JOURNAL OF APPLIED QUANTITATIVE METHODS

Quantitative Methods Inquires

DEVELOPING A PSYCHOMETRIC RULER: AN ALTERNATIVE PRESENTATION OF RASCH MEASUREMENT OUTPUT

Kenneth D ROYAL
PhD, Psychometrician
American Board of Family Medicine

E-mail: kroyal@theabfm.org

Jennifer Ann ELI
PhD, University of Arizona

Abstract:
Rasch measurement is one of the most popular analytical techniques available in the field of psychometrics. Despite the advantages of Rasch measurement, many researchers and consumers of information have noted that interpreting Rasch output can be an arduous task. The purpose of this paper is to respond to this problem by presenting an alternative method for reporting results that is arguably more user-friendly and easily interpretable by consumers of research.

Key words: Psychometric Ruler; Rasch Measurement Output; Rasch model

Rasch measurement is one of the most popular analytical techniques available in the field of psychometrics and is quickly becoming the norm for instrument validation studies. The advantages of Rasch modeling have been well documented in the literature (see Wright and Stone, 1979; Wright and Stone, 1999; Smith, Jr. & Smith, 2004; and Bond & Fox, 2007). Despite the advantages of Rasch measurement, many researchers and consumers of information have acknowledged that there is much room for improvement with regard to output reporting. This is not to say measurement software creators have failed by any means, but being able to interpret Rasch output, such as the "item map" (or Wright Map), can be an arduous task. In the authors quest to more effectively convey the valuable information obtained from Rasch analyses, this work is intended to provide an alternative presentation of Rasch output that is more user-friendly and easily interpreted by consumers of research. Particularly, the authors will produce a psychometric ruler comparable to that of the physical sciences that can be interpreted in the same way.

This article will begin by providing an overview of objective measurement in the social and behavioral sciences, followed by a brief synopsis of Rasch measurement. A discussion of the psychometric ruler will be presented, followed by an explanation as to how readers can produce one from their own Rasch output. A demonstration will be provided on

a universally interesting topic, namely measuring skepticism. A presentation of the psychometric ruler will follow, accompanied by a discussion of how to interpret the results. Strengths, weaknesses and implications of the psychometric ruler will also be discussed.

LITERATURE REVIEW

Objective Measurement, Abstractions and the Imaginary Inch

Psychometrics is the field of study that attempts to measure psychological factors, such as knowledge, abilities, attitudes, personality traits, and so on. In psychometrics, tests, surveys, and other instruments are used to quantify and measure abstract data. When most people think of measurement they think of concrete measures. For example, a person's height in inches appears to be a concrete measure. However, the inch used to measure height is a man-made idea. There is no such thing as a naturally occurring inch. Inches are simply abstractions that have taken on meaning for the purpose of generating a common frame of reference. Thurstone (1931) said:

> The linear continuum which is implied in all measurement is an abstraction… there is a popular fallacy that a unit of measurement is a thing such as a piece of yardstick. This is not so. A unit of measurement is always a process of some kind which can be repeated without modification in the different parts of the measurement continuum (p. 257).

Once a common frame of reference exists, more meaning is available. When a hierarchy of some kind is produced additional meaning is provided. For example, measures from 1-10 imply that 10 is more than 9, 3 is less than 4, and so on. Also, when the distances between the measures are interval in nature, it implies that 4 is twice as much as 2, and 5 is half the amount of 10. It is these properties that many famous researchers have required for objective measurement (see Campbell's (1920) requirement for concatenation, L. L. Thurstone's Law of Comparative Judgement (1927), Guilford's (1936) definition of measurement, and Luce and Tukey's (1964) requirement for conjoint additivity).

Again, consider the example of height. Suppose we take a sample of 100 adults and measure their heights. Perhaps previous research tells us that most adults will fit into the range of 50'' to 80". We do not need to develop a scale that ranges from 1 to 100 inches to describe our sample. We may wish to simply create a ruler that contains the ranges 48" to 84" and determine where within this range the top of each person's head fits on the scale. The range of 48-84" does not mean that we will not encounter people in our sample that are less than 48" tall or greater than 84" tall. This range is simply a useful criterion for measuring the average range of heights for adults. It is important to understand that whatever the range of the scales we used, the meaning of the inch as it relates to height does not change.

When we make measures of mental constructs we must adhere to the same criteria. When we administer a test or a survey, our common frame of reference is its items. Like a ruler, items must be placed along a continuum, a hierarchy ranging from easy to difficult (to endorse). Just as Guttman (1944) realized a test score is ambiguous without understanding the response pattern of the scores represented, we must also realize the probabilistic nature

of the interactions between the persons and items. That is, a more able person always has a greater probability of getting particular items correct than someone who is less able. Conversely, an item that is very difficult will always have a greater probability of being answered incorrectly than a less difficult item. When measuring the mental construct of "ability" an estimate can be established based on the difficulty level of a particular item along with how an individual responds to that item. The same concept can be extended to surveys and non-tests scenarios where one is concerned with measuring a person's agreeability to statements that contain their own varying degree of difficulty to endorse.

Rasch Measurement

The Rasch family of models are the only psychometric models that meet the requirements for objective measurement. Rasch models are logistic, latent trait models of probability for monotonically increasing functions. Unlike statistical models that are developed based on data, Rasch measurement models are static models that are imposed upon data. Rasch models assume the probability of a respondent agreeing with a particular item is a logistic function of the relative distance between the person and item location on a linear continuum. Dichotomous and polytomous versions of the model are available, and can be extended into various scenarios. With survey research, polytomous models are often employed. When a survey utilizes a rating scale that is consistent with regard to the number of response options (i.e., a 5-point rating scale for all items), the Rating Scale Model (Andrich, 1978) would be the appropriate model to apply. The formulae for the Rating Scale Model are presented below:

$$\ln (P_{nij}/P_{ni\ (i-1)}) = B_n - D_i - F_j$$

where,

P_{nij} is the probability that person n encountering item $_i$ is observed in category $_j$,

B_n is the "ability" measure of person $_n$,

D_i is the "difficulty" measure of item $_i$, the point where the highest and lowest categories of the item are equally probable.

F_j is the "calibration" measure of category $_j$ relative to category $_{j-1}$, the point where categories $_{j-1}$ and $_j$ are equally probable relative to the measure of the item. No constraints are placed on the possible values of F_j.

Researchers who employ Rasch analysis techniques are largely concerned with the extent to which observed data match what is expected by the model. An evaluation of fit statistics provides key indicators of how well the data fit the model, helping to establish content validity. With survey data, it is critically important that the rating scale is functioning well. An evaluation of rating scale functioning should include confirmation that response options provide some form of ordering and each response option can be distinguished from all other options, thus illustrating that respondents were able to clearly identify the difference between each rating scale category. These quality control checks ensure both the structural and communicative validity of the rating scale.

Because Rasch measurement is not sample dependent, it is expected that the scale would work in the same manner regardless of the sample. For example, males and females who have the same endorsability level should have the same probability of endorsing an item. Therefore, if results revealed males responded to a particular item differently than females, the item would be exhibiting differential item functioning (DIF), therefore possibly

biasing results. Naturally, items that exhibit DIF should be considered for removal as they impede the production of objective scales. Once all necessary quality control checks have been completed and sufficient evidence for validity exists, items can be mapped to produce a hierarchy which speaks to the construct validity of the measures. It is this hierarchy that will be presented in an alternative manner in this work.

Purpose/Objective

The purpose of this study was to provide a demonstration of Rasch measurement and construct a user-friendly and easily interpretable alternative representation of the psychometric ruler resulting from Rasch measurement output. Although the psychometric ruler presented here is largely metaphorical in nature, it does possess the properties and characteristics of a ruler used in the physical sciences. That is, abstract ideas and mental constructs are plotted along a physical ruler to distinguish the psychometric properties of each item in relation to the other. It is the researchers' intentions that presenting results in the manner presented in this work will aid in the understanding of Rasch measurement output, particularly the output produce from Rasch-based item maps.

METHODOLOGY

Sample and Data Source

In 2010, renowned sociologist Dr. Peter Nardi published a study of magicians' beliefs about the paranormal. He was particularly interested in learning to what extent magicians believed various paranormal phenomena were possible. Nardi hypothesized that magicians would make a very interesting research sample because they are either true believers of paranormal phenomena, or because they are essentially "in on the secrets", the biggest skeptics of all. Nardi administered a web-based survey in various magician Websites, discussion boards, and Internet chat rooms and was able to obtain a sample of 227 responses. The lead researcher contacted Dr. Nardi and requested his data. Dr. Nardi kindly obliged and promptly sent the complete dataset and codebook. It is from this secondary source that the data in this study were obtained.

The Psychometric Ruler

Creating a psychometric ruler involves transforming raw scores to interval measures. Winsteps measurement software (Version 3.69) was used to perform the Rasch analysis in this study (Linacre, 2010). Winsteps software produces measures, called logits, for each person and item in the dataset. In order to create a continuum that is meaningful and easy for interpretation, logits often need to be rescaled. Here, the minimum item logit value was -1.67 and the maximum item logit value was 1.02. A rescaling procedure was conducted that placed the minimum logit at 1 and the maximum logit at 10 on the new scale (although this could easily be presented in the opposite manner should a researcher choose). The formula for the transformation of logits to a scaled score is as follows:

SS = m (Di) + b, where
SS = Scaled Score
m = slope
Di = item difficulty estimate
b = intercept

To convert item difficulty estimates to the present scale (1-10), the following formula was used:

$$\text{Rescaled Logit Value} = (3.3457*Di) + 6.5874$$

All rulers require units of measures that are equidistant throughout the scale. Here, scaled scores constitute the units that would be considered inches on a typical ruler. Scaled scores range from 1-10 in this example. Additionally, within each of these scaled score units are additional units that are increments of 10 (whereas an actual ruler would contain increments of eight). These subunits represent 1/10 of a scaled score. The purpose of using increments of 10 is for easy interpretation, as most people are comfortable with scales that range from 0-10, 1-10, 1-100, 100-1,000, and so on.

RESULTS AND DISCUSSION

Whenever Rasch analyses are performed a series of quality control checks must be performed. Largely, these checks evaluate the extent to which observed data fit the model's expectations. Additional checks evaluate the structure and quality of the rating scale, quality of items, and other diagnostics. However, because the purpose of the present study is not to present content findings from a data analysis, per se, the majority of these critical steps of Rasch analysis will not be presented in this study. Instead, results will focus only on information relevant to the alternative presentation of results as guided by the purpose of this study.

Item Statistics

Item statistics for the original Rasch output are presented in table 1.

Table 1

Item Logit Values

Items	Measure	SE
Channeling (spirit controlling a person in a trance)	1.02	.13
Astrology	.81	.12
Communication with the Dead	.74	.12
Bigfoot (Sasquatch)	.74	.12
Loch Ness Monster	.68	.13
Reincarnation	.44	.11
Clairvoyance (Predict the Future)	.36	.11
Ghosts	.07	.10
Haunted Houses	.07	.10
UFOs	-.01	.10
ESP (Extra Sensory Perception)	-.32	.09
Creationism or Intelligent Design	-.89	.09
Devil	-1.00	.09
Angels	-1.07	.10
Life After Death	-1.67	.11

After performing the rescaling procedure mentioned previously, the rescaled logit values are presented in table 2. Note, Winsteps software has a function that can automatically rescale values without having to use the manual formula presented in the methodology section. Also, it is apparent that identical values in table 1 may appear slightly different in table 2. This is because logit values presented in these tables have been rounded to two decimal places. Full logits values were used when transforming to a score, thus why results differ slightly in table 2.

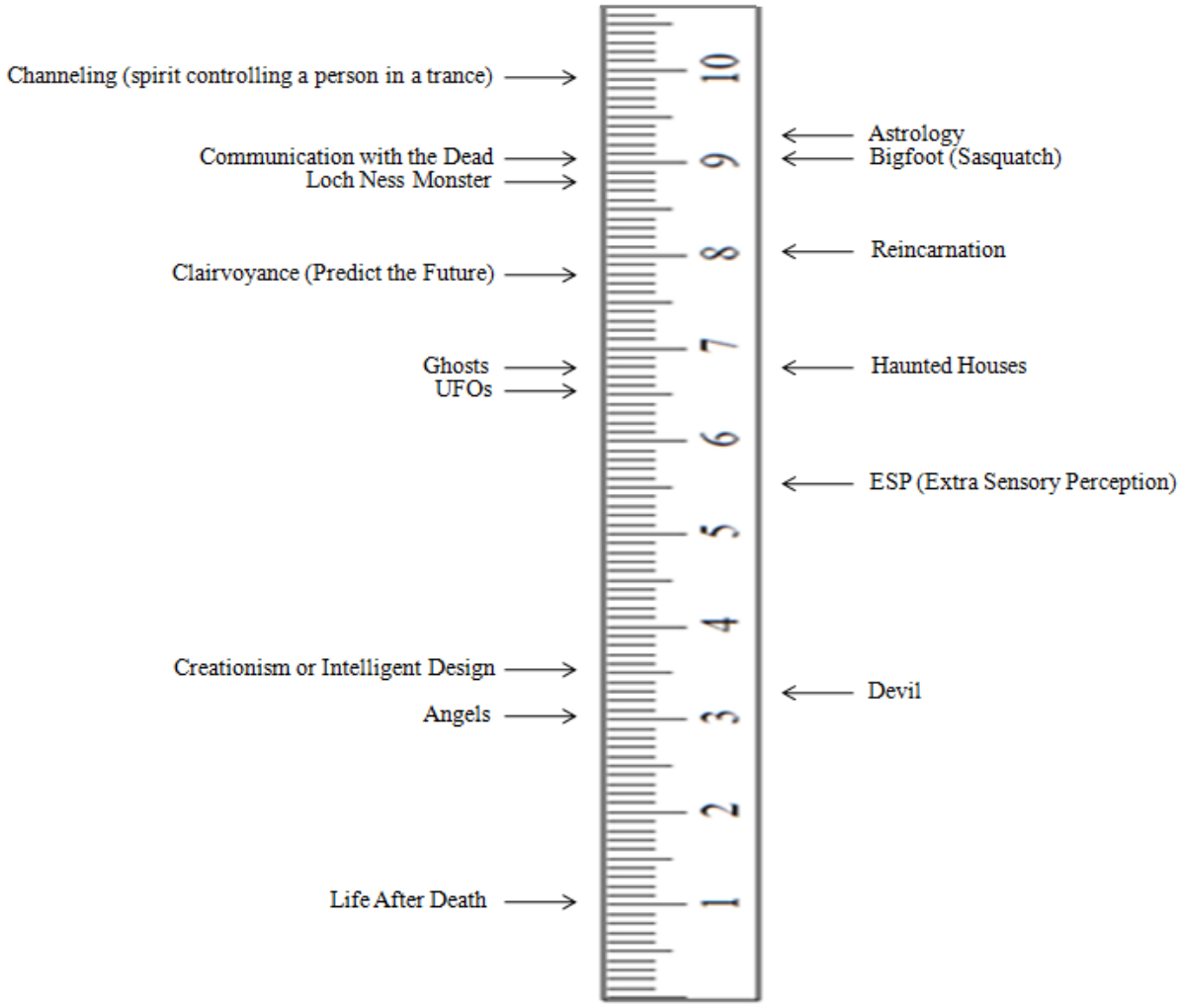
Table 2

Items Rescaled

Items	Measure	SE
Channeling (spirit controlling a person in a trance)	10.00	.45
Astrology	9.30	.40
Communication with the Dead	9.08	.40
Bigfoot (Sasquatch)	9.05	.41
Loch Ness Monster	8.87	.42
Reincarnation	8.05	.37
Clairvoyance (Predict the Future)	7.81	.36
Ghosts	6.84	.34
Haunted Houses	6.83	.33
UFOs	6.56	.34
ESP (Extra Sensory Perception)	5.52	.32
Creationism or Intelligent Design	3.63	.32
Devil	3.25	.32
Angels	3.01	.32
Life After Death	1.01	.36

Psychometric Ruler

As mentioned previously, item logit measures were rescaled to fit a continuum ranging from 1– 10. Although this is purely for metaphorical and illustrative purposes, a physical ruler was created that contained the psychometric values for each item from Table 2. This ruler allows one to visualize the psychometric distance between each item as it relates to the extent to which magicians believed in each of the following (See Figure 1).



To interpret the ruler, first identify the items that appear at the extreme ends. At 1.0, the item *Life After Death* is present. This indicates survey respondents believe this is the easiest item to endorse relative to the others presented on this survey. Located at the other end of the ruler (value of 10.0) is the item *Chaneling (spirit controlling a person in a trance)*. Survey respondents believe this item is the most difficult to endorse (or agree with) relative to the other items presented on this survey. Notice, a hierarchical pattern is present. Items appearing near the bottom of the ruler are indicated to be the easiest to endorse, or in this case believe in, whereas items at the top of the ruler are believed to be the most difficult to endorse (or believe in).

The purpose of this study was to construct a physical ruler for psychological constructs and ideas in order to demonstrate both what is possible in the arena of psychometrics and at the same time produce an alternative presentation of Rasch output results. By examining the ruler one can see that some of the items appearing in close proximity to one another share a conceptual relationship. For example, all items below the "4" mark appear to have a religious or spiritual correlation. Similarly, between 6.5 and 7.0, *haunted houses* and *ghosts* appear in close proximity. Additionally, between (approximately) 8.5 and 9.0 *Bigfoot* and the *Loch Ness Monster* appear close together on the scale. Being able to visualize these conceptual relationships allows readers to better interpret results, and

perhaps develop a more meaningful interpretation of results that are presented solely by numbers.

CONCLUSION & LIMITATIONS

Rulers are such a common symbol in the U.S. society that one would be hard-pressed to find someone who cannot relate to the concept. For this reason, the authors contend that interpreting the psychometric ruler is easy and intuitive. It is the authors' hope that even persons with an aversion to quantitative methods might have an appreciation for the psychometric ruler.

For this demonstration the topic of skepticism was measured. This topic was selected because of its universal appeal, as opposed to a specific content area which may or may not resonate as well with readers. Despite the advantages mentioned in this work, there authors would like to caution that the psychometric ruler is purely metaphorical. Although the psychometric ruler contains many of the properties of a physical ruler (i.e., starting/ending points, interval scaling, precise subscaling, etc.), some elements are not as easily transferrable. For instance, the psychometric ruler ranged from 1 to 10. In actuality, this ruler could have been scaled to any range. Therefore, one cannot say that an item that appears at 10.0 is 10 times greater (or perhaps more intense) than an item that appears at 1.0.

From a Rasch measurement perspective, however, perhaps the greatest limitation of this particular psychometric ruler is only half the information are presented from an actual Rasch measurement software-produced item map. Item maps are particularly useful in presenting the invariant interaction of both persons and items. Probabilities that an individual will correctly answer a test item or endorse a survey item can all be approximated from the item map. The use of the psychometric ruler in this research presents only the item side of the map. Depending upon where the mean of the person distribution falls on the actual item map would determine to what extent persons were able to endorse each item. Although this information is absent in the psychometric ruler, meaningful interpretation of item results can still be made. However, it is pertinent to point out that those who utilize Rasch measurement software will already have full item map output prior to constructing a psychometric ruler like that proposed in this study. Therefore, researchers always have the option to present the results as currently produced by the software, or to produce a psychometric ruler for more user-friendly displays.

An additional, yet minor, limitation of the psychometric ruler (as proposed here) pertains to the nature of difficulty estimates produced from the Rasch analysis. For example, when several items have difficulty estimates that are in close proximity to one another, their proximity on the psychometric ruler will still be very close together even after a re-scaling procedure has been performed. Because all measures have some error associated with them, items that appear at virtually identical locations on the ruler might actually appear in a slightly different order depending on the effects of error. This is inevitable for all measurement. However, in all instances it is good practice to always report both item difficulty measures and standard errors for each item so that readers may better investigate the precision of measurement and have a more informed perspective about the extent to which the psychometric ruler is valid.

In sum, Rasch measurement output has historically been criticized for what some believe to be difficult output to interpret. It is the authors' hope that the preceding demonstration can aid others in their pursuit to conduct better measurement of abstract data and produce more meaningful and user-friendly output for audiences in various arenas.

BIBLIOGRAPHY

1. Andrich, D. **"A Rating Formulation for Ordered Response Categories",** *Psychometrika* 43, 1978, 561-573.

2. Bond, T. G., & Fox, C. M. **Applying the Rasch Model. Fundamental measurement in the human sciences***, 2nd edition*. Mahwah, NJ: Lawrence Erlbaum Associate, 2007.

3. Campbell, N. R. **Physics: The Elements**. London: Cambridge University Press, 1920.

4. Guilford, J. P. **Psychometric methods**. New York: McGraw-Hill, 1936.

5. Guttman, L. **"A Basis for Scaling Qualitative Data",** *American Sociological Review,* 9, 1944, 139-150.

6. Jastak, S., & Wilkinson, G. **"The Wide Range Achievement Test. WRAT3"**, Wilmington, DE: Jastak Assessment Systems, 1993.

7. Linacre, J. M. **Many-faceted Rasch Measurement**. Chicago: MESA Press, 1989.

8. Linacre, J. M. **"Measurement, Meaning and Morality"**, *Rasch Measurement Research Paper #71 (2005):* http://www.rasch.org/memo71.pdf.

9. Linacre, J.M. **WINSTEPS® (Version 3.69.1)**. Computer Software (Beaverton, OR: Winsteps.com, 2010).

10. Luce, R. D., & Tukey, J. W. **"Simultaneous Conjoint Measurement: A New Type of Fundamental Measurement"**, *Journal of Mathematical Psychology* 1, 1964, 1-27.

11. Masters, G. N. & Wright, B. D. **"The Essential Process in a Family of Measurement Models"**, *Psychometrika,* 49, 1984, 529-544.

12. Mallinson, T., Cella, D., Cashy, J. & Holzner, B. **"Giving Meaning to Measure: Linking Self-Reported Fatigue and Function to Performance of Everyday Activities",** *Journal of Pain and Symptom Management* 31, no. 3, 2006, 229-241.

13. Nardi, P. M. Magic, **Skepticism, and Belief: An Empirical Study of What Magicians Believe about the Paranormal**. *Skeptic Magazine*, 15 no. 3, 2010, 58-64.

14. Rasch, G. **Probabilistic models for some intelligence and attainment tests**. Chicago: University of Chicago Press, 1960.

15. Smith, Jr., E. V., & Smith, R. M. **Introduction to Rasch Measurement: Theory, Models and Applications.** Maple Grove, MN: JAM Press, 2004.

16. Stenner, A. J. **Measuring reading comprehension with the Lexile Framework**. Durham, NC: MetaMetrics, Inc., 1996.

17. Thurstone, L. L. **"A Law of Comparative Judgment",** *Psychological Review*, 34, 1927, 273-286.

18. Thurstone, L. L. **"Measurement of Social Attitudes"**, *Journal of Abnormal and Social Psychology* 26, 1931, 249-269.

19. Wright, B. D., & Linacre, J. M. **"Reasonable Mean-Square Fit Values",** *Rasch Measurement Transactions* 8, 1994, 370.

20. Wright, B. D., & Stone, M. H. **Best Test Design**. Chicago: MESA Press, 1979.

21. Wright, B. D., & Stone, M. H. **Measurement Essentials, 2nd edition**. Wilmington, DE: Wide Range, Inc., 1999.