# Assessing the Quality of Your Assessment, Survey, or Other Tool: Rasch Model Interpretation Guide

## Why Use Rasch Modeling?

Rasch modeling is helpful in validating linear interval scales that measure perceptions, skills, abilities, or other traits (DeMars 2010; Rasch, 1980). Several assumptions underpin Rasch modeling and are critical for determining whether it is appropriate to use the model for validation. First, Rasch modeling assumes that each person is being assessed on one trait (e.g., classroom management, instructional planning, collaboration with peers). Second, to assess that trait, each item within a scale needs to be characterized by a difficulty along a continuum from low to high (like a yardstick or ruler). The third assumption is that the likelihood of a person demonstrating a specific trait can be calculated from the difference between the person and item estimates (Bond & Fox, 2007).

Winsteps is heavily referenced in this resource. It might be helpful to review this [video tutorial](https://www.youtube.com/watch?v=FDUYm7ZhXkw) to familiarize yourself with Rasch modeling and Winsteps. The output from Winsteps is largely based on logits, a standardized scale measure, like a standardized *z*-score, where 0 is the mean and 1 logit is one standard deviation.

## List of Winsteps Tables to Examine

Below is a list of tables to examine in Winsteps that can be helpful in targeting areas for improvement for a given scale, using the Rasch model. There are many other tables and figures in Winsteps, but the tables referenced here are a starting point for scale validation.

### Table 1: Wright Map

**The Wright map is useful in examining the item difficulty (ceiling or floor effects) as well as the range of abilities held by the persons represented in the data**. The Wright map provides a comparison of item difficulty and person ability by placing these two principles alongside each other. The results from this output can be used to better understand whether the scale can capture the range of traits held by persons within the dataset.

The Wright map is organized as two vertical distributions: persons are represented on the left side and items on the right (figure 1). Both persons and items are organized from least skilled or easiest at the bottom, to most skilled or difficult at the top. In the red boxes below, you can see that item I0015 is most difficult and item I0011 is the easiest. Likewise, you can see that person 120M2 is the most skilled and person 167M1 is the least skilled.

*Figure 1. Wright Map Example*



The left side of the Wright map shows the mean (indicated by M) as well as two standard deviations (S = one *SD* and T = two *SD*) for person ability. The right side is organized by the same notation, with a mean difficulty level for items as well as two standard deviations. The means and standard deviations for each side are often different. For more information, see [this resource page](https://www.winsteps.com/winman/table1.htm).

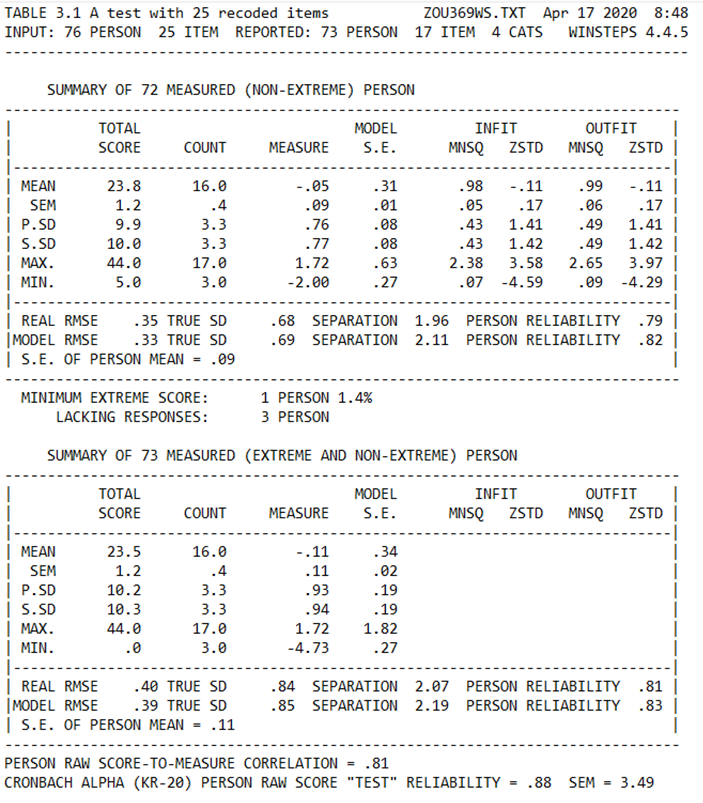
The Wright map can be used to validate the assumption of *monotonicity*, which posits that, as a person’s trait or skill increases, so too should their ability to correctly answer or demonstrate the items contained within the scale. This provides information on possible ceiling and floor effects; whether the scale provides enough items to capture the traits or skills held by persons within the dataset (Linacre, 2021).

### Table 3.1: Summary Statistics

**This table is helpful in validating how many different traits or ability levels the scale can detect as well as in assessing scale reliability**. The summary statistics table provides an overview of overall scale performance based on how many different categories of persons the scale can identify. For example, if a scale is used to assess classroom management ability across three skill levels—*developing*, *proficient*, and *advanced*—it should have a separation value of approximately 3 (figure 2). In this case, a separation of 3 can be interpreted to mean that the scale can distinguish between three skill levels. In the example below, separation is right around 2 for both tables, meaning that there are two categories across all respondents and variates of the Rasch model. The summary statistics table also provides an estimate of reliability for this scale, as measured by KR-20, Cronbach’s alpha, and the generalizability coefficient. For more information, see [this resource page](https://www.winsteps.com/winman/table3_1.htm).

Findings from the summary statistics table can be used to validate the assumption of *sufficiency*, which holds that persons with a similar score or rating on a scale should have the same trait level (Linacre, 2021). That means that items of a similar difficulty level should have a similar probability of being demonstrated with persons with similar levels of a trait (Bond & Fox, 2007). The red box at the bottom of the figure below shows this.

*Figure 2. Summary Statistics Example*



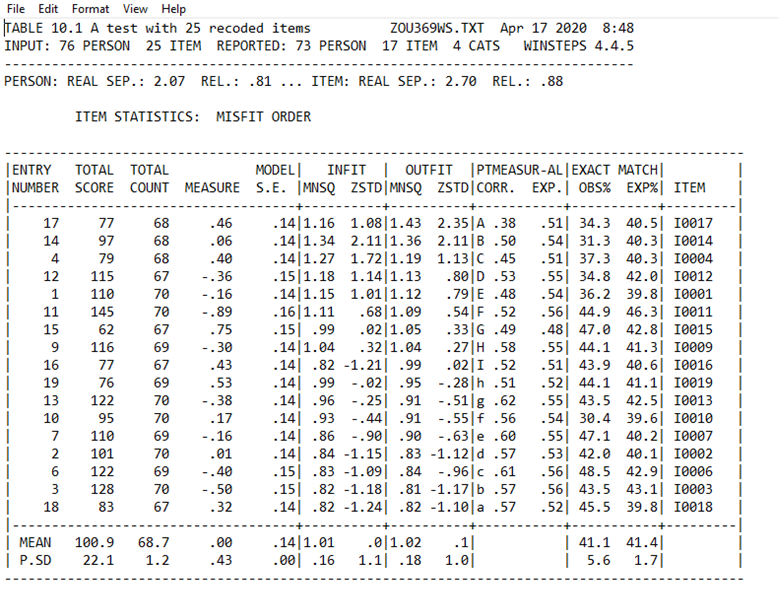
### Table 10. Item Fit

**The item fit table is useful in assessing which items are productive for measurement within the scale and which are redundant or confusing**. The item fit table breaks down which items were demonstrated or answered affirmatively. This can be noted by an item’s position in logits (indicated by the measure column) across the scale. Larger values are associated with an increased level of difficulty, with a mean of 0.

Also captured in the item fit table is *infit*, which is an inlier estimate of redundancy or heterogeneity. Larger infit values are associated with *overfit*, meaning that there is a large amount of noise (confusion) in persons’ responses. On the other hand, items with lower infit values might be redundant and could be removed because they contribute no new information. Ideally, all infit values should be between 0.5 and 1.50.

Outfit is also a useful statistic to examine. This value is associated with unexplained variance. Items with values that fall outside of the range from −2 to 2 might be measuring something different from the other items in the scale (figure 3). The “MODEL S.E.” column shows the level of confidence in the measure—the lower the value, the more confident you can be in the measure. For more information, see [this resource page](https://www.winsteps.com/winman/table10_1.htm).

*Figure 3. Item Fit Example*

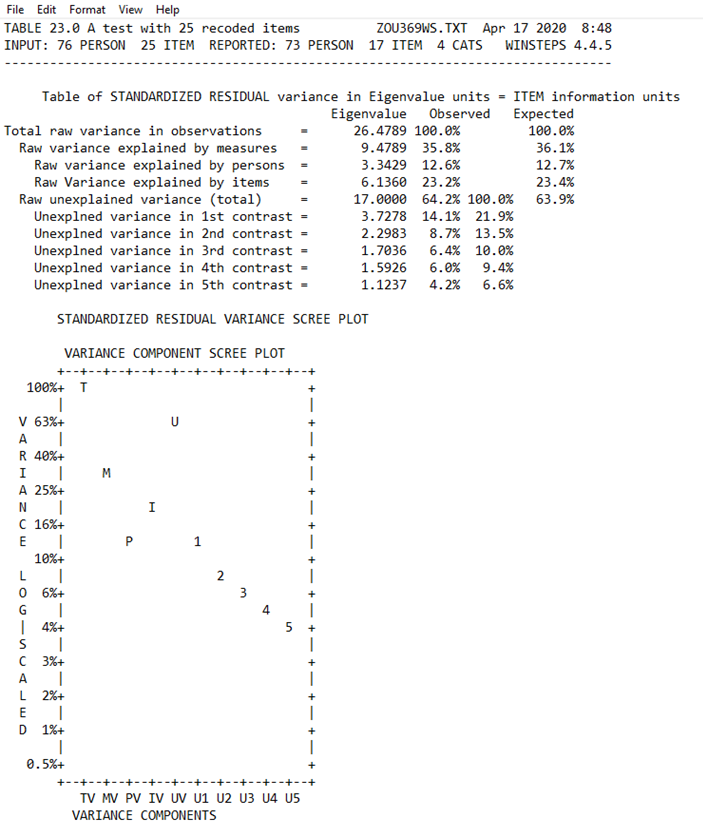


### Table 23: Dimensionality

**The dimensionality tables are useful in assessing whether a scale is measuring a single trait or skill (unidimensional)**. Rasch modeling assumes that items within a single scale measure only one trait, skill, or concept. For this assumption to be validated, at least 40% of the variance or more must be attributable to the first dimension, with an eigenvalue of 2.0. Also, for the first unexplained contrast, eigenvalues must be less than 2.0 and less than 5% of the unexplained variance attributable to the first contrast (figure 4). If this assumption is violated, the scale is likely measuring more than one trait, skill, or concept (Linacre, 2021).

The raw variance explained is provided for both persons and items, as well as the amount of unexplained variance. The unexplained variance is broken down across contrasts (residuals). First, you want to be sure that the raw variance explained by persons and items exceeds that of any single unexplained contrast. Any unexplained contrasts with eigenvalues greater than 2 likely contain enough variance unrelated to the other items in the scale, possibly indicating an additional dimension (Linacre, 2021). To evaluate which items might be contributing to multidimensionality, examine the item clusters for each unexplained contrast. Pay attention to disattenuated correlationswith low values to pinpoint which item is contributing to multidimensionality. For more information, see [this resource page](https://www.winsteps.com/winman/table23_0.htm). Also, [this tutorial video](https://www.youtube.com/watch?v=X4EtIJlPBTU) is very helpful for interpreting output (parts 1 and 2).

*Figure 4. Dimensionality Example*



### Tables 28, 30.1, and 30.2: Differential Item Functioning

**One assumption of Rasch modeling is that different groups are not disadvantaged by the assessment or items in an exam, survey, or aptitude test. Differential item functioning allows us to assess the extent to which a group of items or a single item disproportionately affects the performance of participants from certain groups.** When an item is found to disadvantage a group, it should be revised or removed from the instrument. To start, Table 28 shows the overall difference in the ability of groups within a scale (figure 5). The “MEAN MEASURE” column breaks down the group’s ability, trait, or skill on a scale. This example shows that female students are typically more excited about the topic (science) assessed by this scale. That is, in the “MEAN MEASURE” column, female students are 1.62 logits—equivalent to 1.62 standard deviations—more excited about science than male students.

*Figure 5. Differential Item Functioning Example, Table 28*

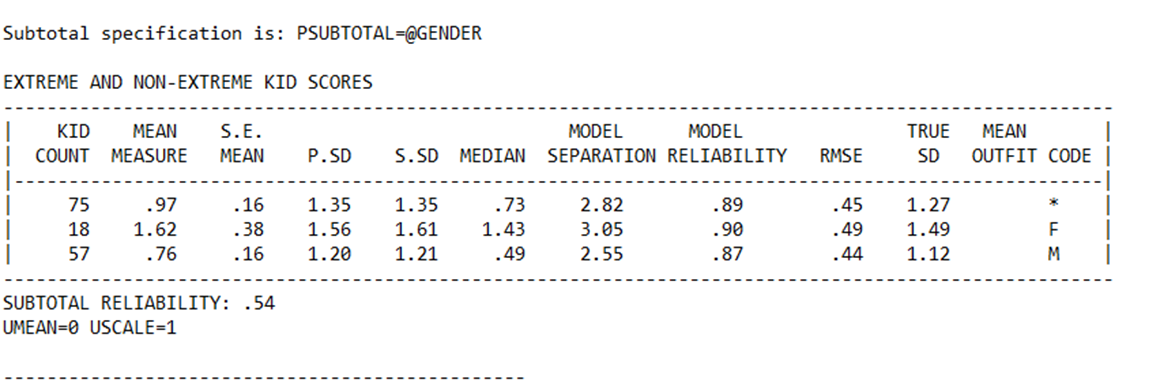


Table 30.1 compares item difficulty between two groups of participants (figure 6). In this case, look at the “DIF CONTRAST” column to assess the size of the difference in logits. Also examine the statistically significant differences, using the “Rasch-Welch t” and “Prob.” (*p* value) columns. Note that the item in the red boxes is more difficult for female students than for male students.

*Figure 6. Differential Item Functioning Example, Table 30.1*

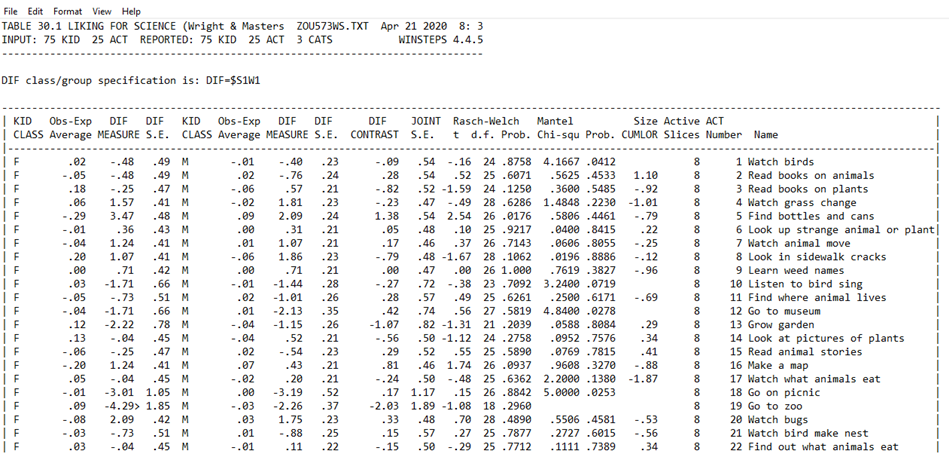
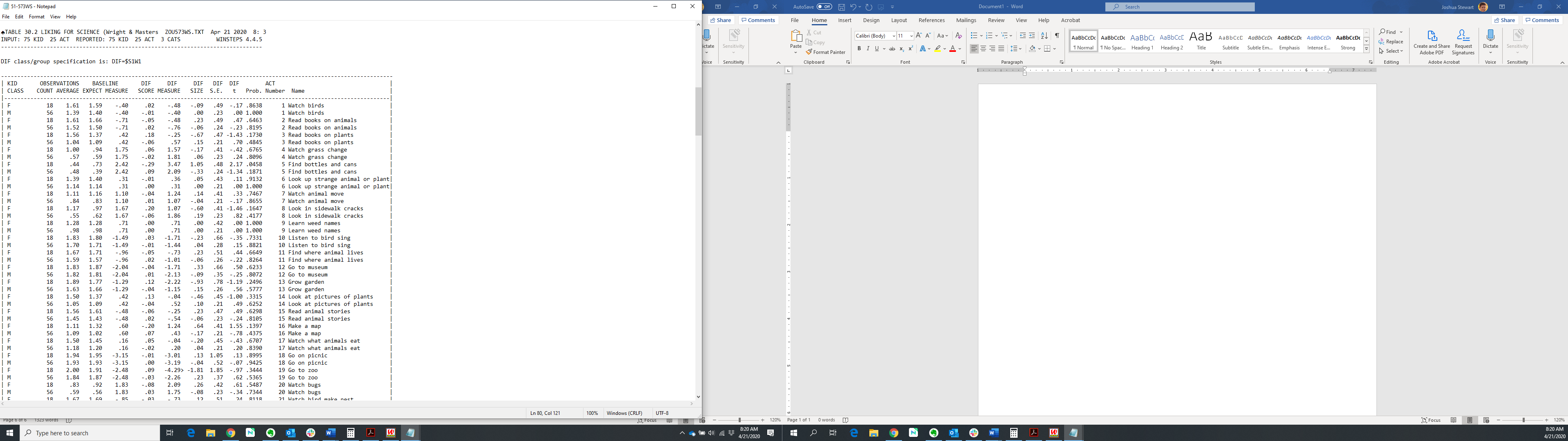


Table 30.2 breaks down items that are disproportionately difficult for members of different groups (figure 7). In this case, pay attention to the “AVERAGE” versus “EXPECT” columns to understand how the item affects these two groups. The “DIF SIZE” column offers the difference of how this item affected these two groups based on their expected and actual performance. Also verify that the item is statistically significant, as evidenced by the “Prob.” and “DIF t” columns. This gives the student’s *t* test statistic and *p* value. In this example, the item is much more difficult (and statistically significant) for female students than for the male students.

*Figure 7. Differential Item Functioning Example, Table 30.2*



For more information on differential item functioning, see [this resource page](https://www.winsteps.com/winman/table31_1.htm) and [this video on how to do a deeper dive in Winsteps](https://www.youtube.com/watch?v=YpM92j0Vq9s).

## References

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

DeMars, C. (2010). *Item response theory*. Oxford University Press.

Linacre, J. M. (2021). *A user’s guide to Winsteps, Ministep: Rasch-Model computer programs* (Program Manual 4.8.0). <https://www.winsteps.com/manuals.htm>

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. University of Chicago Press.

*This handout was prepared under Contract ED-IES-17-C-0005 by Regional Educational Laboratory Central, administered by Marzano Research. The content does not necessarily reflect the views or policies of IES or the U.S. Department of Education, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.*