PSYCHOMETRIC METHODS TO DEVELOP AND TO ANALYZE CLINICAL MEASURES

PSYCHOMETRIC METHODS TO DEVELOP AND TO ANALYZE CLINICAL MEASURES: A COMPARISON AND CONTRAST OF RASCH ANALYSIS AND CLASSICAL TEST THEORY ANALYSIS OF THE PEDSQL^{TM} 4.0 GENERIC CORE SCALES (PARENT-REPORT) IN A CHILDHOOD CANCER SAMPLE

By

LEILA AMIN, BSc., MScOT, OT Reg. (Ont.)

A Thesis
Submitted to the School of Graduate Studies
in Partial Fulfilment of the Requirements
for the Degree
Master of Science

McMaster University

MASTER OF SCIENCE (2011) McMaster University

(Rehabilitation Science) Hamilton, Ontario

TITLE: Psychometric Methods to Develop and to Analyze Clinical Measures: A Comparison and Contrast of Rasch Analysis and Classical Test Theory Analysis of the PedsQL^{TM} 4.0 Generic Core Scales (Parent-report) in a Childhood Cancer Sample

AUTHOR: Leila Amin, BSc., MScOT, OT Reg. (Ont.)

SUPERVISOR: Anne Klassen, DPhil

COMMITTEE: Peter Rosenbaum, MD, FRCP

Ronald Barr, MD

Carol DeMatteo, MSc, OT Reg. (Ont.)

NUMBER OF PAGES: 116

I wish to dedicate this thesis to my sister Lida Amin (1987-2003) who will always be a cancer survivor. Your journey through cancer has changed me forever and inspired me to dedicate my life to making a difference in the lives of other cancer survivors and their families world wide.

ABSTRACT

Traditionally, measures have been developed using Classical Test Theory (CTT). Modern psychometric methods (e.g. Rasch analysis) are being applied to increase understanding of item-level statistics and to aid in interpreting rating scale scores. This thesis aims to compare and contrast psychometric findings for the PedsQL^{TM} 4.0 Generic Core Scales using CTT and Rasch analysis to determine if a Rasch approach provides information that furthers our understanding of scale scores. The assumptions, advantages and limitations of each psychometric paradigm are presented.

Issues that arise when measuring quality of life are discussed to set the stage for a psychometric analysis of the PedsQL^{TM} in a childhood cancer sample. The PedsQL^{TM} measures child health in terms of physical, social, emotional and school function. The parent-report version was used in a Canadian study of 385 parents of children aged 2 to 17 years on active cancer treatment and data was re-analyzed for this thesis. CTT analysis was performed using PASW Statistics and Rasch analysis was performed using Rumm2030.

Internal consistency reliability was higher using CTT ($\alpha = 0.93$) than Rasch analysis (Person Separation Index = 0.78). Rasch analysis item curves showed respondents did not discriminate between response categories and a 3 point scale (vs. 5) was preferred. Item curves also indicated most items were free of bias. There are no equivalent visual representations in CTT of how respondents use response categories or of whether items display bias. Both approaches indicate a large ceiling effect associated with the overall score.

Results challenge internal consistency reliability of the PedsQL^{TM} 4.0. Rasch analysis permits detailed and visually pleasing examination of item-level statistics more effectively than CTT. Research is needed to determine which testing circumstances render Rasch analysis useful and justify time and resources to use both paradigms as complementary tools to maximize understanding of rating scale scores.

ACKNOWLEDGEMENTS

There are several important people in my life that I wish to acknowledge in the preparation of my thesis. First I would like to thank my family for supporting me throughout my studies and for providing me with everything that I needed so that I could put all my energy and focus into the preparation of my thesis. I would like to acknowledge my sister, Paresa Amin, for encouraging me throughout the writing phases of my thesis, for offering her assistance whenever she could help and for always reminding me how proud she was of me for the work I was doing. I would like to acknowledge my boyfriend, Shehzad Moiz, for helping me overcome the challenges I faced as I progressed with writing my thesis. I would also like to acknowledge him for helping me with the parts of my thesis that required more computer skills than I can admit to having. I would like to acknowlege Professor Alan Tennant for his guidance and support during the Rasch analysis course I attendend at the University of Leeds. I would also like to acknowledge Professor Tennant for his continued support with my analysis and for always answering all my questions when I returned home from taking his course. I would like to thank my supervisor, Dr. Anne Klassen, as well as Drs. Lillian Sung and Rob Klaassen for letting me use their data set to re-analyze for my thesis study. Lastly I would like to recognize and acknowledge DF, for making me feel confident that the work I was doing as part of my thesis was making an important contribution to child health and for always nurturing and believing in my potential as an academic and as a clinician.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 3

CHAPTER 4

CHAPTER 5

LIST OF FIGURES

CHAPTER 1

CHAPTER 4

CHAPTER 5

APPENDICES

INTRODUCTION

Evaluation of the reliability, validity and responsiveness of rating scales is essential. Clinicians use rating scale scores to measure health outcomes and to make decisions that directly influence the care of their patients (Hobart, Cano, Zajicek, & Thompson, 2007). Rating scale scores are used as outcome measures in clinical trials that aim to evaluate the efficacy of various treatment approaches. Indeed, the decision to accept or reject a treatment approach is often based on a predetermined change of a score on a rating scale. These scores also have the potential to influence change in health policy and to impact future directions in research.

As a consequence of the underlying theory and method used to develop and to evaluate rating scales, the scores generated by many scales may not satisfy the criteria required for rigorous measurement (Hobart & Cano, 2009). Rating scales used in clinical practice often lack validity making it difficult to interpret accurately the meaning of scores produced, and to decide the extent to which the score should legitimately be used to make a clinical decision.

The importance of rating scale scores highlights the need to evaluate potential limitations in how scores are produced and used, and to suggest methods to overcome any limitations. Exploring the psychometric theories that underlie the development of rating scales provides a method to investigate limitations that may exist in their interpretation and use.

Psychometrics

The study of methods for developing and evaluating rating scales and for analyzing their data is referred to as psychometrics (Hobart & Cano, 2009). The goal of psychometric analysis with respect to rating scales is to establish the extent to which the conceptualization of a variable that cannot be measured directly, such as quality of life, is represented by items on a scale. Different psychometric methods use different kinds of evidence to determine if this goal has been achieved. Traditional psychometric methods are based on Classical Test Theory (CTT) whereas modern psychometric methods are primarily based on Item Response Theory (IRT). The assumptions of each theoretical framework will be presented in this thesis, along with the implications for rating scale development, evaluation and overall score interpretation.

Outline of this Thesis

The purpose of this thesis is to compare and contrast traditional psychometric methods based on CTT to newer psychometric methods based on IRT, in particular the Rasch model. Some measurement theorists believe that the Rasch model should not be classified as an IRT model because the two have distinct properties (Hobart & Cano, 2009); however, for the purpose of this thesis the Rasch model is described as a one-parameter IRT model as suggested by

Streiner & Norman (2008). The assumptions of both CTT and the Rasch model will be explored and the consequences of these assumptions on the development of new measurement tools and the evaluation of existing measurement tools will be discussed (Chapter 1).

In Chapter 2, issues in measurement of adult and pediatric quality of life (QoL) are presented. This thesis is predominantly concerned with 'health-related QoL'; however it is recognized that a variety of terms are often used interchangeably to identify QoL (e.g., health status, functional status) and henceforth the broader term QoL will be used to refer to all such measures given the conceptual overlap among them (Klassen, Strohm, Maurice-Stam, & Grootenhuis, 2009). Issues are highlighted that could be further explored using Rasch analysis as a vehicle to increase understanding of items, scales and overall rating scale scores on QoL measures.

The development history and psychometrics of the PedsQL^{TM} 4.0 Generic Core Scales are presented to explore specifically how a commonly used pediatric QoL measure has been evaluated using a CTT paradigm, and to set the stage for a comparison of methods used in traditional versus modern psychometric analyses (Chapter 3).

The specific methodology used to perform a Rasch analysis is presented in Chapter 4. CTT and Rasch analysis are compared and contrasted particularly with respect to the statistics and procedures used to complete item and scale analyses.

A study is presented in Chapter 5 that uses Rasch analysis to psychometrically evaluate the PedsQL^{TM} 4.0 Generic Core Scales (Parent-report) in a sample of parents of children undergoing cancer treatment. Similarities and differences in the development and evaluation of rating scales using Rasch analysis and CTT are discussed.

Chapter 6 discusses results, strengths and limitations of the study mentioned above. Finally, Chapter 7 presents recommended approaches to address issues in the measurement of pediatric QoL by acknowledging both traditional and modern paradigms of measurement in the development and evaluation of rating scales.

CHAPTER 1

BACKGROUND INFORMATION

Classical Test Theory (CTT)

A test theory is defined as a mathematical representation of the factors influencing scores generated by a rating scale and is characterized by its underlying assumptions (Fan 1998; Hobart & Cano, 2009). CTT describes how errors of measurement can influence the scores obtained with rating scales. The theory is founded on the postulate that a respondent's observed score on the rating scale is a combination of a true score (a theoretical value that is the expected average score an individual would receive if they were repeatedly administered a scale an infinite number of times) and a random error component (Reise & Henson, 2003; Hobart & Cano, 2009).

Random error is inherent in any measurement and varies each time the measure is administered. An important focus of CTT is to identify and to provide strategies that reduce the inherent error on the measurement in question. The relative importance of each error component directs the specific strategy used to improve the validity and reliability of the overall measure (Streiner & Norman, 2008). CTT predominantly focuses on person-level statistics such as means and standard deviations and on test-level statistics such as reliability; however, item-level statistics such as item difficulty and item discrimination also play an important role in the model.

CTT is a useful model that has served as the main theory directing the development and evaluation of rating scales for decades (Hobart & Cano, 2009). For the purpose of this thesis the assumptions of CTT have been simplified to a level that provides a basic understanding of the theory in order to compare and contrast traditional and modern psychometric approaches. CTT assumptions include the following: 1) all items on the scale have equal variances; 2) measurement errors associated with one scale are not correlated with the true scores or measurement errors of another scale; 3) each item, regardless of item difficulty, contributes equally to the final score; and 4) ordinal-level measurement can approximate interval-level measurement (Fan, 1998; Hobart & Cano, 2009; Neumann, Goldie, & Weinstein, 2000; Reise & Henson, 2003). Preference-based instruments are an exception to the fourth point above as these measures do indeed produce interval-level measurement.

Preference-based instruments examine the extent to which respondents value a particular health state. Utilities are numeric measurement produced by such instruments and reflect "an individual's beliefs about the desirableness of a health condition, the willingness to take risks to gain health benefits, and the preferences for time" (Lenert & Kaplan, 2000, p.138). Utilities are used to guide the allocation of resources or to assess cost-effectiveness of various treatment approaches. Knowledge of the value people assign to the health improvement they receive from various interventions can help determine how most effectively to provide people with the outcomes they desire. Preference-based measures developed in a CTT paradigm are the only scales developed using traditional methods that use single and multi-attribute utility functions to generate interval-level measurement from ordinal scores. Instruments that focus on consistently discriminating between respondents (as is the focus of health status rating scales) produce ordinal scores that proponents of CTT believe very closely approximate interval-level measurement (Fan, 1998).

The assumptions listed above direct how rating scale scores are developed and interpreted within a CTT framework and they guide how scores can be used legitimately in clinical practice. An explanation of common criteria evaluated in traditional psychometrics and the implications of the assumptions stated above will be discussed in the following two sections.

Psychometric Criteria Evaluated in a Traditional Paradigm

There are various psychometric criteria reported in the literature to guide the development of sound measurement tools. The Consensus-based Standards for the selection of health Measurement Instruments (COSMIN) study reached international consensus on the properties that should be reported for health-related patient-reported outcomes (Mokkink et al., 2010) and therefore these criteria will be used to guide the critique of the PedsQL^{TM}. Using a CTT framework, a psychometric evaluation can be thought of in terms of scale and item-level analyses.

Item-Level Analyses

Item-level analyses consider the feasibility, scaling success and difficulty for each item. Feasibility is determined by inspecting the percentage of missing values for each item in the scale. Scaling success is determined by examining the number of times an item correlated more strongly with its hypothesized scale construct (represented by the total score of that scale) than with the construct measured by another scale. The item-total correlation value is a reflection of how well items measure what they are intended to be measuring; correlations should be between 0.2 and 0.7 (Streiner & Norman, 2008). Correlations that exceed 0.7 suggest item redundancy, while correlations less than 0.2 suggest the item is measuring an entirely different construct. Item difficulty is determined by inspection of the mean and endorsement frequency for each item (Streiner & Norman, 2008).

Scale-Level Analyses

Scale-level analyses consider the reliability, validity and targeting capacity of the scale.

Reliability

The internal consistency is the most common index of reliability reported in CTT and is commonly referred to as Cronbach's alpha (α). Scales with α values of 0.7 or more are appropriate to use for group-level comparisons (Nunnally & Bernstein, 1994; Pedhazur & Schmelkin, 1991). An α value of 0.9 is recommended before a scale is used for individual-level comparisons (Nunnally & Bernstein, 1994; Pedhazur & Schmelkin, 1991).

Scales that are adequate for group-level analysis have wider confidence intervals around the overall summary score and can only inform clinicians and researchers on the extent to which one group of respondents is statistically different from another (Hobart et al., 2007). Most health measures are norm-referenced and individual-level reliability is not required because the purpose of the measure is to assess the respondent's relative performance on a characteristic being assessed (Streiner & Norman, 2008). In criterion-referenced measures, the respondent's absolute score is the basis of decision-making and therefore reliability at the individual level is required. Furthermore in clinical trials, treatment effects typically vary on an individual basis. Thus, in order to understand the complexities of why individuals undergo different levels and directions of change within a group, individual-level reliability is advantageous (Hobart et al., 2007).

Developing scales with sufficient reliability for individual-level analysis is important in furthering efforts to understand the individual variables that contribute to directions and levels of change within a group (Hobart et al., 2007). Individual-level reliability is not necessarily required when respondents' relative scores are more important than their absolute scores on a measure (Streiner & Norman, 2008). Depending on how the measure is going to be used, other types of reliability that could be assessed include inter-observer and test-retest reliability (Streiner & Norman, 2008).

Inter-observer reliability can be thought of as the agreement in overall ratings on a measure between parent and child respondents. Test-retest reliability is an indication of how reliable a scale is when administered on two separate occasions separated by a time interval sufficiently short that the underlying trait would not have changed (Streiner & Norman, 2008). It is critical to select the appropriate time interval between the first and second administration of the measure; if the time interval is too long the underlying construct may have changed, and if it is too short respondents may remember their responses from the first time they completed the measure (Streiner & Norman, 2008).

Validity

There are several types of validity; the types that are appropriate to consider depend on the purpose of the scale. A scale can be used to describe, to discriminate, to predict or to evaluate change within a sample (Streiner &

Norman, 2008). A scale can serve to achieve a combination of these purposes but it is essential that the appropriate types of validity be assessed for each use of the tool. This assessment is important because a tool that is meant to discriminate may not necessarily be able to evaluate change, to be descriptive or to be predictive (Rosenbaum, 1998).

Face validity is usually a starting point to assessing validity and it is based on a judgment call of whether the content of items seems to measure at face value the construct in question (Streiner & Norman, 2008). Construct validity is an important component of the validation process. It is a reflection of how well the scale measures the intended construct and is often determined by examining a number of hypotheses about how the measure should behave if it actually measures the intended construct. Factor analysis is a technique that can be used to assess construct validity more formally and this technique will be discussed in the methods section of this thesis (Streiner, 1994). Qualitative techniques such as conducting interviews to capture people's perspectives of the construct or conducting an analysis of the content of several measures that assess the same construct are other ways of establishing construct validity (Rajmil et al., 2004).

Known groups validity is a form of construct validation in which the ability of the instrument to produce different scores for groups known to differ on the construct being measured is examined. The known groups method of validation is a reflection of how well the instrument can distinguish between two groups that are known to differ on the attribute being measured, and is important to establish for a scale that is to be used to compare the outcome of a treatment on different groups.

Concurrent validity is another commonly studied form of construct validity and is determined by comparing the results of a measure to a known indicator of the construct to determine how well the results of the scale reflect the construct (Streiner & Norman, 2008).

Targeting

The targeting capacity of the scale is determined by identifying the extent to which the range of the variable measured by the scale matches the range found in the sample. Targeting can be examined at both the item and scale level. At the item level, scores should be distributed evenly across all response categories to achieve good targeting. At the scale level, respondents' scores should span the entire range of the scale such that the mean of the sample lies near the mid-point of the scale and the proportion of respondents that score at the extreme ends of the scale (referred to as the floor and ceiling of the scale) is low. If more than 15% of the sample is scoring the maximum or minimum possible score on a scale this suggests a ceiling effect or floor effect, respectively (for a scale in which a higher score represents less of the characteristic being assessed, as is the case with the PedsQL^{TM}) (Holmes, Bix, & Shea, 1996; Holmes & Shea, 1997).

Applications and Implications of Classical Test Theory

A major implication of CTT arises due to the fact that the individual's ability to endorse an item is not described in relation to their level of the trait in question (Fan, 1998). Instead, CTT considers a pool of respondents and examines the success rate of that particular sample of people on an item. Therefore the scores generated by the rating scale will depend on how much of the trait is possessed by the people in the sample being studied, while how much of the trait the sample has is determined by the norms of the scale (Streiner & Norman, 2008; Fan, 1998). For example, if an individual with average capabilities is measured within a sample of people with severe disability they will likely score in a higher percentile of the sample; if they are measured within a group of people who have no disability they will likely score in a lower percentile of the sample. The term 'circular dependency' is used to describe this limitation (Fan, 1998).

The notion of circular dependency is counterintuitive, as one would expect a measurement tool to be stable and thus independent of both scale and sample. In a CTT framework a respondent's test scores will depend on the particular items that have been administered from the overall test (i.e., measurement is dependent on scale); whereas within a modern psychometric framework, ability scores are independent of the particular choice of test items administered (i.e., measurement is independent of scale). As such, using CTT-developed measures an examinee will have a higher score if the items administered are easier, and a lower score if the items administered are more difficult.

Furthermore, in CTT properties of a measurement tool such as item discrimination, item difficulty, and reliability are dependent on the sample from which they are generated. The item difficulty and discrimination will appear to be higher if assessed in a lower ability and more heterogeneous sample, respectively; whereas these values will be lower if assessed in a higher ability and more homogenous sample. In modern psychometrics, item- and test-level statistics are independent from the sample in which they are assessed.

Proponents of modern psychometric methods assert that scores generated by rating scales developed within a CTT framework should only be used for group-level decision-making, as the variance around the confidence interval for individual measurement can be too wide to produce accurate measurement (Reise & Henson, 2003). Furthermore they assert that raw scores produced from CTT-developed measures (ordinal scores) should not be analyzed using parametric methods (e.g. t-test, ANOVA) because parametric methods should only be used to analyze interval-level data (Hobart & Cano, 2009). Due to their widespread use and acceptability in healthcare measurement, clinicians continue to use scores from CTT-developed measures in patient-level decision-making despite their limitations (Hobart, Cano, Zajicek, & Thompson, 2007).

In CTT, because item and scale statistics only apply to the specific groups of subjects with whom the test was developed, it is necessary to re-establish psychometric properties and to develop new norms when the test is administered to groups that are different than the original sample. It is also necessary to re-norm the scale if any of the items on the scale are altered or deleted to produce a shorter version of the scale or to account for missing responses to an item (Streiner & Norman, 2008). Therefore shorter versions of the rating scale cannot be administered to individuals based on their skill level, and scoring problems arise when data are missing because patients have not answered all of the items on the rating scale. Modern psychometric methods were developed to overcome some of the limitations inherent in traditional psychometric methods (Hobart & Cano, 2009; Schumacker & Smith, 2007).

Item Response Theory (IRT)

IRT is based on the development of mathematical models (item response functions) that generate interval-level measurement from ordinal-level measurement. These models describe the relationship between a person's level of ability and his or her response to the items of a rating scale. There are three types of models, named based on the number of item parameters that are taken into consideration. Item parameters include item discrimination, item difficulty and item guessing. The one-parameter model (also known as the Rasch model) only considers the item difficulty parameter, while the two-parameter model considers item difficulty and discrimination. The three-parameter model was developed specifically for educational testing and includes an additional item guessing parameter to account for students who guess the response to an item (Hobart & Cano, 2009).

There is some debate in the literature about whether the Rasch model is its own separate measurement paradigm or if it can be classified as a one-parameter IRT model (Hobart & Cano, 2009). The primary difference between IRT and Rasch analysis lies in the methodology of carrying out the analysis: in IRT various parameters are applied to the data until a model is found that best fits the data; in Rasch analysis only one model, the Rasch model, is applied and the data are made to fit the model. Therefore, in IRT the data are prioritized and in Rasch analysis the model is prioritized. People who use IRT tend to use the Rasch model because it is the simplest model that can be used to explain data; however, people who use the Rasch model do not tend to use the other two IRT models (Hobart & Cano, 2009).

For the purpose of this thesis I have chosen to present Rasch analysis as a one-parameter IRT model, because the goal is to explore characteristics of the item response function and to understand some of similarities and differences between traditional and modern psychometrics. The study presented in this thesis is analyzed using the one-parameter Rasch model and thus the methodology presented in Chapter 4 reflects procedures that prioritize the model over the data.

The primary assumptions of all IRT models (including the Rasch model) are as follows: 1) unidimensionality and 2) local independence.

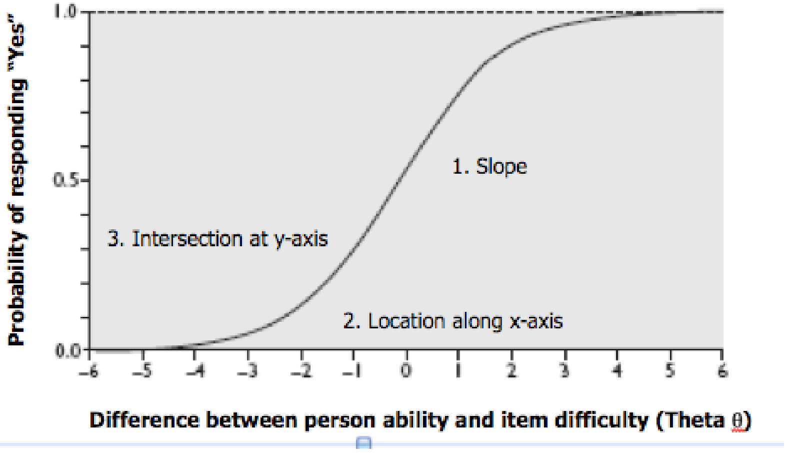
Unidimensionality is a property of a scale in which each item measures only one specific aspect of the overall construct being measured; therefore scores of each item can be summed to produce an overall score. Unidimensionality is a fundamental requirement of the scale in order to achieve construct validity (Streiner & Norman, 2008). Local independence is a property of the items that suggests a person's response to one item does not depend on their responses to any other test items (Reise & Henson, 2003). By testing the assumption of local independence the dimensionality of the scale is also assessed (Schumacker & Smith, 2007; Tennant & Conaghan, 2007; Tennant & Pallant).

Within a Rasch framework, a person's performance on a rating scale is predicted by the degree of the trait being measured, symbolized by the Greek letter theta (θ), and the probability of endorsing an item on the rating scale based on their level of the trait (Streiner & Norman, 2008). The focus in newer psychometric methods has shifted from the relationship between a person's measurement and their observed total score, as done in traditional methods, to the relationship between a person's measurement and the probability of responding to an item (Fan, 1998).

In Rasch, the relationship between a person's performance on any item in the rating scale and the level of the trait they possess is described by an item response function as displayed in Figure 1a. The item response function is one example that reflects the increased attention to item-level statistics versus person-level statistics in modern psychometrics methods (Streiner & Norman, 2008).

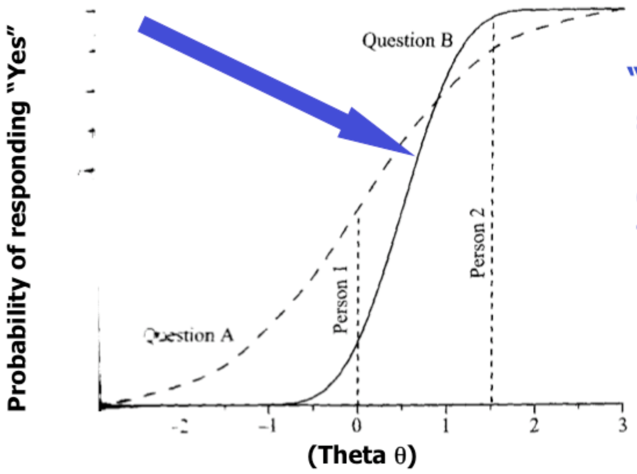
The item response function is an s-shaped curve in which the probability of answering in a positive direction consistently and gradually increases as the amount of the trait increases. Differences in the item response function can occur in three places: 1) the steepness of the slope (Figure 1b); 2) the location of the curve along the x-axis (Figure 1c); and 3) the intersection of the asymptote on the y-axis (Figure 1d) (Streiner & Norman, 2008).

Figure 1a: Item response function



(Streiner & Norman, 2008)

Figure 1b: Item response curves with different slopes



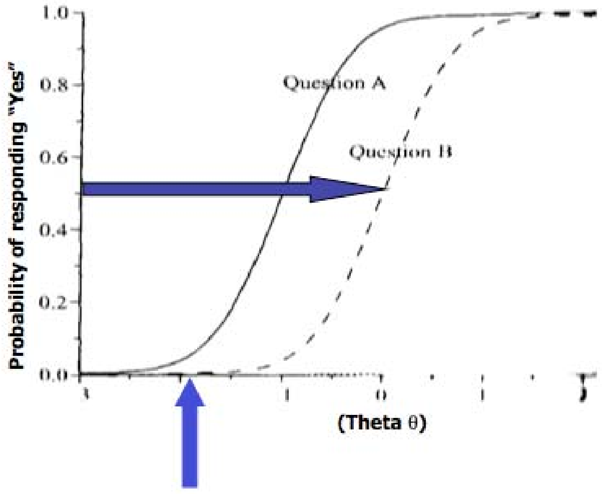
"B" has a steeper slope therefore it is a better discriminator of the trait

(Streiner & Norman, 2008)

The steeper slope of Question B in Figure 1b represents an item that is a better discriminator of the trait being measured: the proportion of people that endorse this item will change rapidly as the amount of the trait being measured increases. Conversely, the flatter slope of Question A reflects an item that is a poorer discriminator of the trait being measured: the proportion of people who endorse this item does not change rapidly as the amount of the trait being measured increases (Streiner & Norman, 2008).

Figure 1c: Item response curves with different item difficulties

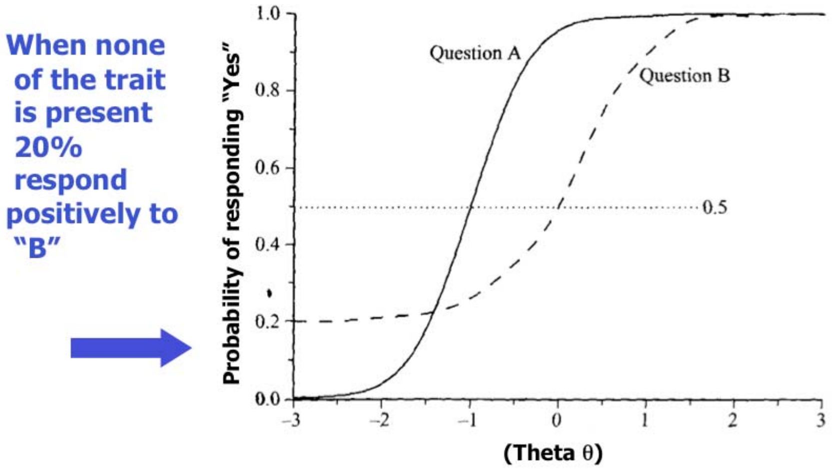
Question B is "harder" than Question A (you need more of the trait to respond positively)



(Streiner & Norman, 2008)

The location of the item response function along the x-axis represents the item's level of difficulty. The further to the right the function is along the x-axis, the further will be the point that represents the 50 percent probability of endorsing the item, and the item is considered to be more difficult (Streiner & Norman, 2008). In Figure 1c, Question B is further along the x-axis than Question A; therefore, Question B is a more difficult item.

Figure 1d: Item response curves with different y-axis intersection


When none of the trait is present 20% respond positively to "B"

(Streiner & Norman, 2008)

The intersection of the function on the y-axis reflects the proportion of people that respond positively to the trait when none of the trait is present. If the curve intersects the y-axis at the point (0,0), no one will respond positively to the trait when none of the trait is present. If the curve intersects at the point (0, 0.2), as is the case for Question B in Figure 1d, 20 percent of people will respond positively to the trait when none of the trait is present (Streiner & Norman, 2008).

In Rasch analysis the estimate of item difficulty is independent of the persons taking the test, and the estimate of person ability is independent of the items they have taken. The ability to separate item and person estimates is a property of the Rasch model that is known as invariance. The degree of precision of Rasch person ability estimates will only vary depending upon the distribution of persons being assessed (Schumacker & Smith, 2007). For example, if a measure of physical function is evaluated in a sample of people who are admitted to a rehabilitation program, then items representing high ability levels (i.e. climbing several flights of stairs) will tend to have large standard errors associated with their estimates, because not many patients would be able to do these. Therefore for the initial development and calibration of a measure, it is best to evaluate the measure using a uniformly distributed sample to ensure items have an equal degree of precision across the contruct being measured.

The required sample size when evaluating or developing a measure depends upon the degree of precision required and how well the sample is targeted to the scale (Pallant & Tennant, 2007; Tennant & Conaghan, 2007; Tennant & Pallant). To achieve 99% confidence in person ability estimates at least 108 cases are required; 243 cases are required to attain the same degree of precision for a poorly targeted measure. For health rating scales the minimum sample size required when evaluating a measure is 250 cases or 20 times the number of items in the measure (whichever is greater). When developing a new measure, 50 cases are usually sufficient to become aware of any serious issues with the measure (Linacre, 1994). In the traditional paradigm smaller sample sizes are generally required for evaluating or developing new measures (Hambleton & Jones, 1993).

Consequences of the assumptions that underpin Rasch will be discussed in the next section to illustrate their influence on scale development, evaluation and interpretation. The psychometric criteria evaluated in a Rasch analysis are presented as part of the methodology in Chapter 4.

Applications and Implications of Rasch Analysis

It is suggested that the shift from studying person-level statistics in traditional psychometric methods to studying item-level statistics in modern methods has potential to increase the validity of rating scales (Hobart, Cano, Zajicek, & Thompson, 2007). To further one's understanding of precisely what a rating scale measures it is helpful to establish a testable theory that describes the trait being measured (Hobart & Cano, 2009). Rasch analysis can be tested empirically, as the characteristics of an item that determine its location on the trait continuum are defined and measurable.

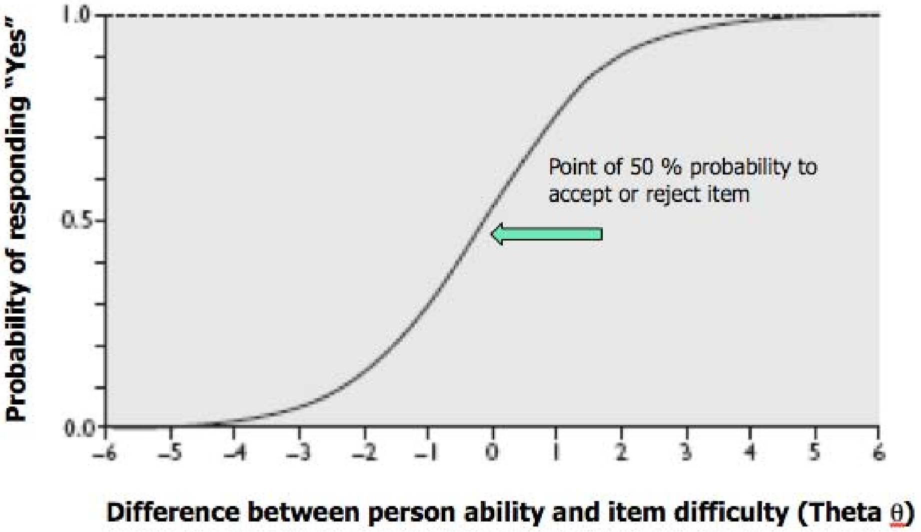
The item response function describes exactly how the score obtained on a rating scale is generated and provides a means to evaluate empirically the extent to which the score actually reflects an individual's level of the trait being measured. It is therefore possible to determine mathematically the extent to which the scores generated by the rating scale measure the level of the trait in question and thus enforce the validity of the scale (Hobart & Cano, 2009). This type of empirical testing is not possible within a CTT framework because the characteristics of an item are not specifically defined using a mathematical function.

The value of validity as well as other core psychometric criteria has been highlighted as an important focus in a recent document published by the US Food and Drug Administration (FDA) (Revicki, 2007). The release of this document represents a milestone in acknowledging the importance of rating scale scores as outcome measures in clinical trials and in clinical decision-making. The document emphasizes the need to ensure a high level of data quality by assessing the completeness of data and the distribution of raw scores (Revicki, 2007). The report also emphasizes the importance of ensuring that items in a scale measure a common underlying construct, that each item contains a similar proportion of information concerning the construct being measured, and that the items are correctly grouped into scales. Lastly, the report points to the need to ensure that the scale is acceptable as a measure for the sample targeted (Revicki, 2007).

Although the FDA document increased attention to the topic of rating scale development and evaluation, it did not offer any suggestions as to how to achieve the above-mentioned psychometric standards. Utilizing modern psychometric methods, such as Rasch analysis, in addition to traditional methods in both the development and evaluation of rating scales may serve as a means of achieving the standards set forth in the FDA document.

The use of Rasch analysis makes possible the generation of true interval-level measurement from ordinal-level scores, addressing one of the widely contested issues with regard to the use of rating scales developed using CTT (Hobart & Cano, 2009; Schumacker & Smith, 2007). Measurement experts who use CTT argue that the ordinal scores produced by rating scales in the traditional paradigm so closely approximate interval-level measurement that the two are essentially the same (Fan, 1998). As mentioned earlier, preference-based measures developed within CTT are the only instruments that actually generate interval level data from ordinal level data through the use of single and multi-attribute utility functions (Neumann et al., 2000). Advocates for modern psychometrics challenge this view and assert that true interval-level measurement can only be produced through the use of Rasch procedures that mathematically convert ordinal raw scores into log odds units. This conversion is based on the point on the item response function at which the probability of endorsing the item is equal to the probability of rejecting that item as displayed in Figure 2 (Hobart, Cano, Zajicek, & Thompson, 2007; Streiner & Norman, 2008).

Figure 2: Point of equal probability to endorse or reject item



(Streiner & Norman, 2008)

In Rasch analysis, because the probability of endorsing an item is unrelated to the probability of answering any other item positively for people with the same amount of the trait, the evaluation of scales and the measurement of people using the scale are independent of both the sample and the scale (Hobart & Cano, 2009). This property of Rasch makes it a useful theory to be applied in conjunction with CTT to enforce the valid interpretation of rating scale scores (Schumacker & Smith, 2007).

Another advantage of Rasch methods includes scale-linking, which is a novel concept that emerged with the development of modern psychometrics. Scale-linking refers to a set of procedures used to ensure respondents' scores across different measures of the same construct can be transformed to the same scale to facilitate the comparison of respondents (Reise & Henson, 2003). These procedures attempt to resolve two important issues that arise when interpreting rating scale scores in the traditional paradigm: non-response, and the ability to compare individuals who took different measures of the same construct (Hobart, Cano, Zajicek, & Thompson, 2007).

Non-response can occur for a variety of reasons, such as lack of understanding or refusal to answer a question based on content (Hobart & Cano, 2009). In CTT the issue of non-response poses a significant problem because in order to compare respondents each must answer an equal number of items (Hambleton & Jones, 2007). Strategies to correct for non-response in CTT include deleting the question or replacing the missing value with the imputed sample mean (Norman & Streiner, 2008). In a Rasch framework missing responses do not cause as much concern because a respondent's level of the trait can be accurately estimated if at least fifty percent of the items have been answered (Reise & Henson, 2003).

A more difficult issue arises when respondents have completed different measures of the same construct and they need to be compared on a common scale. This situation can occur when a measure changes its content over time, when a measure is shortened, when different versions of a measure are used with different age groups, or when a measure is administered in different languages (Reise & Henson, 2003). Scale-linking procedures in Rasch allow investigators to compare respondents who have used different versions of the same scale. This comparison can be made because when data fit the Rasch model it follows that the person estimates (based on the item characteristic curves) are invariant regardless of what items they are based on. Therefore valid comparisons can be made between individuals who have completed different items as long as all the items measure the same construct (Hobart & Cano, 2009; Schumacker & Smith, 2007).

In the following chapters of this thesis, traditional and modern psychometric paradigms will be explored specifically in the context of the most widely used pediatric quality of life (QoL) measure in studies of cancer patients, the PedsQL^{TM} 4.0 Generic Core Scales (Varni, Seid & Rode, 1999). The psychometric properties of the PedsQL^{TM} are examined to provide an understanding of the similarities and differences between traditional and modern psychometric approaches in the evaluation of rating scales, and to highlight some of the additional information afforded by using Rasch analysis in establishing psychometrics.

The next two chapters present measurement issues related to assessment of QoL in pediatrics as well as the history and development of the PedsQL^{TM}. These chapters will provide insight as to why a Rasch analysis of this particular scale is warranted.

CHAPTER 2

QUALITY OF LIFE

Introduction

There is no consensus on the exact definition of quality of life (QoL) despite the growing number of publication that address the topic (Davis et al., 2006; Klassen et al., 2009). QoL is a broad multidimensional concept applied to an individual's status that includes economic welfare; characteristics of the community, such as crime rate and cultural and recreational amenities; characteristics of the environment, such as air and water quality; and health status (Patrick & Erickson, 1993). Measurement of QoL has been noted to be an arduous task due to the plethora of terms that are used to refer to the construct and the difficulty distinguishing it from other related constructs (Leplège & Hunt, 1997).

This confusion highlights the need for researchers and clinicians to pay close attention to the purpose of using a particular QoL measure for a specific patient population and to ensure that the appropriate measure is being used to answer their questions (Rosenbaum, 2009). The purpose of this chapter is to present the main issues that arise when trying to measure QoL in order to set the stage for a critique of how well QoL is measured by the PedsQL^{TM} 4.0 Generic Core Scales.

Defining Quality of Life

The definition of the term QoL has evolved greatly since the recognition that biomedical outcomes alone do not capture all the ways in which an individual is impacted by their illness and treatments (Davis et al., 2006; Drotar, 2004). This recognition has generated interest amongst stakeholders to assess the impact that products, policies, interventions and treatments may have on QoL, and is largely responsible for the creation of what some refer to as a "quality of life industry" (Rosenbaum, 2009). The definition of QoL used in this thesis will reflect that put forth by the World Health Organization (WHO) which states that QoL is an "individual's perception of their position in life in the context of their culture and value systems… and in relation to their goals, expectations and concerns" (WHOQoL Group, 1993, p.153).

Terms that are commonly used interchangeably to address the concept of QoL include health status, health-related QoL, functional status and functional well-being (Drotar, 2004; Leplege & Hunt, 1997; Rosenbaum, 2009). It is also important to address these related concepts but they should not be used mistakenly to refer to QoL. Health status refers to a person's level of wellness and thus a health status measure should include biological, psychological and social functioning domains with items that target performance, capacity, frequency, severity and the presence or absence of symptoms (Drotar, 2004). Health-related QoL is concerned with the opportunities that a person's health status affords, the constraints that it imposes upon the person and the value that a person places on his or her health status (Feeny, William, Mulhern, Barr, & Hudson, 1999). Functional status is defined as the ability to perform daily activities that are essential to meet basic needs (Drotar, 2004) and functional well-being describes how well a person can perform these daily activities. Functional measures typically include domains that address performance in physical, social and emotional functioning with items that tap into the specific activities corresponding to these domains.

This thesis is predominantly concerned with health-related QoL because the PedsQL^{TM} is most commonly cited as a health-related QoL tool. However due to the conceptual overlap between different QoL measures used in a cancer population, the broader term QoL will be employed throughout the thesis (Klassen et al., 2010).

Quality of Life Measures

In healthcare constructs cannot generally be measured by a gold standard criterion. Clinicians and researchers often proceed by administering a series of items, each of which is thought to reflect the underlying construct of interest, and then summing the scores of these items to produce an overall score that represents the construct (Teresi & Fleishman, 2007). QoL is a subjective perception that cannot be observed and is thus represented by the overall score of a series of items that measures manifestations of the elements being assessed.

As a consequence of the challenges associated with defining QoL, it has also been difficult to reach consensus on determining the domains that are important to assess in a QoL measure. For example, in a systematic review of QoL measures used in a cancer population it was found that over 30 domains were assessed in 20 different generic and specific QoL measures (Klassen et al., 2010). Measures that claim to assess QoL in pediatric cancer include a wide variety of domains including autonomy, behaviour, functional status, outlook on life and pain. There was no evidence to suggest differences in the construct being measured between tools that were labelled 'health-related QoL' measures versus 'QoL' measures. The content of these measures was categorized into a conceptual framework and results indicated very few differences between item pools that came from differently labeled QoL measures (Klassen et al., 2010).

Based on the WHO definition, a measure can only assess QoL if it targets a person's perceptions, including their goals and expectations for future abilities, regardless of their current health state. Items should address the importance, satisfaction or feelings the person has toward an issue and not their level of performance or degree of problems in a particular domain (Fayed, Schiariti, Bostan, Cieza, & Klassen, 2011).

Consequences of the Challenges Associated with Quality of Life Measurement

The confusion amongst terms used to describe QoL can be particularly problematic when scores on QoL measures are being used to make comparisons between healthy and diseased populations (Davis et al., 2006). For example, one would expect a greater degree of disparity in health status scores between healthy and diseased populations, whereas these groups may score more closely on a measure of QoL if the items are worded to capture the individual's perception of their functional well-being (Davis et al., 2006; Rosenbaum, Livingston, Palisano, Galuppi, & Russell, 2007). This notion is well expressed by Albrecht and Devlieger (1999) as the 'disability paradox'.

Additionally, in order to identify factors that correlate with higher or lower QoL scores, as is the goal in some studies that examine differences in QoL scores across different levels of illness in people with the same diagnosis, it is of paramount importance that the construct being measured is clearly defined. Confusion around the definition of QoL in these types of studies may lead to erroneous generalizations being made in reference to the impact of therapy on QoL for a particular sub-group of the diagnosis (Davis et al., 2006; Drotar, 2004).

For example, the study by Rosenbaum et al. (2007) illustrates that different measures, exploring different views of QoL can produce very different findings in adolescents with cerebral palsy. In this study QoL and health-related QoL of 203 adolescents with cerebral palsy were assessed using two measures that capture different perspectives of QoL. Participants were classified based on level of gross motor function using the Gross Motor Function Classification System (GMFCS), which ranges from Level 1 (Walks without Limitations) to Level 5 (Transported in a Manual Wheelchair). Subjective accounts of QoL were assessed using the Quality of Life Instrument for People with Developmental Disabilities, which provides an assessment of adolescents' perceptions of the degree to which the important possibilities of his or her life are enjoyed. These subjective accounts of QOL were contrasted to observations of health-related QoL based on parent-reported health status using the Health Utilities Index (HUI), which describes functional status. Results indicated that scores on the HUI only explained a small proportion of variance in subjective QoL. Furthermore it was found that subjective accounts of QoL did not vary significantly based on GMFCS scores, however objective accounts of health-related QoL did vary based on these scores (Rosenbaum et al., 2007).

The findings from the study described above suggest that it would be inappropriate to conclude that severity of cerebral palsy is associated with subjective QoL in adolescents and highlights the importance of understanding and defining the construct being measured in a study. Researchers and clinicians can use the findings on predictors of QoL to identify individuals with expected poor QoL and to target them for additional supportive care interventions; however, to do this the construct of QoL must be clearly defined. The challenges associated with defining and measuring QoL may make it difficult to provide supportive interventions that actually reflect the issues faced by the individual. Furthermore, it may also make it easier to neglect the possibility that other aspects of QoL, such as the value the individual places on the impact of therapy, will not be captured by the measure.

The multiple terms used to describe QoL have produced conflicting results in some studies that attempt to assess QoL longitudinally. For example, in a study to examine the impact of late effects of cancer treatment into adulthood it was found that perceived health-related QoL amongst cancer survivors is equal to or better than that of healthy controls (Pemberger et al., 2005). On the other hand, in a study by Novakovic, Fears, Horowitz, Tucker, & Wexler (1997) using the Karnofsky performance status scale, it was reported that sarcoma cancer survivors scored worse on functional status than healthy controls. The Karnofsky performance status scale is a measure of functional status and not health-related QoL, but because these terms are often used interchangeably to refer to QoL one may conclude that the results of these studies are conflicting. Contradictory conclusions such as the ones presented above may make it difficult for clinicians and researchers to evaluate the impact of treatments on QoL and to make decisions regarding future care for patients.

Generic and Condition-Specific Quality of Life Measures

There are two broad categories of measures used to assess QoL: generic and condition-specific measures. Generic measures enable comparisons to be made across multiple patient groups and they facilitate benchmarking of affected individuals with healthy controls. They are used to facilitate an understanding of how varied demographic or clinical groups differ in their reported QoL scores (Waters et al., 2009). These measures generally have higher specificity within a condition to assess risk factors for low QoL and to monitor treatment outcomes (Varni, Seid, & Rode, 1999). The disadvantage of using generic instruments is that they do not always tap into the specific health concerns that are associated with a particular condition and therefore they may not provide the necessary information clinicians or researchers need to answer their research questions.

Measures specific to individual conditions have items that capture the nuances associated with that particular condition. However, scores on these measures cannot be compared across diagnostic groups and therefore less generalizable conclusions can be drawn from them. Specific measures are more sensitive to change because of their focus on the distinguishing features of a particular condition (Waters et al., 2009). Specific measures are advantageous when the goal is to detect a precise outcome that is not common across multiple diagnoses (Davis et al., 2006). There is some debate in the literature regarding which type of measurement tool is best suited for rigorous QoL measurement.

Use of the Rasch analysis provides some benefit in dealing with the generic versus specific measures debate (Tennant, McKenna, & Hagell, 2004). If scales measure the same overall construct, items from different disease-specific scales can be calibrated on the same scale as the generic using Rasch analysis, given that some items that are common to both scales are employed (Tennant et al., 2004). This approach is currently being used to establish an item bank for disease-specific QoL measures in rheumatic diseases. Combining items in this fashion allows for disease-specific measurement while also permitting comparisons to be made across different diagnoses (McKenna, 2002; Tennant et al., 2004). These combined scales have potential to be used as outcome measures in clinical trials as they allow valid comparisons of QoL to be made across diseases and between healthy and diseased populations.

The issues regarding how to define QoL, and the type of instrument (generic or specific) used to measure it, are applicable to measurement of both adult and pediatric QoL. In the next section, issues specific to the measurement of pediatric QoL will be described.

Pediatric Quality of Life

There are special challenges that must be taken into consideration when measuring QoL in a pediatric population. Two common issues that arise in pediatric QoL measurement are the developmental age of the child being assessed and the use of parent-proxy reports.

It is important that pediatric QoL measures include both child and parent report versions of the scale (Vance, Jenney, Eiser, & Morse, 2001; Varni, 1999). Due to a child's cognitive immaturity, limited social experience and continued dependency, parents may be in a better position to rate some aspects of their child's QoL (or at least offer their perspectives on the child's QoL). Instrument developers are increasingly producing parallel parent versions to complement child report scales; however, the child self-report is the preferred approach of measurement (Davis et al., 2007; Upton, Lawford, & Eiser, 2008).

If for any reason a child is unable to provide a self-report parent-proxy assessments of QoL sometimes can provide the only means of obtaining information about that child's QoL (Meeske, Katz, Palmer, Burwinkle, & Varni, 2004). In a study conducted by Meeske et al. (2004), 95 of the 235 parents surveyed reported that their children could not provide self-report and therefore the parent-proxy report was used in place of the self-report. Due to the increased use of parent report QoL measures, examining concordance of parent and child report scores has become a relevant issue and is therefore discussed in the next section.

Parent-Child Agreement in Quality of Life Scores

Children and parents think about and interpret events differently and thus both perspectives are important to consider in assessing QoL, regardless of concordance of scores (Waters et al., 2009). It is important to consider the purpose of using a parent-proxy measure when reflecting on concordance or discordance of parent and child reports. If the goal is to develop a richer understand of QoL it is appropriate to administer a parent-proxy report even if the reported child-parent concordance for that measure is low. However if the proxy is to stand in place of the child report, concordance of scores is an important issue to be considered (Upton et al., 2008).

Although the parent report can be useful on its own, discordance between child and parent report remains a barrier to the exclusive use of parent report QoL measures in clinical and research settings. Discordance has been documented in QoL assessment of healthy children as well as children with health conditions including asthma, cystic fibrosis, chronic headache, limb deficiencies, arthritis and cancer (Varni, Seid, & Kurtin, 2001). A range of social, health and educational factors can influence differences in parent-child agreement. Lack of agreement also can be due to parents simply not knowing about certain aspects of their child's life, for example their schooling. Children may hide their feelings from their parents and in this case parents' perceptions would not be an accurate representation of their child's QoL.

Davis, Nicolas, Waters, Cook et al. (2007) conducted a qualitative study in which they interviewed parent and child respondents about their thought processes when responding to items on a QoL measure. Their findings suggest parents and children think about and interpret items differently and that they use differing response styles. Children tend to provide extreme scores (highest or lowest rating) and base their responses on one single example more often than parents, who base their responses on several examples in the child's life (Davis et al., 2007). The study concluded that children and parents interpret the meaning of items similarly and thus the discordance in their reported QoL scores can be attributed to different reasoning and response styles as well as the nature of the domain in question (Davis et al., 2007).

Agreement in QoL scores is typically lower for subjective issues such as depression and pain and higher for objective issues such as difficulty with mobility (Varni et al., 2001). A systematic review conducted by Eiser (2001) found that parent-child agreement was at least 0.5 for domains assessing physical function and symptoms (Eiser, 2001). Agreement was lower for more subjective domains assessing social and emotional function; correlations were typically less than 0.3 for these domains (Eiser, 2001).

Conclusions about the relationship between child and parent ratings are compromised by the limitations previously discussed with respect to QoL measurement in adult and pediatric populations alike. The weakness of psychometric properties established for parent-proxy QoL measures relative to self-report measures is an additional limitation specific to QoL measurement in pediatrics. Several studies were excluded from the systematic review conducted by Upton et al. (2008), examining parent-child agreement across several QoL measures, because psychometric properties were not reported for the parent version of the measure or because the studies did not differentiate between psychometric properties of the child and parent versions.

Conclusions from Upton's systematic review (2008) indicate that the PedsQL^{TM} Generic Core Scales is the most commonly used instrument in relation to assessing agreement between parent and child ratings. The parent-proxy version of the PedsQL^{TM} is constructed to provide a direct parallel to the items on the self-report; the only difference between the two is the use of the first person in the child report. The identical nature of the two versions has raised some concern that parents' perceptions are not being adequately captured; however, the similarity facilitates comparison of scores. Comparisons between parent and child reports in some pediatric QoL instruments are hindered due to lack of parallel content in the parent and child versions (Upton et al., 2008).

The next section of this thesis presents a critique of the different measures that are being used to assess pediatric QoL and discusses some applications of these tools.

Critique of Pediatric Quality of Life Tools

Pediatric QoL has emerged as an important outcome in light of the changed emphasis in pediatric healthcare from diagnosis and management of infectious diseases to prevention and control of chronic conditions (Upton et al., 2008). A sound pediatric QoL tool must include domains that measure a child's perception of their social, physical and emotional well-being (Davis et al., 2006). Despite this recognition, items that are appropriately worded to capture a child's perception of functional well-being in these areas are not used in the most common pediatric QoL measures (Davis et al., 2006).

In a study by Rajmil et al., (2003) that examined the similarities and differences between ten common generic pediatric QoL measures, it was found that items that assess physical, psychological and social aspects of health were included in each measure; however the distribution of items in these domains varied substantially. Furthermore, in a systematic review by Davis et al. (2006) it was found that most generic and specific pediatric QoL tools consist of items that assess difficulty, intensity, frequency and severity of physical symptoms or assess problems in activity performance. These findings do not resonate well with the widely held definition of health put forth by the WHO that good health is much more than just the absence of disease (WHOQoL Group, 1993).

Davis et al. (2006) suggests that the weak conceptual underpinnings of most pediatric QoL measures, regardless of their established psychometrics, render these tools as poor outcome measures to evaluate QoL. A rigorous QoL measure must be based on a clear definition of the construct, with a theory that supports the definition being employed, and the measure must include items that reflect a child's perception of an issue.

It is beneficial to use items that focus on positive life aspects in order to decrease the negative feelings a child may experience when responding to a series of items that focus solely on problem areas (Fayed et al., 2011). Waters et al. (2009) suggests that greater involvement of families and children in the development of pediatric QoL tools may serve as a method to ensure the items target relevant life areas and to improve item wording in order to eliminate emotional unease upon responding.

Applications of Pediatric Quality of Life Tools

Pediatric QoL tools are used as outcome measures in epidemiological studies, clinical trials and studies designed to improve a child's performance in specific areas of function; thus, it is crucial to ensure they are measuring the intended construct of interest (Varni, 1999; Varni, Burwinkle, Seid, & Skarr, 2003). To design interventions that sustain and improve QoL it is important to understand the variables that impact and explain different patterns of QoL in pediatrics (Klassen, Anthony, Khan, Sung, & Klaassen, 2011). Items on QoL measures must function free of bias for different sub-groups being assessed and the construct must be defined appropriately in order to produce valid results in these studies.

The demographic factors most commonly taken into account in cancer-specific pediatric QoL studies include age at time of assessment, gender, age at diagnosis and ethnicity (Klassen et al., 2011). Research on these factors has shown that if a child is older at the time of assessment they will report lower QoL scores and vice versa. Furthermore, being female is significantly associated with poorer QoL scores in all domains typically assessed, with the exception of the social function domain.

Limitations in pediatric QoL studies generally include small and heterogeneous sample sizes. Children in the sample typically vary based on a number of factors, which may inflate the true differences between treatment and control groups (Teresi & Fleishman, 2007). To improve overall QoL in pediatrics it is necessary to develop reliable and valid tools that can be used in research to pinpoint the determinants of QoL accurately; these determinants must then be targeted in intervention studies (Klassen et al., 2011). Rasch analysis serves as one means of improving the quality of pediatric QoL measures and can be used to address some of the measurement issues described above.

Applications of Rasch Analysis to Measurement of Quality of Life

The challenges associated with defining QoL, the large array of domains assessed by QoL measures and the degree of content overlap between differently labeled QoL measures, all highlight the need to improve our understanding of how to measure QoL. Rasch analysis can be used as a tool to further examine the properties of items that comprise the various domains found on QoL measures (Rajmil et al., 2004). The use of Rasch analysis can provide additional evidence to support the inclusion or exclusion of items on QoL measures thereby improving our knowledge and understanding of the concepts that comprise QoL. Furthermore Rasch analysis can provide additional information regarding items that function particularly well in terms of psychometric properties such as reliability and validity (Smith, 2001; Svensson, 2001).

It is essential that QoL measures be subject to psychometric testing, in addition to the initial psychometrics established by the developers of scale. This thesis presents the first Rasch analysis of the parent report PedsQL^{TM} 4.0 Generic Core Scales in a sample of parents of children on active cancer treatment. Chapter 5 presents the results of this study, where findings are compared to those obtained when the same data are analyzed using traditional methods.

Using the PedsQL^{TM}, it has been reported that parents typically rate QoL worse than their child with cancer. Insight into how these differences might arise remains limited (Vance et al., 2001; Varni, Burwinkle, Katz, Meeske, & Dickinson, 2002). Parent reports are sometimes used in isolation of child reports to make important decisions regarding future healthcare and therefore psychometric evalution of the parent report is essential (Varni, 1999). It is also important to ensure adequate psychometrics of the parent report as the internal reliability of that report will impact the level of agreement that can be expected between child and parent versions of the scale (Varni, Katz, Seid, Quiggins, & Friedman-Bender, 1998).

The individual reliability of child and parent measures provides a frame of reference for interpreting the agreement between them and if each measure is not individually reliable then high levels of agreement cannot be expected between respondents. Conducting a Rasch analysis of the parent report provides a method to gain further information about the psychometrics of the PedsQL^{TM} from an item-level perspective, which will complement the predominantly test-level and person-level information traditionally available on this tool.

Prior to introducing the above-mentioned study an overview of the development history and psychometrics of the PedsQL^{TM} 4.0 Generic Core Scales is presented.

CHAPTER 3

PEDSQL ^{TM} 4.0 GENERIC CORE SCALES

Development and History

The PedsQL^{TM} is a multi-dimensional child self-report and parent-report tool developed as a generic instrument that can be used with or without disease specific measures (Varni, 1999). The primary purpose of the measure appears to be to discriminate levels of QoL (Varni et al., 2003). The original version of the PedsQL^{TM} was developed in a sample of 291 English speaking pediatric cancer patients (aged 8-18 years) and their parents. Specifically the sample included 179 males and 112 females who were middle class Caucasians, Hispanics, African-Americans, Asians, or Native Americans. It was rationalized that due to the heterogeneity of a cancer diagnosis, it is appropriate to develop a generic QoL measure in a sample of cancer patients (Varni, 1999). Patients from all diagnostic groups were included as were patients at different times during the treatment and survivorship trajectory (i.e., newly diagnosed, on treatment, relapsed disease, recent remission, off-treatment and long-term off-treatment). Exclusion criteria were patients who had co-morbidities.

Item generation involved a five-year, multi-phased process. Items were assembled from the Pediatric Cancer QoL Inventory-32, literature review, interviews with patients and families and discussions with healthcare professionals. It is imperative that children and parents are consulted in the development of items for any pediatric patient-reported outcome measure; however, the exact extent of parent or child involvement must be clear (Waters et al., 2009). No details of the interviews and discussions that took place to develop the PedsQL^{TM} are presented and it is unclear exactly how items were generated from the data collection procedures described (Varni, 1999). Furthermore it is unclear how the developers decided that the content of the parent and child versions of the PedsQL^{TM} and the content of the versions used for different age groups should be identical (Varni, 1999).

Item reduction was an iterative process that involved administering the item pool to new patients and families and interviewing healthcare providers for additional input. The final tool consisted of three core domains: 1) physical (6 items measuring functional status in activities of daily living); 2) psychological (5 items measuring emotional distress); and 3) social (4 items measuring interpersonal function in peer relations). A corresponding symptom-specific cancer module was also developed at this time. Construct validity of the generic and symptom specific PedsQL^{TM} cancer module was assessed using standardized measures of emotional distress, perceived competency and social functioning (Varni, 1999).

The study conducted for this thesis is based on the PedsQL^{TM} 4.0 Generic Core Scales parent-report which can be found in Appendix 1. The PedsQL^{TM} 2.0

and 3.0 versions included additional items as well as a more sensitive scaling range (a 5-point scale was introduced instead of the 4-point scale used in version 1.0). Response options are as follows: Never (score of 0); Almost Never (score of 1); Sometimes (score of 2); Often (score of 3); and Almost Always (score of 4). The 4.0 version, also a 5-point scale, was specifically adapted to measure the core health dimensions delineated by the WHO and thus a fourth domain, school function, was added to the measure (Varni et al., 2003).

The PedsQL^{TM} 4.0 Generic Core Scales consists of 23 items that are applicable for healthy, school and community populations as well as pediatric populations with acute and chronic conditions. There are separate self-report forms for children aged 5-7, 8-12 and 13-18 years and parent-report forms for children aged 2-4, 5-7, 8-12 and 13-18 years. All items are reverse scored and linearly transformed to a 0-100 scale such that a higher score indicates better QoL. In order to account for missing data, scale scores are computed as the sum of items divided by the number of items answered as long as at least fifty percent of the items are answered (Varni, 1999). Three separate summary scores are reported: 1) the physical summary score (sum of items 1-8 in physical function scale); 2) the psychosocial summary score (sum of items 9-23 in the social, emotional and school function scales); and 3) the total score (sum of physical and psychosocial summary scores) (Varni, 1999).

The items on the different forms of the PedsQL^{TM} vary only in the use of developmentally appropriate language and tense. For example, in the self-report for children ages 5-7 years, a 3-point scale is utilized to reflect the developmental level of this age group. The parent-report includes an additional form for the 2-4 year old age group as research suggests children under the age of five are developmentally unable to complete self-report questionnaires (Riley, 2004; Rajmil et al., 2004). The 2-4 year old form includes only 3 items on the school function scale, as the remaining items are not applicable (Varni, 1999).

Critique of PedsQL^{TM} 4.0 Generic Core Scales

The instructions for the PedsQL^{TM} ask how much of a problem each item has been during the past one month, reflecting the predominantly medically-oriented paradigm in clinical care settings and an assumption that problems exist (Fayed et al., 2011). If a child reports no problems in a given domain, their score will be higher on that domain and will suggest a more positive QoL. Domains that assess problems with only a specific activity have the capacity to assess reduced QoL (Waters et al., 2009). The underlying assumption that an absence of problems is equal to a higher QoL has not been empirically tested; in fact, there is research that indicates a high level of ill-being is not the same as a low level of well-being (Leplège & Hunt, 1997; Rosenbaum et al., 2007; Russell, Hudson, Long, & Phipps, 2006).

As mentioned in the previous chapter, problem-focused items can be damaging to a child's self esteem especially if they are already experiencing additional stress due to a health condition or if they report problems in a number of areas (Waters et al., 2009). Phrasing questions in a neutral fashion and asking about positive aspects, such as performance and abilities, will afford stakeholders a better understanding of children's and parents' perspectives of their child's QoL. This positive outlook on QoL is more reflective of how the idea has been defined for the purpose of this thesis.

It is important to consider the theoretical focus and influence behind the domains, items and scoring procedures of a measure when selecting a QoL tool for use in a specific patient population (Waters et al., 2009). Inherent in the PedsQL^{TM} is a perspective that captures functionality and health status rather than health-related QoL, which is what the measure is commonly cited for and used to report. The items do not reflect a subjective perception of well-being and functionality, as required for assessing QoL; rather, items are focused on disability and impairment in specific activities. For example, in the physical function domain, two items ask about the extent of a child's problems with walking and running. Several studies support the notion that function and impairment are distinct entities from subjective experiences of QoL and thus it appears the PedsQL^{TM} does not adequately capture the subjectivity associated with QoL (Smith, Avis, & Assmann, 1999).

A QoL measure must provide a means of assessing a construct that is separate from biomedical function (Waters et al., 2009). Literature has indicated that individuals living with disabilities who have functional impairments can still report high satisfaction with some aspects of their lives (Rosenbaum et al., 2007). This satisfaction would not be captured if a strictly functionality-driven assessment tool such as the PedsQL^{TM} were used to report QoL.

Psychometric Properties of PedsQL^{TM} 4.0 Generic Core Scales

The PedsQL^{TM} satisfies a number of measurement criteria and is the most widely used generic QoL measure in pediatrics (Eiser & Morse, 2001). There are more psychometric data reported for this measure than any other QoL tool (Eiser & Morse, 2001). A complete description of the psychometric properties of the PedsQL^{TM} 4.0 Generic Core Scales can be found in (Varni et al., 2003). A brief summary of the psychometric properties of the PedsQL^{TM} follows.

To date the PedsQL^{TM} 4.0 Generic Core Scales has been evaluated psychometrically from a Classical Test Theory (CTT) standpoint with the exception of three studies (Hill et al., 2007; Kook & Varni, 2008; Lamoureux et al., 2010). The first study evaluated the Korean translation of the PedsQL^{TM} 4.0 in a sample of healthy school age children to establish cross-cultural validity and the second evaluated the validity of the PedsQL^{TM} 4.0 in a sample of pre-school children with refractive errors living in Singapore (Kook & Varni, 2008; Lamoureux et al., 2010). The third study utilized item response theory; however, the purpose of the study was to establish which items on the PedsQL^{TM} 4.0 could be useful in a QoL item bank and not to evaluate the PedsQL^{TM} 4.0 from a modern psychometric standpoint (Hill et al., 2007).

The key studies that established psychometrics of the PedsQL^{TM} 4.0 using CTT are presented below. In the first study conducted on psychometrics, the PedsQL^{TM} 4.0 was administered to 963 children and 1629 parents from three settings: pediatricians' offices, hospitals and specialty clinics. The specialty clinics included orthopedics, rheumatology, and diabetes (Varni et al., 2003). A diverse sample was used that varied by age, ethnicity, gender, availability of health insurance, socio-economic status and health condition (chronic condition, acute condition or healthy).

The internal consistency reliability of the total scale score for the parent and child reports had α values close to 0.9, which is the level required to make individual-level decisions (Bland & Altman, 1997). It is suggested that as a result of this high overall internal consistency the PedsQL^{TM} 4.0 can offer useful information in clinical trials, research, clinical practice, school-health and community populations (Varni et al., 2003). The individual scales that comprise the PedsQL^{TM} 4.0 had lower reliability values suggesting that these scales are more appropriate to be used for group-level comparisons. These findings are sample dependent and thus reliability testing should be conducted again if the scale is being applied to a different population (Hobart & Cano, 2009). Table 1a presents the internal consistency of each individual scale as well as the summary scores of the PedsQL^{TM} 4.0 based on respondent and age group.

Table 1a: Internal consistency reliability of PedsQL^{TM} 4.0 Generic Core Scales

Internal Consistency (α)	Child Report (Ages 5-18)	Parent Report (Ages 2-18)
TS	0.88	0.90
PF	0.80	0.88
PS	0.83	0.86
EF	0.73	0.77
SF	0.71	0.75
ScF	0.68	0.76

TS= total score; PF = physical function summary score; PS= psychosocial function summary score; EF= emotional function score; SF= social function score; ScF= school function score (Varni et al., 2003)

The developers of the PedsQL^{TM} hypothesized that the scale would produce a two-factor solution consisting of a physical and a psychosocial function summary score that could be summed together to represent overall QoL (Varni et al., 1999). Factor analysis of the self- and parent-report data did not support this hypothesized factor structure as results suggest a five-factor solution for both reports.

The self-report factor solution accounted for 52 % of the total variance in the data while the parent-report factor solution accounted for 62% of the total variance (Varni et al., 1999). Streiner's (1994) article on factor analysis states that a scale should account for at least 60 % of the total variance in the data and thus further work is necessary to support the construct validity of the PedsQL^{TM} sub-scales as measuring distinctly unique dimensions of QoL.

Construct validity of the PedsQL^{TM} was also considered using the known groups method. Based on the results of an ANOVA of difference scores the PedsQL^{TM} overall score distinguished between children who were healthy and children who had either an acute or chronic health condition (Varni et al., 1999). Furthermore, PedsQL^{TM} scores were related to accepted indicators of morbidity and illness burden such as days requiring care, fewer days missed from school for children and work for parents, and less impact on work routine and concentration for parents who worked outside the home (Varni et al., 1999). These relationships reflect the concurrent validity of the PedsQL^{TM} in the sample assessed. There is validity evidence (construct, concurrent and known-groups) to support the intent of the PedsQL^{TM} as a descriptive and a discriminatory tool (Varni & Setoguchi, 1992; Varni et al., 1999; Varni et al., 2002; Varni et al., 2003). Based on the literature reviewed about the PedsQL^{TM}, this measure should not be used as a tool to predict change, as this purpose requires evidence of predictive validity assessed in a prospective longitudinal study (Rosenbaum, 1998).

There was a moderate ceiling effect associated with some scales of the PedsQL^{TM} 4.0 as displayed in Table 1b below. Ceiling effects of the self-report version ranged from 5-33% for individual sub-scales in the proportion of the sample with either an acute or chronic condition. Ceiling effects ranged from 12-47% for each sub-scale in the healthy children in the sample. For the parent report, ceiling effects ranged from 5-34% for parents of children with acute and chronic conditions; and from 13-58% for parents of healthy children. Floor effects were minimal in both child and parent reports (Varni et al., 2003).

Table 1b: % Floor and ceiling effects for self and parent reports of the PedsQL^{TM} 4.0

Respondent	Ill		Healthy	
	Ceiling %	Floor %	Ceiling %	Floor %
Child	Total: 1.9 PF: 13.1 PS: 5.2 EF: 22.4 SF: 33.2 ScF: 13.0	Total: 0 PF: 0 PS: 0 EF: 0.3 SF: 0 ScF: 0.3	Total: 7.2 PF: 25.8 PS: 12 EF: 29.8 SF: 47.1 ScF: 23.1	Total: 0 PF: 0 PS: 0 EF: 0.8 SF: 0 ScF: 0.5
Parent	Total: 4.1 PF: 18.1 PS: 5.6 EF: 19.5	Total: 0.2 PF: 2.3 PS: 0.2 EF: 1.4	Total: 10.3 PF: 39.6 PS: 13.8 EF: 29.5	Total: 0 PF: 0 PS: 0 EF: 0.1

	SF: 34.4 ScF: 15.5	SF: 0.5 ScF: 1.7	SF: 58.1 ScF: 34.5	SF: 0 ScF: 0.3

TS= total summary score; PF = physical function summary score; PS= psychosocial function summary score; EF= emotional function score; SF= social function score; ScF= school function score (Varni et al., 2003)

Other important factors to consider in evaluating the psychometric properties of a measure include sensitivity and responsiveness. The term sensitivity is usually used to refer to the ability of a screening tool to detect people who have the condition being screened (Rosenbaum, 1998). In the study discussed below, Varni et al. (2002) uses the term sensitivity to refer to the ability of an instrument to detect small levels of change. The available literature does not indicate evidence of predictive validity; thus it may not be appropriate to evaluate sensitivity (using Varni's definition of the term) or responsiveness of the PedsQL^{TM} scores. A non-statistical factor that can be considered includes the impact of a tool on clinical decision-making.

Sensitivity, responsiveness and clinical impact were assessed in three studies conducted and details of each study can be accessed in the study by Varni et al. (2001). The ability of the PedsQL^{TM} to be sensitive to small group differences among patients with increasing degrees of cardiac disease severity was calculated using ANOVA and the responsiveness of individual-level patient change was calculated using paired t-tests. A third study calculated the impact of PedsQL^{TM} scores on clinical decision-making based on the effect size of the magnitude of change in individual patient overall scores (Varni, Seid, Knight, Uzark, & Szer, 2002).

In the three studies mentioned above, the PedsQL^{TM} was administered to 209 children and 269 parents from pediatric cardiology, orthopedics and rheumatology clinics. Findings suggest that PedsQL^{TM} scores were sensitive to increasing degrees of cardiac disease severity based on the cardiac disease severity rating system developed by the New York Heart Association (i.e., they are discriminative). Greater cardiac disease severity was associated with lower overall PedsQL^{TM} scores (Varni et al., 2002). In the orthopedics setting, PedsQL^{TM} scores were responsive with statistically significant changes in scores from the initial clinic visit for the treatment of a fracture to the subsequent follow-up visit when the child had returned to good health. In the rheumatology clinic, PedsQL^{TM} scores demonstrated an impact on clinical decision-making; when the pediatric rheumatologist examined the completed PedsQL^{TM} instrument at the point of service and made a clinical intervention decision based on the findings, subsequent PedsQL^{TM} scores were significantly higher (Varni et al., 2002). Further studies are recommended in order to generalize these findings to other populations.

This thesis will present a study evaluating the psychometric properties of the parent-report PedsQL^{TM} 4.0 Generic Core Scales in a pediatric cancer population. Self-report data were not collected as part of this study as the aim of the study was to focus on parents' perspectives. To facilitate an understanding of how psychometric properties may differ when using traditional versus modern psychometric frameworks, the next section of this thesis discusses the psychometric properties of the self- and parent-report PedsQL^{TM} in pediatric cancer from a traditional perspective.

Psychometric Properties of the PedsQL^{TM} 4.0 Generic Core Scales in Cancer Patients

The PedsQL^{TM} 4.0 self-and parent-reports have been used extensively as assessment tools in pediatric oncology clinical trials and in clinical practice. The psychometrics of the scale were established for a pediatric cancer population in a study that involved administering the scale in conjunction with the PedsQL^{TM} Multidimensional Fatigue Scale and the PedsQL^{TM} 3.0 Cancer Module in a sample of 339 families (Varni et al., 2002). Two hundred and twenty self-reports were collected from children aged 5-18 and 337 parent proxy-reports were collected from parents of children aged 2-18 years. Child self-report and parent proxy-report were available on 190 parent/child dyads (Varni et al., 2002). The sample included patients from various diagnostic groups and in different phases of treatment. Exclusion criteria included the presence of co-morbidities. Internal consistency was high for all sub-scales and summary scores of the self- and parent-report, as displayed in Table 2 below.

Table 2: Internal consistency reliability of self and parent reports of the PedsQL^{TM} 4.0 in pediatric cancer

Internal Consistency (α)	Self Report (Ages 5-18)	Parent Report (Ages 2-18)
TS	0.88	0.93
PF	0.81	0.89
PS	0.83	0.89
EF	0.73	0.80
SF	0.70	0.73
ScF	0.66	0.77

TS= total summary score; PF = physical function summary score; PS= psychosocial function summary score; EF= emotional function score; SF= social function score; ScF= school function score (Varni et al., 2002)

These results suggest that the internal consistency reliability of the total score for the parent-report is adequate for individual-level patient analysis and that the total score for the self-report is adequate for group-level analysis. The physical function and psychosocial summary scores are acceptable for group-level decision-making for both self- and parent-reports. The scores for the individual scales of the PedsQL^{TM} 4.0 have lower internal consistency reliabilities and should primarily be used for descriptive and exploratory analysis until further studies are conducted.

As discussed in Chapter 2, parent-child agreement (inter-rater reliability) is an important issue to consider in pediatric QoL measurement. Prior to comparing groups of respondents on a measure it is important to ensure that the items comprising the measure operate equivalently across the different groups (Teresi, 2006). If items on the scale do not function equivalently for parent and child respondents it may not be appropriate to make claims of parent-child agreement. Previous literature reviewing parent-child agreement for the PedsQL^{TM} from a traditional standpoint is presented below to facilitate an understanding of how well the parent-report, which is analyzed in this thesis, correlates with the self-report.

Parent-Child Agreement for PedsQL^{TM} 4.0 Generic Core Scales

Upton, Lawford & Eiser (2008) conducted a systematic review of pediatric QoL instruments in relation to parent-child agreement. Their objectives were to evaluate the inter-rater reliability of available parent-child measures, to determine the factors that influence the level of parent-child agreement and to explore the direction of differences in parent and child reports. Their search, conducted from 1999-2006, revealed that parent-child agreement of QoL instruments was evaluated most commonly using the PedsQL^{TM}. Sixteen of the 19 studies included in the review used the PedsQL^{TM}; 7 used the generic core scales, 8 used the generic core scales as well as one of the disease specific modules, and one used the cancer-specific module (Upton et al., 2008).

The 16 studies that used the PedsQL^{TM} and assessed parent-child agreement indicate moderate-good agreement for all sub-scales in the measure. These studies were conducted in several populations including children and parents of children with cancer, epilepsy, attention deficit hyperactivity disorder, rheumatoid diseases, asthma, and heart disease as well as with healthy children and their parents (Upton et al., 2008). Five studies using the PedsQL^{TM} (Felder Puig et al., 2004; Poretti, Grotzer, Ribi, Schönle, & Boltshauser, 2004; Uzark, Jonesa, Burwinkle, & Varni, 2003; Varni et al., 2002; Varni et al., 2002) report higher parent-child agreement for concrete, observable characteristics; four other studies found higher levels of agreement for more subjective characteristics in the psychosocial domains (Eiser, Vance, Horne, Glaser, & Galvin, 2003; Vance et al., 2001; Varni, Burwinkle, Rapoff, Kamps, & Olson, 2004; Varni & Burwinkle, 2006).

Studies that assess the level of agreement between parents and children do not often differentiate the level of agreement based on variables such as age, gender or health condition (Upton et al., 2008). Whether levels of agreement vary based on these factors is a future area of research to be explored. Assessment of item bias using Rasch analysis may serve as a useful method to ensure that items operate equivalently for these different sub-groups within a sample.

Literature examining parent-child agreement in QoL scores suggests a need to provide further evidence supporting the reliability and validity of parent-proxy scales and to systematically investigate variables that may impact the level of parent-child agreement. Rasch analysis may be a useful tool to address this issue. Conducting a Rasch analysis of the parent-proxy version of the PedsQL^{TM} can offer additional credibility for its use as a substitute for self-report when a child is unable to respond. The next chapter of this thesis presents a detailed overview of the procedures and statistics examined in a Rasch analysis.

CHAPTER 4

METHODOLOGY

Rasch Analysis

Rasch analysis can be used in any instance when items from a measure are summed together to form an overall score or sub-scale score. The three main applications of Rasch analysis are in the development of a new measure, the psychometric evaluation of an existing measure, and the creation of item banks for computer adaptive testing (CAT) (Tennant & Conaghan, 2007).

This thesis focuses on the application of Rasch analysis to evaluate psychometrically an existing measure, the parent-report of the PedsQL^{TM} 4.0 Generic Core Scales. The purpose of the analysis is to highlight any additional information that the use of Rasch can afford, as well as to illustrate how its use can address some of the inherent challenges of measuring QoL in pediatrics.

The Rasch measurement model is cited as the most common application of modern psychometric methods in health measurement and thus has been selected as the method to use in evaluating the PedsQL^{TM} for the purpose of this thesis (Tennant et al., 2004). The Rasch model has been used extensively in the education literature for the last 40 years and over the past decade it is being increasingly used in the health science literature (Tennant & Conaghan, 2007).

Rasch analysis is based on the testing of a rating scale against a mathematical measurement model, the Rasch model (Schumacker & Smith, 2007). The Rasch model is a probabilistic form of Guttman scaling and shows what should be expected in responses to items if interval scale measurement is to be achieved. Guttman scaling is a deterministic pattern that presumes a hierarchical ordering of items such that if a respondent has endorsed an item representing a task of average difficulty then all easier items should also be affirmed (Schumacker & Smith, 2007). The Rasch model asserts that if a harder task is endorsed then there is a higher *probability* that an easier task will also be endorsed. From a set of items that are summed to form an overall score, Rasch analysis determines the extent to which the response pattern deviates from the expected pattern of responses that would satisfy the Rasch model (Tennant & Conaghan, 2007).

Rasch analysis is particularly useful in the development of a new measure, as it is possible from the onset to conceptualize a construct and then develop items that are likely to fit model expectations. When used to review the psychometrics of an existing scale, items with poor model fit can be altered and retested, or deleted, or new items can be developed in order to improve targeting of a construct and meet the requirements of the Rasch model. In CAT the use of modern psychometrics makes it possible to estimate person ability levels with any subset of items in an item pool; therefore, it is possible to administer only the items that are required to discriminate amongst individuals, reducing respondent burden (Hays, Morales, & Reise, 2000).

In the health science literature, Rasch analysis is mainly carried out with Winsteps (Linacre, 2007) or Rumm software (Andrich, Lyne, Sheridan, & Luo, 2010); however, many other software packages are available. In this thesis Rumm2030 software, the most up-to-date version, was used. Once data have been entered into the program, the first step involves evaluating overall fit to the Rasch model. The software generates expected scores based on the Rasch model. The differences between observed scores and expected scores based on the Rasch model are examined for each person and item in the analysis. The term 'response residual' refers to what is left over after the portion that accounts for what fits the Rasch model is taken into consideration. The data that fit the Rasch model are referred to as the 'Rasch factor' (Pallant & Tennant, 2007; Tennant & Conaghan, 2007).

Rumm software produces a Chi square (χ^2) fit statistic that indicates the significance of the difference between observed and expected responses across groups representing different ability levels (known as 'class intervals') across the trait being measured (Tennant & Pallant). The software orders all respondents in terms of their ability on the construct being measured and then automatically splits the sample into sub-groups of equivalent sample size in order to approximate ability groups. The number of class intervals depends on the sample size; larger samples will be divided into a greater number of class intervals (Tennant & Pallant).

The χ^2 values for each test are summed to give an overall χ^2 value for the item with the associated degrees of freedom (Tennant & Pallant). A Bonferroni adjustment is applied to the alpha value (set at 0.05 in the software) to account for the influence of summing χ^2 values over multiple tests (Norman & Streiner, 2008). If the alpha value of the overall χ^2 statistic is less than the Bonferroni-adjusted value for alpha, the item is deemed to misfit model expectations significantly.

A second χ^2 value, the item-trait interaction χ^2, reflects how well the property of invariance across the trait being measured is achieved. This value is calculated by summing the χ^2 values across all items in the scale. A significant Bonferroni-adjusted χ^2 value indicates that the hierarchical ordering of the items varies across the trait and that there is substantial deviation from the Rasch model (Pallant & Tennant, 2007; Tennant & Pallant). It is important to note that perfect model fit is rarely achieved; instead, the goal is to figure out the solution that retains the most items, meets the two assumptions of the Rasch model, and therefore *essentially* fits the Rasch model (Hill et al., 2007).

The process of obtaining the solution that maximizes fit to the Rasch model can be thought of as a systematic trial and error process, and each step in this process will be described below. A flow diagram of the Rasch analysis process can be found in Appendix 2. In addition, where there are equivalent procedures established to test various psychometric properties in a CTT paradigm, these will be described to facilitate comparison of the two paradigms in evaluating the PedsQL^{TM} 4.0 Generic Core Scales in the proceeding chapter of this thesis.

Unidimensionality and Local Independence

The first step of the systematic trial and error process in a Rasch analysis involves testing the assumptions of the model to determine what strategies should be tried to improve overall fit. In Rasch, a principal component analysis (PCA) of the residuals is conducted to assess dimensionality by testing the hypothesis that the scale is unidimensional (Schumacker & Smith, 2007).

The first component of a PCA is identification of the primary factor contributing to the variance in the data. In this case, the first factor is essentially the second dimension of the scale, since the first dimension is what has already been accounted for as the Rasch factor. The greatest positive and negative residual loadings on this factor represent the two subsets of items that have the greatest likelihood of producing significantly different person estimates. Responses to any subset of items within a scale should give the same estimate of person ability if the scale is appropriately targeted (Tennant & Conaghan, 2007). If the content of the scale is unidimensional then it follows that the estimate of person ability generated by items will be the same (Tennant & Pallant). If there is multidimensionality in the scale there will be a significant difference in the person ability estimates generated by a subset of items (Tennant & Pallant).

Person ability differences are assessed by a series of independent t-test comparisons of person locations that are estimated from the two subsets of items that have the greatest residual loadings on the first factor of the PCA of residuals. The series of t-tests is carried out in an attempt to challenge the assumption of unidimensionality in the data set. If the proportion of significant t-tests is greater than the proportion allowed to assume unidimensionality (set at 5% of the total number of t-tests conducted) then it is assumed there is multidimensionality in the scale (Tennant & Pallant).

The binomial test of averages can be used to assess the acceptable proportion of significant t-tests given a particular sample size. This procedure uses an exact test of the statistical significance of deviations from a theoretically expected distribution of observations into two categories and it generates a 95% confidence interval around the acceptable proportion of significant t-tests (Tennant & Pallant). It is not necessary to use the binomial test of averages if the proportion of significant t-tests is already less than 5% (Tennant & Pallant).

In traditional psychometrics, a factor analysis is used to assess dimensionality. The results of the factor analysis determine whether or not the pattern of responses on a set of items can be explained by a smaller number of underlying characteristics (Streiner, 1994). If certain items on a questionnaire are more interrelated than others (based on the strength of their correlations), they are grouped into a sub-domain that reflects a component of the overall construct being measured. A PCA is conducted to derive the factors that account for the greatest variance in the scale, and this factor structure typically is conceptually compared to an a priori hypothesized factor structure to confirm dimensionality (Streiner, 1994).

The factors can be thought of as a series of multiple regression equations that represent weighted combinations of all variables in the analysis (Streiner, 1994). The output of a factor analysis describes how much variance is accounted for by each factor (its Eigenvalue) and how much each item loads onto each factor. To determine how many factors to retain, most statistical programs use the Kaiser criterion. This criterion states that each retained factor must account for at least the amount of variance introduced by a single variable; therefore, factors with Eigenvalues less than 1 are ignored (Streiner, 1994).

A factor-loading matrix shows the correlations between each item in the measure and the factors derived by the analysis. Only factor loadings with a correlation of at least 0.3 are taken into consideration (Streiner, 1994). Items may load on more than one factor, in which case they are considered factorially complex variables and should be placed in the factor with the greatest loading. If the complex factor loadings have a difference of less than 0.05 the content of the individual item must be examined to determine where it should be placed (Streiner, 1994). Factorial complexity makes it difficult to understand what construct is being measured by the item and can also mean that the factor is comprised of high scores on some items and low scores on others so attention to item wording may be useful.

Rotations are applied to the variables in a factor analysis to increase the clarity of the hypothetical sub-construct each factor may represent. Orthogonal rotations are often applied because they permit examination of each factor in isolation of the other factors. Each factor retained in the analysis must be comprised of at least three items in order to assess a separate dimension of the scale (Streiner, 1994). Factors with fewer than three items can be discarded. Furthermore a factor analysis can only be used to assess dimensionality if there is a minimum of five subjects per item in the analysis (Streiner, 1994).

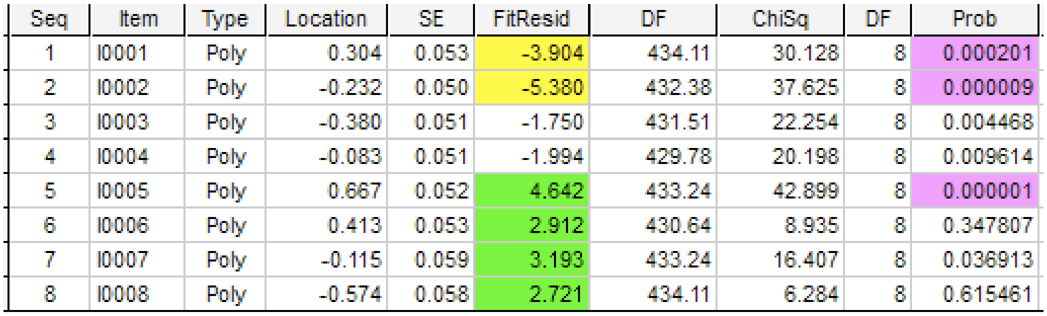
The PCA conducted in a traditional paradigm involves the entire data set and not just on the item residuals, as is the case in a Rasch analysis. Furthermore, the purpose of a PCA in traditional psychometrics is to explore the existence of multiple dimensions or to find support for the existence of hypothesized dimensions; whereas in Rasch analysis, the purpose is to find support for a unidimensional scale.

If data do not fit the Rasch model (as indicated by a significant overall item-trait interaction χ^2) and it has been determined that the scale is made up of more than one dimension, the next step is to examine individually any problematic items and persons. It is important to identify the exact sources responsible for misfit to the model in order to determine the best strategies to correct for the misfit that will permit retention of the greatest number of items.

Individual Item and Person Misfit

Overall misfit to the Rasch model is sometimes attributed to either a particularly problematic respondent or item. Problematic items or respondents are flagged using the "Fit Residual" function in Rumm software as displayed in Figure 3a below.

Figure 3a: Item fit statistics



Seq	Item	Type	Location	SE	FitResid	DF	ChiSq	DF	Prob
1	I0001	Poly	0.304	0.053	-3.904	434.11	30.128	8	0.000201
2	I0002	Poly	-0.232	0.050	-5.380	432.38	37.625	8	0.000009
3	I0003	Poly	-0.380	0.051	-1.750	431.51	22.254	8	0.004468
4	I0004	Poly	-0.083	0.051	-1.994	429.78	20.198	8	0.009614
5	I0005	Poly	0.667	0.052	4.642	433.24	42.899	8	0.000001
6	I0006	Poly	0.413	0.053	2.912	430.64	8.935	8	0.347807
7	I0007	Poly	-0.115	0.059	3.193	433.24	16.407	8	0.036913
8	I0008	Poly	-0.574	0.058	2.721	434.11	6.284	8	0.615461

Probabilities for items that significantly deviate from the Rasch model are highlighted in Fuchsia, flagging the corresponding item as being potentially problematic. Fit statistics are transformed to z-scores such that they represent a normal distribution. If the items and persons displayed fit to the Rasch model perfectly one would expect the overall mean (μ) and standard deviation (SD) (σ) of items and persons to be consistent with that of a normal distribution (μ,σ = 0, ±1). Higher or lower values indicate misfitting items or persons. An item or person is in the acceptable range if its mean and SD lie between -2.5 and 2.5 units. Items that have means below -2.5 are highlighted yellow and items that have means above +2.5 are highlighted green in Figure 1 above.

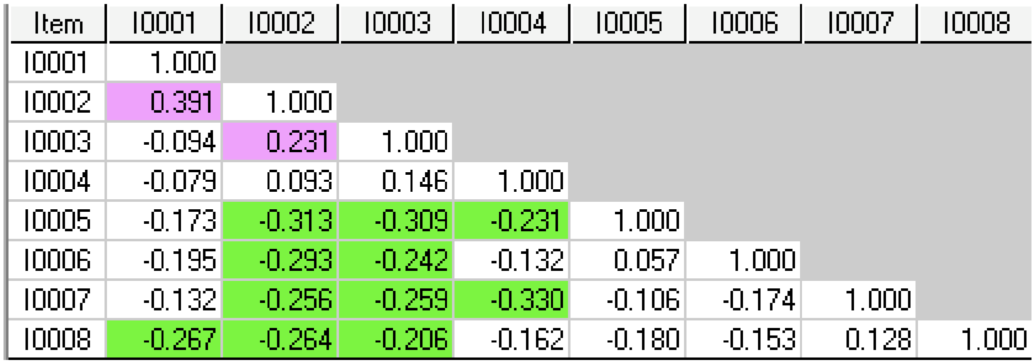
Items on the scale are subject to an additional criterion in order to fit the Rasch model: they must have non-significant Bonferroni-adjusted χ^2 values. Items that have significant χ^2 values are flagged by Rumm software, facilitating their potential removal if subsequent strategies to improve model fit are unsuccessful.

The terms 'item fit' and 'person fit' are not meaningful in the traditional paradigm. In CTT, data are not being examined in relation to an a priori established mathematical model and thus the *concept* of "item and person fit" is important but is examined using a different approach.

CTT approaches are largely non-parametric as they are not based on a mathematical model (Hambleton & Jones, 1993). CTT examines item and person characteristics by looking at descriptive statistics such as means and SDs; however these statistics are not converted to z-scores, reflecting the non-parametric nature of the traditional paradigm. Extreme mean scores on items (i.e., 6.8/7 or 0.2/7 on a 7-point scale) do not provide useful information, as they do not target the majority of the sample (Streiner & Norman, 2008). Items that produce means in these ranges are examined for potential removal from the scale. Items with narrow SDs and with a high proportion of missing data are also flagged as being problematic and are considered for removal from the scale (Streiner & Norman, 2008).

Unique to CTT, item-total correlations (ITCs) are also examined for each item as an indicator of the dimensionality of the scale. ITCs for items in each sub-scale must be between 0.2 and 0.7 to ensure that items within that particular scale correlate more highly to items in that sub-scale than to other items and are not redundant (Streiner & Norman, 2008). Item redundancy is commonly referred to as 'local dependency' in a Rasch analysis and is assessed by examining a map that displays residual correlations of all items in the scale. Clustering of item residuals with item-item correlations greater than 0.2 may indicate local dependency. See Figure 3b for a sample residual correlation map. Items with high residual correlations can be sub-tested to account for their local dependency. Creating a sub-test between highly correlated item residuals corrects for local dependency by combining the scores of dependent items such that they behave as one item (Pallant & Tennant, 2007).

Figure 3b: Residual inter-item correlation matrix



Item	I0001	I0002	I0003	I0004	I0005	I0006	I0007	I0008
I0001	1.000							
I0002	0.391	1.000						
I0003	-0.094	0.231	1.000					
I0004	-0.079	0.093	0.146	1.000				
I0005	-0.173	-0.313	-0.309	-0.231	1.000			
I0006	-0.195	-0.293	-0.242	-0.132	0.057	1.000		
I0007	-0.132	-0.256	-0.259	-0.330	-0.106	-0.174	1.000	
I0008	-0.267	-0.264	-0.206	-0.162	-0.180	-0.153	0.128	1.000

Inter-item correlations with an absolute value greater than 0.2 are highlighted in Figure 3b above. Positively correlated items are highlighted in fuchsia and negatively correlated items are highlighted in green.

The descriptive statistics examined in CTT are also examined in Rasch analysis (although different terminology may be used); however, in Rasch analysis these descriptive statistics are examined more informally and typically are not the parameters that have the greatest influence on item reduction.

In Rasch analysis, when flagging an item for removal, the significance of the χ^2 value for each item is the primary factor taken into consideration. Two other reasons why items could be misfitting the Rasch model include the following: inconsistent use of response options (referred to as threshold disordering), and item bias across groups of respondents (referred to as differential item functioning). These two issues are discussed in the following sections.

Threshold Disordering

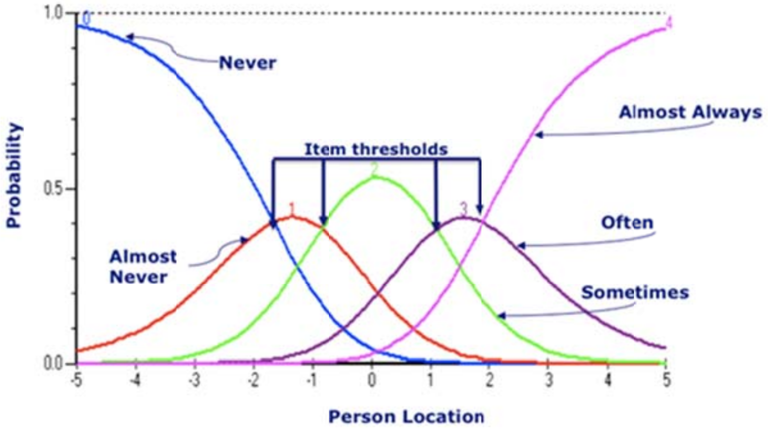
It is possible that respondents do not use the categories of a rating scale in the fashion intended by test developers (Kook & Varni, 2008). To achieve a better understanding of the information afforded by the overall score of a rating scale it is necessary to examine how respondents use response options.

A threshold is the point between two response categories in which the respondent is equally likely to endorse either response option (i.e., the probability of scoring a "0" or a "1" is fifty percent) (Tennant & Pallant). The number of thresholds for each item is equal to one less than the number of viable response options. Thresholds are considered disordered when respondents do not use the categories in a manner consistent with their level of the construct being measured. Disordering of thresholds is an indication that respondents are having difficulty discriminating between the various response options, perhaps because there are too many response options or the labeling of response options is confusing (Tennant & Conaghan, 2007).

In Rasch analysis thresholds are examined using the category probability curves of each item. It is anticipated that respondents with high levels of the construct being measured will endorse high scoring responses on each of the items measuring that construct. This pattern is reflected by an ordered set of response thresholds for each of the items.

In Figure 3c below, the hypothetical item displays ordered thresholds, as each response category (e.g., score of 0: "Never" (blue), score of 1: "Almost never" (red), score of 2: "Sometimes" (green), score of 3: "Often" (purple), and score of 4: "Almost always" (fuchsia)) systematically has a point along the ability continuum (x-axis) where it is the most likely response.

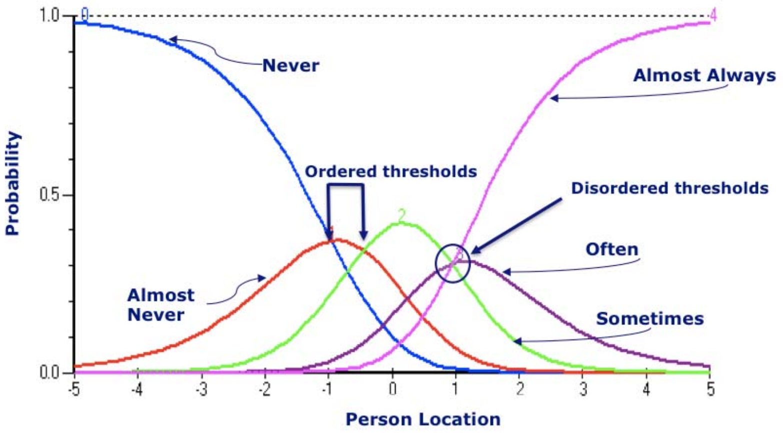
Figure 3c: Example item with ordered thresholds



#

In Figure 3d below, the hypothetical item has disordered thresholds. Responses for the fourth category are inconsistent with what was predicted by the Rasch model; consequently this category is never the most likely option to be endorsed at any point along the underlying the trait.

Figure 3d: Example item with disordered thresholds



The curves displayed above are derived from the item characteristic curves that are a key feature of modern psychometric methods. There are alternative mathematical procedures for examining threshold disordering in CTT, and these are based on the proportion of people who endorse an item. The unique advantage of modern psychometrics software is how easy it is to examine disordered thresholds graphically and subsequently to correct for problems. In CTT, because the main focus is on respondents' total scores, thresholds for each item are not considered significant and therefore are not examined as part of a standard psychometric evaluation.

Correcting disordered thresholds is a strategy to improve both item fit and overall fit to the Rasch model. The strategy involves collapsing adjacent categories for the disordered item and then re-testing overall fit of the data to the Rasch model to determine if there is an improvement. Item-trait interaction probabilities (based on overall χ^2 values) will likely increase after correcting disordered thresholds and this increase indicates less significant misfit from the Rasch model.

Differential Item Functioning

Differential item functioning (DIF) is a form of item bias that occurs when responses to items on rating scales that should otherwise be equivalent differ across groups within the sample (i.e., responses that differ across gender or race when the construct being measured has already been accounted for). It is not valid to interpret apparent group differences on a rating scale as true differences in the construct being measured unless items comprising the measure operate equivalently across the different groups (Teresi, 2006).

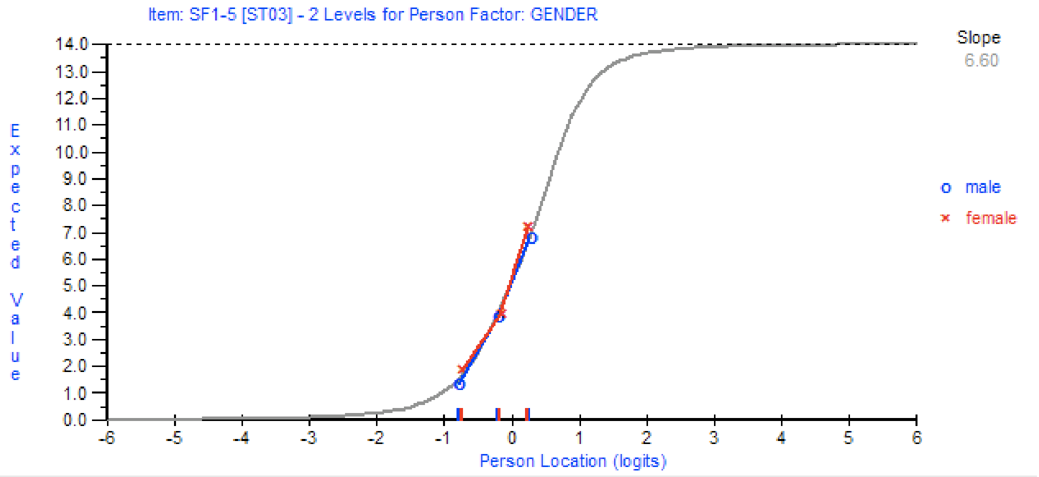
There are two types of DIF, uniform and non-uniform. If DIF is uniform, it is in the same direction across the entire spectrum of the construct being measured. Therefore at all levels of the construct, the likelihood of a specific response to an item is consistently higher or lower for a particular group. If DIF is non-uniform then at higher levels of the construct the likelihood of a specific response to an item may be higher in one sub-group while at lower levels of the construct the likelihood of a specific response to an item may be lower for that same sub-group (Teresi, 2006).

DIF can be assessed in both traditional and modern psychometric paradigms; however, because it is an item-level analysis, DIF is not a focus of the traditional paradigm. The methods used to detect DIF vary depending on which paradigm is used in its assessment. In both paradigms, the methods used to establish DIF involve a prediction of item response based on group membership that simultaneously controls for the underlying construct being measured.

The underlying construct is represented by the total score for a set of items on a scale. Respondents in different groups (e.g., males and females) with the same total score are compared to see it they differ in their responses to each item. If a group difference in responses to the item appears after controlling on an estimate of the underlying construct (the summed rating scale score) then that item is considered to manifest DIF. DIF detection methods differ in the traditional and modern paradigms based on the criteria used to flag it and whether non-uniform DIF can also be assessed (Teresi, 2006).

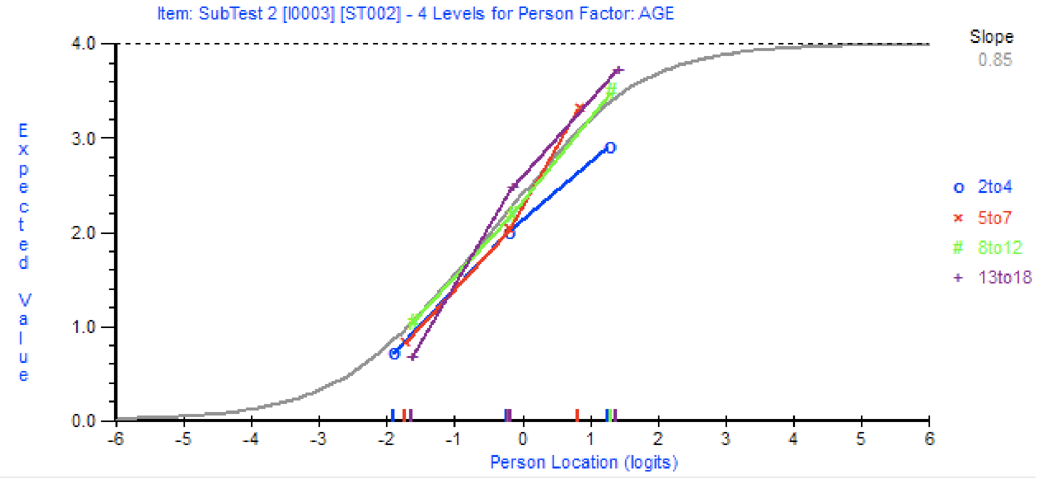
Parametric methods, such as those employed in Rasch analysis, assume the existence of a particular model and therefore examine DIF in terms of specific parameter estimates for that model. In order to evaluate DIF using Rasch all items must have ordered thresholds (or at least every attempt to order them must have been made). DIF is detected using the item characteristic curves. If the item curves for the groups in question (e.g., males and females) have the same difficulty and discrimination parameters the curves will coincide and DIF is not present as illustrated in Figure 3e. The blue curve represents males' responses to an item and the red curve represents females' responses to the same item.

Figure 3e: Item characteristic curves for an item that does not display DIF by gender



If the item curves for the groups in question (e.g. 2-4, 5-7, 8-13, 13-18 year olds) do not coincide (or at least display a significant amount of overlap) DIF is present as illustrated in Figure 3f below. The blue curve represents responses from 2-4 year olds, the red curve 5-7 year olds, the green curve 8-12 year olds and the purple curve 13-18 year olds.

Figure 3f: Item characteristic curves for an item that displays DIF by age

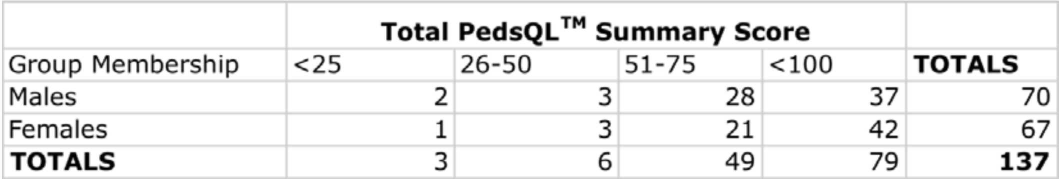


Group differences in discrimination parameters indicate the presence of non-uniform DIF, while differences in difficulty parameters indicate uniform DIF. ANOVA is conducted for each item to compare scores across each level of the person factor in question (e.g., gender) and across different levels of the trait (represented by the different class intervals). Uniform DIF is indicated by a significant main effect for the person factor in the ANOVA and therefore item curves will differ between these groups. Non-uniform DIF is indicated by a significant interaction effect for person factor and class interval. Modern psychometric methods are typically considered to be advantageous for DIF analysis because both uniform and non-uniform DIF can be easily detected by visual inspection of group differences in expected item curves as well as by examining the ANOVA statistics. Furthermore research has shown that parametric methods employed in modern psychometrics are more powerful in examining DIF (Teresi, 2006).

Uniform DIF can be corrected by splitting the data file by sub-groups of the person factor, creating what is known as a subtest, and then calibrating the item with DIF for each group separately (Tennant & Pallant). Non-uniform DIF typically cannot be corrected and it is often necessary to remove the item from the scale. A disadvantage of modern psychometric methods is that the assumptions of local independence and unidimensionality must be reasonably met (therefore implying that thresholds must also be ordered) prior to DIF calculation.

In the traditional paradigm data are not being fit to a pre-determined model, therefore non-parametric methods are employed to assess DIF. The most common method used to calculate DIF involves the use of a contingency table that examines the cross tabulation of item response by group membership for every level of the construct being measured (Teresi, 2006). Table 3 illustrates a hypothetical contingency table examining cross tabulation of item response by gender for different levels of QoL based on overall PedsQL^{TM} summary score.

Table 3: Example of a contingency table



	Total PedsQL^{TM} Summary Score				
Group Membership	<25	26-50	51-75	<100	**TOTALS**
Males	2	3	28	37	70
Females	1	3	21	42	67
TOTALS	3	6	49	79	**137**

The Mantel-Haenszel (M-H) statistic is used to determine if a group difference exists after controlling for the observed summed score of each item as an estimate of the underlying construct (Teresi, 2006). In this method, a common odds ratio, which tests whether the likelihood of item response is the same across groups, is used to establish the magnitude of DIF. The hypothesis is that the M-H statistic is the same for the two groups, when the underlying construct is controlled. DIF is present if there is a significant interaction of item by group (Teresi, 2006). Looking at Table 3 above, if the M-H statistic is the same for males and females who have the same overall summary score (as indicated by group membership) then DIF is not present.

Disadvantages of DIF detection in the traditional paradigm include the inability to detect non-uniform DIF and the non-user friendly procedures used in DIF assessment (complex and time-consuming) (Teresi, 2006). An advantage of using DIF detection methods in the traditional paradigm is that because methods are non-parametric, specific model assumptions do not need to be met prior to valid DIF assessment.

In the traditional paradigm, corrections for item bias are made at the test level (overall test score) instead of at the item level (Nunnally & Bernstein, 1994). The focus on overall test scores versus individual item scores reflects that importance of test-level statistics in the traditional paradigm. There are no formal procedures used to eliminate the cumulative impact of biased items on the overall test score in CTT. If on a test a particular sub-group, for example males, consistently has a higher overall mean score than females, a different scoring procedure will be implemented to account for the bias in overall test scores between genders (Nunnally & Bernstein, 1994; Pedhazur & Schmelkin, 1991).

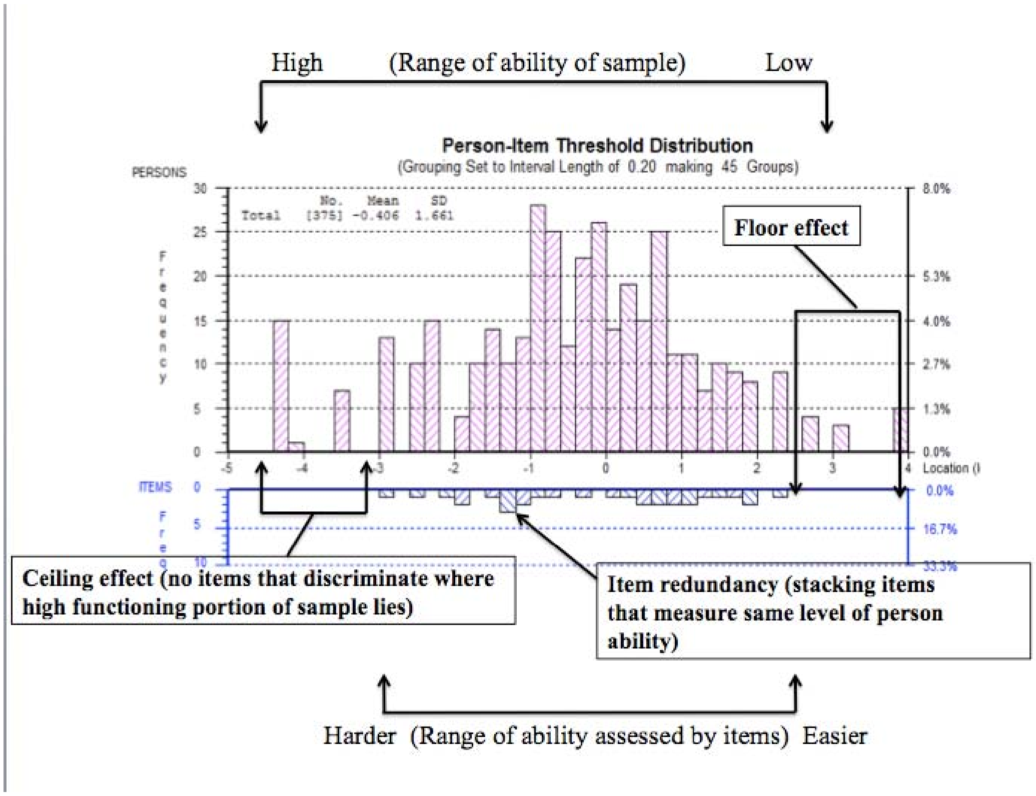
DIF can cancel out at the test level if some items favour one sub-group in the sample and other items favour other sub-groups, or if strategies to rid DIF such as item deletion or creating subtests are used. If DIF does cancel out at the test level it does not necessarily indicate that the items are functioning equivalently for different groups and it is still possible that an adverse impact could result for an individual if a decision is made on the basis of an item that has been shown consistently to have DIF. For this reason it may be important to evaluate item-level bias from a modern perspective as well as total score bias from a traditional perspective.

In modern psychometrics it is especially important to know exactly which items display DIF, as one of the main advantages of modern psychometrics is computer adaptive testing (CAT). Items that display DIF cannot be used in CAT item banks. It is also important to establish DIF with generic measures, such as the PedsQL^{TM} 4.0 Generic Core Scales. The advantage of using a generic measure is to enable comparisons across sub-groups, and ideally the items of a measure must function equivalently for all sub-groups (Teresi, 2006). DIF analysis may be particularly useful to consider when assessing parent-child agreement on a rating scale. If the items of the scale do not function equivalently for parent and child respondents it is not justified to make claims of parent-child agreement for the overall rating scale score, unless a correction factor that account for item bias is employed.

Targeting

It is important to ensure that measures are appropriately targeted at the population being assessed in order to minimize floor and ceiling effects. In CTT targeting is reflected by the percentage of respondents that lie above the ceiling and below the floor of the scale. In Rasch analysis, Rumm software produces a graphical display of how well the set of items spans across the range of person abilities in the sample being measured. The capacity to plot person ability on the same continuum as item difficulty facilitates the identification of persons that lie at the floor or ceiling of the scale as displayed in Figure 3g below.

Figure 3g: Person-item threshold map illustrating ceiling and floor effects



This plot of person ability and item difficulty provides a simple means of determining whether items cover the entire range of the construct and of identifying the items that are redundant and therefore do not increase the explanatory power of the scale (Hobart & Cano, 2009). If items do not cover the entire range of the construct, the plot serves as a visual indication of the level at which items should be added to make the scale a better discrimination tool. A poorly targeted scale would show misalignment or gaps with insufficient items to assess the entire range of the construct where respondents are scoring. A well-targeted scale would have items that assess at all levels of the construct where respondents are scoring.

Traditional psychometric methods do not generate estimates of item locations to understand their relative distances; rather, estimates of item locations are derived from item mean scores (Reise & Henson, 2003). It is argued that item mean scores are not true interval-level measurements as they depend on the distribution of the sample from which they were derived. This dependency may limit the appropriateness of making inferences about the ability of the items on the scale to produce meaningful scores within a CTT framework and of interpreting how well the items target the population in question (Hambleton & Jones, 1993; Reise & Henson, 2003).

In CTT, item redundancy is determined based on the correlations between pairs of items and overlapping content. It should be noted that the correlation between two items does not provide information about the location of the item on the trait continuum and therefore does not provide a means to examine whether items span the entire range of the construct (Reise & Henson, 2003). In traditional methods, item reduction is based on item-total correlations and this may decrease the sensitivity of a scale at the extreme ranges of the construct. If an item lies outside of the normal range (± 2 SD from the mean) it is typically discarded because too few respondents affirm these items (Tennant et al., 2004).

An advantage of using Rasch analysis in scale development is that items at the extreme range of the scale are not excluded; thus there is extended range of coverage of the construct in question. Proponents of Rasch argue that those individuals who lie at the extremes of the scale may be the most important to differentiate and thus items that discriminate at extreme ranges must be included in the measure. To counter this point, proponents of CTT assert that items that are included must discriminate where most of the respondents lie and thus items in extreme ranges are excluded from CTT-developed scales.

Reliability

Reliability is defined as the proportion of variability due to true differences between respondents (Streiner & Norman, 2008). It is an index of the ability of a measurement scale to discriminate consistently between subjects. The goal of reliability testing is not to reveal a difference between groups (as is the goal in most experimental studies), but rather to be able to rank order subjects consistently across variables such as condition, time or rater, depending on which of these factors adds variance to measurement (Streiner & Norman, 2008). The assumptions underlying traditional and modern psychometric theories have implications for how reliability is calculated, the factors that influence it and the indices used to represent it.

In traditional psychometrics, reliability is not a fixed property. If reliability of a scale is evaluated in a heterogeneous sample or if additional items are added to the scale the reliability will increase; conversely, if reliability is evaluated in a more homogenous sample or if items are deleted the reliability decreases. Therefore it is important to report the reliability of the scale in terms of the population for which it was established.

The predominant index used to describe item reliability is the internal consistency, commonly referred to as Cronbach's alpha (α). Scales with α values of 0.7 or greater can be used for group-level comparison whereas an α value of at least 0.9 is recommended for individual level decision-making (Nunnally & Bernstein, 1994; Pedhazur & Schmelkin, 1991). Reliability over specific factors that could potentially be sources of error, such as time or rater, are also taken into consideration in a traditional psychometric analysis; therefore, more than one type of reliability can be reported for a scale. Consideration of the various types of reliabilities (e.g. inter-rater, test re-test) highlights the importance of test-level performance versus item-level performance that is paramount in CTT.

In Rasch analysis, reliability is not dependent on the sample in which it was evaluated and is therefore constant regardless of the sample in which it was evaluated. Furthermore the addition or deletion of items to the scale does not impact the reliability co-efficient. The index analogous to Cronbach's alpha, used to represent reliability in a Rasch analysis, is the Person Separation Index (PSI).

The PSI provides an indication of the power of a measure to discriminate amongst respondents with different levels of the trait being measured. The PSI indicates the degree to which persons can be differentiated into certain groups and its value ranges from 0 to 1. A PSI value of 0.8 is considered acceptable, and represents the ability to differentiate statistically between at least three different ability groups (Schumacker & Smith, 2007). This means that there are three statistically different levels of person ability that can be distinguished by the items on the scale. A value of 0.9 or more would indicate the ability to discriminate between 4 or more groups (Tennant & Pallant).

Both Cronbach's alpha and PSI represent the proportion of total variance; however, the two indices differ in their construction. The PSI is based on estimated locations of person abilities and minimum and maximum scores are excluded in its calculation. The PSI extrapolates values for these extreme scores under the assumption that no additional information can be gained from their inclusion because there is no finite estimate for extreme scores (Schumacker & Smith, 2007). For example, if a scale measures from 0-10, all respondents who score 0 or 10 would be excluded from the sample when calculating the PSI.

Cronbach's alpha is calculated based on correlations between items and not on estimated locations of person abilities; therefore extreme scores are included in its calculation (Bland & Altman, 1997). Due to the omission of extreme scores in the construction of the PSI, when data have a skewed distribution PSI is more constant than the alpha. The error variance for person ability increases as scores become more extreme, therefore exclusion of extreme scores decreases the error variance in the construction of the PSI while there is no effect in the construction of alpha (Schumacker & Smith, 2007).

Cronbach's alpha and PSI are both better indicators of true reliability when the scale is well targeted such that the items and person abilities are well aligned. When there are differences in the two indices it is most likely due to floor or ceiling effects of data, presence of extreme scores and missing data. The PSI is a more useful index to evaluate reliability when the sample in question includes extreme respondents because these respondents are excluded in the calculation as it is assumed they are beyond the realm of being measured with the scale.

The PSI is also useful when there are random missing data in the sample because it can still be calculated despite missing data, whereas alpha is calculated using only those subjects that completed all items in a scale; anyone with missing data is excluded. In traditional psychometrics strategies such as replacing missing values with the sample mean are used to calculate an alpha value in samples with random missing data. Proponents of modern psychometrics argue that these strategies do not compensate accurately for missing data, and thus ways have been developed to overcome the need to have a complete data set in order to calculate reliability.

The presence of locally dependent items can artificially inflate the value of alpha in CTT whereas in Rasch analysis local dependence is taken into account when assessing reliability. Examining the residual correlation matrix for clustering identifies locally dependent items. Absence of a correlation pattern in the residuals would support the assumption of local independence and therefore suggest unidimensionality of the scale. Item residuals that have correlations greater than 0.2 are considered locally dependent and may be another cause of overall misfit to the Rasch model (Tennant & Pallant).

Using Rumm software it is possible to sum scores of dependent items to form subtests (which behave as new items) with a maximum score equal to the sum of the maximum score of the individual items. By creating a subtest it is possible to parcel out the portion of reliability that has been inflated due to local dependency and a more accurate representation of the reliability (taking into consideration artificial inflation of alpha due to locally dependent items) can be obtained for the PSI. For this reason, PSI estimates are typically lower than alpha estimates when using Rumm software (Tennant & Pallant).

The Rasch analysis process can be thought of as an iterative, systematic trial and error process. Each of the measurement properties discussed above must be re-examined as strategies such as correcting disordered thresholds or creating subtests are employed to increase fit to the Rasch model. If it is necessary to delete an item, this decision is often based on which items are most problematic on the greatest number of criteria. For example, an item that individually displays misfit to the model and displays DIF should be deleted before an item that displays only DIF. Other factors such as the clinical importance of the item and the extent to which the item misfits the Rasch model or displays DIF may also be taken into consideration.

Each time a strategy is explored, it is necessary to re-evaluate overall fit to the model and to assess how well the data meet Rasch assumptions prior to attempting another strategy. Thus the process can be cumbersome, time consuming and expensive, but it is useful in ensuring the greatest number of items is retained and the best Rasch solution is produced. Item deletion is reserved as a last resort to improve fit to the Rasch model when everything else has been done to salvage that item. It is not typically feasible to suggest item deletion as an option when conducting an analysis of an already established, widely used scale. However, when developing a new scale, items that display individual misfit to the Rasch model commonly are flagged for removal as a first step in item reduction.

Table 4 below summarizes the statistics discussed thus far and compares how they are reported in traditional and modern psychometric paradigms.

Table 4: Statistics reported in traditional and modern psychometric paradigms

Statistic	**Traditional Paradigm (CTT)**	**Modern Paradigm (Rasch)**
Item feasibility	% Missing items	% Missing items, DIF, Rasch misfit
Item scaling success	Item-total correlations < 0.7 and > 0.2	Examination of spread of item difficulty on person-item plot (must ensure continuum of construct covered)
Item difficulty	Item means, endorsement frequencies	Location of item on "ruler" of item difficulty (based on probability of endorsing item)
Reliability	Cronbach's alpha (> 0.70 acceptable)	Person separation index (> 0.8 acceptable)
Targeting	% Respondents above ceiling and below floor of scale (< 15 % acceptable)	Proportion of respondents above ceiling and below floor of scale as illustrated in person-item plot
Unidimensionality	Factor analysis	Principal component analysis of residuals (left over variance after Rasch factor taken into account)
Local Independence	Item-total correlations	Significance of t-tests between extreme positive and negative loadings on first factor after Rasch factor taken into consideration
Fit to Rasch model	N/A	Bonferroni adjusted Chi square (p-value should be >0.05 to indicate no significant misfit)
Item misfit	Item means (extreme values flagged); and SD (should not be too narrow)	Bonferroni adjusted Chi square for each item and item means and SD (mean/SD must lie between ± 2.5 units)
Person misfit	***	Person means (mean/SD must lie between ± 2.5 units)
Threshold Disordering	***	Category probability curves must be ordered

Differential Item Functioning	***	Significant group differences (determined by ANOVA) between item characteristic curves
Item Redundancy	Item-item correlations (<0.4 acceptable)	Items that stack (have the same difficulty) on the person-item plot

*** = not commonly assessed

Once the Rasch analysis is complete and the data essentially fit the Rasch model, it is possible to export the person estimates obtained as interval-level data into a statistical program for analysis of group and individual differences. This extra step of importing raw data from a CTT developed measure into Rumm software (to fit data to Rasch model and then to tranform raw data into interval-level data) prior to entering data into a statistical package for analysis of differences, serves as a major barrier for some people in the uptake of Rasch analysis in everyday health measurement.

Most health measures are developed using CTT, and raw scores (which are proponents of CTT believe very closely approximate interval level scores) are entered into a statistical package for analysis right away, saving time and resources. Proponents of Rasch analysis believe that if raw scores from CTT measures are going to be analyzed using parametric methods then they must first be fit to the Rasch model and transformed to interval level data; otherwise, it is only appropriate to analyze the data using non-parametric methods. Fan (1998) conducted a study to analyze differences in person estimates obtained using Rasch analysis converted interval level data, rather than raw scores, and did not find conclusive evidence that there are major differences in the overall results produced. Further evidence of the impact of converting ordinal raw scores to interval level data prior to statistical analysis is required to justify the additional time required to perform this transformation.

CHAPTER 5

RASCH ANALYSIS OF THE PEDSQL^{TM} 4.0 GENERIC CORE SCALES (PARENT-REPORT) IN A CHILDHOOD CANCER SAMPLE

Introduction

With advances in treatment approaches, over 80% of children diagnosed with cancer are living into adulthood. These new combinations of treatment place survivors at an increased risk of physical and psychosocial late effects associated with their therapies (Reis et al., 2007). The improvement in survival rates of childhood cancer patients has spiked great interest in the measurement of QoL during active treatment and in determining the predictors of sustained QoL as survivors enter adulthood (Eiser et al., 2003). Traditional outcome measures that focus on health from a purely biomedical perspective are no longer used in isolation, as it is now recognized that it is important to consider subjective well-being or QoL when assessing the impact of high intensity cancer treatment (Vance et al., 2001).

In a study conducted by Klassen et al. (2010) it was found that items and scales that comprise QoL measures in children with cancer vary substantially. The most common generic QoL tool used in the cancer population is the PedsQL^{TM} 4.0 Generic Core Scales (Klassen et al., 2010). The parent-report version of this scale offers an important perspective, especially for children undergoing active treatment, as they are typically too ill or fatigued to provide self-report. Furthermore, many childhood cancers are diagnosed under the age of five when self-report is not appropriate, and in these cases the parent-report becomes the preferred approach of measurement.

There are more psychometric data published for the parent-report version of the PedsQL^{TM} 4.0 Generic Core Scales than for any other QoL tool (Eiser, 2001; Eiser et al., 2003). An understanding is needed of the items and scales of this tool, from both traditional and modern psychometric paradigms, in order to assess how meaningful the summary score is as a measure of QoL.

This thesis presents the first modern psychometric analysis of the parent-report PedsQL^{TM} 4.0 Generic Core Scales in a sample of parents of children on active cancer treatment. The objective of the analysis is to compare and contrast traditional and modern item and scale-level statistics to determine if the theory used to analyze the measure influences these statistics. In addition, the purpose of the study is to determine if the use of Rasch analysis provides further information about the meaning of rating scale scores when used in conjunction with traditional methods, and to offer insight into how information gained from Rasch analysis can be applied to evaluate items and to aid in scale construction.

Two studies have previously examined psychometrics of the PedsQL^{TM} 4.0 Generic Core Scales using modern methods. The first study assessed the

Korean translation of the PedsQL^{TM} 4.0 for cross-cultural validity in a healthy school sample (Kook & Varni, 2008). In the second study the PedsQL^{TM} 4.0 was validated in a sample of preschool children with refractive errors living in Singapore (Lamoureux et al., 2010). A third study used the PedsQL^{TM} 4.0 to demonstrate some of the practical issues that arise when applying modern psychometrics to a measure (Hill et al., 2007). The purpose of this study was to determine which items from the PedsQL^{TM} 4.0 could potentially be considered for use in a separate study aiming to assemble a health-related QoL item bank for computer adaptive testing. The properties of candidate items were assessed to inform item assignment into domains for the item bank and not to evaluate the rigour of the PedsQL^{TM} 4.0 as a QoL tool.

Sample

PedsQL^{TM} 4.0 Generic Core Scales parent-report data were collected as part of a large multi-centre Canadian study of 411 parents of children and adolescents on active cancer treatment. The detailed methods for this study are published (Sung et al., 2008; Sung et al., 2009; Sung et al., 2010). Briefly, parents were recruited from five hospitals between November 2004 and February 2007. Parents were invited to complete a questionnaire booklet that included the PedsQL^{TM} 4.0 Generic Core Scales. Parents of children were included if the child was receiving treatment for any type of cancer, if they were initially diagnosed more than two months before enrolment on the study and if they were not considered palliative. The parent was eligible if they were the primary caregiver (i.e., the person most responsible for the day-to-day care and decision making for the child with cancer) and could read English.

Five hundred and thirteen eligible parents were invited to participate and 411 parents (81%) returned completed questionnaires. Of these parents 385 had children aged 2 to 17 years and had completed the PedsQL^{TM} 4.0 Generic Core Scales on behalf of their child. Findings from the studies in which the scores on the PedsQL^{TM} Generic Core Scales were the outcome of interest were published in two papers as follows: (1) a paper to identify the predictors of poor QoL in pediatric cancer patients receiving chemotherapy; and (2) a paper to describe QoL in children with acute lymphoblastic leukemia (ALL) in different phases of treatment (Sung et al., 2008; Sung et al., 2010).

Methods

For the purpose of this thesis, the PedsQL^{TM} 4.0 Generic Core Scales (parent-report) dataset was analyzed using Rumm2030 software (Andrich et al., 2010) and PASW Statistics (SPSS Inc., 2010). In a Rasch analysis respondents with extreme scores are dropped from the analysis as they do not provide explanatory power; thus the sample size is slightly different for each analysis presented.

Rasch analysis was carried out six times: once on each sub-scale of the PedsQL^{TM} 4.0 (physical function (PF), emotional function (EF), social function (SF) and school function (ScF)) and then on the summary scores (PF (the PF sub-scale), and Psychosocial (PS), which is the summed score of the EF, SF and ScF subscales, and finally on the total summary score (TS), which is the sum of the PF and PS scores. It should be noted that there was a high percentage of missing data (33%) for the ScF scale as many children with cancer typically do not attend school and therefore these items were not applicable. In Rasch analysis in order for sub-scale scores to be legitimately summed to form a summary score, the analysis must show that the summary score reflects a unidimensional scale measuring one higher order construct (Tennant & Conaghan, 2007). Psychometric analysis from a traditional perspective was performed for the PF, PS and TS summary scores. Traditional analysis was not performed for individual sub-scale scores (EF, SF, ScF) because Cronbach's alpha for these scales are reported using Rumm software as part of the Rasch analysis. The traditional analysis performed for the summary scores includes item-item and item-total correlations for each sub-scale.

Results

Rasch Analysis of Total Summary Score

Three hundred eighty-five records were entered into Rumm software for analysis. After exclusion of extreme scores 376 records remained and were divided into nine separate class intervals (automatic grouping of respondents based on ability level by Rumm software) for analysis.

Unidimensionality

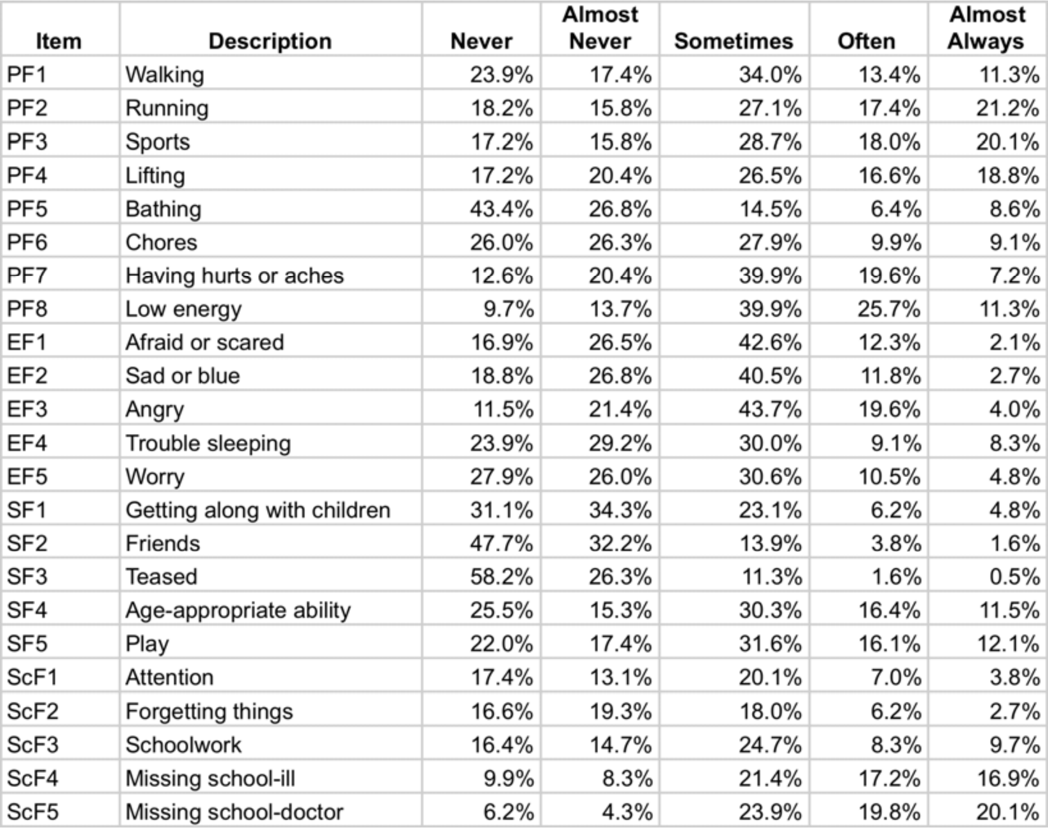
Overall the 23 item PedsQL^{TM} 4.0 scale did not fit the Rasch model (χ^2=381.0, df=184, p<0.001). Multidimensionality was apparent, as 21.72% of t-tests conducted between person estimates from the subset of items with the greatest loadings on the first factor displayed significance. In Rasch analysis person estimates derived from any subset of items on a unidimensional scale must be equivalent. If over 5% of person estimates display statistically significant differences from each other it is assumed there is multidimensionality in the scale.

Threshold Disordering

The PedsQL^{TM} Generic Core Scales (parent-report) asks parents to rank how much of a problem their child has had in four domains (PF, EF, SF and ScF) using a 5-point scale: Never (score of 0); Almost Never (score of 1); Sometimes (score of 2); Often (score of 3) and Almost Always (score of 4) (J. W. Varni et al., 1999). Category frequencies for each item indicate that most people respond using the "Never" "Almost Never" and "Sometimes" response categories. This response pattern is not surprising as previous psychometric studies of the PedsQL^{TM} 4.0

Generic Core Scales report a ceiling effect associated with the scale. The category frequency table (Table 5a) is displayed below.

Table 5a: Category frequencies for PedsQL^{TM} 4.0 Generic Core Scales



Item	Description	Never	Almost Never	Sometimes	Often	Almost Always
PF1	Walking	23.9%	17.4%	34.0%	13.4%	11.3%
PF2	Running	18.2%	15.8%	27.1%	17.4%	21.2%
PF3	Sports	17.2%	15.8%	28.7%	18.0%	20.1%
PF4	Lifting	17.2%	20.4%	26.5%	16.6%	18.8%
PF5	Bathing	43.4%	26.8%	14.5%	6.4%	8.6%
PF6	Chores	26.0%	26.3%	27.9%	9.9%	9.1%
PF7	Having hurts or aches	12.6%	20.4%	39.9%	19.6%	7.2%
PF8	Low energy	9.7%	13.7%	39.9%	25.7%	11.3%
EF1	Afraid or scared	16.9%	26.5%	42.6%	12.3%	2.1%
EF2	Sad or blue	18.8%	26.8%	40.5%	11.8%	2.7%
EF3	Angry	11.5%	21.4%	43.7%	19.6%	4.0%
EF4	Trouble sleeping	23.9%	29.2%	30.0%	9.1%	8.3%
EF5	Worry	27.9%	26.0%	30.6%	10.5%	4.8%
SF1	Getting along with children	31.1%	34.3%	23.1%	6.2%	4.8%
SF2	Friends	47.7%	32.2%	13.9%	3.8%	1.6%
SF3	Teased	58.2%	26.3%	11.3%	1.6%	0.5%
SF4	Age-appropriate ability	25.5%	15.3%	30.3%	16.4%	11.5%
SF5	Play	22.0%	17.4%	31.6%	16.1%	12.1%
ScF1	Attention	17.4%	13.1%	20.1%	7.0%	3.8%
ScF2	Forgetting things	16.6%	19.3%	18.0%	6.2%	2.7%
ScF3	Schoolwork	16.4%	14.7%	24.7%	8.3%	9.7%
ScF4	Missing school-ill	9.9%	8.3%	21.4%	17.2%	16.9%
ScF5	Missing school-doctor	6.2%	4.3%	23.9%	19.8%	20.1%

#

The threshold map displayed in Figure 4a below indicates that 14/23 items on the PedsQL^{TM} have disordered thresholds. Examination of individual category probability curves for disordered items indicates that respondents had difficulty discriminating between the categories "Never" and "Almost Never" (scores of 0 and 1) and "Often" and "Almost Always" (scores of 3 and 4) as displayed in Figure 4b. All disordered items were rescored as "00122" (i.e. collapsing original response categories "Never" and "Almost Never" into one category and collapsing "Often" and "Almost Always" into another category). Collapsing response categories in this fashion revised the original 5-point scale into a 3-point scale by reducing the number of response categories to three as displayed in Figure 4c. Reordering disordered items also improved overall fit to the Rasch model.

Figure 4a: Threshold map of 23 items on PedsQL^{TM} 4.0 Generic Core Scales

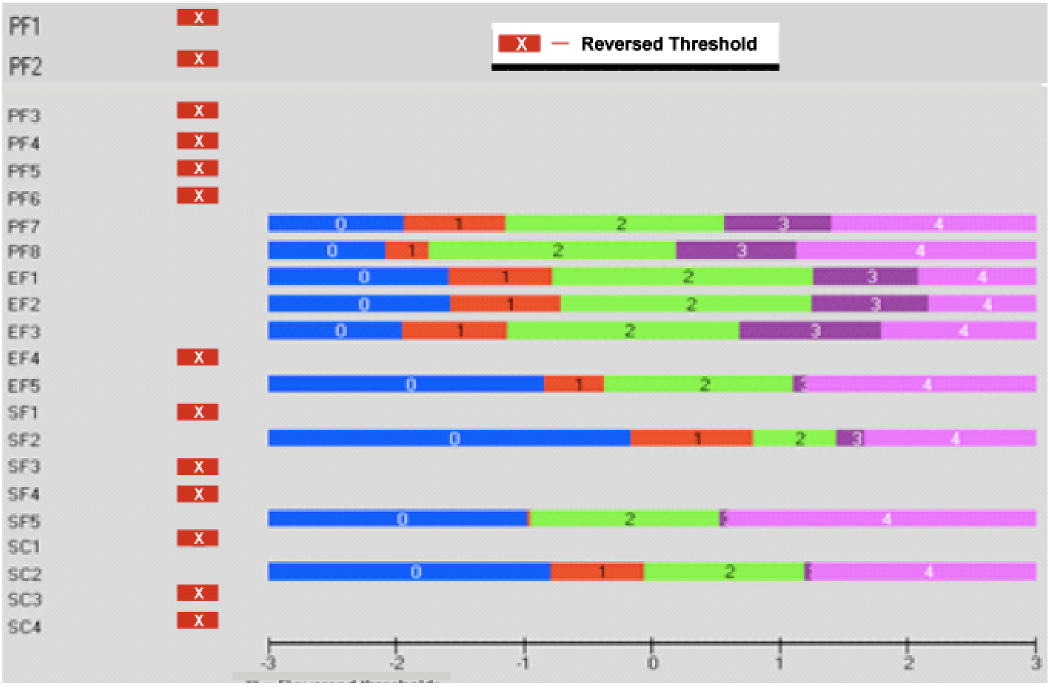


Figure 4b: Category probability curves showing disordered 5-point response options for 14/23 items on the PedsQL^{TM} 4.0

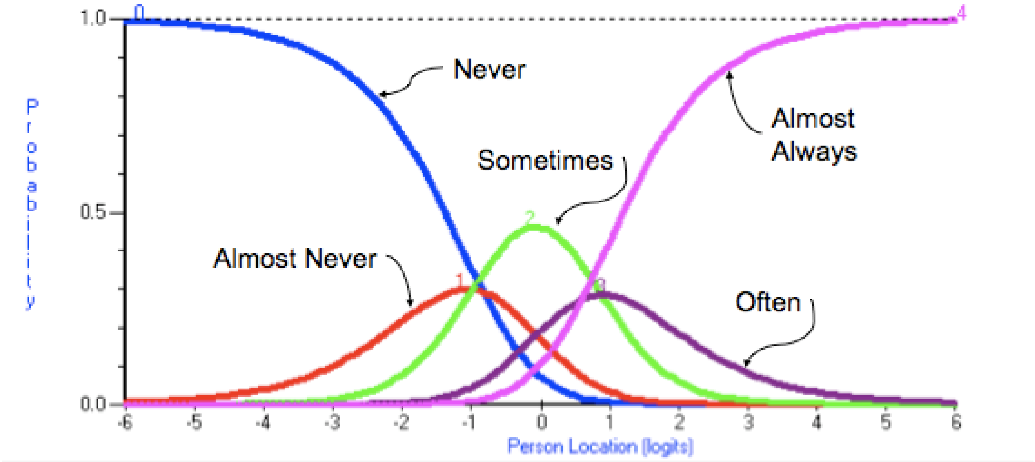


Figure 4c: Corrected category probability curves showing ordered 3-point response options for disordered items on the PedsQL^{TM} 4.0

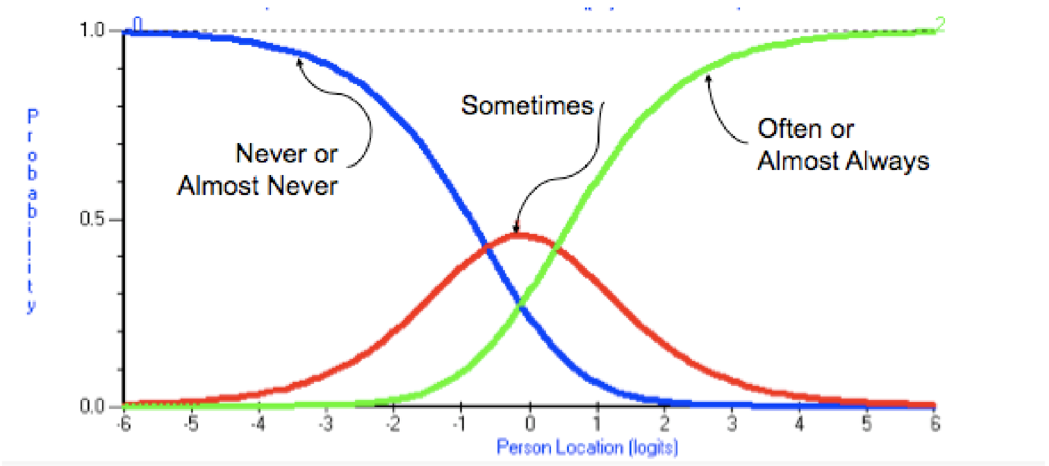
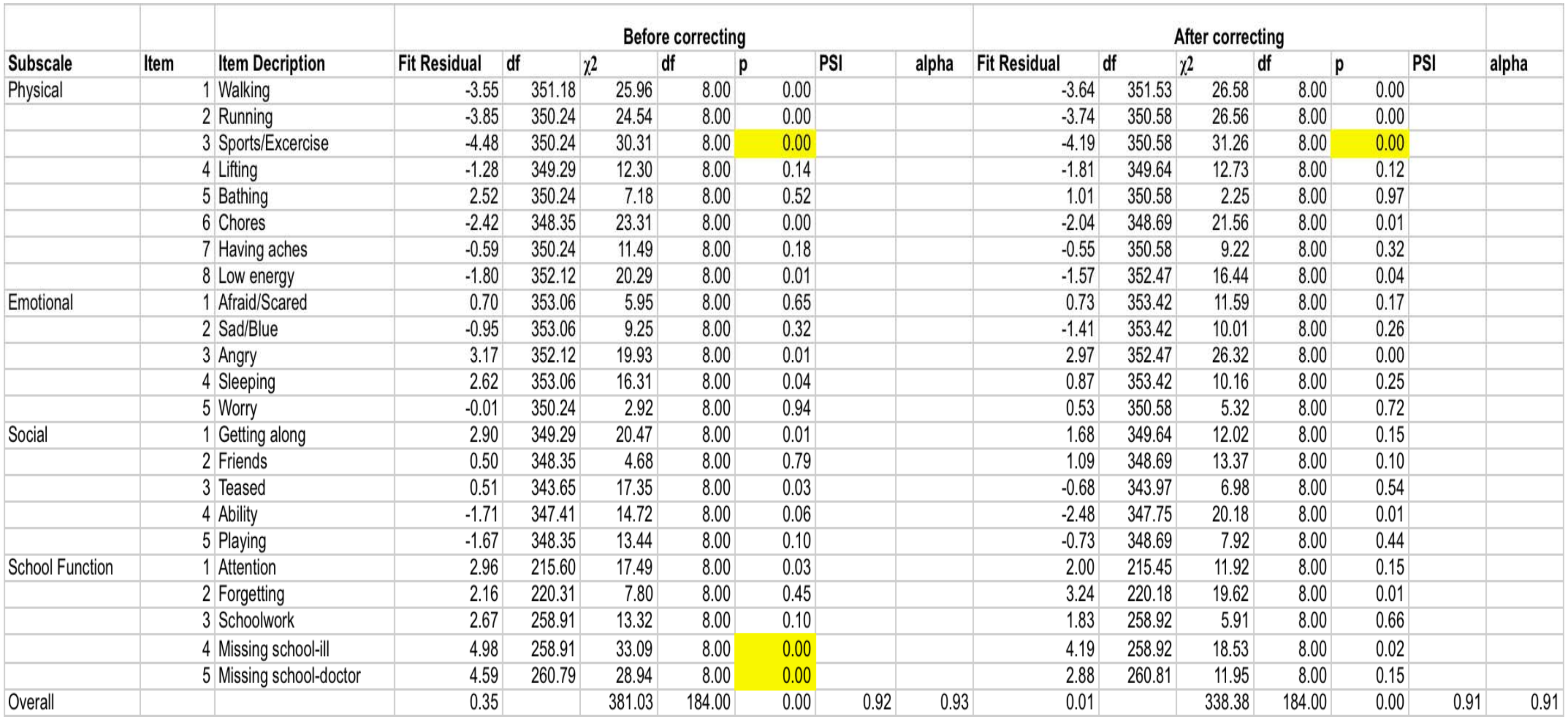


Table 5b below presents fit statistics for each item in the PedsQL^{TM} before and after ordering all thresholds. Item PF5 (Bathing) was the only item in the scale that remained disordered once thresholds were collapsed. Item PF3 (Participating in sports) was the only item that had significant misfit to the Rasch model when thresholds were reordered (χ^2=31.23, df=8, p<0.001). The corresponding p-value for this item is the only one highlighted yellow in the "After Correcting" column in Table 5b, flagging the item as being problematic.

Table 5b: Fit statistics for PedsQL^{TM} items before and after re-ordering thresholds



			Before correcting							After correcting						
Subscale	Item	Item Decription	Fit Residual	df	$\chi2$	df	p	PSI	alpha	Fit Residual	df	$\chi2$	df	p	PSI	alpha
Physical	1	Walking	-3.55	351.18	25.96	8.00	0.00			-3.64	351.53	26.58	8.00	0.00		
	2	Running	-3.85	350.24	24.54	8.00	0.00			-3.74	350.58	26.56	8.00	0.00		
	3	Sports/Excercise	-4.48	350.24	30.31	8.00	0.00			-4.19	350.58	31.26	8.00	0.00		
	4	Lifting	-1.28	349.29	12.30	8.00	0.14			-1.81	349.64	12.73	8.00	0.12		
	5	Bathing	2.52	350.24	7.18	8.00	0.52			1.01	350.58	2.25	8.00	0.97		
	6	Chores	-2.42	348.35	23.31	8.00	0.00			-2.04	348.69	21.56	8.00	0.01		
	7	Having aches	-0.59	350.24	11.49	8.00	0.18			-0.55	350.58	9.22	8.00	0.32		
	8	Low energy	-1.80	352.12	20.29	8.00	0.01			-1.57	352.47	16.44	8.00	0.04		
Emotional	1	Afraid/Scared	0.70	353.06	5.95	8.00	0.65			0.73	353.42	11.59	8.00	0.17		
	2	Sad/Blue	-0.95	353.06	9.25	8.00	0.32			-1.41	353.42	10.01	8.00	0.26		
	3	Angry	3.17	352.12	19.93	8.00	0.01			2.97	352.47	26.32	8.00	0.00		
	4	Sleeping	2.62	353.06	16.31	8.00	0.04			0.87	353.42	10.16	8.00	0.25		
	5	Worry	-0.01	350.24	2.92	8.00	0.94			0.53	350.58	5.32	8.00	0.72		
Social	1	Getting along	2.90	349.29	20.47	8.00	0.01			1.68	349.64	12.02	8.00	0.15		
	2	Friends	0.50	348.35	4.68	8.00	0.79			1.09	348.69	13.37	8.00	0.10		
	3	Teased	0.51	343.65	17.35	8.00	0.03			-0.68	343.97	6.98	8.00	0.54		
	4	Ability	-1.71	347.41	14.72	8.00	0.06			-2.48	347.75	20.18	8.00	0.01		
	5	Playing	-1.67	348.35	13.44	8.00	0.10			-0.73	348.69	7.92	8.00	0.44		
School Function	1	Attention	2.96	215.60	17.49	8.00	0.03			2.00	215.45	11.92	8.00	0.15		
	2	Forgetting	2.16	220.31	7.80	8.00	0.45			3.24	220.18	19.62	8.00	0.01		
	3	Schoolwork	2.67	258.91	13.32	8.00	0.10			1.83	258.92	5.91	8.00	0.66		
	4	Missing school-ill	4.98	258.91	33.09	8.00	0.00			4.19	258.92	18.53	8.00	0.02		
	5	Missing school-doctor	4.59	260.79	28.94	8.00	0.00			2.88	260.81	11.95	8.00	0.15		
Overall			0.35		381.03	184.00	0.00	0.92	0.93	0.01		338.38	184.00	0.00	0.91	0.91

#

Individual Item and Person Misfit

Individual item and person fit statistics are presented as residuals. The overall fit residual is the mean of individual person or item deviations from the Rasch model. Residuals between ± 2.5 logits indicate adequate fit to the Rasch model. Individually, item PF3 (Participating in sports) and items ScF 4 and 5 (Missing school-not feeling well and Missing school-doctor) displayed significant misfit from the Rasch model (χ^2=30.31, df=8, p<0.001; χ^2=33.09, df=8, p<0.001; χ^2=28.94, df=8, p<0.001). The corresponding p-values for these misfitting items are highlighted yellow in Table 5b. The overall item fit residual indicates some item redundancy (μ=0.347, σ=2.709). Individually, respondents did not display a high degree of misfit to the Rasch model. Highly misfitting respondents are generally considered for removal as their aberrant response patterns may skew the entire analysis. The overall person fit residual suggests the sample responded as would be predicted by the Rasch model (μ=-0347, σ=1.693).

Local Dependency

Overall fit to the Rasch model is continuously monitored as strategies are iteratively introduced to achieve unidimensionality and local independence. Thus, since the data did not display a good fit to the Rasch model after correcting disordered thresholds (χ^2=338, df=184, p<0.001), a second strategy to improve fit was attempted. This strategy involved correcting for locally dependent items. Examination of the residual correlation matrix revealed clustering of item residuals with correlations greater than 0.2 into four groups. This patterning of correlated residuals is in accordance with the expected scale structure of the PedsQL^{TM} in which items are divided into four sub-scales (PF, EF, SF and ScF). Items within each scale were sub-tested to account for local dependency.

After accounting for local dependency in the scale by creating sub-tests, unidimensionality was achieved (<5% of t-tests display significance at 0.05 level); overall there was no significant deviation from the Rasch model (χ^2=33.01, df=32, p=0.42).

When all possible attempts have been made to improve fit to the Rasch model (i.e., by correcting disordered thresholds, creating sub-tests of items to account for local dependency within scales and/or deleting individual items that have significant misfit) differential item functioning (DIF) analysis can be conducted. If items are split for a specific sub-group of the sample, as a strategy to account for DIF, it is no longer possible to assess dimensionality of the scale (Pallant & Tennant, 2007). Therefore it is required to have reasonably achieved assumptions of the Rasch model prior to conducting DIF analysis.

Differential Item Functioning

To ensure items have the same meaning for these sub-groups DIF (also referred to as item bias) must be assessed before comparing scores across two or more distinct groups within a sample (Teresi, 2006). In Rasch analysis, an ANOVA of the person-item deviation residuals is conducted to detect presence of DIF. The person attribute for which DIF is being assessed (e.g. age, gender) and the class intervals (groupings of respondents based on their abilities) are the factors used in the ANOVA. Uniform DIF is indicated by a significant main effect ($p<0.01$) for the person factor and non-uniform DIF is indicated by a significant interaction effect of the person factor x class interval ($p<0.01$).

In this study DIF was examined for the child's gender and age. Age was divided using the same categories Varni et al. (1999) used when developing separate parent-report forms for the PedsQL^{TM} (2-4, 5-7, 8-12 and 13-18 years). Rumm2030 software produces a graph to display DIF analysis. If the item curves coincide for various sub-groups of the factor in question (e.g., for males versus females when the factor in question is gender) then the item has the same difficulty parameter for sub-groups of the factor. No significant DIF was detected for gender and marginally significant uniform DIF was detected for age within the sub-test of items PF1-8 as displayed in Figures 4d and 4e below.

Figure 4d: No DIF by gender for sub-tested item PF1-8 (or for any other item)

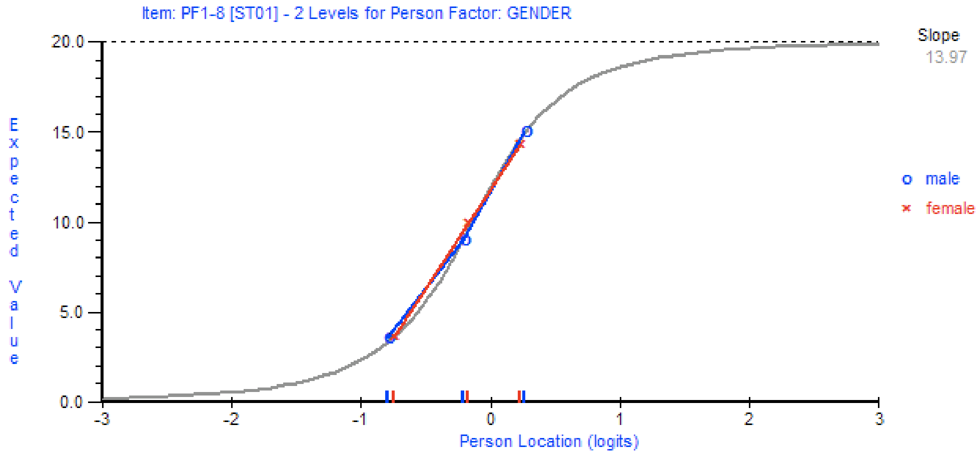
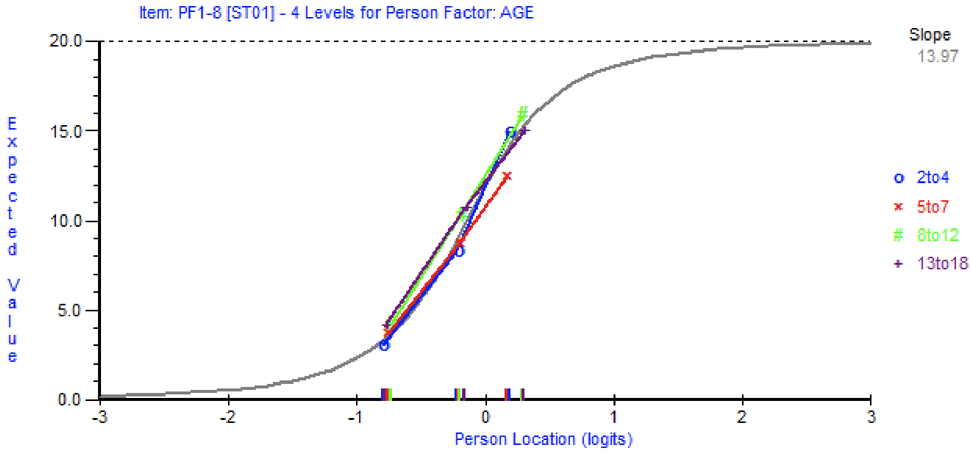


Figure 4e: DIF by age for sub-tested items PF1-8

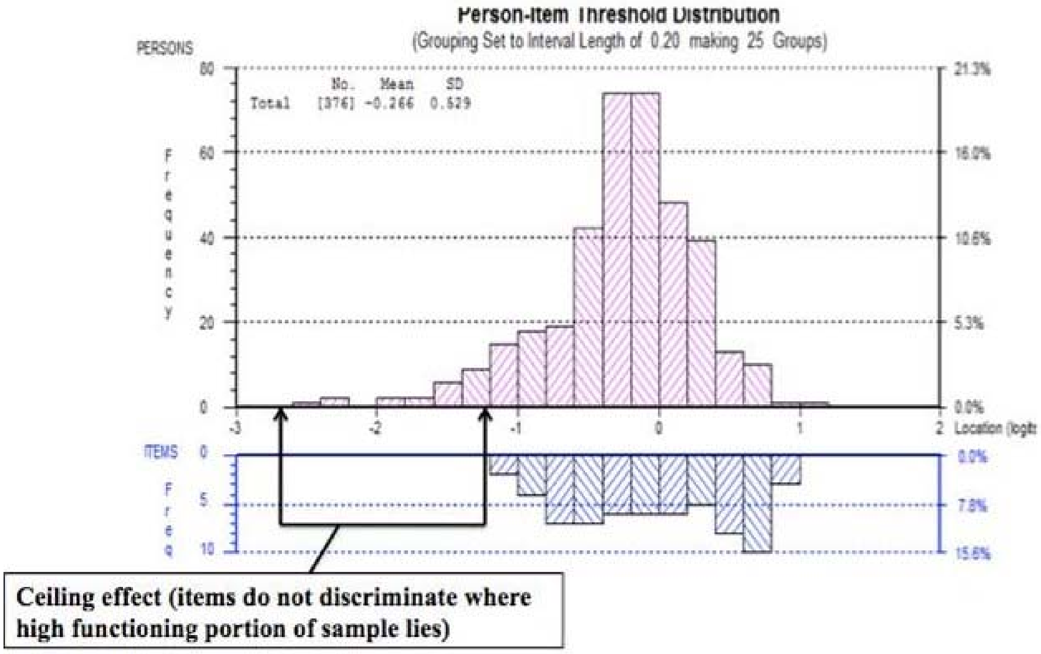


In Figure 4d the item characteristic curves for males and females coincide perfectly indicating the absence of DIF by gender. In Figure 4e the item characteristic curves for the different age groups are very similar but do not coincide perfectly indicating some DIF by age.

Targeting

Rumm2030 software produces person-item threshold distribution maps to assess targeting of a scale. See Figure 4f for person-item threshold map of the PedsQL^{TM}. The mean person ability has a negative value (-0.26), which indicates the sample used in this study has a better QoL than the average QoL that can be assessed by the items on the scale. A positive mean person ability would indicate that our sample had a lower QoL than the average QoL that can be assessed by the items on the scale (a lower raw score on the PedsQL^{TM} indicates higher QoL). A perfectly targeted scale would have a person mean ability value of zero. Examining the person-item threshold map below it appears that items that can discriminate at higher levels of QoL are needed to improve scale targeting, as there is a ceiling effect.

Figure 4f: Person-item threshold map of the PedsQL^{TM} 4.0 (parent-report)

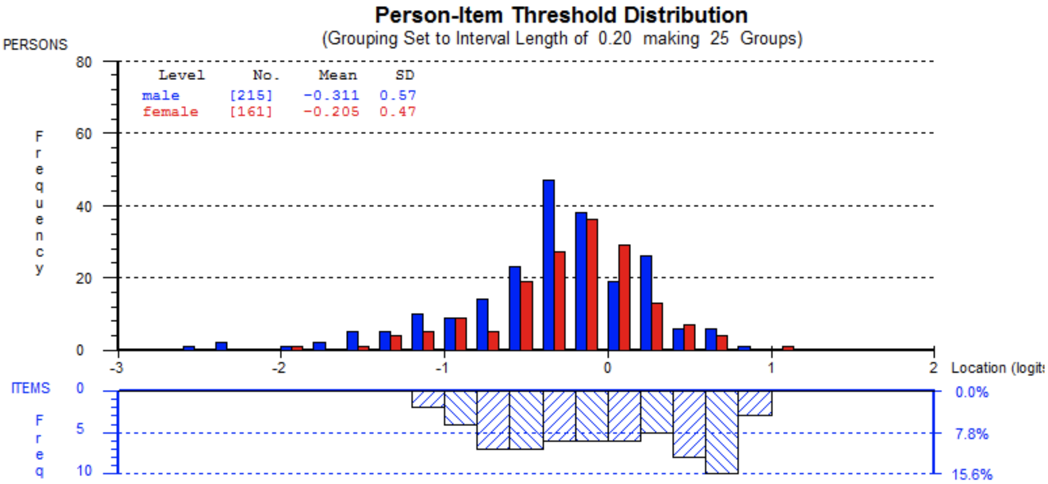


Ceiling effect (items do not discriminate where high functioning portion of sample lies)

#

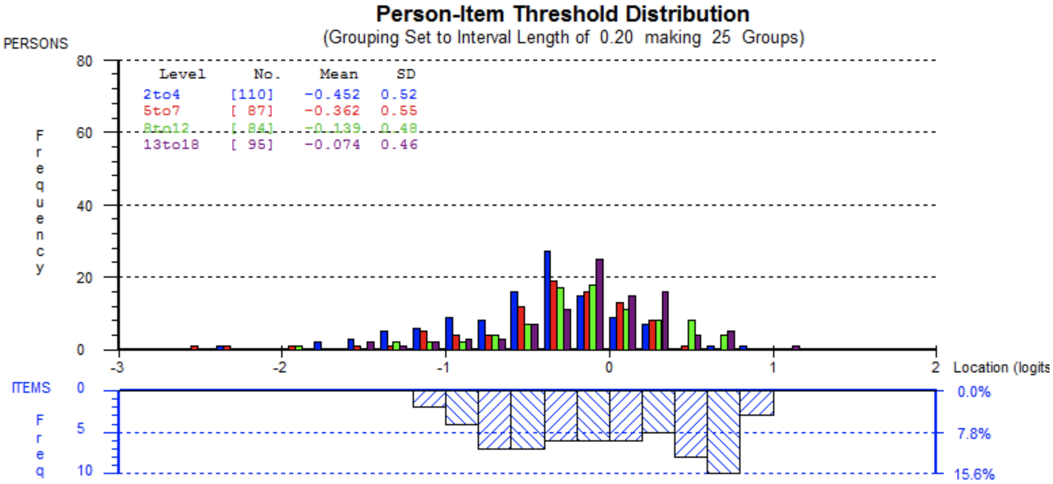
Rumm software can also produce a person-item threshold map that separately displays respondents belonging to various sub-groups, such as age or gender, in order to examine if there is a particular sex or age group that is poorly targeted by the scale. See Figure 4g for person-item map based on gender and Figure 4h for person-item map based on age. Differences in how well the items of a scale target a sub-group are only valid if there is no DIF apparent for that person factor. It is also possible to display individual item locations (for sub-tested items) in the person-item threshold map as displayed in Figure 4i.

Figure 4g: Person-item threshold map divided by gender



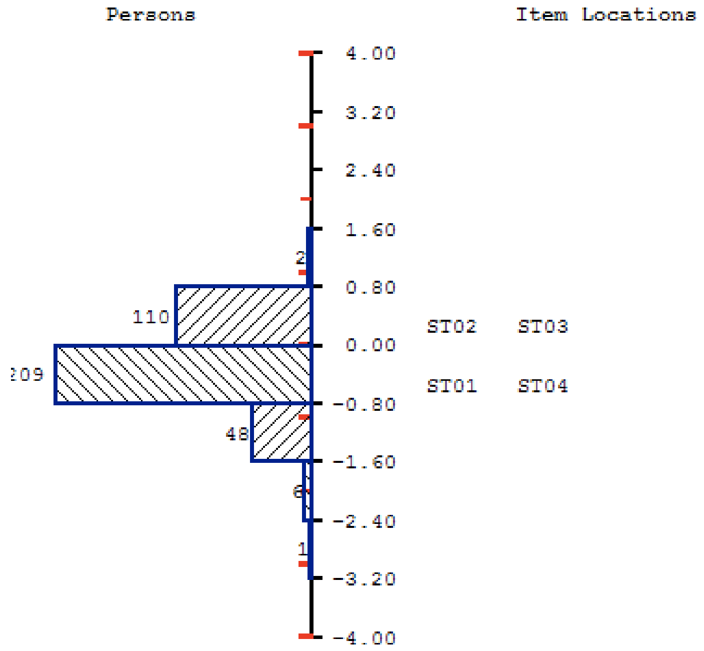
There is no DIF by gender so it is assumed that items are functioning the same for both males and females and person ability comparisons can be made. Examining the person-item map divided based on gender shows that more males lie above the ceiling of the scale than females. Thus there is a systematic difference in the way parents perceive and report QoL for their sons as compared to their daughters.

Figure 4h: Person-item threshold map divided by age



Marginal DIF was detected based on age, therefore items may not function equivalently for the different age groups. It appears that it is mostly the younger age categories (2-4 and 5-7) that lie above the ceiling of the scale.

Figure 4i: Item map displaying locations of sub-tested items



In Figure 4i above sub-tested items PF1-8 (ST01) and ScF1-5 (ST04) as well as sub-tested items EF1-5 (ST02) and SF1-5 (ST03) stack because they have the same level of item difficulty and therefore these items target the same level of person ability. EF1-5 (ST02) and SF1-5 (ST03) appear further along the x-axis therefore these items are considered to be more difficult than PF1-8 (ST01) and ScF1-5 (ST04).

Reliability

The Person Separation Index (PSI) of the total summary score decreases from 0.92 to 0.78 after correcting for local dependency and correcting disordered thresholds to improve fit to the Rasch model. Similarly, Cronbach's alpha (α) decreases from 0.93 to 0.80 when inflation due to local dependency is taken into consideration and disordered thresholds are corrected. See Table 5c for PSI and α values of PF, PS and TS summary scores as well as for the individual sub-scale scores before and after fit to the Rasch model.

Table 5c: PSI and Cronbach's α before and after fit to the Rasch model

		Before Fit to Rasch Model		After Fit to Rasch Model	
Score	**N**	**PSI**	**α**	**PSI**	**α**
Total Summary Score	376	0.91	0.93	0.78	0.80
Physical Function Summary Score	354	0.90	0.92	0.90	0.92
Psychosocial Summary Score	369	0.83	0.88	0.64	0.70
Emotional Function Scale Score	357	0.79	0.81	0.79	0.82
Social Function Scale Score	326	0.74	0.80	0.72	0.79
School Function Scale Score	251	0.77	0.79	0.73	0.70

Rasch Analysis of Physical Function Summary Score

Three hundred eighty-five records were entered in the analysis for the Physical Function scale. After exclusion of extreme scores 354 records remained and were divided by the software into nine separate class intervals for analysis. #

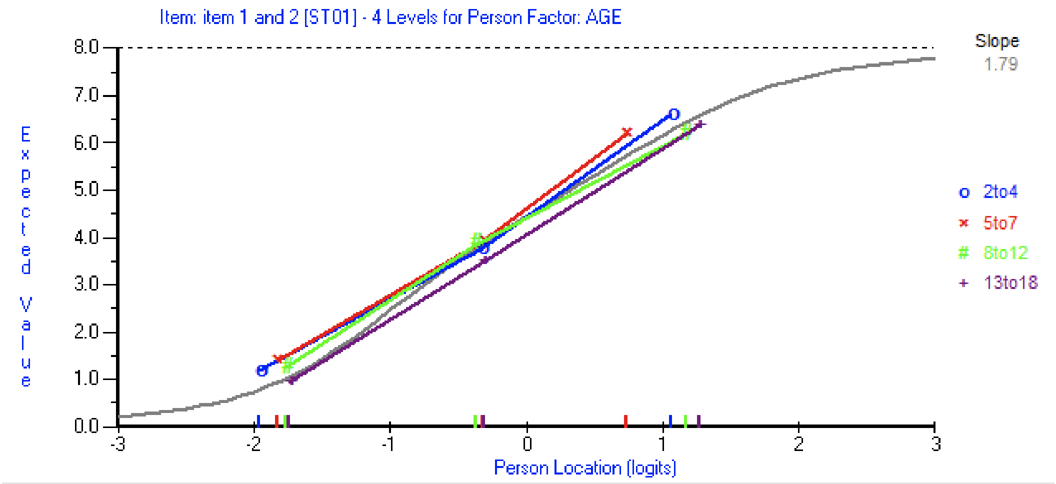
All items in the PF sub-scale had ordered thresholds except PF5 (Bathing). Response categories "Never" and "Almost Never" were collapsed and scored as "0" and response categories "Sometimes" and "Often" were collapsed and scored as "1" for this item. When the analysis was re-run all items in the scale had ordered thresholds. All individual items in the PF sub-scale displayed fit to the Rasch model; however, the overall item fit statistic displayed significant misfit to the Rasch model (χ^2=110, df=64, p<0.001). Using the binomial theorem less than 5% of t-tests conducted between subsets of items on the second factor were significant at the 0.05 level and therefore it is assumed the scale is unidimensional.

To improve fit to the Rasch model, the residual correlation matrix was examined for highly correlated item residuals. PF1 (Walking) and PF2 (running) were highly correlated (0.35) and therefore these items were sub-tested. Accounting for this local dependency, fit to the Rasch model was achieved (χ^2=73 df=56, p>0.05).

Examining the PSI and α values before and after achieving fit to the Rasch model indicates that the reliability of the PF summary score was not inflated by local dependency. It can be concluded that the PF summary score is a reliable estimate of physical function in this sample.

DIF was examined based on gender and age. The initial number of class intervals, which are automatic groupings of respondents based on their ability levels, was set at 9 by Rumm software. The number of class intervals was reduced to 6 for the analysis by gender and 3 for the analysis by age so that there would be at least 20 respondents in each ability group to conduct an ANOVA with sufficient power (Tennant & Pallant,). The sub-test for the walking and running items displayed marginal uniform DIF as displayed in Figure 5a below. There was no DIF based on gender.

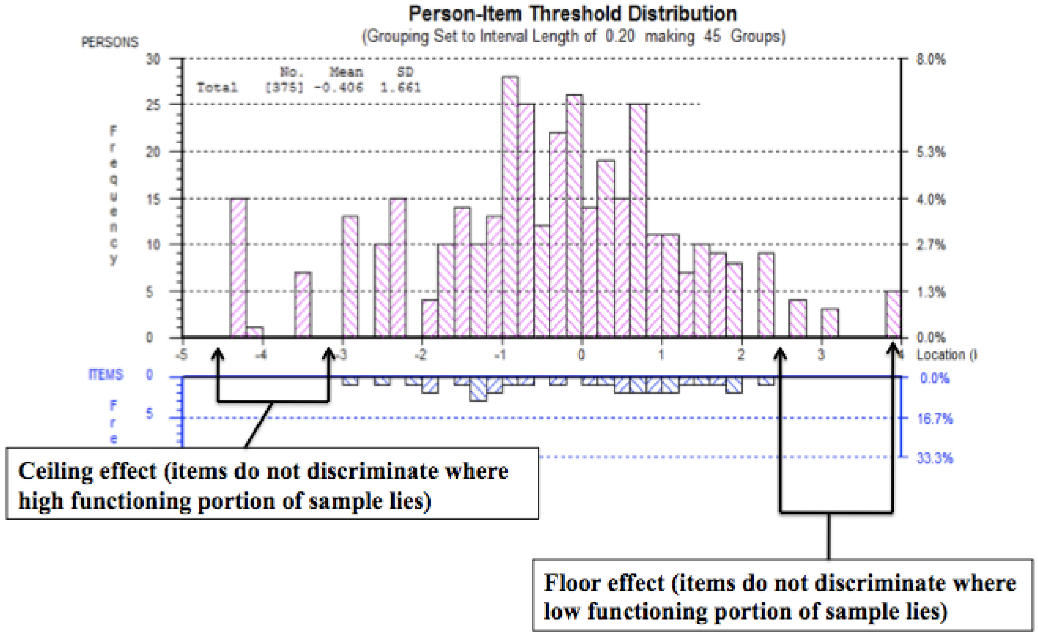
Figure 5a: DIF by age for sub-tested items PF1 (Walking) & PF2 (Running)



In Figure 5a above the item characteristic curves for the various age groups closely overlap indicating that the item difficulty parameter for the sub-tested PF1 and PF2 item are very similar for each age group and that DIF is marginal.

The person-item threshold map in Figure 5b below indicates that there are some people in the sample who lie above the ceiling and below the floor of scale. Therefore the PF scale is poorly targeting the sample measured. The person ability mean is negative (-0.41) indicating that in general the sample assessed in this study has better physical function than that measured by items of the scale.

Figure 5b: Person-item threshold map for physical function summary score



Ceiling effect (items do not discriminate where high functioning portion of sample lies)

Floor effect (items do not discriminate where low functioning portion of sample lies)

Rasch Analysis of Psychosocial Summary Score

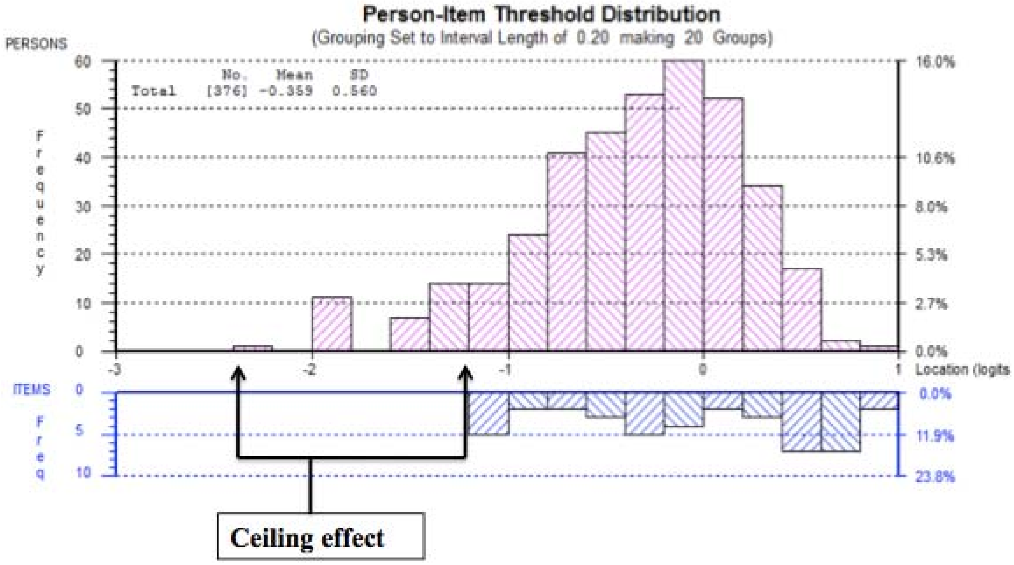
Three hundred eighty-five records were included in the analysis for the Psychosocial (PS) scale, which includes 15 items from the three scales measuring emotional (EF), social (SF) and school function (ScF). After exclusion of extreme scores 369 records remained and were divided by the software into nine separate class intervals for analysis.

Nine items in the PS summary score had disordered thresholds including EF4 (Sleeping), SF1 (Getting along), SF3 (Teased), SF4 (Ability), SF5 (Keeping up with other children), ScF1 (Paying attention), ScF3 (Keeping up with school work), ScF4 (Missing school-ill) and ScF5 (Missing school-doctor). Response categories "Never" and "Almost Never" were collapsed as well as response categories "Sometimes" and "Often" as respondents could not discriminate between these categories. After rescoring, all items in the PS summary score displayed ordered thresholds. All items individually fit the Rasch model ($p > 0.05$). The overall item fit residual displayed significant misfit to the Rasch model (χ^2=174, df=120, $p<0.001$). Multidimensionality was apparent as >5% of t-tests conducted between subsets of item residuals on the second factor displayed significance at the 0.05 level.

Examination of the residual correlation matrix indicates clustering of item residuals into three groups. Items were combined into three domains that correspond to the EF, SF and ScF sub-scales to account for local dependency. This correction achieved overall fit to the Rasch model (χ^2=30, df=24, $p>0.05$) and fulfilled the assumption of unidimensionality (<5% of t-tests displayed significance). The PSI decreased from 0.83 to 0.64 after accounting for local dependency within sub-scales of the PS summary score. Similarly, the α value decreased from 0.88 to 0.70 after accounting for local dependency within sub-scales of the PS summary score. There was no significant DIF detected based on gender or age.

There is a ceiling effect but no floor effect associated with the PS summary score as displayed in the person-item threshold map displayed in Figure 6 below. The person ability mean is -0.36 indicating that the sample in this study has higher psychosocial function than that which can be measured by the items on the scale.

Figure 6: Person-item threshold map for psychosocial summary score



Ceiling effect

Rasch Analysis of Emotional Function Sub-Scale

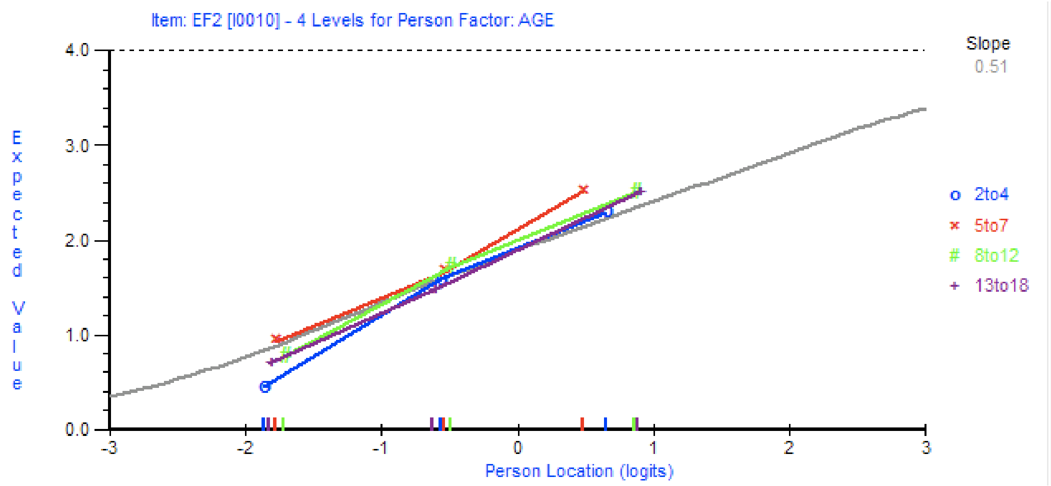
Three hundred eighty-five records were included in the analysis for the Emotional Function sub-scale. After exclusion of extreme scores 357 records remained and were divided by the software into nine separate class intervals for analysis.

All item thresholds were ordered except for EF4 (Sleeping). Response categories "Often" and "Almost Always" were collapsed and scored as "3" to correct threshold ordering. Individual items display fit to the Rasch model ($p>0.05$) except for item EF2 (Feeling sad) ($\chi^2=19.40$, df=2, $p<0.001$). Overall fit to the Rasch model was not achieved ($\chi^2=64.07$, df=40, $p < 0.001$). The EF scale displayed unidimensionality with <5% of t-tests significant at the 0.05 level with 95% confidence using binomial theory. No further strategies were employed to improve fit to the Rasch model as unidimensionality was achieved and the residual correlation matrix did not reveal any highly correlated items.

There is no significant DIF based on gender. EF10 (Sad), EF9 (Afraid), and EF13 (Worry) displayed significant uniform DIF for age as displayed in Figures 7a, 7b and 7c, respectively. The number of class intervals was reduced from nine to three in order to conduct ANOVA with sufficient power (Tennant & Pallant).

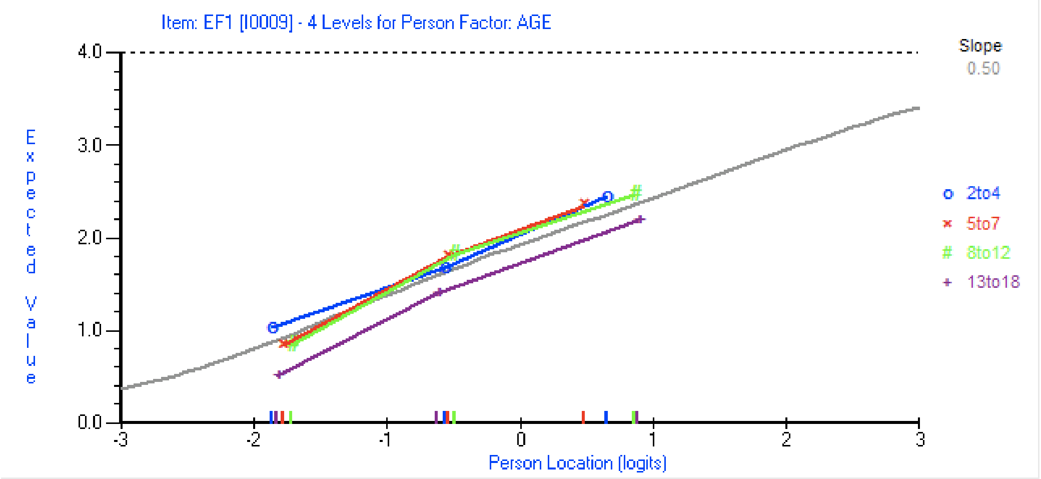
Graphically DIF for the item "Feeling Sad" (Figure 7a) appears marginal. The item curves for each age group closely overlap indicating there is not a high degree of bias based on age for this item. DIF for items "Feeling Afraid" (Figure 7b) and "Worry" (Figure 7c) appear to have more prominent item bias because there is not as much overlap between the item curves for each age group.

Figure 7a: Plot of DIF by age for EF10 (Feeling sad or blue)



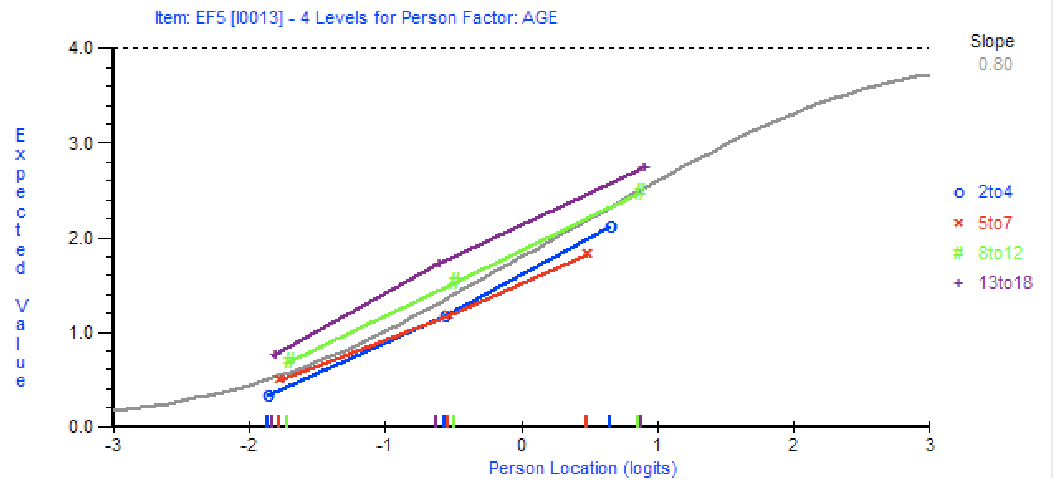
In Figure 7a above there is a high degree of overlap between the item curves for each age group therefore the item EF10 (Feeling sad) is marginally biased based on age.

Figure 7b: Plot of DIF by age for EF9 (Feeling afraid)



In Figure 7b above the item curve for ages 13-18 years (purple line) is beneath the item curves for the other age groups indicating that this item EF9 (Feeling afraid) is an easier item for this age group to endorse. Any significant differences found when comparing results between two groups of respondents would be biased by age for this item and therefore accurate comparisons of scores based on this item cannot be made.

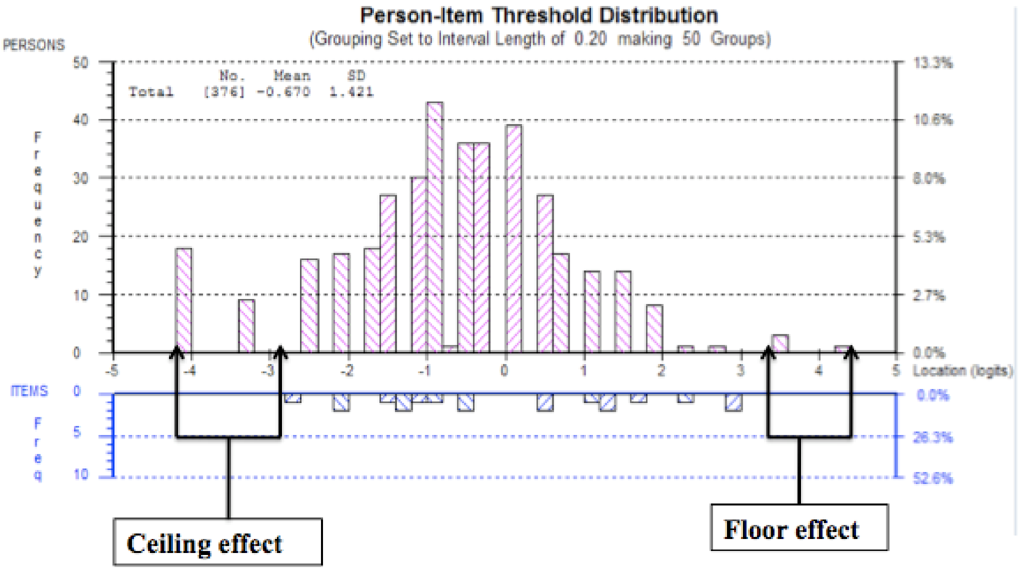
Figure 7c: Plot of DIF by age for EF13 (Worrying about what will happen to him or her)



In Figure 7c above the item curve for 13-18 year olds is significantly higher than the other item curves indicating the item EF13 (Worry) is harder for this age group to endorse. Therefore accurate comparison based on the results of this item cannot be made because they will be biased based on respondents' age.

There are ceiling and floor effects associated with the scale as displayed in Figure 7d. The person ability mean is -0.67 indicating that overall the sample in this study has greater emotional function than that which can be measured by the items on the EF sub-scale.

Figure 7d: Person-item threshold map for emotional function sub-scale



```
         No.   Mean    SD
Total  [376] -0.670  1.421
```

Ceiling effect

Floor effect

The PSI and α values for the EF sub-scale are similar before and after fitting data to the Rasch model. The PSI value remained 0.79 before and after fit to Rasch model and the α value increased from 0.81 to 0.82 after fit to the Rasch model.

Rasch Analysis of Social Function Sub-Scale

Three hundred eighty-five records were included in the Social Function (SF) sub-scale analysis. After exclusion of extreme scores 326 records remained and were divided by the software into nine separate class intervals for analysis.

Overall, fit to the Rasch model was achieved before correcting disordered thresholds (χ^2=41.78, df=40, p>0.05). SF1 (Getting along) and SF4 (Ability) displayed disordered thresholds. Categories "Often" and "Almost Always" were collapsed for SF1 (Getting along) and categories "Never" and "Almost Never" as well as "Often" and "Almost Always" were collapsed for SF4 (Ability) to correct disordering. Fit to the Rasch model was maintained (χ^2=47.67, df=40, p>0.05). Individually, all items displayed fit to the Rasch model before and after correcting for disordered thresholds. The SF sub-scale appears unidimensional with < 5% of t-tests conducted displaying significance at the 0.05 level.

Items SF1 (Getting along) and SF4 (Ability) displayed significant uniform DIF based on age as displayed in Figure 8a and 8b below. No significant DIF was found based on gender. Class intervals were reduced from nine to three to permit sufficient numbers of respondents per group for ANOVA.

Figure 8a: Plot of DIF by age for SF1 (Getting along with other children)

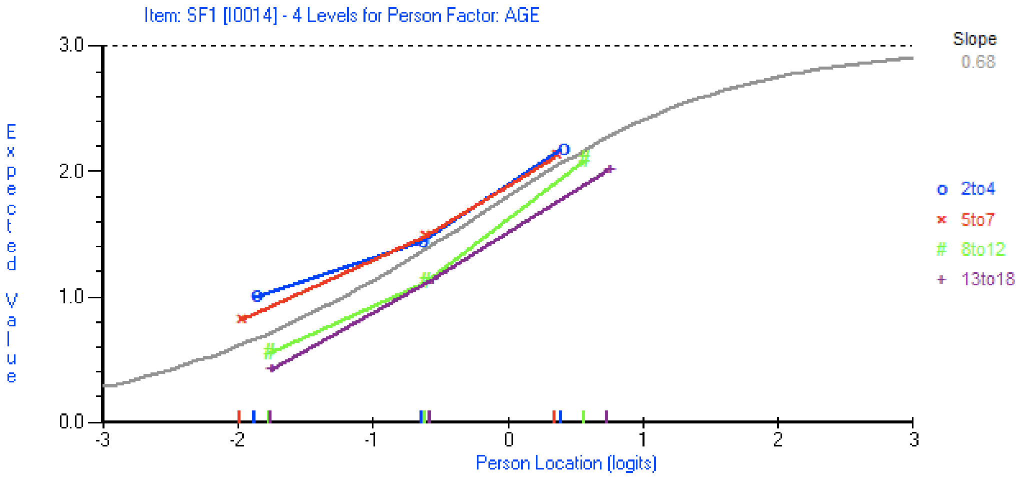
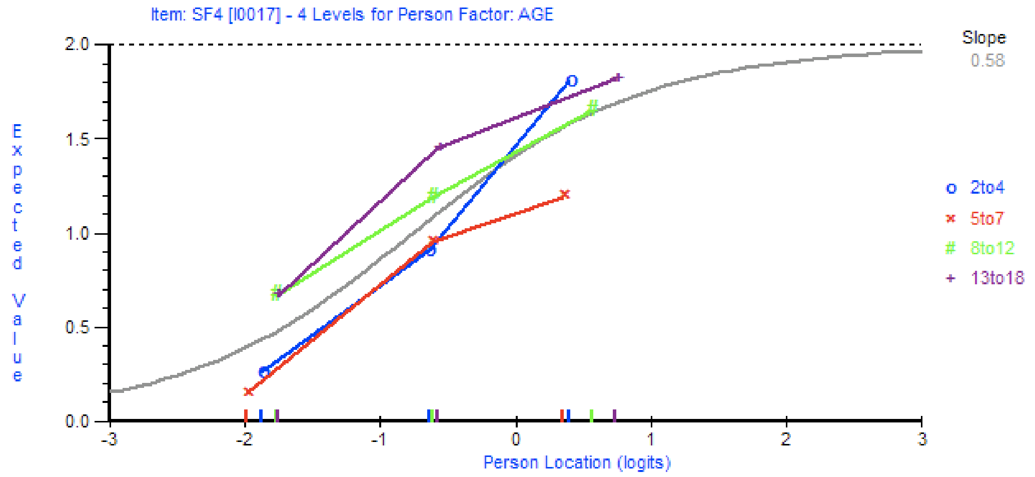


Figure 8b: Plot of DIF by age for SF4 (Not being able to do things other children his/her age can do)

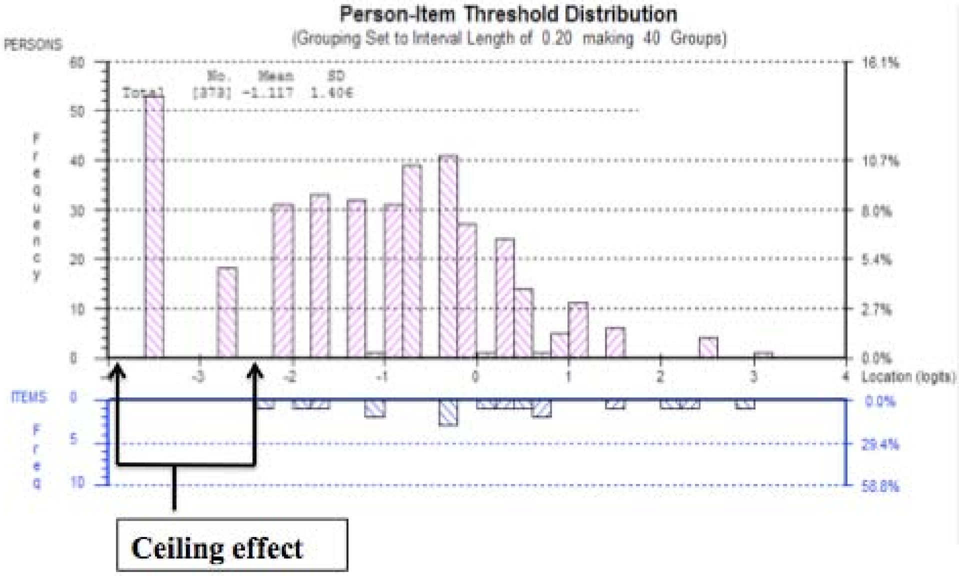


Examining Figures 8a and 8b above it appears that the items "Getting along" and "Ability" function differently for parents of children in the 2-4 and 5-7 age groups as compared to parents of children aged 8-12 and 13-18 years. In Figure 8a the item curves for 2-4 (blue) and 5-7 (red) year olds appear higher than the item curves for 8-12 (green) and 13-18 (purple) year olds. This difference in the item curves indicates that the item SF1 (Getting along with other children) is harder for younger age groups (2-7) and easier for older age groups (8-18). Therefore comparing responses between parents of children in the different age groups is not valid for this item because the item does not function equivalently for various age groups.

In Figure 8b item SF4 (Ability) appears to be harder for older age groups (8-18 years) than for younger age groups (2-7 years). The item curves for 8-12 and 13-18 year olds (green and purple curves) are significantly higher than for 2-4 and 5-7 year olds (red and blue curves). The lack of overlap in the item curves indicates valid comparisons cannot be made because the item is biased based on age. Using different items to assess "Getting along" and "Ability" for children aged 2-4 and 5-7 and children aged 8-12 and 13-18 may be appropriate based on the significant DIF associated with these items in the current SF scale.

There is a ceiling effect associated with the SF sub-scale and a large proportion of respondents lie above the ceiling of the scale as displayed in Figure 8c below. The overall person ability mean is -1.12 indicating that the sample measured has better social function than that which can be assessed by the items on the SF scale.

Figure 8c: Person-item threshold map for social function sub-scale



The PSI and α values for the SF sub-scale are similar before and after fitting data to the Rasch model (the PSI is 0.74 and drops to 0.72; the α value is 0.80 and drops to 0.79). Fit to the Rasch model and unidimensionality was achieved prior to correcting for disordered thresholds so it is expected that the PSI and α values for this sub-scale would remain similar.

Rasch Analysis of School Function Sub-Scale

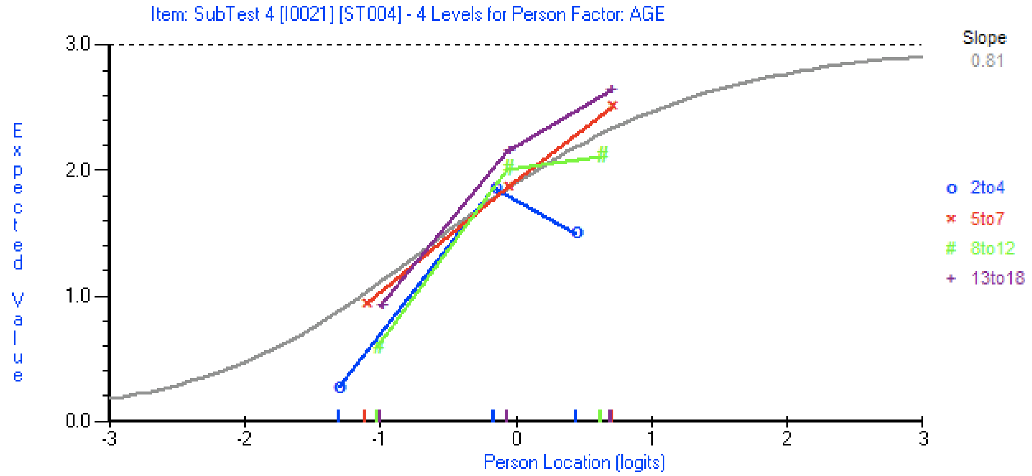
Three hundred eighty-five records were included in the School Function (ScF) sub-scale analysis. After exclusion of extreme scores and participants who had missing data for the entire scale, 251 records remained and were divided by the software into nine separate class intervals for analysis.

Overall items fit the Rasch model (χ^2=32.1, df=40, p>0.05). Individually all items also fit the Rasch model (p>0.05). Item ScF3 (Teased) was the only disordered item in the ScF sub-scale. Response categories "Often" and "Almost always" were collapsed and scored as "3" to correct ordering of thresholds.

The ScF sub-scale displays multidimensionality despite overall fit to the Rasch model (>5% of t-tests display significance at 0.05 level). The residual correlation matrix was examined to determine if there were highly correlated items that could be sub-tested to get rid of local dependency. ScF22 (Missing school-ill) and ScF23 (Missing school-doctor) were correlated above the acceptable level; their correlation was 0.37. These items were sub-tested to account for their local dependency and unidimensionality of the scale was re-assessed using the t-test procedure. Less than 5% of all t-tests displayed significance at the 0.05 level after correcting for local dependency, thus the ScF sub-scale was assumed to be unidimensional.

Item ScF21 (Keeping up with school work) displayed significant uniform DIF based on age as displayed in Figure 9a below. There was no significant DIF detected in the ScF sub-scale based on gender.

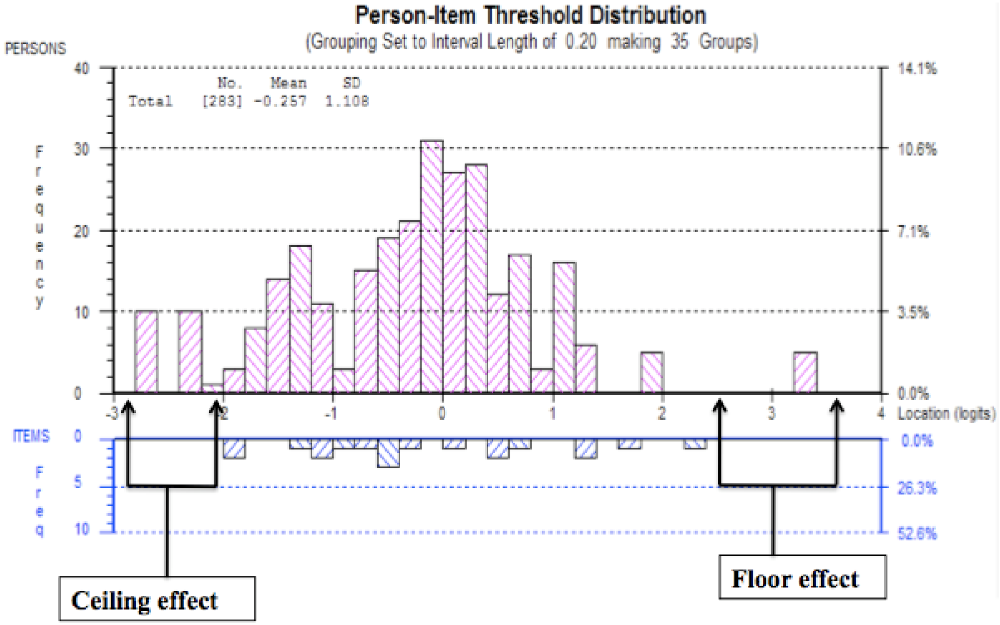
Figure 9a: Plot of DIF by age for ScF21 (Keeping up with school work)



In Figure 9a above the item curves for the various age groups have a higher degree of overlap in the first class interval (between the first and second point on each curve) and a lower degree of overlap in the second class interval (between the second and third point on each curve). Therefore the item ScF21 (Keeping up with school work) displays a greater degree of bias for higher ability respondents than for lower ability respondents. Overall the item curves for the various age groups have a high degree of overlap and therefore the DIF by age is marginal.

There is a floor and ceiling effect associated with the ScF sub-scale as displayed in Figure 9b below. The person ability mean is -0.26 signifying that in general the sample assessed has greater school function than that which can be assessed by the items in the ScF sub-scale.

Figure 9b: Person-item threshold map for school function sub-scale



The PSI and α values are 0.77 and 0.79, respectively, prior to correcting disordered thresholds and accounting for local dependency in items ScF22 and ScF23. After correcting thresholds and local dependency the PSI value remained similar (0.73); however the α value dropped slightly to 0.70 indicating that the α value prior to correcting for local dependency was inflated.

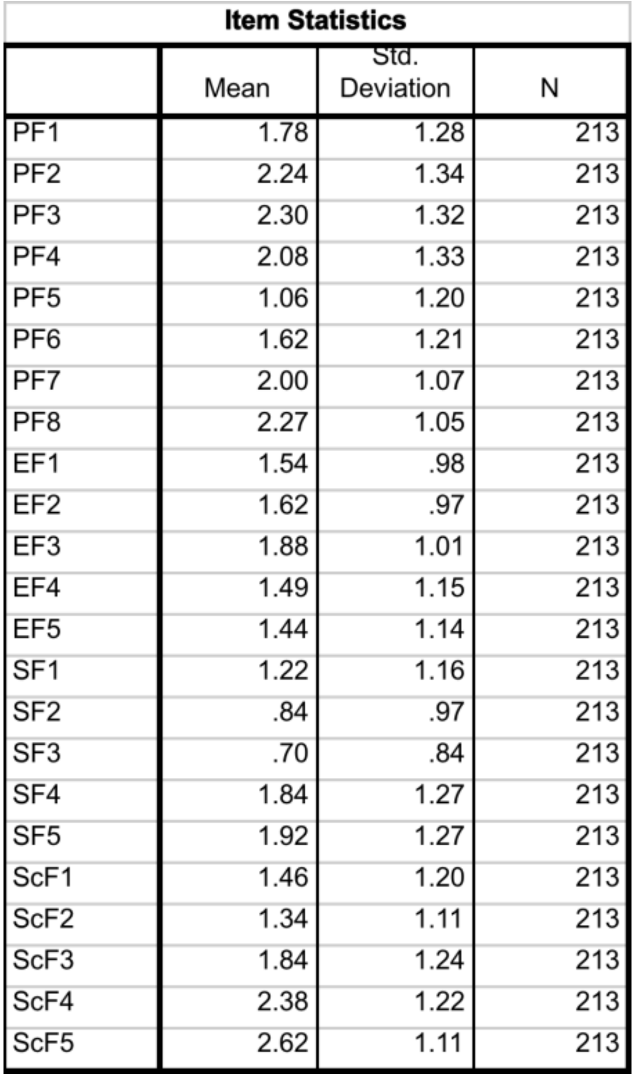
Classical Test Theory Analysis of PedsQL^{TM} 4.0 Generic Core Scales

PASW Statistics, version 18 was used to evaluate the psychometrics of the parent-report PedsQL^{TM} 4.0 Generic Core Scales from a traditional perspective. Descriptive statistics, item-total correlations (ITCs) and internal consistency co-efficient (ICCs) were assessed for the total summary score.

Descriptive Statistics

Mean scores and standard deviations for all items in the PedsQL^{TM} 4.0 are displayed in Table 6a. Mean scores of individual items were examined and those that fell in extreme ranges (between 0-1 or between 3-4) were flagged as not providing information about the sample (items are too difficult or too easy). SF2 (Friends) and SF3 (Teased) had overall means between 0-1 indicating these items are too easy and do not provide discriminatory power in this sample. Standard deviations of items were examined and those that were relatively narrow (<0.8) were flagged. No items fell into this range. Overall the item mean score was 1.72 (min 0.70, max 2.62, SD 0.24) signifying that, generally, items were too easy and that respondents did not endorse a full range of responses. The response categories "Often" and "Almost Always" were not endorsed by many respondents, suggesting a ceiling effect associated with the scale. Category frequencies indicate respondents preferred to endorse "Never" "Almost Never" and "Sometimes" response categories.

Table 6a: Descriptive statistics

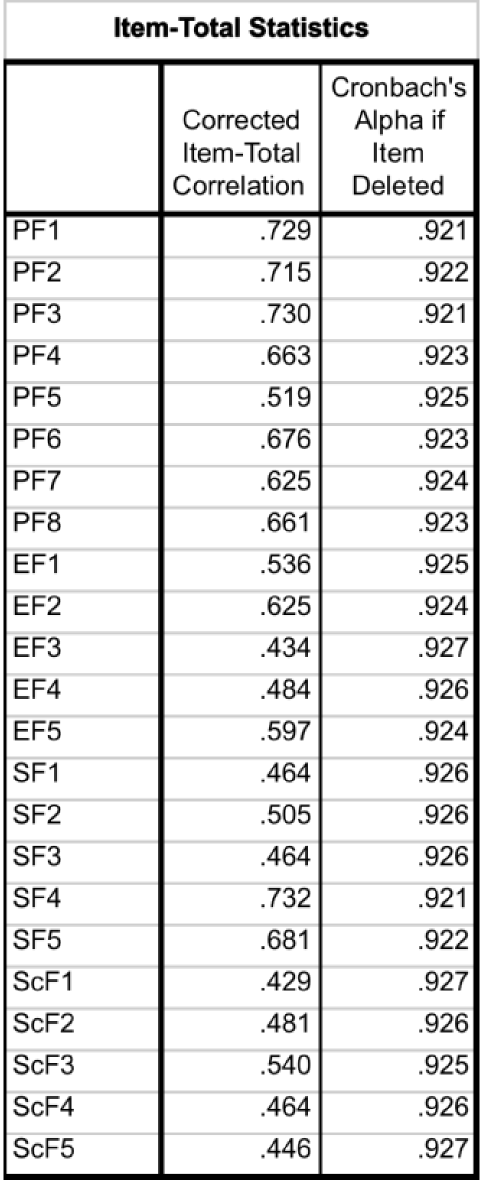


Item Statistics

	Mean	Std. Deviation	N
PF1	1.78	1.28	213
PF2	2.24	1.34	213
PF3	2.30	1.32	213
PF4	2.08	1.33	213
PF5	1.06	1.20	213
PF6	1.62	1.21	213
PF7	2.00	1.07	213
PF8	2.27	1.05	213
EF1	1.54	.98	213
EF2	1.62	.97	213
EF3	1.88	1.01	213
EF4	1.49	1.15	213
EF5	1.44	1.14	213
SF1	1.22	1.16	213
SF2	.84	.97	213
SF3	.70	.84	213
SF4	1.84	1.27	213
SF5	1.92	1.27	213
ScF1	1.46	1.20	213
ScF2	1.34	1.11	213
ScF3	1.84	1.24	213
ScF4	2.38	1.22	213
ScF5	2.62	1.11	213

Cronbach's α for the overall scale is 0.93, which is on the high end of the acceptable range. A very high α value may indicate item redundancy (or local dependency in Rasch terminology). Item-total correlations (ITCs) indicate the extent to which the items on a scale measure a common underlying construct. If ITCs are <0.2 or >0.9 combining items to produce a single score may not be appropriate. See Table 6b for corrected ITCs and Cronbach's α if an item is deleted. All items fell within the acceptable ITC range and Cronbach's α remained constant regardless of which item was deleted indicating that items contribute equally to the internal consistency reliability of the scale.

Table 6b: Corrected ITCs and Cronbach's α if item deleted



Item-Total Statistics

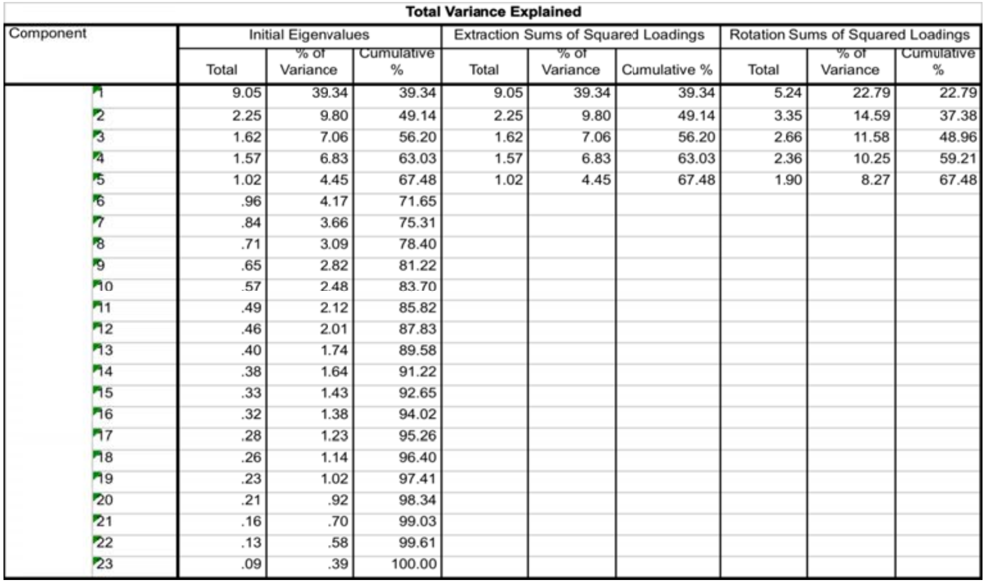
	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
PF1	.729	.921
PF2	.715	.922
PF3	.730	.921
PF4	.663	.923
PF5	.519	.925
PF6	.676	.923
PF7	.625	.924
PF8	.661	.923
EF1	.536	.925
EF2	.625	.924
EF3	.434	.927
EF4	.484	.926
EF5	.597	.924
SF1	.464	.926
SF2	.505	.926
SF3	.464	.926
SF4	.732	.921
SF5	.681	.922
ScF1	.429	.927
ScF2	.481	.926
ScF3	.540	.925
ScF4	.464	.926
ScF5	.446	.927

Factor Analysis

Principal axis factoring was followed by Varimax rotation. Initial Eigenvalues (EVs) were examined using the Kaiser criterion. All EVs <1 were ignored, so that for the item to be included it must explain at least the proportion of variance it introduces into the scale. See Table 6c for EVs and for the percent variance explained by components retained in the analysis. The first five components in the analysis had EVs of 9.05, 2.25, 1.62, 1.57 and 1.02, thereby indicating that the overall scale is comprised of five sub-scales. Together these five factors account for 67.48% of the total variance in the scale, which is over the acceptable level of at least 60% (Streiner, 1994).

The decision to retain five factors is not in agreement with the number of components that would be retained by examination of the Scree plot (Figure 10), which shows a second break at the fourth component, suggesting that the scale is comprised of three sub-dimensions.

Table 6c: Eigenvalues and % variance explained by components

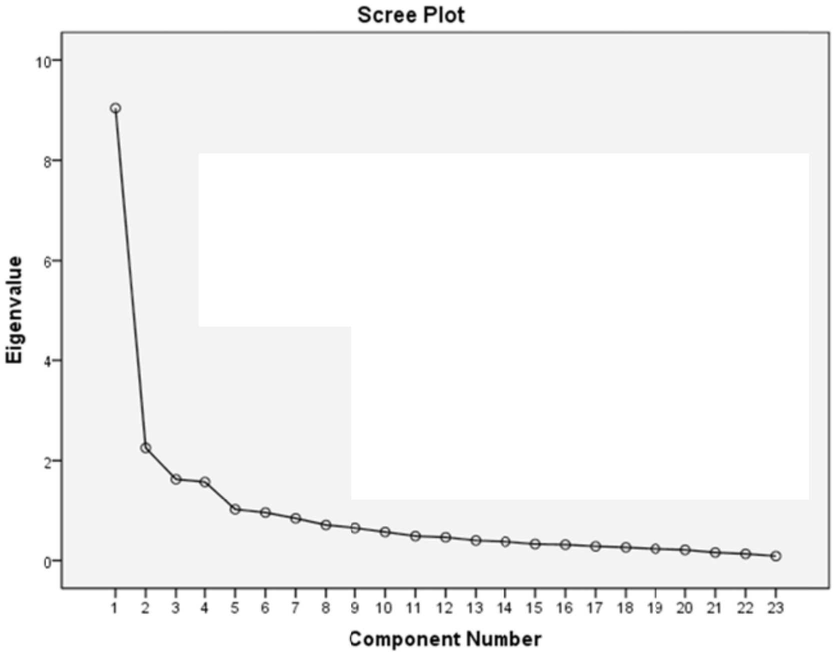


Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	9.05	39.34	39.34	9.05	39.34	39.34	5.24	22.79	22.79
2	2.25	9.80	49.14	2.25	9.80	49.14	3.35	14.59	37.38
3	1.62	7.06	56.20	1.62	7.06	56.20	2.66	11.58	48.96
4	1.57	6.83	63.03	1.57	6.83	63.03	2.36	10.25	59.21
5	1.02	4.45	67.48	1.02	4.45	67.48	1.90	8.27	67.48
6	.96	4.17	71.65						
7	.84	3.66	75.31						
8	.71	3.09	78.40						
9	.65	2.82	81.22						
10	.57	2.48	83.70						
11	.49	2.12	85.82						
12	.46	2.01	87.83						
13	.40	1.74	89.58						
14	.38	1.64	91.22						
15	.33	1.43	92.65						
16	.32	1.38	94.02						
17	.28	1.23	95.26						
18	.26	1.14	96.40						
19	.23	1.02	97.41						
20	.21	.92	98.34						
21	.16	.70	99.03						
22	.13	.58	99.61						
23	.09	.39	100.00						

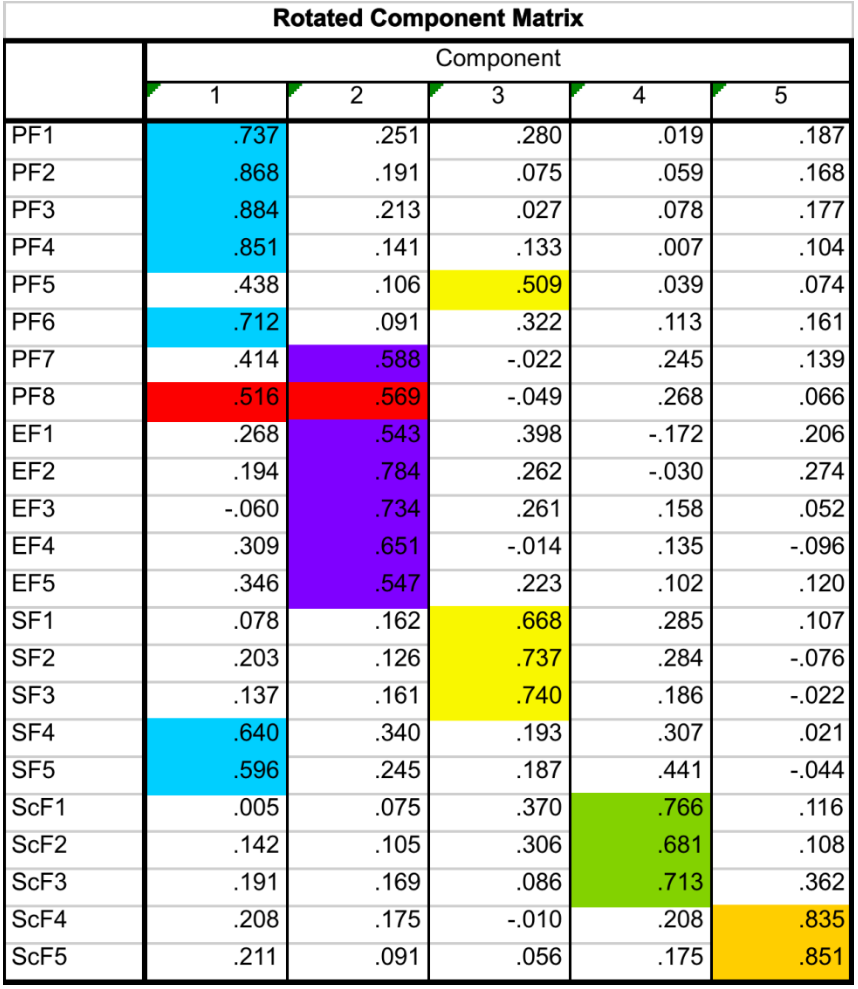
#

Figure 10: Scree plot



The rotated component matrix (Table 6d) was examined to determine how the items load onto each factor.

Table 6d: Rotated component matrix



Rotated Component Matrix

	Component				
	1	2	3	4	5
PF1	.737	.251	.280	.019	.187
PF2	.868	.191	.075	.059	.168
PF3	.884	.213	.027	.078	.177
PF4	.851	.141	.133	.007	.104
PF5	.438	.106	.509	.039	.074
PF6	.712	.091	.322	.113	.161
PF7	.414	.588	-.022	.245	.139
PF8	.516	.569	-.049	.268	.066
EF1	.268	.543	.398	-.172	.206
EF2	.194	.784	.262	-.030	.274
EF3	-.060	.734	.261	.158	.052
EF4	.309	.651	-.014	.135	-.096
EF5	.346	.547	.223	.102	.120
SF1	.078	.162	.668	.285	.107
SF2	.203	.126	.737	.284	-.076
SF3	.137	.161	.740	.186	-.022
SF4	.640	.340	.193	.307	.021
SF5	.596	.245	.187	.441	-.044
ScF1	.005	.075	.370	.766	.116
ScF2	.142	.105	.306	.681	.108
ScF3	.191	.169	.086	.713	.362
ScF4	.208	.175	-.010	.208	.835
ScF5	.211	.091	.056	.175	.851

In Table 6d above, items are highlighted to identify the factor on which they loaded: first factor items are blue; second factor items are purple; third factor items are yellow; fourth factor items are green; and fifth factor items are orange.

All items with the exception of PF8 (Energy), PF5 (Taking a bath or shower) and PF7 (Having hurts or aches) loaded most heavily on one factor. Item PF8 displayed the most factorial complexity (indicated by red highlight in Table 6d) as it loaded almost equally on the first and second factors and it was conceptually unclear where the item best fits. PF5 and PF7 loaded slightly higher on one factor and conceptually both of these items fit best with the construct they loaded on the greatest; thus it was easier to classify these items than it was to classify PF8.

Items SF4 (Age-appropriate ability) and SF5 (Play) loaded more heavily on the physical function factor (highlighted blue in Table 6d) instead of the hypothesized social function factor (highlighted yellow in Table 6d). Item PF5 (Bathing) loaded most heavily on the social function factor with items SF1 (Getting along) SF2 (Friends) and SF3 (Teased), instead of the hypothesized physical function factor.

The school function items that had content related to cognition loaded on the fourth factor (green in Table 6d). The remaining ScF items (Missing school-ill, Missing school-doctor) were the only two items that loaded on the fifth factor (orange in Table 6d). Ideally at least three items must load onto a factor for that factor to be retained so ignoring the fifth factor in this analysis may be justified. In any case, the fifth factor accounts for only 4.4% of the variance so without this factor the scale still explains 63.08% (67.48% - 4.4%) of the variance and is in the acceptable range. The results of this factor analysis justify potential removal of items ScF4 and ScF5 from the scale, as their removal does not appear to greatly sacrifice the content validity of the scale.

Cronbach's α for the total summary score of the measure is 0.93, which is acceptable for individual-level decision-making and reflects the findings of other psychometric studies of the PedsQL^{TM} conducted from a traditional standpoint (Varni et al., 2001; Varni et al., 2002; Varni et al., 2002; Varni et al., 2003; Varni et al., 2004; Varni & Burwinkle, 2006).

CHAPTER 6:

DISCUSSION

Interpretation of Results

In an attempt to compare and contrast findings from traditional and modern psychometric paradigms, as well as to determine if the use of Rasch analysis affords any additional benefit to interpretation of overall rating scale scores, a Rasch analysis of the PedsQL^{TM} 4.0 Generic Core Scales (parent-report) was conducted.

Threshold Ordering

Rasch analysis reveals that 14/23 items on the PedsQL^{TM} have disordered response categories, which means that respondents are not using response options as intended and scores on these items are not meaningful. This disordering may be the result of too many response categories or confusing label options. The PedsQL^{TM} is already established in the literature as a valid and reliable tool and it is the most widely used generic QoL tool in pediatric cancer (Eiser & Morse, 2001; Klassen et al., 2010). For this reason it may not be feasible to suggest collapsing the original 5-point scale to a 3-point scale based on the results of this study, and two other studies that did a Rasch analysis of the PedsQL^{TM} (Kook & Varni, 2008; Lamoureux et al., 2010). Kook & Varni (2008) found that the fourth response category "Almost always" did not work as intended for a number of items and recommended that the scale should be collapsed to a 4-point scale. Findings from Lamoureux et al. (2010) are the same as what was found in this study and suggested that response categories "Never" and "Almost never" as well as "Often" and "Sometimes" did not work as intended and that the scale should be collapsed to a 3-point scale.

The absence of a "Not applicable" response category may have compelled respondents to choose the response option "Never" instead. Using the category "Never" in this case does not accurately reflect the construct being measured, and therefore may have contributed to high levels of disordering in the PedsQL^{TM}. Rasch analysis is suited to generate an overall score based on the number of items that are answered therefore including a "Not applicable" response category would not necessitate the exclusion of respondents. Studies in larger samples should be conducted to determine how well respondents can discriminate between the 5-point response options on the PedsQL^{TM}. The use of qualitative research may aid in improving our understanding of respondents' thought processes when answering questions on the PedsQL^{TM}.

Collapsing categories to correct disordered thresholds on a scale decreases the internal consistency reliability of the scale for that sample, as there are fewer response options between which to discriminate (Streiner & Norman, 2008). Proponents of Rasch analysis defend the decision to collapse response categories despite decreasing the reliability as they believe items that have disordered thresholds do not produce meaningful information (Pallant & Tennant, 2007). Therefore, to ensure each item produces a valid and meaningful score that can be summed to produce an overall score, it is preferable to collapse thresholds at the expense of decreasing internal consistency reliability.

Individual Person and Item Fit

Person and item performance can be assessed based on fit to the Rasch model, in addition to examination of means and SDs as is done in traditional methods. Rumm software can be used to flag problematic items or people that significantly misfit the Rasch model as a way of alerting investigators to perhaps reconsider item wording and score interpretation for this data. There are no equivalent fit statistics employed in traditional methods.

Rasch analysis indicates item PF5 (Bathing) is particularly problematic for respondents; the correction applied to other items with disordered thresholds (collapsing the 5-point scale to a 3-point scale) did not correct the threshold pattern for this item. This item is also a poor discriminator as approximately 50% of respondents endorsed the "Never a problem" response option for this item (in a well-targeted 5-point scale about 20% of respondents should endorse each response option depending on the population being assessed).

Findings obtained using a CTT framework also support the problematic nature of item PF5 (Bathing). This item had a lower mean score than the majority of other items. Furthermore, factor analysis indicates the content of the Bathing item was the only item from the PF scale that did not adequately measure physical function. The Bathing item loaded most heavily on the factor representing Social Function. The unusual way this item loads may be rationalized using findings from the modern psychometric paradigm that indicate severe threshold disordering. Furthermore the wording of this item may also contribute to its unusual factor loading. The item seems to be asking about two separate issues, taking a bath or taking a shower by oneself, therefore it is unclear what the item is actually measuring.

Differential Item Functioning

The item characteristic curves produced by Rumm software permit easy identification of items that display DIF and allow re-assessment of this bias once strategies to correct for DIF are employed. There are equivalent procedures that can be used in traditional methods to assess DIF; however, these also require more advanced mathematical ability and cannot account for non-uniform DIF. Classical methods typically consider bias at the test-score level by using a different scoring formula for different sub-groups of the sample.

Marginal DIF (i.e., item bias) was detected based on age for the sub-test of physical function items in the overall summary score analysis, indicating that these items may not function equivalently for the different age groups. Parents may be better at reporting physical function for older children who can communication with them and inform them of any health concerns. Therefore the physical function item bias may be that younger children, who do not overtly express their health concerns to their parents, have parents who are not as easily able to act as a proxy respondent for their child.

Assessment of DIF is useful to consider in addressing the issue of parent-child agreement of rating scale scores and is an area to be explored in future research. DIF analysis can ensure each item operates equivalently for parent and child respondents and thus permits bias-free estimates of the level of parent-child agreement in reported scores. Item bias may alter how differences in scores are interpreted and thus it is essential that items operate equivalently for both sub-groups prior to making comparisons of agreement between their responses.

Targeting

Overall, the PedsQL^{TM} did not demonstrate good targeting as suggested by the lack of overlap of person ability and item difficulty in the person-item threshold maps for all Rasch analyses conducted. Examining the person-item threshold maps provides a visual means of determining how well the items spread over the range of the construct and is an easy way to determine whether the addition of a particular item improves targeting of the scale.

Poor targeting of the scale was also evident from a CTT perspective as the overall item mean was low (1.72/5). This low mean value signifies that on average the items were too easy for many respondents in the sample. Category frequencies show that few respondents used the "Often" and "Almost always" categories for items and therefore scored at the high end of the continuum of the QoL. Findings from this study that suggest poor targeting are supported by other modern and traditional psychometric studies that have evaluated the properties of the PedsQL^{TM} (Eiser & Eiser, 2007; Kook & Varni, 2008; Lamoureux et al., 2010; Varni et al., 2001; Varni et al., 2002).

In particular, this study and the available literature indicate a large ceiling effect associated with the PF summary score (over 30%) which may suggest that the eight items that comprise the PF sub-scale do not adequately assess physical health-related QoL in a higher functioning group of respondents, so that detecting improvement in this population will thus be difficult. The addition of an item such as the ability to dress oneself, a relatively difficult physical task that requires both gross and fine motor physical function, might improve targeting of the PF scale in a sample of higher functioning children.

A useful feature of Rumm software is the ability to divide person-item threshold maps based on person factors such as age or gender to determine if a particular sub-group of the sample is scoring at the floor or ceiling of the scale. When divided by gender, the person-item threshold map for the overall score shows that more males lie above the ceiling of the scale than do females (Figure 4g). Furthermore, when divided by age, the person-item threshold map indicates that it is mostly the younger age categories (2-4 and 5-7) that lie above the ceiling of the scale (Figure 4h). These findings suggest that the overall score on the PedsQL^{TM} for males in lower age groups should be interpreted with caution, as the items on the scale may not capture improvements in scores.

A stronger ceiling effect is common in generic QoL instruments as they are constructed with the intention of being applicable to a wide range of populations including healthy people. One would expect that cancer patients would score lower on this scale given that they are not well; however, a ceiling effect was still associated with all sub-scale and summary scores in the sample assessed for this study. Factors such as self-worth, social skills, participation and social support have not been studied in relation to QoL in a cancer population and the addition of these items may be useful in improving the range of QoL measured using the PedsQL^{TM}. Conducting qualitative interviews with cancer patients, cancer survivors and their families, to better understand their perspectives on how to measure QoL, may also be a useful exercise to improve targeting of QoL measures (Rajmil et al., 2004).

Reliability

After correcting for local dependency and disordered thresholds to improve fit to the Rasch model, the internal consistency reliability for the TS and PS summary scores dropped below the level acceptable for individual-level analysis, which is α= 0.9 (Nunnally & Bernstein, 1994). The internal consistency reliability of the PF summary score maintained adequacy for individual-level analysis. Prior to fitting data to the Rasch model, the internal consistency reliability values obtained in this study were supportive of values reported in the literature, which indicate the TS, PF and PS summary scores can all be used at the individual-level (Varni et al., 2001; Varni et al., 2002). The findings here would challenge that practice.

The Person Separation Index (PSI) for the TS and PS scores also dropped below the accepted level (set at 0.8) after fitting data to the Rasch model, suggesting that these scores should only be used to differentiate amongst two groups. The PSI for the PF summary score did not drop substantially after fit to the Rasch model and can be used to differentiate amongst four or more ability groups. The α values of individual sub-scale scores were generally adequate for group-level comparison before and after fit to the Rasch model. Similarly, the PSI for sub-scale scores did not change substantially after fit to the Rasch model; however, PSI values were low and could only be used to distinguish between two groups.

Validity

From a Rasch perspective it is only appropriate to sum items on a measure if they form a unidimensional construct. Depending on how a construct is defined it may make sense from a clinical standpoint to sum scores of items even if they do not form a unidimensional construct. The validity of using an overall summary score to represent QoL depends on the definition of QoL being studied. In this thesis the WHO definition of QoL is used, which states that QoL is an "individual's perception of their position in life in the context of their culture and value systems… and in relation to their goals, expectations and concerns" (WHOQoL Group, 1993, p.153). Using this definition a person could have high QoL regardless of their current health state or physical function and therefore it would not be appropriate to suggest that the sum of items in a unidimensional physical and psychosocial scale is an adequate representation of QoL. Based on the critique offered in this thesis, the PedsQL^{TM} is more appropriately classified as a health status instrument rather than a QoL instrument. Regarding the PedsQL^{TM} as a health status measure, it is appropriate to suggest that one could sum the scores of a psychosocial and physical scale as an overall indicator of health status. However, caution must be taken to ensure that the scores of individual sub-scales are also considered as one could get a wide array of varied sub-scale scores summing to the same overall score and therefore it would be difficult to interpret the meaning of the overall summary score.

Comparison of Findings using Classical Test Theory versus Rasch Analysis

From a Rasch analysis standpoint, results of this study generally did not support the internal consistency reliability and validity of the TS, PF or PS summary score of the PedsQL^{TM} 4.0. Items within sub-scales displayed a high degree of local dependency, which inflated internal consistency reliability coefficients and multi-dimensionality was apparent. Correcting thresholds and accounting for local dependency improved fit to the Rasch model and fulfilled the requirement of unidimensionality, which is a pre-requisite for summing items of a measure to produce an overall score.

From a CTT standpoint, there was some support for the validity of the overall summary score. Results of the factor analysis indicate a fifth factor consisting of items ScF4 and ScF5 (Missing school-ill, Missing school-doctor), which does not seem to be measuring school function as hypothesized by the original developers of the PedsQL^{TM}. It is suggested that to retain a factor there should be at least three items that load most heavily on that factor (Streiner, 1994). Therefore it may be justified to ignore the loading of these two items on the fifth factor of this analysis and to consider a four-factor solution as hypothesized by the developers of the PedsQL^{TM} .

Problematic items from a classical perspective were item PF8 (Energy), PF5 (Taking a bath or shower) and PF7 (Having hurts or aches) which all displayed factorial complexity. For PF8 (Energy) the complexity may be due to the fact that both the first factor (physical function) and the second factor (emotional function) relate to one's energy levels and therefore PF8 loaded almost equally on these factors. PF5 (Taking a bath or shower) loaded slightly higher on the third factor (social function) than on the first factor (physical function) where it was hypothesized to load. This unusual item loading may be a result of including two separate functions in a single item. Respondents may be unclear as to whether they are responding based on their ability to shower, which is done standing up, or on their ability to take a bath which can be done lying down or sitting. PF7 loaded slightly higher on the second factor (emotional function) rather than the first factor (physical function) where it was hypothesized to load. The content of this item may fit better with the emotional factor, as having hurts or aches can be upsetting and cause worry, which are both measured by other items on the emotional function sub-scale of the PedsQL^{TM}.

Items ScF4 and ScF5 (Missing school-ill, Missing school-doctor) and PF3 (Participating in sports) were problematic from a Rasch perspective as they displayed significant misfit to the Rasch model. The misfit may be used to explain why items ScF4 and ScF5 loaded on a separate factor in the factor analysis. Therefore analyzing psychometric findings using both paradigms offers a mechanism to better understand or to reinforce findings from one paradigm using the findings from the second paradigm.

Strengths and Limitations

The strengths of this study include the relatively large sample size, inclusion of parents of children treated on high and standard risk care protocols, and inclusion of parents of children from multiple sites in Canada (Hamilton, Toronto, Vancouver, Winnipeg and Kingston). These factors increase the generalizability of findings from this study.

Limitations are that the study was only conducted using parents of cancer patients and furthermore only parents of children who could speak English were included. This shortcoming limits the generalizability of findings to other populations of children with chronic conditions as well as other cultural groups. A further limitation is that parent-reported QoL during active treatment may be subject to inflation due to parental stress during this period (Johnston, Steele, Herrera, & Phipps, 2003). This additional stress may have influenced the psychometric findings obtained in this study.

Self-report data was not obtained as the aim of the intial study for which data was collected was to obtain parent perspectives (Sung et al., 2008; Sung et al., 2010). Some children in the sample were too young to complete a PedsQL^{TM} self-report (children must be over the age of five otherwise only parent-report is available). There was a large percentage of missing data for the school function sub-scale (33%) and therefore the sample size used in the analysis for this sub-scale was lower than for the other scales in both the Rasch and CTT analyses. For the Rasch analysis only respondents that had missing data for every item in the school function scale were excluded (as well as respondents with extreme scores); for the CTT analysis respondents with a missing data point on any item were excluded. Self-report and parent-report both have inherent limitations and thus the use of both to provide complementary information regarding child QoL is ideal for research and for clinical decision-making.

CHAPTER 7

CONCLUSIONS

Summary of Advantages and Disadvantages of Traditional and Modern Psychometrics

The study of traditional and modern psychometrics, such as Classical Test Theory (CTT) and Rasch analysis, provides a framework for the consideration of measurement issues and a platform to guide the interpretation of rating scale scores. Both theories have contributed to our understanding of the criteria required to produce rigorous measures and to minimize the influence of measurement error on rating scale scores so that an accurate portrayal of patient attributes and responses is made. The main caveat associated with CTT is that person and item parameters are dependent on the scale and the sample, respectively, and these dependencies can limit the interpretations made regarding person- and item-level statistics to the construct being measured.

This thesis presents an overview of the similarities and differences between the statistics analyzed in each paradigm and suggests some benefit to using both theories of measurement to guide the development and evaluation of rating scales. In addition the first Rasch analysis of the PedsQL^{TM} 4.0 Generic Core Scales (parent-report) was conducted in a childhood cancer sample in order to explore whether the PedsQL^{TM} is a valid and reliable tool to be used in childhood cancer from a modern standpoint. As well, analyses were conducted in order to determine whether the use of Rasch provides additional information to aid in interpretation of overall rating scale scores.

Traditional methods predominantly focus on test- and person-level statistics and highlight the importance of understanding the relationship between overall test scores and person ability. Limitations of CTT include circular dependency (i.e. respondents' test scores are dependent on the items administered and the properties of the tool are dependent on the sample from which they are generated) and the idea that raw scores generated from CTT measures are not interval-level data and therefore should not be analyzed using parametric statistics. Due to the focus on test-level properties in a CTT-developed measure, it is not possible to administer shorter versions of the same scale because psychometric properties are established for overall test scores and not for individual items. Furthermore, psychometrics must be re-evaluated, and new test norms established, when the test is administered to a sample that is different from the original sample the tool was developed for.

The main advantage of CTT is that data are not fit to a pre-determined mathematical model and thus assumptions do not need to be met prior to applying this framework. CTT methods generally require smaller sample sizes to develop and to evaluate measures and are thus are less time consuming and resource intensive (Hambleton & Jones, 1993).

Modern psychometric theories bring increased attention to item-level statistics and focus on exploring the relationship between item-level scores and person ability. Although CTT does address item-level statistics, these statistics are not a central feature of the paradigm. In Rasch analysis the estimate of item difficulty is independent of the persons taking the test, and the estimate of person ability is independent of the items they have taken. Therfore, respondents in a population can be compared using results from different test items, as person and item statistics are independent of the scale and sample.

Furthermore, because of the focus on item-level statistics and ensuring a tool is reliable and valid at the item-level, if a tool is developed using Rasch analysis it may not be necessary to administer every item on the rating scale. This is often the case if an item bank is available for the construct being measured. Select items, applicable to the individual being assessed, can be administered and the summed score of those items will be reliable and valid. This feature can be particularly appealing to researchers and clinicians who may want to administer only certain items or sub-scales of a rating scale that they feel are more applicable to an individual patient or a parent-proxy.

Disadvantages of Rasch analysis include the need for larger sample sizes (20 times the number of items in a measure if evaluating an existing measure and at least 50 cases to develop a new measure) (Linacre, 1994). A second disadvantage is the additional time and resources required to fit data to the Rasch model and then to transform the data to interval level data prior to analyzing for group differences (Hambleton & Jones, 1993). If the measure used to collect data has already been developed using Rasch analysis then it is more likely that the raw data collected will already fit the Rasch model, but the data must still be transformed to interval level data prior to statistical analysis. Raw data (ordinal level) can only be transformed to interval level data if the data fit the Rasch model. If data do not fit the Rasch model it is not clear from the available measurement literature how to proceed with the analysis, particularly if there is multi-dimensionality in the scale that is not corrected by sub-testing items. This ambiguity serves as a barrier to the widespread use of Rasch analysis in the development and evaluation of measures.

A solution to this barrier may be to do qualitative research at the start of a questionnaire development study to ensure that a meaningful conceptual framework is identified. Items and scales can then be developed from the conceptual framework in order to increase the validity of the measure being developed and perhaps decrease the likelihood of including items that create dimensionality in the scale. This suggestion may improve fit of data to the Rasch model at the onset of the study.

Lastly, the literature reviewed suggests that the additional item-level information gleaned using Rasch analysis benefits primarily those at the extreme ranges of the normal distribution. Therefore, for the majority of respondents, the use of Rasch will not provide additional information to guide the interpretation of overall rating scale scores.

Recommendations

Rasch analysis results bring into question the internal consistency reliability of PedsQL^{TM} 4.0 Generic Core Scales that are established from a traditional standpoint. Published psychometrics of the PedsQL^{TM} 4.0 suggest the overall score is reliable at the individual level; however, when local dependency of items in sub-scales is taken into consideration the overall score has adequate internal consistency reliability for group-level analysis. Rasch analysis reveals respondents can only discriminate between three response options and not five and that there is large ceiling effect associated with all sub-scale and summary scores. Further studies should be conducted in other populations to determine if these findings are replicated. The results of this study challenge the use of the PedsQL^{TM} 4.0 overall summary score as an indicator of QoL (as it is defined by the WHO) or for analysis of individual-level differences in a cancer population. If a fifth version of the PedsQL^{TM} 4.0 were designed it might be beneficial to use a 3-point response option and to add in items that target higher functioning samples and that better reflect a subjective perception of QoL.

Relationships between the information afforded by analysis of item-level statistics versus test- and person-level statistics in a psychometric evaluation of the PedsQL^{TM} 4.0 reveals some benefit in the use of both paradigms as complementary tools to maximize our understanding of rating scale scores. Rasch analysis permits investigators a means of examining item-level statistics in a more detailed and visually pleasing fashion than that possible through the exclusive use of CTT. Category probability curves display response thresholds for each item and permit a means of evaluating whether response categories are being used as they were intended. Person-item threshold maps display how well the items on a measure target the sample being assessed and can be used to easily identify if the addition of a new item improves targeting of the scale for a particular sample. Again, studies should be conducted with other populations to determine if findings are comparable.

It is recommended to include DIF analysis and examination of threshold ordering as part of mainstream traditional psychometric testing. This item-level focus is necessary as DIF may cancel out at the test-level, and if traditional methods are used in isolation it will not be possible to identify which items display bias. Examining item thresholds can provide guidance in regard to the number of categories that respondents are able to differentiate. These additional analyses will contribute to the development of shorter and more robust rating scales as a consequence of the additional focus on item-level analysis. Although it is also possible to explore this information using CTT, the methods are not user friendly, require advanced mathematical abilities and thus are not part of a typical traditional psychometric evaluation.

Lastly, using traditional and modern paradigms together affords the potential to rationalize aberrant findings from one paradigm using the information gathered from the second paradigm. For example, in this thesis Rasch analysis was useful in identifying particular item characteristics, such as significant model misfit, to rationalize why an item may load in an incongruous fashion in a factor analysis. Studies must be conducted in other populations (chronic and healthy) to determine if findings support the results of this study, which was only conducted in cancer patients.

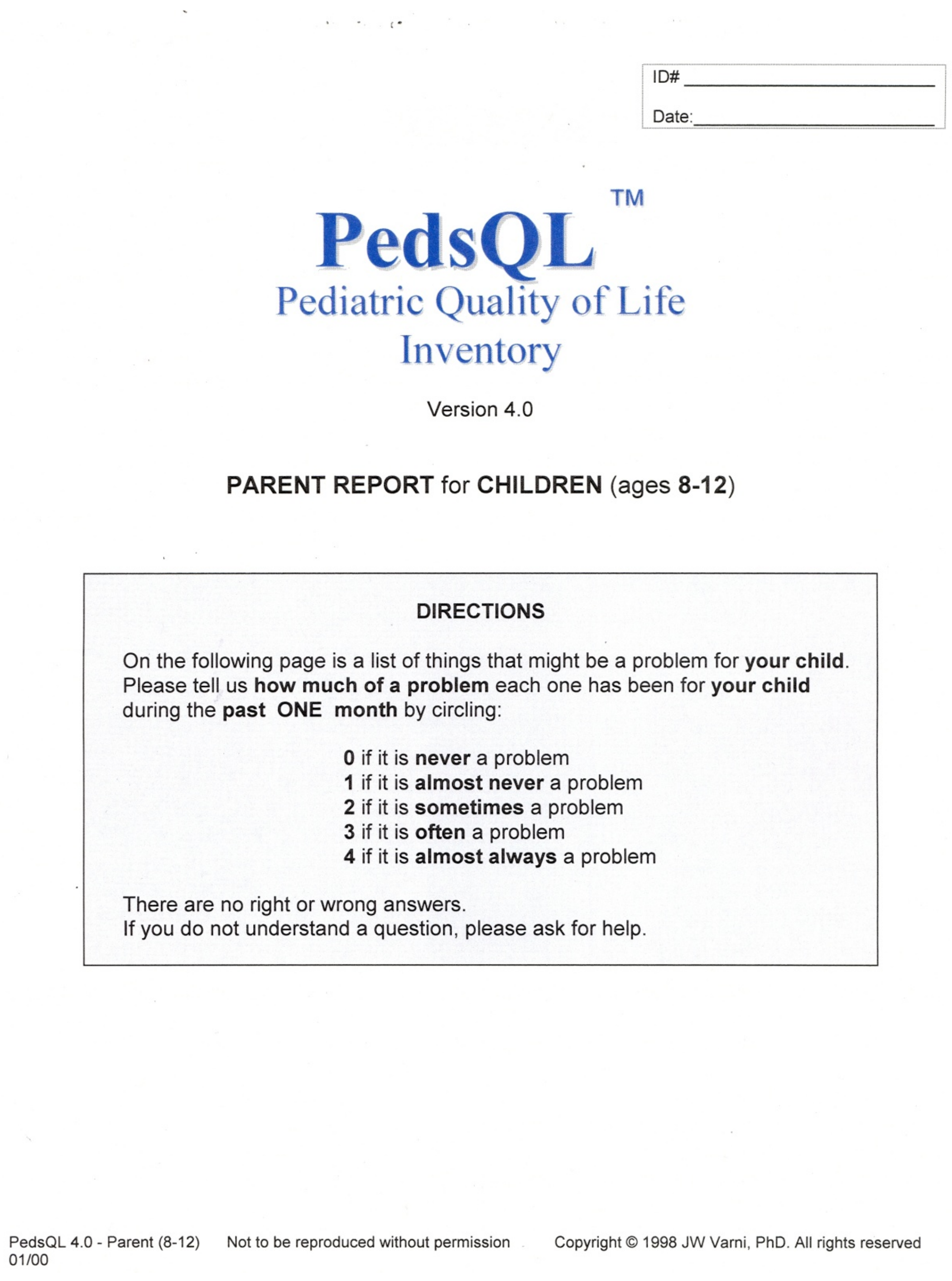
Future Directions

The psychometric theory behind the development of a measure has important implications for the meaning of the overall score obtained from the scale. The assumptions made with regard to how the rating scale score can be interpreted is a direct consequence of the theory used to develop and to evaluate the measure. The use of CTT and Rasch analysis as complementary approaches is warranted to further our understanding of the meaning of a rating scale score. The use of traditional methods continues to predominate as the preferred method to develop and to evaluate rating scales. Further research on specific testing circumstances that would render Rasch analysis as particularly useful in complementing the information available from traditional methods may justify the additional time and resources invested when using both paradigms to develop and evaluate rating scales.

APPENDICES

APPENDIX 1

PedsQL^{TM} 4.0 Generic Core Scales (Parent-report)



ID# ______________________

Date: ______________________

PedsQL™

Pediatric Quality of Life Inventory

Version 4.0

PARENT REPORT for **CHILDREN** (ages **8-12**)

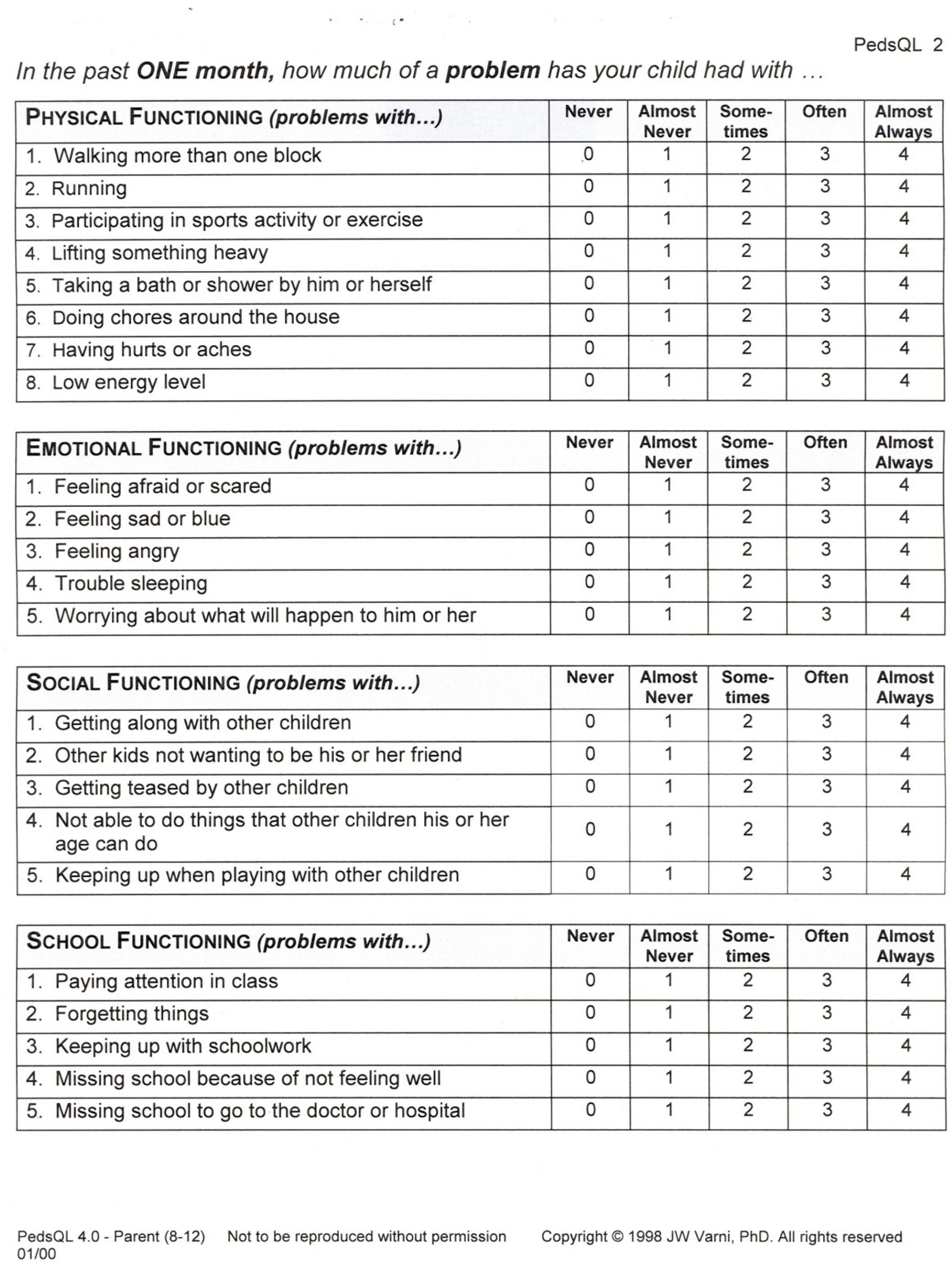
DIRECTIONS

On the following page is a list of things that might be a problem for **your child**. Please tell us **how much of a problem** each one has been for **your child** during the **past ONE month** by circling:

0 if it is **never** a problem
1 if it is **almost never** a problem
2 if it is **sometimes** a problem
3 if it is **often** a problem
4 if it is **almost always** a problem

There are no right or wrong answers.
If you do not understand a question, please ask for help.

PedsQL 4.0 - Parent (8-12) Not to be reproduced without permission
01/00



PedsQL 2

In the past ***ONE month,*** *how much of a* **problem** *has your child had with …*

PHYSICAL FUNCTIONING *(problems with…)*	Never	Almost Never	Some-times	Often	Almost Always
1. Walking more than one block	0	1	2	3	4
2. Running	0	1	2	3	4
3. Participating in sports activity or exercise	0	1	2	3	4
4. Lifting something heavy	0	1	2	3	4
5. Taking a bath or shower by him or herself	0	1	2	3	4
6. Doing chores around the house	0	1	2	3	4
7. Having hurts or aches	0	1	2	3	4
8. Low energy level	0	1	2	3	4

EMOTIONAL FUNCTIONING *(problems with…)*	Never	Almost Never	Some-times	Often	Almost Always
1. Feeling afraid or scared	0	1	2	3	4
2. Feeling sad or blue	0	1	2	3	4
3. Feeling angry	0	1	2	3	4
4. Trouble sleeping	0	1	2	3	4
5. Worrying about what will happen to him or her	0	1	2	3	4

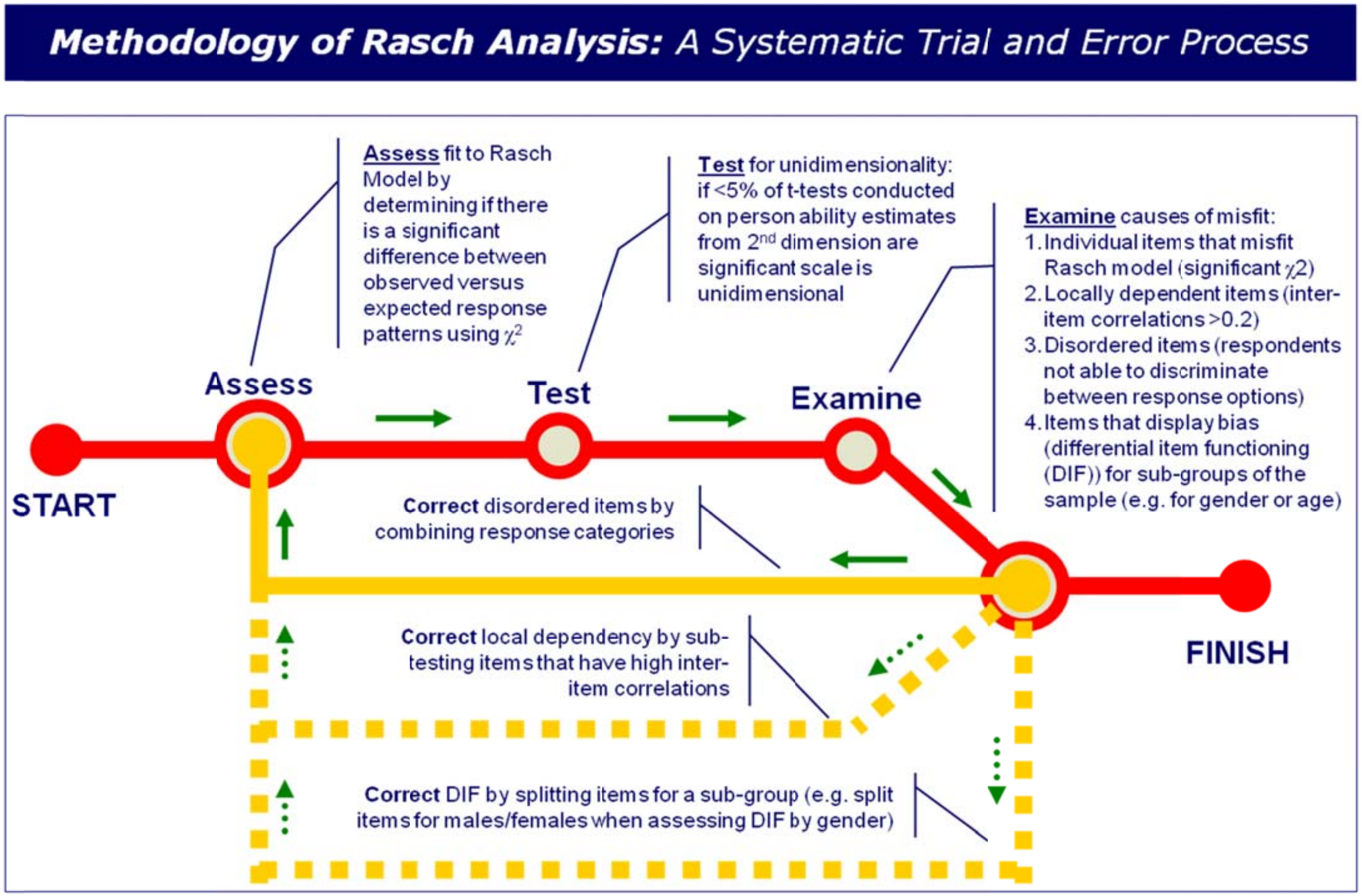
SOCIAL FUNCTIONING *(problems with…)*	Never	Almost Never	Some-times	Often	Almost Always
1. Getting along with other children	0	1	2	3	4
2. Other kids not wanting to be his or her friend	0	1	2	3	4
3. Getting teased by other children	0	1	2	3	4
4. Not able to do things that other children his or her age can do	0	1	2	3	4
5. Keeping up when playing with other children	0	1	2	3	4

SCHOOL FUNCTIONING *(problems with…)*	Never	Almost Never	Some-times	Often	Almost Always
1. Paying attention in class	0	1	2	3	4
2. Forgetting things	0	1	2	3	4
3. Keeping up with schoolwork	0	1	2	3	4
4. Missing school because of not feeling well	0	1	2	3	4
5. Missing school to go to the doctor or hospital	0	1	2	3	4

PedsQL 4.0 - Parent (8-12) Not to be reproduced without permission
01/00

APPENDIX 2

Rasch Analysis Flow Diagram



Methodology of Rasch Analysis: *A Systematic Trial and Error Process*

Assess fit to Rasch Model by determining if there is a significant difference between observed versus expected response patterns using χ^2

Test for unidimensionality: if <5% of t-tests conducted on person ability estimates from 2nd dimension are significant scale is unidimensional

Examine causes of misfit:

1. Individual items that misfit Rasch model (significant χ2)
2. Locally dependent items (inter-item correlations >0.2)
3. Disordered items (respondents not able to discriminate between response options)
4. Items that display bias (differential item functioning (DIF)) for sub-groups of the sample (e.g. for gender or age)

Correct disordered items by combining response categories

Correct local dependency by sub-testing items that have high inter-item correlations

Correct DIF by splitting items for a sub-group (e.g. split items for males/females when assessing DIF by gender)

REFERENCES

Albrecht, G. L., & Devlieger, P. J. (1999). The disability paradox: High quality of life against all odds. *Social Science Medicine, 48*(8), 977-988.

Andrich, D., Lyne, A., Sheridan, B., & Luo, G. (2010). *RUMM2030*. Perth: RUMM Laboratory.

Bland, J. M., & Altman, D. G. (1997). Cronbach's alpha. *British Medical Journal, 314*, 570-572.

Davis, E., Nicolas, C., Waters, E., Cook, K., Gibbs, L., Gosch, A., et al. (2007). Parent-proxy and child self-reported health-related quality of life: Using qualitative methods to explain the discordance. *Quality of Life Research, 16*(5), 863-871.

Davis, E., Waters, E., Mackinnon, A., Reddihough, D., Graham, H., Mehmet-Radji, O., et al. (2006). Paediatric quality of life instruments: A review of the impact of the conceptual framework on outcomes. *Developmental Medicine and Child Neurology, 48*(4), 311-318.

Drotar, D. (2004). Validating measures of pediatric health status, functional status, and health-related quality of life: Key methodological challenges and strategies. *Ambulatory Pediatrics, 4*(4), 358-364.

Eiser, C. (2001). Can parents rate their child's health-related quality of life? results of a systematic review. *Quality of Life Research, 10*, 347-357.

Eiser, C., & Eiser, R. (2007). Mothers’ ratings of quality of life in childhood cancer: Initial optimism predicts improvement over time. *Psychology Health, 22*(5), 535-543.

Eiser, C., & Morse, R. (2001). Quality-of-life measures in chronic diseases of childhood. *Health Technology Assessment, 5*(4), 1-157.

Eiser, C., Vance, Y. H., Horne, B., Glaser, A., & Galvin, H. (2003). The value of the PedsQL^{TM} in assessing quality of life in survivors of childhood cancer. *Child Care Health and Development, 29*(2), 95-102.

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.

Fayed, N., Schiariti, V., Bostan, C., Cieza, A., & Klassen, A. (2011). Health status and QOL instruments used in childhood cancer research: Deciphering conceptual content using world health organization definitions. *Quality of Life Research.*

Feeny, D., William, F., Mulhern, R. K., Barr, R. D., & Hudson, M. (1999). A framework for assessing health□related quality of life among children with cancer. *International Journal of Cancer, 83*(S12), 2-9.

Felder Puig, R., Frey, E., Proksch, K., Varni, J. W., Gadner, H., & Topf, R. (2004). Validation of the german version of the pediatric quality of life inventory (PedsQL^{TM}) in childhood cancer patients off treatment and children with epilepsy. *Quality of Life Research, 13*(1), 223-234.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement, 12*(3), 38-47.

Hays, R. D., Morales, L. S., & Reise, S. P. (2000). Item response theory and health outcomes measurement in the 21st century. *Medical Care, 38*(9), 1128-1142.

Hill, C. D., Edwards, M. C., Thissen, D., Langer, M. M., Wirth, R. J., Burwinkle, T. M., et al. (2007). Practical issues in the application of item response theory: A demonstration using items from the pediatric quality of life inventory (PedsQL^{TM}) 4.0 generic core scales. *Medical Care, 45*(5), S39-S47.

Hobart, J., & Cano, S. (2009). Improving the evaluation of therapeutic interventions in multiple sclerosis: The role of new psychometric methods. *Health Technology Assessment, 13*(12), 1-214.

Hobart, J. C., Cano, S. J., Zajicek, J. P., & Thompson, A. J. (2007). Rating scales as outcome measures for clinical trials in neurology: Problems, solutions, and recommendations. *Lancet Neurology, 6*(12), 1094-1105.

Holmes, W., Bix, B., & Shea, J. (1996). SF-20 score and item distributions in a human immunodeficiency virus-seropositive sample. *Medical Care, 34*(6), 562-569.

Holmes, W. C., & Shea, J. A. (1997). Performance of a new, HIV/AIDS-targeted quality of life (HAT-QoL) instrument in asymptomatic seropositive individuals. *Quality of Life Research, 6*(6), 561-571.

Johnston, C. A., Steele, R. G., Herrera, E. A., & Phipps, S. (2003). Parent and child reporting of negative life events: Discrepancy and agreement across pediatric samples. *Journal of Pediatric Psychology, 28*(8), 579-588.

Klassen, A., Khan, A., Anthony, S., Klaassen, R., & Sung, L. (2010). Conceptualizing quality of life in children with cancer and childhood cancer survivors [Abstract]. *2010 International Society for Quality of Life Research Meeting Abstracts. Quality of Life Research Journal, 1653*(76).

Klassen, A. F., Anthony, S. J., Khan, A., Sung, L., & Klaassen, R. (2011). Identifying determinants of quality of life of children with cancer and childhood cancer survivors: A systematic review. *Supportive Care in Cancer.*

Klassen, A. F., Strohm, S. J., Maurice-Stam, H., & Grootenhuis, M. A. (2009). Quality of life questionnaires for children with cancer and childhood cancer survivors: A review of the development of available measures. *Supportive Care in Cancer, 18*(9), 1207-1217.

Kook, S. H., & Varni, J. W. (2008). Validation of the Korean version of the pediatric quality of life inventory (PedsQL^{TM}) 4.0 generic core scales in school children and adolescents using the rasch model. *Health and Quality of Life Outcomes, 6*(1), 41.

Lamoureux, E. L., Manjula, M., Chang, B., Dirani, M., Kah-Guan, A. E., Chia, A., et al. (2010). Is the pediatric quality of life inventory valid for use in preschool children with refractive errors? *Optometry and Vision Science, 87*(11), 813-822.

Lenert, L., & Kaplan, R. M. (2000). Validity and interpretation of preference-based measures of health-related quality of life. *Medical Care, 38*(9), II-138-II-150.

Leplège, A., & Hunt, S. (1997). The problem of quality of life in medicine. *Journal of the American Medical Association, 278*(1), 47-50.

Linacre, J. M. (2007). *WINSTEPS rasch measurement computer program* (3.64.1 ed.). Chicago, IL: Winsteps.com.

Linacre, J.M. 1994 7:4 p.328. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7*(4), May 27, 2011. Retrieved from http://www.rasch.org/rmt/rmt74m.htm

McKenna, S. P. (2002). Improving the sensitivity of the quality of life in depression scale (QLDS). *Quality of Life Research, 11,* 625.

Meeske, K., Katz, E. R., Palmer, S. N., Burwinkle, T., & Varni, J. W. (2004). Parent proxy–reported health□related quality of life and fatigue in pediatric patients diagnosed with brain tumors and acute lymphoblastic leukemia. *Cancer, 101*(9), 2116-2125.

Mokkink, L. B., Terweea, C. B., Patrickb, D. L., Alonsoc, J., Stratford, P. W., Knola, D. L., et al. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology, 63*(7), 737-745.

Neumann, P. J., Goldie, S. J., & Weinstein, M. C. (2000). Preference-based measures in economic evaluation in health care. *Annual Review of Public Health, 21*(1), 587-611.

Norman, G. R., & Streiner, D. L. (2008). *Biostatistics: The bare essentials* (3rd ed.). Hamilton: BC Decker Inc.

Novakovic, B., Fears, T. R., Horowitz, M. E., Tucker, M. A., & Wexler, L. H. (1997). Late effects of therapy in survivors of ewing's sarcoma family tumors. *Journal of Pediatric Hematology/Oncology, 19*, 220-225.

Nunnally, J. C., & Bernstein, I. R. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

Pallant, J. F., & Tennant, A. (2007). An introduction to the rasch measurement model: An example using the hospital anxiety and depression scale (HADS). *British Journal of Clinical Psychology, 46*(1), 1-18.

Patrick, D. L., & Erickson, J. (1993). *Health status and health policy: Quality of life in health care evaluation and resource allocation*. New York, NY: Oxford University Press.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Erlbaum.

Pemberger, S., Jagsch, R., Frey, E., Felder-Puig, R., Gadner, H., Kryspin-Exner, I., et al. (2005). Quality of life in long-term childhood cancer survivors and the relation of late effects and subjective well-being. *Supportive Care in Cancer, 13*(1), 49-56.

Poretti, A., Grotzer, M. A., Ribi, K., Schönle, E., & Boltshauser, E. (2004). Outcome of craniopharyngioma in children: Long-term complications and quality of life. *Developmental Medicine and Child Neurology, 46*(4), 220-229.

Rajmil, L., Herdman, M., Fernandez de Sanmamed, M. J., Detmar, S., Bruil, J., Ravens-Sieberer, U., et al. (2004). Generic health-related quality of life instruments in children and adolescents: A qualitative analysis of content. *Journal of Adolescent Health, 34*(1), 37-45.

Reis, L. A., Melbert, D., Krapcho, M., Stinchcomb, D. G., Howlader, N. & Horner, M. J. (2007). *SEER cancer statistics review*. Retrieved March 1, 2009, from http://seer.cancer.gov/csr/1975_2005/index.html

Reise, S. P., & Henson, J. M. (2003). A discussion of modern versus traditional psychometrics as applied to personality assessment scales. *Journal of Personality Assessment, 81*(2), 93-103.

Revicki, D. A. (2007). FDA draft guidance and health-outcomes research. *Lancet, 369*(9561), 540-542.

Riley, A.W. (2004). Evidence that school-age children can self-report on their health. *Ambulatory Pediatrics, 4*(4S), 371-376.

Rosenbaum, P. L. (2009). The quality of life for the young adult with neurodisability: Overview and reprise. *Developmental Medicine and Child Neurology, 51*(8), 679-682.

Rosenbaum, P. L. (1998). Screening tests and standardized assessments used to identify and characterize developmental delays. *Seminars in Pediatric Neurology, 5*(1), 27-32.

Rosenbaum, P. L., Livingston, M. H., Palisano, R. J., Galuppi, B. E., & Russell, D. J. (2007). Quality of life and health□related quality of life of adolescents with cerebral palsy. *Developmental Medicine and Child Neurology, 49*(7), 516-521.

Russell, K. M., Hudson, M., Long, A., & Phipps, S. (2006). Assessment of health□related quality of life in children with cancer. *Cancer, 106*(10), 2267-2274.

Schumacker, R. E., & Smith, E. V. (2007). A rasch perspective. *Educational and Psychological Measurement, 67*(3), 394-409.

Smith, K. W., Avis, N. E., & Assmann, S. F. (1999). Distinguishing between quality of life and health status in quality of life research: A meta-analysis. *Quality of Life Research, 8*, 447-459.

Smith, E. V. (2001). Evidence for the reliability of measures and validity of measure interpretation: A rasch measurement perspective. *Journal of Applied Measurement, 2*(3), 281-311.

SPSS Inc. (2010). *PASW statistics* (18th ed.). Chicago, IL: SPSS Inc.

Streiner, D. L. (1994). Figuring out factors: The use and misuse of factor analysis. *Canadian Journal of Psychiatry, 39*(3), 135-140.

Streiner, D. L., & Norman, G. R. (2008). *Health measurement scales: A practical guide to their development and use* (4th ed.). Oxford, UK: Oxford University Press.

Sung, L., Dix, D., Pritchard, S., Yanofsky, R., Dzolganovski, B., Almeida, R., et al. (2008). Identification of paediatric cancer patients with poor quality of life. *British Journal of Cancer, 100*(1), 82-88.

Sung, L., Klaassen, R. J., Dix, D., Pritchard, S., Yanofsky, R., Ethier, M. C., et al. (2009). Parental optimism in poor prognosis pediatric cancers. *Psycho-Oncology, 18*(7), 783-788.

Sung, L., Yanofsky, R., Klaassen, R. J., Dix, D., Pritchard, S., Winick, N., et al. (2010). Quality of life during active treatment for pediatric acute lymphoblastic leukemia. *International Journal of Cancer, 128*(5), 1213-1220.

Svensson, E. (2001). Guidelines to statistical evaluation of data from rating scales and questionnaires. *Journal of Rehabilitation Medicine, 33*(1), 47.

Tennant, A., & Conaghan, P. G. (2007). The rasch measurement model in rheumatology: What is it and why use it? when should it be applied, and what should one look for in a rasch paper? *Arthritis Rheumatism, 57*(8), 1358-1362.

Tennant, A., McKenna, S. P., & Hagell, P. (2004). Application of rasch analysis in the development and application of quality of life instruments. *Value in Health, 7*(1), S22-S26.

Tennant, A., & Pallant, J. F. *Introduction to rasch analysis*. Leeds,UK: Psychometric Laboratory for Health Sciences, Department of Rehabilitation Medicine, The University of Leeds.

Teresi, J. A., & Fleishman, J. A. (2007). Differential item functioning and health assessment. *Quality of Life Research, 16*(S1), 33-42.

Teresi, J. A. (2006). Different approaches to differential item functioning in health applications: Advantages, disadvantages and some neglected topics. *Medical Care, 44*(11), S152-S170.

Upton, P., Lawford, J., & Eiser, C. (2008). Parent–child agreement across child health-related quality of life instruments: A review of the literature. *Quality of Life Research, 17*(6), 895-913.

Uzark, K., Jonesa, K., Burwinkle, T. M., & Varni, J. W. (2003). The pediatric quality of life inventory (TM) in children with heart disease. *Progress in Pediatric Cardiology, 18*(2), 141-148.

Vance, Y. H., Jenney, M. E., Eiser, C., & Morse, R. C. (2001). Issues in measuring quality of life in childhood cancer: Measures, proxies, and parental mental health. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 42*(5), 661-667.

Varni, J. W. (1999). The PedsQL^{TM}: Measurement model for the pediatric quality of life inventory. *Medical Care, 37*(2), 126.

Varni, J. W., Burwinkle, T., Katz, E., Meeske, K., & Dickinson, P. (2002). The PedsQL^{TM} in pediatric cancer: Reliability and validity of the pediatric quality of life inventory generic core scales, multidimensional fatigue scale, and cancer module. *Cancer, 94*(7), 2090-2106.

Varni, J. W., & Burwinkle, T. M. (2006). The PedsQL^{TM} as a patient-reported outcome in children and adolescents with attention-Deficit/Hyperactivity disorder: A population-based study. *Health and Quality of Life Outcomes, 4*(1), 26.

Varni, J. W., Burwinkle, T. M., Rapoff, M. A., Kamps, J. L., & Olson, N. (2004). The PedsQL^{TM} in pediatric asthma: Reliability and validity of the pediatric quality of life inventory generic core scales and asthma module. *Journal of Behavioral Medicine, 27*(3), 297-318.

Varni, J. W., Burwinkle, T. M., Seid, M., & Skarr, D. (2003). The PedsQL^{TM} as a pediatric population health measure: Feasibility, reliability, and validity. *Ambulatory Pediatrics, 3*(6), 329-341.

Varni, J. W., Katz, E. R., Seid, M., Quiggins, D. L., & Friedman-Bender, A. (1998). The pediatric cancer quality of life inventory□32 (PCQL□32). *Cancer, 82*(6), 1184-1196.

Varni, J. W., Seid, M., & Kurtin, P. S. (2001). PedsQL^{TM} 4.0: Reliability and validity of the pediatric quality of life inventory version 4.0 generic core scales in healthy and patient populations. *Medical Care, 39*(8), 800-812.

Varni, J. W., Seid, M., Knight, T. M., Uzark, K., & Szer, I. S. (2002). The PedsQL^{TM} 4.0 generic core scales: Sensitivity, responsiveness, and impact on clinical decision-making. *Journal of Behavioral Medicine, 25*(2), 175-193.

Varni, J. W., Seid, M., & Rode, C. A. (1999). The PedsQL^{TM}: Measurement model for the pediatric quality of life inventory. *Medical Care, 37*(2), 126-139.

Varni, J. W., Seid, M., Smith Knight, T., Burwinkle, T. M., Brown, J., & Szer, I. S. (2002). The PedsQL^{TM} in pediatric rheumatology. Reliability, validity and responsiveness of the pediatric quality of life inventory generic core scales and rheumatology module. *Arthritis and Rheumatism, 46*(3), 714-723.

Varni, J. W., & Setoguchi, Y. (1992). Screening for behavioral and emotional problems in children and adolescents with congenital or acquired limb deficiencies. *Archives of Pediatrics Adolescent Medicine, 146*(1), 103-113.

Waters, E., Davis, E., Ronen, G. M., Rosenbaum, P., Livingston, M., & Saigal, S. (2009). Quality of life instruments for children and adolescents with neurodisabilities: How to choose the appropriate instrument. *Developmental Medicine and Child Neurology, 51*(8), 660-669.

WHOQoL Group. (1993). Study protocol for the world health organization project to develop a quality of life assessment instrument (WHOQoL). *Quality of Life Research, 2*, 153-159.