Cogent Education

ISSN: (Print) 2331-186X (Online) Journal homepage: https://www.tandfonline.com/loi/oaed20

# Rasch analysis: A primer for school psychology researchers and practitioners

William J. Boone & Amity Noltemeyer |

cogent education

EDUCATIONAL PSYCHOLOGY & COUNSELLING | REVIEW ARTICLE

# Rasch analysis: A primer for school psychology researchers and practitioners

William J. Boone[1] and Amity Noltemeyer[1]*

Received: 14 September 2017
Accepted: 10 December 2017
First Published: 18 December 2017

*Corresponding author: Amity Noltemeyer, Department of Educational Psychology, Miami University, 201 McGuffey Hall, Oxford, OH 45056, USA
E-mail: anoltemeyer@miamioh.edu

Reviewing editor:
Gregory Yates, University of South Australia, Australia

Additional information is available at the end of the article

**Abstract:** In order to progress as a field, school psychology research must be informed by effective measurement techniques. One approach to address the need for careful measurement is Rasch analysis. This technique can (a) facilitate the development of instruments that provide useful data, (b) provide data that can be used confidently for both descriptive and parametric statistics, and (c) provide outcome measures that offer clinically meaningful guidance to school psychology researchers and practitioners. In this paper, we first introduce the basic principles of Rasch measurement that undergird the use of Rasch analysis in school psychology. Next, we describe several techniques that can be used to conduct a basic Rasch analysis. In doing so, we use Winsteps software to illustrate the application of these techniques with a single exemplar school psychology rating scale data set. Finally, we provide conclusions and resources to consult for further learning.

**Subjects: Educational Research; School Psychology; Educational Psychology**

**Keywords: Rasch; Winsteps; instrument development; psychometrics; school psychology**

## 1. Introduction

In order to progress as a field, school psychology research must be informed by effective measurement techniques. Suboptimal measurement techniques (e.g. surveys that include redundant items, treating rating scale data as if the scales were linear, very few data quality controls) in school psychology research may contribute to muddled results. This problem may stem from a

## ABOUT THE AUTHORS

William J. Boone is a distinguished research scholar at Miami University (Oxford, Ohio, USA). He earned his PhD from the University of Chicago's Program in Measurement, Evaluation and Statistical Analysis under the direction of Benjamin Wright. Boone is the lead author of the Rasch text "Rasch Analysis in the Human Sciences." He has conducted Rasch analyses for over 25 years. A professional goal of Boone is to expand the use of Rasch methods in social science and medical research.

Amity Noltemeyer is a professor in the School Psychology program at Miami University (Oxford, Ohio, USA). She earned her PhD from Kent State University's School Psychology program. Her research interests include multi-tiered systems of support in schools, resilience, school discipline, and school climate. Noltemeyer has authored many journal articles and book chapters, and co-manages several externally funded grants.

## PUBLIC INTEREST STATEMENT

Research in school psychology should be informed by effective measurement techniques. Rasch analysis (Rasch, 1960) is one way to approach this need for careful measurement. This technique can inform the development of instruments (surveys, tests, etc.) that provide useful data. Also, Rasch analysis can provide outcome measures that offer guidance to school psychology researchers and practitioners. In this paper, we first introduce the basic principles of Rasch measurement that undergird the use of Rasch analysis in school psychology. Next, we describe several techniques that can be used to conduct a basic Rasch analysis. In doing so, we use Winsteps (Linacre, 2017) software to illustrate the application of these techniques with a single school psychology rating scale data set. Finally, we provide conclusions and resources to consult for further learning.

misunderstanding of points made by Stevens (1951) about measurement (see Michell, 1997, 2002 for additional consideration of this issue). One approach to address the need for careful measurement is Rasch analysis (Rasch, 1960). This technique can (a) facilitate the development of instruments that provide useful data, (b) provide data that can be used confidently for both descriptive and parametric statistics, and (c) provide outcome measures that offer clinically meaningful guidance to school psychology researchers and practitioners. In this paper, we first introduce the basic principles of Rasch measurement that undergird the use of Rasch analysis in school psychology. Next, we describe several techniques that can be used to conduct a basic Rasch analysis. In doing so, we use Winsteps (Linacre, 2017) software to illustrate the application of these techniques with a single exemplar school psychology rating scale data set. Finally, we provide conclusions and resources to consult for further learning.

## 2. Rasch measurement

### *2.1. Theory undergirding the Rasch model*

The Rasch model, credited to Danish mathematician Georg Rasch (1960), aims to support true measurement. The mathematics behind the model illustrates the notion that when attempting to measure a single trait, test takers are more likely to correctly answer easy items than difficult items; furthermore, all items are more likely to be correctly answered by people with high ability on the construct being assessed than by those with low ability. Although that description of the Rasch model refers to a dichotomous right/wrong test, the Rasch model has been extended to other data types such as those from rating scales and partial credit models. The need to use Rasch analysis for rating scale data was described in the seminal work, *Rating Scale Analysis* (Wright & Masters, 1982). Data collected from tests and rating scale surveys used in the field of school psychology, to measure one trait, may never fit the Rasch model expectations *exactly*, but the assumption is that they should fit the model well enough to produce useful measures. An example of this broad principle is a thermometer. A thermometer provides a useful measure of a single trait. Although a measure made with this instrument is not an exact assessment of temperature, the thermometer evaluates temperature well enough to provide information that can be used reliably to make decisions.

Rasch measurement is aligned with the idea of "objective measurement"—no matter what construct is being measured, or what measurement instrument is being used, a common metric is used to express results (see Program Committee of the Institute for Objective Measurement, 2000). As an analogy, whether we are measuring a three-inch stick or a three-inch glass, a ruler will tell us that both objects are three inches long. In contrast, if we used a ruler made of rubber, we might not always have information on a common scale. Rather, we might incorrectly conclude that the stick is taller than the glass. Some researchers contend that many of the instruments used in the social sciences use a ruler made of rubber because they actually provide ordinal data but are often claimed to provide quasi-interval data. Because Rasch analysis can provide "measures" expressed on an equal interval scale, school psychologists can develop and use measures to confidently inform decisions.

### *2.2. Exemplars of past applications of the Rasch model*

Rasch measurement has been applied in a variety of ways in education, school psychology, and many other fields. It has been used to (a) develop, evaluate, and improve surveys and tests, and (b) facilitate the computation of Rasch "measures" that lead to data analysis and interpretation of greater confidence (because equal interval data are being utilized). One example of the latter is the Lexile framework (see Stenner, 1996), a measurement system that has been used to provide information on the reading ability of individuals and the difficulty of texts on an equal interval Rasch calibrated scale (Meta Metrics Inc, 2014). Other assessments, such as the Measures of Academic Progress (Northwestern Evaluation Association, 2009), also compute student "measures" using the Rasch model. Furthermore, large-scale international assessment programs such as the Program for International Student Assessment (PISA) and Trends in International Mathematics and Science

Study (TIMSS) have used the Rasch model to work toward achieving invariant measurement across settings and over time.

A review of school psychology journals over the past several years has revealed additional instruments that have been developed, validated, or otherwise studied using the Rasch model, across several cultural contexts (e.g. DiStefano & Morgan, 2010; Wechsler et al., 2010). Despite this emerging research base, some tests and surveys used in school psychology are not developed based upon a strong theory of what it means to measure nor is Rasch measurement used to compute linear scale score measures. However, it is being increasingly recognized that equal interval measures (Rasch measures) may be a way to produce data that can more thoughtfully inform instruction (Vander Ark, 2013).

### *2.3. The logit scale and other advantages of Rasch analysis*

One reason for the use of Rasch measurement techniques is that raw scores are nonlinear and differences between any two consecutive raw scores cannot be assumed to represent equal intervals. See Wright (1992) for a discussion of some of these issues. For example, if Jim achieves a raw score of 20, Sue scores 22, and Jen scores 32 on a Likert-scale social skills assessment where answering Strongly Agree (coded as 4 raw score points) is a higher level of social skills than an answer of Agree (3 raw score points) or Disagree (2 raw score points) or Strongly Disagree (1 raw score point), we can deduce that Jen has a higher level of social skills than Sue and Jim and that Sue has a higher level of social skills than Jim. However, we do not know *how much* higher because the data are ordinal. The jump from Agree to Strongly Agree cannot be assumed to be the same as the jump from Disagree to Agree. One might guess that the difference in social skills between Jen and Sue is greater than the difference between Jim and Sue; however, this conclusion is informed by the erroneous assumption that we are working with equal interval data. All that one knows is that Jen's social skills > Sue's social skills > Jim's social skills. Fortunately for any individual wishing to utilize rating scale surveys to make decisions, these limitations can be addressed through the use of the Rasch model, and Rasch software that uses the Rasch model. Using Rasch techniques allows the data to be expressed on an interval scale, person measures on a logit (interval) scale to be computed, and item measures on the same logit scale to be computed. Details of these and other issues are provided in Best Test Design (Wright & Stone, 1979) and Rating Scale Analysis (Wright & Masters, 1982). In particular, these references provide a "by hand" example of how the raw data are evaluated through the use of the Rasch model.

Why are equal interval measures of importance? Computing "equal interval measures" can allow school psychologists to confidently compare the growth of, and between, schools and students that are located at different portions of a single trait. More specifically, Rasch analysis allows each person's measure to be described using instrument items (e.g. Jen from our example above can be described in terms of her predicted response to specific items) and compared to other respondents (e.g. Jen's predicted response to a specific survey item can be compared to Sue's predicted response to the same survey item). Another benefit of Rasch analysis is that a smaller number of targeted items can provide more reliable measures than a larger number of items (Embretson & Hershberger, 1999); therefore, this technique can be used with small samples. For example, past efforts suggest as few as 30 items administered to 30 respondents can produce useful measures (Linacre, 1994). Using a Rasch approach can also allow school psychologist researchers and practitioners to target instruction/intervention because the expected performance of a person on an item can be inferred from each person's ability measure and the difficulty of items which are expressed on the same scale. This facilitates effective decisions on what the next skill to teach a person should be. Another reason for utilizing Rasch analysis techniques is that when there are different forms of a survey or test, it is possible to express a person's measures on the same scale regardless of which survey or test form was completed by a respondent (Wright & Stone, 1979). Finally, numerous indices can be used to evaluate the measurement functioning of an instrument (e.g. validity, reliability), several of which will be reviewed in subsequent sections of this paper. Resources such as Bond and Fox (2007), Boone, Staver, and Yale (2014), and the extensive Winsteps (Linacre, 2012b) manual provide additional guidance as to the assumptions, promise, and limitations of the Rasch model.

## 3. Rasch analysis techniques

In this section, we first review some preliminary preparations that must be considered before conducting a Rasch analysis. Then, we discuss a number of core Rasch topics: item measures, person measures, Wright Maps, fit statistics, Rasch reliability indices, and some added nuances of rating scale analysis. When covering these topics, we provide an overview of each concept and its utility, rather than step-by-step instructions for how to run an analysis. We do this because of the wealth of information available in other sources that provide this granular level of detail; readers should refer to Boone et al. (2014) and Linacre (2012b) for “how to” information. Finally, throughout this section, we refer to the Resistance to Change in Schools Survey (RCSS) instrument. The second author of the primer presented in this paper developed and piloted this self-report rating scale as a means for demonstrating how Rasch analysis techniques can be operationalized and applied in school psychology.

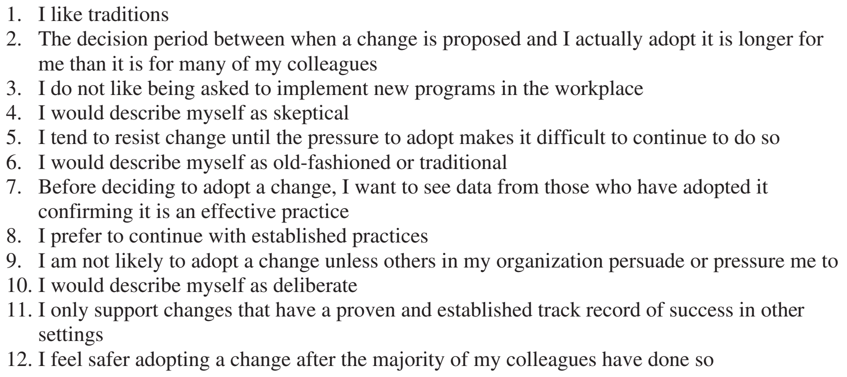
### *3.1. Using an instrument that measures one trait*

Before conducting a Rasch analysis, it is important to ensure that several assumptions are met. First, the instrument should be based on theory concerning the construct. Also, although the Rasch model has been extended to multi-trait measures (e.g. Wu, Adams, Wilson, & Haldane, 2007), for most novice Rasch analysis users it is a best first step to think about how one scale should measure one construct (one variable, one trait). Just as a single metric assessing height, weight, and circumference would not prove helpful for selecting clothing sizes, neither would a one size fits all metric summarizing multiple distinct psychological constructs be useful to inform specific school psychology services. When developing the RCSS items, the second author defined the construct being measured as “resistance to change” using Rogers (2003) diffusion of innovation theory as a guiding framework.

The construct being assessed by an instrument should also range on a continuum from lower levels to higher levels of the construct. When creating items for the RCSS (see Figure 1), the second author initially made a vertical line and placed items on that line in a hypothesized order of those that would be easier to endorse (near the bottom) to those that would be more difficult to endorse (near the top), with the aim of developing items that tapped into a variety of difficulty levels. For example, describing oneself as deliberate was hypothesized to represent a lower level of resistance to change, whereas not adopting change until pressured to do so was hypothesized to represent a higher level. If one cannot make predictions as to the location of items on the continuum, then measurement should not be attempted, for not being able to predict item location reveals a lack of understand regarding the construct of interest. See Wilson (2005) for added details to this issue.

### *3.2. Preparing data for Rasch analysis*

Although data can be entered directly into Winsteps, we recommend importing it from a spreadsheet application (e.g. Excel, SPSS) where rows are structured to represent persons and columns to



1. I like traditions
2. The decision period between when a change is proposed and I actually adopt it is longer for me than it is for many of my colleagues
3. I do not like being asked to implement new programs in the workplace
4. I would describe myself as skeptical
5. I tend to resist change until the pressure to adopt makes it difficult to continue to do so
6. I would describe myself as old-fashioned or traditional
7. Before deciding to adopt a change, I want to see data from those who have adopted it confirming it is an effective practice
8. I prefer to continue with established practices
9. I am not likely to adopt a change unless others in my organization persuade or pressure me to
10. I would describe myself as deliberate
11. I only support changes that have a proven and established track record of success in other settings
12. I feel safer adopting a change after the majority of my colleagues have done so

**Figure 1. Items on the RCSS (for this primer, response options used were: Strongly Disagree, Disagree, Agree, Strongly Agree).**

Source: Winsteps output.

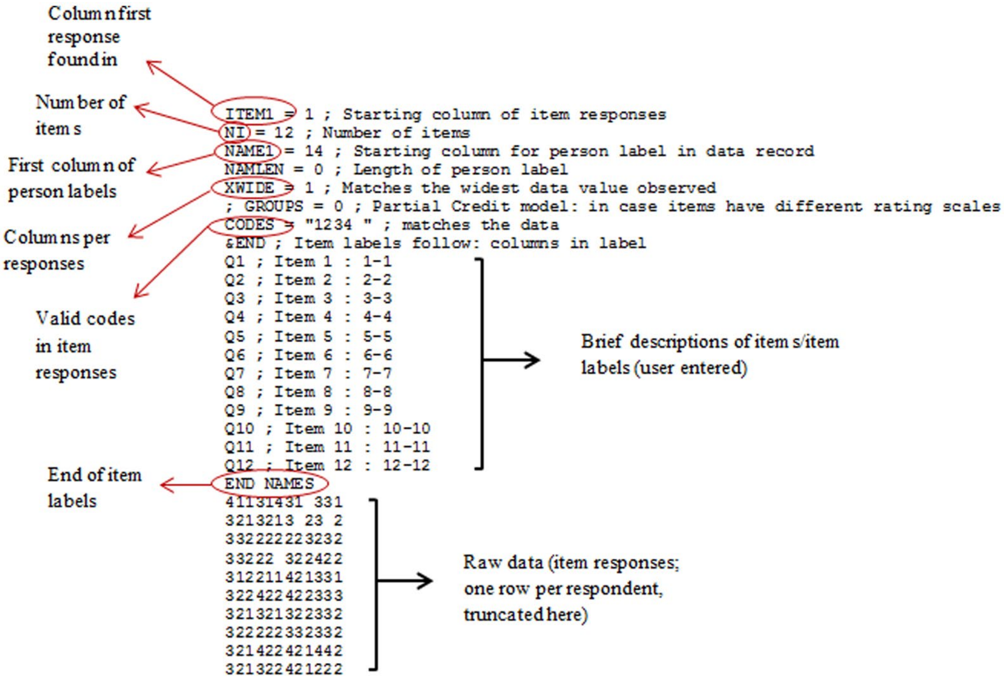
represent items. In addition, we recommend that ordinal or categorical data be assigned numeric values so that the desired performance is the highest number and the least desired performance is the lowest number. For a right–wrong test, a "0" would be an incorrect response and a "1" would be a correct response; in the case of a 4-step Likert scale measuring social skills, a "1" would denote selection of the lowest level of social skills and a "4" would be the highest level of social skills. Missing data should be indicated with a non-numeric value, such as "X" or "." Finally, it is critical to remember to reverse-code any items assessing the construct in the opposite direction. For example, if we added two items to the RCSS to assess openness to change or innovativeness, the Likert scale coding for those items for each respondent would need to be "flipped" (e.g. a 1 would become a 4 and a 4 would become a 1) since these two new items measure the construct of change in the opposite direction of the other items. For the data we evaluated for this primer, we utilized a code of 4 to indicate selection of Strongly Agree, a code of 3 to indicate selection of Agree, a code of 2 to indicate selection of Disagree, and a code of 1 to indicate selection of Strongly Disagree.

After the data have been entered and imported, the next step is to create a "control file" for Winsteps. The control file, which helps the program understand the form of the data, has three parts (Boone et al., 2014): (1) The code that tells Winsteps how to read the data and perform the analyses, (2) the names for each survey item, and (3) the data. See Figure 2 for a portion of the RCSS control file.

### 3.3. Assessing fit

An important consideration within a Rasch framework is "fit." A quality control mechanism, fit, evaluates how well the data conform to the Rasch model. If data deviate greatly from the Rasch model, the causes need to be considered and the misfitting person or item may or may not be removed. It is helpful to consider an analysis of "fit" as a step to investigate if the items of an instrument involve one trait and if the responses of individuals lend themselves to the confident computation and communication of a person measure along a single trait. Two statistics that can be used to assess fit are infit ("inlier-sensitive or information-weighted fit," Linacre, 2012a) and outfit (outlier-sensitive fit, Linacre, 2012a). For the introductory analysis we present, outfit will be used.

Fit statistics are commonly reported in two forms: mean squared (MNSQ) and *z*-standardized (ZSTD). MNSQ is the mean of the squared residuals for an item (Bond & Fox, 2007). In contrast, ZSTD,



```
ITEM1 = 1 ; Starting column of item responses
NI = 12 ; Number of items
NAME1 = 14 ; Starting column for person label in data record
NAMLEN = 0 ; Length of person label
XWIDE = 1 ; Matches the widest data value observed
; GROUPS = 0 ; Partial Credit model: in case items have different rating scales
CODES = "1234 " ; matches the data
&END ; Item labels follow: columns in label
Q1 ; Item 1 : 1-1
Q2 ; Item 2 : 2-2
Q3 ; Item 3 : 3-3
Q4 ; Item 4 : 4-4
Q5 ; Item 5 : 5-5
Q6 ; Item 6 : 6-6
Q7 ; Item 7 : 7-7
Q8 ; Item 8 : 8-8
Q9 ; Item 9 : 9-9
Q10 ; Item 10 : 10-10
Q11 ; Item 11 : 11-11
Q12 ; Item 12 : 12-12
END NAMES
41131431 331
3213213 23 2
332222223232
33222 322422
312211421331
322422422333
321321322332
322222332332
321422421442
321322421222
```

Column first response found in
Number of items
First column of person labels
Columns per responses
Valid codes in item responses
End of item labels
Brief descriptions of items/item labels (user entered)
Raw data (item responses; one row per respondent, truncated here)

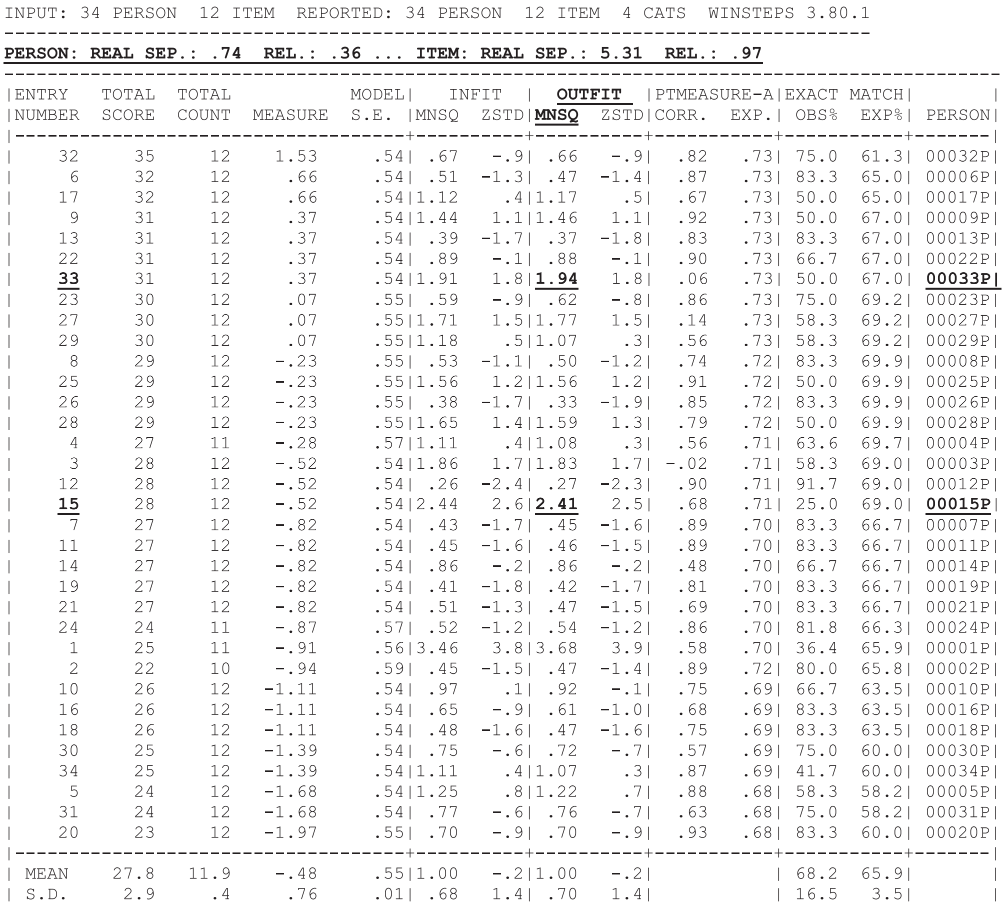
**Figure 2. Portion of the RCSS control file (data are only presented on the first 10 participants to save space). The Winsteps (Linacre, 2017) program can be used to nearly automatically create a control file for researchers.**

Source: Winsteps output.

the standardized form, is a transformation of the mean square value with a sample size correction (Bond & Fox, 2007). Furthermore, whereas ZSTD is sample-size dependent, MNSQ is, “sample-size independent when the noise in the data is spread evently across the population” (Linacre, 2014). Considering the different types and forms of fit statistics, which should be used to evaluate misfit? Although varying recommendations have been presented, Boone et al. (2014) and others recommend examining MNSQ outfit first. Wright and Linacre (1994) suggest that MNSQ values less than 1.4 are acceptable for rating scale data, and values less than 1.3 are acceptable for multiple-choice tests that are not high stakes.

For the beginning researcher, knowledge of some rules of thumb to identify persons and items who do not match up with the goal of computing Rasch measures is helpful. In Figure 3, we present Winsteps Table 17.1, which was produced using the Winsteps control file. For the RCSS, the outfit MNSQ column reveals several individuals who demonstrate some degree of misfit, most notably persons 33 and 15. Figure 4, which reads in the same manner as Figure 3, provides a table from Winsteps that supplies the item outfit MNSQ data. Figure 4 reveals that survey item 6 is the only item flagged as misfitters (using the rule of thumb that MNSQ outfit needs to be greater than 1.3 in order to identify survey items with potential measurement idiosyncrasies).

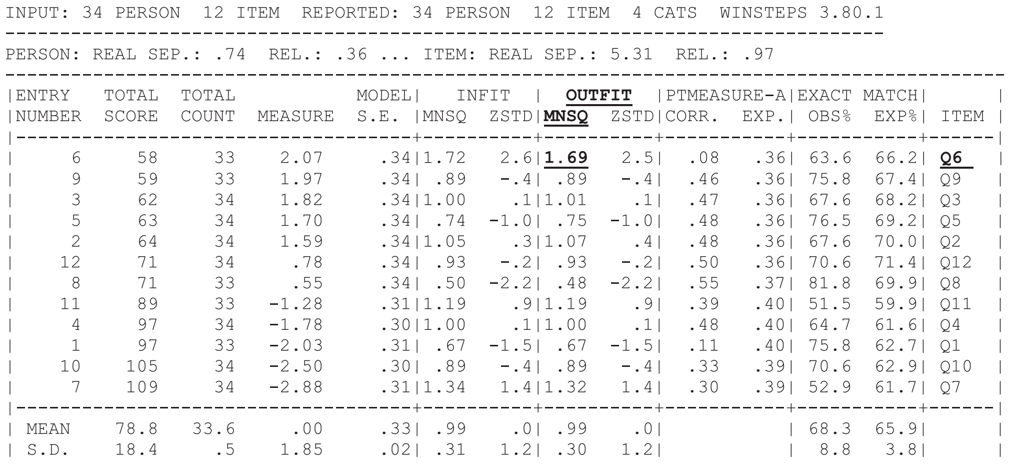
What should be done if a person or item is flagged as potentially misfitting the Rasch model? There is no single set of hard-and-fast rules that must be followed in every circumstance, but here are some potential questions and answers when using misfit criteria. First, it is necessary to consider why misfit might be occurring. Does the misfitting person differ in some way from the target population of interest? Does he or she show a pattern of selecting random answers or guessing? Was the item misunderstood? Could the item measure the trait in a different way than originally intended? Also, it is important to examine the interaction of people and items that might be contributing to



```
INPUT: 34 PERSON  12 ITEM  REPORTED: 34 PERSON  12 ITEM  4 CATS  WINSTEPS 3.80.1
--------------------------------------------------------------------------------
PERSON: REAL SEP.: .74  REL.: .36 ... ITEM: REAL SEP.: 5.31  REL.: .97
-------------------------------------------------------------------------------------------
|ENTRY   TOTAL  TOTAL           MODEL|   INFIT  |  OUTFIT  |PTMEASURE-A|EXACT MATCH|       |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%| PERSON|
|------------------------------------+----------+----------+-----------+-----------+-------|
|    32     35     12     1.53     .54| .67   -.9| .66   -.9|  .82   .73| 75.0  61.3| 00032P|
|     6     32     12      .66     .54| .51  -1.3| .47  -1.4|  .87   .73| 83.3  65.0| 00006P|
|    17     32     12      .66     .54|1.12    .4|1.17    .5|  .67   .73| 50.0  65.0| 00017P|
|     9     31     12      .37     .54|1.44   1.1|1.46   1.1|  .92   .73| 50.0  67.0| 00009P|
|    13     31     12      .37     .54| .39  -1.7| .37  -1.8|  .83   .73| 83.3  67.0| 00013P|
|    22     31     12      .37     .54| .89   -.1| .88   -.1|  .90   .73| 66.7  67.0| 00022P|
|    33     31     12      .37     .54|1.91   1.8|1.94   1.8|  .06   .73| 50.0  67.0| 00033P|
|    23     30     12      .07     .55| .59   -.9| .62   -.8|  .86   .73| 75.0  69.2| 00023P|
|    27     30     12      .07     .55|1.71   1.5|1.77   1.5|  .14   .73| 58.3  69.2| 00027P|
|    29     30     12      .07     .55|1.18    .5|1.07    .3|  .56   .73| 58.3  69.2| 00029P|
|     8     29     12     -.23     .55| .53  -1.1| .50  -1.2|  .74   .72| 83.3  69.9| 00008P|
|    25     29     12     -.23     .55|1.56   1.2|1.56   1.2|  .91   .72| 50.0  69.9| 00025P|
|    26     29     12     -.23     .55| .38  -1.7| .33  -1.9|  .85   .72| 83.3  69.9| 00026P|
|    28     29     12     -.23     .55|1.65   1.4|1.59   1.3|  .79   .72| 50.0  69.9| 00028P|
|     4     27     11     -.28     .57|1.11    .4|1.08    .3|  .56   .71| 63.6  69.7| 00004P|
|     3     28     12     -.52     .54|1.86   1.7|1.83   1.7| -.02   .71| 58.3  69.0| 00003P|
|    12     28     12     -.52     .54| .26  -2.4| .27  -2.3|  .90   .71| 91.7  69.0| 00012P|
|    15     28     12     -.52     .54|2.44   2.6|2.41   2.5|  .68   .71| 25.0  69.0| 00015P|
|     7     27     12     -.82     .54| .43  -1.7| .45  -1.6|  .89   .70| 83.3  66.7| 00007P|
|    11     27     12     -.82     .54| .45  -1.6| .46  -1.5|  .89   .70| 83.3  66.7| 00011P|
|    14     27     12     -.82     .54| .86   -.2| .86   -.2|  .48   .70| 66.7  66.7| 00014P|
|    19     27     12     -.82     .54| .41  -1.8| .42  -1.7|  .81   .70| 83.3  66.7| 00019P|
|    21     27     12     -.82     .54| .51  -1.3| .47  -1.5|  .69   .70| 83.3  66.7| 00021P|
|    24     24     11     -.87     .57| .52  -1.2| .54  -1.2|  .86   .70| 81.8  66.3| 00024P|
|     1     25     11     -.91     .56|3.46   3.8|3.68   3.9|  .58   .70| 36.4  65.9| 00001P|
|     2     22     10     -.94     .59| .45  -1.5| .47  -1.4|  .89   .72| 80.0  65.8| 00002P|
|    10     26     12    -1.11     .54| .97    .1| .92   -.1|  .75   .69| 66.7  63.5| 00010P|
|    16     26     12    -1.11     .54| .65   -.9| .61  -1.0|  .68   .69| 83.3  63.5| 00016P|
|    18     26     12    -1.11     .54| .48  -1.6| .47  -1.6|  .75   .69| 83.3  63.5| 00018P|
|    30     25     12    -1.39     .54| .75   -.6| .72   -.7|  .57   .69| 75.0  60.0| 00030P|
|    34     25     12    -1.39     .54|1.11    .4|1.07    .3|  .87   .69| 41.7  60.0| 00034P|
|     5     24     12    -1.68     .54|1.25    .8|1.22    .7|  .88   .68| 58.3  58.2| 00005P|
|    31     24     12    -1.68     .54| .77   -.6| .76   -.7|  .63   .68| 75.0  58.2| 00031P|
|    20     23     12    -1.97     .55| .70   -.9| .70   -.9|  .93   .68| 83.3  60.0| 00020P|
|------------------------------------+----------+----------+-----------+-----------+-------|
| MEAN    27.8   11.9     -.48     .55|1.00   -.2|1.00   -.2|           | 68.2  65.9|       |
| S.D.     2.9     .4      .76     .01| .68   1.4| .70   1.4|           | 16.5   3.5|       |
```

**Figure 3. Person fit statistics provided in a Winsteps output table. The Outfit MNSQ is provided for each respondent. Additional person statistics of use in a Rasch analysis are also provided.**

Source: Winsteps output.



```
INPUT: 34 PERSON  12 ITEM  REPORTED: 34 PERSON  12 ITEM  4 CATS  WINSTEPS 3.80.1
--------------------------------------------------------------------------------
PERSON: REAL SEP.: .74  REL.: .36 ... ITEM: REAL SEP.: 5.31  REL.: .97
-------------------------------------------------------------------------------------------
|ENTRY   TOTAL  TOTAL          MODEL|   INFIT   |  OUTFIT   |PTMEASURE-A|EXACT MATCH|      |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%| ITEM |
|------------------------------------+----------+----------+-----------+-----------+------|
|     6     58     33     2.07     .34|1.72   2.6|1.69   2.5|  .08   .36| 63.6  66.2| Q6   |
|     9     59     33     1.97     .34| .89   -.4| .89   -.4|  .46   .36| 75.8  67.4| Q9   |
|     3     62     34     1.82     .34|1.00    .1|1.01    .1|  .47   .36| 67.6  68.2| Q3   |
|     5     63     34     1.70     .34| .74  -1.0| .75  -1.0|  .48   .36| 76.5  69.2| Q5   |
|     2     64     34     1.59     .34|1.05    .3|1.07    .4|  .48   .36| 67.6  70.0| Q2   |
|    12     71     34      .78     .34| .93   -.2| .93   -.2|  .50   .36| 70.6  71.4| Q12  |
|     8     71     33      .55     .34| .50  -2.2| .48  -2.2|  .55   .37| 81.8  69.9| Q8   |
|    11     89     33    -1.28     .31|1.19    .9|1.19    .9|  .39   .40| 51.5  59.9| Q11  |
|     4     97     34    -1.78     .30|1.00    .1|1.00    .1|  .48   .40| 64.7  61.6| Q4   |
|     1     97     33    -2.03     .31| .67  -1.5| .67  -1.5|  .11   .40| 75.8  62.7| Q1   |
|    10    105     34    -2.50     .30| .89   -.4| .89   -.4|  .33   .39| 70.6  62.9| Q10  |
|     7    109     34    -2.88     .31|1.34   1.4|1.32   1.4|  .30   .39| 52.9  61.7| Q7   |
|------------------------------------+----------+----------+-----------+-----------+------|
| MEAN    78.8   33.6      .00     .33| .99    .0| .99    .0|           | 68.3  65.9|      |
| S.D.    18.4     .5     1.85     .02| .31   1.2| .30   1.2|           |  8.8   3.8|      |
```

**Figure 4. Item fit statistics provided in a Winsteps output table. The Outfit MNSQ is provided for each item. Additional item statistics of use in a Rasch analysis are also provided.**

Source: Winsteps output.

item misfit. There are tables in Winsteps that allow this level of detail in investigating the reasons for misfitting items and/or respondents (e.g. Winsteps Table 11.1). The important aspect for researchers just beginning to use Rasch is that detailed diagnostics of respondents and items are possible when examining the quality of a data set and evaluating the goal of measuring one single trait. It is often the case, in our experience, that “fit” can be used to highlight items and persons that need to be considered in more detail.

When a respondent misfits, there are many possible causes and steps that can be taken to address the misfit. For example, when a data set is small, it is easy to think a respondent misfits when in fact the misfit is simply partially related to the use of a small data set. Although two respondents misfitted in the current data set, based on our own experience, we would retain these respondents, but experiment with removing the respondents’ answers that caused the misfit and repeat the Rasch analysis. Was it only these odd answers that caused a person to misfit, or is there something else going on? Another step that could be taken would be to remove the respondents from the data set and then repeat the analysis. If data were to be collected a number of times from the same set of respondents, then it might be useful to monitor the fit of these respondents as a function of time point. If an item mistfit for each of three time points, then indeed then the item might best be removed.

### *3.4. Understanding and utilizing person measures and item measures*

Once the control file is created and diagnostics are conducted, a logical next step is to understand and interpret the values reported for person measures and item measures. Person measures are reported in units of logits (often varying from values of −3.00 logits to +3.00 logits, although the original logit measures can be rescaled so that all measures are expressed with positive numbers). Linacre and Wright (1989) provide added details regarding logits for those interested in learning more about logits.

#### *3.4.1. Person measures*

When data are coded and entered in the manner previously recommended, a more positive person measure indicates a person more resistant to change and a more negative person measure indicates a person less resistant to change. Figure 3 provides the “measures” (find the column with the heading MEASURE) for each of the 33 educational professionals who completed the RCSS, arranged from the respondent with the highest measure (1.53 logits) at the top to the person with the lowest measure (−1.97 logits) at the bottom. These measures are expressed on a linear scale; therefore, these measures should be used for any subsequent statistics. For example, if we wished to compare the attitudes of male and female respondents to the survey, we would use the logit measures of all the male respondents and all the female respondents to conduct any statistics. Failing to do so using raw scores would ignore the nonlinear and non-equal interval nature of rating scales.

### *3.4.2. Item measures*

Item measures are also expressed in logit units. With the coding used for our analysis, a higher item measure indicates an item that was harder to agree with (indicating less openness to change), whereas a lower item measure denotes an item that was easier to agree with (indicating less openness to change). Figure 4, which provides item measures, reveals that item 6 is the most difficult item to endorse on the RCSS and item 7 is the easiest item to endorse. More specifically, item 6 is 2.07 logits and item 7 is −2.88 logits. Just as the person measures are expressed using a linear metric (the logit), the item measures are expressed using the same linear logit scale; we discuss the implications of this below when describing Wright Maps.

### *3.5. The Wright Map*

The Wright Map (also named a person–item map) is named after the University of Chicago's Benjamin Wright, who was instrumental in bringing the Rasch measurement model to the attention of researchers in the United States. This map can provide a powerful visual of person–item relationships on an equal interval logit scale. The Wright map helps researchers to (1) assess an instrument's strengths and weaknesses, (2) document the hierarchy of items, (3) compare theory to the observed data, and (4) provide clinical guidance to practitioners (Boone et al., 2014). A part of a Wright Map created with Winsteps for the RCSS data set is provided in Figure 5. For this introductory article, we present only the item part of the Wright Map. Many Wright Maps provide both items and respondents in one plot. We have found that for an introduction to Wright Maps for a rating scale survey, it is helpful just to begin with Wright Maps that only provide item measures.

On the Wright Map from our analysis, each item is plotted. Item measures are listed on the right side of the same vertical line in descending order from the most difficult item to endorse (harder to "agree with" the item) to the easiest item to endorse (easier to "agree with" the item). Using a Wright Map, it is possible to examine whether items are psychometrically redundant, assessing the same level of difficulty on the construct. Such revelations can help guide instrument revision to result in the most effective and parsimonious instrument. For example, if the Wright Map is thought of as presenting items which mark the "cuts" on a meter stick, it can be quickly seen if some valuable cuts are being wasted by the instrument developer. Since there is a limit to the number of cuts (items) that can be posed to a respondent, it makes little sense to present survey items which "cut" the same potion of the single trait. Often, it is far better to have cuts that are distributed along the trait.

When reviewing Wright Maps, it is important to consider whether the ordering of items matches that initially predicted based on theory. If the ordering of items on the Wright Map matches theory, this is evidence that the theoretical construct exists. If some items do not match, we often collect more data to see if the same pattern presents itself. When there is consistent evidence that an item is not located along a trait as predicted, we may revise a theory or review an item and try to consider why the item may be in the wrong location. Is there something wrong with the text of an item? Is it possible the item is not part of the trait being measured? In the latter case, it might be best to remove the item from the analysis and then to repeat the Rasch analysis. We have found that for beginners' use of Rasch, it is often the Wright Map (which provides the item ordering and spacing) which is of great interest.

### *3.6. Reliability and separation*

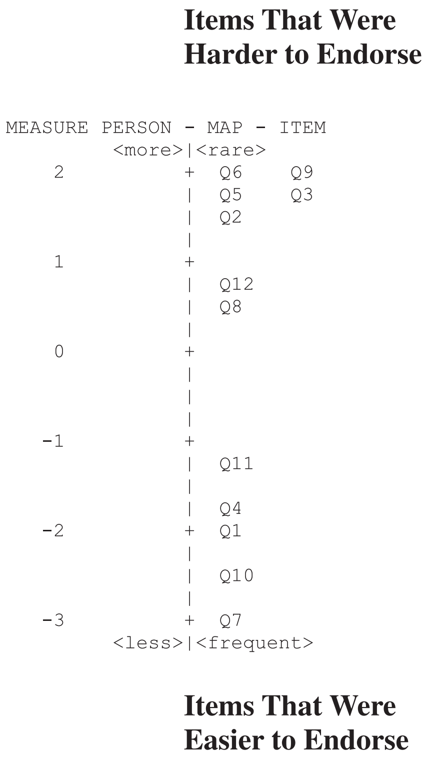
In addition to the analysis of the Wright Map to evaluate the functioning of the instrument and the use of fit statistics to identify items that may best not be retained for use in the instrument, there are added Rasch analysis indices that can also be used to monitor the functioning of an instrument. Some of the Rasch indices include a person reliability index, an item reliability index, an item separation index, and a person separation index. These indices allow for researchers to examine the stability of person and item ordering. These indices represent an improvement over a true score model of computing a KR-20 or Alpha (see Smith, Linacre, & Smith, 2003). The Rasch person reliability and Rasch item reliability values range from 0 to 1 and can be interpreted much like a Cronbach's alpha.

Rasch analysis also provides so-called person separation indices and item separation indices; the former reveal how well a set of items separates persons measured, and the latter reveal how well a sample of people is able to separate the items (Wright & Stone, 1999). Separation index values can range from 0 to infinity, and higher values indicate better separation. Linacre (2012b) suggest that item separation indices of 3 or greater are desirable. In terms of person separation, an index of 1.50 is acceptable, 2.00 is good, and 3.00 is excellent (Duncan, Bode, Lai, & Perera, 2003).

At the top of Figure 3, four indices are presented for the RCSS (underlined and bolded). These data reveal strong item separation (5.31) and item reliability (0.97). These values suggest a sufficient sample to reveal the hierarchy and spacing of items across different samples of similar respondents. However, the person separation index (0.74) reveals a lower level of person separation and the person reliability index reveals a lower level of person reliability (0.36). These two low values may be a reflection of survey items that collected redundant information. Readers are asked to remember our comments about striving to not make "cuts" with items on the same part of the trait since there is a limit to the number of items which can be presented to a respondent. This suggests that the instrument may not have sufficient sensitivity to consistently differentiate between respondents. It could be that more well-targeted items (items between Q8 and Q11) may be needed in the instrument (see Boone et al., 2014). Another issue may be the role that unused (or rarely used) rating scale steps play in the measurement of respondents. It could be that use of a 6 category rating scale (e.g. Strongly Agree, Agree, Barely Agree, Barely Disagree, Disagree, Strongly Disagree) might better help differentiate respondents. This information concerning instrument functioning is one additional example of how instrumentation for use in school psychology can be enhanced through the use of Rasch analysis.

### *3.7. Rating scale step analysis*

The details provided thus far in this brief summary of Rasch analysis techniques are equally appropriate issues to consider for both dichtomous data (e.g. right/wrong tests) and rating scale data. Below we provide some nuances of Rasch techniques that are most easily understood using a rating scale, and in this example, we utilize a 4 step RCSS rating scale (Strongly Agree, Agree, Disagree, Strongly Disagree).



```
MEASURE PERSON - MAP - ITEM
         <more>|<rare>
    2          +  Q6     Q9
               |  Q5     Q3
               |  Q2
               |
    1          +
               |  Q12
               |  Q8
               |
    0          +
               |
               |
               |
   -1          +
               |  Q11
               |
               |  Q4
   -2          +  Q1
               |
               |  Q10
               |
   -3          +  Q7
         <less>|<frequent>
```

**Figure 5. Part of a Wright Map from the RCSS data, from Winsteps. Items are plotted on the linear (interval) logit scale that ranges from −3 logits to 2 logits for this analysis. Items that are harder to agree with are plotted toward the top of the Wright Map. Items that are easier to agree with are plotted toward the base of the Wright Map. Items at the same measure value can be viewed as cutting the "thermometer" at the same spot. Such items, from a measurement perspective, might be redundant.**

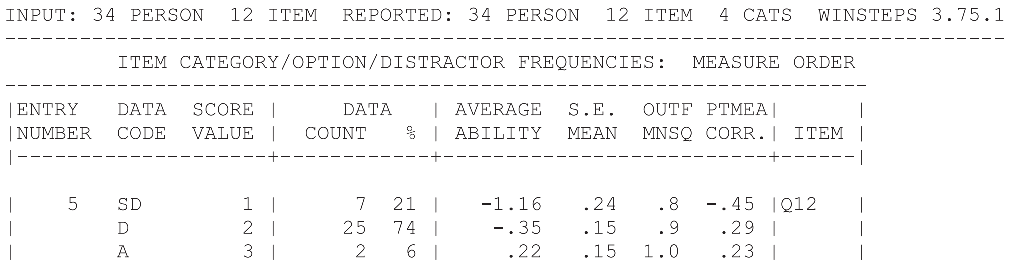
Source: Winsteps output.

Figure 6 illustrates data from the analysis of one of the RCSS items, Q12. The SCORE VALUE, COUNT, and % provide details with respect to what percent of the sample selected a particular rating scale step when answering item Q12. In this case, we can observe that 25 respondents (74% of the sample) selected a 2 (a "Disagree") for this item. Although this breakdown in itself is interesting for all the rating scale steps, it is the AVERAGE ABILITY column which is the most important. This column provides the average Rasch person measure (determined from a respondent completing the entire instrument), which must be compared to the AVERAGE ABILITY values for the respondents who answered the other rating scale steps. In this example, the average ability of all 7 respondents who selected Strongly Disagree (a "1") to item Q12 was −1.16 logits, the average ability of all 25 respondents who selected a Disagree (a "2") to item Q12 was a −0.35, and the average ability of all 2 respondents who selected a Agree (a "3") to item Q12 was 0.22. The importance of these average ability measures is that if the instrument is functioning well, an increase in average ability for each step up (e.g., from a "1" to a "2," from a "2" to a "3") on the scale would be expected. If this pattern is not present for an item, the item may need to be revised or removed. There are also issues associated with sample size to consider, but for purposes of our article we wish to present this basic analysis that considers whether or not the pattern of responses matches that which would be predicted in a well-functioning instrument.

A third step to evaluate the functioning of a rating scale involves the probability curve presented in Figure 7. As discussed by Bond and Fox (2007, p. 224), probability curves "… show the probability of endorsing a given rating scale category for every agreeability-endorsability … difference estimate." Each rating category should have a peak on the curve, revealing that it is the most probable category for some portion of the construct (Bond & Fox, 2007). Figure 7 reveals a well-functioning rating scale and items, with the top trace including 1s, 2s, 3s, and 4s. This suggests that throughout a range of person–item interactions, there are some instances in which each rating scale is "most likely." If it were not the case that a rating scale was most likely (for example, no segment where the 3s are located in the highest trace), this would suggest a rating scale that did not fully maximize the measurement potential of a rating scale survey.

### *3.8. The interplay of items, persons, and rating scales*

Earlier, we introduced readers to the Wright Map, a useful tool for exploring how an instrument's item marks a trait (readers are reminded that many Wright Maps include both persons and items, and it is for this introductory article we have decided to only present the item part of the Wright Map). Figure 8, which is provided using Winsteps for a Rasch analysis, can be used to bring meaning to the "measure" of a respondent or a group of respondents. To understand the power of the table, first note on the far right side of the table—the listing of items from easiest to agree with to hardest to agree with—Q7 is at the base (the easiest to agree with item), and item Q6 is at the top of the table (the hardest to agree with item). This is the same pattern of item difficulty seen in the Wright Map.

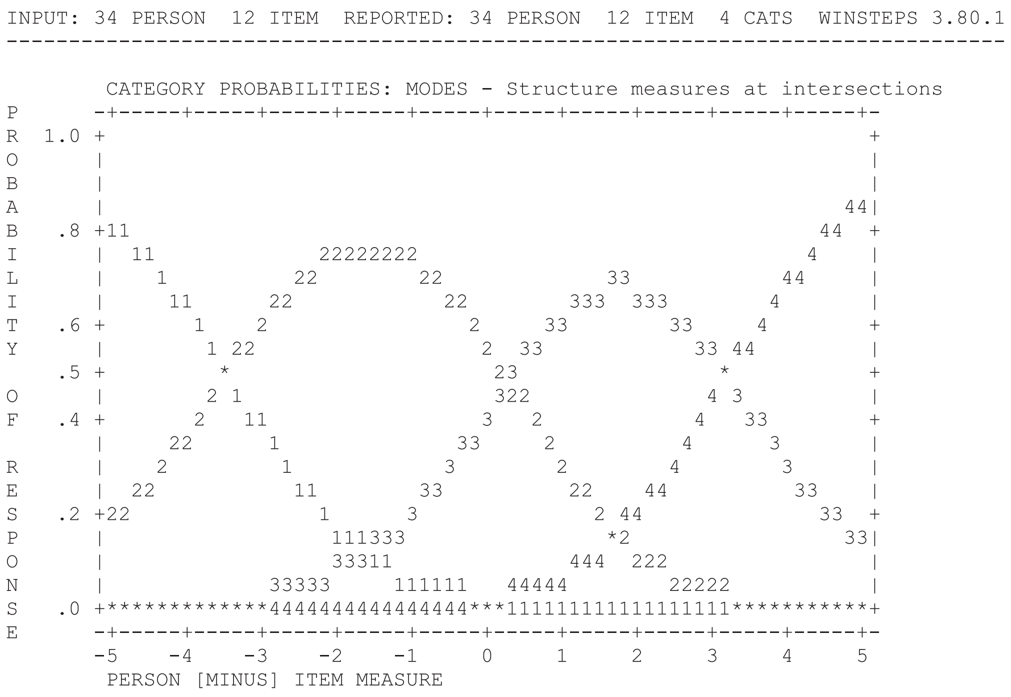


```
INPUT: 34 PERSON  12 ITEM  REPORTED: 34 PERSON  12 ITEM  4 CATS  WINSTEPS 3.75.1
--------------------------------------------------------------------------------
         ITEM CATEGORY/OPTION/DISTRACTOR FREQUENCIES:  MEASURE ORDER
----------------------------------------------------------------------
|ENTRY   DATA  SCORE |     DATA    | AVERAGE  S.E.  OUTF PTMEA|      |
|NUMBER  CODE  VALUE |  COUNT    % | ABILITY  MEAN  MNSQ CORR.| ITEM |
|--------------------+-------------+--------------------------+------|

|    5   SD        1 |      7   21 |   -1.16   .24    .8  -.45 |Q12   |
|        D         2 |     25   74 |    -.35   .15    .9   .29 |      |
|        A         3 |      2    6 |     .22   .15   1.0   .23 |      |
```

**Figure 6. Item option frequencies in measure order for the RCSS. Seven respondents selected "SD" for their answer to Q12, 25 respondents selected "D" for their answer to Q12, and 2 respondents selected "A" for their response to item Q12. The average measure of all seven respondents selecting "SD" was −1.16 logits. Table produced from Winsteps.**

Source: Winsteps output.

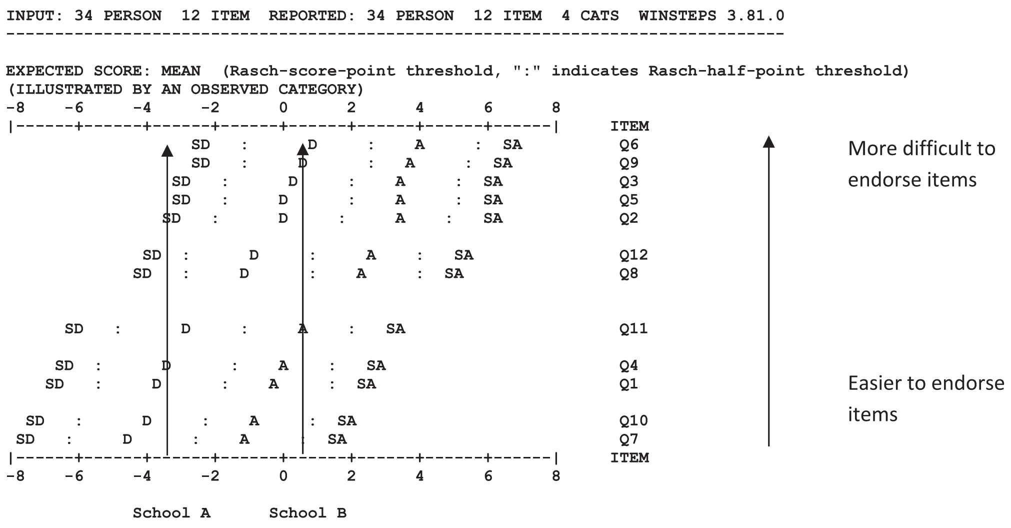


```
INPUT: 34 PERSON  12 ITEM  REPORTED: 34 PERSON  12 ITEM  4 CATS  WINSTEPS 3.80.1
--------------------------------------------------------------------------------

        CATEGORY PROBABILITIES: MODES - Structure measures at intersections
P      -+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-
R  1.0 +                                                             +
O      |                                                             |
B      |                                                             |
A      |                                                           44|
B   .8 +11                                                       44  +
I      |  11             22222222                               4    |
L      |    1          22        22             33            44     |
I      |     11      22            22        333  333        4       |
T   .6 +       1    2                2     33        33     4        +
Y      |        1 22                  2  33            33 44         |
    .5 +         *                     23                *           +
O      |        2 1                    322              4 3          |
F   .4 +       2   11                 3   2            4   33        +
       |     22      1              33     2          4      3       |
R      |    2         1            3        2        4        3      |
E      |  22           11        33          22    44          33    |
S   .2 +22               1      3              2 44              33  +
P      |                  111333                *2                 33|
O      |                  33311              444  222                |
N      |             33333     111111   44444       22222            |
S   .0 +*************4444444444444444***11111111111111111************+
E      -+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-
       -5    -4    -3    -2    -1     0     1     2     3     4     5
        PERSON [MINUS] ITEM MEASURE
```

**Figure 7. Probability curve for the RCSS, from Winsteps. The probability of a response is provided on the vertical axis (from 0 to 1). Each potential response option (1, 2, 3, 4) should be “most probable” for a portion of the horizontal axis.**

Source: Winsteps output.



```
INPUT: 34 PERSON  12 ITEM  REPORTED: 34 PERSON  12 ITEM  4 CATS  WINSTEPS 3.81.0
--------------------------------------------------------------------------------

EXPECTED SCORE: MEAN  (Rasch-score-point threshold, ":" indicates Rasch-half-point threshold)
(ILLUSTRATED BY AN OBSERVED CATEGORY)
-8    -6     -4     -2      0      2      4      6      8
|------+------+------+------+------+------+------+------|     ITEM
                   SD   :      D     :    A     :  SA          Q6
                   SD   :     D      :   A     :  SA           Q9
                 SD   :      D     :    A     :  SA            Q3
                 SD   :     D      :    A     :  SA            Q5
                SD   :      D     :     A    :   SA            Q2

              SD  :      D     :     A    :   SA               Q12
             SD   :     D      :    A     :  SA                Q8


      SD   :      D     :     A    :   SA                      Q11

     SD  :      D      :    A    :   SA                        Q4
    SD   :     D      :    A     :  SA                         Q1

  SD   :      D     :    A     :  SA                           Q10
 SD   :     D      :    A     :  SA                            Q7
|------+------+------+------+------+------+------+------|     ITEM
-8    -6     -4     -2      0      2      4      6      8

             School A      School B
```

**Figure 8. Mean observed categories for RCSS data. The measure of a respondent (or a group of respondents) can be plotted on the horizontal axis, and a vertical line drawn upward. Then, it is possible to predict the response of a respondent (or group of respondents) to each of the instrument items using the rating scale.**

Source: Winsteps output.

Next, note the portions of each row that are marked with a “SD,” a “D,” an “A,” and a “SA.” Second, note the text “School A” and “School B” that has been located at a measure of two possible schools: a School A measure of −3.25 and a School B measure of 0.40 (we are pretending that a sample of the respondents were from School A and a sample were from School B). The vertical lines representing each school illustrate the meaning of each school’s measure. For school A, the regions SD, D, A, and SA crossed by the vertical line for each item show what school A would be predicted to have answered for each and every item. For School A, the predicted answer to Q7 would be Disagree, and for Q6 the predicted answer would be Strongly Disagree. In this plot we show that with Rasch techniques it is possible to plot the location of a group measure (and in fact one could plot the location of each respondent’s measures) and then predict what the answers should be for that “group measure.” Thus, instead of simply reporting a group measure, it is possible to explain the meaning of the group measure. As the items of the survey represent items of differing "agreeability," it should make sense to readers that a respondent with a particular measure will be predicted to have different responses to survey items as a function of where the item lies on the trait.

## 4. Conclusion

Rasch measurement can be used to inform school psychology research and practice by improving the quality of instrumentation functioning, allowing researchers to conduct a detailed analysis of data quality prior to parametric statistics, helping "measures" to be computed for use in statistical tests and enabling the researcher to better communicate test/survey performance. Each respondent's measure (and the average measure of any subgroup of respondents) can be described using items from an instrument to provide context. Instead of groups of respondents simply being compared through a statistical test of significance and effect size, the meaning of the differences can be explained using items (e.g. School A is statistically more resistant to change than School B, and the meaning of this difference is that School A is much more likely to answer Strongly Disagree to an item than School B).

This article merely introduces a few basic Rasch techniques and demonstrates their application with the RCSS; however, it is important to note that more advanced approaches are available. As one example, Rasch can be used to develop different versions of an instrument so it can be targeted to respondents (e.g. two versions of the RCSS can be developed—one for schools highly resistant to change and schools highly supportive of change), but all respondents regardless of form completed can be expressed on the same measurement scale. By presenting a targeted form to a group of respondents, relevant items can be administered, and in the end, a person measure of greater precision can be computed.

Readers interested in learning more about these and other Rasch applications are referred to a variety of websites including the Rasch measurement Special Interest Group: http://raschsig.org/index.html, the Institute for Objective Measurement, Inc.: http://www.rasch.org/, and the Winsteps website: http://www.winsteps.com. In addition, several books can extend the information presented in this article, including Bond and Fox (2007).

**Acknowledgments**
We would like to thank an anonymous reviewer for comments regarding this manuscript. Figures 2-8 are adapted from Winsteps output.

**Funding**
The authors received no direct funding for this research.

**Author details**
William J. Boone[1]
E-mail: boonewj@miamioh.edu
Amity Noltemeyer[1]
E-mail: anoltemeyer@miamioh.edu
[1] Department of Educational Psychology, Miami University, 201 McGuffey Hall, Oxford, OH 45056, USA.

**Citation information**
Cite this article as: Rasch analysis: A primer for school psychology researchers and practitioners, William J. Boone & Amity Noltemeyer, *Cogent Education* (2017), 4: 1416898.

**References**

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Dordrecht: Springer.10.1007/978-94-007-6857-4

DiStefano, C., & Morgan, G. B. (2010). Evaluation of the BESS TRS-CA using the Rasch rating scale model. *School Psychology Quarterly, 25*(4), 202–212. doi: 10.1037/a0021509

Duncan, P. W., Bode, R. K., Lai, S. M., & Perera, S. (2003). Rasch analysis of a new stroke-specific outcome scale: The stroke impact scale. *Archives of Physical Medicine and Rehabilitation, 84*, 950–963. doi:10.1016/S0003-9993(03)00035-2

Embretson, S. E., & Hershberger, S. L. (Eds.). (1999). *The new rules of measurement. What every psychologists and educator should know*. Mahwah, NJ: Lawrence, Erlbaum.

Meta Metrics Inc. (2014). What is a lexile measure? *The Lexile Framework for Reading: Matching Readers with Text*. Retrieved from https://www.lexile.com/about-lexile/lexile-overview/

Linacre, J. M. (1994). *Sample size and item calibration [or person measure] stability*. Retrieved from www.rasch.org/rmt/rmt74m.htm

Linacre, J. M. (2012a). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*(2), 878.

Linacre, J. M. (2012b). *A user guide to Winsteps Ministep Rasch model computer programs: Program manual 3.75.0.* Retrieved from http://www.winsteps.com/a/winsteps-manual.pdf

Linacre, J. M. (2014, June 29). *Infit mean square or infit zstd* [From Rasch Measurement Forum discussion board]. Retrieved from raschforumboards.net

Linacre, J. M. (2017). *Winsteps® Rasch measurement computer program*. Beaverton, OR: Winsteps.com.

Linacre, J. M., & Wright, B. D. (1989). The "length" of a logit. *Rasch Measurement Transactions, 3*(2), 54–55.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology, 88*, 355–383.https//doi.org/10.1111/bjop.1997.88.issue-3

Michell, J. (2002). Steven's theory of scales of measurement and its place in modern psychology. *Australian Journal of Psychology, 54*(2), 99–104.https//doi.org/10.1080/00049530210001706563

Northwestern Evaluation Association. (2009). *Measures of academic progress* [Measurement instrument]. Portland, OR: Author.

Program Committee of the Institute for Objective Measurement. (2000). *Definition of objective measurement*. Retrieved from http://www.rasch.org/define.htm

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedagogiske Institut.

Rogers, E. M. (2003). *Diffusion of innovations* (5th ed.). New York, NY: Free Press.

Smith, R. M., Linacre, J. M., & Smith, Jr., E. V. (2003). Guidelines for manuscripts. *Journal of Applied Measurement, 4*, 198–204.

Stenner, J. A. (1996). *Measuring reading comprehension with the Lexile framework.* Paper presented at the North American Conference on Adolescent/Adult Literacy, Washington, DC. Abstract retrieved from http://eric.ed.gov/?id=ED435977

Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1–49). New York, NY: John Wiley.

Vander Ark, T. (2013, June 5). *Vander Ark on innovation: A proposal for better growth measures* [Web log post]. Retrieved from http://blogs.edweek.org/edweek/on_innovation/2013/06/a_proposal_for_better_growth_measures.html

Wechsler, S. M., Nunes, C. S., Schelini, P. W., Pasian, S. R., Homsi, S. V., Moretti, L., & Anache, A. A. (2010). Brazilian adaptation of the Woodcock-Johnson III cognitive tests. *School Psychology International, 31*(4), 409–421. doi:10.1177/0143034310377165

Wilson, M. R. (2005). *Constructing measures: An items response modeling approach*. Mahwah, NJ: Lawrence Erlbaum.

Wright, B. D. (1992). Raw scores are not linear measures: Rasch vs. classical test theory CTT comparison. *Rasch Measurement Transactions, 6*(1), 208.

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

Wright, B. D., & Masters, G. (1982). *Rating scale analysis*. Chicago, IL: MESA Press.

Wright, B., & Stone, M. (1979). *Best test design*. Chicago, IL: Mesa Press.

Wright, B., & Stone, M. (1999). *Measurement essentials* (2nd ed.). Wilmington, DE: Wide Range,.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. R. (2007) *ConQuest generalized item response modeling software* (Version 2.0) [Statistical analysis software]. Melbourne: Australian Council for Educational Research.