



Technical Paper

The Informed Design and Validation of Quality of Life Questionnaires and Rating Scales Using Rasch analysis

**Frances Denny
Principal Statistician
Exploristics Ltd**

Introduction

Quality of life (QoL) is a term used to describe the physical, mental and social well being of an individual [1]. This concept was first introduced by Oncologists who realised that clinical measures alone did not provide adequate information on patients well being while coping with a disease or undertaking treatment. QoL is now a recognised, valid measure of well-being and widely used in other medical fields. Indeed, most clinical trials include QoL alongside other clinical measures of response.

Typical QoL measurements are derived from responses to a questionnaire (or instrument). These instruments comprise a list of questions (or items) which either assess general aspects of health or are disease specific. They are completed either by the individual or by an interviewer. Patients are given a choice of responses in the form of either dichotomous (e.g. yes or no) or ordinal (e.g. very easy, easy, hard, very hard) and in some cases a mixture of various response scales.

There are numerous QoL instruments in use today. Generic instruments, such as the Short Form 36 (SF-36), assess the patient's ability to complete various activities in everyday life [2]. Disease specific instruments, such as the QLQ-C30 developed by the European Organisation for Research and Treatment of Cancer (EORTC) [3], evaluates patients ability to complete everyday tasks and the side effects that they might experience while coping with their disease and subsequent treatment.

The nature of response scales (i.e. ordinal rather than continuous) can present a problem for the analysis. Within the medical arena, Rasch analysis has become a frequently used tool for analysing dichotomous and ordinal data generated through responses to QoL instruments. In basic terms, Rasch analysis converts dichotomous and ordinal observations into a linear, continuous scale. In other words, it translates qualitative analysis into a quantitative approach [4]. This scaling allows the quantitative evaluation of many attributes of the QoL measure such as the multi-dimensional interaction between people, probes, prompts, raters, test items, tasks, etc [4].

Rasch analysis has two main applications. It can be applied to data collected in a pilot study to create and validate a QoL instrument. In this instance it is useful for item banking, and as an indication of whether the instrument sufficiently captures the QoL of a particular set of patients. Rasch analysis is more frequently applied retrospectively to data after a clinical trial has been completed. In this case, the various properties of a questionnaire such as content and construct validity, item redundancy, reliability and a domain structure can be explored, enabling a thorough understanding of patient responses.

The following section describes how some of the additional outputs from Rasch analysis can provide information beyond that of a simple analysis of the QoL data during the creation and validation of a new instrument or to understand and quantify responses following a clinical study.

1. Construct and content validity

The basic principal of Rasch analysis is to find the log-odds (logit) ratio of each person choosing one response category over the previous for each item using maximum likelihood estimation. As a result, an estimate for each of the model parameters that is person, item and response category in the dataset is obtained. Therefore if there are one hundred people answering ten items with a three point response scale there will be one hundred person estimates, ten item estimates and three category estimates. The person estimates are usually ranked in order where the person with the greatest ability is placed at the top and the person with the weakest ability at the bottom. A similar hierarchy is also created for the item estimates where the most difficult item is placed at the top decreasing to the easiest item at the bottom. As the estimates are linear, they can also be taken to represent the value of the person/item for comparison with other similar questionnaires so as to determine the construct validity.

The example used throughout this paper is where seven hundred and sixty-six study participants were administered an instrument comprising of twenty items with a five point response scale [5]. As the same response scale is employed across all items, the Rasch Rating Scale model is the most appropriate to model the data in our example. Table 1 describes the item hierarchy of the instrument. Item 14 was considered by Rasch analysis as the hardest while item 9 was determined to be the easiest.

Table 1. Item measures and fit statistics

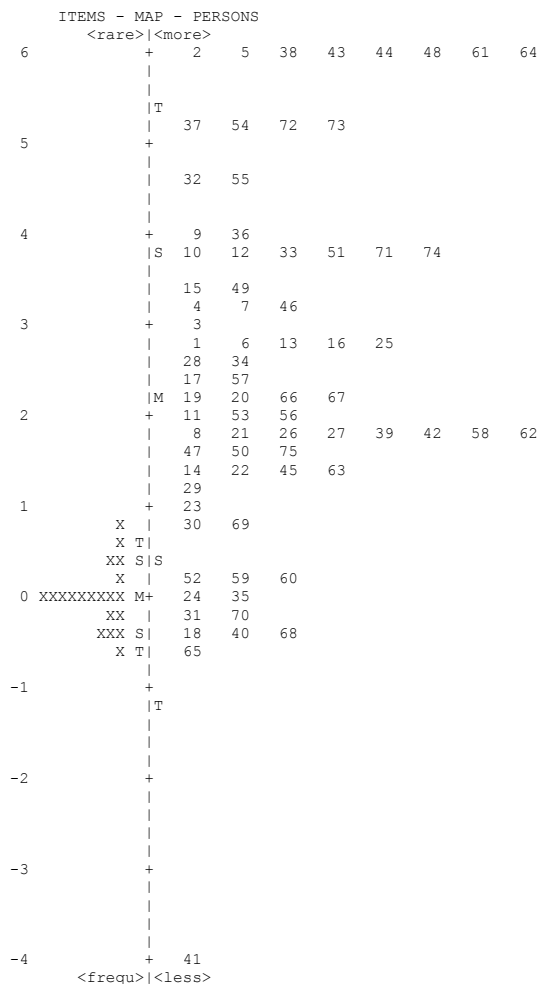
Item	Item Number	Measure	S.E	Infit MnSq	Outfit MnSq
I felt terrified	14	0.78	0.18	0.86	0.91
I had a racing or pounding heart	6*	0.68	0.18	1.16	1.42
I felt like I needed help for my anxiety	3	0.36	0.18	0.77	0.61
I found it hard to focus on anything other than my anxiety	15*	0.32	0.19	1.53	1.45
I avoided public places or activities	11	0.15	0.20	1.15	1.19
I felt upset	5	0.07	0.18	0.66	0.47
Many situations made me worry	19	0.05	0.18	1.08	1.17
I felt something awful would happen	13	0.04	0.18	1.12	1.03
I felt fidgety	12	0.00	0.19	1.19	0.91
I had difficulty sleeping	20	-0.02	0.19	0.67	0.60
I felt frightened	1	-0.05	0.19	0.77	0.62
I felt anxious	2	-0.07	0.19	0.67	0.47
I had sudden feelings of panic	8	-0.08	0.20	1.10	1.00
I was concerned about my mental health	4*	-0.09	0.19	1.18	1.64
I had trouble paying attention	10	-0.11	0.19	1.21	1.09
I felt indecisive	18	-0.14	0.20	0.50	0.49
I was anxious if my normal routine was disturbed	7	-0.38	0.20	1.32	1.37
I felt nervous	17*	-0.41	0.20	1.49	2.20
I had twitching or trembling muscles	16	-0.47	0.20	0.85	0.68
I was easily startled	9	-0.63	0.21	0.63	0.72

As both the person and item estimates are on the same scale which is an interval scale, a person-item map (Figure 1) can be used to gauge person ability against item difficulty to determine content validity [6]. The scale is taken as the axis in the middle of the map with persons and items on opposite sides and located according to their ability and difficulty estimates [6]. Item-person maps also exist, whereby the main focus is on the item difficulty

which is useful in showing whether items target the population properly or whether they are too easy/hard [6] and can be used to determine construct validity.

Figure 1 is a person-item map of our example population. Person abilities are represented by Patient ID numbers on the right hand side of the axis along the logit scale while item difficulties are represented by the symbol “X” on the left hand side. When the items on the instrument are representative of the disease condition, the person abilities and item difficulties should be normally distributed. As illustrated in Figure 1 the item difficulties are distributed fairly normally around the mean measure (M). However the person abilities are positively skewed. The person abilities are probably skewed because person responses may have been more positive towards items than expected thus indicating a ceiling effect. This suggests that the instrument might not be able to detect changes in patients attitudes or able to discriminate between different population subgroups.

Figure 1. Person-item map



X	Item difficulties
No.s	<ul style="list-style-type: none"> Person abilities (RHS) Values on the logit scale (LHS)
M	Location of the mean measure
S	One sample standard deviation from mean
T	Two sample standard deviations from mean

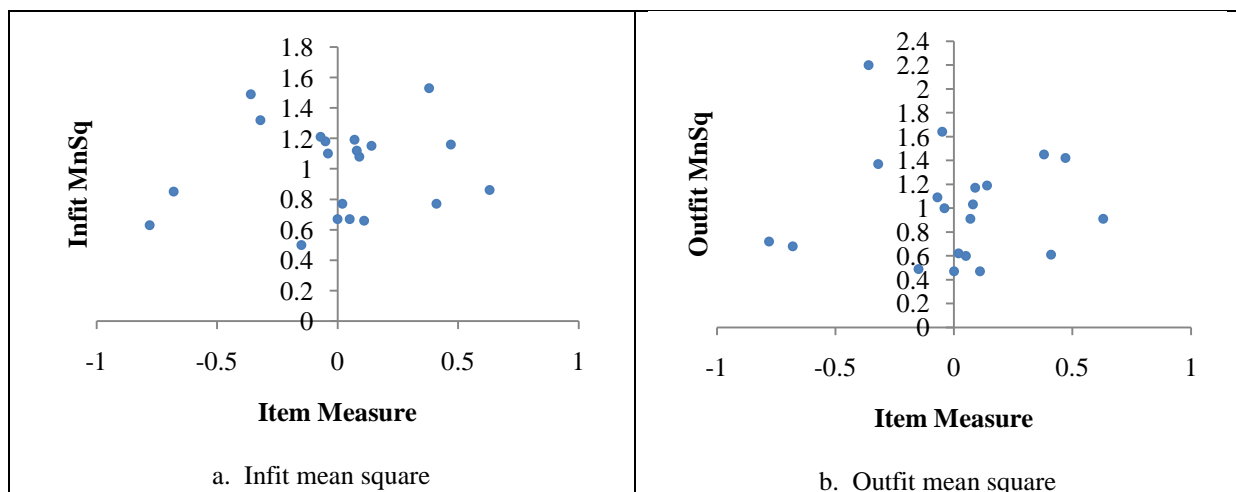
2. Item redundancy

An observed and expected score is derived for each parameter. The difference between the observed and expected measure for each parameter is taken to be the model residual for that parameter. Rasch analysis also generates two goodness-of-fit statistics for each person, item

and response category estimate. These are referred to as the mean-square infit and outfit statistics. Outfit statistics are the summation of squared standardised model residuals. Infit statistics are the summation of squared model residuals which are first weighted by the variance. High outfit statistics are due to persons giving an unexpected response to a given item. Lower than optimal fit statistics occur when a person gives either one of the extreme responses to all items. Guidelines of what are considered to be acceptable infit and outfit statistics are given in Bond and Fox [6]. For example, the optimal lower and upper limits are 0.7 and 1.4 respectively for a four point response scale.

In the example in this paper there are four items with an outfit mean square above the 1.4 limit (marked by an asterisk in Table 1). This suggests item redundancy. To test this theory, the items should be removed and the analysis repeated. A visual representation of these fit statistics can be seen through infit and outfit plots (Figure 2). Those fit statistics outside the optimal levels can be clearly identified as outliers.

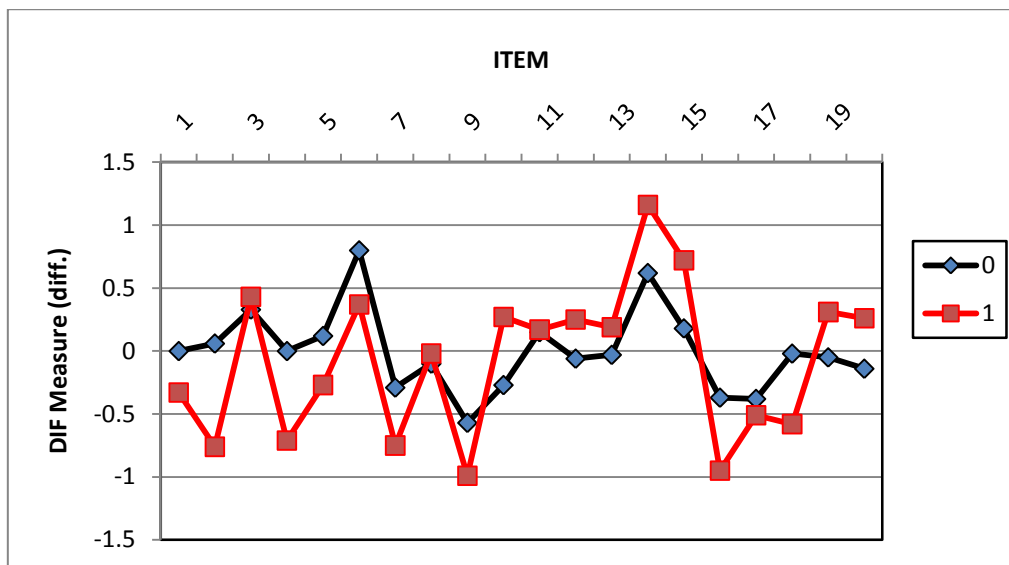
Figure 2. Infit and outfit plots



If subgroups exist within a sample population, differential item functioning (DIF) can be employed to assess each subgroup to see how one group scores against the other. This is another useful tool in identifying any item redundancy. If differences exist between the groups this could indicate that a group may be performing better/worse than another or if an item is considered harder/easier by one group in comparison to another. Differential person functioning (DPT) can also be performed.

A DIF plot for our example (Figure 3) clearly demonstrates that the males (represented in black) and females (represented in red) rate the difficulty of the items differently. Those items which were identified as redundant by the fit statistics (items 4, 6, 15 and 17) are also rated differently by the subgroups.

Figure 3. Differential item functioning



3. Response scale

Several statistics are produced for the response categories, the frequency of each response chosen, the expected and observed average response, an infit and outfit statistic and a category threshold estimate. When an appropriate response scale has been used the frequencies should increase monotonically as the person and item estimates increase and the observed measures should be close to the expected measures. In addition, infit and outfit statistics should lie within the optimal levels for the response scale used. Finally, the category threshold estimates, which indicate the point at which one category is likely to be chosen over another, should also advance monotonically with every response category. All this information combined will help determine if the response scale is optimal.

Table 2 shows that the response categories are monotonically increasing, the observed and expected measures are closely related and the fit statistics are generally between the optimal levels. However there is some disordering with the threshold estimates (marked with asterisk) therefore suggesting the response scale should possibly be reconsidered. One possible suggestion would be to combine response categories 2 and 3 together so that there are four response categories rather than the original five. The analysis should then be reapplied to determine any improvement.

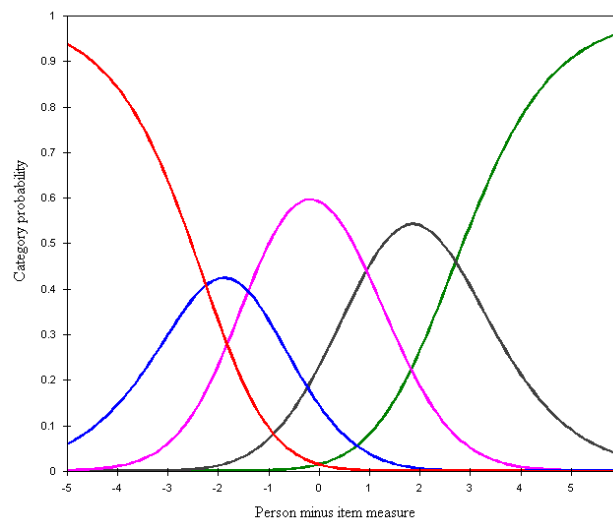
Table 2. Category summary statistics

Response Category	Observed Count	Observed Average	Expected Average	Infit Mean Square	Outfit Mean Square	Threshold Value
1	2	-2.15	-2.04	0.88	0.79	-
2	2	0.09	-0.11	1.29	1.11	-0.73
3	7	0.25	0.52	0.68	0.52	-1.02*
4	18	1.45	1.63	0.70	0.44	0.16
5	38	3.12	2.99	0.63	0.80	1.60

A category probability curve (CPC) is also generated for every item. A CPC illustrates the probability of a person with a given ability choosing each of the response categories when responding to an item with a given difficulty. The CPCs are very useful in visually determining if the most appropriate response scale is being employed.

Figure 4 illustrates the CPC of the hardest item on the instrument, item 14. Although there is even separation between response categories, and the extreme response categories behave in the normal manner, that is being the most probable and asymptotic to one, the middle response categories don't increase monotonically with the person-item trait. Response category four should be more probable than response category three which in turn should be more probable than response category two. However as Figure 4 shows, response category three is more likely to be chosen. Therefore we would conclude that the response scale should be redefined.

Figure 4. Category probability curve of item 14



4. Separation and reliability coefficients

The level of separation between person abilities across an analysis population is described by a separation index where a low value is considered to be better [7]. Similarly a reliability coefficient in Rasch theory is a measure of the spread of person ability within the population [7]. Most researchers use Cronbach's alpha to measure reliability but the coefficient put forward by Rasch theorists can be used as an alternative. The WINSTEPS user guide gives some idea of what a respectable reliability coefficient for a given response scale should be [8].

For our example, the separation index is 2.55, which suggests that there are approximately three statistically distinct levels of person ability. As this value is quite low, it also suggests good separation between the person estimates. The reliability coefficient is high at 0.87 indicating excellent instrument reliability. While this suggests a level of confidence in the instrument it is important to reinvestigate these properties when redundant items have been removed and response categories 2 and 3 are combined. Only then, can the instrument be determined as valid and reliable.

5. Domain structure

A factorial analysis, namely principal component analysis (PCA) is required to determine if the instrument is uni- or multidimensional and if it is performing as it should. PCA is performed on residuals as they are normally distributed, independent of each other and therefore uncorrelated. PCA of our example indicates that there are two domains in the questionnaire.

Table 3 describes the segregation of items in the instrument into the two domains. Those items in domain 1 are considered to be stronger indications of anxiety than those in domain 2.

Table 3. Factor loadings from Principal Component Analysis

Item Number	Rasch Factor Loadings (sorted by factor loading)	
	Domain 1	Domain 2
1	0.47	
2	0.40	
3	0.48	
4	0.02	
5	0.10	
6	0.52	
7	0.25	
8		-0.49
9	0.55	
10		-0.23
11		-0.45
12		-0.31
13		-0.25
14	0.23	
15		-0.51
16	0.23	
17		-0.33
18	0.44	
19		-0.35
20	0.16	

Summary

Rasch analysis is a useful tool for analysing dichotomous and ordinal data particularly that which arises from patient responses to QoL instruments. This paper has presented how the outputs produced by the analysis can be used to assess the validity and reliability of an instrument while also identifying item redundancy and a domain structure. It also demonstrated that Rasch analysis can be used to identify differences between population subgroups. By comparison with more traditional techniques employed to identify an instrument's properties, Rasch analysis is considered superior as it can analyse response data without including clinical measures. Rasch analysis can be applied to most QoL questionnaires as well as many ratings scales. Exploristics offers the opportunity to analyse ordinal data using Rasch analysis as part of the development or further refinement of an instrument.

References

- [1] What is Quality of Life? Fallowfield L. What is...? series. Hayward Medical Communications, 2009.
- [2] SF-36. Available: www.sf-36.org, 2009.
- [3] European Organisation for Research and Treatment of Cancer. Available: www.eortc.be/home/qol/QLGhistory.htm, 2009.
- [4] WINSTEPS software. Available, www.winsteps.com.
- [5] The R Project for Statistical Computing. Available, www.r-project.org.
- [6] Bond TG, Fox CM. Applying the Rasch model-Fundamental measurement in the human science. Lawrence Erlbaum Associates, Inc: Mahwah, New Jersey, 2001; p235-260.
- [7] Smith EV, Smith RM (Eds). Introduction to Rasch Measurement. JAM Press: Maple Grove, Minnesota, 2004.
- [8] Linacre JM. A User's Guide to WINSTEPS: Rasch-model Computer Programs. Available: www.winsteps.com, 1991-2005.