Supplementary Document

A Guide To Understanding The Rasch Model

Two views of measurement will be addressed: Classical Test Theory (CTT) and Item response Theory (IRT)

Introduction:

Stevens stated in 1946 that "measurement…. is defined as the assignment of numerals to objects or events according to rules. The fact that numerals can be assigned under different rules leads to different kinds of scales and different kinds of measurement" [1]. In this seminal paper he described the 4 known types of data that form the fundamental parts of measurements. Nominal and Ordinal data are considered descriptive data, have no numerical value, and are considered the lowest data forms in terms of measurement precision. On the other hand, interval and ratio data are quantitative numerical data, and statistical analyses using interval or ratio data are most meaningful since the numerical meaning of the numbers is maintained throughout their measurement range. Unfortunately, most of the available measurement tools in neuromuscular disorders have been developed based on Classical Test Theory (CTT), in essence being ordinal based, generally creating multi-item total scores or sub-scores which complicate their applicability, since they are ordinal measures [2]. CTT is based on the assumption that a given score reflects an underlying phenomenon or construct of interest. This observed score [OS] (usually a sum of different items) is expressed with the following formula: OS= True Score + Error of Measurement [3].

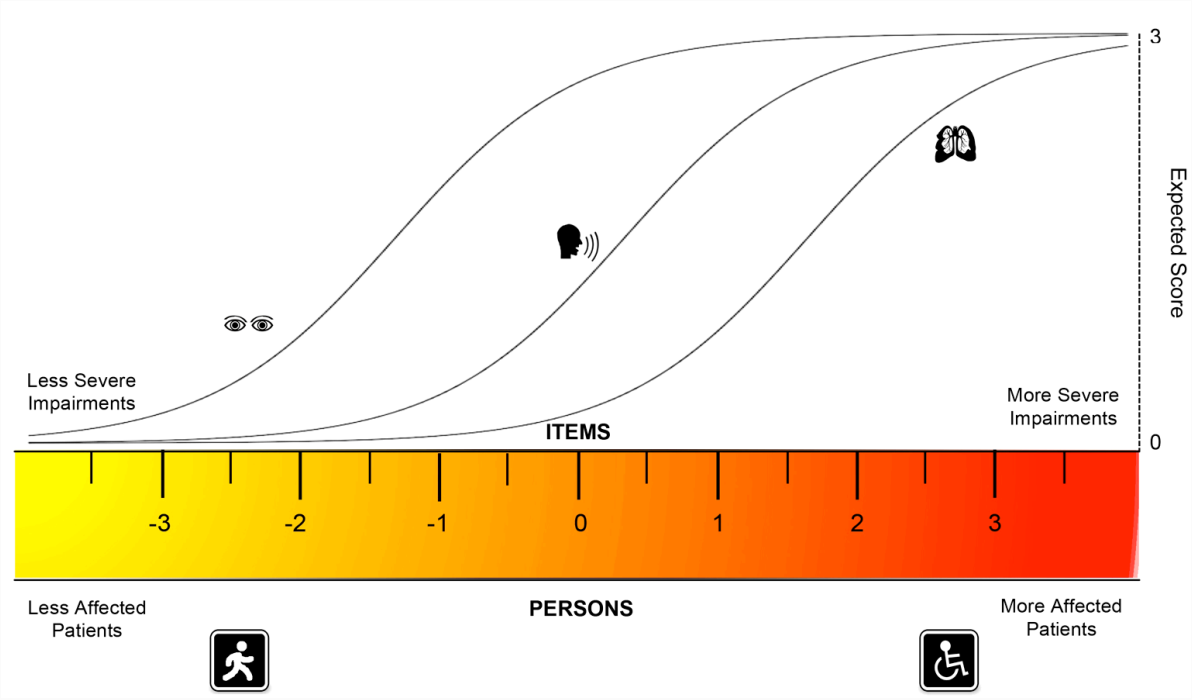
Item Response Theory (IRT) is a probabilistic approach to measurement, where the response patterns to individual items on a given measure are used to calculate a sample-independent metric

of the construct of interest. Through the probabilistic approach, collected data in patients (generally using ordinal based measures) can be transformed to a higher level of measurement precision, in essence by creating interval measures.

The Rasch Model [4,5]**:**

The Rasch model is an especial case within the IRT models, since it is aimed at developing tools that are at the interval level, where one unit is equivalent across the scale. This model was developed by Georg Rasch and it is based on aligning the persons and the items along the same metric (Figure 1). In this example, we will be using disease severity in Myasthenia Gravis as the construct of interest. By definition, the metric is in logits, with a mean of 0, theoretically extending both ways to ± infinity, although commonly ranges are much narrower.

Figure 1. Example of a Rasch-based Tool of Impairments in Myasthenia Gravis.



In this case, we see that persons with more disease severity (graphed below the ruler) are at the right side of the scale and people with less disease severity are towards the left. The items measuring impairments are located above the ruler are and these are ordered according to their "difficulty", in this case easier items are reflecting less severe impairments.

The term “item difficulty” simply reflects how many people endorsed the item. In our example, we expect that more respondents endorse items reflecting less impairment, and we will call them less difficult items. In figure 1, we theoretically organized ocular items as being lower in difficulty than respiratory items.

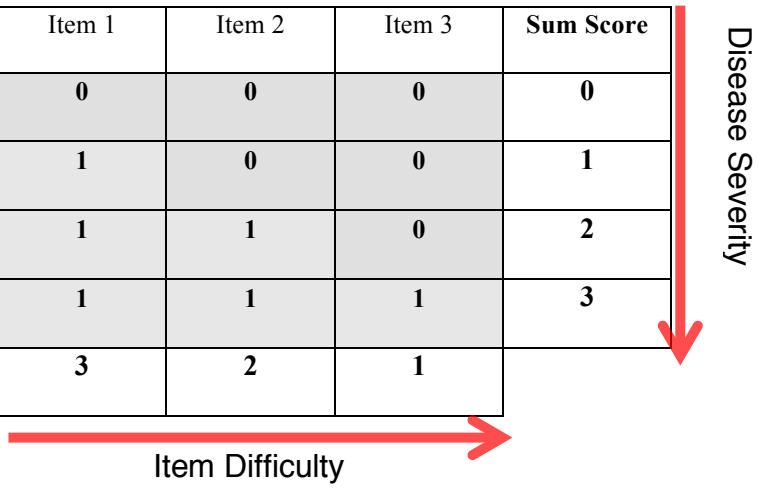
In this model, the probability of endorsing or receiving a higher score on a given item depends **ONLY** on the person’s location on the metric (e.g. disease severity) and on the difficulty of the item. All items, having different degrees of difficulty to accomplish, are expected to have the same ability to differentiate between respondents so that they can be summed up, and this is visually represented by the parallel item characteristic curves between the items in Figure 1.

As the persons increase in severity of illness, they are more likely to have higher score (reflecting more disease severity) in subsequently “more difficult” items or items reflecting more impairments. On the contrary, someone with low disease severity would probably have lower scores in the more difficult items, reflecting less disease severity.

By analyzing the response patterns of the respondents, we can organize the persons and the items on the logit ruler. Table 1 shows an example of a theoretical response pattern of 3 items with a yes/no (1/0) options, in patients with different levels of disease severity, and this explains how the items and the persons are organized along the ruler.

Table 1. Theoretical Pattern of Responses for Tool with 3 Items and 4 respondents.

	Item 1	Item 2	Item 3	**Sum Score**
Person 1	**0**	**0**	**0**	**0**
Person 2	**1**	**0**	**0**	**1**
Person 3	**1**	**1**	**0**	**2**
Person 4	**1**	**1**	**1**	**3**
Persons Endorsing the Item	**3**	**2**	**1**	



The vertical arrow reflects the persons with increasing amounts of disease severity and higher total scores. Person 1 has a sum score of 0 and is the less affected person, whereas Person 4 has a sum score of 3 and is the most affected. The bottom row shows the number of persons endorsing each item. In this case, most respondents endorse Item 1, so this is the "easiest" item, reflecting less impairment. Item 3 is the one reflecting more impairment or "more difficult".

Rasch Model Requirements:

The Rasch model has several requirements (read: "check-points"), which have to be fulfilled to be use a tool as an interval-level scale [6,7]. These are described below:

Invariance: This refers to the *item hierarchy*, which should remain constant across the latent trait. This is assessed with model fit statistics described below. Invariance is also expected regarding the *sample*, where response patterns between different populations (i.e ocular/generalized; male/female) are expected to be the same. When patterns are different, it is called Differential Item functioning (DIF) [6,8] .

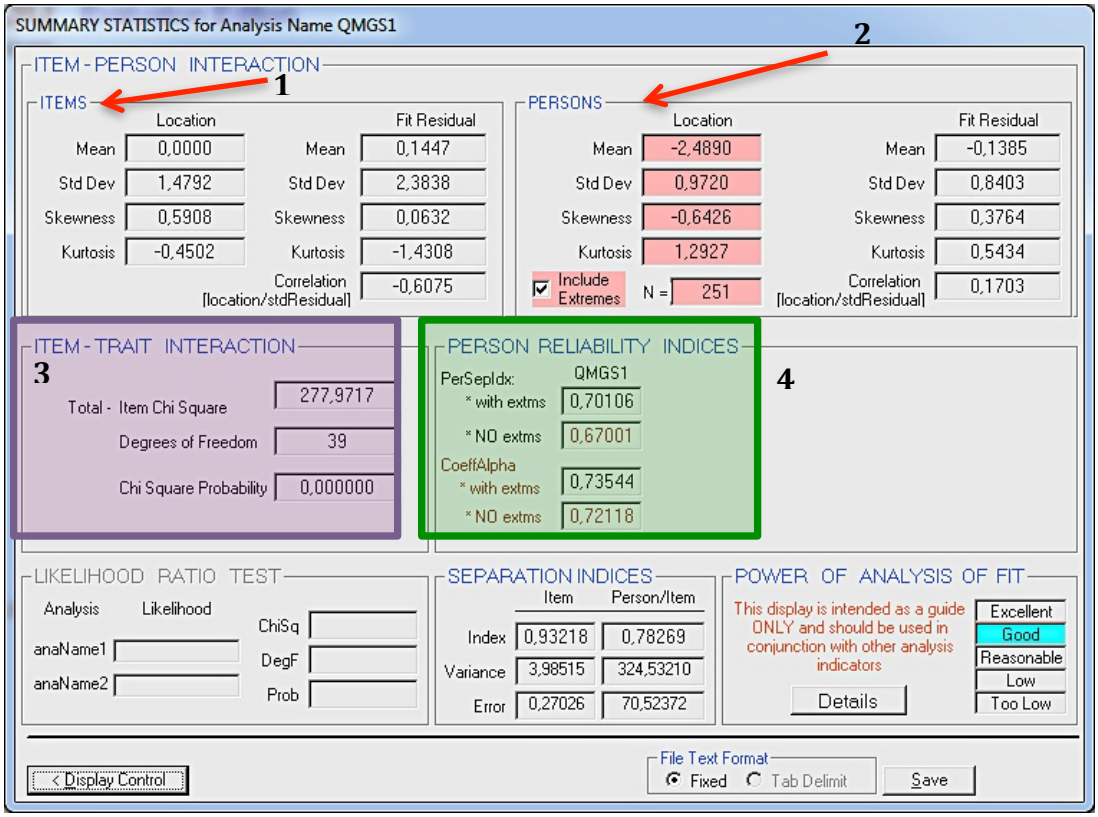
Unidimensionality: A *unique* underlying trait (e.g. disease severity) is being measured.

The assumptions described above are checked through several steps and statistics. We will go step-by-step using the Quantitative Myasthenia Gravis Scale (QMGS) [9] and the Myasthenia Gravis Composite (MGC) [10]as examples, and using the Rasch Unidimensional Measurement Model (RUMM 2030) [11]software.

Summary Statistics:

They are related to the persons, the items and the interaction between the item and the latent trait. Figure 2 is a screen-shot of the global statistic window for the MQGS analysis.

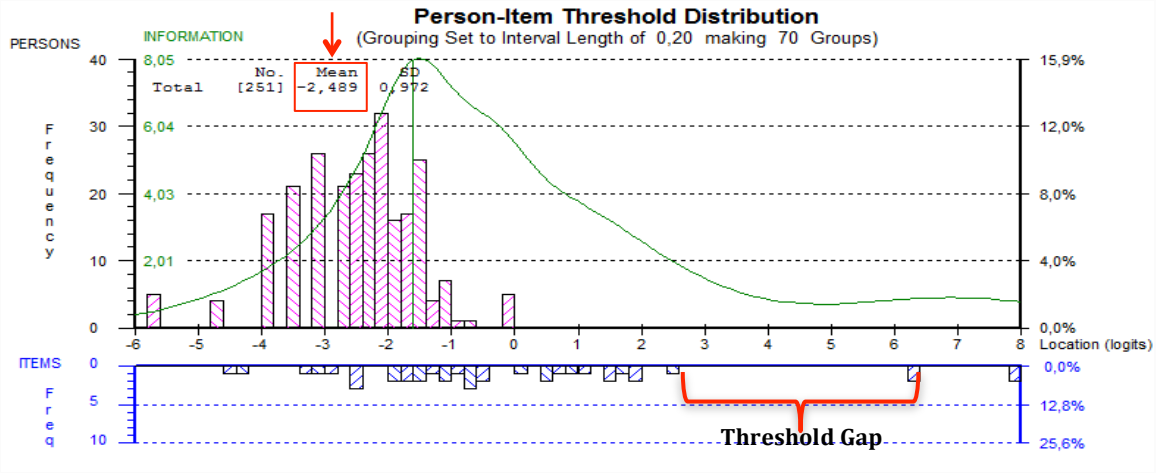
Figure 2. Screen-shot of Summary Statistics



1. Items: Mean location and SD. By default, items are fit to a mean of 0.
2. Persons: Mean location and SD. The expected mean location is around 0 for a well-targeted measure.
3. Item-Trait Interaction: This represents the overall fit to the model.
4. These are the reliability and discrimination indices.

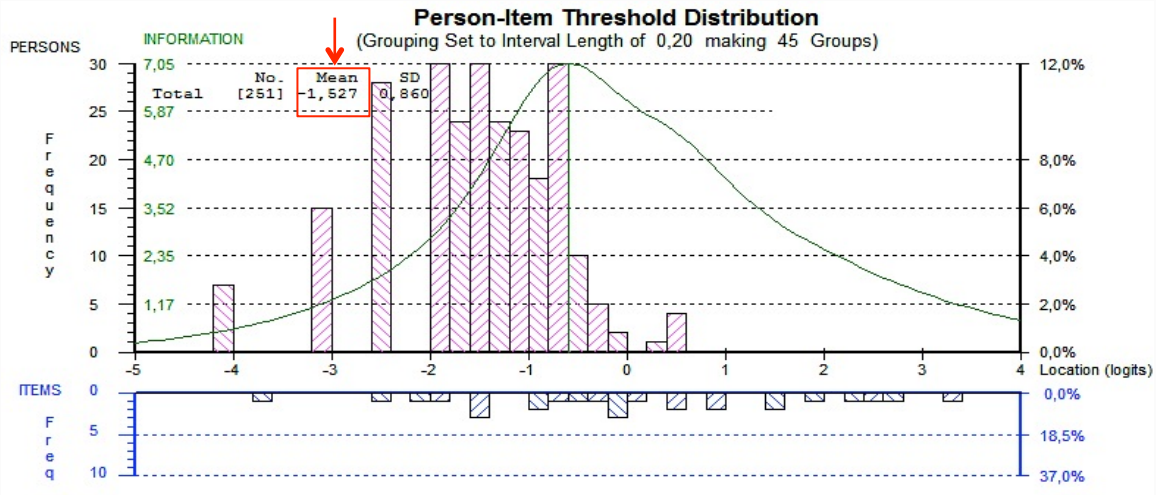
Person-Threshold Maps: The persons' locations can be plotted against the item thresholds, to visually analyze if the tool targeting is adequate. Ideally, the mean location should be close to 0, and there should not be gaps between thresholds. Figure 3 shows a poorly targeted measure, whereas Figure 4 depicts a better-targeted tool.

Figure 3. Person-Thresholds Map for the QMGS



The upper part (magenta) shows the persons distributions and below on blue, the item thresholds distributions. Ideally, the thresholds should form a continuum with few or no gaps. In this case, there are wide gaps towards the right end of the spectrum (more difficulty) indicating poor targeting in that range.
The green line is the information of the tool, indicating the locations where the tool is more precise. The test information is independent of the frequency of the persons' distributions.

Figure 4. Person-Thresholds Map for a Better Targeted Tool



In this example, there are fewer gaps in between the item thresholds, and the persons' mean location is closer to 0 than in Figure 3.

Item-Trait Statistic: This refers to the item hierarchy invariance previously described, and consists of a Chi-squared statistic, where significant p-values indicate item hierarchy variance and hence poor model fit.

Person Reliability Indices: The Person Separation Index (PSI) is a measure of discrimination (how well can the measure distinguish between at least 2 groups of patients). Cronbach's alpha, is a measure of internal consistency, and reflects correlations between items. For both, minimal values are 0.7, with ideal values > 0.8 for groups and >0.9 for individual use[12].

Individual Item Statistics:

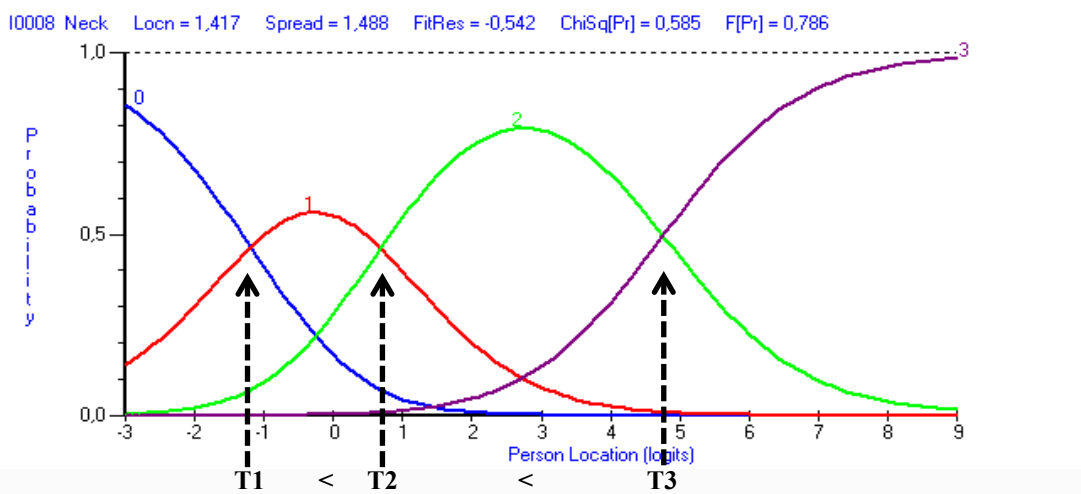
For each item, the residuals between the expected and observed responses are calculated. Residuals ≤ -2.5 and ≥ 2.5 are considered markers of poor fit, indicating that the item responses are too predictable (e.g caused by redundancy) or not predictable at all (not explaining the underlying construct). Chi-squared and ANOVA statistics can also be used, where significant p-values (after Bonferroni correction) are markers of deviation and therefore of poor fit [6,13]. Within the item statistics, we also obtain the location for each item.

Threshold Analysis [6,14]**:**

In the case of polytomous items, like in the MGC and QMGS where each item has an ordered Likert-type response, the probability of achieving higher scores depends on the person's location. We expect that with lower disease severity, the score on an item will be lower, and the expected score will increase as the underlying disease severity increases. Thresholds are the locations where the probability of endorsing adjacent response options (e.g. 2 or 3) (intersections of two adjacent response options) is equal. Logically, these thresholds should be ordered, with higher

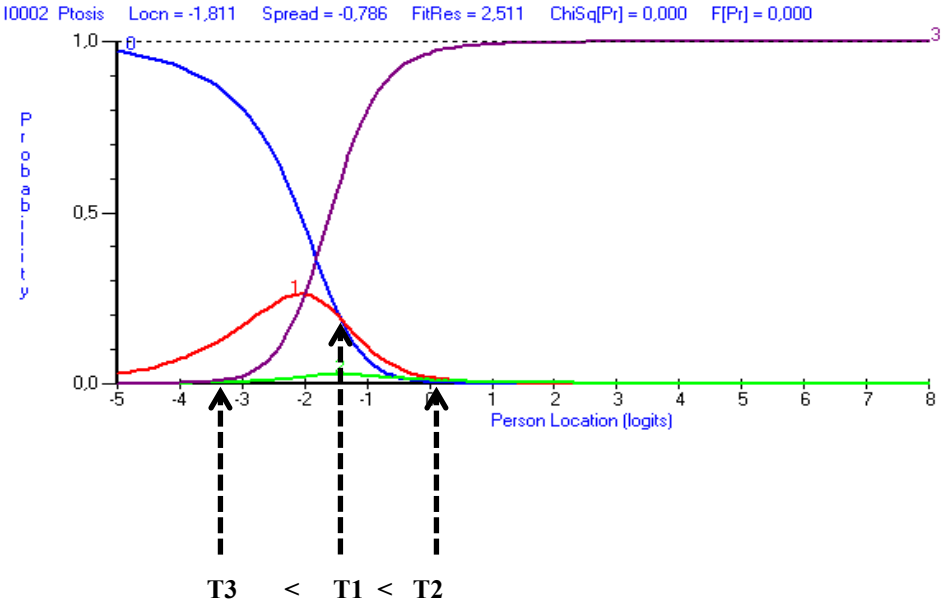
values as disease severity increases (Figure 6). Disordered thresholds usually occur when there are too many response options, or where there is confusing wording of the response options, making it hard for the respondents to select between adjacent options. Figure 7 shows an example of disordered thresholds

Figure 6. Example of Ordered Thresholds.



As disease severity increases, the probability of achieving higher scores does too. T1< T2< T3, indicating that the thresholds are properly ordered, and thus, the response pattern of this item follows the expectations of the model.
T1= Threshold between scores 0-1 T2= Threshold between scores 1-2 T3= Threshold between scores 2-3

Figure 7. Example of Disordered Thresholds



In this case, T3< T1< T2, this means that the probabilities of scoring 0 or 1 (T1) become equal with HIGER levels of disease severity than when scoring 2 and 3 (T3), which is not logical.

T1= Threshold between scores 0-1 T2= Threshold between scores 1-2 T3= Threshold between scores 2-3

Local independence: This refers to the assumption that the response to one item should not depend on the other. When this assumption is violated, we are in the presence of local dependence (LD), which can inflate the reliability coefficients or provide changes of which the magnitude does not necessarily reflect reality. This occurs when there are redundant items, or items referring to a previous question. To assess this, we analyze the correlation matrix of the residuals for the items. A correlation ≥0.2 is considered a marker of LD [15]. Table 2 shows the correlation matrix for the MQGS items.

Table 2. Residual Correlation Matrix of the QMGS

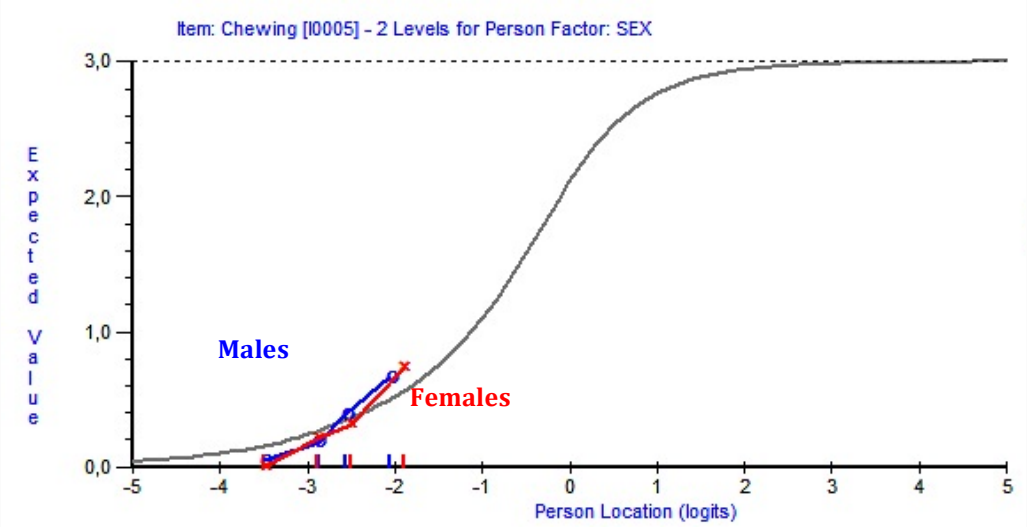
Item	Diplopia	Ptosis	Facial	Swallow	Speech	R.Arm	L.Arm	FVC	R.Grip	L.Grip	Head	R.Leg
Diplopia												
Ptosis	0.08											
Facial	-0.03	-0.04										
Swallow	-0.06	-0.03	-0.12									
Speech	-0.05	-0.02	0.00	0.06								
R.Arm	-0.23	-0.41	-0.06	0.02	-0.05							
L.Arm	-0.23	-0.41	-0.12	0.03	-0.05	**0.92**						
FVC	-0.19	-0.07	-0.07	-0.10	-0.09	-0.32	-0.35					
R. Grip	-0.13	-0.06	-0.18	0.00	0.01	-0.09	-0.11	0.03				
L. Grip	-0.10	-0.11	-0.14	0.04	-0.06	-0.11	-0.10	-0.02	**0.53**			
Head	-0.22	-0.36	-0.03	-0.01	-0.02	0.08	0.10	-0.21	-0.14	-0.11		
R.Leg	-0.27	-0.42	-0.09	0.02	-0.11	**0.31**	**0.35**	-0.25	-0.27	-0.24	**0.39**	
L. Leg	-0.33	-0.40	-0.06	-0.01	-0.11	**0.33**	**0.37**	-0.27	-0.24	-0.19	**0.40**	**0.77**

Bolded values are residual correlations ≥ 0.2, indicating local dependence. The correlations between the Right and Left arm and the Right and Left leg are particularly high, indicating redundancy (highlighted).

Differential Item Functioning:

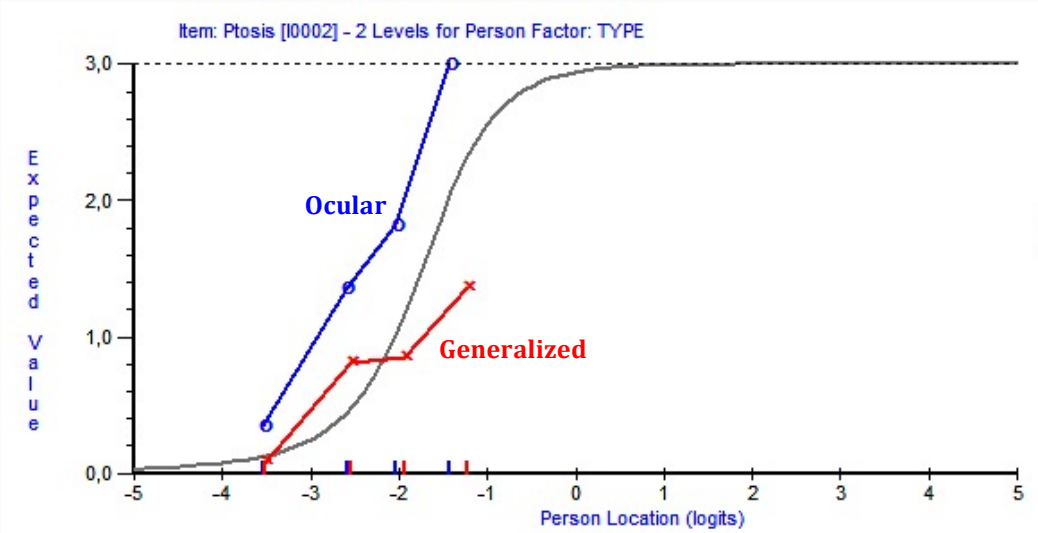
DIF is found when two populations, such as males and females having the same degree of ability (read: disease severity), respond differently to an item. This means that at the same level of the underlying trait (e.g disease severity), the probability of endorsing an item differs between groups, violating the invariance assumption. One way to deal with is to delete the item, or to split the item so its scores will be different for the 2 sub-groups. The latter might be problematic if the tool is aimed at assessing the subpopulations as a whole and in general items without DIF are preferred. DIF is analyzed visually, through item characteristic curves (ICCs) for the whole sample as well as for the subpopulations, and statistically by ANOVA among other tests [6,8]. Figure 8 shows an example of No DIF whereas Figure 9 depicts an item with DIF.

Figure 8. Example of Item with No DIF



In Figure 8, the red and blue lines represent females and males, respectively. Both groups have very similar and superimposed Item Characteristic Curves (ICCs). This means that patients respond to this item (MGC chewing) in the same manner in both groups, indicating sample independence for this item. ANOVA p value was not significant.

Figure 9. Example of Item with DIF



In figure 9, the two groups are separated, indicating that the response patterns differ. The red line, representing patients with generalized MG, is shifted towards the right compared to the blue line (pure ocular MG). This indicates that with the same underlying levels of disease severity, ocular patients tend to score higher in this item (QMGS ptosis). ANOVA p-value was significant after Bonferroni correction.

Summary:

The strict assumptions of the Rasch model are frequently hard to be fulfilled by tools built by CTT methods, and usually Rasch calibration is done using large item banks. The Rasch method unravels more scale's related deficits otherwise not seen through CTT approach or IRT methods. Rasch modeling provides valuable information on the items, even when a Rasch-built measure at the interval level scale is not possible, for instance because a construct is by definition multidimensional.

In summary, the Rasch model is a strong model based on logical assumptions on response patterns based on the underlying ability of the respondent and the difficulty of the presented items. Rasch-built measures have the properties of invariance, where a unit of measurement remains constant across the scale. Rasch techniques are also helpful to improve the quality of items being recruited ordinally.

REFERENCES

1. Stevens SS (1946) On the Theory of Scales of Measurement. Science 103: 677–680. doi:10.1126/science.103.2684.677.
2. Merbitz C, Morris J, Grip JC (1989) Ordinal scales and foundations of misinference. Arch Phys Med Rehabil 70: 308–312.
3. Wilson M (2005) Constructing Measures. Routledge Academic.
4. Rasch G (1966) An item analysis which takes individual differences into account. Br J Math Stat Psychol 19: 49–57.
5. Rasch G (1981) Probabilistic Models for Some Intelligence and Attainment Tests. Univ of Chicago Pr (Tx). 199 p.
6. Pallant JF, Tennant A (2010) An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS). Br J Clin Psychol 46: 1–18. doi:10.1348/014466506X96931.
7. van Nes SI, Vanhoutte EK, van Doorn PA, Hermans M, Bakkers M, et al. (2011) Rasch-built Overall Disability Scale (R-ODS) for immune-mediated peripheral neuropathies. Supplementary material. Neurology 76: 337–345. doi:10.1212/WNL.0b013e318208824b.
8. Holland PW, Wainer H (1993) Differential item functioning.
9. Barohn RJ, McIntire D, Herbelin L, Wolfe GI, Nations S, et al. (1998) Reliability testing of the quantitative myasthenia gravis score. Ann N Y Acad Sci 841: 769–772.
10. Burns TM, Conaway MR, Cutter GR, Sanders DB, Muscle Study Group (2008) Construction of an efficient evaluative instrument for myasthenia gravis: the MG composite. Muscle Nerve 38: 1553–1562. doi:10.1002/mus.21185.
11. Andrich D, Sheridan B, Luo G (2010) *Rasch models for measurement: RUMM2030*. RUMM Laboratory.
12. Nunnally JC Jr (1994) Psychometric Theory. McGraw Hill.734 p.
13. Tennant A, Conaghan PG (2007) The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? Arthritis Rheum 57: 1358–1362. doi:10.1002/art.23108.
14. Shaw F, Wright B, Linacre JM (1992) Disordered Steps? Rasch Measurement Transactions 6: 225.
15. Lundgren Nilsson Å, Tennant A (2011) Past and present issues in Rasch analysis: the functional independence measure (FIM™) revisited. J Rehabil Med 43: 884–891. doi:10.2340/16501977-0871.