AN ABSTRACT OF THE THESIS OF

Isaac J. Washburn for the degree of Master of Science in Human Development and Family Studies presented on February 12, 2009.
Title: Rasch Modeling in Family Studies: Modification of the Relationship Assessment Scale.

Abstract approved:

Alan C. Acock

Measurement is at the heart of any good research project. Although Classical test theory is the most widely used measurement theory, many alternative theories have emerged in the last hundred years. Rasch modeling is one such theory and by taking a different approach to measurement holds the possibility of enhancing measurement development in the field of family science. To test this idea, a common scale of relationship quality, the Relationship Assessment Scale (RAS), was administered and checked for reliability using both classical test theory and Rasch modeling. The scale was shown to be reliable by classical test theory standards, but Rasch modeling showed improvement was possible. The RAS was modified to make it fit the Rasch model and administered a second time to a new sample. The modified-RAS was then tested for reliability using both classical test theory and Rasch modeling. The modified scale was still reliable by classical test theory standards and now reliable by Rasch modeling standards as well. Rasch modeling offers a wealth of information on what works and what does not work in our scale.

Rasch Modeling in Family Studies: Modification of the Relationship Assessment Scale

by
Isaac J. Washburn

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Master of Science

Presented February 12, 2009
Commencement June 2009

Master of Science thesis of Isaac J. Washburn presented on February 12, 2009.

APPROVED:

_____

Major Professor, representing Human Development and Family Studies

_____

Chair of the Department of Human Development and Family Sciences

_____

Dean of the Graduate School

I understand that my thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my thesis to any reader upon request.

_____

Isaac J. Washburn, Author

ACKNOWLEDGEMENTS

This work is the result of many patient people. I need to thank my wife, who is a very good motivator, my committee, for all of their direction, and my fellow graduate students, who had to listen to parts of this thesis several times. Most of all I would like to acknowledge my parents who from a very young age taught me to love learning. Everything I know is built on what all of you have taught me.

CONTRIBUTION OF AUTHORS

Parts of the introduction and results section used in the manuscript were prepared by Isaac Washburn, Chris Dogaru, and Alan Acock and presented at the National Council on Family Relations

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Rasch Modeling in Family Studies:
Modification of the Relationship Assessment Scale

Introduction

When we consider the pace of scientific development across many fields of study, the most rapid periods of development often follow an improvement in measurement. Although many different measurement theories exist, Rasch modeling has the potential to offer substantial advantages over classical measurement theory for developing scales. The development of better measures would only increase our ability to explain family related outcomes and processes.

Measurement error is a serious issue in the research done in family science or any science that depends on statistical results. Measurement error can affect the relationship between variables and decrease the correlation between variables (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003). Measurement error at its worst means a scale is not even measuring what it was intended to measure. To avoid these problems we need to assure that our measures are both reliable and valid. Too often researchers give little or no thought to measurement development or assume that someone else has already tested the measure for reliability and validity and so do not test the scale on their own data for reliability and validity. The common statistical methods we use simply take unreliable and invalid data and give us unreliable and invalid results (Wright, 2006).

The Relationship Assessment Scale (Hendricks, 1988) will be utilized in comparing Rasch modeling and Classical Test Theory. The Relationship Assessment Scale exceeds the usual standards of classical test theory, but it will be shown that there is room for improvement using Rasch modeling.

Classical Test Theory

Classical test theory was developed in response to this problem of getting reliable measures. The main assumption of classical test theory, the theory of true and error scores, has led to several different estimates for the reliability coefficient (Ghiselli, Campbell, & Zedeck, 1981). The theory simply states that the observed score on a test or survey is equal to the true score plus the measurement error.

$$X_i = X_t + e_i$$

X_i is a person's observed score
X_t is a person's true score
e_i is the measurement error

At the same time, classical test theory assumes that the measurement errors are "normally and uniformly distributed in persons, have an expected value of zero and are uncorrelated to all other variables" (Embretson, 5). Cronbach's alpha, Kuder-Richardson formula 20 and Guttman's L_1, L_2, and L_3 are examples of estimates of the reliability coefficient. The idea behind classical test theory is that if we can get measures that are highly correlated over parallel tests then we have a reliable measure. It is important to remember that reliability is simply the ability to consistently get the same score on a measure. Reliability does not guarantee that we are measuring what we think we are measuring, or that we are getting data that would answer research questions. We need to verify the validity of our measures to answer these other questions about our measure.

Cronbach's alpha and a simple factor analysis are commonly considered by most publications in the field of family science to be sufficient in establishing reliability and validity. Although these methods have been empirically shown to be

helpful in determining reliability and validity of our measures, alternative measurement theories may give more information about the reliability and validity of our measures. To really understand what Rasch modeling adds to the process of measurement development it is important to understand the tools commonly used under classical test theory. For this paper will focus on the most common of these tools: Cronbach's alpha.

Reliability coefficient.

The problem of measurement is not a new concern for the field of family science or behavioral sciences in general. Classical test theory has been the dominate model of reliability since approximately 1904 (Ghiselli, Campbell, & Zedeck, 1981). It is usually accepted that Spearmen and Yule were originally responsible for the theory of true and error scores (Ghiselli, Campbell, & Zedeck), the major assumption of classical test theory.

A large part of the success of classical test theory is the fact the theory's assumptions are easily understood and accepted. The theory of true and error scores is an easily understood concept that leads to easily understood tests of reliability. Based off the idea of true and error scores, the reliability coefficient is simply the correlation between parallel tests. This means that the reliability coefficient is our ability to reliably obtain a score on a test. This reliability coefficient can be estimated in a number of ways depending on what your question of concern is. As mentioned earlier, Cronbach's alpha, Kuder-Richardson formula 20 and Guttman's L_1, L_2, and L_3 are all estimates of the reliability coefficient that require only a single administration of a measure to get an estimate of the reliability coefficient.

Each one of these methods has their own strengths and weaknesses, but given the almost exclusive use of Cronbach's alpha in the literature, I will focus on it as the major method of reliability estimation. The major strength of Cronbach's alpha is that it can estimate the reliability coefficient of a measure from a single administration. This strength can easily be seen in the equation for the standardized Cronbach's alpha.

$$\alpha = \frac{N \cdot \bar{r}}{1 + (N-1) \cdot \bar{r}}$$

Here N is equal to the number of items and r is the average inter-item correlation among the items. All we need is a test with a sufficient number of items and a large enough inter-item correlation and we can get a reliable measure. The weakness of Cronbach's alpha under the assumptions of classical test theory is that each item of the test must be considered a parallel test of what you are trying to measure. This means that all N items must have the same mean and standard deviation (Ghiselli, Campbell, & Zedeck, 1981). They must also be correlated with each other to the same degree and are correlated with any other variable to the same degree as well. Different assumptions do exist for Cronbach's Alpha in different measurement theories, but these are the assumptions for classical test theory. Unfortunately, when reporting alpha reliability few people explicitly state under which assumptions they are using Cronbach's alpha. The reliance on Cronbach's alpha has led to some possibly unforeseen consequences. We can see from the equation for Cronbach's alpha that we either need a lot of items, a high inter-item correlation or both. The problems with Cronbach's alpha can easily be seen in figures 1 and 2.

Figure 1. Limits on alpha when few items are available to measure a variable

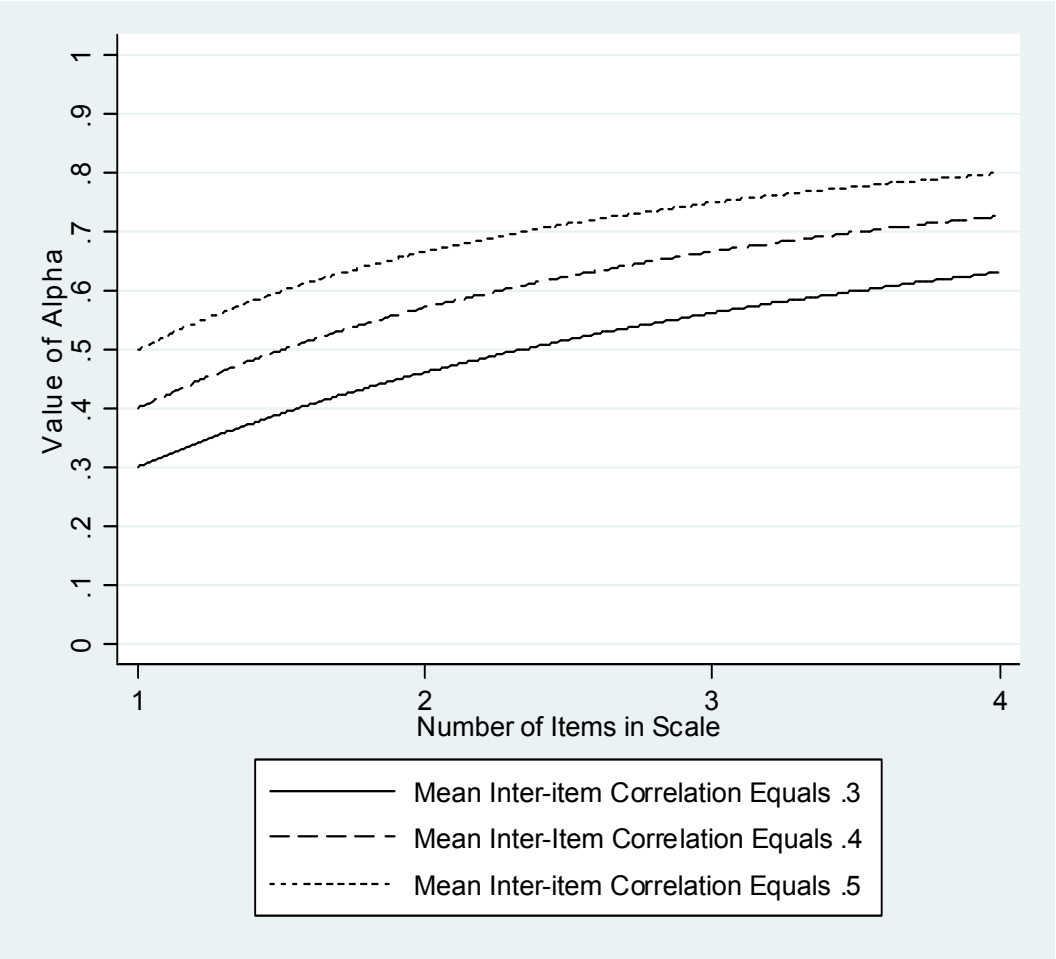
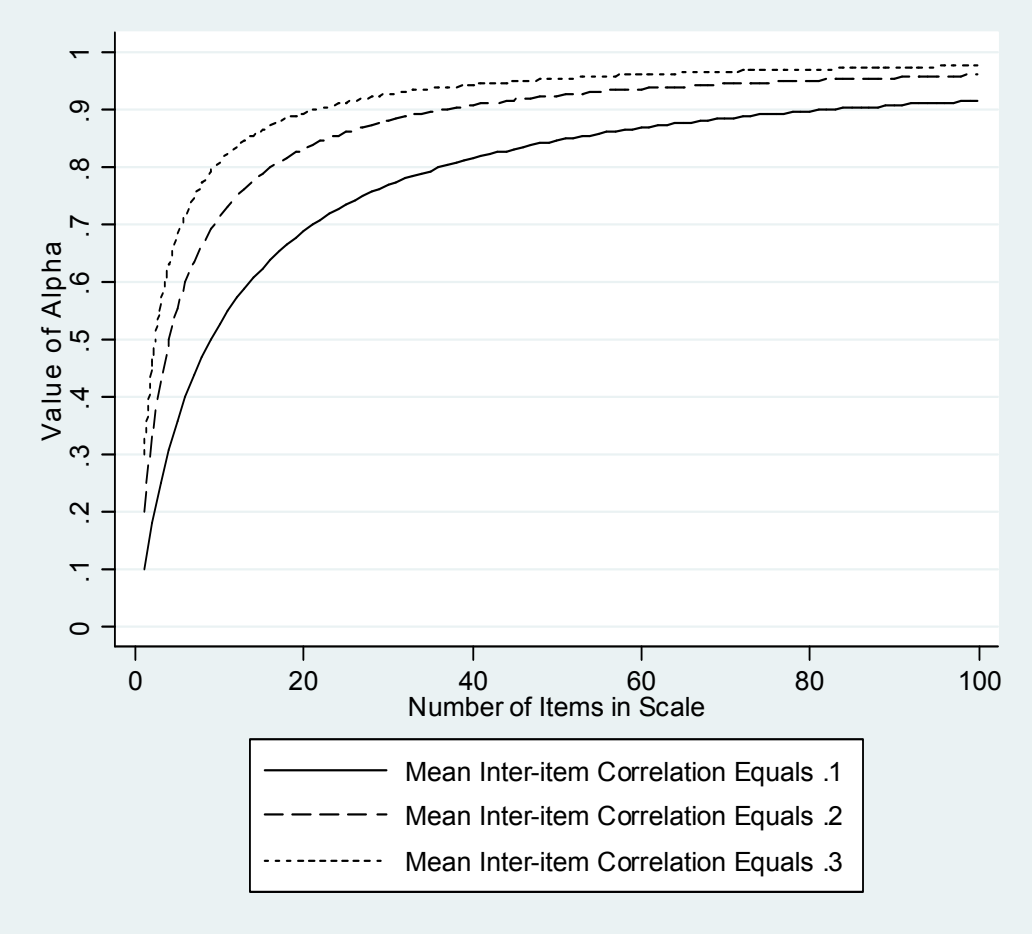


Figure 1 clearly shows what often happens in large studies measuring a large amount of variables with the same survey. We see that even with a reasonable inter-item correlation we have low reliability because of the lack of items or parallel tests to estimate the reliability coefficient. With a small number of items, Cronbach's alpha is really measuring the inter-item correlation between items. We see that in national datasets that purport to measure something with two or three variables, the inter-item correlation would have to be very high for a reliable measure. As the number of items increase, the alpha becomes inflated and the inter-item correlation becomes less important. In Figure 2, we can see the problem with having too many items. With

enough items we can inflate the estimate of the reliability coefficient even though the inter-item correlation is low. Large surveys that measure a single construct are taking advantage of this mathematical fact. Yet a mean inter-item correlation of .1 means a mean change in r-squared of just .01. That suggests that the items share on average one percent common variance, not something that suggests a single construct.

Figure 2. Alpha with a large number of items



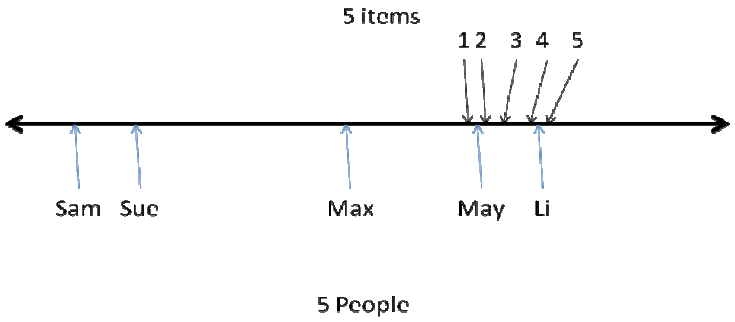
We can see in figure 2 that a survey with one-hundred questions would have to have items with no correlation or negative correlations to get an alpha of less that .8. Reliability in this is case is really a matter of asking enough questions or having enough parallel tests (since this is how classical test theory sees each item). This may

seem like a solution, simply ask enough questions and you have a reliable measure. However, anyone that has ever participated in a long phone survey knows, longer surveys are not always an option. In fact we my have only a few moments of time with each of the subjects in our study.

In family science, we are often restricted to the number of items that we can include in our measures. The people we deal with often do not have the time, patience, or interest to answer a survey with a large number of items, particularly if several different constructs are being measured. This limit on the number of items that are reasonable for a study leaves a high inter-item correlation as our only hope of getting a reliable measure. This is not as big of a problem as it may seem, the answer is simply part of the assumptions of classical test theory: questions need to have similar means and standard deviations. A researcher with little space for questions simply needs to ask questions that are homogenous in content. This leads to a new problem, homogenous items work to limit the variability measurable.

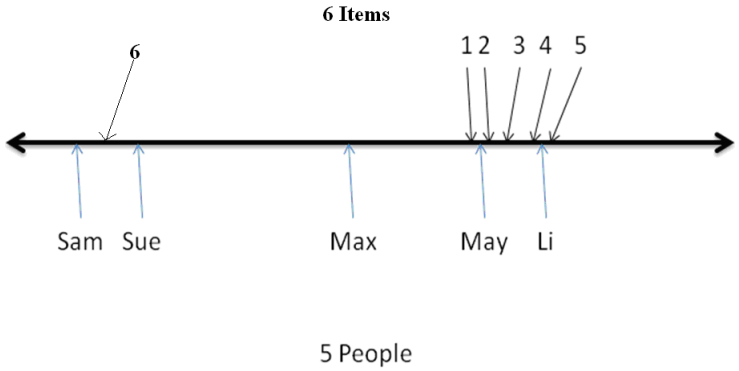
Washburn, Dogaru, and Acock (2007) give an example of this problem in their presentation at the National Council on Family Relations conference. Consider a scale measuring attitudes toward abortion. The range of attitudes would go from acceptance of abortion in any situation to condemnation of abortion in any situation. Measuring this true range of abortion attitudes would be the goal of any scale developed for this population. Homogeneous items that have similar means and standard deviations would focus on only one region of the spectrum of attitudes. Therefore the scale would only differentiate one end of the range of possible attitudes. We can see this more clearly in figure 3.

Figure 3. When items are tapping only a narrow range of a variable



We see that our five items would clearly separate May and Li, but Sam and Sue, who are just as different as May and Li, would probably respond the same to our five questions as would Max who is quite different from everyone else. If we add one more item (Figure 4), we see that now we would be able to get a difference between Sam and Sue. This item would have a different mean and possibly a different standard deviation from the other items. Since under classical test theory, all parallel tests need to have the same mean and standard deviation, and for Cronbach's alpha each item is taken as a parallel test, we would have to question our estimate of the reliability coefficient in this situation.

Figure 4. Adding just one more item



A second solution would be a separate test for people at either end of the spectrum. Questions one through five deals with people on the same end as May and

Li and a second set of questions would deal with people on the same end as Sam and Sue. This might work, but from figure 4 we see that even this solution would leave Max stuck in the middle and would not differentiate people close to Max. It would also be difficult to know beforehand where subjects lie on the spectrum you want to measure. Now we see the problem with using alpha as a measure of reliability, it accurately measures how consistent our measure is, but not how well it differentiates people. Since explaining variance is often the reason we test people, a reliable test that does not capture all of the possible variance in responses is not very useful and could attenuate any real differences between subjects.

Rasch Modeling

The history of Rasch modeling.

Rasch Modeling was developed by Georg Rasch in the 1950s from his work on assessing slow readers and assessing the intelligence of soldiers (Andersen & Olsen, 2001). Rasch's background in mathematics led him to doubt the use of ordinal scales in statistical analysis. He saw the simplicity of measurement in the physical sciences and wondered if a measurement theory could be developed in the social sciences that had the same properties. In particular, Rasch was interested in two ideas, separation and specific objectivity (Andersen & Olsen). Separation occurs when the estimate of person scores and item difficulties can be estimated independently of each other. That means you can estimate either the person scores or the item difficulties without estimating the other (Rost, 2001). This is achieved using conditional maximum likelihood, leading to the two estimates becoming mathematically isolated. The idea of separation led directly to what Rasch called specific objectivity. Specific objectivity

deals with the ability of a measure to be independent of what it measured and vice versa. For example, the metric of a ruler does not depend on what you measure and the height of an individual does not depend on what ruler you use. Rasch sought to give the social sciences the same properties in measurement.

Rasch's first attempts to equate the scores of slow learners that used different tests led him to use the Poisson distribution (Wright, 2006). The Poisson distribution allowed the estimates of the difficulty of the reading tests to be separated from the estimates of the people's reading ability. As Rasch moved on to work with intelligence tests he decided that the logit transformation did a better good of simplifying the process and it produced an interval level scale for the results (Andersen & Olden. 2001). It was for these intelligence tests that the first Rasch model was used. Rasch presented the Dichotomous Rasch Model in 1960 in his paper describing the model and how he used it to improve the intelligence tests that the army used. Since its inception where it was only applicable to dichotomous data, Rasch models have become increasingly capable of handling complex ordinal scales and producing simple interval scales. The family of Rasch models has grown considerably since that time and now covers a wide range of survey types (Rost, 2001). Rasch models are also used to measure a variety of constructs from depression (Pallant & Tennant, 2007;Tang, Wong, Chiu, H., & Ungvari, 2007) to exercise (Safrit, Cohen, & Costa, 1989) to driver license tests. The rating scale model that allows for polytomous data, fully developed by Andrich (1978), is the Rasch model that will be used throughout this paper.

Fitting the Rasch model.

The rating scale model for polytomous data (from hereon simply the Rasch model) allows ordinal scales to be tested for fit to the Rasch model. For a measure to fit the Rasch model it must adequately separate people across whatever construct is being measured. The ideal scale would have items of all difficulty levels. As we will see, this is one area where maximizing alpha using classical measurement theory works against us and Rasch modeling works for us.

Rasch modeling does not disregard alpha or factor analysis. Alpha and factor analysis are merely seen as part of a much larger set of tools. In fact, you can get an idea of the factors of your scale through most programs that do Rasch modeling. Winsteps (the program we use for our application) estimates the proportion of variance in the first principal component and each of the possible secondary components. This is because just as in classical test theory, the assumption of unidimensionality is important in Rasch modeling.

The term “difficulty” is borrowed from the educational literature where Rasch modeling was initially implemented in the U.S. Generally, the term refers to a particular level of the construct we are trying to measure, meaning the level at which the item has a 50 percent chance of being endorsed. For example imagine we have a scale measuring beliefs about abortion, the difficulty of endorsing an item that says abortion is always an option will be “harder” to endorse than one that says abortion is only ok to save the mother or child. More people would be able to endorse the second item, giving it a lower difficulty than the first item. Just like difficulty, a person’s score, their “ability”, represents their place in the spectrum of whatever you are measuring. In the case with our abortion measure, people with a higher score, would

be more accepting of abortion then people with a lower score. A person's "ability" and an item's "difficulty" are measured on the same logit scale.

Running data through the Rasch model initially provides the participants score on your measure as a logit. This can be rescaled to any mean and standard deviation as is done with standard tests (GRE, SAT, and so on). Mathematically, the logit is the natural logarithm of the odds of endorsing an item (or a response option). If p is the probability of an event (and 1-p is the probability that the event doesn't occur), the odds is the ratio $p/(1-p)$, and the logit is the logarithm of the odds:

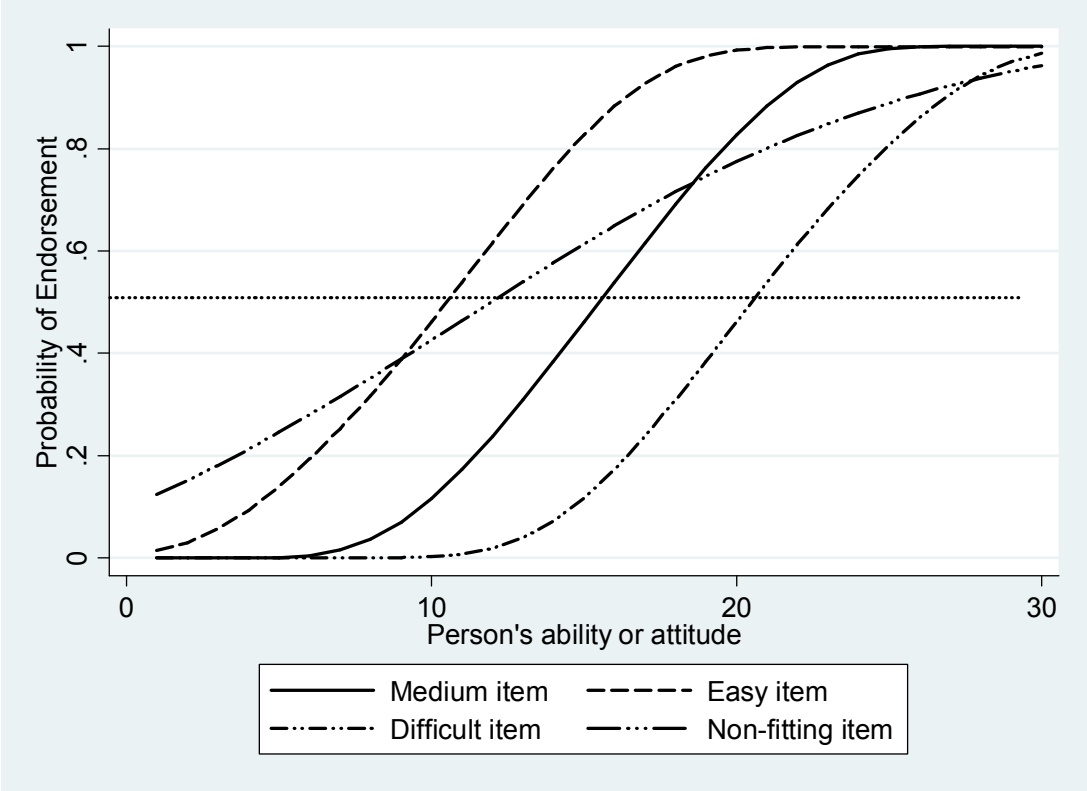
$$\text{logit} = \ln\left(\frac{p}{1-p}\right)$$

The logit can easily be converted to any scale, but it is useful to understand what a logit is. This is easiest to understand in the case of a right or wrong answer. Ideally, you would like to have a scale that can differentiate people between -3 logits and +3 logits. These values correspond to an extremely easy item where 95% of the people select the correct answer to an extremely difficult item where only 5% select the correct answer. In order to differentiate people across this range from those who can endorse only the easiest items to those who can endorse the most difficult item, it is important to have items that cover this range of difficulty.

We can apply this way of thinking to measuring family variables. If we only ask questions that are easy for most people to endorse, then we will not do a good job of getting the true distribution of scores on the variable. This is the problem that we addressed with figure 3 and 4. A range of items with varying difficulties must be asked in order to get a true distribution of scores on a variable.

Another important issue that needs to be considered in measurement is the requirement of many statistical techniques that the variables are measured on an interval or continuous scale. The raw score of most of the instruments used in social sciences do not meet this requirement. A Likert scale, for instance, is an ordinal scale, not an interval one. We can tell that “agree” is greater than “disagree”, but we cannot say that the distance between “disagree” and “strongly disagree” is equal to the distance between “disagree” and “agree”. A goal of the Rasch modeling technique is to transform the ordinal measures into true interval ones. This is achieved by fitting the data to the Rasch Model. If the data fits the Rasch Model, then the difficulty of the items and the ability of the participants on an interval logit scale.

Figure 5. Single Parameter Model and Interval Level Measurement



We have said that Rasch modelers argue that Rasch models produce interval level measurements. It is important to clarify what this means. Figure 5 presents the probability of endorsing a response for four items of varying difficulty. A Rasch model is known as a single parameter model. This parameter differentiates the items on difficulty. Three of the items in Figure 7 fit a one parameter model. The solid line is an item of medium difficulty and there is a dashed easy item to the left of it. A person who has less ability (or a less favorable attitude or belief) is less likely to endorse the medium item than a person with more ability. This is true across the range of ability except where the lines converge on a zero probability or a probability of 1.0.

The line on the right marked by a dash and two dots is a difficult item. A person who has a 50% chance of endorsing this item clearly has more ability than a person who has a 50% chance of endorsing the medium difficulty item or the easy item. A horizontal line marked by dots represent a 50% chance of getting an item correct. We can see how the three good items vary in difficulty. A person with an overall ability of about 10 would have a 50% chance of getting the easy item correct. A person would need an overall ability of around 16 to get the medium difficulty right 50% of the time and a person would need an overall ability of about 21 to get the difficult item correct 50% of the time.

Figure 5 has a non-fitting item marked by a dash and three dots. This item would be rejected by a Rash model because a single parameter representing item difficulty is insufficient to distinguish it from the other items. At first it looks like we could say the non-fitting item is be between the easy and medium item requiring an overall ability of about 16 to have a 50% chance of getting it correct. However, if we

look at people who have low ability, say around a score of 5, they are more likely to get the non-fitting item correct and if we look at people who have high ability, say a score of 20, they are less likely to get the non-fitting item correct than either the easy or medium difficulty items. By excluding non-fitting items, Rasch models can justify saying their scores have interval level properties, but if we include non-fitting items this will not be the case.

Since the goal of Rasch modeling is to get data that fits the model, not a model that fits the data, a wide range of diagnostic tools are available to see what about your participants and your measure do not fit the Rasch model as well as measures of reliability and validity. The goal of Rasch modeling to get the data (and hence your measure) to fit a specific model often flies in the face of conventional statistical analysis. Item Response Theory (IRT) is similar to Rasch modeling in that the goal is to estimate both item difficulty and person ability, as well as other parameters. The one parameter IRT model is mathematically identical to the Rasch model and so people often consider IRT and Rasch modeling the same technique. However, IRT differs from Rasch modeling in that IRT has several models to choose from (with added parameters beyond just item difficulty and person ability) and they choose that model that fits the data best. Although this is more in line with modern statistical modeling, anything beyond the basic IRT model (that only estimates item difficulty and person ability) loses the properties that Rasch felt were so important: interval level scales, separation, and specific objectivity.

Item Response Theory, as distinct from Rasch modeling, is usually implemented as a three parameter model. IRT adds a third parameter that represents

the slope of the function. The slope of the three "good" items, see figure 5, in terms of a Rash model, are the same. However, the slope of the non-fitting item is much flatter than the good items. IRT can also add a "guessing" parameter. This may be important in educational testing situations where there is a substantial chance of getting the item correct just by guessing. Also, some individual items may be easier to guess right than others. The non-fitting item has about a 15% chance of being answered correctly, even if a person has no true ability. The guessing parameter is less important when we are measuring social attitudes because we do not need to worry about guessing a correct answer. IRT advocates feel their approach is best because they do everything they can to develop a model that fits the items. Rasch advocates feel their approach is superior because they have an interval level model that they test on the items. By limiting the final scales to items that fit their model, they have an interval level model. A secondary advantage of the Rasch approach is that it is statistically a much more parsimonious approach (one parameter rather than two or three). Because of this simplicity, Rash models can be estimated on very small samples, as few as 30-50 people, where three parameter IRT models are strictly large sample procedures.

Reliability and Separation.

Rasch modeling has two summary measures of reliability, namely, person reliability and item reliability. These two statistics are the major measures of reliability that is giving by fitting the Rasch model. Reliability, as defined in Rasch modeling, is the observed variance minus the measurement error divided by the observed variance. The observed variance minus the measurement error is an estimate of the real variance.

Reliability = (Observed Variance – Measurement Error) / Observed Variance

The person reliability, according to Linacre (2009), is "equivalent to the traditional 'test' reliability". It represents how likely we will be able to get the same ordering of individuals using a repeated test. It is a measure of replicability. High values mean that a person estimated to have a high score has a high probability of actually having a high true score.

We can increase person reliability by adding people who have high or low scores to our sample, the greater the sample variance on the dimension being measured, the greater the person reliability. Adding additional items will also increase person reliability. Just like with alpha, the more items there are in our scale the greater the person reliability will be. Increasing the number of response options for each item also increases person reliability. Four ordinal response options will give you more variance than 3 ordinal response options. It is important to remember that response options can create problems as well. Take a study of school yard bullying, asking kindergarteners about how often they pick on other children leave you with underutilized response options. These underutilized response options will hurt your fit to the Rasch model and decrease person reliability. The number of response options must be balanced between getting the most possible differentiation as well as not creating unused response options. Selecting items that cover the full range of difficulties can also increase person reliability. Ideally, you select items that match your sample distribution.

In addition to these factors that increase person reliability is the fact that reliability depends very little on the sample size—number of items affects person reliability; not number of people. Rasch modeling can do a good job assessing person

reliability when we have a small sample, e.g., 30-50 people. This is extremely important in developing scales and pretesting scales where multiple iterations of a measure will be developed and tested.

Rasch modeling's item reliability does not have the equivalent in classical measurement theory. Item reliability measures how well you are estimating the difficulty of each item. Like person reliability, it is a measure of our ability to get the same ranking of items (as opposed to people) given a different sample. It is less important than person reliability for most survey research applications. We can increase item reliability by increasing the variance in the difficulty of items. That is, have items that are relatively less difficult to endorse and other items that are more difficult to endorse. Increasing your sample size will also increase item reliability. Having more people gives you more information with which to score the items on their difficulty—number of people affects item reliability; not number of items.

Along with reliability, the person and item separation are also given, which are functions of the person and item reliability. This separation is different from the idea of separation of the estimates of person ability and item difficulty that is fundamental to Rasch modeling. Item (or person) separation is how well your measure separates individual items (or people) from each other. Or another way to think about separation is how many groups can the people (or items) be separated into easily. Separation is the square root of the reliability divided by one minus the reliability.

$$\text{Separation} = \text{square root}\left(\text{Reliability} / (1 - \text{Reliability})\right)$$

A person or item separation of over 2 is considered a good fit, giving us at least two distinct groups of people. These measure how well our people or items are being

distributed. A value that is a lot less than 2 for person reliability means that most of your people are lumped together, i.e., undifferentiated. This will mean you have less variance to use to predict an outcome variable or less variance to explain if your measure is of an outcome variable. The same reasoning applies to item separation. When there is little item separation, this means that most of the items are redundant, i.e., stacked on top of each other at some point along the scale. They are mostly too easy, or too hard, or clustered in the middle (see figure 3). Whichever of these is the case, the items will be less than ideal for differentiating people.

Both item and person reliability are based on only non-extreme items and participants. Non-extreme items are items with at least one participant answering different from all of the rest. For example, if you asked if participants liked their mother and everyone said yes, than that is an extreme item and it did nothing to differentiate people in your sample. In this case, we know that this is an "easy" item to endorse but we do not know how easy since no one answered no. We have a floor affect for this item. Extreme participants are similar; they endorse the lowest (or highest) response option for every item. Participants who endorse the most positive response for every question on how much they like their parents are indistinguishable from each other. We know they like their parents but we do not really know how much. We have a ceiling affect. Both the floor affect and ceiling affect for items and participants means we cannot use them in fitting the Rasch model. Programs may estimate the reliability and separation including extreme items and participants but these are often arbitrary because the values given for the item difficulty or participant ability are arbitrary.

The software package Winsteps (the program we use for this analysis), does another split when estimating reliability and separation. Winsteps presents a "real" reliability and a "model" reliability. The model reliability is based off of the best possible estimate of the standard error for the reliability, while the real reliability is based off an inflated standard error. The real reliability inflates the standard error according to the data misfit to the Rasch model (Linacre, 2008). The actual standard error is between these two estimates, so the real and model reliability act as upper and lower bounds of the reliability of the measure. Linacre (2009) (the developer of Winsteps) suggests using the real reliability as you develop your measure and the model reliability when you are satisfied with the your data's fit to the Rasch model.

Fit statistics.

The reliability and separation statistics are global fit statistics, but they do not tell you anything about the individual items or participants. The major model fit statistics provided to check if your data fits the Rasch model at the person and item level are the infit and outfit statistics. The infit and outfit indicate how accurately a particular item or participant fit the Rasch model. The infit statistic ("*in*lier-sensitive *fit*") is a weighted fit, and is more sensitive to unexpected patterns of responses by persons on items that are targeted to the person (that is, items that match person's ability), and vice-versa. If a person has an overall logit of 1.0, then the infit statistics are comparing his or her response to items at close to this degree of difficulty. The outfit statistic ("*out*lier-sensitive *fit*") is a fit statistic more sensitive to unexpected patterns of answers by persons on items that are relatively very easy or very hard for them. Both measures of fit are based on the chi-square statistic. The outfit and infit

statistics are usually reported as both mean-square fit statistics, and as standardized fit statistics (ZSTD). The ZSTD for the infit tests how well we can differentiate this person from those people near this person and the ZSTD for the outfit tells us the same for people who are farther away. The same is true of item infit and outfit. These are interpreted as *z*-tests and a value greater than 2.0 or less than -2.0 suggests a problem with the fit to the Rasch model.

There are several reasons why an item has an outfit ZSTD higher than 2.0. That item could be measuring a separate construct than the rest of the survey. Negatively worded items that need to be reversed coded sometimes do not fit because they represent a methods factor. The item might be badly worded and so people are coming up with multiple interpretations of the same items. It is possible that a few people answered that question in a highly unexpected way thus creating the misfit. Your sample may contain a subgroup that responds differently. “Is your spouse supportive?” may mean something very differently to incarcerated men than it means to professional women. The reason an item or person has an infit or outfit ZSTD lower than -2.0 is simply redundant information. The person (or item) is not adding anything that contributes to separating the items (or people). This is not a real problem for participants misfit, but it can be a problem for item misfit.

The Relationship Assessment Scale

In order to make the comparison between classical reliability and validity tests and the information gathered through Rasch modeling, I selected a frequently used ordinal scale survey on relationship quality that has been rigorously tested for both reliability and validity using classical test theory methods. The Relationship

Assessment Scale was developed by Susan Hendricks (1988) as a tool to measure relationship quality. The scale was developed to be short, while still being reliable and valid and has been shown to be highly correlated with longer measures of relationship quality. The short nature of the scale is important in clinical settings, or online surveys and was kept in mind when the scale was modified to fit the Rasch model in the second phase of the study.

Table 1. Means and Standard Deviations for Relationship Assessment Scale

	(N = 125)	
Item number & content	Mean	SD
1. How well does your partner meet your needs?	4.22	0.87
2. In general, how satisfied are you with your relationship?	4.26	0.92
3. How good is your relationship compared to most?	4.28	0.91
4. How often do you wish you hadn't gotten into this relationship?	4.13	0.97
5. To what extent has your relationship met your original expectations?	3.94	1.08
6. How much do you love your partner?	4.79	0.53
7. How many problems are there in your relationship?	3.51	1.13

Note: Scores could range from 1 (low satisfaction) to 5 (high satisfaction). Items 4 and 7 are reverse coded.

The Relationship Assessment Scale has found great use in the clinical setting (Vaughn & Matyastik Baier, 1999) and has been shown to be both a reliable and a valid measure of relationship satisfaction (Hendricks, 1988, Hendrick, Dicke, & Hendrick, 1998, Vaughn & Matyastik Baier, 1999). A search of Goggle scholar shows that at least 60 other articles or books have cited the Relationship Assessment Scale.

The Relationship Assessment Scale has shown itself to be a useful measure over the 20 years since it was first published.

The Relationship Assessment Scale is a seven-question Likert-scale measure that deals with a single factor or global construct. Each item had five possible responses, though each item had slightly different response options. The original sample that Hendricks used to test the measure was a group of undergraduate students in a dating relationship (N=125). Table 1 gives the means and standard deviations from the original publishing of the measure (Hendricks, 1988). We can see from the means in Table 1 that this measure has similar means and standard deviations for each item, a requirement of classical test theory. The second thing we notice about the means is that most of the items have a mean between four and five, suggesting a ceiling affect for the sample that this measure was tested on.

The items also showed good inter-item correlation, see Table 2, and the mean inter-item correlation for the original sample was .49. Hendrick's reported both a standardized $\alpha = .87$ and a unstandardized $\alpha = .86$, both are acceptable alpha's for classical test theory. Hendrick (1988) also reports the results of a second test of the measure and the Relationship Assessment Scale gave comparable results for the second test of the measure as well.

Table 2. Item Correlations for the Items of the Relationship Assessment Scale

Item	1	2	3	4	5	6	7
1	1.00	.67	.48	.34	.35	.44	.21
2		1.00	.51	.43	.35	.34	.26
3			1.00	.38	.27	.34	.35
4				1.00	.31	.39	.34
5					1.00	.39	.17
6						1.00	.21
7							1.00

The Relationship Assessment Scale was also tested for validity using several different tests. First a principal-components factor analysis was run on the data and a single factor was found to explain 46 percent of the variance of the data. The factor loadings for the seven questions are in Table 3. All of the loadings were in the moderate range from .49 to .79, suggesting the measure is unidimensional. The principle-components factor analysis deals with construct validity of the measure, but the measure was also tested for predictive validity and tested against other longer, know measures of relationship satisfaction.

Table 3. Principal-Components Factor Analysis for the Relationship Assessment Scale

Item and Content	(N = 125)
1. How well does your partner meet your needs?	.77
2. In general, how satisfied are you with your relationship?	.79
3. How good is your relationship compared to most?	.72
4. How often do you wish you hadn't gotten into this relationship?	.67
5. To what extent has your relationship met your original expectations?	.58
6. How much do you love your partner?	.66
7. How many problems are there in your relationship	.49

The Dyadic Adjustment Scale (Spanier, 1976) is a multidimensional measure that has become common as a global measure of martial satisfaction (Vaughn & Matyastik Baier, 1999). Hendrick (1988) slightly modified the Dyadic Adjustment Scale to make it more suitable for dating relationships and then tested a sample of college students (57 couples) to see if the Relationship Assessment Scale and the

Dyadic Adjustment Scale were correlated. The Relationship Assessment Scale and the score on the total score on the Dyadic Adjustment Scale had a correlation of .80 for the sample and .83 for the Relationship Assessment Scale and the satisfaction subscale in the Dyadic Adjustment scale. A small subsample of the couples were contacted later to see if they had stayed together or broken up. The Relationship Assessment Scale correctly predicted 91 percent of the couples that were together and 86 percent of the couples that were apart.

Ten years later Hendrick, Dicke, and Hendrick (1998) published a collection of research that had been done on validating the Relationship Assessment scale since the original publication. The Relationship Assessment Scale was again compared to the Dyadic Adjustment Scale with a larger more diverse sample and similar results were found. The Relationship Assessment Scale was also compared to the Kansas Marital Satisfaction Scale and a correlation of .74 for women and .64 for men was found for the two scales.

A test-retest reliability for the Relationship Assessment Scale was done during this time as well (Hendrick, Dicke, & Hendrick, 1998). The scale was given once and then re-administered six-seven weeks later. The test-retest reliability was found to be .85. Given the change that can happen in college age relationships, a .85 test-retest reliability is an impressive number. The work of Hendrick and others over the course of ten years solidifies the value, reliability, and validity of the Relationship Assessment Scale by the standards of classical test theory.

Purpose And Hypotheses

In comparing classical measurement theory and Rasch modeling I sought to answer two major questions:

1. What extra information does the Rasch model give over classical testing theory in testing the Relationship Assessment Scale for both reliability and validity?
2. Does modifying the Relationship Assessment Scale using Rasch modeling give us better reliability in classical test theory as well as better fit to the Rasch model?

The answer to the first question is simply an empirical one, based on actually checking a measure for reliability using both theories as guidance. Based on the differences between Rasch modeling and classical test theory I do not foresee an improvement in classical test theory reliability by fitting the Rasch model. This is based primarily on the fact that they are attempting to do different things: Rasch modeling seeks to create items with different means and classical test theory requires the items to the same means.

Method

I will illustrate the advantages Rasch modeling offers by evaluating an existing scale. Hendrick (1988) presented a scale to measure relationship satisfaction in a quick and reliable fashion. We selected this scale because it is published and has excellent strength based on classical measurement theory. Most alternative examples we considered were far weaker than Hendrick's Relationship Assessment Scale and therefore lacked the ability to show the added information provided by Rasch modeling. We wanted to see what Rasch modeling could offer to a scale that was already carefully developed.

First, the Relationship Assessment Scale was checked for reliability and validity using both classical test theory (to verify Hendricks' results) and then the measure was fit to the Rasch model. The Relationship Assessment Scale was then modified using Rasch modeling theory to make it fit the Rasch model better. The modified Relationship Assessment Scale was then again checked using a different sample for reliability and validity using both classical test theory and Rasch modeling. The modified Relationship Assessment Scale fit the Rasch model sufficiently that further modifications were not needed.

Participants

The two samples that were collected for this study were quite different. A trait of Rasch modeling that takes advantage of objective specificity. Allowing measure development to be separated from samples selected.

Sample 1

We administered her scale to 149 undergraduates taking a general requirement course in physics at a west coast university. Students participating in our survey included those who reported being in a romantic relationship, Hendrick used this screen in her study as well. The sample is also similar to the sample that Hendrick used in first testing the Relationship Assessment Scale's reliability and validity.

Sample 2

The second sample was 133 participants collected through advertisements on internet discussion boards. The participants, based on those that frequent the discussion boards, are young professions. They are mainly Romanian but speak

English. The filter of being in a romantic relationship was again used here. The second sample tended to be older, have more females, and was in relationships longer.

Data Analysis

The data obtained on the Relationship Assessment Scale was checked for reliability using Stata for classical test theory and Winsteps for fitting to the Rasch Model. The Modified Relationship Assessment Scale was also checked for reliability using Stata for classical test theory and Winsteps for fitting to the Rasch model. Winsteps was also used in both cases to check slight modifications to both the RAS and modified-RAS.

Results

Original RAS

Table 4. Means and Standard Deviations for Two Datasets

	1988 Dataset, Study 1 ($N = 125$)		2007 Dataset ($N = 149$)	
Item number & content	Mean	SD	Mean	SD
1. How well does your partner meet your needs?	4.22	0.87	4.03†	0.97
2. In general, how satisfied are you with your relationship?	4.26	0.92	4.04†	0.97
3. How good is your relationship compared to most?	4.28	0.91	4.15^{ns}	0.99
4. How often do you wish you hadn't gotten into this relationship?	4.13	0.97	3.88†	1.35
5. To what extent has your relationship met your original expectations?	3.94	1.08	3.94^{ns}	1.18
6. How much do you love your partner?	4.79	0.53	4.21***	1.26
7. How many problems are there in your relationship?	3.51	1.13	3.54^{ns}	1.29

Note: Scores could range from 1 (low satisfaction) to 5 (high satisfaction). Items 4 and 7 are reverse coded.
† $p < .10$; *$p < .05$; ** $p < .01$; ***$p < .001$

Our descriptive statistics matched the original descriptive statistics obtained by Hendricks (1989) quite closely. Table 1 reports the means and standard deviations for each item and they are reasonably similar for both groups. Hendrick's reported both a standardized $\alpha = .87$ (unstandardized $\alpha = .86$) and a mean inter-item correlation of .49 which is consistent with our standardized $\alpha = .89$ (unstandardized $\alpha = .88$) and mean inter-item correlation of .53.

We also had similar substantial inter-item correlations in both datasets. A principal component factor analysis resulted in a single dominant first principal component for both datasets. Hendrick found that a principal component could account for 46 percent of the variance, for our sample the principal component accounted for 67 percent of the variance. Hendrick did several demonstrations of validity for her scale that we will not repeat here.

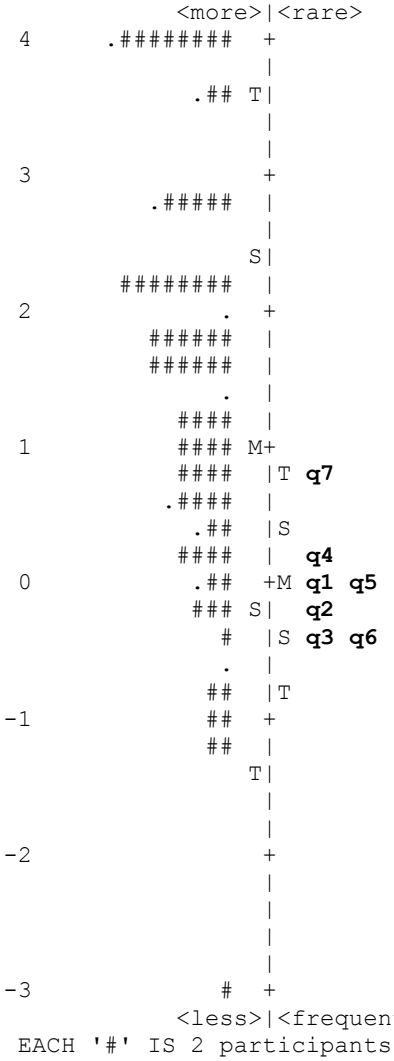
The variable map is our first look at how well the RAS fits the Rasch model. The variable map is the real version of Figure 3 and appears in Figure 6. At the extreme left of Figure 6 are the logit values on which the interval scale is based. Here they range from -3 to 4. The vertical axis separates the distribution of individuals from the distribution of items.

The note at the bottom of the table tells us that the #'s are two people who are at the same point of the scale. At the top of the distribution of people that there are 17 people who have a reported logit of 4 (greatest relationship satisfaction). There are two individuals with the lowest score of -3 logits. A careful examination of other output in Winsteps tells us that theses 17 people picked the most positive response options and the two people picked the most negative response to all 7 items. Because

they all did this, Winsteps has no way of differentiating how satisfied each of them are. In this case, the ceiling effect masks potentially important variation among 17 of our 149 participants and a floor effect masked differences between these two individuals.

On the left side of the vertical line there is an “M” which is the mean score for the people in our group. This is a logit of about 1.0. When the mean logit for the participants is greater than 0, this suggests that our scale is lacking in difficulty for reasons that will be discussed later. There is also an “S” and a “T” both above and below the “M”. These represent one (S) and two (T) standard deviations above/below the mean.

Figure 6. Variable Map for Relationship Satisfaction Scale



To the right of the vertical line are the same symbols (M, S, T), but this time they refer to logit values for the items. By default, Winsteps forces the mean logit for the items to be zero such that the item with exactly average difficulty has this neutral score. The “S” and “T” for the items are closely packed around the mean indicating that there is little variance in the item difficulties. All of the items are fairly similar in difficulty. The least difficult item has a logit a bit below zero and the most difficult has a logit a bit below one.

Figure 7. Summary Statistics for RAS as generated by Winsteps

```
SUMMARY OF 131 MEASURED (NON-EXTREME) PERSONS
+-------------------------------------------------------------------------------+
|           RAW                          MODEL          INFIT        OUTFIT     |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD  |
|-------------------------------------------------------------------------------|
| MEAN      26.9       7.0        1.10     .54      1.01    -.1    .99    -.1  |
| S.D.       5.6        .2        1.22     .16       .83    1.3    .81    1.3  |
| MAX.      34.0       7.0        3.54    1.05      4.78    3.6   4.60    3.6  |
| MIN.       8.0       6.0       -2.91     .36       .14   -2.8    .14   -2.8  |
|-------------------------------------------------------------------------------|
| REAL RMSE    .63  ADJ.SD    1.05  SEPARATION  1.68  PERSON RELIABILITY  .74  |
|MODEL RMSE    .56  ADJ.SD    1.08  SEPARATION  1.92  PERSON RELIABILITY  .79  |
| S.E. OF PERSON MEAN = .11                                                     |
+-------------------------------------------------------------------------------+
  MAXIMUM EXTREME SCORE:     17 PERSONS
  MINIMUM EXTREME SCORE:      1 PERSONS
        VALID RESPONSES:  99.6%

     SUMMARY OF 149 MEASURED (EXTREME AND NON-EXTREME) PERSONS
+-------------------------------------------------------------------------------+
|           RAW                          MODEL          INFIT        OUTFIT     |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD  |
|-------------------------------------------------------------------------------|
| MEAN      27.7       7.0        1.49     .70                                  |
| S.D.       6.1        .2        1.70     .45                                  |
| MAX.      35.0       7.0        4.81    1.85                                  |
| MIN.       7.0       6.0       -4.02     .36                                  |
|-------------------------------------------------------------------------------|
| REAL RMSE    .87  ADJ.SD    1.47  SEPARATION  1.69  PERSON RELIABILITY  .74  |
|MODEL RMSE    .83  ADJ.SD    1.49  SEPARATION  1.79  PERSON RELIABILITY  .76  |
| S.E. OF PERSON MEAN = .14                                                     |
+-------------------------------------------------------------------------------+
PERSON RAW SCORE-TO-MEASURE CORRELATION = .92 (approximate due to missing data)
CRONBACH ALPHA (KR-20) PERSON RAW SCORE RELIABILITY = .88 (approximate due to
missing data)

     SUMMARY OF 7 MEASURED (NON-EXTREME) ITEMS
+-------------------------------------------------------------------------------+
|           RAW                          MODEL          INFIT        OUTFIT     |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD  |
|-------------------------------------------------------------------------------|
| MEAN     503.0     130.4         .00     .11      1.02    -.2    .99    -.4  |
| S.D.      28.3        .9         .36     .01       .45    3.1    .38    2.6  |
| MAX.     533.0     131.0         .74     .12      1.84    4.8   1.52    3.4  |
| MIN.     441.0     129.0        -.46     .10       .59   -3.5    .59   -3.2  |
|-------------------------------------------------------------------------------|
| REAL RMSE    .12  ADJ.SD     .34  SEPARATION  2.76  ITEM   RELIABILITY  .88  |
|MODEL RMSE    .11  ADJ.SD     .35  SEPARATION  3.08  ITEM   RELIABILITY  .90  |
| S.E. OF ITEM MEAN = .15                                                       |
+-------------------------------------------------------------------------------+
UMEAN=.000 USCALE=1.000

ITEM RAW SCORE-TO-MEASURE CORRELATION = -.99 (approximate due to missing data)
913 DATA POINTS. APPROXIMATE LOG-LIKELIHOOD CHI-SQUARE: 1911.19
```

The variable map gives us a good first look at how our data fit the Rasch model and how reliable our measure is, but it does not give us actual estimates on how reliable our measure is. The summary statistics give us a quick overview of how our measure is fitting the Rasch model (see figure 7). As mentioned above, Rasch modeling cannot model people with perfect scores, either all right or all wrong (in our case all 5's or all 1's.), so the model is estimated using only the non-extreme scores

and an approximation is made to include extreme scores. Although we did not have any items where only the extreme values were selected, it is possible to have extreme items and a similar process would take place for those.

For our analysis of the Relationship Assessment Scale, we see in figure 7 that the measure has an estimated person reliability of .74 for the 131 non-extreme people who took our measure. Although .74 person reliability is not a horrible reliability, we see that the person separation is 1.68, so improvement is definitely possible. At the bottom of figure 7 we find the fit statistics for the items, a item reliability of .88 and a item separation of 2.76. The summary statistics in Winsteps also gives the Cronbach's alpha ($\alpha = .88$) and the raw score to measure (Rasch logit) correlation ($r = .92$). This last statistic is simply how correlated the interval level scale is to the original ordinal level scale.

The min and max of the ZSTD for our dataset for the people outfit is -2.8 to 3.6 and the min and max of the ZSTD for the item misfit is -3.2 to 3.4. This suggests that we may have people and items that are not fitting the Rasch model.

Winsteps provides a table summarizing information for each person in the dataset (see figure 8). This figure is taken directly from Winsteps output and so it needs some explanation on the content found inside of it.

Figure 8. Person Fit Statistics For RAS as Generated by Winsteps

ENTRY NUMBER	RAW SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEA CORR.	EXACT MATCH OBS%	EXACT MATCH EXP%	PERSON
77	31	7	1.85	.59	4.78	3.6	4.60	3.6	A .09	28.6	48.8	078314
5	30	7	1.53	.54	4.01	3.2	3.86	3.1	B .05	14.3	45.0	005507
12	30	7	1.53	.54	3.60	2.9	3.43	2.8	C .42	14.3	45.0	012405
36	32	7	2.24	.66	3.36	2.6	3.22	2.5	D .09	71.4	62.7	036404
6	25	7	.43	.41	2.81	2.4	3.20	2.7	E-.04	28.6	44.0	006204
67	25	7	.43	.41	2.96	2.6	2.82	2.4	F .61	.0	44.0	068504
116	23	7	.11	.39	2.84	2.6	2.90	2.6	G-.64	14.3	36.9	117404
19	15	7	-.98	.38	2.89	3.1	2.65	2.7	H .39	.0	28.3	019204
13	26	7	.61	.43	2.26	1.9	2.69	2.2	I-.49	57.1	47.0	013204
140	21	7	-.18	.37	2.68	2.7	2.64	2.6	J .54	14.3	34.2	141506
66	18	7	-.58	.36	2.29	2.4	2.65	2.8	K-.66	.0	31.9	067205
25	27	6	2.09	.67	2.64	2.0	2.03	1.5	L .84	66.7	56.3	025404
68	32	7	2.24	.66	2.01	1.5	2.46	1.9	M-.68	57.1	62.7	069204
20	16	7	-.84	.37	2.35	2.5	2.27	2.3	N .18	14.3	29.6	020105
96	22	7	-.04	.38	2.19	2.0	2.12	1.9	O-.50	28.6	35.4	097304
129	15	7	-.98	.38	2.12	2.1	2.08	2.0	P .16	.0	28.3	130507
149	27	7	.80	.45	1.93	1.5	2.01	1.6	Q-.55	42.9	47.7	150105
50	13	7	-1.29	.41	1.76	1.4	1.54	1.1	R .42	14.3	34.4	050505
7	17	7	-.71	.36	1.65	1.4	1.61	1.3	S .32	.0	31.5	007104
BETTER FITTING OMITTED												
16	23	7	.11	.39	.40	-1.4	.41	-1.3	t .42	28.6	36.9	016306
145	22	7	-.04	.38	.40	-1.5	.39	-1.4	s .69	57.1	35.4	146404
90	27	7	.80	.45	.31	-1.5	.35	-1.3	r .78	71.4	47.7	091505
93	17	7	-.71	.36	.34	-2.0	.33	-2.0	q .54	57.1	31.5	094414
42	30	7	1.53	.54	.27	-1.6	.31	-1.5	p .69	71.4	45.0	042405
92	27	7	.80	.45	.28	-1.6	.31	-1.5	o .85	71.4	47.7	093506
132	27	7	.80	.45	.28	-1.6	.31	-1.5	n .85	71.4	47.7	133404
43	25	7	.43	.41	.26	-1.8	.30	-1.6	m .36	85.7	44.0	043107
148	25	7	.43	.41	.26	-1.8	.30	-1.6	l .36	85.7	44.0	149504
135	28	7	1.02	.48	.28	-1.6	.28	-1.6	k .60	85.7	47.1	136204
105	22	7	-.04	.38	.23	-2.2	.24	-2.1	j .12	57.1	35.4	106294
126	25	6	1.37	.55	.19	-1.8	.22	-1.7	i .55	83.3	47.6	127504
4	23	7	.11	.39	.21	-2.2	.21	-2.2	h .82	85.7	36.9	004209
82	28	7	1.02	.48	.19	-2.0	.21	-1.9	g .88	85.7	47.1	083509
88	28	7	1.02	.48	.19	-2.0	.21	-1.9	f .88	85.7	47.1	089304
136	26	7	.61	.43	.18	-2.1	.18	-2.1	e .57	85.7	47.0	137404
47	21	7	-.18	.37	.15	-2.8	.15	-2.8	d .00	85.7	34.2	047404
60	21	7	-.18	.37	.15	-2.8	.15	-2.8	c .00	85.7	34.2	061104
91	21	7	-.18	.37	.15	-2.8	.15	-2.8	b .00	85.7	34.2	092504
100	22	7	-.04	.38	.14	-2.8	.14	-2.7	a .52	85.7	35.4	101505
MEAN	27.7	7.0	1.49	.70	1.01	-.1	.99	-.1		54.1	49.7	
S.D.	6.1	.2	1.70	.45	.83	1.3	.81	1.3		24.6	14.1	

The relationship assessment scale had 5 possible responses for each item so with 7 items the possible range for the total raw score is 7 to 35. Winsteps reports on the people with the biggest positive ZSTD scores and the people with the biggest negative ZSTD scores. The overall mean raw score total was 27.7 and most of our problematic people have scores that are either 30 or above, or 25 or below. The most miss-fitting people are found at the top of the list, with participants that are fitting too

well (adding redundant information in separating the items) listed at the bottom. We are not concerned with the people adding redundant information, but we are concerned with those that are most fitting the Rasch Model.

Although this figure gives us an idea if our misfit to the Rasch model is because of individual participants, it does not give us enough information to make decisions about removing participants. To give us a better idea of how our misfitting people are responding to our items we can look at what Winsteps calls a keyform. The keyform provides each individual person's measure, their MNSQ infit and outfit, as well as their responses to the seven items. The keyform also gives us information about which responses where expected by the model and which were not. The information we obtain through the keyform is best explained through example.

Figure 9. Keyform of Person 77

```
KEY: .1.=OBSERVED, 1=EXPECTED, (1)=OBSERVED, BUT VERY UNEXPECTED.

NUMBER - NAME ------------------ MEASURE - INFIT (MNSQ) OUTFIT - S.E.
     77  078314                     1.85     4.8   A    4.6      .59

-3     -2      -1       0       1       2       3       4
|-------+-------+-------+-------+-------+-------+-------|  NUM   ITEM
                                        4             .5.      7  q7
                                        4       .5.            4  q4
     (1)                                4                      5  q5
                                        4     .5.              1  q1
                                        4     .5.              2  q2
|-------+-------+-------+-------+-------+-------+-------|  NUM   ITEM
-3     -2      -1       0       1       2       3       4
```

Person 77 (who had a ZSTD score larger than 2, see figure 9) had only one very unexpected response on item 5 about their expectations being met in the relationship. Person 77 gave an answer of 1 to this item (extremely dissatisfied) but based on the other responses given we expected 77 to answer with a 4 (agree). There are some other items (q7, q4, q1, & q2) where 77 answered with a 5 (strongly agree),

but this is close to what we expected based on the overall score for this person. We would want to ask 77 why the answer to q5 was extremely dissatisfied. Not all of the questions asked appear in the keyform; any response that matched the models expectations are omitted for space as a keyform is generated for every single participant that took your measure. Using the person misfit table and the keyforms a picture starts to develop around where the measure fits the Rasch model and where it does not.

Just like for the people in the study, we can look at each individual item and get a better understanding of which items are fitting the model and which items are not. In figure 10 see that item 7, "how many problems are there in your relationship" was the most difficult item with a logit of .74. and item 6, "how much do you love your partner" was the easiest item with a logit of -.46.

Figure 10. Item Fits Statistics for RAS as Generated by Winsteps

ENTRY NUMBER	RAW SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEA CORR.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
6	533	129	-.46	.12	1.84	4.8	1.50	3.0	A .61	47.3	54.5	q6
5	501	131	.08	.11	1.46	3.1	1.52	3.4	B .62	48.9	49.0	q5
7	441	131	.74	.10	1.10	.8	1.15	1.1	C .73	33.6	43.4	q7
4	492	131	.18	.11	.92	-.6	.87	-1.0	D .76	56.5	48.0	q4
1	514	131	-.09	.11	.59	-3.5	.67	-2.6	c .74	61.1	50.0	q1
2	516	131	-.11	.11	.65	-3.0	.62	-3.1	b .76	65.6	50.1	q2
3	524	129	-.34	.12	.60	-3.3	.59	-3.2	a .77	65.1	53.1	q3
MEAN	503.0	130.4	.00	.11	1.02	-.2	.99	-.4		54.0	49.7	
S.D.	28.3	.9	.36	.01	.45	3.1	.38	2.6		10.7	3.3	

We found two items with an outfit ZSTD of greater than 2 and three items with an outfit ZSTD of less than -2. The positive *z*-scores are interpreted quite differently than the negative z-scores. An item with an outfit ZSTD of less than -2 means that item is redundant, i.e., it is not differentiating between people that another item is not already differentiating between. These redundant items can artificially improve model

fit and can unnecessarily lengthen our survey. These items might be candidates for dropping if we want to shorten our questionnaire. Still, they are not hurting our overall reliability and some of them may be helpful. Also, we would want to drop them one at a time because they may be redundant with each other and if we dropped all of them at once we would lose whatever the set of items differentiated. Those items with an outfit ZSTD above two mean that for some reason they are not fitting the Rasch model and are hurting the overall fit to the Rasch model.

Likert type items are often treated as interval level measures even though they were designed to be ordinal level. Rasch modelers contend that the logit score produced is a true interval level measure. This is achieved by identifying the item difficulty and by allowing different response options to vary. Winsteps provides output on each item and all possible responses to that item. From my analysis of the relationship assessment scale I found a problem with question 5. This can be seen in figure 11.

There are 11 people who selected the first response option for item 5 and their average logit for the entire scale was a -.78 placing them low on relationship satisfaction. We would like our response options to cover a substantial range and a rule of thumb is that there should be about a 1 logit shift in the average score as you move up one response category. (The shift could be less with more response options and greater with fewer response options.) Consider the 7 people who selected the response option 2, they had an average logit on the combined scale of .27.

Figure 11. Average Measure per Response Option for RAS

ENTRY NUMBER	DATA CODE	SCORE VALUE	DATA COUNT	DATA %	AVERAGE MEASURE	S.E. MEAN	OUTF MNSQ	PTMEA CORR.	ITEM	
6	A	1	11	7	-1.07	.41	1.8	-.42	q6_love	1
		2	7	5	-.11	.30	1.7	-.21		2
		3	17	12	.40	.16	1.1	-.23		3
		4	17	12	.89	.25	1.0	-.13		4
		5	95	65	2.21	.16	1.3	.57		5
		MISSING ***	2	1*	.99	.38		-.03		
5	B	1	11	7	-.78	.51	2.8	-.38	q5_expect	1
		2	7	5	.27	.45	2.7	-.16		2
		3	20	13	.26*	.19	.8	-.28		3
		4	53	36	1.28	.12	1.0	-.09		4
		5	58	39	2.68	.22	1.5	.56		5
7	C	1	13	9	-1.04	.34	.6	-.46	q7_problem	1
		2	19	13	.23	.21	1.2	-.28		2
		3	38	26	.90	.14	1.0	-.20		3
		4	33	22	1.79	.16	1.0	.09		4
		5	46	31	3.00	.23	1.5	.59		5
4	D	1	16	11	-.96	.31	1.3	-.50	q4_wish	1
		2	9	6	-.24	.19	.6	-.26		2
		3	21	14	.41	.11	.4	-.26		3
		4	34	23	1.21	.10	.4	-.09		4
		5	69	46	2.75	.17	1.0	.69		5
1	c	1	6	4	-1.84	.54	.4	-.40	q1_needs	1
		2	5	3	-.97	.15	.1	-.27		2
		3	17	11	-.13	.12	.2	-.34		3
		4	72	48	1.35	.10	1.0	-.08		4
		5	49	33	2.91	.23	1.1	.59		5
2	b	1	5	3	-1.94	.68	.6	-.37	q2_satisfied	1
		2	6	4	-.53	.18	.4	-.24		2
		3	19	13	-.10	.17	.5	-.36		3
		4	67	45	1.17	.09	.7	-.17		4
		5	52	35	3.04	.20	.9	.67		5
3	a	1	2	1	-2.50	1.52	.7	-.27	q3_compare	1
		2	10	7	-1.01	.25	.4	-.39		2
		3	20	14	-.02	.12	.4	-.35		3
		4	47	32	.99	.10	.6	-.20		4
		5	68	46	2.75	.17	.8	.69		5
		MISSING ***	2	1*	1.89	.20		.03		

This is about half a logit higher. It is okay that it is higher because they should be more satisfied with their relationship than those people who picked response option 1. Ideally, if these categories are really different, the difference would be closer to a full logit. What happens with those who picked the medium response option, 3? Their combined score should be higher than those who picked response option 2, but it is actually slightly lower, their combined score was .26. This suggests that people with

higher relationship satisfaction would choose option 2 before option 3.This is a serious problem with this item.

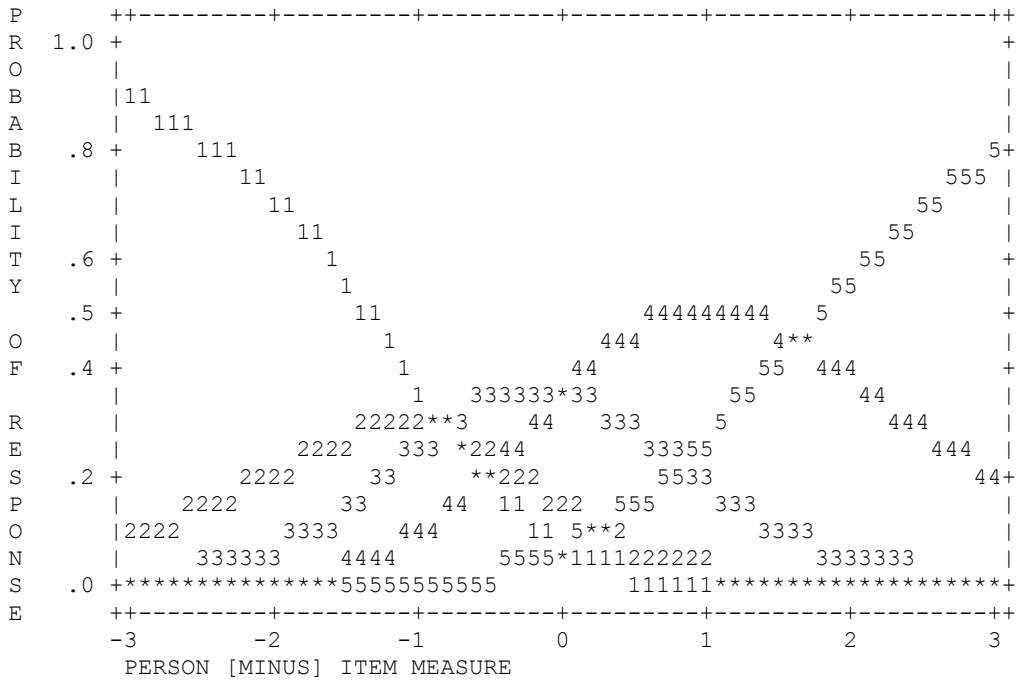
There was very little missing data in this survey. Where there is missing data, Winsteps gives you the overall logit for people who skipped a particular item. With this survey the one person who skipped item 6 had and overall logit of .99 compared to the two people who skipped item 3 who had an overall logit of 1.89. This information can be very useful in evaluating how problematic missing values are. If it turns out that those skipping items are systematically very high or very low on their overall logit, then assumptions about data missing at random are questionable and assumptions about data missing completely at random are untenable (Acock, 2005).

Examining item 1, partner meets your needs, we see a very good differentiation. Those selecting the first response option had an overall logit of -1.84, those selecting the second response option had an overall logit of -.97, those picking the third response option had -.13, those picking the fourth had 1.35, and those picking the fifth response option had an overall logit of 2.91. Even though the item mapping showed this item in the middle on item difficulty, it is still valuable for differentiating individuals. This illustrates how the wealth of information Winsteps provides can lead to what seem like contradictory recommendations. We would want to keep this item because of the range of difficulties the response options give us.

This scale uses five response options. Winsteps provides information that lets us decide if this is the right number of options and if the labels we picked are working as desired. While Winsteps provides information about each item's response options, Winsteps also provides comparable information for the response options overall.

Figure 12 is the graph of response probabilities and is another piece of information Winsteps provides about response options. The curve represented by "11" that starts high on the vertical axis (between 0.9 and 1.0 on the left) and swoops down is the distribution of the likelihood of a person selecting the first option (extremely dissatisfied) based on his or her overall score (horizontal axis) The higher your overall score, the less likely you are to pick response option 1.

Figure 12. Category Probabilities Plot for 5 Response Options of RAS



Thus, a person who has an overall logit of -4 (extremely dissatisfied with his/her relationship) has a 95% chance of endorsing this response option. By contrast, a person who has an overall logit of +4 has about a 90% chance of endorsing the fifth response option (extremely satisfied) and no chance of selecting the first response option. These two response options are valuable because they clearly differentiate between those who are dissatisfied and those who are satisfied with their relationship.

The third response option is in the middle and the fourth is just to its right. The most problematic response option is the second. This option "never sees the light." By this we mean that there is no overall score where this is the most likely response. When there is a response option that is never the most likely choice, regardless of a person overall score, then it is possible the option might be dropped. This would simplify the scale and probably would not hurt the overall reliability.

Figure 13. Thresholds for the Rating Scale for the RAS

CATEGORY LABEL	SCORE	OBSERVED COUNT	%	OBSVD AVRGE	SAMPLE EXPECT	INFIT MNSQ	OUTFIT MNSQ	STRUCTURE CALIBRATN	CATEGORY MEASURE	
1	1	57	6	-.93	-1.03	1.18	1.49	NONE	(-2.34)	1
2	2	63	7	-.37	-.32	.96	1.01	-.76	-1.09	2
3	3	152	17	.23	.37	.74	.64	-.87	-.19	3
4	4	323	35	1.26	1.18	.70	.80	.01	.98	4
5	5	318	35	2.00	2.03	1.35	1.16	1.62	(2.85)	5
MISSING		4	0	1.84						

Along with this figure, Winsteps provides a table of statistics that gives a better idea about how our rating scale was used, figure 13. In particular we are interested in the column labeled Structure Calibration. This tells us at which logit value do we move from more likely to answer 1 to more likely to answer 2 and so forth. These thresholds give us an idea if our set of response options is being fully utilized. Figure 12 is simply a graphical representation of that same idea. In our analysis of the relationship assessment scale we saw right away we have a problem with point 2 of our five-point rating scale.

The threshold for point 3 (-.87) actually comes before point 2 (-.76). Never at any time are people most likely to pick response 2. From minus infinity to a logit of -.86 they are mostly likely to pick the first response option (extremely dissatisfied).

From -.87 to a logit of .00 they are most likely to pick option three. From .01 to 1.61 they are mostly likely to pick option four. From 1.62 to plus infinity they are most likely to pick option 5. This gives strong evidence that a five response options might not be the best choice for this survey or that the five response options were not labeled clearly enough.

Modification of the Survey

To facilitate movement from the results of the Relationship Assessment Scale to the modified versions, here is a brief overview of the modifications we made to Hendrick's survey (original and modified survey included in appendix). The first step was to remove any items that might be hurting the fit to a Rasch model. The first item that we removed dealt with how a person's relationship compared to others. This item had an outfit ZSTD of -3.2 and in reviewing the actual paper surveys, we found that several people were confused about who "others" were (their personal other relationships, or the relationships of other people). The fit statistic showed it to be redundant and responses written on the surveys themselves showed it to be slightly confusing. The second item and final item we removed completely was "How well does your partner meet your needs". This had similar difficulty as another item ("To what extent has your relationship met your original expectations?") and an outfit ZSTD that showed it was also redundant.

The second step was to deal with the negatively worded items. Negatively worded items are commonly used to catch people who are not being honest, or just to force people to think more deeply about the questions. However, in the Rasch modeling literature negatively worded items are often seen as problematic because

they might not be measuring the same construct as positively worded items. The solution of reverse coding the items is common, but the question remains how often is the reverse coded item really the same as item that is positively worded. Like many rules in Rasch modeling, this is not written in stone and there are times when you may not be able to get away from negatively worded items. In our case we were able to reword the two negatively worded items, but later as we attempted to add items we were stuck with an item we thought only worked negatively worded, though we later removed the item as it improved fit to the Rasch Model (see appendix C for more).

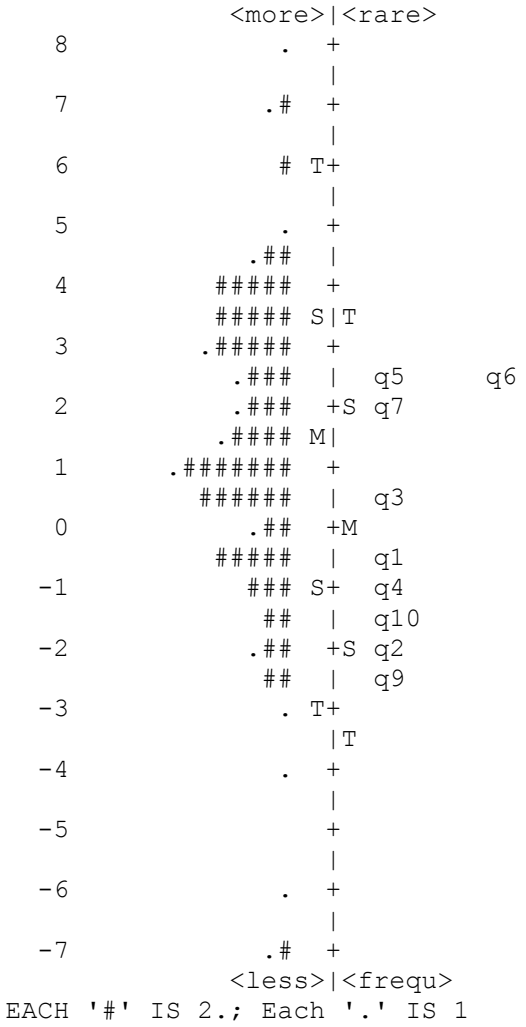
In total we took three items and copied them exactly as they were in Hendrick's survey, we reworded two items that were negativity worded and we removed two items completely. Two of the items that we copied verbatim had outfit ZSTD above 2, but we kept them because we thought the problem was in the response options and not the question, I will discuss the response options later. We were then left with a survey of five items, all positively worded. The next step was to add more items without making the survey too long, a concern expressed in the original survey development. The added items would increase person reliability in two ways: the simple fact that more items means more reliability and new items of greater or lesser difficulty than our original items would help to better differentiate people. The new items would also have an affect on item reliability: more items decreases our ability to replicate item order, but new items of greater or lesser difficulty would increase our ability to replicate item order. The new items are a win-win situation for the person reliability, but if our items do not cover a range of difficulties it will hurt reliability.

Finally we decided we needed to change the response options for all nine items on our modified survey. We saw in Figure 6 that the second response category was not being utilized. We ran two different models, a 3-point scale and a 4-point scale. The 4-point scale gave us better variability than did the 3-point scale. Moreover, a 4-points scale forces the participants not to select a neutral level. A 4-point response scale also improves reliability by increasing variability. At the same time we made sure that we labeled all of the response categories for each of the items. This is important because it helps to minimize the problem found with item 5 of the original survey where we have two ordinal items flipped in the interval scale.

Modified RAS

The modified Relationship Assessment Scale (modified-RAS) was checked for reliability using both alpha and fitting the Rasch model. The modified-RAS had a unstandardized alpha of .92 with a mean inter-item correlation of .34. Figure 15 gives us the variable map for our modified Relationship Assessment Scale. We see that we our people range from around -6 to around 8 logits and our items range from around -3 around 2 logits. This gives around a 5 logit spread for our items, as opposed to around a 2 logit spread for the original items. We also see that we only have 4 possible extreme participants (those that either answered all 4's or 1's). This suggests that we addressed the ceiling affect that we had in the RAS

Figure 14. Variable Map for Modified-RAS



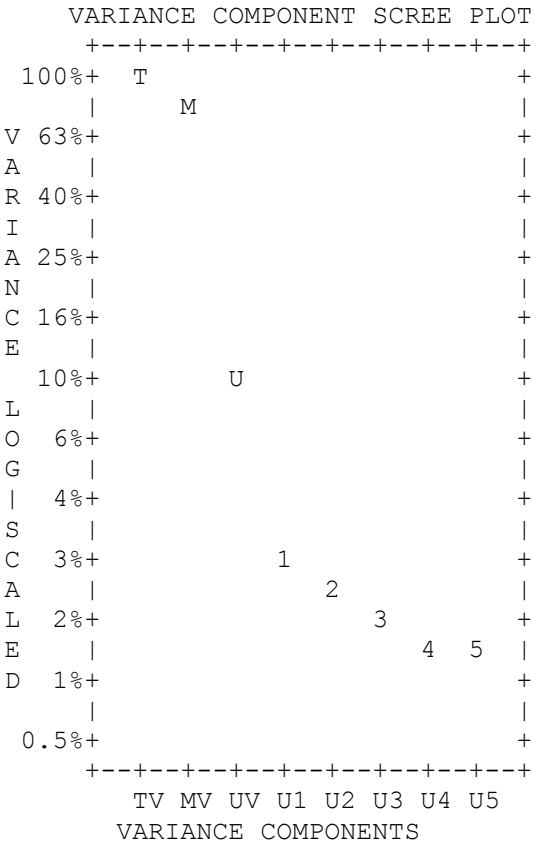
Looking at table 5 we find out reliability estimates. We now have a person reliability of .88 and person separation of 2.71, both better than before. We have also increased our Cronbach's alpha from .88 to .92, so we have a better fitting model by classical test theory standards as well. We can also see from table 5 that from our outfit ZSTD we have items with fit statistics from -2.9 to 2.5. We have at least one item that could possibly be considered redundant (ZSTD = -2.9) and at least one item that may not be fitting the Rasch model (ZSTD = 2.5).

Table 5. Summary Statistics For Modified-RAS.

	Alpha	Reliability	Separation	Infit ZSTD		Outfit ZSTD	
				Min	Max	Min	Max
Person	.92	.88	2.71	-2.4	4.5	-2.4	4.3
Item		.99	9.63	-3.3	3.1	-2.9	2.5

We can also use Winsteps to check the factor structure of our scale. The principal component explains 87.4 percent of the variance in the observations. We can see from the scree plot (figure 14.) that we have a classic example of a principle component, the large drop of from the variance explained by the scale (M) and the 1st contrast, followed in quick succession by the remaining contrasts.

Figure 15. Scree Plot for Principe Component Analysis



Looking at the item fit statistics in table 6, there are two items that might cause a problem. Item 5 has a outfit ZSTD of 2.5 and item 6 has a outfit ZSTD of -2.90. Item

5 and 6 also have very similar item difficulties, but by looking at table 7 we see that their response options cover different ranges. The rest of the items all fall within acceptable ranges for fit to the Rasch model. We see from table 6 that the items in the modified RAS have a range of -2.57 logits to 2.50 logits.

Table 6. Item Fit Statistics for Modified-RAS.

	RAW		MODEL	INFIT		OUTFIT	
ITEM	SCORE	MEASURE	S.E.	MNSQ	ZSTD	MNSQ	ZSTD
q1	380	-0.36	0.18	1.00	0.10	0.92	-0.50
q2	424	-1.84	0.19	1.01	0.20	0.86	-0.60
q3	357	0.37	0.18	0.96	-0.30	1.05	0.40
q4	403	-1.12	0.18	1.19	1.50	1.13	0.80
q5	286	2.50	0.18	1.42	3.10	1.41	2.50
q6	290	2.38	0.18	0.64	-3.30	0.62	-2.90
q7	294	2.21	0.18	0.75	-2.10	0.74	-1.90
q9	440	-2.57	0.20	1.15	1.20	0.97	0.00
q10	413	-1.57	0.19	0.91	-0.70	0.85	-0.70
MEAN	365.2	0.00	0.18	1.01	0.00	0.95	-0.30
S.D.	57.8	1.85	0.01	0.22	1.80	0.22	1.50

Table 7. Average Measure for Response Options for Each Item in the Modified-RAS

	DATA	DATA		AVERAGE		DATA	DATA		AVERAGE
ITEM	CODE	COUNT	%	MEASURE	ITEM	CODE	COUNT	%	MEASURE
q1	1	10	8	-3.30	q6	1	16	12	-3.21
	2	22	17	-1.58		2	72	55	0.73
	3	71	53	1.56		3	39	30	3.35
	4	30	23	4.18		4	5	4	6.88
q2	1	8	6	-5.27	q7	1	18	14	-2.98
	2	12	9	-1.45		2	64	49	0.75
	3	53	40	0.64		3	41	31	3.03
	4	60	45	3.23		4	8	6	6.19
q3	1	9	7	-4.21	q8	Deleted			
	2	35	26	-0.63					
	3	71	53	2.12					
	4	18	14	4.29					
q4	1	3	2	-8.36	q9	1	5	4	-5.98
	2	20	15	-1.43		2	8	6	-2.87
	3	73	55	1.37		3	50	38	0.51
	4	37	28	3.29		4	69	52	2.81
q5	1	21	16	-1.79	q10	1	5	4	-6.69
	2	66	50	0.86		2	15	11	-1.33
	3	40	30	2.94		3	63	48	0.78
	4	5	4	5.68		4	49	37	3.48

Looking at Table 7 again, none of the items have response options that are out of order and all items cover a good range with their different response options. We have a range of difficulty from -8.36 to endorse option 1 on question 4 to 6.88 to endorse option 4 on question 6.

Although table 7 gives us an idea how are response options are working for each item; figure 16 shows us graphically how are response options are doing as a single rating scale across the survey. A good range exists for each response option to be the most likely to be endorsed as a person increases in their satisfaction with their relationship.

Figure 16. Category Probabilities Plot for 4 Response Options of Modified-RAS

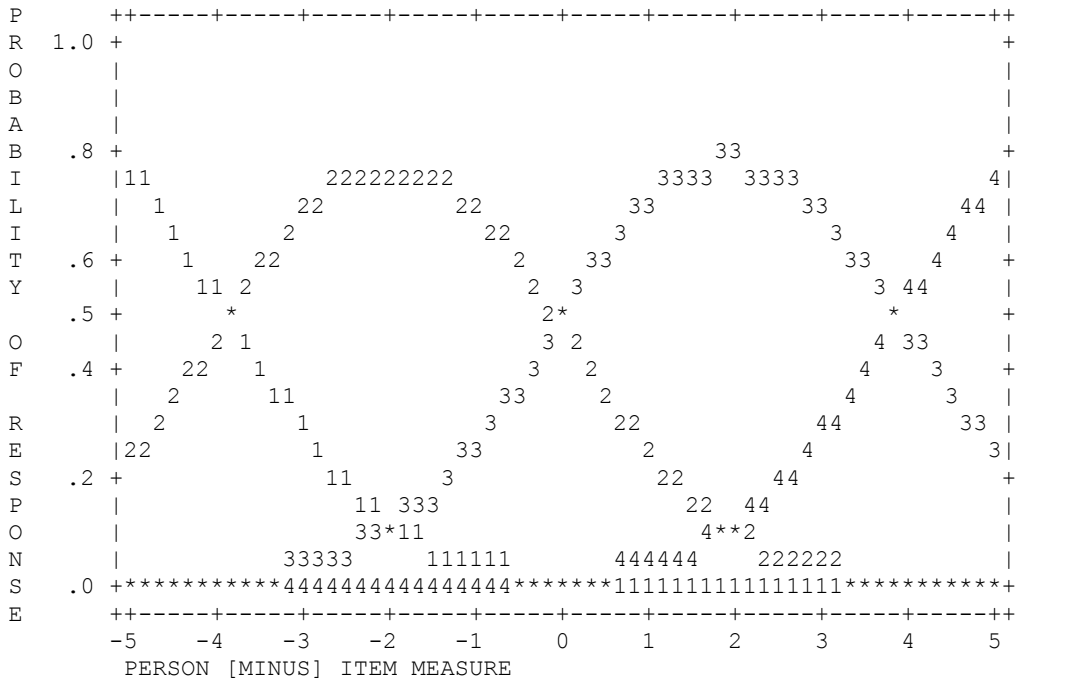


Table 8 gives us the thresholds (called structure calibration) for the response options on our survey. Response option 2 takes over as the most likely to be selected at -3.85 logits on the participants scale of relationship satisfaction. Option 3 takes over

at -.03 logits and finally option 4 takes over at 3.88 logits. We are getting clear differentiation between the thresholds of our response options. This will help us to be confident that our scale is truly interval. With good reliability and separation and not items that are not fitting the model we can ignore the person fit statistics.

Table 8. Thresholds for the Rating Scale for the Mod-RAS

CATEGORY LABEL	OBSERVED COUNT	%	OBSVD AVRGE	STRUCTURE CALIBRATN
1	68	6	-3.08	NONE
2	314	27	-1.22	-3.85
3	501	43	1.93	-0.03
4	272	23	4.73	3.88

OBSERVED AVERAGE is mean of measures in category. It is not a parameter estimate.

Discussion

Good measurement is vital to good research and Rasch modeling gives new tools and insight into measurement that classical measurement theory lacks. I have presented a published measure of relationship satisfaction and through fitting the measure to the Rasch model shown how the measure could be improved. Based on classical measurement theory the RAS is an excellent scale. It is highly reliable, the mean correlation of items is substantial, and there is a dominant first principal component supporting its unidimensionality. However, a major problem is apparent. The RAS seems to have a ceiling affect. On average the participants scored a 3.95 out of 5. Seventy-five percent of participants had an average score of 3.5 or higher. Clearly most of the participants were very satisfied with their relationships, but we are unable to distinguish clearly between individuals in this group of people with high satisfaction. This is common when items are all of similar difficulty and are easy to

endorse. A floor affect would have the opposite affect of creating an inability to distinguish participants at lower levels of ability or in our case satisfaction.

The relationship satisfaction scale does not ask items that cover the full range of difficulty. Each item has a very similar mean and standard deviation on a 1-5 point scale suggesting that the distribution of answers to each item is similar. Some of the items also appear to be very similar or even redundant. Adding redundant items will simplify the factor structure and necessarily increase alpha, increasing validity and reliability by classical test theory standards. Redundant items will not however differentiate survey participants except for those in the narrow range of the redundant items. Someone using Classical test theory would end here, having no tools to determine the difficulty of their questions beyond simple means and standard deviations. However, having determined alpha and a single factor construct, Rasch modeling provides tools to look at each item in our measure and each person that took our measure.

The variable map (figure 6 for RAS and figure 15 for Modified RAS) is a quick glance on the distribution of our items and participants. Looking at the variable map for the RAS we noticed that both item 1 and item 5 as well as item 3 and item 6 have the same difficulty. Item 4 and item 2 also have very similar difficulties. Our observation about the means of our items is made explicit in the variable map and we see how similar the items really are. We can ask any of these items but once asked the other items do not have much independent information we can use to differentiate relationship satisfaction for this sample. The implications of not being able to differentiate people with extremely high or low scores on a variable can be dramatic in

a science that emphasizes prediction. Because of the ceiling affect, more difficult items to endorse need to be added to distinguish people at the uppers regions of relationship satisfaction. Something seen with our descriptive statistics but is once again made explicit through fitting the Rasch Model. The scale is still useful in its present format, but could be strengthened with a few additional items that are harder to endorse. As a minimum, we would like items that cover at least two logits of difficulty.

Looking at figures 10 there are 3 items (item 1, 2, and 3) with a outfit ZSTD under -2 indicating that about half of the items in the measure are redundant. Two items (item 5 and 6) have a outfit ZSTD above 2 suggesting some miss-fit to the Rasch model. Only item 4 and 7 seem to be giving independent information that is differentiating people according to the Rasch model. This information gives us an idea of why they measure is not fitting the Rasch model, but in determining what to do about the items we need more information about the measure.

The redundant items (1, 2, and 3) are simply that and rewording, elimination, or even doing nothing are possible things to do with these items. Different wording might make the items easier or more difficult, thereby differentiating them from the other items. If length of the measure was not a barrier, elimination would shorten the measure and not hurt our fit to the Rasch model. Again if length of the measure was not a barrier, the items could simply be left and new items of lesser and greater difficulty could be added. A combination of these actions was taken in modifying the RAS.

Items five and six had an outfit ZSTD greater than 2, indicating that they are not fitting the Rasch model. This could be for several reasons: the items discriminates

participants at different levels of ability, an individual or group of participants are answering the item in a different way, or your response options are causing the trouble. The first reason, different discrimination, goes to the very heart of Rasch modeling. Item difficulty is estimated when a measure is fitted to the Rasch model and this estimates are supposed to hold for all levels of ability. If an item is more difficult than other at low levels of difficulty and less difficult than another items at higher levels of difficulty then our item is not fitting the Rasch model. Any item that exhibits this problem needs to be addressed.

Sometimes an item is not fitting the Rasch model because of the responses of a single individual or group of participants. It could be that these participants did not understand the measure (foreign language, illiteracy, etc.), they lied on some of the questions, or are not in your target population (children being asked questions about adult relationship satisfaction). In this case, the best method is simply to remove the participants from your dataset and rerun your analysis. With proper demographics, those participants that might not understand or not even be in your target population can easily be found and removed. Those participants that are lying may be harder to determine and I will discuss them later.

The final problem deals with the response options. Winsteps has the capability of combining points in your response options, giving you various combinations to find response options that work. At the same time, Rasch modeling does not require every question in the survey to have the number of response options. This is called a partial credit model in Rasch modeling. For example, you might wish to ask a series of yes and no questions followed by a series of Likert scale questions, as long as all of the

questions are dealing with the same construct, the partial credit model handles it just fine. (An example of combining points of a scale and partial credit model in Winsteps is provided in the appendix.)

Examination of figure 12, and the related statistical information suggest that the second response option could be eliminated. We might even consider reducing our five response options to just three, namely, 1, 3, 5. We could do this and re-run Winsteps to see if this is justified. When you have more response options than your target population can differentiate, Winsteps can make this quite clear. This is especially important when administering items to a population that has limited language skills. Another reason for the problem with our rating scale could be that the second and fourth response option was not labeled by Hendrick, leaving participants to decide what they were. This could have result in confusion about the second response category. Since a scale has to be strictly ordinal to fit the Rasch model it is always important to label all possible response categories.

To decide if an item should be removed or should remain in the survey you need to asses the value of that question, rethink the wording and or do some qualitative work on your survey. The higher than |2| for an outfit ZSTD for an item misfit rule is used as a convenient cutoff, but it is not a hard fast rule; theory and good follow-up work should decide if an item is kept or rejected. The RAS had two items that had outfit ZSTD higher than 2, both items were retailed with some rewording and changes to the response options.

Although our items could use some work, it is important to remember that participants can also hurt the measures fit to the Rasch model. Fitting the Rasch model

provided several ways to examine the participants fit to the Rasch model. Winsteps provided a table of miss-fitting persons and keyforms that allow us to check for strange patterns of answers from our participants. Checking this out for all of the people who have poor fits, we could develop a series of questions to ask and we might find this helpful in revising our scale. This is tool to determine where the misfit really is, in your items or in your people.

In the previous example person 77 seemed to be miss-fitting the Rasch model if we combine this with other graphs and tables from Winsteps we may decide to remove this person from our data set. We also might decide to remove or modify the particular item that person 77 gave unexpected answer to instead of removing person 77. It all depends on if many participants are giving unexpected answers to that item or is person 77 unique in this way. Recall that the purpose of Rasch modeling is to fit the data to the model, not fit the model to the data. In attempting to develop a survey it may be necessary to remove people or items to get a good fit to the Rasch model. Now that being said, it is dangerous to remove a person or item without some theory, qualitative work, or strong evidence from a Rasch program like Winsteps backing the removal up. Just because person 77 has one very unexpected response is not grounds for removal of either the person or the item. However, if you do make the decision to remove a person or item from your dataset and rerun your model, Winsteps is able to easily accommodate that and help you evaluate hour decision (see appendix B).

Table 9 gives a quick overview of the changes made to the RAS. The modified-RAS is still short and to the point, something the original author sought and now has a greater range of difficulty in items. The modified-RAS is not a perfect

measure, but by fitting the Rasch model, the scale is better at differentiating between people across the spectrum of relationship satisfaction and increased reliability as well.

Table 9. Question Content Differences between RAS and Mod-RAS

RAS Question Content	Modified-RAS Question Content
How well does your partner meet your needs?	--
In general, how satisfied are you with your relationship?	In general, how satisfied are you with your relationship?
How good is your relationship compared to most?	--
How often do you wish you hadn't gotten into this relationship?	Are you glad you got into this relationship?†
To what extent has your relationship met your original expectations?	To what extent has your relationship met your original expectations?
How much do you love your partner?	How much do you love your partner?
How many problems are there in your relationship?	My relationship is problem free?†
--	Do you think of your relationship as perfect?
--	Do you think of yourself as the happiest couple in world?
--	Do you like spending time with you partner?
--	Do you think your relationship has changed your life for the better?

Note: The Mod-RAS had four possible response options; the RAS had five possible response options.

† These two items are re-wordings of the original questions in order to make the items positively worded

Conclusion

Measurement is vital to any statistical analysis in the field of human development and family science. Without proper measurement there is no way we can trust any statistical analysis we do. The use of classical test theory allows researchers to get a good idea of the reliability of their measure, but it also causes unwanted consequences. Cronbach's alpha is a great tool, but it has its limitations, limitations researchers often forget or ignore. Researchers who do not know the explicit limitations or assumptions of alpha are liable to consider their measure reliable based solely on alpha, but in reality, alpha may simply mask real problems with a measure.

Rasch modeling expands upon the amount of information we obtain from classical test theory. Rasch modeling allows researchers to have a comprehensive look at their measure. It avoids troublesome assumptions of classical test theory and focuses on both differentiation and replicability. Rasch modeling also seeks to obtain fundamental traits of measurement (Karabatsos, 2004). In particular, the ability to produce interval level scales from ordinal level data is important in a word of statistical modeling that assumes interval level data. This is not to say that Rasch modeling is the perfect solution to all measurement problems. People have even argued against the claim that Rasch produces interval level data (Barrett, 1999; Karabatsos, 2004). With that said, it is important to address the limitations of Rasch modeling and this comparison to classical test theory.

The beginning of this paper addressed the weaknesses associated with classical test theory and its application of alpha; it is fitting to end with the limitations of Rasch modeling and this study in particular. Because of the focus on reliability in this paper,

the importance of validity in measurement was ignored to a certain degree. Rasch modeling does a better good of creating a measure that differentiates people than does alpha, but the issue of predicative validity was ignored. Hendricks (1989) showed clearly that the RAS is able to predict with high probability those couples that would stay together or break up. The modified-RAS was not checked for predictive ability as this is an area that Rasch modeling cannot help simplify measurement development. It could be assumed that a measure that differentiates better would have better predictive ability, but this was not tested for in this study and so must be left to others to discern.

A second issue with Rasch modeling, although not really a weakness, is that is really requires measurement development to be upfront and explicit. Although Rasch modeling can be used in secondary data analysis, ultimately is best used to develop measures before actual data collection takes place. Data collected with measures development using classical test theory is unlikely to have a great fit to the Rasch model. Much of the information gathered by fitting the Rasch model is unusable if researchers are simply using Rasch modeling post hoc.

It is important to note that this study only covered the basic applications of Rasch modeling to measurement development. Several very important aspects of Rasch modeling were not covered by this comparison because of its focus on reliability. Rasch modeling, as presented here, focused on item and person differentiation, but another aspect of fitting the Rasch model is the ability to check for discrimination of items by demographic variables. In our case, Winsteps can determine if some questions have different difficulties for men or women. If a question does have significantly different difficulties by gender then we are not

measuring men and women on the same scale and so results of men and women would not be comparable. Fitting the Rasch model allows us to determine if our measure is discriminating against a group of people based on demographic variables; a very important ability given the current trend to include more underrepresented populations.

Another aspect of Rasch modeling that was not covered in this study is the ability to equate test scores even when different versions are administered. Often when a measure is administered a second time the participant's score may reflect more of learning the measure then actual change in the participant. Yet if a different version of the test is administered how do now that the scores are comparable? Rasch modeling allows for tests to be equated and therefore the scores for each test can be scaled so that they end up having the same interval level scale. Taking two incomparable measures and making them comparable, all the while, avoiding test-retest error. This idea is often taken a step farther by creating Computer Adaptive Testing (CAT). The current incarnation of the GRE is a perfect example. People taking the GRE take very different tests and yet the results are all on the same scale. This allows people to take the GRE multiple times without introducing error, it shortens the length of the test by making explicit the difficulty of each question, and discourages cheating. In the field of family science, the ability to give different versions of the same measure could help when measuring spouses (who might talk about a single measure or copy responses), interventions where the same constructs need to be measured more than once or even in classrooms to limit cheating by students. The ability to give shorter measures is also important, because it allows researchers to measure more in a shorter time.

Rasch modeling opens a window into measurement development that classical test theory did not. Rasch modeling allows researchers interested in measurement development to really dig deep into measures and work on covering the broadest range of variability possible in a population. Rasch modeling is not the answer to all measurement problems, but it is a step in the right direction and helps to make better more explicit decisions about measurement.

References

Acock, A. C. (2005). Working with missing values. *Journal of Marriage and the Family, 67,* 1012-1028.

Andersen, E. B., & Olsen, L. W. (2001). Essays on Item Response Theory. A. Boomsma, M. A. J. van Duijn, T. A. B. Snijders (Ed.). *The Life of Georg Rasch as a Mathematician and as a Statistician* (pp 3-24). New York: Springer-Verlag.

Andrich (1978). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement, 2*, 581-594.

Barrett, P. (1999). Rasch scaling: Yet another incomprehensible test theory or something far more dangerous? In BPS Millennium Conference: Beyond Psychometrics. Retrieved January 10, 2009, from http://www.pbarrett.net/presentations/BPS-rasch_98.pdf.

Embretson, S. E. (2006). The new rules of measurement: What every psychologist and educator should know. S. E. Embretson & S. L. Hershberger (Ed.) *Issues in the assessment of cognitive abilities* (pp 1-16). Mahwah, N.J: Lawrence Erlbaum Associates

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences*. A Series of books in psychology. San Francisco: W.H. Freeman.

Hendrick, S. S. (1988). A generic measure of relationship satisfaction. *Journal of Marriage and the Family, 50*, 93-98.

Hendrick, S. S., Dicke, A., & Hendrick, C. (1998). The relationship assessment scale. *Journal of Social and Personal Relationships, 15*, 137-142.

Karabatsos, G. (2004). Introduction to Rasch Measurement. E. V. Smith Jr & R. M. Smith (Ed.). *The rasch model, additive conjoint measurement, and new models of probabilistic measurement theory* (pp 630-664). Maple Grove, Minnesota: JAM Press.

Linacre. M. (n.d.). Reliability and separation of measures: Winsteps help. In Reliability and separation of measures. Retrieved January 10, 2008, from http://www.winsteps.com/winman/index.htm?reliability.htm.

Pallant, J. & Tennant, A. (2007). An introduction to the rasch measurement model: An example using the hospital anxiety and depression scale (HADS). *British Journal of Clinical Psychology, 48*, 1-18.

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*, 879-903.

Rost, J. (2001). Essays on Item Response Theory. A. Boomsma, M. A. J. van Duijn, T. A. B. Snijders (Ed.). *The growing family of rasch models* (pp 25-42). New York: Springer-Verlag.

Safrit, M. Cohen, A., Costa, M. (1989). Item response theory and the measurement of motor behavior. *Research Quarterly Exercise and Sport, 60*, 325-35.

Spanier, G. B. (1976).Measuring dyadic adjustment: New scales for assessing the quality of marriage and similar dyads. *Journal of Marriage and the Family, 38*, 15-25.

Tang, W., Wong, E., Chiu, H., Ungvari, G. (2007). Rasch analysis of the scoring scheme of the HADS depression subscale in chinese stroke patients. *Psychiatry Research, 160*, 97-103.

Vaughn, M. J. & Matyastik Baier, M. E. (1999). Reliability and validity of the relationship assessment scale. *The American Journal of Family Therapy, 27*, 137-174.

Washburn, I. J., Dogaru, C., & Acock, A. C. (2007, November). *Can rasch modeling help our measurement*. Paper presented at the annual meeting of the National Council on Family Relations, Pittsburgh, PA.

Wright, B. D. (2006). The new rules of measurement: What every psychologist and educator should know. S. E. Embretson & S. L. Hershberger (Ed.) *Fundamental measurement for psychology* (pp 65-104). Mahwah, N.J: Lawrence Erlbaum Associates

APPENDICES

Appendix A

RELATIONSHIP ASSESSMENT SCALE - Original

Please mark on the answer sheet the letter for each item which best answers that item for you.

How well does your partner meet your needs?

A	B	C	D	E
Poorly		Average		Extremely well

In general, how satisfied are you with your relationship?

A	B	C	D	E
Unsatisfied		Average		Extremely satisfied

How good is your relationship compared to most?

A	B	C	D	E
Poor		Average		Excellent

How often do you wish you hadn't gotten in this relationship?

A	B	C	D	E
Never		Average		Very often

To what extent has your relationship met your original expectations:

A	B	C	D	E
Hardly at all		Average		Completely

How much do you love your partner?

A	B	C	D	E
Not much		Average		Very much

How many problems are there in your relationship?

A	B	C	D	E
Very few		Average		Very many

NOTE: Items 4 and 7 are reverse scored. A=1, B=2, C=3, D=4, E=5. You add up the items and divide by 7 to get a mean score.

MODIFIED RELATIONSHIP ASSESSMENT SCALE

Please mark on the answer sheet the number for each item which best answers that item for you.

In general, how satisfied are you with your relationship?

1	2	3	4
Extremely Unsatisfied	Unsatisfied	Satisfied	Extremely satisfied

Are you glad you got into this relationship?

1	2	3	4
Strongly disagree	Disagree	Agree	Strongly Agree

To what extent has your relationship met your original expectations:

1	2	3	4
A lot worse than expected	Worse than expected	Better than expected	A lot better than expected

How much do you love your partner?

1	2	3	4
Not at all	Not much	A lot	Completely

My relationship is problem free?

1	2	3	4
Strongly disagree	Disagree	Agree	Strongly Agree

Do you think of your relationship as perfect?

1	2	3	4
Strongly disagree	Disagree	Agree	Strongly Agree

Do you think of yourself as the happiest couple in world?

1	2	3	4
Strongly disagree	Disagree	Agree	Strongly Agree

Do you like spending time with you partner?

1	2	3	4
Never	Infrequently	Often	Always

Do you think your relationship has changed your life for the better?

1	2	3	4
Strongly disagree	Disagree	Agree	Strongly Agree

Removed Item:

Do you ever think of other people as possible romantic interests?

1	2	3	4
Never	Infrequently	Often	Always

Appendix B

Several programs perform Rasch modeling. I used the Winsteps program because it has extensive capabilities to provide rich diagnostic information, it is inexpensive, it has regularly updates, and it is easy to use, at least for people who use Stata. For Stata users there is a single user written command. Here is an example:

```
raschcvt momt701a-momt701h mcaseid, outfile(a:\problems) ///
  id(mcaseid) min(1) max(4) xwide(1)
```

After you reduce your dataset to the variables you will include in your scale, momt701a to momt701h along with an id variable, you run this command. The command must specify the name of the id variable, id(mcaseid), the minimum score on the scale, min(1), the maximum score on the scale, max(4). We have also had it use a format where each score has a single column, xwide(1). This command creates two files. In this example, problems.dat is a data file in the format Winsteps understands and problems.con is a program file that you run in Winsteps (after editing out the top two lines). This example produces a Winsteps program that looks extremely complex to somebody who is new to Winsteps. However, the program works without any modification.

For those who use SAS or SPSS, the Winsteps program can read your SAS or SPSS dataset. This is quite handy, but you then have to write the Winsteps program. Winsteps has a menu system that many feel is not yet very easy to use. We have included the control file we used (created by the above Stata command) with some explanation for SAS and SPSS users. The author of Winsteps, Linacre, just sent the three of us a beta version of a new Winsteps that supports Stata the same way it

supports SAS and SPSS. The Stata command is still useful because it sets up the Winsteps data and program file with one line of code.

The Control File in Winsteps for our Rasch Model. The control file created by Stats starts with a list of the variables and their length; this is not necessary and so was not included in the example control file. Any thing that starts with a ";" is a comment and will not be run by Winsteps, in our example all comments are in italics as well. Some of the comments given here were taken from www.winsteps.com in the help section.

```
TITLE=Rasch analysis of original RAS. ; this is the title of the model and will appear on any output tables.
DATA=\\onid-fs\washburi\MastersThesis\ras_data_orginal.dta.dat ;this tells Winsteps where your data is.
ITEM1=1    ; item responses start in first field
 NI=7      ; there are 7 responses, i.e., 7 response fields
 NAME1=8   ; the person name is in the 8th field
 DELIMITER = SPACE ; the field delimiters are spaces
 XWIDE=1   ; values are right-aligned, 1 characters wide.
 CODES=12345 ; the valid codes.
MAXPAG=60 ; set 60 lines per page
PRCOMP=S ; Principle component analyzes the standardized residuals.
MUCON=0 ; maximum number of JMLE (UCON) iterations
LCONV= .001 ; logit change at convergence
RCONV= .1 ; score residual at convergence
GROUPS= ;determines rating scale or partial credit model.
HLINES=Y ; heading lines in output files.
;This remaining commands deal with output tables and exporting to other software
        which can be done directly in the Winsteps menus.
;PSELECT= ?????1*
```

```
;TABLES=111101100111111000001111
;ISFILE=\\onid-fs\washburi\MastersThesis\ras_data_orginal.dta.isf
;IFILE=\\onid-fs\washburi\MastersThesis\ras_data_orginal.dta.IFL
;PFILE=\\onid-fs\washburi\MastersThesis\ras_data_orginal.dta.PFL
;XFILE=\\onid-fs\washburi\MastersThesis\ras_data_orginal.dta.XFL
&END
```

;These are the labels that will appear in the Winsteps output tables and exported files

```
q1 how well does your partner meet your needs?
q2 in general, how satisfied are you with your relationship?
q3 how good is your relationship compared to most?
q4 how often do you wish you hadn't gotten in this relationship?
q5 to what extent has your relationship met your original
expectations:
q6 how much do you love your partner?
q7 how many problems are there in your relationship?
END LABELS
```

;the variable names, this are mainly used in the exported files.

```
q1_needs
q2_satisfied
q3_compare
q4_wish
q5_expect
q6_love
q7_problem
END NAMES
```

Partial Credit Model. In the control file for Winsteps is the command groups=, by default this has nothing after it and Winsteps runs a Rasch model with a single rating scale. By changing the command to `GROUPS=0`, you get a partial credit model allowing every item in your survey to have a different rating scale. You still go to "output tables / rating (partial credit) scale" to see how the rating scale for each item fit.

Combining Points of Scale. In the control file for Winsteps after the command `CODES=` (which tells Winsteps what the rating scale will be) add `NEWSCORE=` with the

new values that you want them to take. For example, in our survey we had `CODES=12345` and to combine the second and third category we add `NEWSCORE=12234`. This will now give us a four-point scale for all of our items.

Removing a person from the dataset. In the control file for Winsteps simply add the following to your command file.

```
PDFILE=*
(the entry number of the person to delete)
*
```

For example: in our survey if I wanted to delete person 078, I see they have an entry number of 77 and so I add to my control file:

```
PDFILE=*
77
*
```

The asterisk means a list will follow and you must have a hard return after each entry you wish to delete.

Removing an item from the dataset. This is the same as removing a person, but the command is now

```
IDFILE=*
(the entry number of the item to delete)
*
```

Exporting Person/Item files to another program:

The procedure for exporting person/item measures is just a matter of selecting the options you want.

- Click on the "`Output Files`" pull down menu in the main screen of Winsteps

- Click on Person `File PFILE=` (or Item File `IFILE=` for item measures)
- Choose how to display the file, what file format you want, whether to include column headings, and whether to make it permanent.
- Click OK and you have the exported file containing either the person or item measure plus infit and outfit statistics for each person or item.

  Winsteps can export directly to SPSS, excel or a simple text file, but we know of no way to export directly to SAS or Stata.

Guide to retrieve tables presented through Winsteps pull down menus:

Table 5 = `Output Tables / Variable Map`

Table 6,7,8 = `Output Tables / Summary Statistics`

Table 9 = `Diagnosis / Dimensionality Map`

Table 10 = `Output Tables / Person (column): Fit Order`

Table 11, 13 = `Output Tables / Item (column): Fit Order`

Figure 4 = `Output Tables / Persons Keyforms: misfit order`

Figure 5 = `Output Tables / Persons Responses`

Figure 6, table 12 = `Diagnosis / Category Function`

Figure 8-9 = `Diagnosis / Construct KeyMap`

Appendix C

Figure 1.

SUMMARY OF 130 MEASURED (NON-EXTREME) PERSONS

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	27.3	9.9	.73	.57	1.03	-.2	1.08	-.2
S.D.	4.9	.5	1.50	.05	.82	1.5	1.11	1.6
MAX.	37.0	10.0	4.25	.77	5.80	4.9	9.90	7.6
MIN.	13.0	5.0	-4.22	.54	.13	-3.2	.13	-3.3

REAL RMSE	.66	ADJ.SD	1.35	SEPARATION	2.05	PERSON RELIABILITY .81
MODEL RMSE	.57	ADJ.SD	1.39	SEPARATION	2.43	PERSON RELIABILITY .86

S.E. OF PERSON MEAN = .13

MAXIMUM EXTREME SCORE: 1 PERSONS
MINIMUM EXTREME SCORE: 2 PERSONS
VALID RESPONSES: 99.5%

SUMMARY OF 133 MEASURED (EXTREME AND NON-EXTREME) PERSONS

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	27.1	9.9	.66	.60				
S.D.	5.4	.4	1.85	.20				
MAX.	40.0	10.0	7.21	1.89				
MIN.	10.0	5.0	-7.10	.54				

REAL RMSE	.71	ADJ.SD	1.71	SEPARATION	2.42	PERSON RELIABILITY .85
MODEL RMSE	.63	ADJ.SD	1.74	SEPARATION	2.76	PERSON RELIABILITY .88

S.E. OF PERSON MEAN = .16

PERSON RAW SCORE-TO-MEASURE CORRELATION = .96 (approximate due to missing data)
CRONBACH ALPHA (KR-20) PERSON RAW SCORE RELIABILITY = .88 (approximate due to missing data)

SUMMARY OF 10 MEASURED (NON-EXTREME) ITEMS

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	354.8	129.3	.00	.16	.99	-.9	1.34	-1.0
S.D.	64.6	.6	1.51	.01	.65	3.9	1.76	3.9
MAX.	441.0	130.0	2.38	.17	2.89	9.9	6.61	9.9
MIN.	252.0	128.0	-2.16	.15	.47	-5.5	.47	-5.2

REAL RMSE	.17	ADJ.SD	1.50	SEPARATION	8.76	ITEM RELIABILITY .99
MODEL RMSE	.16	ADJ.SD	1.50	SEPARATION	9.55	ITEM RELIABILITY .99

S.E. OF ITEM MEAN = .50

UMEAN=.000 USCALE=1.000
ITEM RAW SCORE-TO-MEASURE CORRELATION = -1.00 (approximate due to missing data)
1293 DATA POINTS. APPROXIMATE LOG-LIKELIHOOD CHI-SQUARE: 2124.73

From the first table (labeled summary of 130 (non-extreme) persons) we see that we have a person reliability of .81 and a person separability of 2.05. This shows that we have a better fit to the Rasch model for people than did the original, that only had a person reliability of .74 and a person reliability of 1.68. We also improved the fit

for the items (last table labeled summary of 10 (non-extreme) items), the modified version has an item reliability of .99 and item separation of 8.76 as compared to the original measure that had an item reliability of .88 and a item separation of 2.76. We can see that our modification of the Relationship Assessment Scale has increased the fit to the Rasch model. If we look again at the last table in the figure, at the minimum and maximum for the outfit ZSTD, we see that at least one item has an outfit ZSTD statistic of 9.9. This suggests that more work needs to be done, since at least one item is not fitting the Rasch model and eliminating this item may increase reliability.

Table 1 is the fits statistics for the items. Table 1 gives us the information on how well each item in our model is fitting the Rasch model. From the table we can clearly see that item 8 is not fitting in with the other items.

Figure 2.

```
TABLE 10.1 \\onid-fs\washburi\MastersThesis\ras_d ZOU769WS.TXT Aug 26 12:44 2008
INPUT: 133 PERSONS  10 ITEMS  MEASURED: 133 PERSONS  10 ITEMS  40 CATS      3.62.1
-------------------------------------------------------------------------------
PERSON: REAL SEP.: 2.08  REL.: .81 ... ITEM: REAL SEP.: 7.67  REL.: .98

         ITEM STATISTICS:  MISFIT ORDER

+-------------------------------------------------------------------------------------+
|ENTRY   RAW                  MODEL|   INFIT  |  OUTFIT  |PTMEA|EXACT MATCH|       |
|NUMBER SCORE COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.| OBS%  EXP%| ITEM  G |
|----------------------------------+----------+----------+-----+-----------+-------|
|     8  252    129    2.21     .16|2.96   9.9|4.64   9.9|A-.05| 38.8  61.7| q8    0 |
|     5  287    129    1.42     .16|1.06    .5|1.05    .4|B .64| 65.1  62.4| q5    0 |
|     4  404    130   -1.50     .17| .99   -.1| .95   -.4|C .68| 66.2  64.7| q4    0 |
|     9  441    129   -1.66     .16| .81  -1.4| .71  -1.8|D .73| 72.1  67.0| q9    0 |
|     3  358    130    -.06     .15| .78  -1.9| .78  -1.8|E .75| 73.1  63.5| q3    0 |
|    10  414    129   -1.28     .16| .73  -2.3| .70  -2.4|e .77| 71.3  63.6| q10   0 |
|     1  381    130    -.39     .15| .70  -2.6| .73  -2.3|d .78| 76.2  61.2| q1    0 |
|     2  425    130   -1.11     .15| .64  -3.0| .57  -3.2|c .79| 72.3  62.2| q2    0 |
|     6  291    129    1.29     .16| .59  -3.8| .58  -3.9|b .81| 78.3  65.0| q6    0 |
|     7  295    128    1.09     .15| .58  -4.0| .58  -4.0|a .81| 74.2  61.6| q7    0 |
|----------------------------------+----------+----------+-----+-----------+-------|
| MEAN   354.8  129.3   .00     .16| .98   -.9|1.13   -.9|     | 68.7  63.3|       |
| S.D.    64.6     .6  1.33     .01| .67   3.8|1.18   3.9|     | 10.7   1.7|       |
+-------------------------------------------------------------------------------------+
```

The next table gives us a good idea why item 8 is misfitting. This table gives us the average score of people who chose that response. Item 8 was the only reverse coded item that we left in the survey, the response options were flipped before running

Winsteps so the order of difficulty for the response options should be from 1 being easiest to 4 being hardest. In table 2 we see that for item 8 this is not the case. Response option 4 is the most difficult to endorse, but option 1 is actually the third hardest to endorse. This is a serious misfit to the model and based off of this information alone warrants serious thought about removing this item. I decided that given the warnings about reverse coded items (Fox), I would remove this item and rerun the analysis.

Figure 3.

TABLE 10.3 \\onid-fs\washburi\MastersThesis\ras_d ZOU769WS.TXT Aug 26 12:44 2008
INPUT: 133 PERSONS 10 ITEMS MEASURED: 133 PERSONS 10 ITEMS 40 CATS 3.62.1

ITEM CATEGORY/OPTION/DISTRACTOR FREQUENCIES: MISFIT ORDER

ENTRY NUMBER	DATA CODE	SCORE VALUE	DATA COUNT	%	AVERAGE MEASURE	S.E. MEAN	OUTF MNSQ	PTMEA CORR.	ITEM	
8	A	1	38	29	.64	.35	3.2	.05	q8	1
		2	65	49	.52*	.18	1.9	.01		2
		3	26	20	.18*	.31	4.2	-.09		3
		4	3	2	1.17	3.18	10.0	.06		4
		MISSING ***	1	1*	.87			.02		