2018

Applying Polytomous Rasch Analysis To Validate Parenting Related Scales

Lei Huang
University of Alabama at Birmingham

APPLYING POLYTOMOUS RASCH ANALYSIS TO VALIDATE PARENTING RELATED SCALES

by

LEI HUANG

ROBIN GAINES LANZI, COMMITTEE CHAIR
CONNIE KOHLER, ADVISOR
SCOTT SNYDER, MENTOR
PETER HENDRICKS
DAVID REDDEN

A DISSERTATION

Submitted to the graduate faculty of The University of Alabama at Birmingham,
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

BIRMINGHAM, ALABAMA

2018

APPLYING POLYTOMOUS RASCH ANALYSIS TO VALIDATE PARENTING RELATED SCALES

LEI HUANG

HEALTH EDUCATION AND HEALTH PROMOTION

ABSTRACT

This study applies Rasch analysis to validate two scales in the Parenting for the First-time project: Beck Depression Inventory (BDI-II) and Parenting Stress Index (PSI). The main purpose of this study is to evaluate quality of the two scales by using Rasch analysis to examine the unidimensionality, reliability, fit statistics, group and time invariance, and optimal response categories. Meanwhile, this study also compares the results of Rasch model and Classical Test Theory (CTT) to assess the advantages of Rasch analysis.

Result: Both Rach analysis and CTT show the evidence of unidimensionality of BDI-II in the three administrative periods: Prenatal, 6 month, and 12 month. For parenting stress scale, the results of both analysis shows that there are two dimensions: Childrearing (CRI) and Self Stress (SSI). Overall, the Rasch person reliabilities index are less than Cronbach alphas for the three scales (BDI-II, CRI, and SSI) in each administrative period.

Most of items in BDI-II do not remain invariant across two age groups (<19 and >=19). Item 29, 31, 32, 33 (misfit items) of CRI cannot keep invariant across the age groups. No DIF items are found in SSI scale. Differential Test Function (DTF) analysis shows that, except for several items, BDI-II roughly functions similarly over the three administration periods. Both DTF analysis for CRI and SSI also shows that the two

measures are time invariant (except for one CRI item showing outside the 95% boundary of DTF chart).

Response category optimization shows that collapsing one response level may generate a better reliability statistic for BDI-II. Although collapsing one response level increases the person reliability for CRI and SSI, higher person reliability may not result in a "good" category probability curve. The small increase in reliability is less important than the scales performing in an acceptable manner.

Conclusion: Rasch analysis is a complementary and alternative method of classical test theory (CTT) for evaluating the quality of a measure. In this study, both Rasch and CTT presented similar results in term of reliability and validity. However, Rasch analysis provides more detailed information on person ability, item difficulty, targeting, and misfitting items to improve instrument design.

Keywords: Rasch analysis, Classical test theory, BDI-II depression, Parenting stress, Unidimensionality, Reliability, Differential item function, Item test function, Response category optimization

ACKNOWLEDGMENTS

This dissertation was completed with considerable guidance and assistance. I would like to express my deepest gratitude to my mentors and my family.

First, I would like to express my sincere appreciation to my advisor, Dr. Connie Kohler, who provided me unconditional support during my academic career at UAB. For me, Dr. Kohler is not only my advisor, but also a mother figure. No words can describe the amount of guidance and assistance I got from Dr. Kohler. What I learned from Dr. Kohler was not only a positive attitude toward my academic career, but also an optimistic attitude toward my life and people around me.

Second, I would like to extend my appreciation to my mentor Dr. Scott Snyder, who suggested the topic of my dissertation. Since then, Dr. Snyder sacrificed his valuable time and efforts to train me for mastering the knowledge and the skills in this special field. His consistent encouragement and suggestions prompted me to explore and enhance my research ability and interest in this study.

Third, I would like to thank my chair Dr. Robin Lanzi for her valuable suggestions on my dissertation. Special thanks to Dr. Lanzi and Dr. Kristi Guest for allowing me access to the Parenting for the first Time data. Also, I would like to express my gratitude to Dr. Peter Hendricks and Dr. David Redden for their advice, time, and for serving on my committee.

Last, I would like to thank my wonderful family, especially my husband, Xiaoyong Lei, who took care of our two kids during this process. I appreciate his love, patience, and support. I could not have completed this journey without him.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table *Page*

LIST OF FIGURES

Figure *Page*

CHAPTER 1
INTRODUCTION

Overview

The purpose of this research is to demonstrate and evaluate the application of Rasch measurement modeling to existing measures used in the field of behavioral science and health. It's focus is to apply the Rasch analysis to two specific measures taken in a longitudinal study of healthy parenting and to compare the Rasch Model outcomes to those that are obtained by Classical Test Theory method (CTT), such as reliability and unidimensionality.

Measurement of health-related concepts is a critical component for understanding health problems, making proper health policy decisions, and monitoring the services of medical and health care (McDowell, 2005). Since unhealthy behavior has been cited as the cause of much illness and death, reliable and valid measurements are essential in the health behavior field to understand why people behave in healthy or unhealthy ways, and therefore, will be useful for planning and evaluating health behavior interventions and programs. For example, through estimation of population characteristics such as level of self-efficacy or health related expectancies; researchers can develop and evaluate health promotion programs that are targeted to that population.

Measurement scales in behavioral science and health must be evaluated in terms of their psychometric properties including unidimensionality, interval level measurement and invariance across samples and time. A useful measurement, (which Wright (1997) attributes to Thurstone (1931), should match certain requirements: 1) Unidimensionality: the scales measure only one attribute; 2) Linearity: measurement is a interval continuum; 3) Invariance: the continuum of the measurement will not change in different settings; 4) Sample-free calibration: the measurement will not be affected by the selected samples; 5) Test-free measurement: the individual score should not be affected by excluding questions in certain levels. The last two requirements indicate that a measure should be independent of person and estimate.

Before we apply a measure to specific data, we should have some assurance that the measure conforms to those assumptions that the scales are interval, invariant across group and time, and independent of person and estimate. These properties are particularly critical for repeated measures in longitudinal or measures tested in cross sectional research studies. However, Classic Test Theory is limited in its capacity to verify these properties. A stronger way to examine measures in the field may be to subject them to Rasch model procedures. Rasch measurement modeling is thought to be superior to CTT because it can assure us that a scale posseses these properties once it conforms to the model (Huisingh, Snyder, McGwin & Owsley, 2018).

Using data from a longitudinal study of new mothers' parenting practices, the study applies Rasch analysis to examine if two measures in the study, Beck Depression Index (BDI-II) and Parenting Stress Index (PSI), fit Thurstone's criteria. In addition, this study will also compare the Rasch analysis and CTT to find if: 1) Rasch analysis can be

an alternative to CTT when constructing and validating measures, and 2) Rasch analysis could provide more information than CTT in term of person and item estimates, and scale category optimization.

By showing the approach and results of Rasch analysis for the two measures, the study contributes to our knowledge by demonstrating how the two repeated measures function in the longitudinal study, i.e., by checking the performance of each item, and provides insights on how to revise the measures to improve the quality of the scales. Measures with test-retest stability are crucial for longitudinal studies which examine the relation between depression and/or stress and other parenting characteristics, and provide guidance for future intervention projects.

Rationale and significance

The basis of Classic Test Theory (CTT) is the idea that a true score could be the theoretical foundation for developing reliable measures (Wilson, Allen, & Li, 2006). The traditional method perceives that the raw scores obtained from an instrument (e.g. score from Likert scales) possess the characteristics of a mathematical measure (e.g. ruler) with equal interval and infinity. However, critics (e.g.Wright, 1997) argue that such raw scores cannot be treated as linear measures.

First, the ordinal scales are not infinite. Wright (1997) argued that the response format in a measure which begins at "none right" to "all right" make the scores bounded. Second, the interval between the scales may not be equal. Researchers have to be aware that we cannot assume that ordinal data are linear and use it directly for parametric

statistics (Boone, Staver, &Yale, 2014). Third, the index of latent traits can be confounded by the difficulty of the instrument itself and the ability of respondents (Wilson et al., 2006). For example, in a very difficult math test/or a math question, the person who has higher ability may not be able to score highly enough to discriminate him from the person who has low ability. The raw score bias tends to favor "questions" in the middle of a test, and the magnitude of the bias depends on the distribution of item difficulty (Wright, 1997), thus the raw score cannot be sample free and test free.

Because of the inherent drawbacks of CTT, it is necessary to reexamine the psychometric integrity of existing measures using a Rasch model, with the probability based analysis. A Rasch model estimates the expected score based on a probability function, and then it transforms the probabilities to logit for both the person and item, so that the logit transformations can be used in statistical models as linear variables. These methods can solve the inherent problems of CTT discussed above and assure that the developed scale have: 1) Equal interval and linear measures; 2) Comparable person and items calibrations; 3) Invariance across group and time.

Depression highly correlates with negative maternal behavior (Lovejoy, Graczyk, O'Hare, & Neuman, 2000). It is also a predictor of parenting self efficacy (Smith, T., 2015). The Beck Depression Index (BDI-II) is a commonly used measure for assessing the severity of depression in clinic and research settings (Siegert, Tennant, & Tuner-Strokes, 2010). Because of its sound psychometric reliability, the BDI-II is regarded as a cost effective instrument, and applied broadly worldwide (Wang & Gorenstein, 2013).

BDI-II has been validated repeatedly, including with Rasch analysis. For example, Siegert, Tennant, & Turner-Stokes (2010) applied Rasch model in a

neurological rehabilitation sample to examine item fit, DIF, and using Rasch principle analysis to assess unidimensionality. Lerdal, Kottorp, Gay, Grov, & Lee (2014) examined the item fit and DIF in stroke survivors. Lambert et al. (2015) use Rasch analysis to compare the equality of cut-off point in several depression measures (including BDI-II).

However, to the researcher's knowledge, Rasch analysis has not been applied to validate the reliability and validity of BDI-II with a sample of mothers with newborn babies. In addition, the researcher has not found a study which conducted the analysis to address the measurement time invariance in a longitudinal study and optimize response categories for BDI-II.

Parenting stress was linked to negative parenting characteristics, unhealthy parenting styles, and use of harsh discipline (Haskett, Ahern, Ward, & Allaire, 2006). It has a negative correlation with parenting self efficacy (Jackson, 2000), which is the best predictor of parenting styles (Sanders & Woolley, 2005). The Parenting stress index-Short Form (PSI-SF), designed by Abidin (1990), is one of the most common and widely used instruments of its type (Lee, Gopalan, & Harrington, 2016), which has been applied in variety of settings (Haskett, Ahern, ward, & Allaire, 2006).

PSI-SF has been applied for a variety of research settings, and was validated among different samples using CTT (Haskett, Ahern, ward, & Allaire, 2006). However, limited research has examined its psychometric characteristics in the population of mothers with a new baby. The author found only one research (Puma, 2007) , which applied Rasch analysis for this specific population. This study examined Rasch assumptions, and evaluated structure invariance by using confirmative factor analysis,

and evaluated the properties of the scalling categories of PSI. However, Puma's study did not apply Principle component analysis of Residual, the method of Rasch anlayais for evaluating unidimensionality. In addition, the instrument were administrated in 14, 24, and 36 month after the babies were born. Therefore, this study conducted a supplemental investigation to reexamine this measure at prenatal or 6 and 12 month after the babies were born.

It is a longstanding debate about the optimal number of response options to maximize reliability and validity (Jones & Loe, 2013). On one hand, the number of categories should generate enough variance for acquiring good reliability. On the other hand, more scale categories may increase response burden, affect respondent's cognitive motivation, and further increase the response errors (Alwin & Krosnick, 1991). For a rating scale instrument, five to seven categories are commonly used to evaluate psychometric characteristics (Lietz, 2010), and instruments with more than 10 categories may not be as effective as an instrument with fewer response categories (Jones & Loe, 2013).

One of the advantages of a Rasch model approach over CTT is that it provides a way to help researchers to optimize or minimize the number of response categories without affecting the reliability of the measure. This approach is really meaningful not only because it can save much administration cost, but also because it can minimize response burden, and therefore improve the quality of responses and the measure.

In summary, this study provided a comprehensive Rasch analysis to validate the two parenting related measures. Specially, the study examined the measure variance across group and time for the two measures in the longitudinal samples. The results

obtained from this study are especially meaningful for improving the quality of repeated measures in longitudinal projects, and thus provide assurance with regard to reusing the two measures repeatedly in research, which evaluates the relation between depression/stress and other parenting variables.

Statement of Purpose

Overall this study was conducted to validated two repeated measures (BDI-II and PSI) in a longitudinal project among mothers with new babies. This study has three primary research aims.

Aim 1: To check if the two measures meet the assumption of the Rasch model

1) Will Beck Depression Inventory-II and Parenting Stress measures comply with the unidimensionality and local independence assumption?
2) Will item estimates for the two scales remain invariant across demographic characteristics (e.g. age) and/or over time?

Aim 2: To check if Rasch model and CTT yield equivalent results when examine the unidemensionality and reliability.

1) What is the difference between Rasch principal component analysis and Principal component analysis when examining the domains of the two measures?
2) Is there a meaningful difference between Rasch analysis and CTT in their abilities to assess reliability of these measures in this sample?

Aim 3: Optimize the response categories for the two measures using Rasch analysis

CHAPTER 2

REVIEW OF RELATED LITERATURE

Overview

This chapter provides a comprehensive review of the literature relevant to the present study. First, a brief introduction to Rasch model and its basic statistics, second, a review of the family members of Rasch model based on different score formats (e,g. dichotomous and polytomous), third, the advantages of Rasch model over CTT are discussed, and fourth, psychometric analysis of BDI-II and PSI are reviewed.

Introduction of Rasch model

Basics of Rasch model

The fundamental difference between CTT and Rasch is that Rasch model is a probability model, it estimates the probability of a respondent responding "correctly" to an item. Meanwhile, Rasch model estimates two parameters: ability and difficulty.

Ability is the latent trait of a respondent. As the Rasch model was established for education tests, the construct 'ability' originally referred to how well an examinee performs on an academic exam, for example, a math exam. In a math exam, the student who gets more correct answers or gets more "1"(if "1" represents "Correct" in a dataset)

has a higher ability in math than a student who gets fewer correct answers (Fewer "1", or more "0", if "0" represents "Incorrect").

The test format for the math exam above is the simplest scoring format: dichotomous scoring in which each item is dichotomously scored, and a correct answer will be coded as"1" and incorrect answer will be coded as "0". It can be easily extended to an attitude measure with dichotomous item format, such as the following items borrowed from a dental health attitude study (Yildiz & Dogan, 2011):

I worry about color of my teeth.

I have noticed some white sticky deposits on my teeth.

I brush each of my teeth carefully.

I often check my teeth in a mirror after brushing.

I have used a dye to see how clean my teeth are.

For the items above, the responders would check "agree" or "disagree". An "Agree" response gets one point ("1"), and a "Disagree" response gets zero points ("0"). If a respondent pays more attention to dental health, he/she can surely acquire more points from the answers, and his/her "ability" to get answers "correct" (i.e. more "1") is higher.

In a math exam, students answer all questions with their own math ability. Some questions are very difficult, therefore few students can answer them correctly, while some questions are very easy, so most student can give right answers. Therefore, easy questions get more points from students, and difficult questions get less points from students. The difficulty of questions can be ordered in this way. There are questions that are too difficult or too easy, that no one can answer, or everyone can answer. A teacher may be

concerned that those outlying questions may not be able to discriminate a student's ability, and the teacher may need to remove them from the pool of questions.

Back to the items for dental health attitude, if an item gets a lot of "Agree" from a group of respondents, we may say the "difficulty" of the item is low, and an item which gets a lot of "disagree" has higher "difficulty". A researcher certainly does not want an item to have a very high/low difficulty, as this indicates the item cannot contribute much to discriminating his/her respondents, and therefore affecting the reliability of his/her measure.

Rasch probability model

The fundamental difference between CTT and Rasch model is that CTT uses true score to evaluate a construct, while Rasch model is based on logistic distribution to estimate the probability of a respondent's "success" on one item. For dichotomous test format, the "success" means the chance of coding "1".

Suppose the test format is dichotomous, the equation of the Rasch model can be written as below (as Embretson & Reise, 2000) to estimate the probability of "success (1)":

$$P = (X_{is} = 1|\theta_s, \beta_i) = \frac{e^{(\theta_s - \beta_i)}}{1+e^{(\theta_s - \beta_i)}} \qquad (2.1)$$

s: person s

i: item *i*

ɵ: trait level (ability)

β: item difficulty

The equation is exactly the same as the traditional logistic model if w_{is} represents $\theta_s - \beta_i$ (as Embretson & Reise, 2000):

$$P = (X_{is} = 1|w_{is}) = \frac{e^{w_{is}}}{1+e^{w_{is}}}$$

The person parameter (ability) also refers to trait level, as it is the estimated latent trait for respondents. From the equation above, we can see that the Rasch model only need estimate one parameter, although in application, both trait level and item parameter (difficulty) are unknown and should be estimated. The Rasch model is also called one parameter Item Response Model, as it includes estimates of only item difficulty in the model in relation to latent trait.

Item characteristic curve

Item characteristic curve (ICC) is the basic feature for item response theory (including the Rasch model). The estimation of model parameters and a construct's evaluation depend on the curve.

In Rasch model or item response theory, each item has its own item characteristic curve (ICC). A typical ICC in Rasch would look like "Figure 1" below where "ability" increases along the trait continuum:

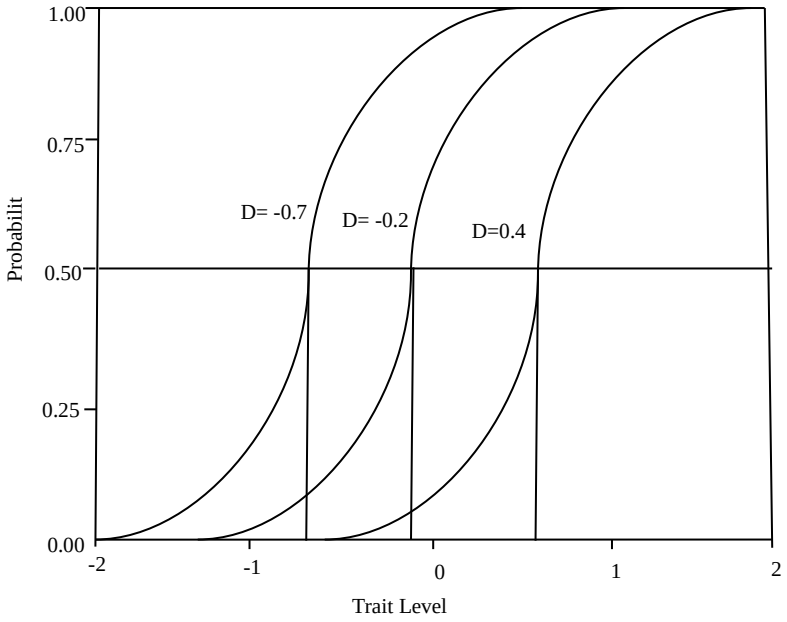


Figure 1. Item characteristic curves

Baker (2001) briefly introduces the procedure of how to find the ICC that best fits the observed proportions of the correct responses: first, the observed probability of the correct responses for each respondent will be plotted along with the ability axis for each item, then the model applies the maximum likelihood function to find a fitted curve to best describe the observed proportions. The maximum likelihood estimation utilizes a complicated iterative procedure to find the optimal estimates. This complicated procedure would be achieved easily by computer. By applying the same procedure, all items in a measure can acquire an ICC. As the Rasch model assumes that all items have the same ability to discriminate respondents, the shapes of all ICC will be same (see Figure 1), but the positions in the trait continuum are different. In Figure 1, item 3(b=0.4) can discriminate respondents who have a higher ability/trait level, while item 1(b=-0.7) is able to discriminate respondents with a lower ability/trait level.

Item parameter estimation (Difficulty)

Item difficulty and person ability can be estimated separately (Bond & Fox, 2007), which means that we can estimate item parameter without knowing person parameters. The estimation of parameters can also be achieved by the complicated computing procedure, Maximum Likelihood Estimation. There are three popular methods for parameter estimation (more details can be seen in Embertson &Resise, 2000), Joint Maximum Likelihood (JML), Marginal Maximum Likelihood (MML), and Conditional Maximum Likelihood (CML).

MML handles the unknown trait level as the expectation of response probability; this is estimated by an expectation/maximization (EM) algorithm, an iterative procedure, which estimates the optimal expectation by treating the observed data as a sample from a population. Then an iterative searching process will be run to find optimal item parameters based on the probability expectations, which can maximize the likelihood of "correct" answers.

Both JML and CML treat person parameters as "fixed" values. JML pre-arranges trait level to respondents and then use an iterative procedure to estimate item and person parameters. During the process, the provisional trait levels and item parameters are improved sequentially until the procedure finds the most optimal estimate for the item.

In CML, trait levels are first "fixed" by the response probability, and then the optimal item parameters are searched iteratively. As CML requires sufficient statistics available to estimate trait level, it can be only used in the Rasch model. Because the Rasch model only contains one parameter (difficulty), therefore, total score or response probability provides sufficient information for estimating trait level (Embretson & Reise, 2000).

CML is more efficient to apply for the Rasch model. Eggen (2000) compared the efficiency and loss of information among the three estimation methods, he claimed that CML may lose very small amount of information comparing with MML and JML, but the efficiencies are larger than 93%. In his study, if the test has 20 or more items, the efficiencies will be larger than 99%. Meanwhile, the CML has an advantage that there is no distribution assumption on trait levels (Embretson & Reise, 2000).

Person parameter estimation (Ability)

In CML, person parameters (ability) are estimated after the estimation of item parameters by maximum likelihood scoring. The probability of item response in Figure 1 will be sum up to likelihood as below (as Thompson, 2009).

$$\boldsymbol{L\,u|\theta_j = \prod_{i=1}^{n} P_{ij}^{u_{ij}} Q_{ij}^{1-u_{ij}}} \qquad (2.2)$$

u: is a response vector ($u_{j1,}u_{j2,}\ldots,u_{jn}$)

i: item i

j: person j

L: Likelihood for person *j* to have an item response

P: probability of "correct" answer on *i*th item

Q: probability of 'incorrect' answer on *i*th item

Suppose a person has an item response on 3 items (1, 0, 1), by the estimation of ICCs, the conditional likelihood would be multiplied as $P_1(\theta)Q_2(\theta)\,P_3(\theta)$. Thompson (2009) described how to estimate a respondent's ability (trait) based on his or her response pattern: First, the product of conditional probabilities $P_1(\theta)Q_2(\theta)\,P_3(\theta)$ will be calculated over the varying hypothetical ability (trait) level; then the likelihood function

curve (similar as normal distribution curve) will be depicted which spreads along the continuum of trait level; This curve has both left and right tail approaching to 0, and a peak(maximum likelihood); Eventually the ability (trait) corresponds to the peak (maximum likelihood) will be selected as the ability level of the respondent.

However, likelihood function has a problem of not being able to estimate the trait level for respondents who endorse or not endorse all items. To solve the problem, log likelihood transformation is applied; the log likelihood uses log function to transform *L* in equation 2.2, so that the log likelihood becomes an addition function of ICCs instead of multiplication, then an iterative procedure such as Newton-Raphson is conducted to find the most optimal estimate of trait level (Embretson & Reise, 2000).

Fit statistics

Item fit and person fit are the fit statistics in Rasch model to evaluate "model fit", in other word, to find outlier items or outlier persons. Misfit is the specific term to describe outliers (Yu, 2013), which means that an observation is far away from what we expect from model estimation. An item misfit may happen when an item does not measure the concept we are trying to measure, for example, an item of European history is included in an American history (Yu, 2013). Person misfit reveals incongruent response pattern with the exception of our models, it may be caused by cheating or guessing.

The purposes of fit statistics calculation are to (Reise, 1990): 1) verify item observation complying to model estimation; 2) identify respondents with incongruent response pattern from model. Four kinds of Fit statistics were proposed, two are chi-

square fit statistics (Wright & Panchapakesan, 1969), and the other two are their transformation (Wright & Panchapakesan, 1969).

Simply specifying, fit statistics evaluate the discrepancy between observation and model estimation. The discrepancy can be measured by standardized residual (as Wang & Chen, 2005):

$$Z_{ni} = Y_{ni}/\sqrt{W_{ni}}$$

Y_{ni} is the discrepancy between true score and expected score

W_{ni} is the variance of Y_{ni}

Unweighted mean square error (MNSQ) would be computed separately for an item or a person (noted as Embretson & Reise, 2000):

Item fit=$\sum \frac{{Z_{ni}}^2}{I}$

Person fit=$\sum \frac{{Z_{ni}}^2}{N}$

The unweighted fit statistic is also called outfit mean squared (OUT.MSQ).

Unweighted fit statistics are sensitive to abnormal responses (Wang &Chen, 2005), for example, cheating and guessing (Linacre, 2002b). Wright and Panchapakesan (1969) proposed weighted MNSQ by adding weights to items or respondents. The weighted MNSQ is also called Infit statistics. Infit statistics weigh MNSQ by their own variances, as extreme observations have larger variance than targeted observations. The calculation of Infit statistics put more weights on the observations that is not extreme and less weights on extreme observations (Smith, Conrad, Chang, & Plazza, 2002).

Suppose we are creating a histogram to find the distribution of respondents' trait level, persons who are in the extreme left or right sides will get less weighting than persons in the middle, as we care more about persons without extreme scores (Yu, 2013).

Therefore, Infit statistics are more sensitive to response patterns which are not congruent to pattern expectation (Linacre, 2002b), for example, an item which may affects construct validity.

Both Infit.MSQ and outfit.MSQ can be transformed to *t* statistics with an approximate unit normal distribution (Smith, et al., 2002). It is transformed by Wilson-Hilferty cube root transformation (Smith, et al., 2002) to diminish the effect of sample size. They are called infit and outfit standardized residual (Infit.MS and Outfit.MS). One of the advantages to use *t* statistics is to compare the extent of agreement between Infit and Outfit mean square (Yu, 2013).

If observed data conform to the model, the mean square fit statistics (infit and outfit) should be near 1, and the *t* statistics should be near 0 with a standard deviation near 1 (Bond & Fox, 2007). MSQ less than 1.0 or *t* statistics less than 0 indicates observations are too predictable that there is less variation than modeled (all easy item correct or all difficult incorrect (Guttman-style response), while MSQ> 1.0 or *t* statistics >0 indicates unpredictability (Bond & Fox, 2007; Linacre, 2002b; Yu, 2013). Bond and Fox (2007) listed the guideline for misfit: MSQ>1.3 or *t* statistics>2.0 means underfit, while MSQ<0.75 or <-2.0 means overfit. Misfit of Outfit can help us to diagnose outliers (overfit for imputed response, and underfit for guessing and careless mistakes), while misfit of Infit diagnose s incongruent response patterns, such as overfit for Guttman pattern and underfit for alternative construct (Linacre, 2002b).

Unidimensionality

Unidimensionality is the fundamental requirement for Rasch application. The reasons for test unidimensionality are (Smith, E., 2002): 1) It is generally agreed that

items in a measure should evaluate one attribute; 2) multidimensionality will bias item and person estimates for both true score theory and item response theory.

Factor analysis is a commonly used method to detect multidimensionality in CTT; however, it has some problems. First, factor analysis is a linear model, it assumes that data are normal distributed, this assumption cannot be held when applying the Rasch model (Smith, E., 2002), as the Rasch model asserts that only log odds transformations can be treated as normal distribution (Wright, 1996). Second, Wright (1996) pointed out that factor size and loadings may not be reproduced when applying the measure in a new sample; it will lead to factorial invariance and contradictory re-analysis. In addition, several studies have found that linear principle component analysis will overestimate the number of factors of a measurement (Smith, E., 2002).

Wright (1996) proposed to use fit statistics and Principle Component Analysis of Standardized Residual (RPCA) to evaluate unidimensionality. It is also called Rasch Factor analysis (Bond & Fox, 2007) by some researchers. E. Smith (2002) compared the effectiveness of RPCA and Principle Component Analysis (PCA) of raw data. The study showed that, if the components are not highly correlated, PCA was better to detect multidimensionality, while Rasch's fit statistic was better to detect the departure from unidimensionality when majority of items contribute to one component.

Wright (1996) suggested to identify the first factor of a measure by examining the t-transformed standardized Infit and Outfit statistics (Z-MSQ). Items with misfit statistics (>=2) reveal that these items may be belong to second factor, and need further investigation. When it is necessary, an iterative procedure can be applied to investigate the misfit items by analyzing those subscales to identify the second factor.

Applying factor analysis with Rasch standardized residuals is based on the assumption that Rasch residuals would represent random noise and independent of each other if the data conform to the model (Smith, E., 2002). Therefore, the purpose of RPCA is to test the hypothesis that all variance is originated from one latent variable; any existence of substantive common factors would indicate the departure from unidimensionality (Linacre, 1998).

A Rasch fit analysis in conjunction with RPCA can provide more insight to find the items that contribute to multidimensionality (Smith, E., 2002). A procedure suggested (Smith, E., 2002) to apply Rasch factor analysis includes: 1) use traditional statistics to identify problematic items, such as reversed coding; 2) examine fit statistics to identify misfit items and person; 3) iteratively conduct RPCA following the above steps.

Local independence

Local independence of items is another assumption of the Rasch model, which requires that items in a test should not be related to each other (Baghaei, 1998). That is, the success or failure on any items should not be affected by the success or failure of any other items (Bond & Fox, 2007). Local independence not only includes unidimensionality, but also goes beyond it (Wright, 1996). For example, suppose there are two identical math questions in an exam, this will not affect dimensionality. However, examinees are expected to answer correctly or incorrectly for both of the two questions, in other words, the probability of failure or success of one question conditioned on another, which violate the local independence assumption (Wright, 1996).

The problems of local independence not only were addressed in the Rasch model, but also in CTT (Baghaei, 2008). As Baghaei (1998) pointed out, local item dependence (LID) will affect the accuracy when measuring a construct. First, the group of LID items may act as a dimension to disturb the dimension identification of a measure (even if LID effect is small, the measures still partially reflect existence of LID). Second, LID will result in small standard errors, which gives a fake impression of the precision and inflates reliability. Therefore, lack of local independence can be a major threat to construct validity (Linacre, 2009).

Reliability

The Rasch model provides two kinds of reliability estimates: Person reliability and Item reliability.

Person reliability can be interpreted as traditional Cronbach's alpha (Boone, Staver, & Yale, 2014), in which, values closer to 1 indicates more internal consistency. The logic behind the calculation of the two reliabilities is similar (Clauser & Linacte, 1999). The basic rationale for both methods is: observed variance = true variance + error variance, and reliability is the ratio between true measure variance and observed variance (Clauser & Linacte, 1999).

Nonlinear raw score are used to calculate Cronbach's alpha, and the formula for calculating is (adopted from Clauser & Linacte, 1999):

$$R_\alpha = (\frac{k}{k-1})(1 - \frac{\sum \sigma_i^2}{\sigma^2})$$

k is the observations numbers, and σ_i^2 is the raw score variance for *ith item* across examinees, and σ^2 is the raw score variance across examinees.

Rasch model uses its linear measures to acquire reliability estimation. The formula is (Clauser & Linacte, 1999):

$$R_R = \frac{\sum(Measure\ Standard\ Error)^2/N}{Variance\ of\ Observed\ Measures}$$

Linacre (1997) pointed out that Cronbach's alpha "is an index of the repeatability of raw scores, misinterpreted as linear measures". It usually overestimates the reliability of a measure and gives tester an illusion that it is test independent. In fact, Rasch reliability is more conservative than Cronbach's alpha, and it has real test independence.

Person reliability depends on (Linacre, 2012):1) Sample ability variance. 2) Length of test (and rating scale length); 3) Number of categories per item; 4) Sample-item targeting. Wide ability range, longer tests, more item categories, and better item targeting improve person reliability (Linacre, 2012).

Item reliability indicates the stability of item placement when applying a measurement to another sample with same sample size (Bond & Fox, 2007). Low item reliability reveals that the sample size is not big enough to estimate precisely the item's location on the measurement (Linacre, 2012). Meanwhile, it depends primarily on Item difficulty and person sample size (Linacre, 2012), that is, wide difficulty range and large person sample size leads to high item reliability.

Separation

The item and person separation index in Rasch analysis consists of statistics to evaluate the precision of a measurement and distinguish item difficulties or person abilities. Person separation indicates how well the measure can separate respondents, while item separation tells us how a sample can separate the items (Wright & Stone,

1999). Separation is the ratio of the person or item's adjusted standard deviation (Adjusted S.D) to the square-root of the average error variance (RMSE) (Linacre, 2012). Adjusted S.D is the square root of "true" variance. The relation between SEPARATION and RELIABILITY is (Linacre, 2012):

$\text{RELIABILITY} = \text{SEPARATION}^2/(1+\text{SEPARATION}^2)$ or

$\text{SEPARATION} = (\text{RELIABILITY}/(1-\text{RELIABILITY}))^{0.5}$

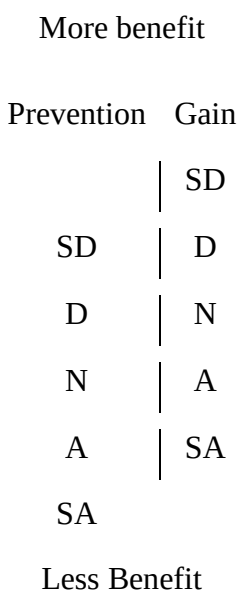
Fisher(1992) believed that a person separation index >=1.5 indicates an acceptable level of separation (reliability>=0.7), while some researchers (e.g. Duncan, Lai, Bode, Perera, & De la Rosa, 2003; Las Hayas, Quintana, Padierna, Bilbao, & Muñoz, 2010) use 2.0 as threshold, as a value of 2.0 is comparable to a reliability of 0.80.

Polytomous Rasch model

The previous section introduced the basics of Rasch model by using the original dichotomous Rasch model for illustration. However, for instrument building or validation, the most frequently used response formats are ordered categories, in which integers are assigned to categories successively. A popular format to assess attitude is the Likert scale, and one of its typical expression can be: SD (strongly disagree), D (disagree), N (neutral), A (agree), and SA (strongly agree).

In CTT, we assume that the interval between two ordinal item categories is equal. However, the Rasch model does not support this assumption. The position of each category of items along the logit continuum may not be able to align the response options with each other. For example, suppose there are two items about perceived benefits of breast cancer screening (questions adopted from Frankenfield, 2009): 1) Doing breast

self-exams prevents future problems for me. 2) I have a lot to gain by doing breast self-exams. The position of categories may look like this:



From the example above, we can see that the two items may not parallel each other, for example, "SD" in the Prevention item is actually equal to the level of "D" in Gain. The actual position of each item and its categories will be depicted by polytomous Rasch model.

Partial credit model (PCM)

Masters (1988) proposed a partial credit model for measures with more than two categories. This partial credit model assumes that people choose a response category by taking successive steps. These procedures may be like the process of solving a math problem, for example, $(1.5/0.3\text{-}2)^2$. The first step is to find $1.5/0.3 = 5$, the second step is $5\text{-}3 = 2$, and the third step is $2^2 = 4$. In the math problem, each step has its own difficulty.

How far the person can go (i.e. which step he/she can finish) depends on the difficulty of each step).

Similarly, for an attitude item with ordered categories such as SD (strongly disagree), D (disagree), N (neutral), A (agree), and SA (strongly agree). There are three steps:

0----------1----------2----------3----------4

SD D N A SA

step1 step2 step3 step4

If a person chooses A (agree), he/she has already considered between SD and D, D and N, N and A, A and SA. He/she favors D over SD, N over D, A over N, but rejects SA. The probability of choosing one of the response categories is expressed as (Masters, 1982, 1988):

$$\pi_{xni} = \frac{exp \sum_{j=0}^{x}(\beta_n - \delta_{ij})}{\sum_{k=0}^{mi} exp \sum_{j=0}^{x}(\beta_n - \delta_{ij})} \quad x = 0, 1, \ldots m_i$$

where $exp \sum_{j=0}^{x}(\beta_n - \delta_{ij}) \equiv 0$.

In the partial credit model, ICC may be depicted similar as dichotomous model for each step. However, these probability curves are based on a given step being conditional on the previous step. Suppose that there are three response categories (i.e. two steps of response), there two ICC curves for $\frac{\pi_{2ni}}{\pi_{1ni}+\pi_{2ni}}$ and $\frac{\pi_{3ni}}{\pi_{2ni}+\pi_{2ni}}$ would be estimated accordingly. The expected probability of each category, π_{1ni}, π_{2ni}, and π_{3ni} would be transformed from the two ICC curves previously described, creating three Category probability curves. The intersections between each two response category curves are σ_{i1} and σ_{i2}, difficulties for the two steps. A person with an ability of less than σ_{i1} is more

likely to choose category 1, a person with ability between σ_{i1} and σ_{i2} is more likely to choose category 2, and a person with ability larger than σ_{i2} is more likely to choose category 2.

Rating scales model (RSM)

Although the partial credit model was established originally as an extension of the rating scales model proposed by (Andrich, 1978), the rating scale model is actually a special case of the partial credit model (Masters, 1982). In the rating scale model, the difficulties of steps do not vary greatly. The σ_{ij} could be expressing as $\sigma_{ij} = \sigma_i + \tau_j$ and the partial credit model becomes:

$$\pi_{xni} = \frac{exp\sum_{j=0}^{x}[\beta_n - (\delta_i + \tau_j)]}{\sum_{k=0}^{mi} exp\sum_{j=0}^{x}[\beta_n - (\delta_i + \tau_j)]} \quad x = 0, 1, \ldots m_i$$

Therefore, in the rating scale model, the estimated parameter is β_n for each person, σ_i for each item, and τ_1, τ_2,.., τ_m for m+1 rating categories (Masters, 1982). Similarly as with the partial credit model, category probability curves can be produced.

By using Category probability curves, both the partial credit model and the rating scale model can use the same formula to acquire expected scores:

$$E(x) = \sum_x xp_x(\theta)$$

Suppose a person has estimated probabilities of 0.18, 0.44, 0.23 for checking response categories 0, 1, 2 respectively, the expected score for this respondent would be 0×0.18+1×0.44+2×0.38=1.2.

Difference between PCM and RSM

Although RSM can be derived from PCM, they are different on a crucial aspect: RSM assumes that the relative difficulties of the steps vary little across items (τ_1, τ_2,.., τ_m are the same for all item response steps), while PCM does not have this assumption , which means, in PCM, τ_1, τ_2,.., τ_m differs across different items (Embretson & Reise, 2000).

In other word, in RSM, all items share the same scale structure, while in PCM, each model has its own scale structure (Linacre, 2000), which leads to an increased number of parameters being estimated. Although more parameters imply that a model fits the data better, it will affect its ability to estimate and communicate (Linacre, 2000).

First, a more complex model may have no meaning for inference. Wright (1998) illustrated that a model perfectly fit to the selected sample data provides an unreasonable estimation for new data compared to a model with fewer parameters. In addition, the estimation is not robust for PCM if there are less than ten observations in a category, although this is usually not a problem for RSM (Linacre, 2000).

Second, it is easier to communicate with others when explaining a model with the same structure across items than making them imagine that each item has its unique structure (Wright, 1998). In fact, in an instrument designed with the same scale structure across items, an item with its own structure may be aberrant (e.g. because of a wording problem) and need to be excluded from the instrument (Linacre, 2000).

Overall, as Linacre (2000) suggested: if items have the same rating scale, we can use RSM, if items have a different rating scale, PCM is preferred.

Advantages of Rasch model over Classical Test Theory

The fundamental difference between CTT and Rasch analysis is that CTT relies on the raw score to evaluate the variance between variances of items and total variance among observations. Instead, the Rasch model estimates the expected score based on a probability function, and then it transforms the probabilities to logit, so that the logit transformations can be used in statistical models as linear variables.

These methods can solve several inherent problems of CTT (see Snyder & Sheehan, 1992; Zhu, Timm, & Ainsworth, 2001; Bond & Fox, 2007): 1) CTT makes the untenable assumption that items with rating scales (e.g. Likert scale) are internally equal; 2) developers usually set up category numbers by prior knowledge instead of empirical determination; 3) items and respondents cannot be calibrated along the same continuum; 4) Item difficulty and subject ability are not mutual independent with each other; 4) CTT is often sample and item dependent; 5) missing data will be deleted when applying CTT.

Equal interval and linear measures

Researchers may imagine that raw test data can be used to measure abilities as a meter-stick, but actually this meter-stick may not have equal intervals for measuring (Boone et al., 2014). Bond & Fox (2007) provided an example to illustrate the argument (see Bond & Fox 2007, p24). In their example, raw percentage and log odds of relative abilities of individuals are shown in a figure. The raw percentage is the rate that a person answered the questions correctly, and it seems that, by adding extra points, the leap in ability from 45% to 55% is equal to the leap from 85% to 95%, or 5% to 15%. However,

when checking the estimated log odds (Rasch ability), they do not have such an equal gap. The major problem in CTT is that we can only infer the order of persons or items from the raw score, but cannot acquire accurate relative distances between them (Bond and Fox, 2007).

Calibrate person and items

In CTT, the order of people is identified by the total score of items, but items cannot calibrate along the same continuum, therefore, we do not have enough information to evaluate the item discrimination ability. However, the Rasch model utilizes difficulty and ability to identify the relative location of persons and items. Because difficulty and ability are logit odds, they can be depicted along the same continuum. This method provides an alternative way to observe the spread of a measure.

A Wright map is one of methods to visualize the location of persons and items. The Wright map was named by Mark Wilson (2011) in honor of the contribution of Benjamin Wright to Rasch measurement. The Wright map displays all persons and items along a common vertical scale (Wilson, 2011)---logit odds with 0 as the middle point (i.e.50% chance to choose).

The Wright map tells us about both the respondents and the items a single picture (See Figure 2). The left side of the scale is the distribution of respondents, while the right side is the distribution of the items. The upper of left side represents "more able" respondents, and the lower left side represents "less able" respondents. Similarly, the upper right is for "more difficult" items, and the lower right side is for "less difficult" items.

Therefore, the ability of respondents in the upper left side is “better” or “smarter” than the respondents on the lower left, while the items on the upper right side are tougher than items in the lower left side. In another way, items on the upper right side are more “difficult” for respondents who are in the lower left side to answer, and items on the lower right side are “easier” for respondents who are in the upper left side to answer.

In Figure 2, we can see that samples are near a normal distribution, and there are quite amount of items are below the average ability of persons, which indicates that this measure is a little bit “easier” for the respondents, and more “difficult” items should be put in the measure to discriminate the “more able” respondents better.

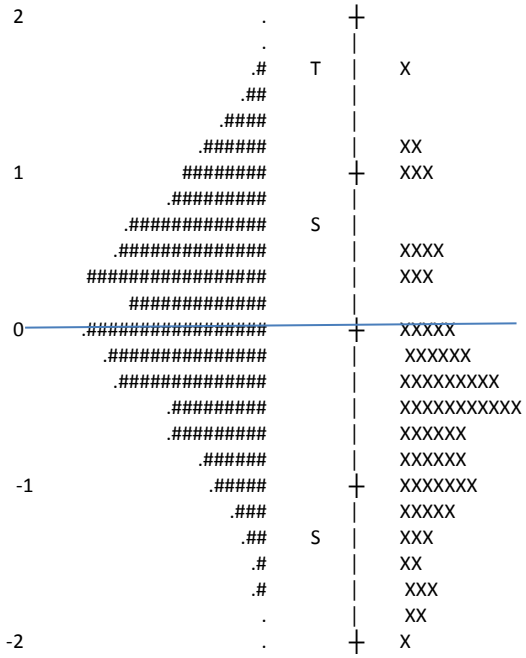


Figure 2. Example of Wright Map

Determine response categories and minimum item sets

It is a longstanding debate about the optimal number of response options to maximize reliability and validity (Jones & Loe, 2013). For a rating scale instrument, five

to seven categories are commonly used to evaluate psychometric characteristics (Lietz, 2010). Researchers believe that an increased number of categories will improve reliability and validity of an instrument, although some claimed that instruments with more categories (>10) may not be as effective as instrument with fewer response categories (Jones & Loe, 2013).

In CTT, there is neither a statistical method nor an index to evaluate the effectiveness of response structure other than the correlation index, such as Cronbach alpha and item-total correlation index. However, the Rasch polytomous models have an inherent parameter τ_m, a parameter for calculating the probability of steps between two categories. Andrich (1996) provided an example to show, by checking item ICC and τ_m, how a five-category item was collapsed down to three categories for a better fitting the model.

The separation index is also used for checking item structure. A higher item separation index means better categorization, and higher person separation can distinguish among respondents better (Zhu et al., 2001). Zhu et al. (2001) examined the response categories in an instrument to measure exercise perseverance and barriers. By comparing item and person separation indices, he found that three categories had higher item- and person-separation than the five categories originally designed. Then, the three-category method was chosen as the optimal categorization for further construct analysis in his study.

The separation index is also a reference to determine the length of an instrument. In CTT, the item selection procedure attempts to retain the items that can capture the most variability in the sample. However, there is not an explicit method for selecting

optimal items (Mallinson, Stelmack, & Velozo, 2004). The separation index is very helpful to evaluate how well an instrument can separate respondents and distinguish items, as it reflects the precision of a test. One advantage of the separation index is that it remains invariant across tests (Mallinson et al., 2004).

Invariance

Invariance is a fundamental aspect of measurements, and researchers are seeking a measures which is sample variant and item invariant (Engelhard, 1989). Sample invariant calibration of items means that the estimated locations of items on an attitude measure remain unchanged across subgroups of respondents and items, while item invariant measurement of individuals means that we can acquire invariant person ability estimates no matter which items are chosen from item bank (Engelhard, 1989). It is a unique feature of Rasch model that a person's ability is invariant with respect to a specific examination, and item difficulty is invariant with respect to a specific sample (Bond & Fox, 2007).

Baker (2001) illustrated how a measure can be group invariant and item invariant in his book. Suppose we draw two groups with different abilities from a sample, we should acquire same item parameters for the two groups, as item parameter is independent of the ability of respondents, this is group invariant. For item invariant, suppose we draw two set of items with different average difficulty, and apply them to respondents, we should obtain same ability parameters for each respondent in the two tests.

If an item cannot retain group invariance, it is not appropriate and should be excluded from the measure (Baker, 2001). The analysis of the group invariance is called Differential item Function (DIF), and it is a useful technique for analyzing data of measurement (Boone et al., 2014). DIF reflects the difference of item response function across groups, which means that, in same ability level, the probability of choosing item response is different across groups. If an item displays variance across groups, it violates the unidimensionality assumption and will affect the construct validity (Tennant et al., 2004).

The following is a simplified Wright map to illustrate what is DIF. It is similar as the example provided by (Boone et al., 2014), which depicted the item positions across gender group.

Female	Males
| Q1 Q4	
| Q3 Q6 Q7	
|Q9 Q10	|Q1
| Q5 Q8 Q2	| Q3 Q6 Q7
	|Q9 Q10
	|Q4
	|Q5 Q8 Q2

Although female may tend to give more “difficult” response than male, this difference does not generate DIF, because the order of items shifts down in a similar

distance except for item 4. Item 4 shows different relative positions across gender group, it is the sign of DIF, and item 4 needs further investigation.

A practical method (Tennant et al., 2004) to detect DIF statistically is to apply two-way ANOVA for comparing the standardized residual of the observed scores among respondents. This method will detect two kinds of DIFs: uniform and non-uniform DIF. Uniform DIF is the constant difference between groups on the item function, while non-uniform DIF is the difference across the trait. Uniform DIF occurs when, for example, males consistently score higher than female on an item, while nonuniform DIF occurs when female score a higher response to low level items, but score a lower response to high level items (Amin et al., 2012).

Another challenge for researchers who attempt to develop instruments for longitudinal study is that the instrument may not assess the same domain in different time point of administration. The investigation of invariance of items across time will allow researchers to interpret and compare respondent's latent traits in a meaningful way (Brown, 2016). A Rasch model can be used to assess item invariance and equivalence over time. For example, Brown (2016) applied a Rasch model to assess the time invariance of a WURSS-21 scale, and found strong evidence to support the assumption of invariance.

Missing data

A challenge of evaluating an attitude measure is the presence of missing data (Hohensinn & Kubinger, 2011). There are several non-Rasch techniques to deal with missing data, however, as Boone et al. (2014) pointed out, those traditional techniques

are questionable: first, an easy way to confront missing data may be throwing out the data. If the data sample is large, and missing data are random, this practice may not affect the evaluation of a measurement. However, if the dataset is small and the missing data are not random, it will affect the accuracy of evaluation and bias the results. Second, some researchers may use a person's raw mean or item mean from an ordinal rating to replace the missing data, based on the untenable assumption that the scale has equal intervals and all items share the same weight. For a dichotomous scale, studies have shown (Hohensinn & Kubinger, 2011; Shin, 2009) that treating missing data as incorrect answers will leads to more bias than treating them as un-administered.

The Rasch model offers a way to solve the problem. The Rasch model is very robust when facing missing data; it can estimate difficulty and ability utilizing an incomplete data matrix, although it may affect the precision of estimation (Bond & Fox, 2007). This is because the Rasch model is a sample and item independent measurement, it does not require missing data being imputed or omitted, and it can compute expected values for missing observation without bias (Linacre, 2012).

Measures

Beck Depression Inventory-II (BDI-II)

Depression highly correlates with negative maternal behavior (Lovejoy, Graczyk, O'Hare, & Neuman, 2000). It is also a predictor of parenting self-efficacy (Smith,T. 2015). BDI-II is a commonly used measure for assessing the severity of depression (Siegert, et al., 2010), with wide use in clinic and research settings. One of the reasons for its popularity is that it has robust reliability and validity.

Wang and Gorenstein (2013) conducted a thorough literature review on the application of BDI-II, and reviewed 118 articles in variety of countries and samples. Among the articles which reported a reliability coefficient, the coefficient ranged from 0.73 to 0.96, and most were around 0.90. In addition, most of the articles reported similar factor structures for this measure. Because of its sound psychometric reliability, the BDI-II is regarded as a cost effective instrument, and applied broadly worldwide (Wang & Gorenstein, 2013).

However, the factor structure detected by CTT studies may be controversial from the studies of Rasch analysis. Wang and Gorenstein (2013) found that the articles they reviewed reported 2 or 3 dimensions for BDI-II, while several Rasch researchers claimed that BDI-II is unidimensional. For example, Siegert, Tennant, and Turner-Strokes (2010) examined BDI-II in a neurological rehabilitation sample, and concluded that the BDI-II demonstrated unidimensionality with several misfit items: Crying, sleep pattern, and lost interest in sex. Lambert et al. (2015) also claimed that BDI-II was unidimensional in a sample of cancer patients.

The application of Rasch analysis for analyzing BDI-II is mainly focused on the item fit, detection of group variance, and assessment of unidimensionality. For example, Siegert,et al. (2010) applied the Rasch model to examine item fit, group invariance, and used Rasch factor analysis to assess unidimensionality. Lerdal, Kottorp, Gay, Grov, and Lee (2014) examined the item fit and DIF in stroke survivors. Lambert et al (2015) used Rasch analysis to compare the equality of cut-off points of several depression measures (including BDI-II). There are few published reports conducting more complex analyses, such as, response category optimization or evaluating time invariance.

To the researcher's knowledge, Rasch analysis has not been applied to examine the reliability and validity of BDI-II with a sample of mothers with newborn babies. This study will apply Rasch analysis to validate the BDI-II to find if the BDI-II scale meets Rasch model's assumptions: 1) Unidimensionalty; 2) Group Invariance; 3) Time invariance. In addition, this study will evaluate the efficacy of the number of response categories and optimize the response categories.

Parenting Stress Index- Short Form (PSI-SF)

Parenting stress is linked to negative parenting characteristics, unhealthy parenting styles, and use of harsh discipline (Haskett, Ahern, Ward, & Allaire, 2006). It has a negative correlation with parenting self-efficacy (Jackson, 2000), which is the best predictor of parenting styles (Sanders & Woolley, 2005). The Parenting Stress Index-Short Form (PSI-SF), designed by Abidin (1990), is one of the most common and widely used instruments (Lee, Gopalan, & Harrington, 2016), which has been applied in a variety of settings (Haskett, Ahern, ward, & Allaire, 2006).

The PSI originally was designed to include three subscales (Abidin, 1990): Parental Distress, Parent-Child Dysfunctional Interaction, and Difficult Child. However, some studies found that the PSI only has two dimensions. For example, Haskett, Ahern, and Ward (2006) conducted a confirmatory factor analysis for PSI-SF in a sample of parents (mother and father) with children aged 4 to 10, and found strong evidence of a two-factor model (Personal Distress and Childrearing Stress). Perez-Padilla, Menendez, and Lozano (2015) also categorized this measure in two dimensions: Personal stress, and childrearing stress.

Although PSI-SF has been applied for a variety of research settings, and was analyzed among different samples, there is infrequent application of Rasch analysis to reexamine its psychometric characteristics. Puma (2007) applied a longitudinal Rasch analysis for the population of mothers with babies when the babies were 14, 24, and 36 month old. Puma (2007) examined the Rasch assumptions, evaluated structure invariance by using confirmative factor analysis, and evaluated the properties of the scaling categories of PSI. However, Puma's study did not apply principle component analysis of residual, the method employed in Rasch analysis for evaluating unidimensionality. Therefore, the current study conducted a supplemental investigation to reexamine this measure.

In summary, this study provided a comprehensive Rasch analysis to evaluate the two parenting related measures: BDI-II and PSI-SF. Specifically, the study examined the psychometric properties of the two measures in the longitudinal samples. This study not only focuses on assessing fit statistics, unidimensionality, but also extends to evaluate invariance and optimize scale categories. In addition, the study compared Rasch model and CTT to see if the two methods can yield equivalent results when examine the unidemensionality and reliability.

CHAPTER 3

METHODS

This study is a secondary data analysis of the Parenting for the First-time Project (Parenting for the First-time, 2001), a longitudinal study of the social and cultural contexts of the transition to parenting, and the impact of teen parenting programs on that transition. This study applies Classical Test Theory (CTT) and Rasch analysis using data from two scales measuring Depression (BDI-II) and Parenting Stress (PSI) to evaluate whether the scales meet Rasch assumptions and optimize scale categories.

Participants

Pregnant participants in the third trimester were recruited from four different sites (South Bend, Indiana; Washington D.C.; Birmingham, AL; and Kansas City, Kansas) from 2001 to 2007 at nine time points: prenatal, and at 4, 6, 8, 12, 18, 24, 30, and 36 months after birth.

Respondents were recruited from primary care providers and school-age mothers' programs. The sample size is 684 from three groups: adolescent (under the age of 19, n=389), adult low resource (older than 21 years of age and less than 2 years of college, n=168), and adult high resource (older than 21 years of age and more than 2 years of college, n=127); mothers ranged in age from 14 to 37 years (Smith, T., 2015).

The Beck Depression measure (BDI-II) was administrated at prenatal, 6, 12, 24, and 36 months after birth. The Parenting Stress measure (PSI) was administrated at 6, 12, 24, and 36 months after birth. To ensure that this study had a sufficient number of participants for the time invariance analysis, participants were selected had taken the three surveys at prenatal, 6, and 12 month if the survey had been administrated. Therefore, the number of participants in this study was 357 for Beck Depression measure, and 359 for Parenting Stress measure.

Instrument

Beck Depression Inventory-II (BDI-II)

The Beck Depression Inventory-II (BDI-II) is used to detect the existence and severity of symptoms of depression. This measure includes 21 self-report items with four-point scales ranging from 0 to 3. A higher number indicates greater severity of symptoms of depression. For example, BDI-II item responses ranging from “I don't feel I am being punished”, “I feel I may be punished”, “I expect to be punished”, and “I feel I am being punished”. The maximum total raw score is 63. There are three levels of depression classified by the total score of items (Smith, T., 2015): Minimal (0-13), Mild (20-28), and Severe (>29). In this study, the Beck measure had Cronbach alpha coefficients of 0.85, 0.88, and 0.89 respectively across three administrations: Prenatal, 6-Month, and 12-Month. Siegert, et al. (2010) applied a Rasch model in a neurological rehabilitation sample to examine Beck item fit, DIF, and unidimensionality. Despite three items (Change in sleeping pattern, Changes in appetite and Loss of interest in sex) which did not comply with the Rasch assumptions, they claimed that the remaining items for the

measure can form a unidimentional construct. Appendix A presents the questionnaire of BDI-II.

Parenting Stress Index- Short Form (PSI-SF)

The PSI-SF is used to measure the parental stress of parents who have children up to 12 years old (Smith, T., 2015). Most of the items in the measure have a five point Likert scale ranging from "strongly agree" to "strongly disagree". For example, one of the items states "I often have the feeling that I cannot handle things very well (1-5)". For items with this kind of format, "Strongly Agree (1)" indicates a severe symptom of parental stress, while "Strongly Disagree (5)" indicates a minimal symptom of parental stress. Three items do not have the same Likert scales as the others. Two items, item 22 and item 32, are multiple-choice questions. The respondents are asked to choose one response from five choices. Item 33 asks the parent to count the number of things which their child does that bother them, and choose one of the five categories that have been offered by the questionnaire (e.g. dawdles, refuses to listen, overactive, etc.). Appendix A presents the questionnaire of PSI.

The PSI originally was designed to include three subscales (Abidin, 1990): Parental Distress, Parent-Child Dysfunctional Interaction, and Difficult Child. However, Haskett, et al. (2006) conducted a confirmatory factor analysis for PSI-SF in a sample of parents (mother and father) with children aged from 4 to 10, and found strong evidence of a two-factor model (Personal Distress and Childrearing Stress). Puma (2007) identified two PSI sub-dimensions (Parental Distress and Parent-Child Dysfunctional Interaction) using Confirmative Factor Analysis, and evaluated the structure invariance across

different time points. However, this study claimed that Rasch's unidimensionality assumption was not violated after the exploratory factor analysis.

Data Analysis

Model validation, person, and item fit

Both the Rating Scales Model (RSM) and Partial Credit Model (PCM) can be used for polytomous scales. Linacre (2000) suggested applying RSM for measures with same rating scales among items. Therefore, RSM was conducted to evaluate the data and report fit statistics. Meanwhile, Cronbach's alpha and Item Total Correlation from CTT were also calculated.

Bond and Fox (2007) listed the guideline for misfit of an item: MSQ>1.3 or *t* statistics>2.0 means underfit, while MSQ<0.75 or *t* statistics <-2.0 means overfit. Misfit items were deleted to see if this procedure improved the reliability of the measure, however, deleting these unfit items decreased the reliability of the measures. Since this study is dealing with an existing measurement, items with misfit statistics were retained to keep more information as suggested by Wright and Linacre (1994).

Unidimensionality

Principle component analysis (PCA) and Principal Component Analysis of standardized residual (RPCA) were conducted to evaluate unidimensionality and detect sub-domains for the two measures. The RPCA conducted principle component analysis on standardized residuals instead of raw scores used in PCA. In the RPCA procedure, if the variance explained by the measure exceeds 50%, and an eigenvalue of the

unexplained variance in the first residual factor is less than 2, the unidimensionality assumption is supported (Chen et al., 2012; Linacre, 2012).

Local independence

Local independence of items is another assumption of the Rasch model, which requires that items in a test should not be related to each other (Baghaei, 1998). That is, the success or failure on any items should not be affected by the success or failure of any other items (Bond & Fox, 2007). The violation of the local independence assumption may lead to an inaccurate estimation of person latent trait and item difficulty (Fendrich, Smith, Pollack, & Mackesy-amiti, 2009). Items were expected to have no strong association outside of the latent trait (Wright, 1996). In this study, residual correlations between pairs of items were calculated, and correlations less than 0.3 support the assumption of local independence (Hamilton & Chesworth, 2013).

Differential item functioning across age group

When items display inconsistent response patterns for a particular subgroup of the sample, it reveals a violation of group invariance. Differential Item Functioning was used to detect the violation (Amin et al., 2012). Each item was examined for DIF across two age groups: mothers who were younger than 18 years old (n=185), and mothers who were older than 19 years old (n=172).

The study reports DIF contrast, Welch, and DIF Mantel Haenszel statistics for each item on the two measures. Meanwhile, two way analyses of variance (ANOVA) respectively for each item in the measures also were computed. This ANOVA method was introduced by Tennant et al. (2004).

Each person was categorized into one of four groups according to their ability estimates (e.g. very high, high, low, very low) based on percentile (i.e. 25^{th}, 50^{th}, 75^{th}). This group variable served as an independent variable together with one of the respondent characteristics, such as age or gender (Hamilton & Chesworth, 2013). The dependent variable is the standardized residual given by (Tennant et al. 2004):

$$Z_{ni} = \frac{X_{ni} - E[X_{ni}]}{\sqrt{V[X_{ni}]}}$$

DIF contrast is the difference of the difficulty of the item between the two groups. DIF contrast that is between 0.43 and 0.64 with statistical significance ($p<.05$) can be considered as slight to moderate violation of assumption of group invariance, and >=0.64 can be considerate as large violation (Zwick, Thayer, & Lewis, 1999).

Time invariance

To test the time invariance assumption, Differential Test Functioning (DTF) analysis tested the invariance of item difficulty between two time points. The DTF analysis indicates if the items of a measure function the same way between tests administrated at two different time points, and compares the two sets of difficulties acquired from the two administrations (Linacre, 2012). For BDI-II, DTF analyses were conducted for Prenatal vs. 6 Month, and 6 Month vs. 12 Month; and for PST, DTF analysis was conducted or 6 Month vs. 12 Month.

Reliability, Separation, Person Ability, and Item Difficulty

Three types of reliability statistics were reported: Rasch person reliability, item reliability, and Cronbach's alpha.

Rasch person reliability is equivalent to the traditional test reliability, Cronbach's alpha. The value of Rasch person reliability indicates if the measure can discriminate the samples into enough levels based on their ability (Linacre, 2012): 0.9 = 3 or 4 levels; 0.8 = 2 or 3 levels; 0.5 = 1 or 2 levels. Low person reliability reveals that the range of a person trait is small, or an instrument is too short. Item reliability has no equivalence in traditional testing, and it highly depends on item difficulty variance and sample size (Linacre, 2012).

Person and item separation coefficients were reported to evaluate the ability of a measure in classifying respondents and verifying item hierarchy (Linacre, 2012). Low person separation is defined as a person separation coefficient <2 or a person reliability <0.8. Low item separation is defined as an item separation coefficient <3, or item reliability <0.9. Low person separation indicates that the instrument needs more items to discriminate between high and low ability performers, while low item separation shows that the sample size is not large enough to confirm the item difficulty hierarchy (Linacre, 2012).

Targeting

The Rasch model utilizes item difficulty and person ability to identify the relative location of persons and items, which enable calibration of both person and item in the same logit continuum centered at 0. If the range of item difficulties and person abilities do not share the same spread, it indicates that the sample was not well targeted by the items in the measure, and this will affect the measure's competence to estimate accurately the location of the respondents along the trait continuum (Salzberger, 2003).

Three methods were used in this study to address targeting. First, mean item difficulty and person ability were compared to find if there are large deviations between item difficulties and person abilities. The deviation between 1 and 2 can be considered fair, between .5 and 1 can be considered good, between 0.5 and .25 can be very good, and < .25 can be considered excellent (Fisher, 2007). Second, all person and item estimates (person ability and item difficulty) were depicted in a Wright map. A Wright map is a useful tool to visualize the spread of samples and items, that is, the distribution of sample across items. Third, operational range maps were examined to check how well the range of item difficulty matched the range of person ability. Acceptable range of a measure should be free of ceiling and floor effect (Lo, et al, 2015). Ceiling effect was defined as >15% respondents' ability greater than the highest threshold of item, and floor effect was that >15% respondents' ability lower than the lowest threshold of item (Lo, et al, 2015). Ceiling and floor effect can lower the reliability of a measure to discriminate respondents (Lo, et al, 2015).

Optimization of response category

A Rasch model was run to acquire relevant statistics, including observed count, average measures, outfit mean square, and structure measures, to evaluate if the response categories function well and the necessity of optimization by following Linacre's guideline (2002a):

1) Checking observed to see if there was abnormal observation distribution.
2) Checking average measures to see if they are disordered. The average measure is the average ability across a particular response category. Average measures are expected to increase with higher category labels, for example, the average

measure for category labels should be: “Totally agree” > “Agree” > ”Disagree” > ”Totally disagree”. A disordered average measure implies that number on the rating scale does not correspond to a higher level of the construct (Smith,Wakely, de Kruif, & Swartz, 2003).

3) Checking outfit mean squares of the response categories with a value >2.0. Higher outfit statistics of a particular category is a sign that the category has not been used as expected. This may be caused by a rarely chosen category, the confused perception of respondents on the meaning of a particular category, and the relation between two adjacent categories (Smith, et al., 2003).
4) Checking step measures to identify disordered thresholds. Ordered step measures are expected, which means that the step measure in the third step is larger than in the second step, and then also larger than the first step. Ordered thresholds imply that when one moves up to a higher level of the construct, the higher level of response category, in turn, become the most probable response (Smith et al., 2003).
5) Checking if the distance between two structure measures is larger than 1.0. If not, it means that the scale is not equivalent across the response categories. This criterion (1.0) is only work for five category rating scales.
6) Checking if distance between two structure measures is less than 5.0. If not, more categories may be needed to be added into the middle of the two response categories.

Linacre (1999) suggested combining the disordered categories once the statistical measures do not meet the criteria. After examining the statistical measures and

combining adjacent categories, a Rasch analysis was run again to evaluate the improvement of the scale, including average measure, fit statistics, and reliability. A category probability curve was depicted to assist in the optimization process.

The analytical approaches used to evaluate the assumptions of Rasch scaling are presented below:

1. Unidimensionality: PCA and RPCA.
2. Items perform as expected by Rasch model: Examination of infit and outfit statistics.
3. Categories of the rating scale behave as expected: Examination of observed count, average measure, structure measure, and visual examination of category probability plots.
4. The measure can discriminate respondents into distinct ability levels: Person reliability and separation index
5. The range of items is well-matched to the range of abilities of subjects: Examination of mean item difficulty and person ability, visual examination of Wright Map, and visual examination of Operational range map.
6. The items on the scale perform consistently for different age groups of subjects and across different time points (invariance): Differential Item Function (DIF) analysis and Differential Test Function (DTF) analysis.

CHAPTER 4

RESULTS

This chapter presents the results separately for each measure (Beck Depression Inventory, Parenting Stress Index) with the following topics: model fit and reliability, unidimensionality and local independence, differential item function analysis, targeting, time invariance, and category ordering.

Beck Depression Inventory-II (BDI-II)

Model fit and Reliability

Table 1 presents the person and item statistics (Model infit and outfit, mean person ability, mean item difficulty, person and item reliability, person separation and item separation, and Cronbach's α) for the BDI-II items. The internal consistency (Cronbach's α) is high with 0.85, 0.88, and 0.89 for prenatal, 6 month and 12 month respectively. Rasch person reliability (the index of internal consistency) is lower than the Cronbach's α, with 0.79, 0.75, and 0.71 for the scales of prenatal, 6 month, and 12 month respectively.

Model infit and outfit are used to evaluate the overall fit of the data with the Rasch model, and are expected to be near 1. For BDI-II, the model infit and outfit are 1.1, 1.1, and 1.09 for the three time periods respectively, indicating the overall fit of the data with the Rasch model. Person reliability and person separation indices evaluate the

Table 1

Summary of person and item statistics for BDI-II

	Item infit	Item outfit	Mean person ability	Mean item difficulty	Person reliability	Person separation	Item reliability	Item separation	Cronbach alpha
Prenatal	1.1	0.97	-1.85(1.00)	0.00(1.06)	0.79	1.96	0.99	9.27	0.85
6 Month	1.1	0.96	-2.4(1.32)	0.00(0.84)	0.75	1.73	0.98	6.56	0.88
12 Month	1.09	0.96	-2.87(1.52)	0.00(0.86)	0.71	1.58	0.98	6.29	0.89

measure's competence to distinguish between high and low performers. In this analysis the person separation indices are 0.79/1.96, 0.75/1.73, and 0.71/1.58 for the three time periods, indicating that this instrument can only discriminate two strata of participants basing on their personal traits (i.e. depression). That is, the measure may not be sensitive to discriminate subjects. Item separation/reliability was 9.27/0.99, 6.56/0.98, and 6.29/0.98 across the three periods, suggesting that the instrument discriminates approximately 6 to 9 levels of difficulty among the items.

Several items had infit values that do not match the response pattern predicted by the Rasch model. Item 6 (Being punished) has infit.MSQ in excess of 1.5 across the three time periods (Table 2). Other items with abnormal infit index (>1.5) are item 21(Sex) in the prenatal measure, item 9 (Suicide) in the 6 month measure, and item10 (Crying) in the 6 month and 12 month measures. As noted, this suggests that these items cannot match the response pattern predicted by the Rasch model. However, excluding these abnormal items did not increase the Rasch person reliability. Therefore, those items were retained for further analysis. Meanwhile, the item total correlation analysis shows that item18 (Appetite) and item 21(Sex) have a correlation coefficient of less than 0.3 in the prenatal assessment. But in 6 month and 12 month assessments, all item-total correlation values are larger than 0.3.

The mean logit item difficulty is 0.0 for all the three time periods, and the mean person-ability indices were -1.85, -2.4, and -2.87. The deviations between the average item difficulty and average person ability are larger than 2, which exceeds the criteria of fair targeting (i.e.<2) suggested by Fisher (2007). This indicates that those participants, on average, tend to choose the lower level of scale. The observed count in table 4

Table 2

Item difficulty, Infit MSQ, and Total correlation statistics for BDI-II

Items		Prenatal			6 month			12 Month		
#/Questions		Difficulty	Infit MSQ	Item total correlation	Difficulty	Infit MSQ	Item total correlation	Difficulty	Infit MSQ	Item total correlation
Item1	Sadness	0.39	1.00	0.497	0.34	0.93	0.538	0.23	0.96	0.558
Item2	Discouragement	0.73	1.09	0.385	0.29	1.04	0.531	0.2	0.98	0.507
Item3	Failure	0.8	1.28	0.479	0.42	1.21	0.559	0.07	1.00	0.597
Item4	Loss of pleasure	-0.03	0.78	0.515	-0.2	0.96	0.488	-0.17	0.84	0.585
Item5	Guilty feeling	0.28	0.9	0.421	0.06	0.85	0.493	0.14	0.87	0.523
Item6	Being Punished	1.02	2.00*	0.309	0.81	1.61*	0.553	1.13	1.79*	0.387
Item7	Loss of confidence	0.5	1.39	0.495	0.38	1.08	0.562	0.28	1.06	0.539
Item8	Self criticism	0.34	1.19	0.531	0.13	1.06	0.58	-0.09	1.23	0.515
Item9	Suicidal	2.75	1.11	0.397	2.31	1.53*	0.466	2.44	1.16	0.447
Item10	Crying	-0.71	1.29	0.459	-0.37	1.58*	0.486	0.14	1.66*	0.393
Item11	Anxiety	-0.48	1.12	0.497	0.24	1.28	0.347	6.00	1.13	0.424
Item12	Loss of interest	-0.2	1.18	0.469	-0.26	1.1	0.601	-0.17	0.95	0.565
Item13	Indecisiveness	0.18	0.97	0.498	0.32	0.97	0.578	0.11	0.95	0.57
Item14	Worthlessness	1.77	1.05	0.512	1.19	1.1	0.616	1.22	1.28	0.512
Item15	Loss of Energy	-1.04	0.66	0.468	-0.67	0.65	0.568	-0.79	0.74	0.517
Item16	Change in sleep	-1.62	0.76	0.432	-1.63	1.07	0.288	-1.61	0.94	0.47
Item17	Irritation	-0.51	0.84	0.497	0.04	1.03	0.491	0.09	0.96	0.516
Item18	Change in appetite	-1.59	1.21	0.292*	-1.19	1.27	0.328	-1.2	1.26	0.435
Item19	Concentration	-0.32	0.96	0.446	-0.49	0.74	0.643	-0.46	0.94	0.509
Item20	Tiredness	-1.18	0.77	0.423	-0.67	0.83	0.51	-0.78	0.92	0.579
Item21	Lost interest to sex	-1.08	1.58*	0.240*	-1.05	1.31	0.397	-1.05	1.31	0.412

** Infit MSQ>1.5, or item total correlation<0.3*

confirmed this tendency; nearly 90% of scale categories checked by participants were 0 and 1. Category 2 and Category 3 only account for around 10%. These patterns are similar across the three time periods.

Unidimensionality and Local Independence

Table 3 presents the results of the conventional Principal Component Analysis (PCA) and the Rasch Principal Component Analysis (RPCA). In the PCA, the first factor of the measure explained around 27%, 33%, 32% of raw variance respectively across the three administrative periods. Except item 18 and 21, all other items have loadings larger than 0.35 on the first factor. Items with large than 0.40 loading on other factors are item 15, 18, and 21 in the prenatal test, items 16 and 18 in the 6 month, and items 18, 20, 21 in the 12 month tests.

In RPCA, the Rasch dimension (all items/first factor) explained 39%, 38%, and 36% of raw variance across the three administrative periods. The variance explained by the items is around 4 times larger than the variance explained by the 1st contrast (possible second dimension). The eigenvalues for the 1st contrast are around 2 (2.3 or 2.2) across the three periods, which shows that the second dimension may only be comprised of two items.

Meanwhile, Principle factor analysis (PFA) conducted for the three assessments detected only one dimension (Eigenvalues of second factor were less than 1). Therefore, although the variance explained by the measure does not exceed the threshold (i.e. 50%), since the possible second dimension only had two items, and considering the variances

Table 3

PCA and RPCA analysis for BDI-II

	Principle component analysis		Rasch principle component analysis				
	Factor1 Eigenvalue (proportion)	Item loading on other factors (>0.4)	Raw variance explained by measures	Unexplained variance(total)	Unexplained variance in 1st contrast	Explained by items	Possible Item of second dimension
Prenatal	5.73 (27%) All items have loading larger than 0.35 on Factor1 except item 18, 21	15,18,21	13.5 (39.1%)	21.0 (60.9%)	2.3 (6.8%)	10.3 (30.0%)	15,20
6 Month	6.99(33%) All items have loading larger than 0.35 on Factor1	16, 18	12.8 (37.9%)	21.0 (62.1%)	2.2 (6.6%)	7.5 (22.2%)	16,18
12 Month	6.81(32%) All items have loading larger than 0.40 on Factor1	18,20,21	11.6 (35.6%)	21.0 (64.4%)	2.2 (6.6%)	6.8 (21.0%)	18,20

53

explained by the items is much larger than the variance explained by the 1st contrast, the Rasch analysis supports the assumption that the BDI-II measure has only one dimension.

In addition, the correlations of the residual between pairs of the item were examined for the three periods separately. Pairs with a correlation larger than 0.3 were not found across the three periods. This indicates that the assumption of local independence is not violated by the measures.

Overall, the results of PCA, RPCA, and local independence analyses support the unidimensionality assumption of the BDI-II measure.

Differential Item Function (DIF) analysis

The measures in the Parenting study were developed with mothers who were younger than 18 years old (n=185), and mothers who were older than 19 years old (n=172). The assumption of group invariance supposes that the scales should perform similarly between the two groups, although the two groups may have different levels of depression. The purpose of DIF analysis is to examine the potential problems related to the assumption of invariance.

Table 4 reports several statistics related to DIF analysis: 1) DIF contrast, 2) Welch p value, 3) Mantel Haenszel *p* value, and 4) *p* value of ANOVA. DIF contrast, Welch, and DIF Mantel Haenszel statistics are generated by Winsteps , the software conducting Rasch analysis. Using ANOVA analysis to detect DIF was introduced by Tennant and Pallant (2007).

The Welch method is a model that estimates the difference between the item difficulties for two groups, by keeping everything else constant, while the Mantel-

Table 4

Differentiation Item Function Analysis for BDI-II (>=19 vs. <19)

	Prenatal				6 Month				12 Month			
Item#	DIF Contrast	Welch	Mantel Haenszel	Anova DIF	DIF Contrast	Welch	Mantel Haenszel	Anova DIF	DIF Contrast	Welch	Mantel Haenszel	Anova DIF
1	0.55	0.008	0.007	0.013					0.5	0.04	0.11	
3					0.6	0.02			0.59	0.01	0.01	0.024
5					0.48	0.03	0.006	0.010				
6	1.14	0.00	0.012	0.006					0.84	0.01	0.02	
7									-0.44	0.07		0.04
8					-0.42	0.05	0.04	0.025	-0.57	0.01	0.02	0.0073
9					1.2	0.04						
10					0.53	0.007			1.2	0.00	0.004	0.0007
12	0.46	0.01	0.03	0.024	0.41	0.04						
14									-0.49		0.04	
15	-0.33	0.02	0.004	0.024	-0.48	0.005	0.001	0.005	-0.42	0.024	0.013	0.034
16					-0.5	0.008						
20	-0.43	0.003	0.004	0.004	-0.53	0.002	0.005	0.006	-0.57	0.002	0.005	0.0023
21	-0.32	0.03										

Note: The statistics in Welch, Mantel Haenszel and Anova DIF are *p* value. The table only show statistics with $p<0.05$.

Haenszel method estimates the difference from cross-tables of observations of the two groups (Linacre, 2012).

Tennant and Pallant (2007) introduced the use of ANOVA in analyzing the DIF. In this model, the residual is the outcome, and person ability and the group variable are the two independent variables. Table 5 shows most items have $p<0.05$ in any of the three tests.

DIF contrast is the difference of the difficulty of the item between the two groups. The significant items ($p<0.05$) with an absolute DIF contrast that is larger than 0.64 was item 6 (1.14) in the prenatal assessment, item 9 (1.2) in the 6 month, and item10 (1.2) in the 12 month. Those items are also reported misfit (Table 2). Excluding these items from corresponding Rasch analysis did not improve the person reliability and separation of the measure.

Considering that there are nearly half of the items which exceeds the criteria of group invariance, BDI-II violates the assumption of group invariance.

Targeting

To examine if the measure has fair targeting,, the researcher looks at several statistics with regard to item and person means. The mean logit item difficulties are 0.0 for the three assessment periods, and the mean person abilities are -1.85, -2.4, and -2.87. The deviations between average item difficulty and person ability are larger than 2 in the 6 month and 12 month assessments, which is not “fair” (Fisher, 2007), indicating that the competency of this measure to target the ability of the respondents is not very good in the 6 and 12 month data.

The Wright map reflects the person's ability level compared to the average difficulty of the items. Figure 3 provides the Wright person-item map on the logit scale. The figure depicts items and respondents with “High depression” on the top of the continuum, and those with “Low depression” on the bottom. On average, the items are most often located between -1.5 to 0.5, while the respondents are most often located between -5 to -1. Using 6 month data as an example (Figure 3), the upper categories of the half items extend to difficulty/acceptance logit levels of between 0 and 3. Half of the persons have no items falling within the person spread with the ability logit ranging between -2 to -5. Overall, the items were too “hard” compared to the person’s ability (stress level), so subjects seldom endorse the higher level of scale categories.

Close examination of the meaning of items reveals that the items which overlap with respondents are item 15, 16, 18, 20, and item 21, they are questions about energy, sleep, appetite, fatigue, and sex. Therefore, items querying the physical condition of the respondents contribute more to discriminate the depression level of respondents.

The Wright map only provided a rough picture on how items and samples located along the logit continuum, as the measure was a polynomial scale; thus, we need to check the map of operational range to examine the performance of each scale category. Figure 4 presents the operational range for BDI-II for the three administration periods. The effective range for samples are -4 to 1 (5 logit span), -4.5 to 2.1 (7.6 logit span), and -5 to 1.5 (6.5 logit span) respectively for prenatal, 6 and 12 month assessments. Item 16 (Change in sleep pattern) is the “easiest” item with lowest difficulty logit, and item 9 (Suicidal) is the most “difficult” item with highest logit for all of the three administrative periods. Figure 4 shows that “easy” items can discriminate the sample better than

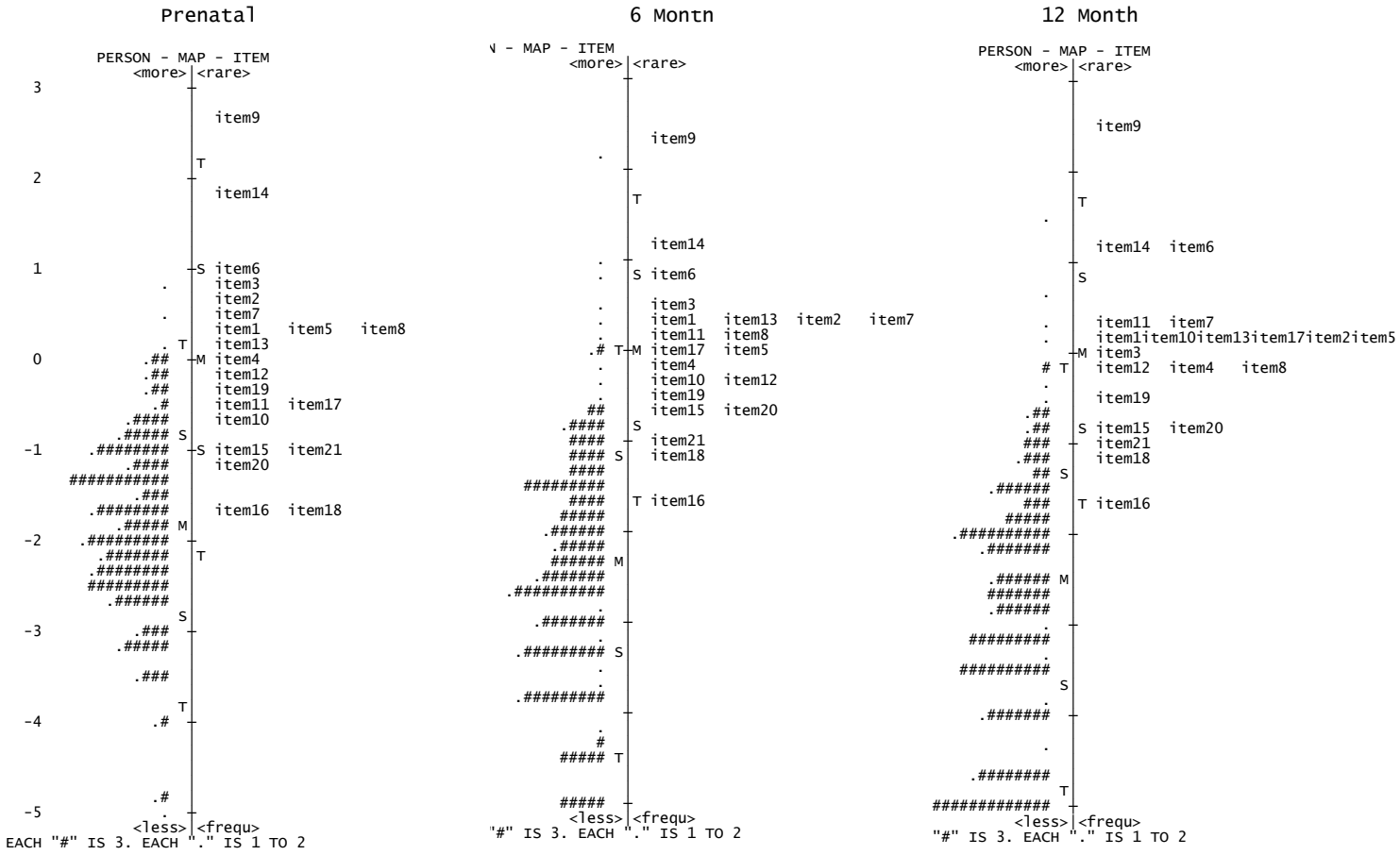
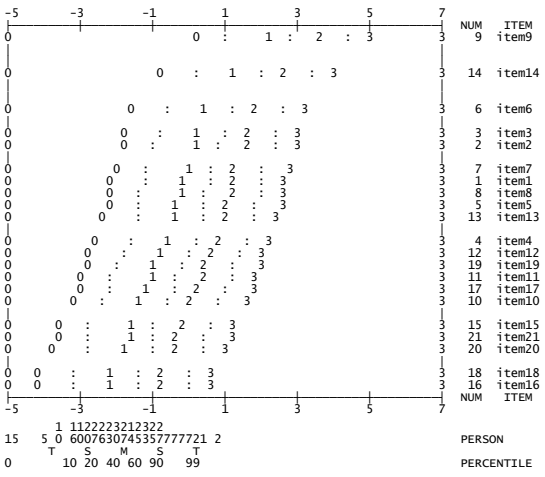
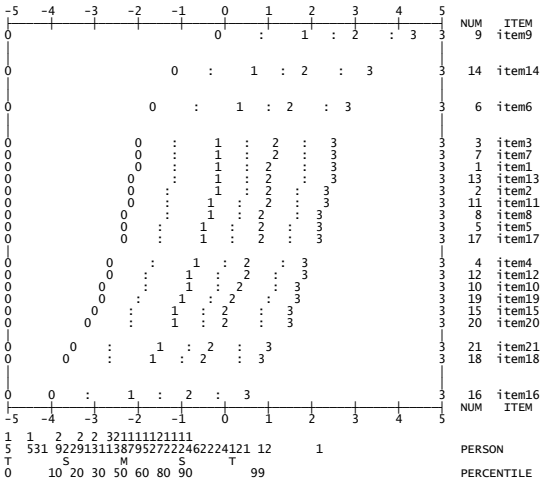


Figure 3. Wright person-item map on the logit scale for BDI-II

Prenatal



6 Month



12 Month

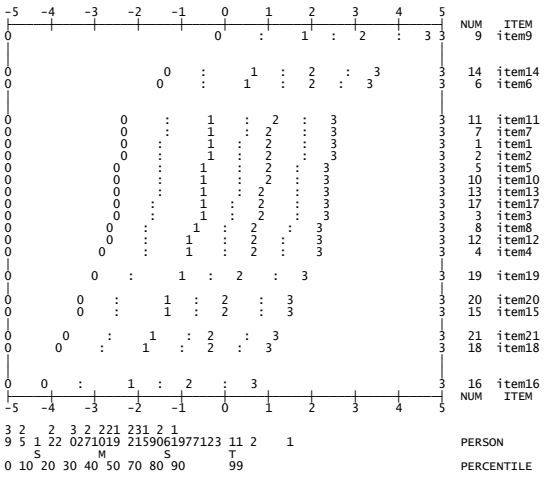


Figure 4. Operational range map for BDI-II

"difficult" items. Overall, the items have a broader coverage over ability logits with a range from -5 to 5.

The next step is to detect if the operational range is free of ceiling and floor effects to be an acceptable range. Ceiling effect was defined as >15% respondents' ability greater than the highest threshold of item, and floor effect was that >15% respondents' ability lower than the lowest threshold of item (Lo, et al, 2015). Ceiling and floor effects can lower the reliability of a measure to discriminate respondents (Lo, et al, 2015).

Figure shows that, there is less than 5% person ability that is below the lowest item difficulty threshold in the prenatal assessment, showing that the range is acceptable. However, there are 10% and 20% person ability that are lower than the lowest item threshold in the 6 month and 12 month assessments, raising a red flag that the operational range in the two administrations may not be acceptable, especially in the12 month assessment.

Time invariance

Figure 5 presents the results of Differential Test Functioning (DTF) analysis for prenatal vs. 6 month data, and 6 month vs. 12 month data. The DTF analyses determine if the items of a measure function the same way between two tests, and compare the two sets of difficulties (Linacre, 2012). The Figure (Figure 5) depicts the relative locations of the difficulty logit between two administration periods. Overall, most of the items remain inside the boundary of the two 95% confidence bands (item 20, 16 and 21 roughly near the upper band), suggesting that most of items function the same way across time. However, in Figure 5a, item 11(Anxiety) and 17(Irritation) locate outside the lower

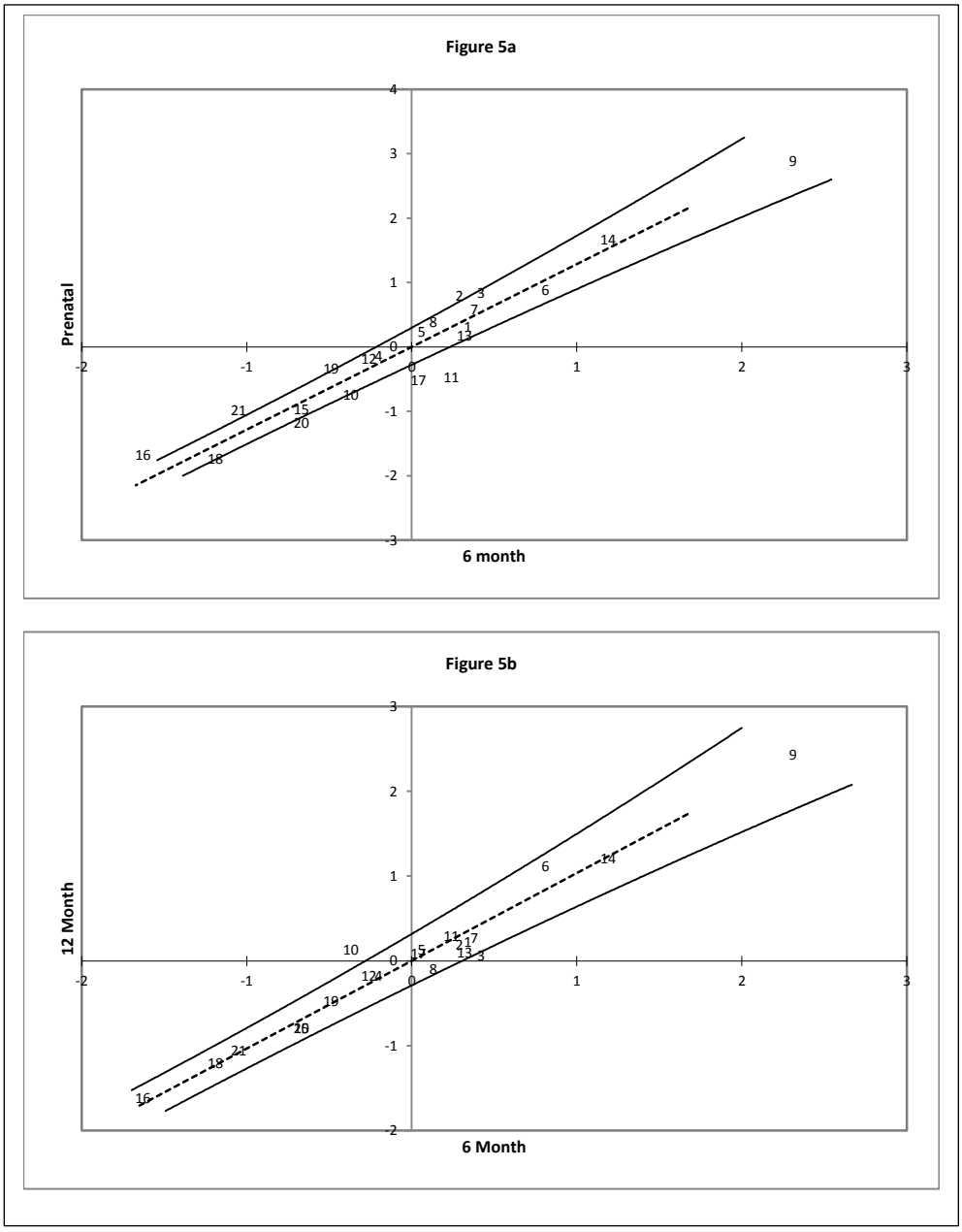


Figure 5. Differential Test Function Analysis for BDI-II

confidence bands, indicating that they are "harder" in the prenatal administration than in the six month. That is, subjects more often choose a lower category for these items in the prenatal administration than in the six month (Please note, for BDI-II, lower category means lower level of depression). Meanwhile, Item10 (Crying) in Figure 5b locates outside the upper confidence bands, suggesting this item was "easier" at 12 months than at six months (subjects are more often to choose higher category/higher depression level for this item at 12 months than at 6 months).

Overall, despite several items exceeding the expectation, the measures remained invariant across the three time points.

Category Ordering

Table 5 shows that the observed count of response category 0 and 1 account for nearly 90% of the occurrences of the categories. The occurrence of response category 0 and 1 add up to 88%, 93% and 95% respectively across the administration periods, indicating that the participants tend to choose these two levels. Average measures (mean of person ability-item difficulty) increase as expected along the three administration periods. The structure measures (threshold between categories) in prenatal and 12-month periods increase along the category levels, and decrease along the levels in the 6-month analyses. The infit or outfit mean square statistics of category 3 for all of the three periods are larger than 1.5, indicating that this is a problematic category whose values are far away from expected.

Figure 6 presents the three category probability curves for BDI-II. The category probability curves for the three periods all showed that categories 2 and 3 had the same

Table 5

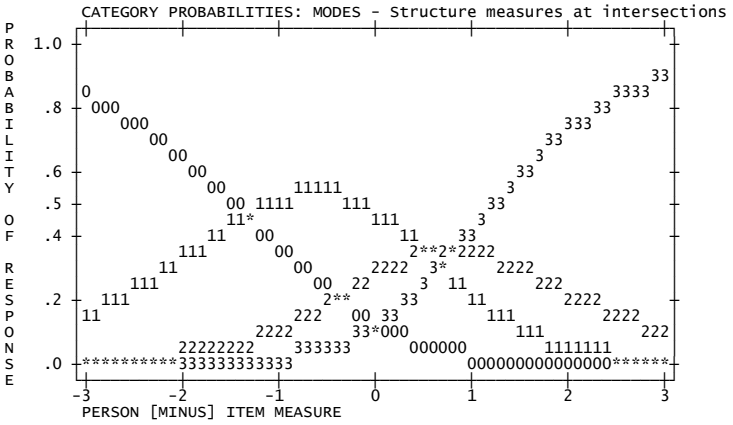
Category scale statistics for BDI-II

	Prenatal				6 Month				12 Month			
#	Observed count (%)	Average Measure	Structure measure	Infit/Outfit MNSQ	Observed count (%)	Average Measure	Structure measure	Infit/Outfit MNSQ	Observed count (%)	Average Measure	Structure measure	Infit/Outfit MNSQ
0	4126(55%)	-2.61		.97/.98	5030(68%)	-2.84		.99/1.01	5264(71%)	-3.06		1.00/.98
1	2441(33%)	-1.11	-1.35	.91/.66	1871(25%)	-1.36	-1.2	.93/.69	1756(24%)	-1.54	-1.38	.92/.74
2	636(8%)	-0.42	0.54	1.07/1.12	347(5%)	-0.61	0.66	1.09/1.23	301(4%)	-0.68	0.6	1.03/1.12
3	284(4%)	-0.03	0.81	1.38/1.69	190(3%)	-0.02	0.54	1.46/1.85	114(2%)	-0.34	0.78	1.56/2.26

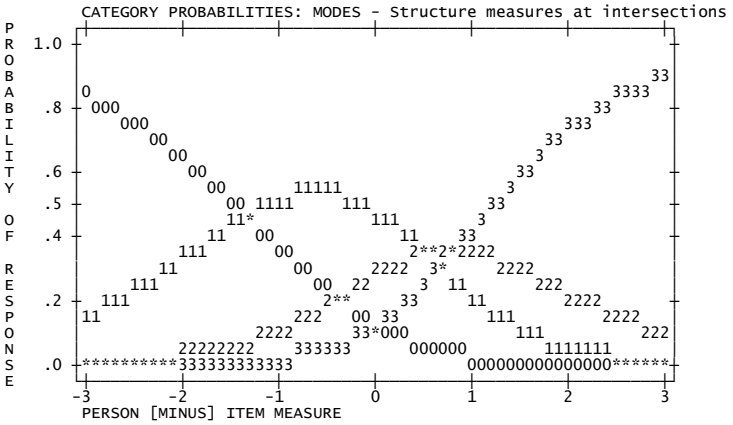
Category optimization

Scale	Person Reliability	Person Separation	Person Reliability	Person Separation	Person Reliability	Person Separation
0123	0.79	1.96	0.75	1.75	0.71	1.58
0133	0.79	1.96	0.72	1.59	0.69	1.49
0122	0.81	2.07	0.77	1.85	0.73	1.66

Prenatal



6 Month



12 Month

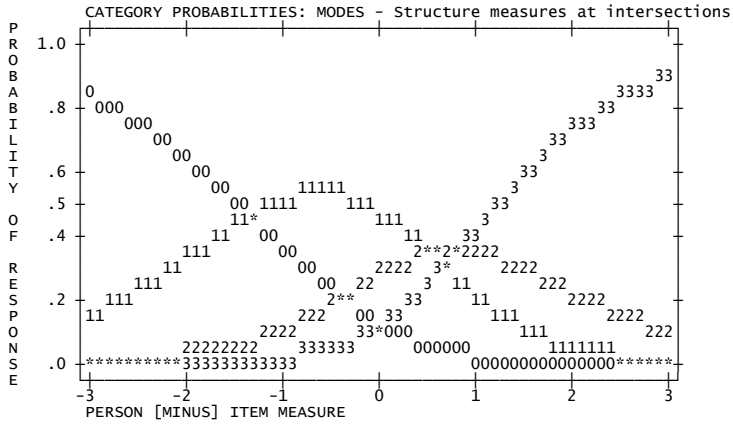


Figure 6. Category probability curves for BDI-II (0123)

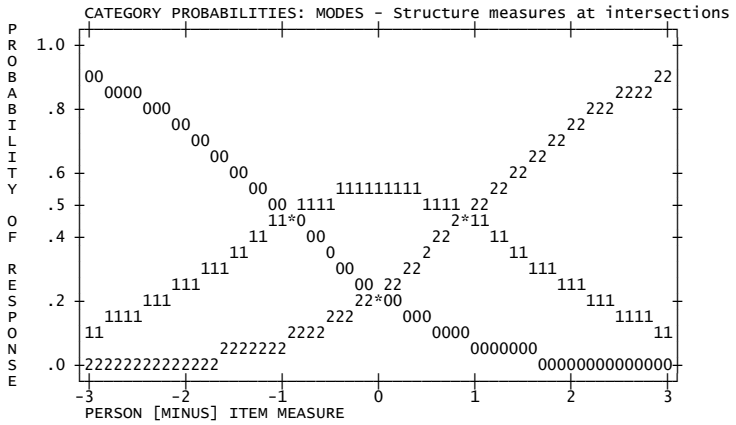
likelihood of being selected along the continuum for respondents with <-1 logit. While as expected, the respondents with lower ability (less depression) should have a higher probability to choose category 2 than category 3. This pattern of response suggests that these items may function better with a three-point format than a four-point format.

By observing the category probability curves, it seems that there are two ways to collapse the level of categories for improvement: 1) combining level 2 to 3 (0133), or 2) combining level 3 to 2 (0122). Table 4 presents person reliability after collapsing the categories. The collapsed response "0133" generated lower person reliability and person separation indices than the original pattern. The reliability/separation was 0.79/1.96, 0.72/1.59, and 0.69/1.49 respectively, while the original reliability/separation was 0.79/1.96, 0.75/1.75, and 0.71/1.58. However, collapsed response "0122" results in a higher person reliability and separation index for all of the three periods, and the person reliability and separation become: 0.81/2.07, 0.77/1.85, and 0.73/1.66 respectively for the three periods. Figure 7 presents the response category probability curves for "0122", which shows the improved probability estimation along the trait level.

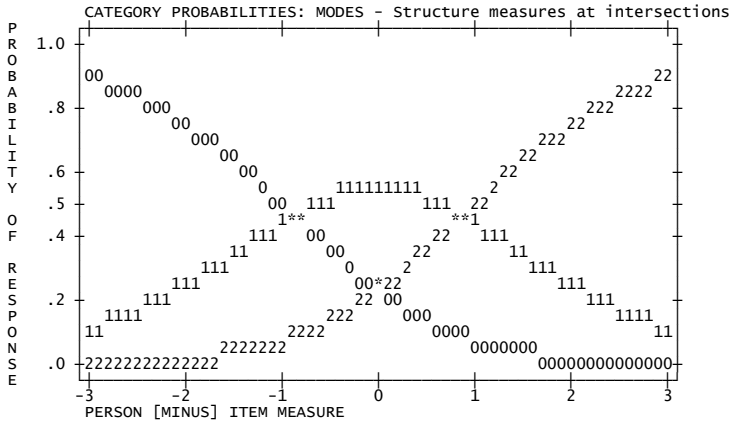
Parenting Stress Index (PSI)

A Principle Component Analysis (PCA) was conducted for the 36 items of the PSI measure of stress. The first factor explained 25.9% and 29.5% of total variance for 6 month and 12 month adminsitrations respectively. (This measure was not used in the prenatal assessment.) The results of factor pattern in the 6 month data clearly show that item1 to item 12 are loading on a second factor. Appendix C provides the factor pattern for the PCA analysis. Rasch residual principle component analysis (RPCA) for both six

Prenatal



6 Month



12 Month

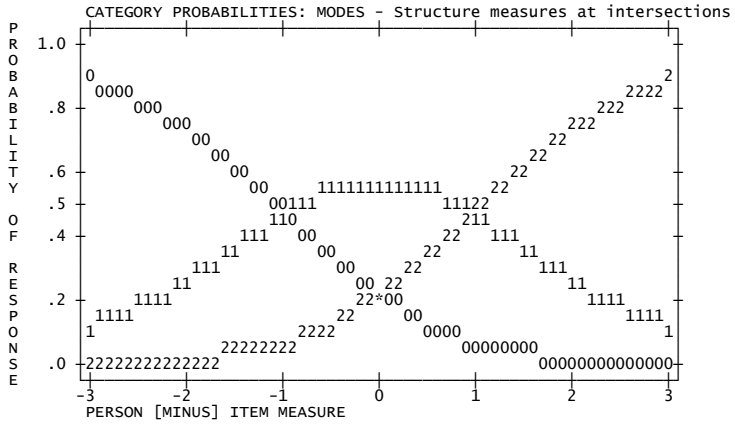


Figure 7. Category probability curves for BDI-II (0122)

and 12 months demonstrates that the raw invariance explained by the measure are around 37%, and the 1st contrast (possible second dimension) may have five items (Eigenvalue=4.9), which is enough to comprise a second dimension. The results of Principle factor analysis (PFA) also revealed two factor structure (The eigenvalue of the second factor is larger than 2). Appendix C presented the rotated factor pattern for PSI.

Checking the questions in the instrument, it appears that the questions in Item13 to Item 36 are about the stress related to the child; each of the questions has the key word "Child/Children". For example, Item 20 is: "My child is not able to do as much as I expected." Meanwhile, the questions from Item 1to Item12 are about the "feeling" of the respondents. For example, Item1 asked: "I often have the feeling that I cannot handle things very well." Therefore, the PSI has two constructs: childrearing stress and self-stress.

Some studies also identify two dimensions for the PSI when checking its psychometric properties. For example, Perez-Padilla, Menendez, & Lozano (2015) identified two factors: Personal stress and childrearing stress among mothers with children younger than 12 years old, and used the two constructs separately to evaluate their relation with parental stress and locus of control. Haskett, Ahern, Ward, & Allaire, (2006) also reported that there is a strong evidence of two factor structure, and find that childrearing stress is a significant predictor of a parental history of abuse. Therefore, to align with the assumption of unidimensionality, Rasch analysis was conducted for the two constructs separately (Childrearing Stress (CRI) and Self Stress (SSI)). The results for the two measures are presented one by one.

Childrearing Stress (CRI)

Model fit and Reliability

Table 6 presents the fit statistics of CRI items. The internal consistency (Cronbach's α) is high with 0.88 and 0.89 for 6 month and 12 month assessments respectively. Rasch person reliabilities are lower than Cronbach's α; they are 0.79, and 0.82 for 6 month, and 12 month assessments respectively.

Item infit/outfit are 1.07/1.17, and 1.08/1.67 for 6 month and 12 month periods respectively, indicating that the overall fit of the Rasch model is good. Person separation indices are 1.93 and 2.14 for the two periods, suggesting that CRI can only discriminate two strata of participants. Item separation /reliability were 0.98/7.14 and 0.98/7.91 across the two periods, which means that this measure can discriminate approximately 7 or 8 levels of difficulty.

Items (Table 7) which have infit.MSQ in excess of 1.5 across the two periods are 29 (react strongly), 31(hard to establish schedule), 32 (hard to get/stop kids do things), and 33(number of things bother). Item13(Make me feel good and item 22 (Good parents) have infit.MSQ which is larger than 1.5 in 6 months data.

Most of items had the response pattern with five-point Likert scale: "Strongly Agree", "Agree", "Not sure", "Disagree", and "Strongly Disagree". However, items 22, 32, 33 are three multiple choice questions each with 5 response categories. This may be a reason why their fit statistics are out of range. Item total correlations showed abnormal items similar to the infit statistics: items 22, 29, 31, 32, 33 are less than 0.3 in both the 6 month and 12 month periods. Excluding these abnormal items did not increase the Rasch person reliability, however, the Cronbach alpha increased.

Table 6

Summary of person and item statistics for Child Rearing Index (CRI)

	Item Infit	Item outfit	Mean person ability	Mean item difficulty	Person Reliability	Person Separation	Item Reliability	Item Separation	Cronbach alpha
6 Month	1.07	1.17	1.74 (1.07)	0.00 (0.61)	0.79	1.93	0.98	7.14	0.88
12 Month	1.08	1.13	1.67 (1.01)	0.00 (0.65)	0.82	2.14	0.98	7.91	0.89

Table 7

Item difficulty, Infit, and Total correlation statistics for CRI

		6 month			12 Month		
item #		Difficulty	Infit MSQ	Item total correlation	Difficulty	Infit MSQ	Item total correlation
13	Make me feel good	0.33	1.35	0.51	0.01	1.70*	0.42
14	Close to me	-0.53	0.88	0.59	-0.6	0.90	0.61
15	Smile at me	0.10	1.33	0.48	-0.29	1.14	0.55
16	Being appreciated	-0.24	0.89	0.62	-0.39	0.93	0.62
17	Laugh/giggle	-0.34	0.84	0.62	-0.67	0.88	0.59
18	Learn quickly	-0.38	0.74	0.65	-0.38	0.81	0.61
19	Smile often	-0.60	0.70	0.62	-0.67	0.71	0.67
20	Do much as expected	-0.35	0.77	0.61	-0.53	0.83	0.59
21	Get use to new things	-0.06	0.95	0.51	-0.05	0.71	0.60
22	Good parent	0.12	1.38	0.22	0.16	1.48*	0.17*
23	Warm feeling	0.21	1.12	0.49	-0.09	1.13	0.49
24	Mean because of kids behavior	-0.46	0.72	0.65	-0.35	0.81	0.61
25	Cry often	-0.03	0.74	0.60	-0.06	0.79	0.67
26	Wake in bad mood	-0.22	0.90	0.54	-0.18	0.79	0.63
27	Moody	0.05	0.83	0.64	-0.01	0.88	0.62
28	Bother me a lot	0.09	0.96	0.55	0.26	0.95	0.60
29	React strongly	1.80	1.96*	0.23*	1.88	1.82*	0.28*
30	Upset easily	0.31	0.90	0.53	0.61	1.18	0.47
31	Hard to establish schedule	0.72	1.56*	0.23*	0.7	1.58*	0.26*
32	Hard to get/stop kids do things	1.22	1.55*	0.12*	1.55	1.34	0.21*
33	Number of things bother	-1.36	2.14*	0.20*	-0.73	1.84*	0.21*
34	Child does thing bother	0.22	0.91	0.50	0.48	1.05	0.54
35	Child is problem	-0.52	0.62	0.69	-0.53	0.69	0.68
36	Child demand more	-0.07	1.07	0.43	-0.12	0.89	0.57

** Infit MSQ>1.5 or Item total correlation<0.3*

The fit statistics reveals that there are at least 5 items which did not perform as the Rasch model expects, and three of them have non-Likert scale response format.

Unidimensionality and Local Independence

The results of the conventional Principal Component Analysis (PCA) and Rasch Principal Component Analysis (RPCA) are presented in Table 8. In PCA, The first factor of the CRI measure explains 33.3% and 33.5% of the variance respectively for six month and 12 month assessments. All items have loadings larger than 0.45 on the first factor except for items 22, 29, 30, 31, 32, and 33. Items with larger than 0.40 loading on second factors are items 31 and 32 in the six month, and items 28, 29, 30, 31 in the 12 month assessments.

In RPCA, the Rasch dimension (all items/first factor) explains 36.1% and 40.9% of raw variance (lower than the criteria of 50%) in six month and 12 month assessments, respectively. The variance explained by items is around three times larger than the variance explained by the 1st contrast (possible second dimension). The eigenvalue for the 1st contrast is around 3 (3.1 or 3.8) across the two periods, which shows that the second dimension may comprise of 3 items, and the possible items for the second dimension are items 29, 31, 32, which are also misfit items. Meanwhile, Principle factor analysis only revealed one dimension (the eigenvalue of second factor is less than 2), providing another evidence that CRI is unidimensional.

Overall, most of the items load on one factor, but some items especially items with multiple response choices (item 22, 31, 32) did not function as expected. In addition, local independence analysis does not reveal any pair of items which has a correlation

Table 8

Principle component analysis for CRI

	Principle component analysis (PCA)		Rasch principle component analysis				
	Factor1 Eigenvalue (proportion)	Item loading on second factors (>0.4)	Raw variance explained by measures	Unexplained variance(total)	Unexplained variance in 1st contrast	Explained by items	Possible Item of second dimension
6 Month	8.0(33.5%) items have loading larger than 0.40 on Factor1 except 22,29,31,32,33	31, 32	13.6 (36.1%)	24.0 (63.9%)	3.1 (8.3%)	7.6 (20.3%)	29, 31, 34
12 Month	8.4(33.5%) items have loading larger than 0.40 on Factor1 except 22,29,31,32,33	28, 29, 30, 31	16.6(40.9%)	24.0(59.1%)	3.8 (9.4%)	8.9 (22.0%)	29, 30, 32, 31

larger than 0.3 during the two administration period. This is the supportive evidence that the assumption of local independence is not violated by CRI.

Differential Item Function (DIF) Analysis

Table 9 reports the statistics related to the DIF analysis of CRI. The items with an absolute DIF contrast that is larger than 0.6 are item 31 (0.62), item 32 (0.63), item 33 (0.98), in six month data, and item 31 (0.61) and item 32 (0.62) in 12 month data. Items 31 and 32 are significant in all of the three tests in both six month and 12 month assessments. Again, excluding these items from corresponding Rasch analysis did not improve the person reliability and separation of the measure. The results illustrated that those misfit items (Item 31, 32, 33) are also the items violating the assumptions of group invariance.

Targeting

The mean logit item difficulty (Table 9) is 0.0 for the two periods, and the mean person ability is 1.74, and 1.67 respectively. The deviations between average item difficulty and person ability is "fair" (<2), indicating that the measure's competence of targeting the sample is fair (Fisher, 2007). The observed count in table 10 shows that the count of categories 4 and 5 account for nearly 80% of the total number of counts. Categories 1, 2, and 3 only account for around 20%. All of the three periods have the same response pattern. These results show that participants, on average, tend to choose the higher level of scale categories (e.g. Strongly Disagree). According the meaning of the questions, a higher level of scale categories represents lower level of stress.

Table 9

Differentiation Item Function Analysis for CRI (>=19 vs. <19)

	6 Month				12 Month			
Item#	DIF Contrast	Welch	Mantel Haenszel	Anova DIF	DIF Contrast	Welch	Mantel Haenszel	Anova DIF
14					-0.52	0.0058		0.021
31	0.62	0.000	0.0005	0.001	0.61	0.000	0.000	0.001
32	0.63	0.001	0.0342	0.030	0.62	0.000	0.004	0.001
33	-0.98	0.003		0.028				

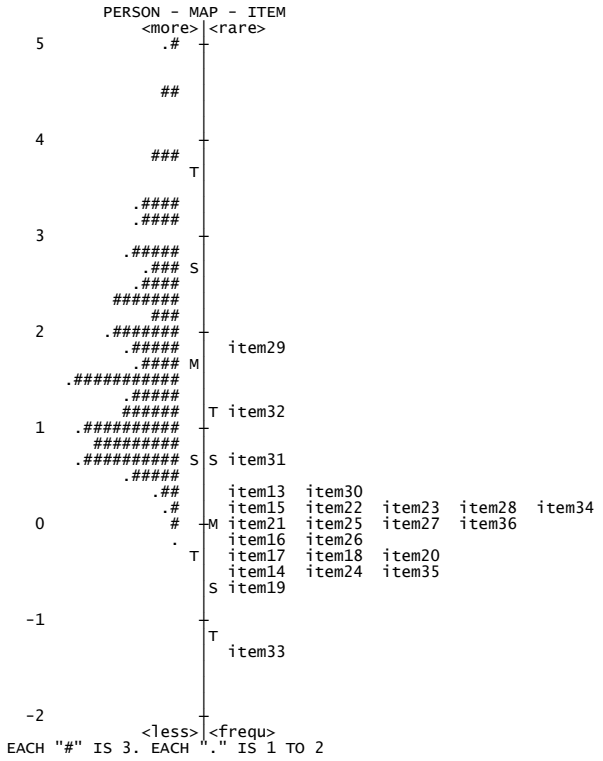
Note: The statistics in Welch, Mantel Haenszel and Anova DIF are *p* value. The table only show statistics with *p<0.05 and DIF contrast >0.43*

For example, item14 asked: "My child rarely does things for me that make me feel good." The answer "strongly disagree' means that the parent was not stressed by this feeling.

Figure 8 provides the Wright person-item map on the logit scale. The figure depicts items and respondents with "high ability/low stress" on the top of the continuum, and those with "low ability/high stress" on the bottom. On average, the items are most often located between -1 to -0.5, while the respondents are most often located between -1 to 3 with a skewed bell shape. Using 12 month data as an example (Figure 11), most of the items overlap with the difficulty between -0.8 and 0.8 logit, and there are two items (19, 32) located between 1.5 and 2 logit. Overall, the items were too "easy" compared to the person's ability (stress level), which shows that subjects seldom endorse the low level of scale categories.

To examine the distribution of scale response categories, Figure 9 presents the operational range for CRI. Overall, the items can cover the ability range from -4 to 5. However, the samples nest between -0.5 to 5, and -1 to 5 for the two periods, which are

6 Month



12 Month

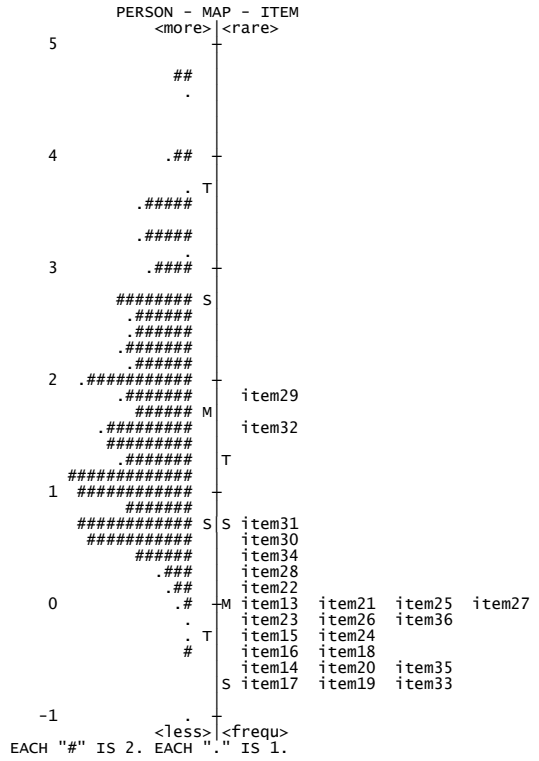
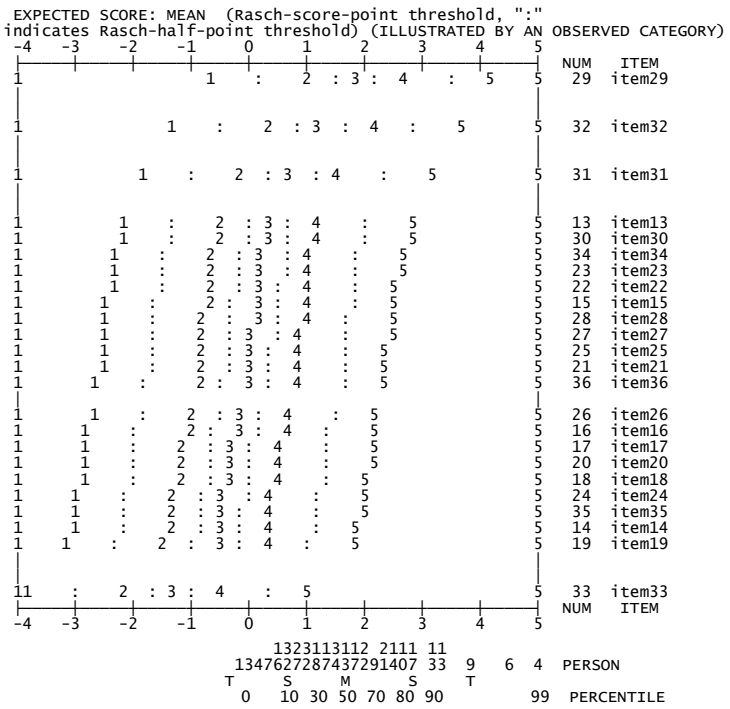


Figure 8. Wright person-item map on the logit scale for CRI

6 Month



12 Month

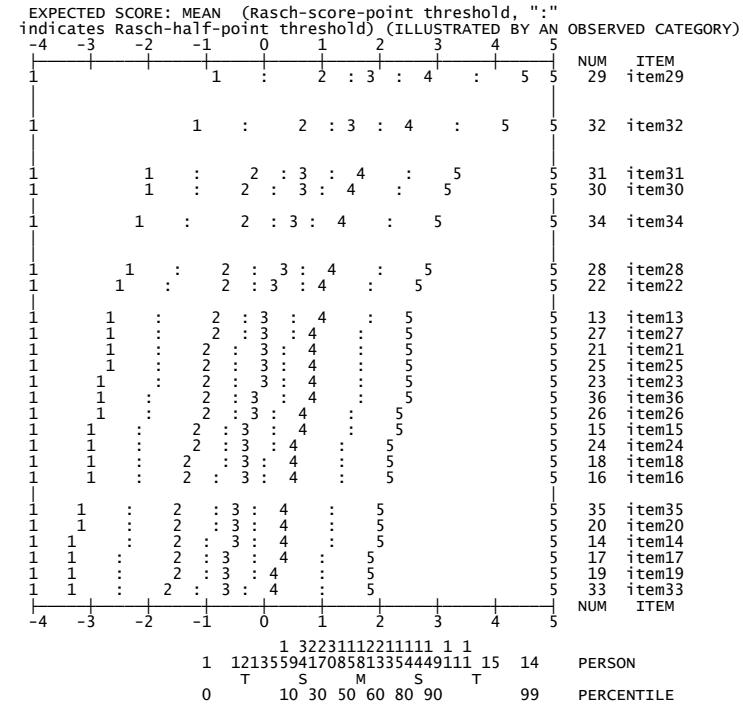


Figure 9. Operational range map for CRI

the interval with “high ability” or “less stress”. The sample seldom chooses the “1” or “2” category. Item 29, 32, 31 are the items which may cover the sample better than other items, however, samples still rarely choose lower level of these items, and those items have been reported as misfit. Except for several extreme cases, person ability is all contained in the range of item difficulties, which means the measure has acceptable range. However, the low level of scale categories (i.e. 1, 2) did not contribute a lot to discriminating the subjects.

Time invariance

Figure 10 presents the results of Differential Test Functioning (DTF) analysis: six month vs. 12 month. Overall, most of the items remain inside the boundary of the two 95% confidence bands, suggesting that most items function the same way across time. Item 13, 23, 15 are located slightly outside the higher confidence band, indicating that they are “harder” in the 6 month assessment than in the12 month. In other words, respondents more often choose a lower category for these items in the six month assessment than in the 12 month. For CRI, a higher category means a low level of stress. Items 32, 33 are below the lower band, indicating that those items are easier in the six month assessment than the 12 month. They are also misfit items. Overall, despite several misfit items, this measure remains time invariant.

Category Ordering

Table 10 shows that the observed count of response categories “4” and “5” account for more than 80% of the occurrences of the categories. The occurrence of category “4” and “5” add up to 85% and 83% respectively in the six month and 12 month

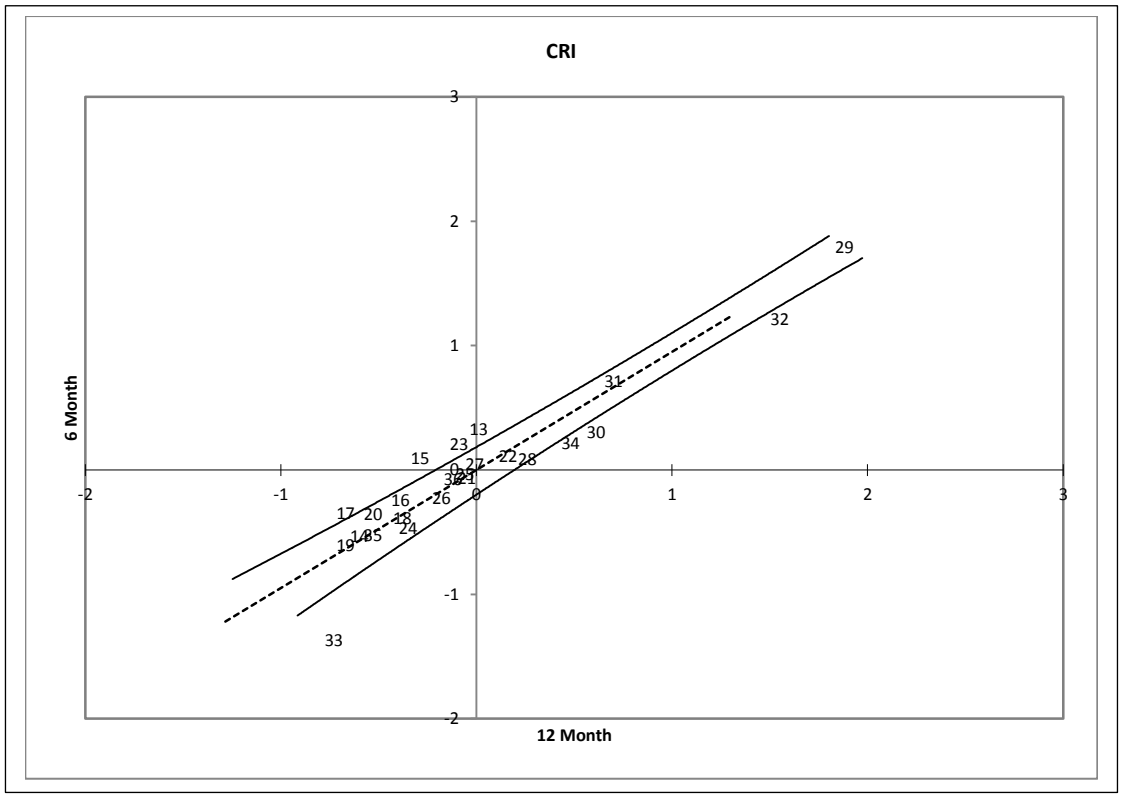


Figure 10. Differential Test Function Analysis for CRI

Table 10

Category scale statistics for CRI

	6 Month				12 Month			
#	Observed count (%)	Average Measure	Structure measure	Infit/Outfit MNSQ	Observed count (%)	Average Measure	Structure measure	Infit/Outfit MNSQ
1	161 (2%)	0.55		2.05/3.75	187 (2%)	0.12		1.94/2.86
2	584 (7%)	0.41	-1.31	1.20/1.60	627 (8%)	0.30	-1.47	1.22/1.53
3	529 (6%)	0.83	0.63	1.08/1.30	563 (7%)	0.73	0.63	1.04/1.17
4	2603 (30%)	1.05	-0.52	0.91/0.51	2699 (32%)	1.13	-0.49	1.00/0.63
5	4672 (55%)	2.37	1.21	0.83/0.92	4363 (51%)	2.40	1.33	0.85/0.95

Category optimization

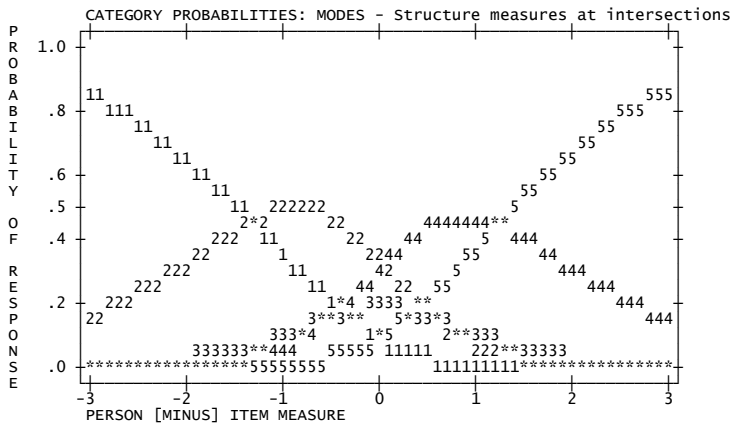
Scale	Person Reliability	Person Separation	Person Reliability	Person Separation
12345	0.79	1.93	0.82	2.14
12335	0.87	2.59	0.89	2.83
12445	0.80	1.98	0.82	2.13

data, which indicates that respondents tend to choose higher level of categories. Average measures (mean of person ability-item difficulty) increase along the two administration periods as expected. The structure measures (threshold between categories) increase between category "2" and "3" (-1.31 to 0.63), then decrease between category "3" and "4" (0.63 to -0.52). This pattern, which occurs in both six month and 12 month data, is not what the Rasch model expected. The infit or outfit mean square statistics of category 1 for both of the administration periods are larger than 1.5, indicating that there is a problematic category which does not function as expected.

Figure 11 presents the category probability curves of CRI. Both of the category probability curves for six month and 12 month show that category "3" is the problematic category. The curve for category "3" does not have a peak as the curves of other categories. In addition, category "3" has less probability of being selected by respondents with average ability (0) than categories "2" and "4". While as expected, the respondents with average ability should have a higher probability of choosing category "3". This pattern of response suggests that these items may function better with a four-point format than a five-point format.

By observing the category probability curves, the research identifies two ways to collapse the level of categories for improvement: 1) Combing level 4 to 3 (12335); or 2) combing level 3 to 4 (12445). Both of the collapsed responses generate higher person reliability and person separation indices than the original pattern, except the separation of the collapsed response "12445" in the 12 month data (2.14 vs. 2.13). The reliability/separation of the collapsed response "12335" is 0.87/2.59 and 0.89/2.83 for six month and 12 month assessments respectively. The reliability/separation of the collapsed

6 Month



12 Month

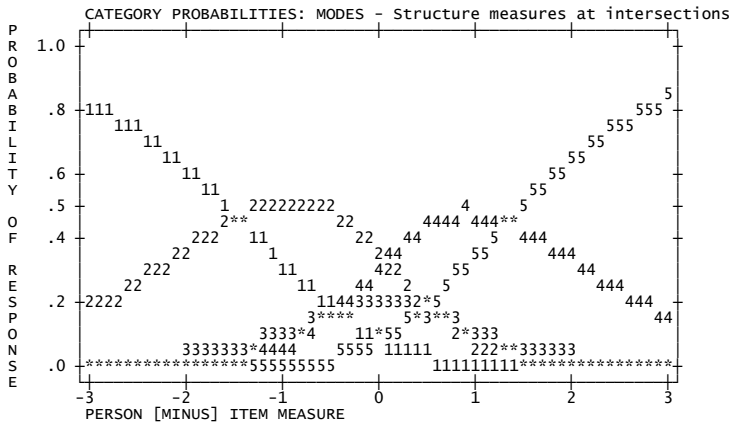


Figure 11. Category probability curves for CRI (12345)

response “12445” is 0.80/1.98 and 0.82/2.13 for six month and 12 month data, while the original reliability/separation was 0.79/1.93, and 0.82/2.14. The four-point collapsed response “12335”has larger person reliability and separation than other formats.

However, the category probability curves for “12335” do not function as Rasch model expected (Figure 12). By checking the meaning of the categories, the researcher found that category “3” represents “not sure”. This category may not align with other categories to be treated as ordinal. Therefore, the researcher conducted the Rasch analysis on CRI by converting all “3” category to missing values. The new analysis reports 0.77 and 0.79 person reliability for the two periods, which are even lower that the reliability of original format (12345). However, the new analysis presents better optimized category probability curves which estimate the category probability as expected (Figure 13).

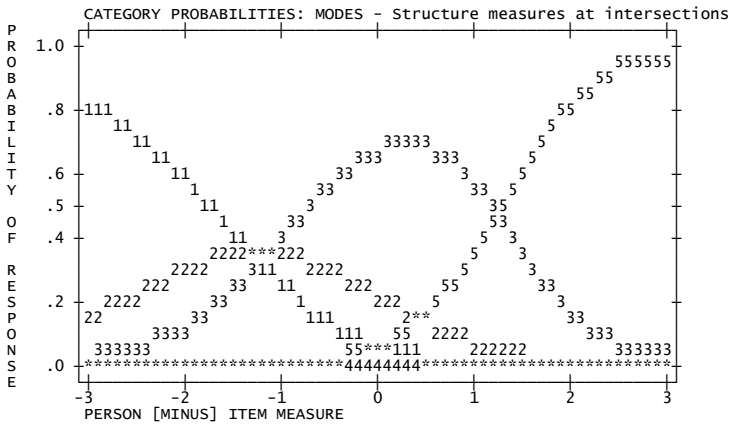
Self Stress (SSI)

Model fit and Reliability

Table 11 presents the fit statistics for SSI items. The internal consistency (Cronbach’s α) is good with 0.85 for both the six month and 12 month data respectively. Rasch person reliability numbers are lower than Cronbach alphas; they are 0.81 and 0.79 for six month and 12 month data respectively.

Item infit/outfit are 1.02/1.10 and 1.01/1.02 for six month and 12 month periods respectively, indicating the Rasch model fits the data. Person separation indices are 2.08 and 1.92 for the two periods, suggesting that SSI can only discriminate two strata of participants. Item separation /reliability are 0.98/6.82 and 0.98/6.43 respectively for the

6 Month



12 Month

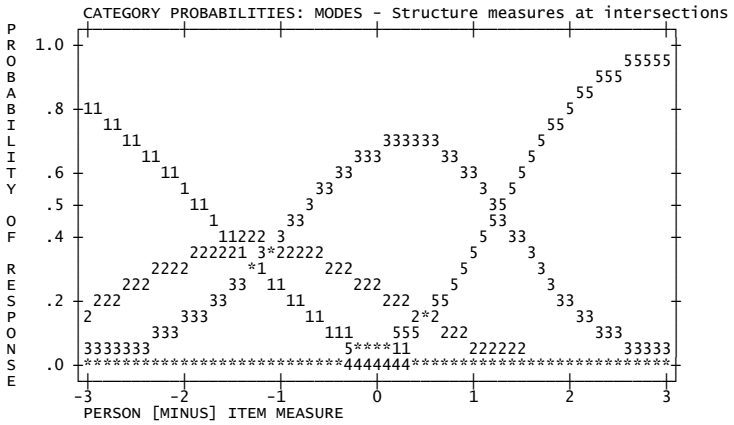
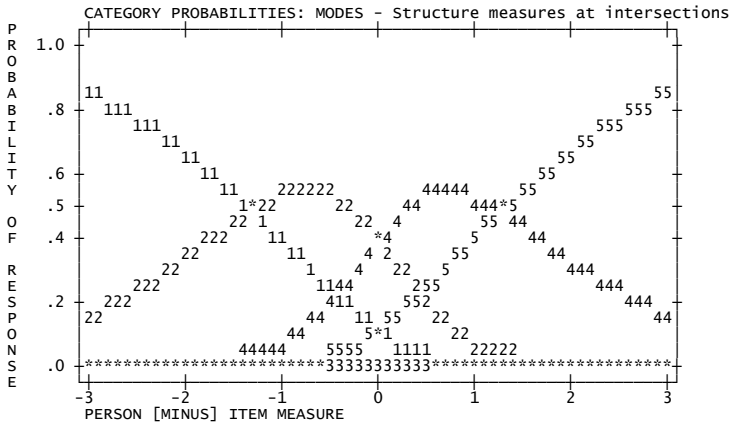


Figure 12. Category probability curves for CRI (12335)

6 Month



12 Month

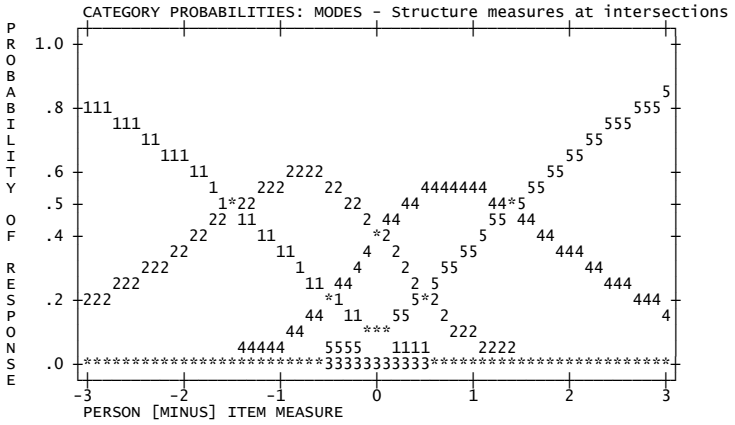


Figure 13. Category probability curves for CRI (12445)

Table 11

Summary of person and item statistics for Self-Stress Index (SSI)

	Item infit	Item outfit	Mean person ability	Mean item difficulty	Person reliability	Person separation	Item reliability	Item separation	Cronbach alpha
6 Month	1.02	1.1	0.85(1.25)	0.00(0.42)	0.81	2.08	0.98	6.82	0.85
12 Month	1.01	1.12	1.01(1.24)	0.00(0.42)	0.79	1.92	0.98	6.43	0.85

two periods, which means that this measure can discriminate approximately six or seven levels of difficulty.

No items were found with infit.MSQ greater than 1.5 (Table 12).There is no item with item-total correlations less than 0.3 in both six month and 12 month periods. The mean logit item difficulty is 0.0 for the two periods, and the mean person ability is 0.85, and 1.01 respectively (Table 11). The mean person ability is much closer to mean difficulty than it was for the CRI. Like the CRI, this result also indicates that those participants tend to choose the higher level of scale categories (e.g. Strongly Disagree), but the tendency is not as strong as with CRI, although they are derived from same measure (i.e. PSI). In addition, the observed count in Table 14 shows that the count of categories "4" and "5" account for nearly 70% of the total number of counts. Categories 1, 2, and 3 only account for around 30%. The patterns are similar for both six month and 12 month data.

Unidimensionality and Local Independence

The results of the conventional Principal Component Analysis (PCA) and Rasch Principal Component Analysis (RPCA) are presented in Table 13. In PCA, The first factor of the SSI measure explains 38.5% and 39.3% of the variance respectively in six month and 12 month periods. All items had loadings larger than 0.40 for the two periods.

In RPCA, the Rasch dimension (all items/first factor) explains 42.0% and 41.4% of raw variance in six month and 12 month assessments. The variance explained by items is around three times larger than the variance explained by the 1st contrast (possible second dimension). The eigenvalue for the 1st contrast is around 1.9 and 2.0 for the two

Table 12

Item difficulty, Infit, and Total correlation statistics for SSI

	6 month			12 Month		
item #	Difficulty	Infit MSQ	Item total correlation	Difficulty	Infit MSQ	Item total correlation
1	0.15	0.89	0.533	0.04	0.91	0.523
2	0.87	1.33	0.435	0.84	1.37	0.423
3	-0.24	0.81	0.604	-0.33	0.79	0.615
4	0.19	1.05	0.495	0.32	1.08	0.509
5	0.09	0.91	0.564	0.11	0.83	0.632
6	-0.16	1.12	0.461	-0.02	1.21	0.441
7	0.5	0.90	0.601	0.51	1.01	0.536
8	-0.61	1.28	0.466	-0.54	1.17	0.461
9	-0.49	1.00	0.541	-0.59	0.86	0.586
10	-0.58	0.95	0.501	-0.46	0.99	0.501
11	0.22	1.08	0.510	0.19	1.12	0.499
12	0.04	0.95	0.566	-0.07	0.83	0.635

administration periods (which shows that the second dimension may be comprised of 2 items), which are not enough to comprise the second dimension. Principle factor analysis (PFA) also confirmed the one-dimension structure. No correlation of standardized residual of the items is found that is larger than 0.3. Overall, the assumption of unidimensionality and local independence is not violated by SSI.

Differential Item Function (DIF) analysis

There are not any items with significant DIF contrast that is larger than 0.43. This measure remains invariant across groups.

Table 13

Principle component analysis for SSI

	Principle component analysis(PCA)		Rasch principle component analysis				
Month	Factor1 Eigenvalue (proportion)	Item loading on second factors (>0.4)	Raw variance explained by measures	Unexplained variance(total)	Unexplained variance in 1st contrast	Explained by items	Possible Item of second dimension
6 Month	4.6 (38.5%) items have loading larger than 0.40 on Factor1	4, 5	8.7 (42.0%)	12.0 (58.0%)	1.9 (9.4%)	5.0 (24.0%)	4, 5
12 Month	4.7 (39.3%) items have loading larger than 0.40 on Factor1	2	8.5 (41.4%)	12.0 (56.8%)	2.0 (9.8%)	4.6 (22.3%)	2, 4

Targeting

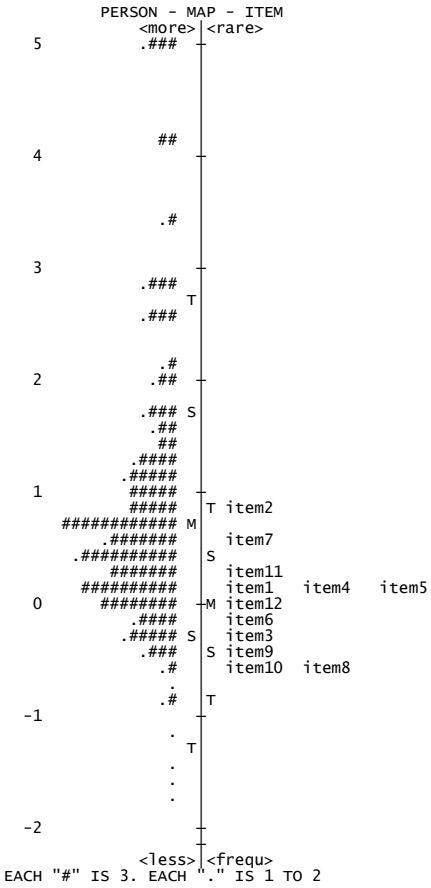
The mean logit item difficulty is 0.0 for the two periods, and the mean person ability is 0.85, and 1.01 (Table 11). The deviation of mean difficulty and ability is around 1 which is "good" for targeting according to Fisher (2007). Figure 14 provides the Wright person-item map on the logit scale. As with the CRI, the figure depicts items and respondents with "high ability/low stress" on the top of the continuum, and those with "low ability/high stress" on the bottom. In both six and 12 month data, the items are most often located between -1 to 1, while the respondents are most often located between -1 to 2 with a bell shape. The items are a little bit "easy" compared to the person's ability (stress level). However, all of the items are located in the continuum of person ability.

The effective ranges of items are from -4 to 5, and the ranges for samples are from -2 to 5 (Figure 15). Although the items are still a little bit "easy" for samples (category 1 is rarely chosen), no ceiling or floor effect was detected. The range of measure is acceptable. In fact, both the Wright map and operational range map show that the range of item difficulty well-matched the range of person ability.

Time invariance

Figure 16 presents the results of Differential Test Functioning (DTF) analysis for 6 month vs. 12 month administrations. All of the items remain inside the boundary of the two 95% confidence bands, suggesting SSI measure function the same way across time. The result of DTF analysis shows that the measure is invariant across the two time points.

6 Month



12 Month

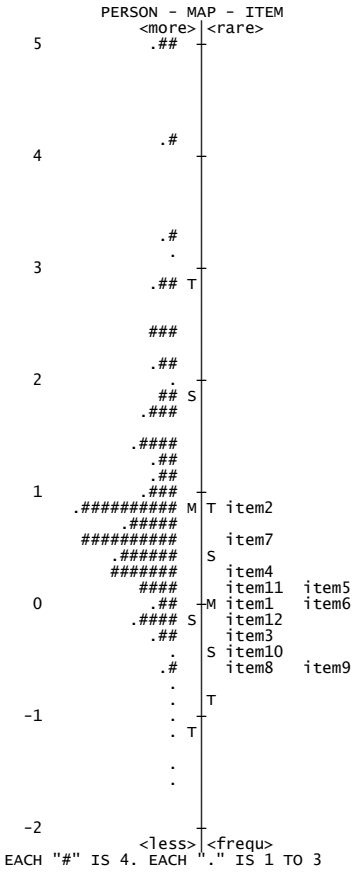
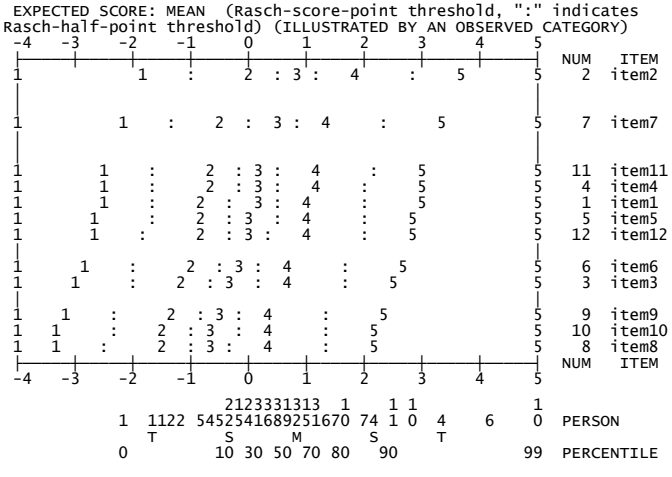


Figure 14. Wright person-item map on the logit scale for SSI

6 Month



12 Month

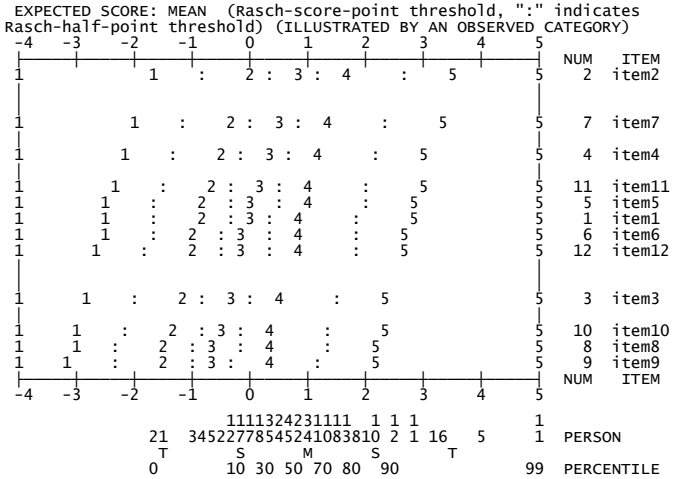


Figure 15. Operational range map for SSI

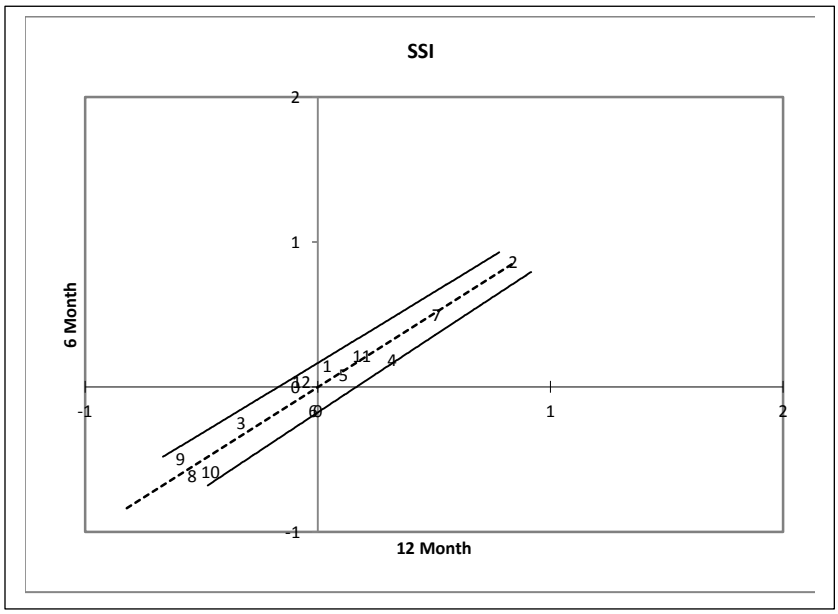


Figure 16. Differential Test Function Analysis for SSI

Category Ordering

Table 14 shows that observed count of response categories "4" and "5" account for about 70% of the occurrences of the categories. The occurrence of categories "4" and "5" add up to 67% and 73% respectively in six month and 12 month data, showing the tendency of respondents to choose higher levels of response categories. The average measure increases along the level of categories as expected. The structure measures increase between categories "2" and "3" (-1.52 to 1.01 in six month, -1.33 to 0.82 in 12 month), decrease between categories "3" and "4" (1.01 to -1.14 in six month, 0.82 to -1.07), and then increase again between categories "4" and "5" (-1.14 to 1.66 in six month, -1.07 to 1.58 in 12 month). The structure measures do not increased along the level of categories as expected. The infit or outfit mean square statistics for category "1" in 12 month is larger than 1.3. These results show that it could be possible to collapse categories for improvement.

Figure 17 presents the category probability curves of the SSI. Both of the category probability curves for six month and 12 month data indicate that the probability for a respondent to choose category 3 is minimal, as it is hardly to observe the occurrence of category 3 in the graph. This may demonstrate that respondents do not treat the "Not sure" as a middle category; instead, they treat it as "Unknown". The structure measures do not increase along the level of category, as the respondents do not treat them as increasing ordinal interval. Therefore, the researcher decided to collapse the categories into a four-point format to see if it can improve the person reliability and separation.

By observing the category probability curves, the researcher identifies two ways to collapse the level of categories for improvement: 1) Combing level 4 to 3 (12335), or

Table 14

Category scale statistics for SSI

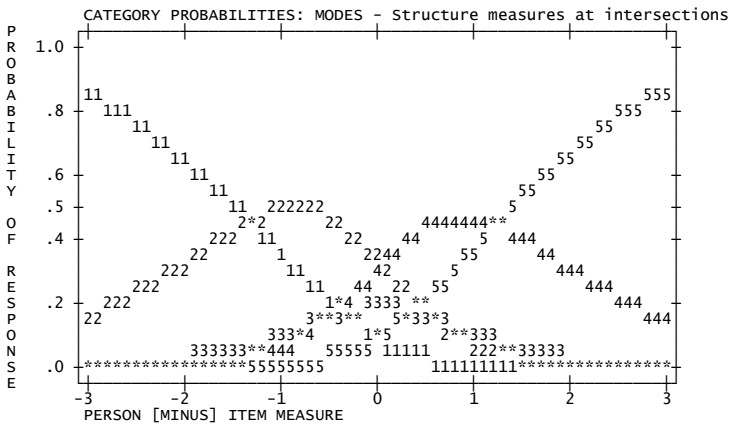
	6 Month				12 Month			
	Observed count (%)	Average Measure	Structure measure	Infit/Outfit MNSQ	Observed count (%)	Average Measure	Structure measure	Infit/Outfit MNSQ
1	242 (6%)	-0.49		1.14/1.43	211 (5%)	-0.34		1.25/1.65
2	809 (19%)	-0.04	-1.52	1.08/1.23	618 (14%)	0.02	-1.33	1.08/1.11
3	326 (8%)	0.22	1.01	1.08/1.19	323 (8%)	0.32	0.82	1.00/1.18
4	1729 (40%)	0.69	-1.14	0.95/0.83	1743 (41%)	0.77	-1.07	0.92/0.84
5	1173 (27%)	1.77	1.66	0.91/0.97	1368 (32%)	1.80	1.58	0.89/0.95

Category optimization

Scale	Person Reliability	Person Separation	Person Reliability	Person Separation
12345	0.81	2.08	0.79	1.92
12335	0.82	2.11	0.82	2.10
12445	0.81	2.05	0.79	1.92

94

6 Month



12 Month

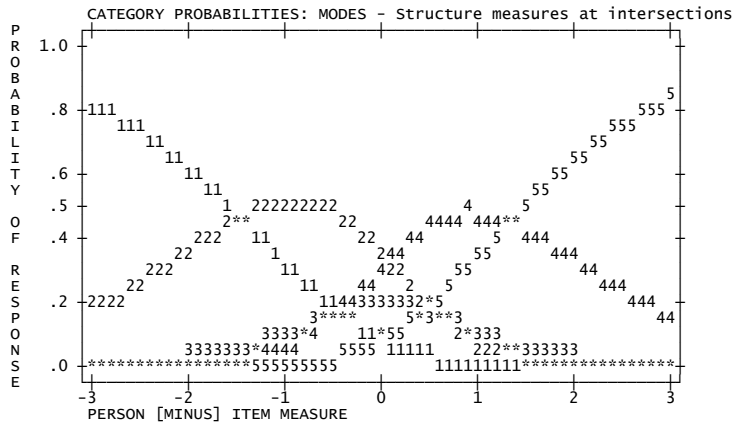


Figure 17. Category probability curves for SSI (12345)

2) combing level 3 to 4 (12445). The collapsed format "12445" generates similar reliability and separation as the original format, 0.81/2.05 in six month, and 0.79/1.92 in 12 month (the original reliability/separation was 0.81/2.08 and 0.79/1.92.) But the collapsed format "12335" does increase the person reliability and separation in both administration periods, and its reliability/separation are 0.82/2.11 and 0.82/2.10 for six month and 12 month periods. The four-point collapsed response "12335" has a larger person reliability and separation than the other format.

However, the category probability curves for "12335" did not perform well (Figure 18), so the researcher decided to covert the "3" (not sure) to missing value. The new analysis reports 0.80 and 0.78 person reliability for the two periods, which are slightly lower than the original formats (0.81 and 0.79). But, the new analysis reports better category probability curves as expected (Figure 19).

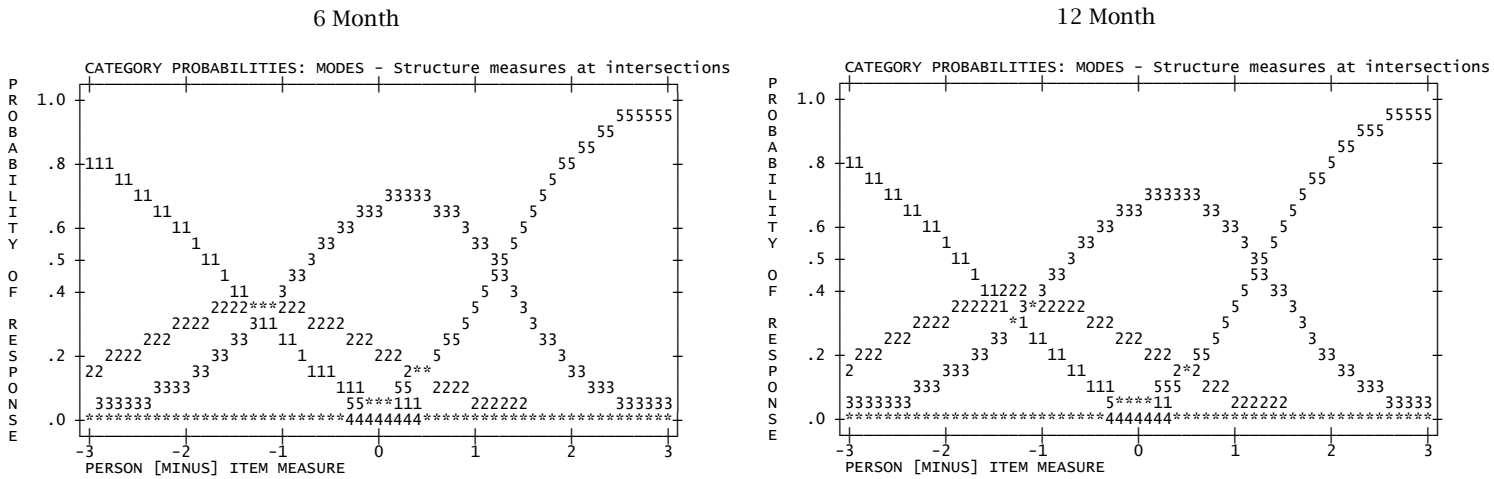


Figure 18. Category probability curves for SSI (12335)

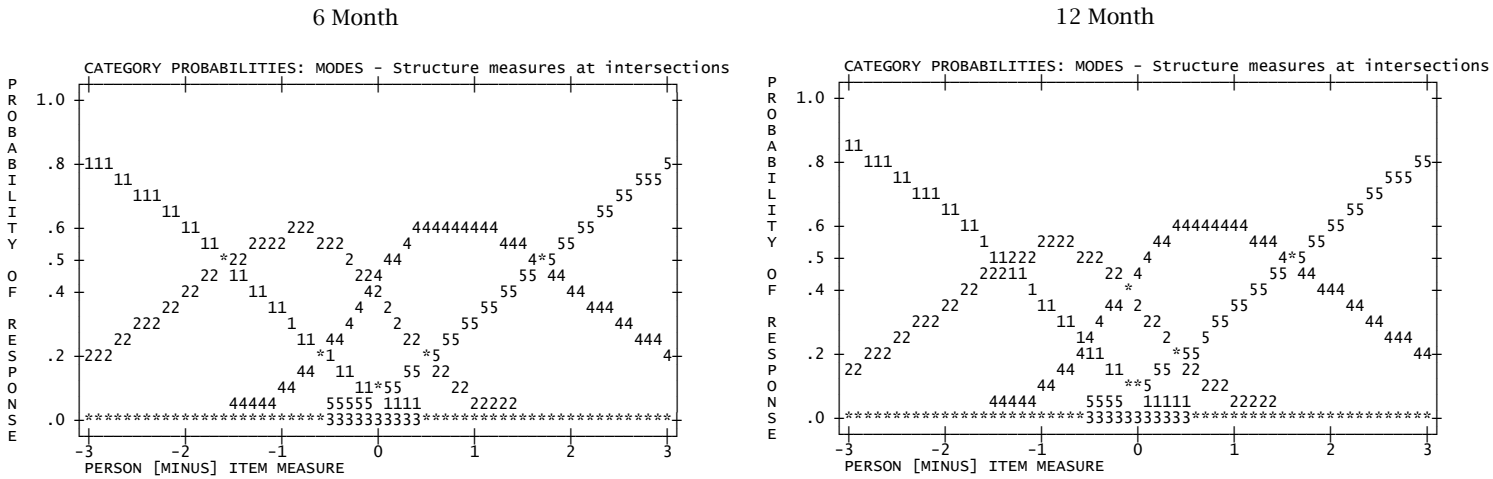


Figure 19. Category probability curves for SSI (12445)

CHAPTER 5

DISCUSSION AND CONCLUSION

In this chapter, each of the research questions will be reviewed, and the results along with interesting findings will be briefly summarized. Strengths, limitation, and implications of the study will be discussed, and conclusions will be given.

Aim 1: To check if the two measures meet the assumption of the Rasch model

Beck Depression Inventory (BDI-II)

For BDI-II, the principal component analysis (Table 3) shows that almost all of the BDI-II items had strong correlation with the first factor. Most of the items had loading greater than 0.4 on the first factor. Meanwhile, the Rasch principal component analysis (Table 3) reports that about 36% of the variance of the standardized residual can be explained by BDI-II measures, and this is slightly higher than the results obtained with the traditional PCA. The eigenvalues of the 1st contrast (second dimension) are around 2, which is not strong evidence to conclude the existence of a second dimension. Principle factor analysis also revealed only one dimension. BDI-II does not violate the unidimensionality assumption. In addition, as no correlation of standardized residual between pairs of items is larger than 0.3, the local independence assumption is not violated.

To evaluate the invariance across group, Differential Item Function (DIF) analysis was conducted for two age groups (<19 and >=19). The results reveal that the BDI-II has several items with significant absolute DIF contrast >0.43, which is a moderate or large violation of invariance (Zwick, Thayer, & Lewis, 1999). They are item 1, 6, 12, and 20 in prenatal period; item 3, 5, 9, 10, 15, 16, 20 in 6 month; item 1, 3, 6, 7, 8, 10, 14, 15, and 20. Most of them are significant in one of the three DIF statistics: Welch, Mantel Haenszel, and ANOVA DIF. Considering that there are close to half of the items with absolute DIF contrast exceeding the criteria (0.43), the BDI-II does not remain invariant across the age group.

Differential Test Function (DTF) analysis was conducted to examine the assumption of time invariance. The result shows that most items are located within the 95% confidence band except items 11(restless or agitated) and 17 (irritatable) in prenatal versus 6 month assessments, and item 10 (Crying) in the 6 versus 12 month assessments. All of the items are related to the issue of emotional control. This seems to indicate that subjects may change their perception regarding emotional control over time.

These results indicate that the BDI II meets the Rasch model assumption of unidimensionality, but not that of invariance. The implications of this are discussed in next section.

Parenting stress index (PSI)

PCA, RPCA, and PFA for the Parenting Stress Index reveal strong evidence for having two sub-domains. PFA reports a factor with 12 items which all related to personal stress questions, and another factor has items about childrearing questions. This result

resembles that in the study conducted by Haskett, Ahern, Ward, and Allaire (2006) and Perez-Padilla, Menendez, & Lozano (2015), which supported the two dimesions of the PSI. Additionally, Rasch PCA cannot support the unidimensionality, as the eigenvalues of the 1st contrast for the three administrative periods are about 5, giving clear evidence of the existence of second dimension. Therefore, the two measures (Childrearing stress and Self stress) are discussed separately below.

Childrearing Stress Index (CRI)

Both of the results of PCA and PPCA clearly identified one dimension for the CRI. RPCA detected a low possibility of existence of second dimension. The PCA indicated that all items with loading >4.0 belong on factor 1 except items 22, 29, 31, 32, 33. Those items have been reported as misfit items in the Rasch analysis across the three time points. Among them, item 22, 32 and 33 are not on a Likert scale, and so respondents may not cognize those questions in a similar way as they do those on a Likert scale.

In the DIF analysis for CRI, item 31(hard to establish schedule) and item 32 (hard to get/stop kids to do things) are significant and with DIF contrast >0.43 across the two periods. Item 33 (number of things bothering) had significant DIF contrast (>0.64) in the 6 month administration. These items (29, 31, 32, and 33) are also misfit items (See Table 7). The DTF analysis (Figure 10) also detected that item 33 (Number of things bothering) was the item that violates the assumption of time invariance. Overall, CRI does not violate the assumption of unidimensionality and invariance assumptions, However, the PCA, PPCA, DIF, DTF analysis repeatedly detected the same abnormal items, thus the

revision of those questions for future research are necessary to improve the quality of the measure.

Self Stress Index (SSI)

Both PCA and RPCA support the assumption of unidimenionality of the SSI. All item have loading >4.0 in first factor in PCA. In RPCA, The eigenvalue for the 1st contrast is around 2, which is not enough for a second dimension. No items have a DIF contrast large than 4.0, and no items appeared outside of the 95% boundary in the DTF analysis. The results of DIF and DTF show that the measure is invariant across age group and time. Thus, both Rasch assumptions are met for the SSI.

Aim 2*:* To check if Rasch model and CTT yield equivalent results

Cronbach's α versus Rasch person reliability

All of the Rasch person reliability indexes for the measures in this study are lower than their corresponding Cronbach's alpha coefficients. For example, the Cronbach's alpha is 0.89 for the 12 month BDI-II administration, but the Rasch person reliability is only 0.71, and person separation is only 1.58 (Table 1). Higher Cronbach's alphas are not related to higher person reliability. For example, Cronbach's α for the CRI is about 0.88 in the 6 month assessment, while for SSI it's around 0.85. Using the Rasch model, however, the person reliability for the two measures are similar: 0.79 and 0.81 (see Table 6 and Table 11). Overall, the Rasch person reliability index seems more conservative than Cronbach's α when assessing reliability. Lincare (1997) believed that Cronbach's α "is an index of the repeatability of raw scores, misinterpreted as linear measures", and it usually overestimates the reliability of a measure.

To some extent, the value of person reliability may reflect if a construct targets a sample well or not. For example, the Cronbach's α for SSI is around 0.85, which is similar to BDI-II, the person reliability of SSI (0.79/0.82) is larger than BDI-II's person reliability (0.71/0.75/0.79). By observing the Wright map and operational range for the two constructs, it appears that the items for the SSI cover the continuum of person ability better than do the BDI-II items. Thus, SSI targets the sample better than BDI-II, and this may explain why SSI has higher person reliability than BDI-II. Linacre (2012) pointed out that larger sample variance and good sample-item targeting can improve the person reliability of a measure.

PCA versus RPCA dimensionality

For BDI-II, PCA shows that almost all of the BDI-II items had strong correlations with the first factors (most of items had loadings greater than 0.4 on the first factors). RPCA reports that the 1st contrast (possible second dimension) is comprised of 2 items, which are not enough to indicate the existence of second dimension. PCA also reveals that the variances that can be explained by the the first factor are around 30%, while RPCA reports that about 36% of variance of the standardized residual can be explained by BDI-II measures (this is slightly higher than the results of the traditional PCA). Meanwhile, the PCA finds that the items which load on other factors (with loading >0.4) are items 15 (loss of energy), 16 (change of sleeping pattern), 18 (loss of appetite), 20 (fatigue), or 21(loss of interest to sex). PRCA detects a similar dimension structure as PCA for BDI-II; it reports that the possible items for the second dimension are 15, 16, 18, and 20, and item 21 is misfit with the infit.MSQ larger than 1.5in prenatal data collection

periods. Both PCA and RPCA analysis indicated similar dimension structures across the three administrative periods.

The conclusion of this study that BDI-II has only one dimension may be controversial. Wang and Gorenstein (2013) conducted a comprehensive literature review on the research studying BDI-II's psychometric properties. They found that the articles they reviewed reported 2 or 3 dimensions for BDI II. However, several Rasch studies claim the BDI-II is unidimensional. For example, Siegert, Tennant, and Turner-Strokes (2010) examined BDI-II in a neurological rehabilitation sample, and concluded that the BDI-II demonstrated unidimensionality with several misfit items: Crying, sleep pattern, and lost interest to sex. Their result was similar to the results of PCA and RPCA in this study. Lambert et al. (2015) also claimed that BDI-II was unidimensional in a sample with cancer, and they had not found misfit items.

For Childrearing Stress (CRI), PCA shows that the variances that can be explained by the first factor are about 33.5% for both administration periods: six month and 12 month (see Table 8). RPCA reports that about 37% of variance is explained by the measure. The items in PCA loading on second factors are item 28 (Bother me a lot), 29 (React strongly), 30 (Upset easily), 31(Hard to establish schedule), or 32 (Hard to get/stop kids to do things), while RPCA reports that it could be 29, 30, 31, 32, 34 (Child does things bother me) in both administration periods. Among those items, item 29, 3132, and 33 are misfit items. This result shows that PCA and RPCA present a similar picture when finding the possible second dimension for CRI.

For Self Stress (SSI), all of the items in first factor in PCA analysis have loading >0.4 (see Table 13). The items with loading >0.4 on the second factor are item 2 (Give up

life for children), 4 (Unable to do new things), and 5 (Unable to do things I like). Rasch analysis reports similar results showing that there is only one dimension for this construct, and the possible items for 1st contrast is 2, 4 or 5

Overall, the PCA and RPCA can reveal similar dimension structures. The items lying on a possible second dimension are only slightly different between the two analyses. In addition, it is interesting to note that the possible second dimension includes the misfit items. This shows that Wright's method (1996) to allocate misfit items into a second dimension seems tenable. Smith (2002) suggested that using iterative RPCA to identify dimensions by examining fit statistics in each iterative step.

Aim 3*:* Optimize the response categories for the measures using Rasch analysis

Category optimization analysis shows that all of the constructs can collapse into one category for improving person reliability. After collapsing BDI-II's category from "0123" to "0122", the person reliabilities increased about 0.02 points with well-shaped category probability curves. This result shows that more response categories may not be able to guarantee a better reliability; sometimes fewer response categories may be more reliable or efficient for administration, when the scales perform in a monotonically increasing manner.

The person reliabilities for CRI increase about 0.07 points after the categories are collapsed from "01234" to "12335", which is a big improvement. The person reliabilities for SSI also increase (0.01 and 0.02). However, the collapsed categories do not perform as expected because the category "3" (not sure) may not perform as an ordinal level variable. The collapsed categories "12445" perform as the model expected and generate

"better" category probability curves than "12335". Meanwhile, treating category "3" as a missing value does not increase person reliability. However, it does generate better functioned category probability curves. Therefore, excluding option "3" may be another way of optimizing the response categories. The small increase in reliability is less important than the scales performing in an acceptable manner.

In summary, Rach analysis does provide a way for choosing optimal response categories and improve the function of scale categories.

Other Findings

In addition to the above results, there are some interesting findings regarding the measures themselves. BDI-II, the measure designed for clinical samples originally, may not be appropriate to evaluate the respondents in the parenting study. The respondents in this study are nested in a range of lower depression, and the BDI-II is not competent enough to discriminate them. A set of special scales may be needed for evaluating depressive symptoms in this kind of population who are homogenous with regard to same sex, similar age, and facing similar life event-- a new baby.

Although CRI has an acceptable of range for targeting subjects, those samples are nested in a range of lower stress, which indicates that respondents tend to choose the higher levels of the response categories (Disagree and Strongly Disagree). Items getting unanimous "Disagree" or "Strongly Disagree" from respondents may not have enough competence to discriminate subjects. The tendency to choose higher level of response categories may be caused by the questions within the instruments, which ask if those parents think that their child will be difficult for them. Those questions may not be

appropriate for the subjects in this study, as they are expecting a baby, or just having a baby. The joy of having a baby may overcome any uneasy feelings, which makes them tend to disagree the questions about child in the measure.

It is very interesting to notice that, although CRI and SSI are from the same instrument, SSI is the better measure. It has fair reliability, and targets the samples very well (See Figure 14 and Figure 15). DIF and DTF also show that this scale functions similarly across age groups and over time. Besides the targeting problem of CRI, CRI also has several special items with different response formats (i.e. item 22, 32, 33. See Appendix B), those items unavoidably are identified by Rasch analysis as misfit items. This indicates that those response formats are not ordinal as the designer assumed. Researchers have to be cautious in using different response formats in one instrument and should allocate a value for the response category that is the same as other formats.

Strengths of Study

This is the first time that a Rasch analysis has been applied to validate the two measures in the Parenting for the First-time Project. This study has conducted a comprehensive analysis in term of the two instruments, including unidimensionality, reliability, group and time invariance, and category optimization.

Compared with classical test theory, the results of the study provide much more detail on the relative distribution of person ability and item difficulty so that we may understand and evaluate more precisely the competence of the measure to discriminate. In addition, the study also examined the distribution of each response category, which provides firsthand information for researchers to reexamine the way of designing the

questionnaire and asking questions. Furthermore, the unique aspect of Rasch analysis allows us to evaluate the function of a measure across groups and over time to ensure the quality of a measure when administered among different populations and over time.

Limitations of Study and Future Research

This study has two main limitations. First, because of the small number of measures, this study did not conduct the Rasch analysis as some researchers suggested (Duku et al., 2013; Wright & Linacre, 1994; Yu, 2013) to rerun the model several times by deleting the misfit items. In this study of measurement validation, deleting misfit items would decrease the person reliability; therefore, this study only interprets the results of the first round of Rasch analysis. However, the researcher would like to suggest use of the iterative steps for a study which is going to develop an instrument with a large amount items chosen from an item bank.

Second, because of the space and time restriction, this study did not conduct further Rasch analysis for optimized response categories. In fact, there are may be more choices for collapsing response categories by observing the category probability curves. Future research can be done to compare more varieties of collapsed response categories for better optimization. Meanwhile, further Rasch analysis can be done for the optimized measures for detailed results in terms of fit statistics, item difficulty, and person reliability. It would be very interesting to evaluate the difference of the item and person estimated between scales with original and optimized response categories.

There is a need for a valid Depression measurement designed specifically for the population of mother with new babies. Although BDI-II has been applied widely in non-

clinical and clinical samples (Wang & Gorenstein, 2013), it had some problems in discriminating the population in this study, and in meeting the invariance assumptions. Researchers need to be cautious when applying this measure repeatedly in a population with homogeneous characteristics. A new depression measure which targets this population should be designed and carefully validated by Rasch analysis.

Several items in CRI are misfit because the format of the responses is different from other items' Likert scale. In this study, deleting those misfit items decreased the reliability of the measure, so those items were retained in the study. However, one reason of their "bad" performance may due to the violation of one of Rating Scale Model assumptions: which assumes that the step measures are equal across the items. Therefore, it is necessary to conduct a Rasch analysis using Partial credit model (PCM) for further investigation. If the PCM still identify them as misfit, new questions should be designed to replace them by using Likert scales, and should be validated by Rasch analysis. In addition, as the sample in this study tend to disagree with the questions in both CRI and SSI, annotation the wording of the response categories needs to be revised to minimize the tendency, and then be validated by Rasch analysis.

Through the DIF analysis, this study found that nearly half of BDI-II items violated the assumption of invariance across age groups. It has been argued that evidence of differential item function can be considered a violation of unidimensionality, although those items which display DIF may fit the model (Tennant et al., 2004). Therefore, although only three of them were identified as misfit by Rasch analysis, the problems with DIF probably indicates that the unidimensionality is not stable across groups. The nonequivalent understanding on this construct across age group may be confound by

sociodemographic status, for example, level of education. The original study (Smith, T., 2015) divided mothers into high (more than 2 years of college) and low (less than 2 years of college) resource status. Low resource included all adolescents and a subset (n=168) of the adults. In this analysis, both low and high resource mother were combined into one "adult" group and compared to one "adolescent" group. Considering the possible threats to the assumption of unidimesionality by the DIF items, subsequent research should be conducted to examine the cause and impact of the DIFs.

Implications

The primary implication of this study is a change in the perspective on how to use the Beck Depression Inventory (BDI-II) and Parenting Stress Index (PSI) to detect these traits among the population of mothers with new-born babies. This study shows that the BDI-II measure does not performs well in this sample. This sample shows homogeneity when answering certain questions in this measure. And some questions failed to discriminate this sample, which can be attributed to the decreased reliability of the measure. This implies that researchers need to be more cautious when applying a commonly used measure among a group with homogenous characteristics, as the questions designed for a general population may not work for a special group. More comprehensive understanding regarding the special needs and thoughts of this group of mothers is essential for revision and addition of items in a commonly used measure. A suggestion, given the result of the study, is for researchers to conduct a pilot study or administer a focus group to gain in-depth insights regarding depression and stress in the research population.

This study also provides evidence that the Parenting Stress Index (PSI) contains two sub-domains: Childrearing Stress Index (CRI) and Self Stress Index (SSI). Perez-Padilla, Menendez, & Lozano (2015) reported there were two domains in PSI when studying a sample of at-risk mothers. This finding will contribute to the discussion about how to summate the rating scales of the PSI and how to establish the relationship between PSI and other parenting traits. The two dimension finding suggests that it is better to include two variables (CRI and SSI) in the model when studying the relation between parenting stress and other parenting factors, and summate the score for the two constructs separately.

Both Rasch analysis and Classical Test Theory show that there are several items in the CRI that may not contribute to discriminating the sample. Replacing those items to improve the reliability of this construct is highly recommended based on the results of this study. In addition, this study also found that SSI is a robust measure by examining fit statistics, item test function, and group variance. The availability of this robust measure will improve the quality of the research on relationships between SSI and other variables (e.g. self-efficacy) that may affect parenting style and skills. This kind of research will be helpful for developing intervention projects aimed at minimizing parenting stress.

Researchers may find it is a dilemma to choose the number of scale categories for their measure. On one hand, the number of categories should generate enough variance for acquiring good reliability. On the other hand, more scale categories may increase response burden, affect respondent's cognitive motivation, and further increase the response errors (Alwin & Krosnick, 1991). This study provides evidence for researchers to administer a parenting stress measure with fewer response scale categories than

originally designed. The scale categories optimization analysis shows that four scale categories may function as reliably as the five scale categories. This finding is really meaningful not only because a study can save administration costs, but also because it can minimize response burden, and therefore improve the quality of responses and validity of the measure.

In summary, this study not only checks Rasch fit statistics, reliability, and validity, but also examines the distribution of each response category, items, and samples. These firsthand information results are very useful. First, by scrutinizing those abnormal (unfit) items, researchers can get a chance to re-edit those questions to make them more appropriate to the respondents; Second, researchers can examine the relative location between items and samples, and design or select questions that can cover the sample along the trait continuum. Third, by optimizing the scale categorizes, researchers can minimize the respondent's burden and save cost by administrating questionnaire with fewer scale categories.

Conclusion

Rasch analysis is a complementary method to classical test theory (CTT) for evaluating the quality of a measure. In this study, both Rasch and CTT presented similar results in term of reliability and validity. Compared with CTT, Rasch analysis is more conservative in reliability evaluation. Both PCA and RPCA present similar results when assessing unidimensinality. In this study, both of the methods function well when assessing reliability and unidimensionality.

However, Rasch analysis has more advantages over CTT when checking other Thurstone's requirement, such as, linearity, invariance, sample free, and test free, etc. By converting the ordinal score into probability and logistic score, Rasch analysis can help researchers to compare linear scores among different settings, groups, and samples. In addition, Rasch analysis provides more detailed information on person ability, item difficulty, targeting, and misfits, which are helpful for researchers to design suitable questionnaires for the targeted population by considering and checking fit statistics, targeting, and coverage of a measure on the population.

REFERENCES

Abidin, R. R. (1990). *Parenting Stress Index (PSI)*: Pediatric Psychology Press Charlottesville, VA.

Alwin, D. F., & Krosnick, J. A. (1991). The Reliability of Survey Attitude Measurement:The Influence of Question and Respondent Attributes. *Sociological Methods & Research, 20*(1), 139-181. doi: 10.1177/0049124191020001005

Amin, L., Rosenbaum, P., Barr, R., Sung, L., Klaassen, R. J., Dix, D. B., & Klassen, A. (2012). Rasch analysis of the PedsQL: an increased understanding of the properties of a rating scale. *Journal of Clinical Epidemiology, 65*(10), 1117-1123. doi: http://dx.doi.org/10.1016/j.jclinepi.2012.04.014

Andrich, D. (1978). Application of a Psychometric Rating Model to Ordered Categories Which Are Scored with Successive Integers. *Applied Psychological Measurement, 2*(4), 581-594. doi: 10.1177/014662167800200413

Andrich, D. (1996). Category ordering and their utility. *Rasch Measurement Transactions, 9*(4).

Baghaei, P. (1998). Local dependency and Rasch measures. *Rasch Measurement Transactions, 21*(3).

Baker, F. B. (2001). *The basics of item response theory*: ERIC.

Bond, T., & Fox, C. (2007). *Applying the rasch model*: Lawrence Erlbaum Associates London.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences* (Vol. 402): Springer.

Brown, R. L. (2016). Rasch analysis of the WURSS-21 dimensional validation and assessment of invariance. *Journal of Lung, Pulmonary & Respiratory Research, 3*(2).

Chen, H.-f., Wu, C.-y., Lin, K.-c., Chen, H.-c., Chen, C. P. C., & Chen, C.-k. (2012). Rasch Validation of the Streamlined Wolf Motor Function Test in People With Chronic Stroke and Subacute Stroke. *Physical Therapy, 92*(8), 1017-1026.

Clauser, B., & Linacte, J. M. (1999). Relating Cronbach and Rasch reliabilities. *Rasch Measurement Transactions, 13*(2).

Duku, E., Vaillancourt, T., Szatmari, P., Georgiades, S., Zwaigenbaum, L., Smith, I. M., . . . Bennett, T. (2013). Investigating the Measurement Properties of the Social Responsiveness Scale in Preschool Children with Autism Spectrum Disorders. *Journal of Autism and Developmental Disorders, 43*(4), 860-868. doi: http://dx.doi.org/10.1007/s10803-012-1627-4

Duncan, P., Lai, S., Bode, R., Perera, S., & De la Rosa, J. (2003). Stroke Impact Scale-16. A brief assessment of physical function. *Neurology, 60*. doi: 10.1212/01.wnl.0000041493.65665.d6

Eggen, T. J. H. M. (2000). On the loss of information in conditional maximum likelihood estimation of item parameters. *Psychometrika, 65*(3), 337-362. doi: 10.1007/bf02296150

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*: Lawrence Erlbaum Associates London.

Engelhard Jr, G. (1989). *Historical Views of the Concept of Invariance and Measurement Theory in the Behavioral Sciences*: ERIC.

Fendrich, M., Smith, E. V., Pollack, L. M., & Mackesy-amiti, M. E. (2009). Measuring Sexual Risk for HIV: A Rasch Scaling Approach. *Archives of Sexual Behavior, 38*(6), 922-935. doi: http://dx.doi.org/10.1007/s10508-008-9385-2

Fisher, W. P. (2007). Rating Scale Instrument Quality Criteria. *Rasch Measurement Transactions, 21*(1), 1095.

Frankenfield, K. M. (2009). Health belief model of breast cancer screening for female college students. Retrieved from http://commons.emich.edu/cgi/viewcontent.cgi?article=1257&context=theses

Hamilton, C. B., & Chesworth, B. M. (2013). A Rasch-Validated Version of the Upper Extremity Functional Index for Interval-Level Measurement of Upper Extremity Function. *Physical Therapy, 93*(11), 1507-1519.

Haskett, M. E., Ahern, L. S., Ward, C. S., & Allaire, J. C. (2006). Factor structure and validity of the parenting stress index-short form. *Journal of Clinical Child & Adolescent Psychology, 35*(2), 302-312.

Hohensinn, C., & Kubinger, K. D. (2011). On the impact of missing values on item fit and the model validness of the Rasch model. *Psychological Test and Assessment Modeling, 53*(3), 380-393.

Huisingh, C. E., Snyder, S., McGwin Jr., G., & Owsley, C. (2018). A Survey of Older Drivers' Attitudes about Instrument Cluster Designs in Vehicles (Unpublished). Department of Ophthalmology. University of Alabama at Birmingham.

Jackson, A. P. (2000). Maternal Self-Efficacy and Children's Influence on Stress and Parenting Among Single Black Mothers in Poverty. *Journal of Family Issues, 21*(1), 3-16. doi: 10.1177/019251300021001001

Jones, W. P., & Loe, S. A. (2013). Optimal Number of Questionnaire Response Categories. *SAGE Open, 3*(2), 10. doi: DOI: 10.1177/2158244013489691

Lambert, S. D., Clover, K., Pallant, J. F., Britton, B., King, M. T., Mitchell, A. J., & Carter, G. (2015). Making Sense of Variations in Prevalence Estimates of Depression in Cancer: A Co-Calibration of Commonly Used Depression Scales Using Rasch Analysis. *J Natl Compr Canc Netw, 13*(10), 1203-1211.

Las Hayas, C., Quintana, J. M., Padierna, J. A., Bilbao, A., & Munoz, P. (2010). Use of Rasch methodology to develop a short version of the health related quality of life for eating disorders questionnaire: a prospective study. *Health Qual Life Outcomes, 8*, 29. doi: 10.1186/1477-7525-8-29

Lee, S. J., Gopalan, G., & Harrington, D. (2016). Validation of the Parenting Stress Index–Short Form With Minority Caregivers. *Research on social work practice, 26*(4), 429-440.

Lerdal, A., Kottorp, A., Gay, C. L., Grov, E. K., & Lee, K. A. (2014). Rasch analysis of the Beck Depression Inventory-II in stroke survivors: a cross-sectional study. *J Affect Disord, 158*, 48-52. doi: 10.1016/j.jad.2014.01.013

Lietz, P. (2010). Research into questionnaire design. *International Journal of Market Research, 52*(2), 249-272.

Linacre, J. M. (1997). KR-20 / Cronbach Alpha or Rasch person reliability: Which tells the "truth"? *Rasch Measurement Transactions, 11*(3).

Linacre, J. M. (1998). Structure in Rasch residuals: Why principal components analysis (PCA)? *Rasch Measurement Transactions, 12*(2).

Linacre, J. M. (1999). Investigating rating scale utility. *J Outcome Meas, 3*.

Linacre, J. M. (2000). Comparing "Partial Credit Models" (PCM) and "Rating Scale Models" (RSM). *14*(3).

Linacre, J. M. (2002a). Optimizing Rating Scale Category Effectiveness. *Journal of Applied Measurement, 3*(1).

Linacre, J. M. (2002b). What do Infit and Outfit, Mean-square and Standardized mean? . *Rasch Measurement Transactions, 16*(2).

Linacre, J. M. (2009). Local independence and residual covariance: A study of Olympic figure skating ratings. *Journal of applied measurement, 10*(2), 157-169.

Linacre, J. M. (2012). *A User's Guide to W I N S T E P S: M I N I S T E P Rasch-Model Computer Programs*

Lo, C., Liang, W.-M., Hang, L.-W., Wu, T.-C., Chang, Y.-J., & Chang, C.-H. (2015). A psychometric assessment of the St. George's respiratory questionnaire in patients with COPD using rasch model analysis. *Health and Quality of Life Outcomes, 13*(1), 131. doi: 10.1186/s12955-015-0320-7

Lovejoy, M. C., Graczyk, P. A., O'Hare, E., & Neuman, G. (2000). Maternal depression and parenting behavior. *Clinical Psychology Review, 20*(5), 561-592. doi: http://dx.doi.org/10.1016/S0272-7358(98)00100-7

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174. doi: 10.1007/bf02296272

Masters, G. N. (1988). The analysis of partial credit scoring. *Applied Measurement in Education, 1*(4), 279-297.

McDowell, I. (2005). *Measuring health: A guide to rating scales and questionnaires* (3rd ed.): Oxford University Press.

Parenting for the First Time. (2001). Retrieved from http://shaw.nd.edu/research-projects/all-research-projects/parenting-for-the-first-time/

Perez-Padilla, J., Menendez, S., & Lozano, O. (2015). Validity of the Parenting Stress Index Short Form in a Sample of At-Risk Mothers. *Eval Rev, 39*(4), 428-446. doi: 10.1177/0193841x15600859

Puma, J. E. (2007). The psychometric functioning of a modified short form of the parenting stress index: Implications for clinical practice and research (Doctoral dissertation). Retrieved from ProQuest database.

Reise, S. P. (1990). A comparison of item-and person-fit methods of assessing model-data fit in IRT. *Applied Psychological Measurement, 14*(2), 127-137.

Salzberger, & T. (2003). When gaps can be bridged. *Rasch Measurement Transactions, 17*(1), 2.

Sanders, M. R., & Woolley, M. (2005). The relationship between maternal self-efficacy and parenting practices: Implications for parent training. *Child: care, health and development, 31*(1), 65-73.

Shin, S.-H. (2009). How to treat omitted responses in Rasch model-based equating. *Practical Assessment Research & Evaluation, 14*(1).

Siegert, R. J., Tennant, A., & Turner-Stokes, L. (2010). Rasch analysis of the Beck Depression Inventory-II in a neurological rehabilitation sample. *Disabil Rehabil, 32*(1), 8-17. doi: 10.3109/09638280902971398

Smith, E. V. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *J Appl Meas, 3*(2), 205-231.

Smith, E. V., Conrad, K. M., Chang, K., & Piazza, J. (2002). An introduction to Rasch measurement for scale development and person assessment. *J Nurs Meas, 10*(3), 189-206.

Smith, E. V., Wakely, M. B., de Kruif, R. E. L., & Swartz, C. W. (2003). Optimizing Rating Scales for Self-Efficacy (and Other) Research. *Educational and Psychological Measurement, 63*(3), 369-391. doi: 10.1177/0013164403063003002

Smith, T. L. (2015). The influence of personal, interpersonal, and community factors on the parenting self-efficacy of first time mothers (Doctoral dissertation). Retrieved from ProQuest database.

Snyder, S., & Sheehan, R. (1992). The Rasch Measurement Model: An Introduction. *Journal of Early Intervention, 16*(1), 8.

Tennant, A., & Pallant, J. F. (2007). DIF matters: A practical approach to test if Differential Item Functioning makes a difference. *Rasch Measurement Transactions, 20*(4), 4.

Tennant, A., Penta, M., Tesio, L., Grimby, G., Thonnard, J. L., Slade, A., . . . Phillips, S. (2004). Assessing and adjusting for cross-cultural validity of impairment and activity limitation scales through differential item functioning within the framework of the Rasch model: the PRO-ESOR project. *Med Care, 42*(1 Suppl), I37-48. doi: 10.1097/01.mlr.0000103529.63132.77

Thompson, N. A. (2009). Ability Estimation with Item Response Theory. Retrieved from http://www.assess.com/docs/Thompson_(2009)_Ability_estimation_with_IRT.pdf

Thurstone, L. L. (1931). Measurement ofsocial attitudes. *Journal of Abnormaland Social Psychology, 26*.

Wang, W.-C., & Chen, C.-T. (2005). Item parameter recovery, standard error estimates, and fit statistics of the WINSTEPS program for the family of Rasch models. *Educational and Psychological Measurement, 65*(3), 376-404.

Wang, Y.-P., & Gorenstein, C. (2013). Psychometric properties of the Beck Depression Inventory-II: a comprehensive review. *Revista Brasileira de Psiquiatria, 35*(4), 416-431.

Wilson, M. (2011). Some Notes on the Term: "Wright Map". *Rasch Measurement Transactions, 25*(3).

Wilson, M., Allen, D. D., & Li, J. C. (2006). Improving measurement in health education and health behavior research using item response modeling: introducing item response modeling. *Health Educ Res, 21 Suppl 1*, i4-18. doi: 10.1093/her/cyl108

Wright, B. D. (1996). Comparing Rasch measurement and factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal, 3*(1), 3-24.

Wright, B. D. (1997). A History of Social Science Measurement. *Educational Measurement: Issues and Practice, 16*(4), 33-45. doi: 10.1111/j.1745-3992.1997.tb00606.x

Wright, B. D. (1998). Model selection: Rating Scale Model (RSM) or Partial Credit Model (PCM)? *Rasch Measurement Transactions, 12*(3).

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370.

Wright, B. D., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological measurement, 29*(1), 23-48.

Wright, B. D., & Stone, M. H. (1999). *Measurement essentials*: Wilmington. Wide Range Inc.

Yildiz, S., & Dogan, B. (2011). Self reported dental health attitudes and behaviour of dental students in Turkey. *Eur J Dent, 5*(3), 253-259.

Yu, C. H. (2013). A simple guide to the Item Response Theory(IRT) and Rasch Modeling. Retrieved from http://www.creative-wisdom.com/computer/sas/IRT.pdf

Zhu, W., Timm, G., & Ainsworth, B. (2001). Rasch calibration and optimal categorization of an instrument measuring women's exercise perseverance and barriers. *Research Quarterly for Exercise and Sport, 72*(2), 104-116.

Zwick, R., Thayer, D. T., & Lewis, C. (1999). An Empirical Bayes Approach to Mantel-Haenszel DIF Analysis. *Journal of Educational Measurement, 36*(1), 28.

APPENDIX A

BECK INSTRUMENT

Beck Inventory (Prenatal)

This questionnaire consists of 21 groups of statements. Please read each group of statements carefully, and then pick out the <u>one statement</u> in each group that best describes the way you have been feeling during the <u>past two weeks</u>, <u>including today</u>. Put a check beside the statement you have picked. If several statements in the group seem to apply equally well, check the highest number for that group. Be sure you do not choose more than one statement for any group, including Item I16 or Item I18.
Present <u>TURQUOISE, YELLOW, GREEN, BLUE, and ORANGE</u> cards

I1. 0 □ I do not feel sad.
1 □ I feel sad much of the time.
2 □ I am sad all the time.
3 □ I am so sad or unhappy that I can't stand it.

I2. 0 □ I am not discouraged about my future.
1 □ I feel more discouraged about my future than I used to be.
2 □ I do not expect things to work out for me.
3 □ I feel my future is hopeless and will only get worse.

I3. 0 □ I do not feel like a failure.
1 □ I have failed more than I should have.
2 □ As I look back, I see a lot of failures.
3 □ I feel I am a total failure as a person.

I4. 0 □ I get as much pleasure as I ever did from the things I enjoy.
1 □ I don't enjoy things as much as I used to.
2 □ I get very little pleasure from the things I used to enjoy.
3 □ I can't get any pleasure from the things I used to enjoy.

I5. 0 □ I don't feel particularly guilty.
1 □ I feel guilty over many things I have done or should have done.
2 □ I feel quite guilty most of the time.
3 □ I feel guilty all of the time.

I6. 0 □ I don't feel I am being punished.
1 □ I feel I may be punished.
2 □ I expect to be punished.
3 □ I feel I am being punished.

I7. 0 □ I feel the same about myself as ever.
1 □ I have lost confidence in myself.
2 □ I am disappointed in myself.
3 □ I dislike myself.

I8. 0 □ I don't criticize or blame myself more than usual.
1 □ I am more critical of myself than I used to be.
2 □ I criticize myself for all of my faults.
3 □ I blame myself for everything bad that happens.

I9. 0 □ I don't have any thoughts of killing myself.
1 □ I have thoughts of killing myself, but I would not carry them out.
2 □ I would like to kill myself.
3 □ I would kill myself if I had the chance.

I10. 0 □ I don't cry any more than I used to.
1 □ I cry more than I used to.
2 □ I cry over every little thing.
3 □ I feel like crying, but I can't.

I11. 0 □ I am no more restless or wound up than usual.
1 □ I feel more restless or wound up than usual.
2 □ I am so restless or agitated that it's hard to stay still.
3 □ I am so restless or agitated that I have to keep moving or doing something.

I12. 0 □ I have not lost interest in other people or activities.
1 □ I am less interested in other people or things than before.
2 □ I have lost most of my interest in other people or things.
3 □ It's hard to get interested in anything.

I13. 0 □ I make decisions about as well as ever.
1 □ I find it more difficult to make decisions than usual.
2 □ I have much greater difficulty in making decisions than I used to.
3 □ I have trouble making any decisions.

I14. 0 □ I do not feel I am worthless.

1 □ I don't consider myself as worthwhile and useful as I used to.

2 □ I feel more worthless as compared to other people.

3 □ I feel utterly worthless.

I15. 0 □ I have as much energy as ever.

1 □ I have less energy than I used to have.

2 □ I don't have enough energy to do very much.

3 □ I don't have enough energy to do anything.

I16. 0 □ I have not experienced any change in my sleeping pattern.

1 □ I sleep somewhat more than usual.

2 □ I sleep somewhat less than usual.

3 □ I sleep a lot more than usual.

4 □ I sleep a lot less than usual.

5 □ I sleep most of the day.

6 □ I wake up 1-2 hours early and can' t get back to sleep.

I17. 0 □ I am no more irritable than usual.

1 □ I am more irritable than usual.

2 □ I am much more irritable than usual.

3 □ I am irritable all the time.

I18. 0 □ I have not experienced any change in my appetite.

1 □ My appetite is somewhat less than usual.

2 □ My appetite is somewhat greater than usual.

3 □ My appetite is much less than before.

4 □ My appetite is much greater than usual.

5 □ I have no appetite at all.

6 □ I crave food all the time.

I19. 0 □ I can concentrate as well as ever.

1 □ I can't concentrate as well as usual.

2 □ It’s hard to keep my mind on anything for long.

3 □ I find I can’t concentrate on anything.

I20. 0 □ I am no more tired or fatigued than usual.

1 □ I get more tired or fatigued more easily than usual.

2 □ I am too tired or fatigued to do a lot of the things I used to do.

3 □ I am too tired or fatigued to do most of the things I used to do.

I21. 0 □ I have not noticed any recent change in my interest in sex.

1 □ I am less interested in sex than I used to be.

2 □ I am much less interested in sex now.

3 □ I have lost interest in sex completely.

APPENDIX B

PARENTING STRESS INDEX

PSI – Short Form

This questionnaire contains 36 statements. Read each statement carefully. For each statement circle the response that best represents your opinion.

Circle the SA if you <u>strongly agree</u> with the statement.

Circle the A if you <u>agree</u> with the statement.

Circle the NS if you are <u>not sure</u>.

Circle the D if you <u>disagree</u> with the statement.

Circle the SD if you <u>strongly disagree</u> with the statement.

For example, if you sometimes enjoy going to the movies, you would circle A in response to the following statement:

I enjoy going to the movies. **SA** **(A)** **NS** **D** **SD**

While you may not find a response that exactly states your feelings, please choose the response that comes closest to describing how you feel. YOUR FIRST REACTION TO EACH QUESTION SHOULD BE YOUR ANSWER. Choose only one response for each statement, and respond to all statements.

SA = Strongly Agree A = Agree NS = Not Sure D = Disagree SD = Strongly Disagree

		(1)	(2)	(3)	(4)	
I1.	I often have the feeling that I cannot handle things very well.	SA	A	NS	D	SD
I2.	I find myself giving up more of my life to meet my children's needs than I ever expected.	SA	A	NS	D	SD
I3.	I feel trapped by my responsibilities as a parent.	SA	A	NS	D	SD
I4.	Since having this child, I have been unable to do new and different things.	SA	A	NS	D	SD
I5.	Since having a child, I feel that I am almost never able to do things that I like to do.	SA	A	NS	D	SD
I6.	I am unhappy with the last purchase of clothing I made for myself.	SA	A	NS	D	SD
I7.	There are quite a few things that bother me about my life.	SA	A	NS	D	SD
I8.	Having a child has caused more problems than I expected in my relationship with my spouse (male/female friend).	SA	A	NS	D	SD
I9.	I feel alone and without friends.	SA	A	NS	D	SD
I10.	When I go to a party, I usually expect not to enjoy myself.	SA	A	NS	D	SD
I11.	I am not as interested in people as I used to be.	SA	A	NS	D	SD
I12.	I don't enjoy things as I used to.	SA	A	NS	D	SD
I13.	My child rarely does things for me that make me feel good.	SA	A	NS	D	SD
I14.	Most times I feel that my child does not like me and does not want to be close to me.	SA	A	NS	D	SD
I15.	My child smiles at me much less than I expected.	SA	A	NS	D	SD
I16.	When I do things for my child, I get the feeling that my efforts are not appreciated very much.	SA	A	NS	D	SD
I17.	When playing, my child doesn't often giggle or laugh.	SA	A	NS	D	SD

I18.	My child doesn't seem to learn as quickly as most children.	SA	A	NS	D	SD
I19.	My child doesn't seem to smile as much as most children.	SA	A	NS	D	SD
I20.	My child is not able to do as much as I expected.	SA	A	NS	D	SD
I21.	It takes a long time and it is very hard for my child to get used to new things.	SA	A	NS	D	SD
I22.	For the next statement, choose your response from the choices "1" to "5" below. I feel that I am: 1) not very good at being a parent 2) a person who has some trouble being a parent 3) an average parent 4) a better than average parent **5)** a very good parent	1	2	3	4	5
I23.	I expected to have closer and warmer feelings for my child than I do and this bothers me.	SA	A	NS	D	SD
I24.	Sometimes my child does things to bother me just to be mean.	SA	A	NS	D	SD
I25.	My child seems to cry or fuss more often than most children.	SA	A	NS	D	SD
I26.	My child generally wakes up in a bad mood.	SA	A	NS	D	SD
I27.	I feel that my child is very moody and easily upset.	SA	A	NS	D	SD
I28.	My child does a few things which bother me a great deal.	SA	A	NS	D	SD
I29.	My child reacts very strongly when something happens that my child doesn't like.	SA	A	NS	D	SD
I30.	My child gets upset easily over the smallest thing.	SA	A	NS	D	SD
I31.	My child's sleeping or eating schedule was much harder to establish than I expected.	SA	A	NS	D	SD

I32.	For the next statement, choose your response from the choices "1" to "5" below. I have found that getting my child to do something or stop doing something is: 1) much harder than I expected 2) somewhat harder than I expected 3) about as hard as I expected 4) somewhat easier than I expected 5) much easier than I expected	1 2 3 4 5
I33.	For the next statement, choose your response from the choices "10+" to "1-3." Think carefully and count the number of things which your child does that bother you. For example: dawdles, refuses to listen, overactive, cries, interrupts, fights, whines, etc.	10+ 8-9 6-7 4-5 1-3
I34.	There are some things my child does that really bother me a lot.	SA A NS D SD
I35.	My child turned out to be more of a problem than I had expected.	SA A NS D SD
I36.	My child makes more demands on me than most children.	SA A NS D SD

APPENDIX C

ROTATED FACTOR PATTERN OF PSI

Rotated Factor Pattern				
	6 Month		12 Month	
	Factor1	Factor2	Factor1	Factor2
item1	0.19398	0.51741	0.24220	0.47141
item2	0.01123	0.45330	0.05651	0.47277
item3	0.20394	0.64195	0.16730	0.57510
item4	0.11557	0.55067	0.11071	0.52291
item5	0.14321	0.62315	0.18960	0.62782
item6	0.11189	0.49338	0.15331	0.42622
item7	0.00417	0.65677	0.08809	0.59857
item8	0.18208	0.50783	0.21369	0.45059
item9	0.19121	0.49709	0.30792	0.52514
item10	0.22048	0.47919	0.28418	0.39744
item11	0.10875	0.54897	0.16708	0.47187
item12	0.16923	0.60437	0.20436	0.58671
item13	0.48505	0.21065	0.43793	0.11219
item14	0.61144	0.25722	0.63378	0.21004
item15	0.46486	0.24634	0.64659	0.15382
item16	0.57764	0.35439	0.65866	0.26228
item17	0.68706	0.10146	0.66053	0.11442
item18	0.72759	0.05078	0.71286	0.16674
item19	0.72921	0.01933	0.79338	0.11030
item20	0.69356	0.07552	0.68192	0.10658
item21	0.54149	0.24721	0.62812	0.25878
item22	0.12314	0.20016	0.06152	0.27590
item23	0.48489	0.31989	0.48248	0.17983
item24	0.67704	0.21085	0.61461	0.24663
item25	0.57150	0.22851	0.64108	0.27198
item26	0.55003	0.03731	0.63933	0.19720
item27	0.61260	0.14629	0.58689	0.26797
item28	0.51735	0.15879	0.46672	0.42959
item29	0.20269	0.14851	0.09043	0.42653
item30	0.54643	0.09812	0.37781	0.38554
item31	0.18330	0.13307	0.18202	0.23610
item32	0.06534	0.13547	0.05817	0.24601
item33	0.12286	0.02157	0.12892	0.16263
item34	0.40857	0.28169	0.37195	0.41396
item35	0.64293	0.21528	0.60118	0.28297
item36	0.40731	0.2285	0.43929	0.36627

APPENDIX D

INSTITUTIONAL REVIEW BOARD APPROVAL FORM

THE UNIVERSITY OF ALABAMA AT BIRMINGHAM

Institutional Review Board for Human Use

Exemption Designation
Identification and Certification of Research
Projects Involving Human Subjects

UAB's Institutional Review Boards for Human Use (IRBs) have an approved Federalwide Assurance with the Office for Human Research Protections (OHRP). The Assurance number is FWA00005960 and it expires on November 8, 2021. The UAB IRBs are also in compliance with 21 CFR Parts 50 and 56.

Principal Investigator:	HUANG, LEI
Co-Investigator(s):	
Protocol Number:	**E161202007**
Protocol Title:	*Applying Rasch Analysis to validate a Parenting Self-efficacy Scale*

The above project was reviewed on 12/19/14. The review was conducted in accordance with UAB's Assurance of Compliance approved by the Department of Health and Human Services. This project qualifies as an exemption as defined in 45CFR46.101(b), paragraph 4.

This project received EXEMPT review.

Date IRB Designation Issued: 12/19/14

Margie Lawson, CIP

Designated Reviewer
Chair Designee

Investigators please note:

Any modifications in the study methodology, protocol and/or consent form/information sheet must be submitted for review to the IRB prior to implementation.

470 Administration Building
701 20th Street South
205.934.3789
Fax 205.934.1301
irb@uab.edu

The University of
Alabama at Birmingham
Mailing Address:
AB 470
1720 2ND AVE S
BIRMINGHAM AL 35294-0104

APPENDIX E

APPROVAL TO ACCESS DATASET FORM

THE UNIVERSITY OF ALABAMA AT BIRMINGHAM

Civitan-Sparks Clinics

Touching Lives, Changing Futures

November 16, 2016

Dear IRB,

I, as the PI of the Predicting and Preventing Neglect in Teen Mothers (IRB Protocol X010625008), verify that Lei Huang has my permission to access the data from this project. The data will be placed in a de-identified database and only the variables of interest as listed in the IRB application will be available for analysis for Lei Huang's dissertation research.

Kristi C. Guest, Ph.D.
UAB Assistant Professor, Department of Psychology

Community Health Services Bldg 20
930 20th Street South, Suite 101
205.934.5471
Fax 205.975.2380
www.uab.edu/civitansparks

The University of
Alabama at Birmingham
Mailing Address:
CH19 307
1720 2nd AVE S
BIRMINGHAM AL 35294-2041