’s number; 2.7183

β_n = person’s ability

δ_i = item or task difficulty

However there are some important differences. As Shaw (1991) explains:

This approach seems to imply that the Rasch model is just a stripped-down version of more complicated models which “must be better” because they account for more of the “presumed reality” of traditional test theory. Quite apart from Occam’s razor (that entities are not multiplied beyond necessity), this interpretation is shallow in an essential way. That the Rasch model can be reached by simplifying more complicated models has nothing to do with its genesis or rationale, or with the theory of measurement. (p. 131)

The Rasch model is based on measurement principles that provide sample-free item calibrations and test-free person measures on a common linear scale that can be analyzed statistically. Introducing the parameters for item discrimination and guessing violates these principles of measurement as outlined by Shaw (1991) that:

1. the measures of objects be free of the particulars of the agents used to estimate these measures and the calibrations of agents be free of the particulars of the objects used to estimate these calibrations.
2. the measures of objects and calibrations of agents function according to the rules of arithmetic on a common scale so they can be analyzed statistically.
3. linear combinations of measures and calibrations correspond to plausible concatenations of objects and agents. (p. 131)

The mathematical elegance (simplicity) of the model allows for superior estimation capabilities. This makes Rasch, with its sole parameter of item difficulty, a more viable proposition for practical testing. In Rasch model thinking, the model is superior and data which does not fit the model is discarded.

Rasch Analysis

Rasch analysis is a statistical procedure used to transform ordinal-scaled measures into interval-scaled measures that provide good reliability and acceptable quantitative validity measured with fit characteristics. A primary advantage of using Rasch analysis is that the interval scaling scheme establishes standardized distances between points, thus allowing for more accurate interpretation of the levels measured. Items are distributed according to their difficulty and subjects are distributed according to their abilities. This results in a single linear scale that represents the underlying trait in question. Rasch analysis also evaluates item fit and thereby helps to determine which items are most useful in

assessing the construct under discussion. It reduces item redundancy and can shorten measurement tests to reduce the time needed for test administration and scoring.

Rasch techniques can provide psychometric information that was previously unavailable with CTT techniques. Without converting the data into an interval scale, clinicians might mistakenly treat a participant's total score as a sum of equal-valued items. After the data are converted, a researcher can utilize Rasch analysis to assess several psychometric characteristics, including unidimensionality, item hierarchy, and person reliability and separation statistics (Pomeranz, Byers, Moorhouse, Velozo & Spitznagel, 2008). Patient ability (from least to most able) and item difficulty (from least to most difficult) can be calibrated into a common underlying scale measured in logits (log odds units) (Davidson, Keating & Eyres, 2004).

Rasch Analysis looks at data fit and examines the agreement between the model's predicted responses and the observed responses. Fit statistics are provided that highlight poorly constructed items or indicate that some items do not measure the desired attribute. The researcher looks at how well the data fits the Rasch Model rather than the conventional approach of how well the model fits the data. "The Rasch model is a mathematical description of how fundamental measurement should operate with social/psychological variables. Its task is not to account for the data at hand, but rather to specify what kinds of data conform to the strict prescriptions of fundamental measurement" (Bond & Fox, 2007, p. 235). Rasch models include dichotomous and polytomous models.

Dichotomous Model

The model from which all other Rasch models have grown is the dichotomous model which is expressed in the logarithmic odds on success form as follows:

$$\ln\left(\frac{P_{ni}}{1 - P_{ni}}\right) = B_n - D_i,$$

Where P_{ni} is the probability of examinee n correctly answering item i ,

$1 - P_{ni}$ is the probability of examinee n incorrectly answering item i ,

B_n is the proficiency level of examinee n , and

D_i is the difficulty level for item i .

This is applied to dichotomous or yes/no data to obtain examinee proficiency and item difficulty.

Rating Scale Model

The first of the polytomous models is an extension of the dichotomous model for items that have more than 2 response choices such as the ODI, DPQ, SFS and the NDI. The Rasch Rating Scale model (RSM) (Andrich, 1978) is the recommended model. It can be expressed as follows:

$$\ln\left(\frac{P_{nik}}{1 - P_{ni(k-1)}}\right) = B_n - D_i - F_k$$

where P_{nik} is the probability of examinee n scoring at level k on scale i ,

$1 - P_{ni(k-1)}$ is the probability of examinee n scoring at level $k-1$ on scale i ,

B_n is the proficiency level of examinee n , and

D_i is the difficulty level for item i .

F_k is the difficulty of the step from level $k-1$ to k .

Essentially it is the dichotomous model with thresholds added between steps. Likert scales are often presented in a format such as SD – strongly disagree, D- Disagree, N – Neutral, A –

agree, SA – Strongly Agree or on scales such as the one found on the Spinal Function Sort which ranges from Able to Unable with varying degrees of reduced ability in between. A five item scale such as this would be modeled as having 4 thresholds. Rasch analysis would determine if there are this many distinct thresholds or if categories could be collapsed.

Other instruments might be modeled as a rating scale such as the ODI (version 1) (Fairbank et al., 1980) where there are 6 statements in each section as shown below:

Section 3—Lifting

1. I can lift heavy weights without extra pain.
2. I can lift heavy weights but it gives extra pain.
3. Pain prevents me from lifting heavy weights off the floor, but I can manage if they are conveniently positioned, e.g. on a table.
4. Pain prevents me from lifting heavy weights but I can manage light to medium weights if they are conveniently positioned.
5. I can lift only very light weights.
6. I cannot lift or carry anything at all.

The assumption is that the statements are a series of uniformly increasing steps.

Within Rasch modeling, instruments are assessed with regard to unidimensionality, fit and differential item functioning (DIF). Unidimensionality is the extent to which all items on a given scale are measuring the same construct or latent trait. This unidimensionality or local independence is a requirement of all Rasch models. Fit statistics estimate how much each item adheres to the modeled expectations and indicate if each item on the instrument contributes to the measurement of that single latent trait. Infit statistics give more weight to the abilities of persons closer to the item value. The outfit statistic is unweighted and therefore is more sensitive to outliers in the data.

Overall fit is the extent to which the data for the class intervals fits the Rasch model. It is tested with a chi square statistic where a chi square value larger than the selected alpha value e.g. $p > .05$ indicates no deviation of data from expected. The person separation index is an estimate of the spread of respondents on the variable. It is the adjusted person standard deviation divided by the average measurement error. The person separation index provides information regarding the number of groups the test can discriminate amongst (Wright & Masters, 1982).

Differential Item Functioning (DIF) is a method for detecting test items that function differently across subgroups of examinees as delineated by such parameters as age or gender. Uniform DIF is when the items perform similarly across all groups; failure to do so indicates item bias.

Partial Credit Model

The partial credit model (Wright & Masters, 1982) is a version of the rating scale model wherein the threshold estimates, including the number of estimates, are free to vary from item to item. This may be the better method of analysis for at least some of this data as the item series barely suggest equal spacing between choices. It is expressed by the formula:

$$\ln \left(\frac{P_{nik}}{1 - P_{ni(k-1)}} \right) = B_n - D_{ik}$$

where P_{nik} is the probability of examinee n scoring at level k on scale i ,

$1 - P_{ni(k-1)}$ is the probability of examinee n scoring at level $k-1$ on scale i ,

B_n is the proficiency level of examinee n , and

D_{ik} replaces $D_i + F_k$ in the Rating Scale equation where D_i is the difficulty level for item i and F_k is the difficulty level for item i .

The replacement of $D_i + F_k$ with D_{ik} signifies that in the Partial Credit model each set of threshold estimates is related uniquely to its own item instead of for the entire set of items. This model is robust in situations such as the DPQ where there are differing lengths of visual analogue scale as this model does not require the same number of response categories for each item. It allows for an empirical test of whether the distances between response choices are constant. The other three measures, the NDI, ODI and SFS, do have consistent numbers of response categories so the data may fit the rating scale model for the individual analyses of each instrument. However, while the numeric values are the same, for the ODI and NDI each item has varied response choices making the Partial Credit model a superior model for these instruments as well as the DPQ. Therefore, the rating scale model will be used with only the SFS and the partial credit model will be used with the ODI, DPQ and NDI. This is necessary because in Rasch modeling the first tenet is that the data fits the model.

Research Questions

1. Does each instrument measure one unidimensional trait?
2. Do all of the items within an instrument fit (belong) on that instrument?
3. Are the scale categories within items appropriate?

Limitation – these instruments will not be examined for their ability to operate as one unified whole i.e. test equating.

Significance of Proposed Study

An improved scoring system of each of the instruments would result in better sensitivity and specificity of test scores. This would give the examiner more confidence in the scores obtained from these measures.

CHAPTER THREE: DESIGN AND METHODOLOGY

Subjects

An intact data set for 298 individuals who had participated in a Functional Capacity Evaluation at Central Interior Disability Management Services (CIDMS) between 1998 and 2008 was supplied by CIDMS once the University of Northern British Columbia Ethics Committee approval was obtained. Authorization to use this data (pending ethics approval) was obtained from CIDMS and a letter authorizing the use of the data is attached. (Appendix 5) A copy of the “Consent to Evaluate” that was signed by each subject prior to assessment is attached (Appendix 6). Ages of participants ranged from 21 to 64 year old, as depicted in Figure 1, both genders were represented (125 females and 173 males). Table 1 reports the physical occupational demands of the participants which ranged from Sedentary to Very Heavy according to the DOT (1991) classification found in Appendix 7.

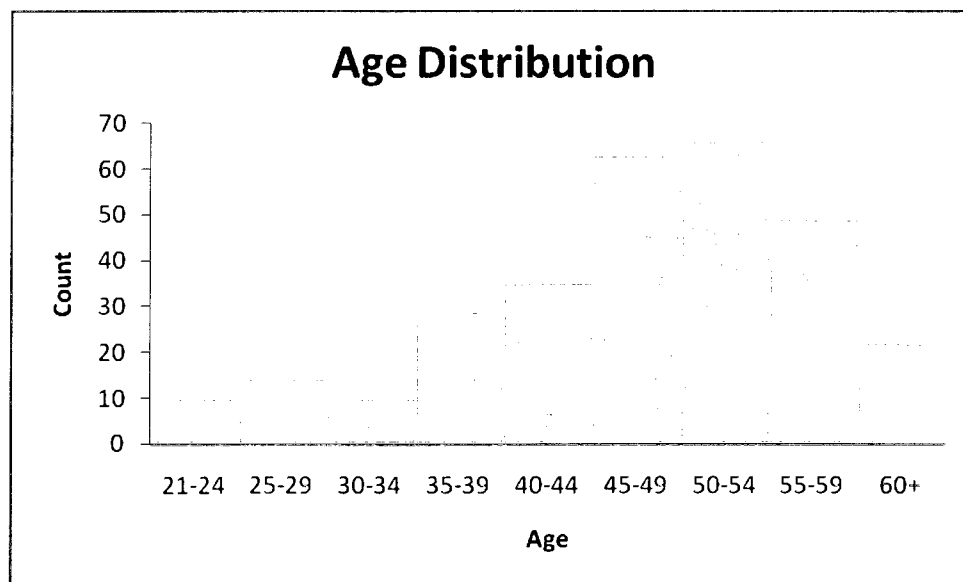


Figure 1 Age distribution of study sample

Table 1 *Physical Occupational Demands*

Demand Level	Description	Participants
No job	Unemployed	1
Sedentary	Material handling up to 10 pounds on an occasional basis, up to 1/3 of the day*. Office-type work.	15
Light	Material Handling up to 20 pounds on an occasional basis, up to 1/3 of the day*. Laboratory Workers, Lumber Graders, Teachers, etc.	69
Medium	Material Handling of 21 to 50 lbs. on an occasional basis.* Mill Workers, Care Aides, etc.	137
Heavy	Material Handling of 51 to 100 lbs. on an occasional basis.* Electrician, Manual Laborer, etc.	40
Very Heavy	Material Handling of 100+ lbs. on an occasional basis.* Trades such as Millwright, Heavy Duty Mechanic, Planer Operator, etc.	36

*Detailed physical demands in Appendix 7.

Diagnoses ranged from multiple soft tissue injuries following motor vehicle accident to chronic conditions such as fibromyalgia. The criteria for administering the questionnaires to these clients was based on their identification, on the Ransford Pain Drawing (Ransford, Cairns & Mooney, 1979), of pain in the neck or back. Their response determined which questionnaire was appropriate; either the Neck Disability Index or the Oswestry Disability Index. If they identified pain in both areas, they were given both. For pain in other areas, they were asked to complete the Dallas Pain Questionnaire and if they identified a loss of ability to perform activities of daily living, they were asked to complete the Spinal Function Sort.

Instrumentation

The four instruments administered were the Oswestry Disability Index, Dallas Pain Questionnaire, Spinal Function Sort, and the Neck Disability Index. These instruments were chosen as they are part of the standard Matheson FCE battery (Roy Matheson and

Associates, 2006). Full descriptions can be found in the Literature review. Samples of each instrument can be found in Appendices 1-4.

Procedures

A full data set was obtained from Central Interior Disability Management Services in Excel format. There were 298 lines of data with each line representing data obtained for one person on one testing occasion. Some individuals were tested on more than one occasion and in that situation, this was identified. Information regarding gender, age and diagnosis was also provided. The file for each client was retained by CIDMS and only the completed data set without any patient names or other specific identifiers was provided to this researcher. The criterion for inclusion in this study was that the file contained any of the four completed instruments –Oswestry Disability Index, Spinal Function Sort, Dallas Pain Questionnaire and/or Neck Disability Index .

Data Analysis

The Oswestry Disability Index and the Neck Disability Index have been analyzed with the Rasch Partial Credit Model using the WINSTEPS computer program (Linacre, 2009). The Partial Credit Model was selected for the analysis of these two questionnaires because, although the numeric values of the rating scale were the same for all items, the individual response choices differed. Wright and Masters (1982) advise the use of the Partial Credit Model in these situations. In contrast, the Dallas Pain Questionnaire and the Spinal Function Sort can be and were analyzed using the Rating Scale Model as the scale for each item is identical. The DPQ does have varying lengths of scale and while according to Andrich (1978) the Rating Scale Model is robust in this situation, I found that since the items

with the longer scales were functioning poorly, it was useful to do the initial analysis using the Partial Credit Model. Once the items were rescaled to equal lengths, the Rating Scale model was used. While the original intent of this work was to equate the four instruments, Linacre (2009) outlines conditions that must exist to equate tests. First, the tests must meet the criteria of unidimensionality with good item fit and ordered thresholds. He states that the latent variable has to be “invariant” across the instruments to be equated or linked. If these criteria are not met, equating should not be attempted.

CHAPTER FOUR: RESULTS

The Oswestry Disability Index

The Oswestry Disability Index (ODI) is a perceived disability scale consisting of ten items. Each item has 6 possible responses in statement form which are arranged in ascending order from least impairment to most impairment (Appendix 1).

Data collected from 133 patients were analyzed using WINSTEPS Version 3.68.0 (Linacre 2009). The partial credit model, which treats the category structure of each item separately, was used because, although the numeric values of the rating scale for each item are the same, the response choices differ. (Wright & Masters, 1982). Rasch analysis uses two types of fit statistics, infit and outfit, to analyze the internal validity of items in the scale. When investigating data, Linacre (2009) recommends the following approach to assessing the results. Negative point-measure or point-biserial correlations should be investigated first and if negative correlations are noted look for miskeys and data entry errors. If all point-measure correlations are positive, investigate outfit before infit, mean square before z-scores and high values before low values.

Positive point-measure correlations indicate that the expectation that individuals with high amounts of the latent trait will score in the higher range on this item, is met. Outfit statistics are more responsive to outliers and high outfit mean-squares can simply be the result of a few random responses by low performers. Infit mean-squares are more responsive to inliers and are sensitive to responses in which estimated person ability values are similar to item difficulty values. High infit mean-squares indicate that the items are mis-performing on the targeted population. Mean squares indicate the amount of distortion in the measurement

system. These infit and outfit statistics are reported as mean squares (MNSQ) and standardized z-scores (ZSTD). While a mean square of 1 and a z-score of 0 is ideal, Wright et al (1994) found that for rating scales, a range of .6 to 1.4 for the infit and outfit mean squares is reasonable. The expected value is 1.0 and values greater than that indicate unpredictability or noise in the data; values less than 1.0 indicate overfit or redundancy of the item. Linacre (2009) states that “if mean-squares are acceptable, the ZSTD can be ignored”.

Diagnostic Measures

The initial ODI data run showed the results in Table 2.

Table 2 *Diagnostic Measures for the Oswestry Disability Index*

Item	Pt – Measure (<i>r</i>)	OUTFIT MNSQ.	INFIT MNSQ.
1. Pain	.63	1.26	1.32
2. Self care†	.68	0.66	0.62
3. Lifting	.59	1.08	1.01
4. Walking	.69	0.83	0.85
5. Sitting†	.69	0.64	0.65
6. Standing	.66	0.82	0.87
7. Sleeping*	.61	1.43	1.48
8. Sex Life*	.67	1.39	1.41
9. Social Life	.71	0.87	0.90
10. Travelling	.60	0.94	0.89

*Equal to or greater than Linacre's 1.4 upper limit.

†Equal to or less than Linacre's 0.6 lower limit.

All correlations are between .59 and .71 which is acceptable. All are positive indicating that each item is positively associated with the measure. The item "sleeping" has an outfit mean square of 1.48. The items "sleeping" and "sex life" have problematic mean square infit statistics of 1.48 and 1.43. "Pain" is the next highest but remains within the .6 to 1.4 range recommended by Linacre. The significance of these mean squares is expressed as a ZSTD and values outside the range of -2 to +2 are associated with $p < .05$. The ZSTD associated with these mean squares were 3.5 and 3.0 respectively.

Once overall item statistics have been evaluated, items can be examined using the Modal Probability Curves. This pictorial representation illustrates the category function and boundaries for each item. The modal perspective on category boundaries on the latent variable identifies a mode being between intersections of the category probability curves. This simplifies inference about which category is most likely to be observed to any item at any point along the latent variable. For example in figure 2, Category 1 “I can stand as long as I want but it gives me extra pain” is the most probable response ($\cong .42$) for a range of -3 to -2 logits below the item’s mean difficulty. When all categories are modal, they look like the graph below for item 6 “standing”. Their thresholds are ordered and the performers with the least amount of the latent trait (disability) are most likely to choose “0” and the performers with the most disability are most likely to choose 5. Note that all categories must have a range of difficulty at which the category is modal.

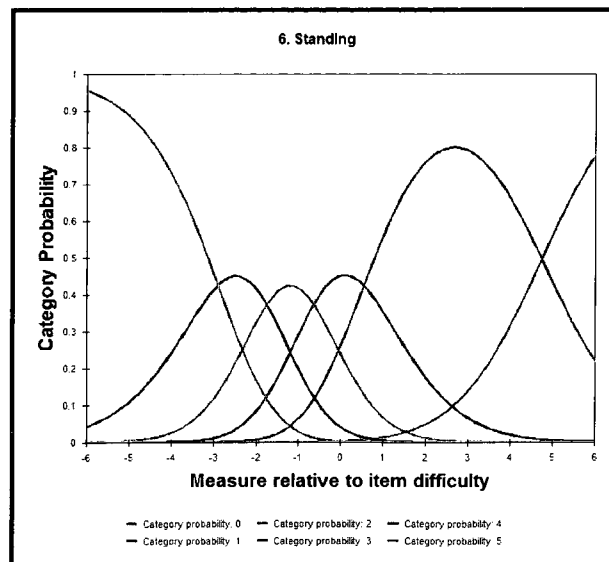


Figure 2 Standing item #6 Oswestry Disability Index.

Figure 2 illustrates the probability of responding to any particular category (y-axis) given the differences in estimates between person ability and item difficulty/ endorsability (x-axis). For example, if a person's ability was 1 logit lower than the difficulty of the item (-1 on the x-axis), the probability of endorsing a "0", "4" or "5" would be close to zero, or endorsing a "1" or "3" would be close to 0.22 and of endorsing a "2" would be close to 0.42. This person is most likely to endorse Category 2 on this item. For the person with higher ability estimates such as +1 to +5 on the x-axis, the most likely response is a 4. The graph shows that each response category is the most probable for some level of the variable.

While item 6 "Standing" performed well and fit the Rasch principles, several others did not. Item 1- Pain (Figure 3) did not function as well as item 6.

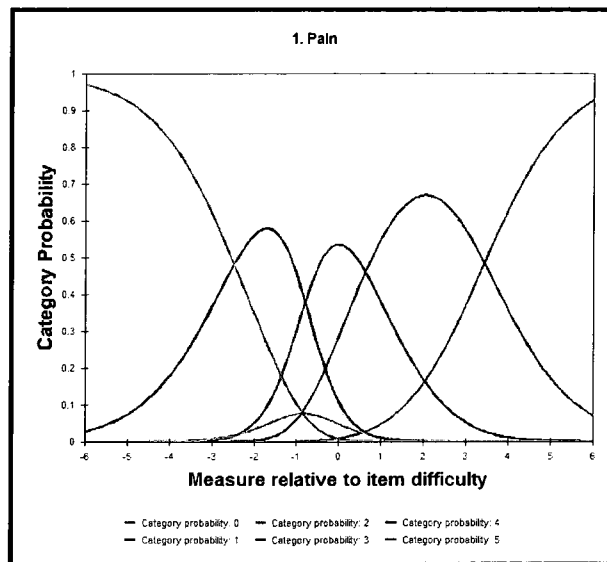


Figure 3 Pain item #1 Oswestry Disability Index

In the graph depicted in Figure 3, for the individual whose ability is one logit lower than the item difficulty i.e. -1 on the x-axis, the probability of endorsing a "0" "2" "4" and "5" is close to 0 whereas the probability of endorsing "1" or "3" is approximately 0.40. Since the

categories are supposed to be ordered with increasing amounts of the latent variable, this demonstrates disordered thresholds. The Rasch-Andrich thresholds are the intersections between adjacent categories (Andrich, 1978). Thresholds for categories 1-2 and 2-3 are disordered. Andrich (1978) is adamant that “disordered thresholds are a violation of the principles underlying the Rasch Model and must be eliminated”. The category intervals on the latent variable must correspond with the modal intervals of the categories. Since category 2 is never modal, it must be removed. Similarly category 2 did not function for the Items “Walking” – Pain prevents me from walking more than ½ mile and “Sex Life” – My sex life is nearly normal but it is very painful. Category 1 did not function for “Sitting” – I can only sit in my favorite chair for as long as I like, “Sleeping” – I can sleep well only by using tablets and “Social Life” My social life is normal but increases the degree of pain.

To correct the disordered scale problems outlined herein, the ODI was rescaled from six categories to five. Each item was assessed individually to determine which two categories should be collapsed into one and is seen in Table 3. The rescaled instrument performed well.

Table 3 *Oswestry Categories Collapsed*

ITEM	MALFUNCTIONING CATEGORY	COMBINED WITH	NEW COLLAPSED CATEGORY
1	Painkillers give complete relief from pain	Painkillers give moderate relief from pain	Painkillers give moderate to complete relief from pain
2	I do not get dressed, wash with difficulty and stay in bed	I need help in most aspects of self care	I need help in most aspects of self care
3	Pain prevents me from lifting heavy weights but I can manage light to medium weights if they are conveniently positioned	Pain prevents me from lifting heavy weights from the floor but I can manage if they are conveniently positioned	Pain prevents me from lifting medium to heavy weights from the floor but I can manage if they are conveniently positioned
4	Pain prevents me from walking more than ½ mile	Pain prevents me from walking more than ¼ mile	Pain prevents me from walking more than short distances
5	I can sit in my favorite chair as long as I like	I can sit in any chair as long as I like	I can sit as long as I like
6	Pain prevents me from standing at all	Pain prevents me from standing more than 10 minutes	Pain prevents me from standing more than 10 minutes
7	I can sleep well only by using tablets	Even when I take tablets I have less than 6 hours sleep	Even when I take tablets I have less than 6 hours sleep
8	My sex life is nearly normal but is very painful	My sex life is normal but causes some extra pain	My sex life is normal but causes some extra pain
9	My social life is normal but increases the degree of pain	Pain has no significant effect on my social life apart from limiting my more energetic interests such as dancing etc.	Pain has no significant effect on my social life apart from limiting my more energetic interests such as dancing etc.
10	Pain prevents me from traveling except to the doctor or hospital	Pain restricts me to short necessary journeys of less than 30 minutes	Pain restricts me to short necessary journeys of less than 30 minutes

Table 4 *Diagnostic Measures for Rescaled Oswestry Disability Index*

ITEM	PT-MEASURE (<i>r</i>)	OUTFIT MNSQ	INFIT MNSQ
1. Pain	.61	1.09	1.10
2. Self care	.71	0.79	0.79
3. Lifting	.59	1.25	1.26
4. Walking	.70	0.75	0.77
5. Sitting	.66	0.90	0.92
6. Standing	.68	1.00	1.05
7. Sleeping	.59	1.14	1.15
8. Sex Life	.64	1.12	1.04
9. Social Life	.71	0.84	0.84
10. Travelling	.62	1.09	1.13

The diagnostic measures outlined in Table 4 show positive point measure correlations and that the infit and outfit MNSQ for “Sleeping” and “Sex Life” have improved from 1.48 and 1.43 to 1.04 and 1.15 and the ZSTD has improved from 3.5 and 3.0 to 1.0 and 0.8, indicating improvement in function of these items. With the rescaled categories, “Pain” is functioning better with an infit MNSQ of 1.10 reduced from 1.32. and a ZSTD reduced from 2.0 to 0.7 and the Modal Probability Curves for this item are shown in Figure 4.

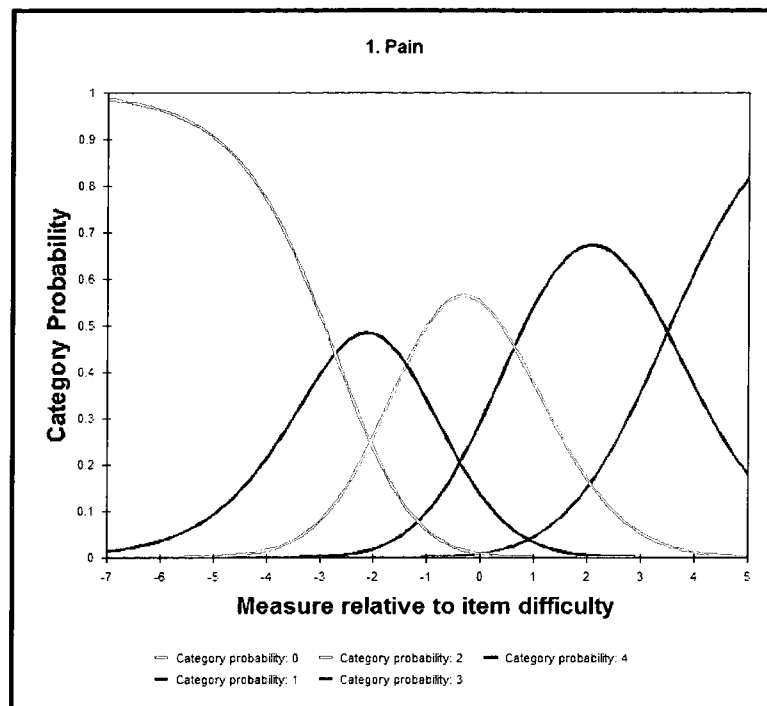


Figure 4 Rescaled pain item #1

Now each category is functioning properly with ordered thresholds. For the example of a person with 1 logit less ability than the item difficulty, the probability of endorsing “0”, “3” or “4” approaches 0 whereas the probability of selecting “1” is .35 or “2” is .5. The same was observed for the other items which previously demonstrated disordered thresholds.

Item Difficulty

To examine item difficulty, the revised scale was assessed with the Rasch Rating Scale model; see Figure 5 . When examining item difficulties, coverage for a wide range of abilities should be evident. The items are arranged on a common scale from easiest to endorse to most difficult. This allows us to see the order of ease of endorsement of the items. It is interesting to note that the easiest to endorse item was standing – i.e. standing tolerance is affected first with back pain. The item most resistant to endorsement was “selfcare”

indicating that individuals with back pain do not perceive a loss of ability to perform self care activities as readily. The order of the categories from easiest to most difficult to endorse was Standing, Lifting, Travelling, Pain, Social Life, Sitting, Sex Life, Walking, Sleeping and Self Care. This is consistent with function observed in clients with back pain in the clinical setting.

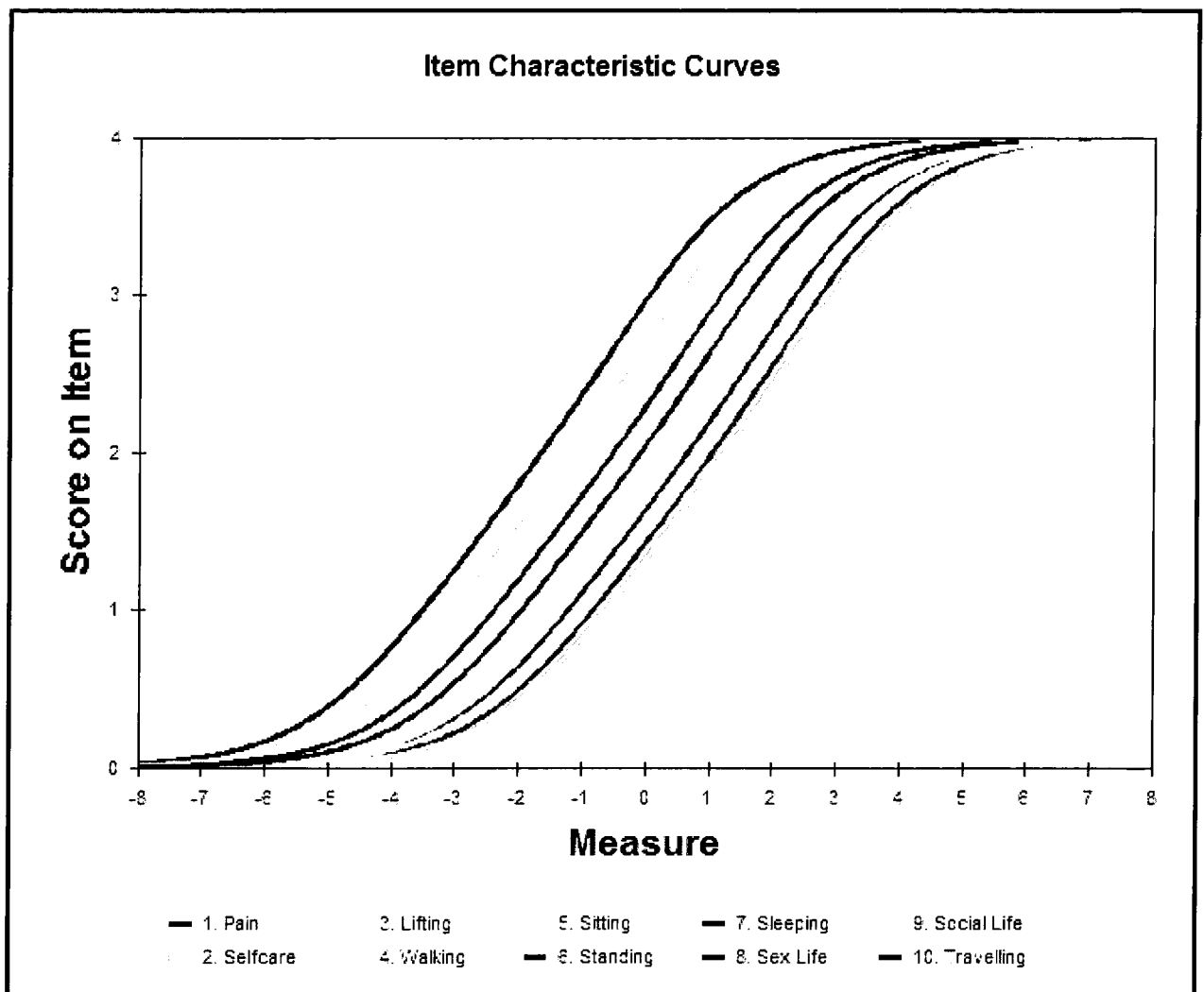


Figure 5 Item characteristic curves depicting item difficulty on the rescaled ODI

Differential Item Functioning (DIF)

Differential Item Functioning (DIF) is an indicator of possible bias and is the result of a lack of invariance across testing situations. For instance one sub-group (i.e. males) with a given level of a latent trait responds differently to an item compared with another subgroup (i.e. females) with a similar level of the latent trait. This was investigated with respect to gender and age and no DIF was found. The results for gender are depicted in Figure 6.

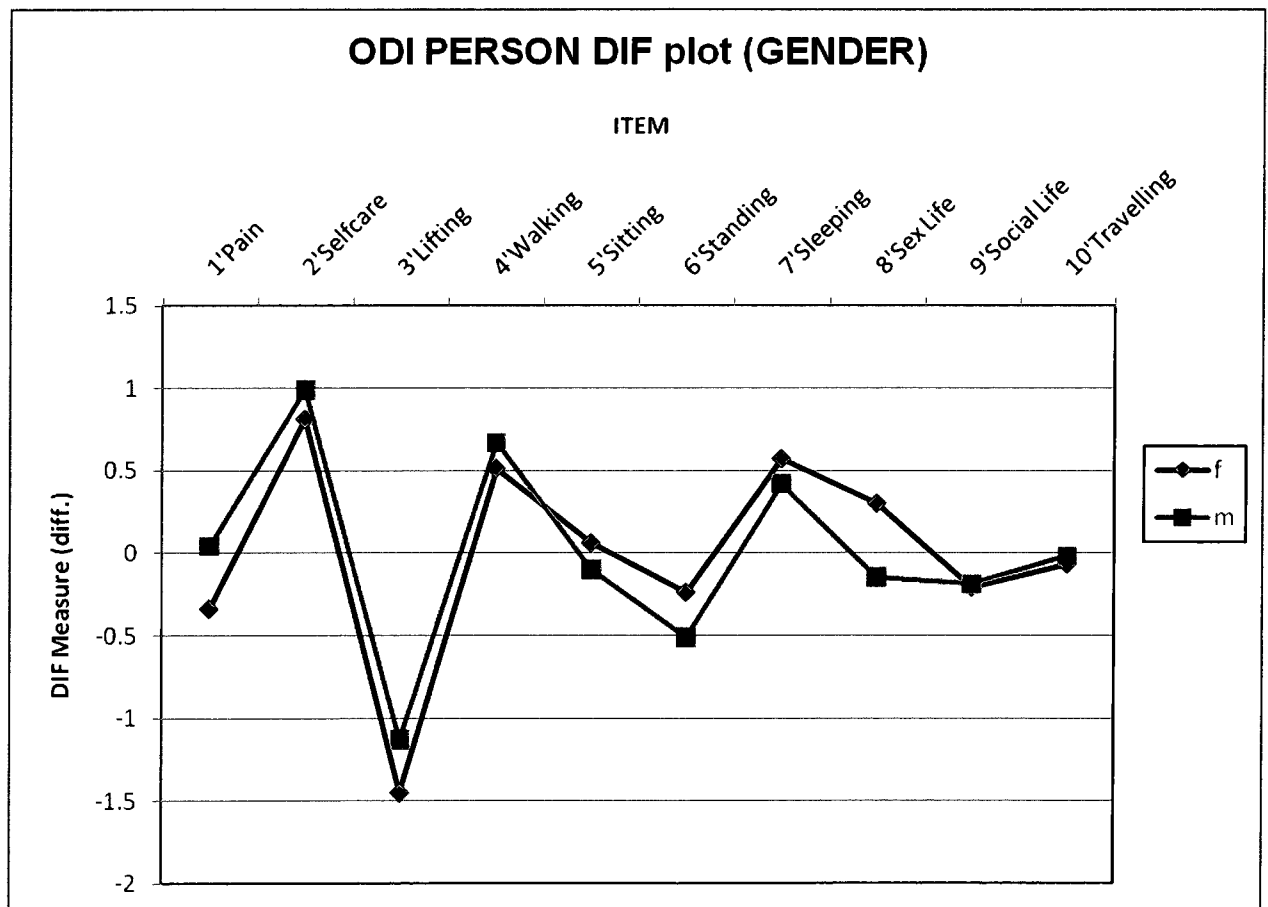


Figure 6 DIF by Gender for the ODI

Dallas Pain Questionnaire (DPQ)

The Dallas Pain Questionnaire is a popular measure for use with individuals suffering from chronic low back pain. For the purposes of this study, only the responses to the “daily activities” and “work/leisure activities” which comprise the functional activities factor of this instrument are considered.

Data collected from 241 participants were analyzed using the Rasch partial credit model. Andrich’s rating scale model is reported to be robust with scales of varying lengths but in this case, the items with the largest scales demonstrated more than one disordered threshold and therefore initial analysis with the Master’s Partial Credit was done. Once the items were functioning better, the scale was assessed with Andrich’s Rating Scale Model.

Table 5 *Diagnostic Measures for the Dallas Pain Questionnaire*

Item	PT-MEASURE (<i>r</i>)	OUTFIT MNSQ.	INFIT MNSQ.
1. Pain	.58	1.33	1.30
2. Self care	.64	0.95	0.95
3. Lifting*	.48	1.45	1.21
4. Walking	.61	1.08	0.99
5. Sitting	.61	0.99	1.01
6. Standing	.65	0.76	0.84
7. Sleeping	.62	0.93	0.95
8. Social Life	.70	0.84	0.86
9. Travelling	.68	0.91	0.92
10. Vocational	.55	0.92	1.02

*Equal to or greater than Linacre's 1.4 upper limit.

Diagnostic Measures for the DPQ

In the initial analysis, all items fit the model except Item 3 "lifting" - which was misfitting with an outfit mean square of 1.45 and a corresponding ZSTD of 3.4. The Category Probability Curves for all items showed disordered thresholds for at least one level of endorsement for each item. Lifting is shown in figure 7. It can be observed that there is no point on the latent variable that category probability 1 or 2 is the most likely to be selected.

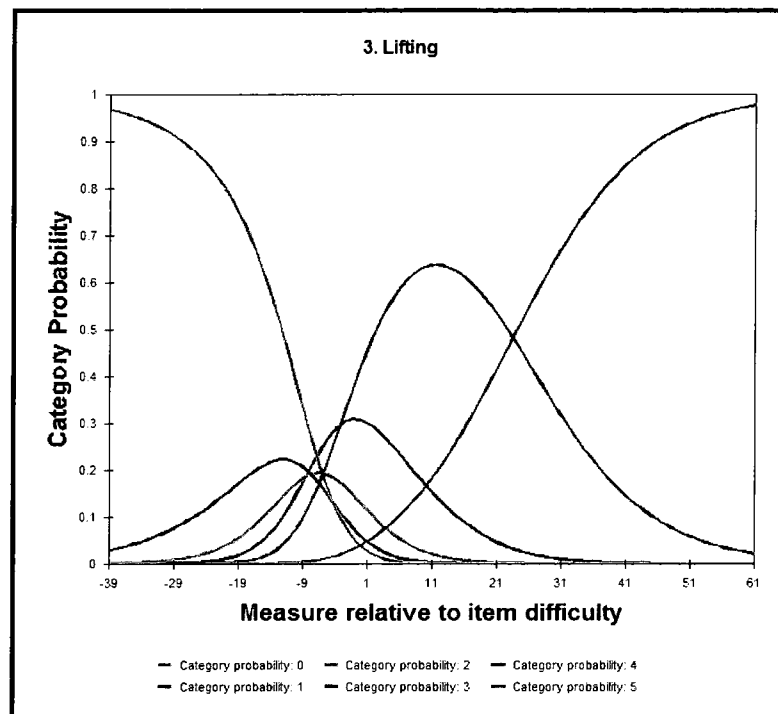


Figure 7 DPQ Item 3 Lifting Category Probability Curve

For the items in the “work/leisure” activities, where there is a seven or eight point scale, the Category Probability Curves demonstrate disordered thresholds on several categories. The “vocational” item is shown in Figure 8. In this study, 144 people selected the final category next to “I cannot work”.

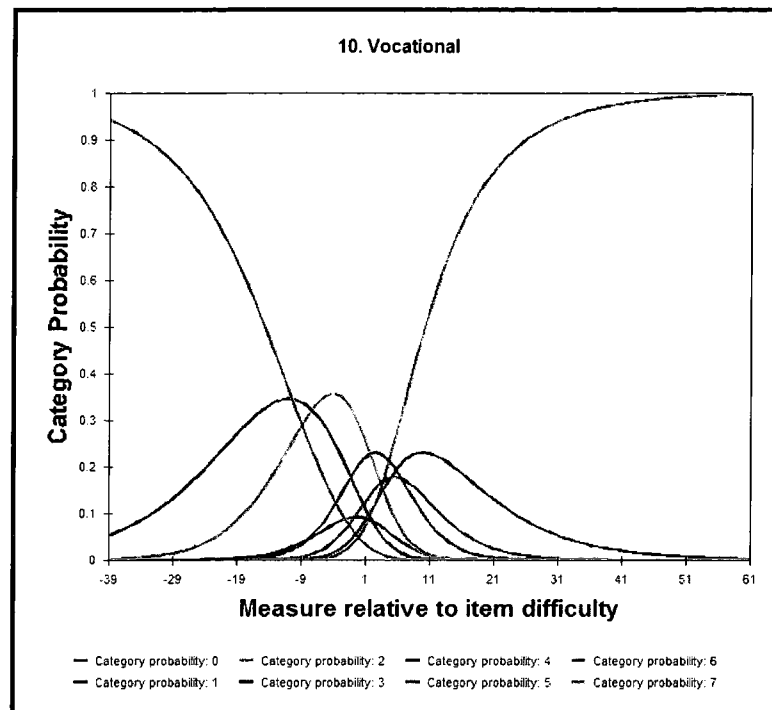


Figure 8 Category probability curve of DPQ item 10 “Vocational”

An attempt was made to create a five point scale for all items. For the first six items, the “0” and “1” categories were collapsed. Item 7 “Sleeping” consisted of five response options on the original DPQ, so it was left as is. The final three items, Item 8 “Social Life”, Item 9 “Traveling” and Item 10 “Vocational” were modified by collapsing adjacent categories “0” and “1”, “2” and “3”, “4” and “5”, “6” and “7” were not collapsed. This improved the scale and with further analysis using the Partial Credit Model, there were no items with mean squares outside the range of .6 to 1.4 however in the first six items, category “1” demonstrated a disordered threshold as did category “7” on the final three items. It was then determined that a four point scale might work better. A four point scale was created by collapsing the new category “1” with category “2” for items the first seven items and by collapsing categories “6” and “7” in the final three items. This corrected the disordered thresholds in items 2 through 10 but Item 1 “Pain” remained problematic. Since the rating

scale now met the criteria for use of the Andrich Rating Scale Model, analysis using this model where one set of threshold values are applied to all the items on the test was undertaken. Once this analysis was done, it was noted that “pain” with an outfit mean square of 1.66 and a ZSTD of 6.3 as well as an infit mean square of 1.72 and a ZSTD of 7.0, did not fit the scale.

The DPQ and Oswestry Disability Index have essentially the same items – the ODI using progressive response choices and the DPQ using a visual analogue scale. White and Velozo (2002) found that the “pain” item did not fit in the Oswestry Disability Index and postulated that it was because “ the responses to the pain item relate to the use of pain medications differing from the responses of the other items that all relate to function (physical, social)” (p. 825). While I did not find as White and Velozo had, that the pain item did not fit the scale on the ODI, it is clear with the fit statistics as reported, it did not fit on this scale. The use of pain medication is not the issue in the DPQ; it is more likely that this item does not fit because pain is a symptom and the other items relate to function/ability. The “pain” item was removed leaving a 9 point scale.

The Category Probability Curve for Item 2 “Personal Care” is shown in Figure 9. In the rating scale model, the same set of threshold values are applied to all items, so all item category probability curves are as depicted below. Now each category is functioning properly with ordered thresholds. For the example of a person with 1 logit less ability than the item difficulty, the probability of endorsing “3” approaches 0 whereas the probability of selecting “0” or “2” is .22 . Category 1 is the most likely choice for individuals in this group with a probability of .54.

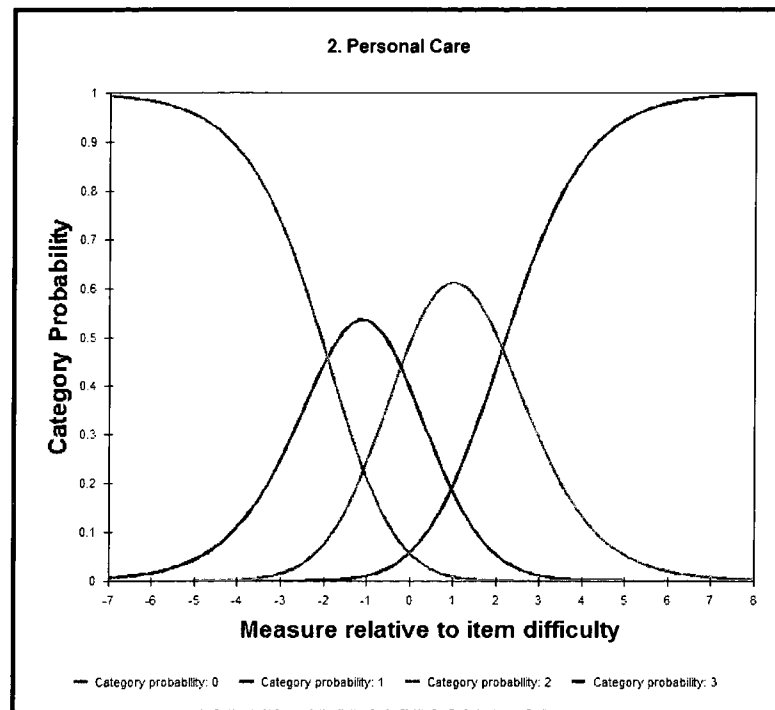


Figure 9 Category probability curve for the 4-point DPQ rating scale

An improvement from 5.86 to 10.93 in the item separation index was seen following the transformation of the original DPQ to a four point scale with the “Pain” item removed.

Item Difficulty

Item difficulty is illustrated with the Item Characteristic Curves shown in Figure 9. In contrast to the results on the ODI for similar items, the “Vocational” item was easiest to endorse, followed by “Social Life” and then “Lifting”. “Standing”, which had been the easiest to endorse on the ODI, was fifth on the DPQ. “Personal Care” was hardest to endorse on both scales. Results from test equating could prove interesting.

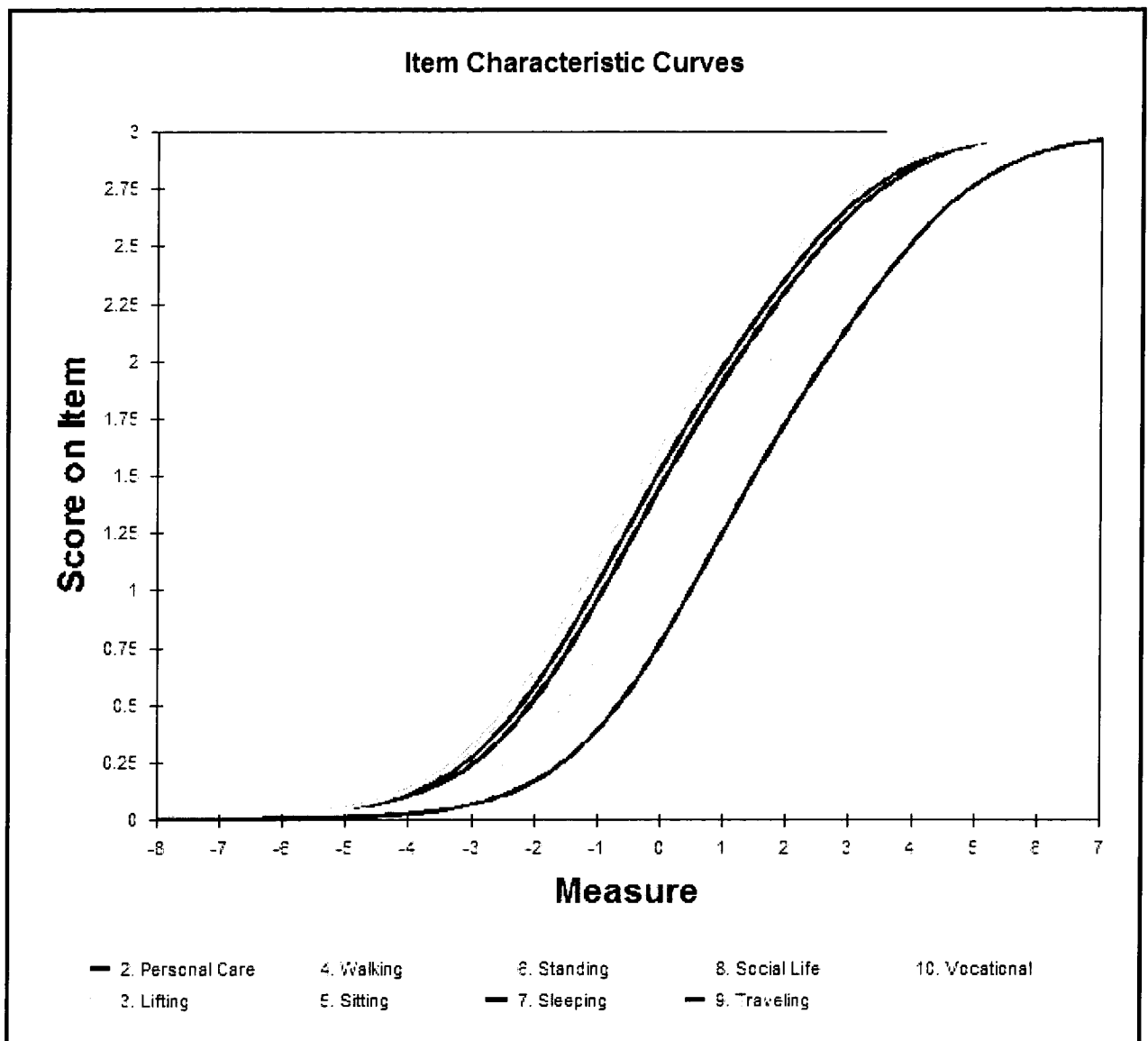


Figure 10 Item characteristic curves depicting item endorsement for the 4-point DPQ

Differential Item Functioning (DIF)

DIF analysis showed no significant difference in scale function for men or women.

PACT Spinal Function Sort

The PACT Spinal Function Sort is a Pictorial Activity Sort consisting of 48 Likert Scale items with two items repeated as a reliability check. The test-taker grades his or her ability to perform each task on a five point scale from “able” to “unable”. Category scores are added and multiplied by factors from one to four and then summed to give a total score out of a maximum of 200. The result is presented as the individual’s Rating of Perceived Capacity (RPC) which is based on the DOT Physical Demand Characteristics of Work outlined in Appendix 7. Data for 260 respondents were analysed using the WINSTEPS computer program. Since this is a true Likert scale with each item having an equal number of response options, Andrich’s Rating Scale Model was the appropriate model.

The initial analysis revealed six misfitting items with mean squares greater than 1.4. These are shown in Table 6.

Table 6 *Spinal Function Sort Misfitting Items*

ITEM	PT-MEASURE (<i>r</i>)	OUTFIT MNSQ	INFIT MNSQ
02. Retrieve/tool/floor	.59	1.75	1.27
49. Paint brush/eye level	.60	1.74	1.39
21. Light Bulb Overhead	.62	1.66	1.46
19. Wash Dishes Sink	.62	1.64	1.23
37. Climb Step Ladder	.62	1.59	1.50
22. Install Face Plate	.65	1.43	1.03

Diagnostic Measures for the SFS

It should also be noted that Item 49 “Paint brush at eye level” is the duplicate of Item 17 included as a reliability check. Item 17 showed mean squares of 1.04 and 1.01 with associated ZSTDs of 0.4 and 0.1 whereas Item 49 had mean squares of 1.39 and 1.74 with associated ZSTD of 2.7 and 4.1. The other reliability check, Items 6 and 50 “Place/retrieve 5# weight waist to overhead” performed differently as well which questions their use as a reliability check. Item difficulty varied for “Place/retrieve 5#...” as well. Most respondents found it easier to endorse this item when it appeared as item 50 after responding to questions regarding weights ranging from 20 to 100 lbs. rather than as item 6 where it was anchored by items of a light nature.

The six items shown in Table 6 were removed and the data were analysed again. In this second analysis, three new items demonstrated mean squares above 1.4. Each time the offending items were removed, new ones cropped up. The attempt to reduce items to represent a unidimensional scale was a failure.

Item Difficulty

With the number of items on the SFS, it is not feasible to look at the ICCs to assess item difficulty. Another pictorial method is the Item Person Map as depicted in Figure 11.

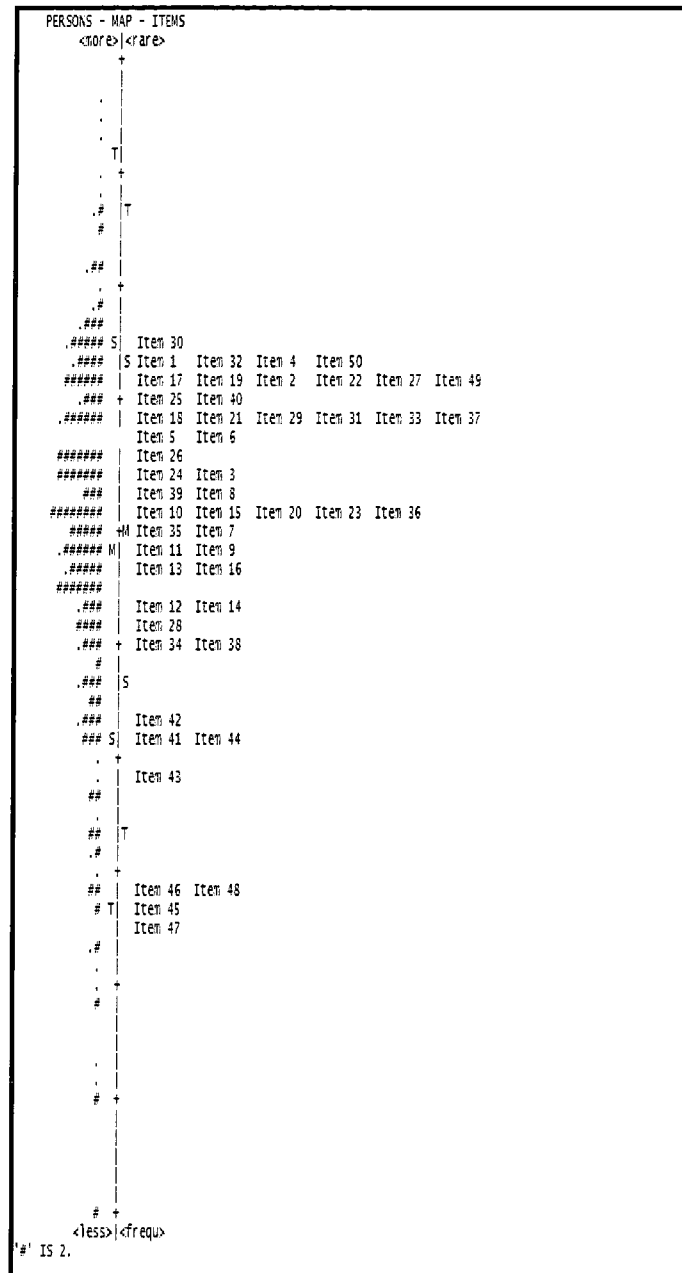


Figure 11 Item Person Map for Spinal Function Sort

In Figure 11, items are displayed on the right and people on the left. Each “#” represents two people. The items are ranked from most difficult on the bottom to easiest on the top. It is not surprising to see items “41” to “48” at the bottom. These related to material handling of weights in the 50 to 100 lbs. range. Item 30 “Get into driver’s seat” is the easiest. The Items “6” and “50” despite being identical are at differing levels of difficulty.

Differential Item Function (DIF) for the SFS

Most items were problematic due to their outfit mean squares. Outfit mean squares are sensitive to off-target responses by persons on items that are at the subject’s ability level or on-target responses to items that are distant to the subject’s ability level. Removal of 20 individuals from the data set based on an outfit mean square and ZSTD greater than 2.0 did not improve the functioning of the items in Table 6. To determine if the items were functioning differently for subgroups, a differential item function (DIF) analysis was undertaken. Figure 12 visualizes the differences in item DIF size by gender.

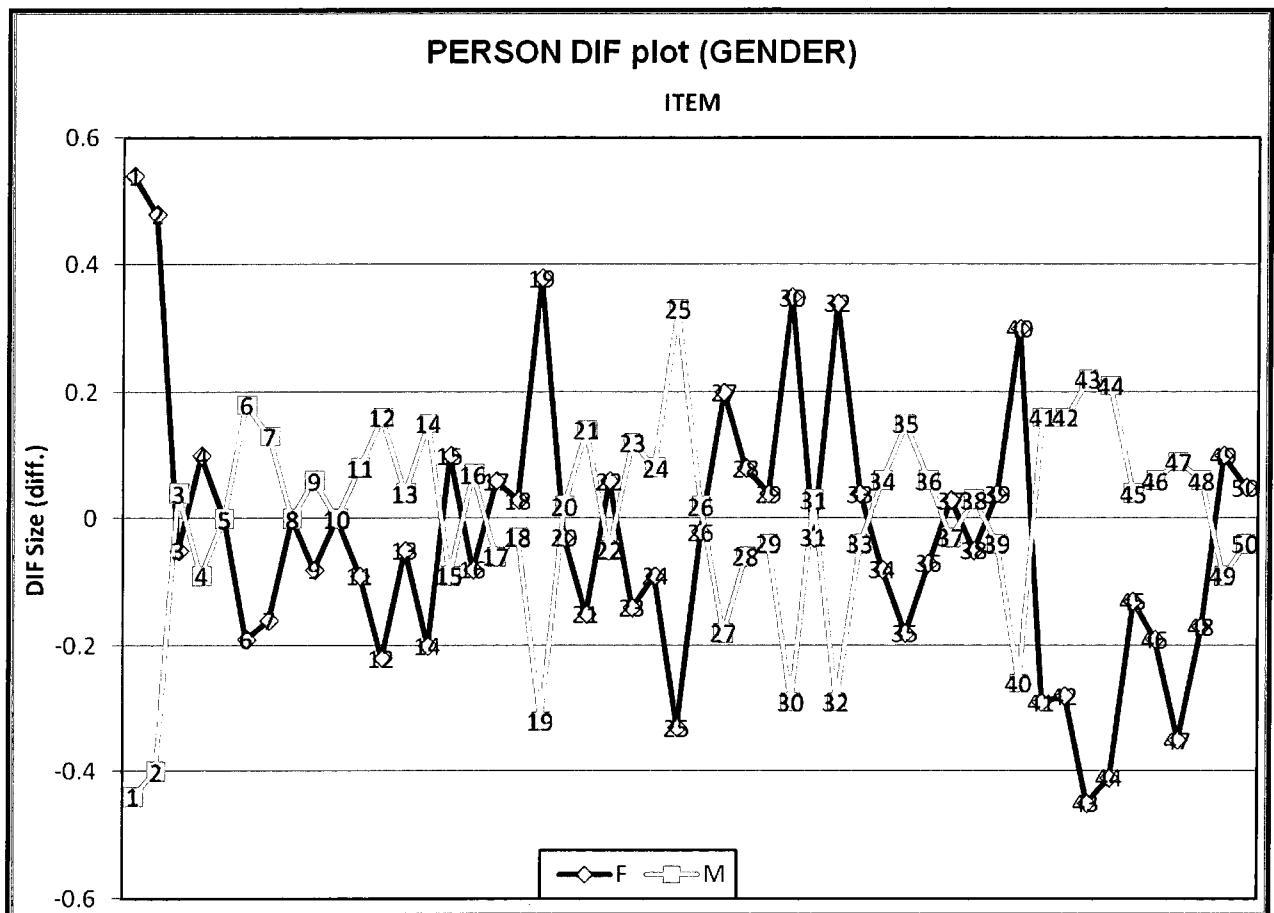


Figure 12 Spinal Function Sort Item DIF size by gender.

Figure 12 shows the size of the item DIF in logits for each group relative to the difficulty of each item. It shows that the items do function differently by gender. This plot shows that for Item #19 "Wash Dishes at Sink" women reported a higher estimation of their ability at this task than the men did. Similarly, for the medium to heavy material handling tasks (items 41 to 48), the men estimated higher ability than the women did. In fact, a pattern emerged during the analysis where males reported lower abilities on traditionally female tasks such as dishwashing (item 19) and kitchen floor sweeping (item 40) despite these tasks being less physically demanding than other tasks that they indicated they were capable of performing. It was interesting to note that the same men, who were restricted in ability with regard to

sweeping with a kitchen broom, reported a higher perception of ability when sweeping the push broom which is commonly used in industrial settings as well as in the garage and workshop. One item that contradicted this trend was Item 3 “push and pull a vacuum cleaner”. In this item a man is depicted performing this task. These results may be specific to the population tested. Many of the participants were from more traditional cultures and the bulk of the participants were over 40 years old. The sample of males under 40 was not large enough for a comparison. All participants were from the northern interior of the British Columbia. This instrument may perform more effectively in other geographic regions.

Neck Disability Index

The 10-item Neck Disability Index (NDI) is the most widely used measure for assessing the effect of neck pain on activities of daily living. It was developed by Vernon and Mior (1991) by adapting five of the scales from the Oswestry Disability Index – Pain, Self Care, Lifting, Sleeping, Travelling (Driving) and developing five new scales identified by literature review and consultation with clinicians. There was minimal input from patients in the development of this scale. Scoring is done exactly as for the ODI with the same resultant disability categories.

Data collected from 76 participants were evaluated using the WINSTEPS program in a manner similar to the analysis of the ODI. The initial NDI data is reported in Table 7.

Table 7 *Diagnostic Measures for NDI*

Item	PT-MEASURE (<i>r</i>)	INFIT MNSQ.	OUTFIT MNSQ.
1. Pain	.74	0.76	0.76
2. Self care	.69	0.96	0.96
3. Lifting*	.36	1.41	1.85
4. Reading	.70	0.71	0.70
5. Headache*	.54	1.41	1.40
6. Concentration†	.78	0.66	0.74
7. Work	.53	1.20	1.20
8. Driving†	.76	0.63	0.63
9. Sleeping	.62	1.03	1.00
10. Recreation	.53	1.13	1.18

*Equal to or greater than Linacre's 1.4 upper limit.

†Equal to or less than Linacre's 0.6 lower limit.

Diagnostic Measures for the NDI

The point measure correlations range from .36 to .78. At .36 lifting correlates poorly with the overall measure. It can be noted that the items "Lifting" and "Headache" have outfit and infit mean squares higher than 1.4 with corresponding ZSTD greater than 2. Both items display disordered thresholds when the Modal Probability curves are displayed.

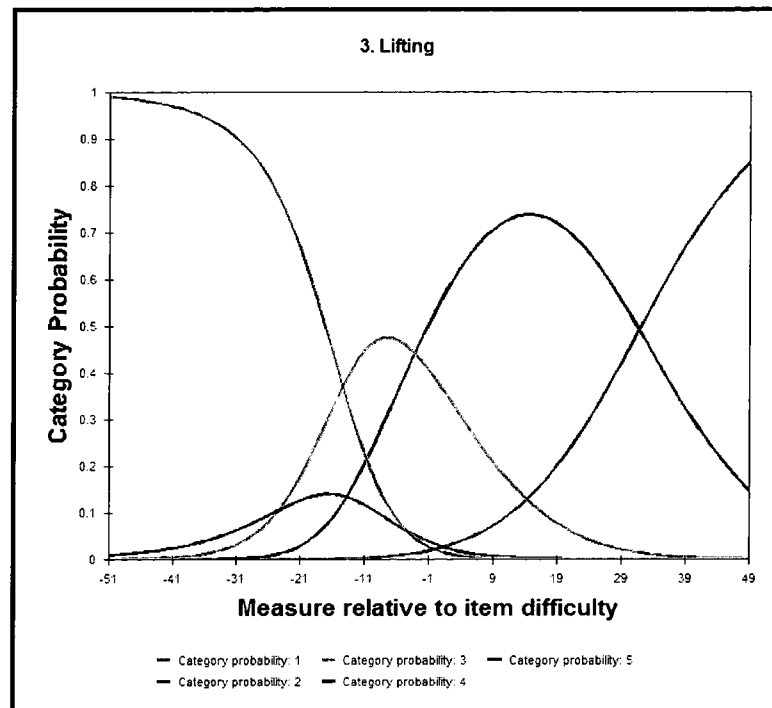


Figure 13 Category Probability Curves NDI “Lifting” item #3

In Figure 13 for the “Lifting” item, no one has chosen category 0 - “I can lift heavy weights without extra pain”. Category 2 “Pain prevents me from lifting heavy weights from the floor...” shows disordered thresholds. On the “Headache” Item of the Neck Disability Index depicted in figure 14, it can be noted that category 2, “I have moderate headaches which come infrequently” is not functioning well. The fourth category, “I have severe headaches which come frequently” is also not functioning well.

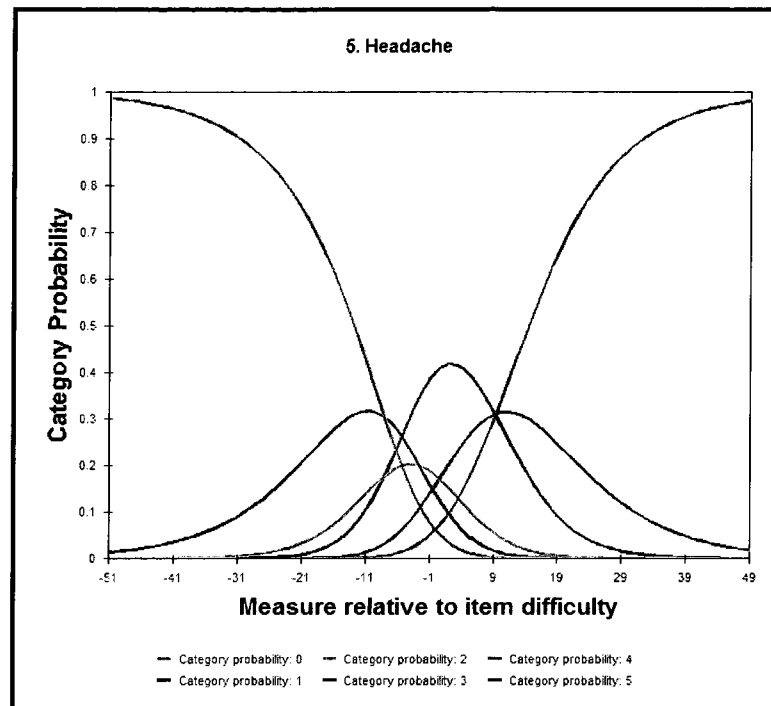


Figure 14 Category probability curves for the NDI “Headache” item #5

Several unsuccessful attempts were made to rescale the NDI in a way that would allow the retention of the “Lifting” and “Headache” items. Examination of the items themselves show that the response category labels for each are poorly worded and contain more than one concept, for example, severity and frequency of headache are contained in the same choice. Items should contain a single statement that allows for degrees of endorsement.

With the above-mentioned items removed, a new 8-item NDI scale remained. For this new scale, person reliability improved from .82 to .87 and item reliability improved from .95 to .96. However, disordered categories were seen in the “work”, “sleeping” and “reading” items which differ from the findings of van der Velde et al., (2009) where they found disordered thresholds for “personal care”, “work” and “recreation” items. With this

small data set, many of the choices had fewer than the 10 observations suggested by Linacre (2009) so meaningful rescaling could not be performed.

Item Difficulty

Item difficulty is presented in Figure 15. Similar to the findings on the ODI and the DPQ, personal care is the most difficult to endorse. Otherwise, item difficulty varied depending on the instrument used.

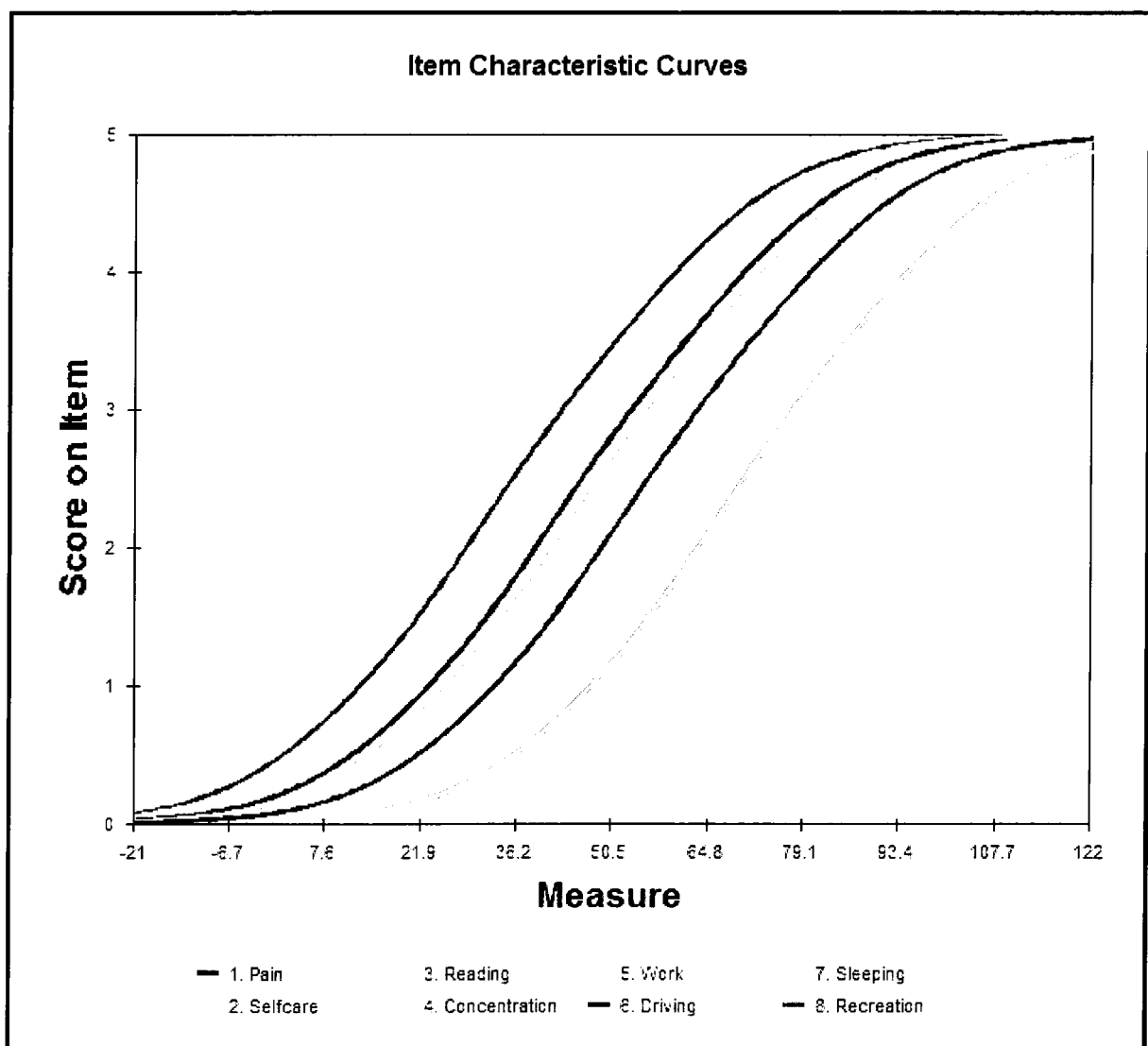


Figure 15 Item Difficulty for the NDI

Differential Item Functioning

There was no DIF by gender as illustrated by the chart in Figure 16. While there was a difference in “Driving”, it was not significant for this sample.

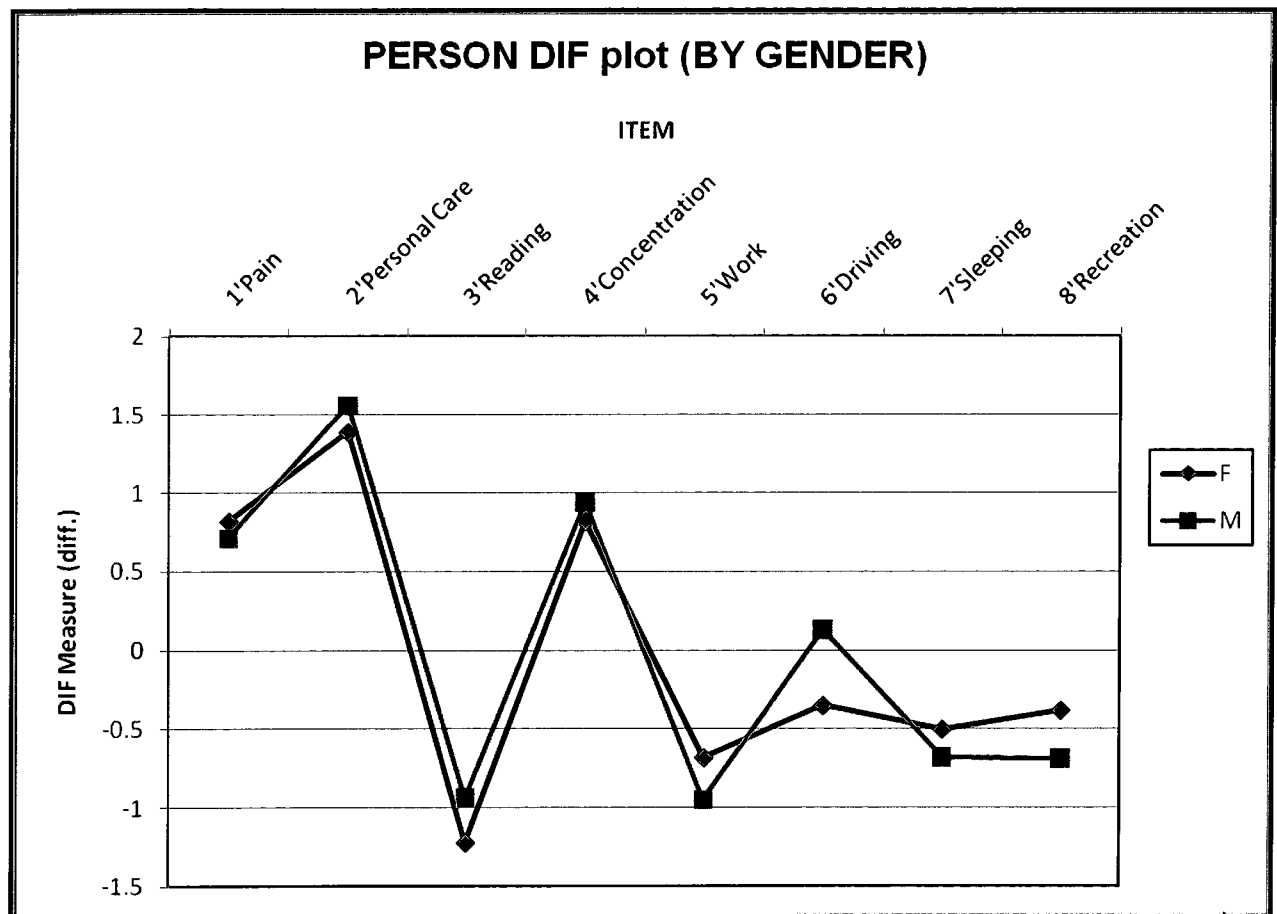


Figure 16 DIF by Gender for the NDI

In the analysis of misfitting persons, a pattern emerged indicating that individuals with chronic conditions or multiple injuries were more likely to be misfitting. Sample size was not large enough to divide the participants by diagnosis and run a DIF analysis but this does suggest that the instrument may function differently with different diagnoses.

CHAPTER FIVE: DISCUSSION

The purpose of this study was to evaluate the psychometric properties of four of the self-report instruments commonly used in Functional Capacity Evaluations to answer the question “Are the client’s subjective reports reliable?” This is important as these measurement instruments are currently being used by Work Capacity Evaluators to determine reliability of the client’s subjective reports as compared to demonstrated abilities. This can potentially affect the continuation of disability benefits and allotment of funds for rehabilitation as well as monetary awards in litigation. Rasch Modeling, a prescriptive method offering distribution-free person ability and item difficulty estimates on a linear latent variable, was used to evaluate the ODI, DPQ, SFS and NDI with surprising results. As a clinician who has used these instruments for the past ten years, I was astounded by the poor performance of the Spinal Function Sort and Neck Disability Index. The Dallas Pain Questionnaire fared somewhat better and the Oswestry Disability Index was the best performer.

The Oswestry Disability Index (ODI)

Three previous studies have reported on the measurement properties of the Oswestry Disability Index based on a Rasch analysis. Although these studies are discussed in the Literature Review, I provide brief summaries of each previous study for ease of comparison with my study. Davidson (2008) compared three variations of the ODI, Version 1, Version 2 and the Revised ODI, using a convenience sample of 100 individuals, 40% with a duration of current episode greater than six months and 63% who had experienced more than five

previous episodes of back pain. She reported that while Versions 1 and 2 met the criteria for unidimensionality with her population, the revised version did not.

Page et al. (2002) reported that with a sample of 95 patients and a mean time since initial symptom presentation of 2.3 ± 0.6 wk., the revised version of the ODI did not demonstrate unidimensionality. The authors postulated that:

although item 1 purports to measure LBP (low back pain) disability intensity, it does so by asking patients to what extent painkillers reduce their LBP disability. Many patients in our clinic did not take painkillers, although they reported substantial LBP disability intensity. Moreover, for many patients, the extent to which painkillers reduce LBP is not a direct way to functionally assess LBP disability intensity.

The first item, “Pain Intensity”, was removed and disordered thresholds were addressed by collapsing categories “2” with “3” and “4” with “5”. The authors reported improved precision of the instrument.

White and Velozo (2002) reported on a sample of 942 patients with low back pain presenting for physical therapy, 70% of whom had experienced symptoms for less than six months and 50% of whom were working at their regular place of employment. They found that all items except the Item 1 “Pain Intensity” fit the Rasch model and that “construct validity of the scale using the Rasch model, required the structure of the rating scale to be modified from six response levels to four.”

In my study, using a sample of 133 patients presenting for functional capacity evaluation, I found that all the items on the ODI Version 1 fit the Rasch Model when the criteria of Mean square fit indices between 0.6 and 1.4 were applied. Granted, the fit indices for items 7 and 8, “sleeping” and “sex life”, respectively were marginally “noisy” with values ranging from 1.39 to 1.48. I considered this to be insufficient evidence to warrant removal of these items, particularly as scale category ordering had not yet been addressed. Two items

exhibited fit values ranging between 0.62 to 0.65; constrained but acceptable values. In contrast to the results of Page et al. (2002) and White and Velozo (2002), Item 1 “Pain Intensity” fit well even before category reduction. To correct disordered thresholds, the scale was revised from six categories to five on an item by item basis. Again, this decision to reduce to five categories differs from that of Page et al. (2002) and White and Velozo (2002). However these authors employed a common aggregation of categories across all items. My approach was more pragmatic and data-driven as categories were collapsed on an item by item basis. I judged this to be an appropriate strategy as each item had different descriptors for each scale point – more Partial Credit than Rating Scale. The revised instrument met the criteria for unidimensionality with no borderline indices, either high or low.

Rasch analysis purports to be a sample-free measurement model, so this variation in whether the scale including Item 1 fits or does not fit the model was initially of concern to me. One possible explanation is that the latent trait of perceived disability might be different in acute vs. chronic or working vs. not working populations. Davidson and I both examined populations which were largely chronic i.e. more than five recurrences or longer than six months duration of symptoms whereas White and Velozo (2002) and Page et al. (2002) report on patients with shorter durations of symptoms who were presenting for physiotherapy treatment. Almost 100% of my participants were taking some form of pain medication and pain management was a significant part of their daily activities – often the most significant part. In the clinical setting, my colleagues and I have noted that pain rather than ability becomes the limiting factor for participation of individuals who have chronic pain. Further assessment of the ODI with the two populations would be recommended.

Dallas Pain Questionnaire (DPQ)

There are no published works to date on Rasch analysis of the DPQ. Lawlis et al. (1989) developed this instrument with varying scale lengths which they justified by differential weighting of items. They indicated that in pilot studies, some of the items seemed to impact the construct of perceived disability more than others. They gave the ones with the highest impact more segments to reflect this weighting. I do not understand how giving an individual more selection options in a given item accomplishes the authors' goal as outlined above. Initial Rasch analysis of this instrument revealed a lack of unidimensionality of the scale and disordered thresholds particularly in the items with more segments. When the scale was reduced to 9 items with 4 categories each, the DPQ demonstrated satisfactory internal reliability and construct validity as indicated by the Rasch analysis.

In contrast to the ODI findings where the pain item did fit the Rasch model, in this case the pain item did not function as part of a unidimensional scale. One explanation could be that while the ODI item refers to pain management, the DPQ simply asks the individual to indicate the level of pain they are experiencing from "No Pain" at one end of the scale to "Worst Imaginable Pain" at the other.

Ozguler et al. (2002) had changed the scale to 10 segments reporting that this new homogeneous instrument functioned well to measure the impact of spinal pain on behavior. While uniform scale length was definitely a step in the right direction, ten segments might be excessive. In my analysis, I found that the scale functioned well with 4 segments.

Spinal Function Sort (SFS)

For me, the poor performance of the SFS when analyzed with the Rasch model was the most disappointing. Within the rehabilitation field, the creator of the SFS, Leonard Matheson, is a well-respected researcher who was a pioneer in functional capacity evaluation and who has contributed a great deal to the advancement of the field. The problems with the SFS – lack of unidimensionality, disordered thresholds and lack of local independence, could not be overcome by eliminating items or persons or rescaling items. This was the only instrument that demonstrated DIF where items functioned differently by gender. It appears that men do not think that they can do dishes even if they are physically capable of much heavier tasks and likewise women do not like to use any tools. These types of task items seem to measure gender roles more than ability. The SFS does not meet the requirement of local independence, a basic tenet of the Rasch model. The order of presentation of the items from lighter to heavier tasks and the clustering of similar tasks influences the response for one item with the response on the similar item. The reliability check which consists of two items repeated at the end of the test is ineffective. Item difficulty of Item 6 and 50 “place or retrieve a 5 lb weight between waist and overhead” was different depending on where it was placed in the test. Respondents score it as an easier task after completing the questions related to material handling of 50 to 100 lbs. In my experience, respondents quite often remember that this item has been presented earlier in the test and flip back to check their previous answer.

Neck Disability Index (NDI)

Since the conception of this project, a Rasch analysis of the NDI has been published. Van der Welde et al. (2009) reported on a sample of 521 trial subjects fit to the Rasch model. They reported on a lack of fit of the data to the model and disordered response thresholds in “personal care”, “lifting”, “headaches”, “work” and “recreation”. They eliminated two items, “headaches” and “lifting” and developed an eight item scale that demonstrated unidimensionality. They chose not to address the disordered thresholds as:

this would have precluded the possibility of providing a straightforward exchange between the everyday summed ordinal score and its corresponding interval score. Furthermore, collapsing the scale would have resulted in a varying number of categories across items, which represents a considerable change from the original design of the NDI scale.

They suggest that the disordered response thresholds be examined in other samples to see if the problem is generic.

In my analysis, I also found that “lifting” and “headache” did not form part of a unidimensional construct relating to perceived disability. I found disordered thresholds albeit in different items but I also found that I was unable to effectively collapse categories to improve these items. I believe that the disordered thresholds cannot be fixed because the instrument itself, despite its popularity and wide-spread use, is poorly designed. The response categories are confusing to respondents as they often contain more than one concept such as pain and function. For instance, in the driving category, two adjacent categories present as follows. “I can drive my car as long as I want with moderate pain in my neck.” “I can’t drive my car as long as I want because of moderate pain in my neck.” It was noted that the response categories designed to measure the highest levels of neck disability such as “I cannot read at all” were rarely or never endorsed by participants.

Conclusion

The self-report questionnaires included as part of the Matheson Functional Capacity Evaluation Software package did not meet expectations. The original intent of this work had been to equate these measures to see if questionnaire selection and administration could be streamlined but this was not possible due to the poor performance of these instruments when analyzed with a modern psychometric approach, Rasch Modeling. As Linacre (2009) says “if tests don’t make sense separately, they won’t make sense together.”

Limitations of Design

The focus of this measurement thesis is on the internal validity of the tools. There is no content validity from experts in the field in relation to scale reduction decisions. There is no linkage between patient scores to function in the job setting. No follow up with subjects is possible.

Recommendations for Practitioners

These instruments should never be used to replace clinical observation and judgment. Scores should be used as a contribution toward decisions regarding symptom magnification but only as small part of a bigger picture. Clinicians should become familiar with the measurement properties of any instrument they use and critically evaluate the methods used to obtain the results. Better instruments may be available for use as part of the Functional Capacity Evaluation. It is the responsibility of the clinician to use reliable and valid instruments when measuring and reporting on symptom magnification.

Recommendations for Future Research

Oswestry Disability Index

Pilot testing of the scale as outlined in Appendix 8 as well as testing the new instrument with acute and chronic back pain populations.

Dallas Pain Questionnaire

Since the new scale length was developed by collapsing categories post hoc, further research is required to test the proposed new four-segment, nine-item homogeneous scale.

Spinal Function Sort

A better model for an activity sort such as this would be as a computer administered test where pictures are displayed in a random order and gender neutrality is maintained. A test battery for women that depicts all tasks being performed by women and likewise a similar test battery for men would reduce or eliminate bias by matching the test to the gender of the test taker. Further effort to identify tasks that depict an activity with given physical demands that is equally likely to be performed by both genders would improve the instrument or the development of parallel tests specific to gender could be another route to go. With the advancement of technology, computer tests can be easily adapted to fit any given situation.

Neck Disability Index

Although DIF by gender was not significant in my sample nor in the much larger sample used by van der Velde et al. (2009), it would be interesting to see if the NDI functions differently by diagnosis. Van der Velde et al. (2009) excluded patients with neck pain that was not mechanical in nature. They also excluded clients with third-party liability or

compensation claims as well as individuals with co-existing problems. My sample was too small to divide by diagnosis but analysis of acute vs. chronic, multiple areas (i.e. neck and back or neck and shoulder) vs. single area (neck) would be recommended.

I was unable to correct the disordered thresholds in this instrument despite numerous attempts to collapse the categories. The NDI is often confusing to respondents and response categories contain more than one concept. It may be unsalvageable. There are a plethora of other measurement tools for neck disability so it would be worthwhile to further investigate alternatives to the NDI.

REFERENCES

- Ackelman, B.H., & Lindgren, U. (2002). Validity and reliability of a modified version of the neck disability index. *Journal of Rehabilitation Medicine, Vol. 34*: 284–287.
- Andrich, D. (1978). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement, 2*(4), 581-594.
- Beurskens, A., de Vet, H., & Köke, A. (1996, April). Responsiveness of functional status in low back pain: a comparison of different instruments. *Pain, 65*(1), 71-76.
- Birnbaum, A. (1968) "Some Latent Trait Models and Their Use in Inferring an Examinee's Ability," in F. M. Lord and M. R. Novick, *Statistical Theories of Mental Test Scores*, Reading, MA: Addison–Wesley.
- Bond, T.G. & Fox, C.M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.
- Christensen, F.B., Laursen, M., Gelineck, J., Hansen, E.S. and Bünger, C.E. (2001). Posterolateral spinal fusion at unintended levels due to bone-graft migration. *Acta Orthopaedica Scandinavica Vol 72 (4)*: 354–358
- Cleland JA, Childs JD, Whitman JM. (2008). Psychometric properties of the Neck Disability Index and numeric pain rating scale in patients with mechanical neck pain. *Archives of Physical Medicine & Rehabilitation; 89*:69-74.
- Cleland JA, Childs JD, Whitman JM. (2008). Response to Vernon H. Letter to the Editor re Psychometric properties of the Neck Disability Index and numeric pain rating scale in patients with mechanical neck pain. *Archives of Physical Medicine & Rehabilitation; 89*:1415-1416.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart & Winston.
- Cronbach, Lee J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334.
- Davidson, M. (2008). Rasch analysis of three versions of the Oswestry Disability Questionnaire. *Manual Therapy, 13*(3), 222-231.
- Davidson, M.B., Keating, J.L., Eyres, S. (2004). A low back-specific version of the SF-36 Physical Functioning Scale. *Spine Vol. 25*(5) 586-594.

- Fairbank J.C., Couper J, Davies J.B., O'Brien, J.P. (1980). The Oswestry low back pain questionnaire. *Physiotherapy Vol 66*: 271–3.
- Fairbank, J.C.T. & Pynsent, P.B. (2000). The Oswestry disability index. *Spine, Vol. 25, No. 2*: 2940-2952.
- Fan, X. (1998). Item response theory and classical test theory: an empirical comparison of their item-person statistics. *Educational And Psychological Measurement, 58(3)*, 357-381.
- Gibson, L. & Strong, J. (1996). The reliability and validity of a measure of perceived functional capacity for work in chronic back pain. *Journal of Occupational Rehabilitation, Vol. 6, No. 3*: 159-175.
- Greenough, C.G. & Fraser, R.D. (1992). Assessment of outcome in patients with low-back pain. *Spine, Vol. 17*: 36-41.
- Grönblad, M., Hupli, M., Wennerstrand, P., Järvinen, E., Lukinmaa, A., Kouri, J., et al. (1993). Intercorrelation and test-retest reliability of the Pain Disability Index (PDI) and the Oswestry Disability Questionnaire (ODQ) and their correlation with pain intensity in low back pain patients. *The Clinical Journal of Pain, 9(3)*, 189-195.
- Hains F., Waalen J; Mior S (1998). Psychometric properties of the neck disability index. *Journal of Manipulative and Physiological Therapeutics, Vol. 21 (2)*, 75-80.
- Hudson-Cook N., Tomes-Nicholson K., Breen A. A. revised Oswestry disability questionnaire. In: Roland M, Jenner JR, eds. *Back Pain: New Approaches to Rehabilitation and Education*. (pp. 187–204). Manchester: Manchester University Press.
- Huskisson, E.C. (1974). Measurement of pain. *Lancet, Vol. 2*, 1127-1131.
- Isernhagen, S., (1995). Contemporary issues in functional capacity evaluation, in S. Isernhagen (Ed.): *The Comprehensive Guide to Work Injury Management*, (pp. 410–429). Gaithersburg: Aspen Publishers.
- Karabatsos, G., (1999) *Axiomatic measurement theory as a basis for model selection in item-response theory*. Paper presented at the 32nd annual conference of the Society for Mathematical Psychology, Santa Cruz, CA.
- Kopec, J., Esdaile, J., Abrahamowicz, M., Abenhaim, L., Wood-Dauphinee, S., Lamping, D., et al. (1995). The Quebec Back Pain Disability Scale: measurement properties. *Spine, 20(3)*, 341-352.

- Lawlis, G.F., Cuenas, R., Selby, D., and McCoy, C.E. (1989). The development of the Dallas Pain Questionnaire. An assessment of the impact of spinal pain on behavior. *Spine May; Vol. 14 (5)*, pp. 511-6.
- Linacre, J.M. (2009). Winsteps (Version 3.68.0) [Computer Software]. Chicago: Winsteps.com.
- Lord FM (1952) A theory of test scores. Psychometric Monograph No 7. Psychometric Society, New York.
- Marty, M., Blotman, F., Avouac, B., Rozenberg, S., & Valat, J. (1998). Validation of the French version of the Dallas Pain Questionnaire in chronic low back pain patients. *Revue du Rhumatisme (English Ed.)*, 65(2), 126-134.
- Matheson, L.N. (1988). How do you know he tried his best? *Journal of Industrial Rehabilitation Quarterly*, 1, 10-12
- Matheson L., Mayer J., Mooney V., Sarkin A., Dreisinger T., Verna J., Leggett S (2008). A method to provide a more efficient and reliable measure of self-report physical work capacity for patients with spinal pain. *Journal of Occupational Rehabilitation*, Vol. 18 (1): 46-57
- Matheson L.N. & Matheson M.L., Grant J. (1993). Development of a measure of perceived functional ability. *Journal of Occupational Rehabilitation*, Vol. 3: 15-30.
- Matheson, L.N. (2004). History, design characteristics, and uses of the pictorial activity and task sorts. *Journal of Occupational Rehabilitation*, Vol. 14, No. 3, 175-195.
- Matheson, L.N., & Matheson, M.L. (1989). *Spinal function sort*. Rancho Santa Margarita, CA: Performance Assessment and Capacity Testing.
- Matheson, Roy & Associates (2006). *The functional capacity evaluation certification program manual*. Keene, NH: Published in-house.
- Melzack, R. (1975). The McGill Pain Questionnaire: major properties and scoring methods. *Pain*, Vol. 1(3), 277-299.
- Ozguler, A., Gueguen, A., Leclerc, A., Landre, M., Piciotti, M., Le Gall, S., Morel-Fatio, M., and Boureau, F. (2002). Using the Dallas Pain Questionnaire to classify individuals with low back pain in a working population. *Spine Vol. 27, No. 16*: 1783-1789.
- Page, S., Shawaryn, M., Cernich, A., & Linacre, J. (2002, November). Scaling of the Revised Oswestry Low Back Pain Questionnaire. *Archives of Physical Medicine & Rehabilitation*, 83(11), 1579-1584.

- Pollard, C.A. (1984). Preliminary validity study of the pain disability index. *Journal of Perceptual Motor Skills*, Vol. 59(3), 974.
- Pomeranz, J.L., Byers, K.L., Moorhouse, M.D., Velozo C.A., Spitznagel R.J., (2008). Rasch analysis as a technique to examine the psychometric properties of a career ability placement survey subtest. *Rehabilitation Counseling Bulletin*, Jul; 51 (4): 251-9.
- Ransford, A., Cairns, D., & Mooney, V. (1979). The pain drawing as an aid to the psychological evaluation of patients with low back pain. *Spine*, Vol. 1, 127-134.
- Robinson R.C., Kishino N., Matheson L.N., Woods S., Hoffman K., Unterberg J., Pearson C., Adams L., Gatchel R. (2003). Improvement in postoperative and nonoperative spinal patients on a self-report measure of disability: The Spinal Function Sort (SFS). *Journal of Occupational Rehabilitation*, Vol. 13, No. 2, 107-113.
- Roche, G., Ponthieux, A., Parot-Shinkel, E., Jousset, N., Bontoux, L., Dubus, V., Penneau-Fontbonne, D., Roquelaure, Y., Legrand, E., Colin, D., Richard, I., Fanello, S., (2007). Comparison of a functional restoration program with active individual physical therapy for patients with chronic low back *pain*: a randomized controlled trial. *Archives of Physical Medicine and Rehabilitation*, Vol. 88 (10), 1229-35.
- Shaw, F. (1991). Descriptive IRT vs. Prescriptive Rasch, *Rasch Measurement Transactions*, Vol. 5:1, 131.
- Stevens, S.S. (1946). On the theory of scales of measurement. *Science*, 103, 677-680.
- Streiner, D.L. & Norman, G.R. (1989). *Health Measurement Scales: A Practical Guide to Their Development and Use*. New York: Oxford University Press, Inc. pages. 64-65.
- Strong, J., Ashton, R., & Large, R. (1994). Function and the patient with chronic low back pain. *The Clinical Journal Of Pain*, 10(3), 191-196.
- Thew, Kimberley A. (2007). An examination of the perceptions of functional capacity evaluations in Prince George, British Columbia: A case study. M.A. dissertation, University of Northern British Columbia (Canada).
- Thurstone, L.L. (1927). The unit of measurement in educational scales. *Journal of Educational Psychology*, 16, 433-451.
- Tsutsumi, A., Iwata, N., Wakita, T., Kumagai, R., Noguchi, H. & Kawakami, N. (2008). Improving the Measurement Accuracy of the Effort-Reward Imbalance Scales, *International Journal of Behavioral Medicine*, 15:2, 109-119.

- van der Velde, G., Beaton, D., Hogg-Johnston, S., Hurwitz, E., & Tennant, A. (2009). Rasch analysis provides new insights into the measurement properties of the neck disability index. *Arthritis & Rheumatism*, 61(4), 544-551.
- Vernon H., Mior S. (1991). The neck disability index: A study of reliability and validity. *Journal of Manipulative Physiological Therapeutics*, Vol 14: 409-15.
- Vernon, H., (2008). Letter to the Editor in Response to Cleland et al, Psychometric properties of the Neck Disability Index and numeric pain rating scale in patients with mechanical neck pain. . *Archives of Physical Medicine & Rehabilitation*, 89:1414-1415.
- Walsh, Thom (2000). Point of view re: the Oswestry disability Index, *Spine*, Vol. 25, No. 2: 2953.
- White, L., Velozo, C. (2002). The use of Rasch measurement to improve the Oswestry classification scheme. *Archives of Physical Medicine & Rehabilitation*, 83(6), 822-831.
- Wright, B.D., Linacre M., Gustafsson, J-E. and Mrtin-Loff, P. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8(3), 370. Available from <http://www.rasch.org/rmt/rmt83.htm> accessed March 29, 2009
- Wright BD, Masters GN. (1982). *Rating scale analysis*. Chicago: Mesa Press.

Appendix 1 – Oswestry Disability Index – Version 1.0

This questionnaire has been designed to give the doctor information as to how your back pain has affected your ability to manage in every day life. Please answer every section, and mark in each section only the *one* box which applies to you. We realize you may consider that two of the statements in any one section relate to you, but please just *mark the box which most closely describes your problem*.

Section 1 – Pain Intensity

- ☐ I can tolerate the pain I have without having to use painkillers
- ☐ The pain is bad but I manage without taking painkillers
- ☐ Painkillers give complete relief from pain.
- ☐ Painkillers give moderate relief
- ☐ Painkillers give very little relief from pain.
- ☐ Painkillers have no effect on the pain and I do not use them.

Section 2 – Personal Care (Washing Dressing etc)

- ☐ I can look after myself normally without causing extra pain.
- ☐ I can look after myself but it causes extra pain.
- ☐ It is painful to look after myself and I am slow and careful
- ☐ I need some help but manage most of my personal care
- ☐ I need help every day in most aspects of self-care
- ☐ I do not get dressed, wash with difficulty and stay in bed

Section 3 – Lifting

- ☐ I can lift heavy weights without extra pain
- ☐ I can lift heavy weights but it gives me extra pain
- ☐ Pain prevents me from lifting heavy weights off the floor, but I can manage if they are conveniently positioned e.g. on a table.
- ☐ Pain prevents me from lifting heavy weights but I can manage light to medium weights if they are conveniently positioned.
- ☐ I can lift only very light weights
- ☐ I cannot lift or carry anything at all.

Section 4 – Walking

- ☐ Pain does not prevent me from walking any distance
- ☐ Pain prevents me walking more than 1 mile.
- ☐ Pain prevents me walking more than ½ mile.
- ☐ Pain prevents me walking more than ¼ mile
- ☐ I can only walk using a stick or crutches
- ☐ I am in bed most of the time and have to crawl to the toilet.

Section 5 – Sitting

- ☐ I can sit in any chair as long as I like
- ☐ I can only sit in my favorite chair as long as I like
- ☐ Pain prevents me from sitting more than 1 hour
- ☐ Pain prevents me from sitting more than ½ hour.
- ☐ Pain prevents me from sitting more than 10 minutes
- ☐ Pain prevents me from sitting at all.

Section 6 – Standing

- ☐ I can stand as long as I want without extra pain
- ☐ I can stand as long as I want but it gives me extra pain
- ☐ Pain prevents me from standing for more than 1 hour
- ☐ Pain prevents me from standing for more than 30 minutes
- ☐ Pain prevents me from standing for more than 10 minutes
- ☐ Pain prevents me from standing at all.

Section 7 – Sleeping

- ☐ Pain does not prevent me from sleeping
- ☐ I can sleep well only by using tablets
- ☐ Even when I take tablets I have less than 6 hours of sleep
- ☐ Even when I take tablets I have less than 4 hours of sleep
- ☐ Even when I take tablets I have less than 2 hours of sleep
- ☐ Pain prevents me from sleeping at all.

Section 8 – Sex Life

- ☐ My sex life is normal and causes no extra pain
- ☐ My sex life is normal but causes some extra pain.
- ☐ My sex life is nearly normal but is very painful
- ☐ My sex life is severely restricted by pain.
- ☐ My sex life is nearly absent because of pain
- ☐ Pain prevents any sex life at all.

Section 9 – Social Life

- ☐ My social life is normal and gives me no extra pain
- ☐ My social life is normal but increases the degree of pain
- ☐ Pain has no significant effect on my social life apart from limiting my more energetic interests such as dancing etc.
- ☐ Pain has restricted my social life and I do not go out as often
- ☐ Pain has restricted my social life to my home.
- ☐ I have no social life because of pain.

Section 10 – Travelling

- ☐ I can travel anywhere without extra pain
- ☐ I can travel anywhere but it gives me extra pain
- ☐ Pain is bad but I manage journeys over 2 hours
- ☐ Pain restricts me to journeys of less than 1 hour
- ☐ Pain restricts me to short necessary journeys of less than 30 minutes
- ☐ Pain prevents me from traveling except to the doctor or hospital.

Name _____ Date _____

Appendix 2 – Dallas Pain Questionnaire

Instructions

Mark an "X" along the line that expresses your thoughts from 0% to 100% in each section. Reach each statement carefully. There are words to help you with each statement. If you need help, please ask.

Section I: Pain Intensity

To what degree do you rely on pain medications or pain relieving substances for you to be comfortable?

None
0% (_____ : _____ : _____ : _____ : _____ : _____) 100%
Some All the time

Section II: Personal Care

How much does pain interfere with your personal care (getting out of bed, teeth brushing, dressing, etc.)?

None
(no pain)
0% (_____ : _____ : _____ : _____ : _____ : _____) 100%
Some I cannot get out of bed

Section III: Lifting

How much limitation do you notice in lifting?

None
(I can lift as I did)
0% (_____ : _____ : _____ : _____ : _____ : _____) 100%
Some I cannot lift anything

Section IV: Walking

Compared to how far you could walk before your injury or back trouble, how much does pain restrict your walking now?

I can walk the same Almost the same Very Little I cannot walk
0% (_____ : _____ : _____ : _____ : _____ : _____) 100%

Section V: Sitting

Back pain limits my sitting in a chair to:

None, pain same as before Some I cannot sit at all
0% (_____ : _____ : _____ : _____ : _____ : _____) 100%

Section VI: Standing

How much does your pain interfere with your tolerance to stand for long periods?

None
Same as before
0% (_____ : _____ : _____ : _____ : _____ : _____) 100%

Some

I cannot
stand

Section VII: Sleeping

How much does pain interfere with your sleeping?

None
Same as before
0% (_____ : _____ : _____ : _____ : _____ : _____) 100%

Some

I cannot
sleep at all

(_____ X 3 = _____ % Daily Activities Interference)

Section VIII: Social Life

How much does pain interfere with your social life (dancing, games, going out, eating with friends, etc.)?

None
Same as before
0% (_____ : _____ : _____ : _____ : _____ : _____) 100%

Some

No activities
total loss

Section IX: Traveling

How much does pain interfere with traveling in a car?

None
Same as before
0% (_____ : _____ : _____ : _____ : _____ : _____) 100%

Some

Cannot
travel

Section X: Vocational

How much does pain interfere with your job?

None
No interference
0% (_____ : _____ : _____ : _____ : _____ : _____) 100%

Some

I cannot
work

(_____ X 5 = _____ % Work/Leisure Activities Interference)

Instructions

Page 3

This is a test of your **current** ability to perform work tasks. There are 50 drawings of work tasks in this booklet. Each drawing has a short description of a work task. Look at each drawing and read the description. On the separate answer sheet, indicate your current level of ability to perform the task in the written description. You do not have to do the task exactly as the drawing. The drawing is meant to help explain the written task description.

If you can perform the task with no difficulty, circle #1, "Able".

Able		Restricted		Unable	
1	2	3	4	5	?

If you cannot perform the task at all, circle #5, "Unable".

Able		Restricted		Unable	
1	2	3	4	5	?

If you can perform the task, but you have some difficulty, circle #2, #3, or #4, "Restricted".

Able		Restricted		Unable	
1	2	3	4	5	?

Be sure to circle only one number. If you circle #2, this would indicate that you are only slightly restricted.

Able		Restricted		Unable	
1	2	3	4	5	?

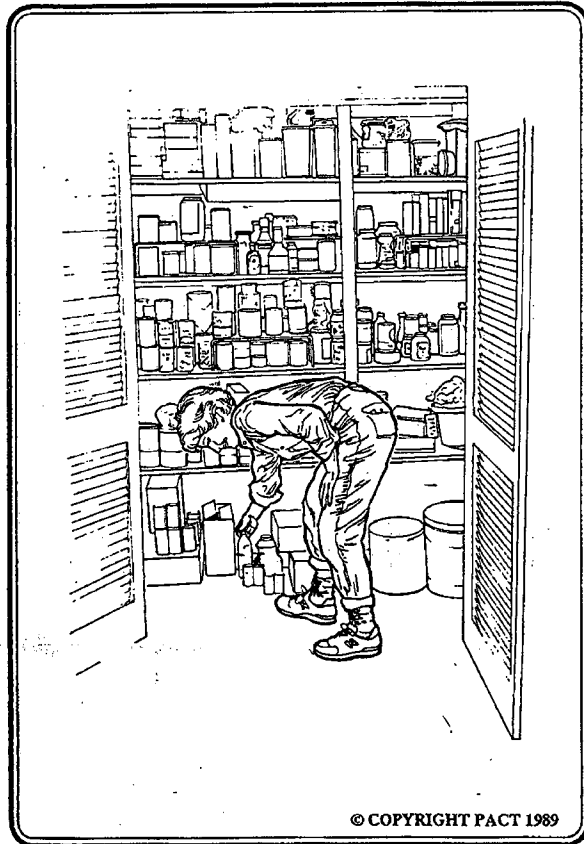
If you circle #4, this would indicate that you are very restricted, almost unable to perform the task.

Able		Restricted		Unable	
1	2	3	4	5	?

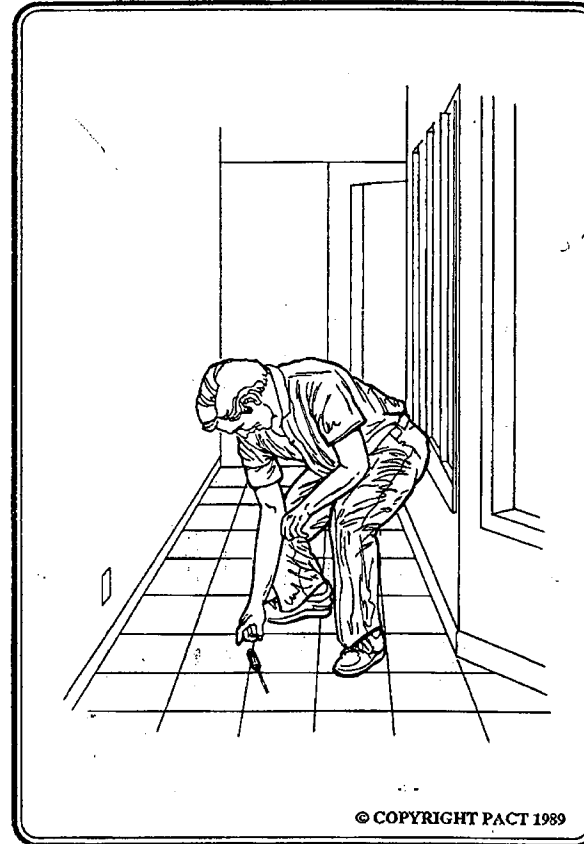
If you don't know whether or not you can perform the task, circle "?". which stands for "I don't know".

Able		Restricted		Unable	
1	2	3	4	5	?

Work quickly. Do not spend too much time on one drawing. Your first impression is usually the best.



1. Place a glass bottle on the floor.



2. Retrieve a small tool from the floor.

.....
Work quickly. Do not spend too much time on any one item. Your first impression is usually the best.

P.A.C.T. SPINAL FUNCTION SORT

© Copyright 1993 Performance Assessment and Capacity Testing

	Able	Restricted	Unable ?		Able	Restricted	Unable ?
3	1	2	3	4	5	?	?
4	1	2	3	4	5	?	?
5	1	2	3	4	5	?	?
6	1	2	3	4	5	?	?
7	1	2	3	4	5	?	?
8	1	2	3	4	5	?	?
9	1	2	3	4	5	?	?
10	1	2	3	4	5	?	?
11	1	2	3	4	5	?	?
12	1	2	3	4	5	?	?
13	1	2	3	4	5	?	?
14	1	2	3	4	5	?	?
15	1	2	3	4	5	?	?
16	1	2	3	4	5	?	?
17	1	2	3	4	5	?	?
18	1	2	3	4	5	?	?
19	1	2	3	4	5	?	?
20	1	2	3	4	5	?	?
21	1	2	3	4	5	?	?
22	1	2	3	4	5	?	?
23	1	2	3	4	5	?	?
24	1	2	3	4	5	?	?
25	1	2	3	4	5	?	?
26	1	2	3	4	5	?	?
27	1	2	3	4	5	?	?
28	1	2	3	4	5	?	?
29	1	2	3	4	5	?	?
30	1	2	3	4	5	?	?
31	1	2	3	4	5	?	?
32	1	2	3	4	5	?	?
33	1	2	3	4	5	?	?
34	1	2	3	4	5	?	?
35	1	2	3	4	5	?	?
36	1	2	3	4	5	?	?
37	1	2	3	4	5	?	?
38	1	2	3	4	5	?	?
39	1	2	3	4	5	?	?
40	1	2	3	4	5	?	?
41	1	2	3	4	5	?	?
42	1	2	3	4	5	?	?
43	1	2	3	4	5	?	?
44	1	2	3	4	5	?	?
45	1	2	3	4	5	?	?
46	1	2	3	4	5	?	?
47	1	2	3	4	5	?	?
48	1	2	3	4	5	?	?
49	1	2	3	4	5	?	?
50	1	2	3	4	5	?	?

Name: _____		Date: _____		ER: _____	
FCE at:		PDC Sedentary		PDC Light	
RPC Score:		100 - 110		125 - 135	
		PDC Medium		PDC Heavy	
		165 - 175		180 - 190	
				PDC Very Heavy	
				196 +	

Ds	DK	Int
0-2	0-3	1
0-2	4+	2
3-4	na	3
5+	na	4

Appendix 4 – Neck Disability Index

This questionnaire has been designed to give your therapist information as to how your neck pain has affected you in your everyday life activities. Please answer each section, marking only ONE box which best describes your status today.

Section 1 – Pain Intensity

- ☐ I have no pain at the moment
- ☐ The pain is very mild at the moment
- ☐ The pain is moderate at the moment
- ☐ The pain is fairly severe at the moment
- ☐ The pain is very severe at the moment
- ☐ The pain is the worst imaginable at the moment

Section 2 – Personal Care (Washing Dressing etc)

- ☐ I can look after myself normally without causing extra pain.
- ☐ I can look after myself normally but it causes me extra pain.
- ☐ It is painful to look after myself and I am slow and careful
- ☐ I need some help but manage most of my personal care
- ☐ I need help every day in most aspects of self-care
- ☐ I do not get dressed, wash with difficulty and stay in bed

Section 3 – Lifting

- ☐ I can lift heavy weights without extra pain
- ☐ I can lift heavy weights but it gives me extra pain
- ☐ Pain prevents me from lifting heavy weights off the floor, but I can manage if they are conveniently positioned e.g. on a table.
- ☐ Pain prevents me from lifting heavy weights but I can manage light to medium weights if they are conveniently positioned.
- ☐ I can lift only very light weights
- ☐ I cannot lift or carry anything at all.

Section 4 – Reading

- ☐ I can read as much as I want to with no pain in my neck
- ☐ I can read as much as I want to with slight pain in my neck
- ☐ I can read as much as I want to with moderate pain in my neck
- ☐ I can't read as much as I want because of moderate pain in my neck
- ☐ I can hardly read at all because of severe pain in my neck
- ☐ I cannot read at all

Section 5 – Headache

- ☐ I have no headache at all
- ☐ I have slight headaches which come infrequently
- ☐ I have moderate headaches which come infrequently
- ☐ I have moderate headaches which come frequently
- ☐ I have severe headaches which come frequently
- ☐ I have headaches almost all the time

Section 6 – Concentration

- ☐ I can concentrate fully when I want to with no difficulty
- ☐ I can concentrate fully when I want to with slight difficulty
- ☐ I have a fair degree of difficulty in concentrating when I want to.
- ☐ I have a lot of difficulty in concentrating when I want to.
- ☐ I have a great deal of difficulty in concentrating when I want to.
- ☐ I cannot concentrate at all.

Section 7 – Work

- ☐ I can do as much work as I want to
- ☐ I can only do my usual work but no more
- ☐ I can do most of my usual work, but no more
- ☐ I cannot do my usual work
- ☐ I can hardly do any work at all
- ☐ I can't do any work at all

Section 8 – Driving

- ☐ I can drive my car without any neck pain
- ☐ I can drive my car as long as I want with slight pain in my neck
- ☐ I can drive my car as long as I want with moderate pain in my neck
- ☐ I can't drive my car as long as I want because of moderate pain in my neck
- ☐ I can hardly drive at all because of severe pain in my neck
- ☐ I can't drive my car at all

Section 9 – Sleeping

- ☐ I have no trouble sleeping
- ☐ My sleep is slightly disturbed (less than 1 hour sleep loss)
- ☐ My sleep is mildly disturbed (1-2 hour sleep loss)
- ☐ My sleep is moderately disturbed (2-3 hours sleep loss)
- ☐ My sleep is greatly disturbed (3-5 hours sleep loss)
- ☐ My sleep is completely disturbed (5-7 hours sleep loss)

Section 10 – Recreation

- ☐ I am able to engage in all my recreational activities with no neck pain at all
- ☐ I am able to engage in all my recreational activities with some pain in my neck
- ☐ I am able to engage in most but not all of my usual recreational activities because of pain in my neck
- ☐ I am able to engage in a few of my usual recreational activities because of pain in my neck
- ☐ I can hardly do any recreational activities because of pain in my neck
- ☐ I can't do any recreational activities at all.

Comments: _____

Name _____

Date _____

Appendix 5 – Letter of Consent

Phone: (250)564-3077
Fax: (250)564-3008
Email: lois.lochhead@cidms.com

210-1811 Victoria Street
Prince George, BC
V2L 2L6



Central Interior Disability Management Services



October 24, 2008

University of Northern British Columbia
3333 University Way
Prince George, BC
V2N 4Z9

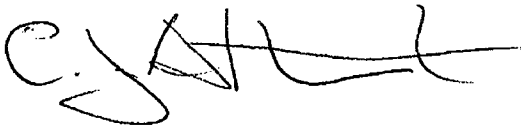
Attention : Research Ethics Committee

Dear Sirs:

As an officer of Central Interior Disability Management Services, I have authorized your student, Lois Lochhead, Student Number 200002080, to use data collected at our facility for research related to her Master's Thesis entitled *Assessment of Perceived Functional Ability: Using Rasch Analysis to Evaluate the Measurement Properties of Four Perceived Pain & Disability Scales*.

An intact data set will be provided to Ms. Lochhead once Ethics Approval has been obtained from UNBC. The data was collected by our staff and entered into an Excel spreadsheet in house. This data set contains no personal identifiers of the clients who took part in Functional Capacity Evaluations at our facility between 1998 and 2008. Each client signed a consent to evaluate and a sample of this form has been provided to Ms. Lochhead. Original material remains in the offices of Central Interior Disability Management Services.

If you have any questions or concerns, please do not hesitate to contact me.



Charles J. Attwater
Human Resources Manager

Appendix 6 – Consent to Evaluate

Phone: (250)564-3077
 Fax: (250)564-3008
 Email: lois.lockhead@ciams.com



210-1811 Victoria Street
 Prince George, BC
 V2L 2L6

Central Interior Disability Management Services

Informed Consent for Functional Capacity Evaluation

Explanation of the Functional Capacity Evaluation

A Functional Capacity Evaluation is a test of your ability to safely perform the physical demands of work. The activity intensity of the test will begin at a level that should be non-stressful and will be advanced in stages depending on your tolerance. You will never be forced to perform an activity you feel would place you at risk. Your evaluator will monitor your safety during the evaluation.

Your Responsibilities

To ensure your safety and the validity of this evaluation, it is your responsibility to fully disclose information you have pertaining to your past and current health. Further, it is your responsibility to give your best effort at all times during the evaluation without hurting yourself.

Inquiries

You are encouraged to ask questions about the procedures used in the evaluation and in the estimation of functional capacity that results from the evaluation. If you have any concerns or questions please ask us for further explanations at any time during the evaluation.

Freedom of Consent

Your participation in this evaluation is voluntary. You are free to deny consent or stop the evaluation at any time.

Release of Reports

By signing this form, you give Central Interior Disability Management Services permission to send the results of this evaluation to your insurance carrier, physician, rehabilitation counselor, employer (if work related), and legal representatives (if any).

Research

Data obtained during this assessment may be used for research purposes. Neither your name nor any personal health information will be released as part of the data set.

"I have read this form and I understand the evaluation procedures that I will perform. I consent to participate in the evaluations procedures that I will perform. I consent to participate in this evaluation."

Name: _____

Telephone Number _____

Doctor's Name _____

Date _____

Signature of Participant _____

Witness _____

Appendix 7 – DOT Physical Demand Characteristics of Work

PHYSICAL DEMAND CHARACTERISTICS OF WORK

1993 Leonard N. Matheson, PhD

PHYSICAL DEMAND LEVEL	OCCASIONAL 0 - 33% of the workday	FREQUENT 34 - 66% of the workday	CONSTANT 67 - 100% of the workday	Typical Energy Required
SEDENTARY	10 lbs.	Negligible	Negligible	1.5 - 2.1 METS
LIGHT	20 lbs.	10 lbs. and/or Walk/Stand/Push/Pull of Arm/Leg controls	Negligible and/or Push/Pull of Arm/Leg controls while seated	2.2 - 3.5 METS
MEDIUM	20 to 50 lbs.	10 to 25 lbs.	10 lbs.	3.6 - 6.3 METS
HEAVY	50 to 100 lbs.	25 to 50 lbs.	10 to 20 lbs.	6.4 - 7.5 METS
VERY HEAVY	Over 100 lbs.	Over 50 lbs.	Over 20 lbs.	Over 7.5 METS

Appendix 8 – Oswestry Disability Index – Rescaled

This questionnaire has been designed to give the doctor information as to how your back pain has affected your ability to manage in every day life. Please answer every section, and mark in each section only the *one* box which applies to you. We realize you may consider that two of the statements in any one section relate to you, but please just *mark the box which most closely describes your problem*.

Section 1 – Pain Intensity

- ☐ I can tolerate the pain I have without having to use painkillers
- ☐ The pain is bad but I manage without taking painkillers
- ☐ Painkillers give moderate to complete relief from pain.
- ☐ Painkillers give very little relief from pain.
- ☐ Painkillers have no effect on the pain and I do not use them.

Section 2 – Personal Care (Washing Dressing etc)

- ☐ I can look after myself normally without causing extra pain.
- ☐ I can look after myself but it causes extra pain.
- ☐ It is painful to look after myself and I am slow and careful
- ☐ I need some help but manage most of my personal care
- ☐ I need help every day in most aspects of self-care

Section 3 – Lifting

- ☐ I can lift heavy weights without extra pain
- ☐ I can lift heavy weights but it gives me extra pain
- ☐ Pain prevents me from lifting heavy weights off the floor, but I can manage if they are conveniently positioned e.g. on a table.
- ☐ I can only lift light weights
- ☐ I cannot lift or carry anything at all.

Section 4 – Walking

- ☐ Pain does not prevent me from walking any distance
- ☐ Pain prevents me walking more than 1 mile.
- ☐ Pain prevents me walking more than short distances
- ☐ I can only walk using a stick or crutches
- ☐ I am in bed most of the time and have to crawl to the toilet.

Section 5 – Sitting

- ☐ I can sit as long as I like
- ☐ Pain prevents me from sitting more than 1 hour
- ☐ Pain prevents me from sitting more than ½ hour.
- ☐ Pain prevents me from sitting more than 10 minutes
- ☐ Pain prevents me from sitting at all.

Section 6 – Standing

- ☐ I can stand as long as I want without extra pain
- ☐ I can stand as long as I want but it gives me extra pain
- ☐ Pain prevents me from standing for more than 1 hour
- ☐ Pain prevents me from standing for more than 30 minutes
- ☐ Pain prevents me from standing for more than 10 minutes

Section 7 – Sleeping

- ☐ Pain does not prevent me from sleeping
- ☐ Even when I take tablets I have less than 6 hours of sleep
- ☐ Even when I take tablets I have less than 4 hours of sleep
- ☐ Even when I take tablets I have less than 2 hours of sleep
- ☐ Pain prevents me from sleeping at all.

Section 8 – Sex Life

- ☐ My sex life is normal and causes no extra pain
- ☐ My sex life is nearly normal but causes extra pain.
- ☐ My sex life is significantly restricted by pain.
- ☐ My sex life is nearly absent because of pain
- ☐ Pain prevents any sex life at all.

Section 9 – Social Life

- ☐ My social life is normal and gives me no extra pain
- ☐ Pain has no significant effect on my social life apart from limiting my more energetic interests such as dancing etc.
- ☐ Pain has restricted my social life and I do not go out as often
- ☐ Pain has restricted my social life to my home.
- ☐ I have no social life because of pain.

Section 10 – Travelling

- ☐ I can travel anywhere without extra pain
- ☐ I can travel anywhere but it gives me extra pain
- ☐ Pain is bad but I manage journeys over 2 hours
- ☐ Pain restricts me to journeys of less than 1 hour
- ☐ Pain restricts me to short necessary journeys of less than 30 minutes

Name _____ Date _____