A critique of Rasch analysis for the development of patient-centred outcomes: an illustrative example using the Dyspnoea-12

Journal:	*Journal of Advanced Nursing*
Manuscript ID:	JAN-2010-0213
Manuscript Type:	Manuscript/Short Report
Keywords:	

A critique of Rasch analysis for the development of health-related instruments: an illustrative example using the Dyspnoea-12

Abstract

Background: The development of questionnaires has traditionally involved the application of classical test theory (CTT). More recently Rasch analysis has gained momentum as a robust application of 'modern' psychometric testing for the development of new instruments and the refinement of existing ones.

Aim: This paper is a report of the application of Rasch analysis to the development and refinement of the Dyspnoea-12 questionnaire; an instrument that measures breathlessness severity using descriptor items. The aim is to provide a critique and working example of Rasch analysis techniques.

Method: 358 patients with a cardiopulmonary disease responded to an initial list of 81 items. Hierarchical modeling reduced the list to 34 items. Subsequent Rasch analysis was used to informed decisions regarding further item removal and fit to the unidimensional model. This paper presents the application of Rasch analysis to these 34 items.

Results: 22 items failed to reach the requirements of the Rasch model and were removed.

Conclusion: This paper provides a working example of Rasch analysis. We have presented the steps involved in reducing and refining a large item-set by identifying those items which possessed the most reliable measurement properties. We have provided nurse researchers with an alternative to CTT when developing or refining questionnaires that measure patient-centred outcomes.

What is already known about this topic

- The majority of health-related questionnaires have been developed using classical test theory
- Rasch analysis provides a robust technique for the development of new health-related questionnaires or the refinement of others

What this paper adds

- This papers demonstrates that Rasch analysis is a viable option for questionnaire development and refinement.
- A detailed description of the processes involved with the application of Rasch methodology is provided
- This paper provides nurses with a method for critiquing the robustness of other questionnaires developed using Rasch analysis.

Implications for practice and/or policy

- It is vital that measures of disease severity are developed and refined using robust psychometric techniques; this paper should inform the future development and critique of health-related questionnaires.

Key words: psychometrics, Rasch analysis, outcomes, breathlessness

INTRODUCTION

Reliable and valid questionnaires for the measurement of patient reported outcomes (PROs) are an important aspect for research and clinical practice. PROs are latent constructs in the sense that they cannot be measured directly but only through measurable indicators, such as the patient's self-report of disease symptoms. An observer cannot estimate symptoms; they are subjective and experienced only by the patient.

The development of questionnaires has traditionally involved the application of classical test theory (CTT) (Nunnally, 1978). This approach involves the assessment of item-total correlations to identify redundant items and testing the questionnaire's dimensionality using factor analytic techniques (Watson and Thompson 2006). More recently in medical and rehabilitation fields, Rasch analysis (Rasch, 1960) has gained momentum as a robust application of 'modern' psychometric testing for the development of new instruments and the refinement of existing ones. In the nursing literature, Rasch analysis has received comparatively little attention (Watson & Thompson, 2006; Rattray & Jones, 2007). A search of all Journal of Advanced Nursing (JAN) papers published since 1990 to 8 February 2010 using the search term 'factor analysis' or 'Rasch' was conducted. The term 'factor analysis' identified 227 papers that referred to this technique, whereas two papers were identified using the search term 'Rasch'; only one of these reported the application of Rasch to the development of a scale (Gilworth, et al. 2007) and the other was a theoretical paper (van Alphen, et al., 1994).

We previously developed the Dyspnoea-12, an instrument that quantifies breathlessness severity using 12 descriptor items (Yorke et al. 2010) Research has identified that, like pain perception, breathlessness consists of a sensory-quality as well as an emotive response (Wilson & Jones, 1991; Yorke, 2008). The Dyspnoea-12 provides an overall score for breathlessness severity that captures these different aspects. It was developed using Rasch analysis (Rasch, 1960). This paper describes that application of Rasch to refine and reduce a set of items to form the Dyspnoea-12, a unidimensional scale. The aim is to provide a working example of Rasch techniques and in doing so, attempts to demystify the complexities of this psychometric approach.

Background

Despite the prevalence of CTT being utilised in scale development, it is important to note that CTT has a number of limitations, all of which could have implications upon the scales that have been derived under that methodology. Briefly, CTT is limited in that the scale scores derived are of an ordinal nature, but they are often treated as interval-level data. This means that CTT-based scales may be prone to differential sensitivity at the centre, relative to the extremes, of the score range (DeVellis, 2006). Another limitation is that the evaluations of scales are dependent upon the sample on which they have been tested against, and also, the measurement of people is dependent upon the set of items with which they have been measured (Hobart & Cano, 2009). This means that different samples with different variances will not yield equivalent data or data that can easily be compared across samples (DeVellis, 2006). Hobart and Cano (2009) identified that Rasch analysis represents a logical progression from CTT, as it attempts to improve the scientific quality of the theory underpinning rating scales.

The Rasch model was developed by the Danish mathematician Georg Rasch within the realm of educational psychology (Rasch, 1960). Its primary function is to test how well items within an instrument conform to a unidimensional model. In other words, it checks if the underlying construct being measured has a single dimension on which all of the questionnaire items rely. This is a key concept for instruments where a total summated score is calculated (Hagquist et al., 2009). Historically, in Rasch measurement the position that each item and person occupies on this dimension is termed its item 'difficulty' and person 'ability'. This is because much of the early work was carried out in the field of education using multiple choice exam questions. These terms are also applied, for example, when quantifying physical ability levels in the field of rehabilitation. In symptom measurement, such as breathlessness, 'severity' is a better term and will now be used throughout this paper.

There are two key features of the relationship between a person's symptom severity and that expressed by an item in a questionnaire. First, the observed response is dependent on the difference between patient severity and the severity of the item. Second, the model is probabilistic as uncertainty (a theory-based probability) surrounds the expected response, consistent with the real life situation (Tesio, et al, 2007). The Rasch model assumes that the probability that a person will affirm an item is a logistic function of the *difference* between the person's ability [θ] (in terms of the breathlessness severity level of the person) and the difficulty of the question [β] (again, in terms of the breathlessness severity level represented by the item), and only a function of that difference. For an explanation of the equation, broken down into its component parts, please see Figure 1.

Insert figure 1

The Rasch model tests that items measuring a lower severity are more likely to be endorsed by patients with a higher level of severity (as determined by the responses to all items combined). The converse is true when a person's severity is less than that for the item (Borsboom, 2005). This is a property of scales that is commonly termed Guttman scaling (Guttman, 1944). In Rasch analysis the response patterns obtained are tested against what is expected, so it is a probabilistic form of Guttman scaling (Pallant & Tennant 2007). The resulting severity estimates for items and respondents are reported in 'logits' – the log-odds of responding to an item (Figure 2).

Insert figure 2

It also offers an alternative where the limitations of CTT can be overcome; Rasch analysis can provide a transformation of an ordinal score into a linear, interval-level variable, given fit of data to Rasch model expectations (Tennant & Conaghan, 2007). This means that the problem of differential sensitivity can be overcome. Also, the Rasch model has the advantageous property of invariance, meaning that the item and person parameters can be estimated independently of each other (Andrich, 2004). To put this another way; the measurement of people is not dependent upon the sampling distribution of the set of items with which they have been measured, and the difficulty estimates of items on a scale are not dependent upon the distribution of the sample on which they have been derived (Hobart & Cano, 2009). For a more complete account of the development of modern test theory in the health sciences, along with its advantages over CTT, the reader is directed towards Hobart & Cano (2009).

Rasch analysis enables an examination of the contribution of items individually as they are added and removed from the item set. This enables the selection of items during questionnaire development phase that provide maximum measurement precision. If data fit model expectations then a fundamental assumption is that each item contributes reliably to the measurement of the single underlying construct. If an item set meets the criteria of invariance and items form a unidimensional scale, then a summated score for the concept being measured can be legitimately obtained.

The Study

Aims

To describe and critique the application of Rasch analysis to the development of the Dyspnoea-12.

Participants

Participants were recruited from out-patient clinics from three NHS Trusts. Participants completed an 81-item list during their clinic visit (n=275, 77%); the remainder completed them at home for return within two weeks. Their baseline characteristics are shown in Table 1. The study received ethical approval from the Local Research Ethics Committee and all participants provided written consent to participate.

Insert table 1

Dyspnoea-12

Details of the overall development of Dyspnoea-12 are described elsewhere (Yorke et al, 2010). Briefly, a pool of 81-items was arranged as a questionnaire that asked patients to respond to each one reflecting their current experience of being breathless.

Response options were ‘none’, ‘mild’, ‘moderate’ or ‘severe’. Items were removed if more than 50% of the sample responded ‘none’ or demonstrated bias associated with age. Thirty-four items survived this process and were further reduced and refined using Rasch analysis (Rasch, 1960). The remainder of this paper reflects the application of Rasch analysis to these 34 items which was reduced to a final 12-item set; Dyspnoea-12.

The Application of Rasch Analysis

In this study, analyses were performed using Rasch Unidimensional Measurement Model (RUMM2020) (Andrich et al., 2003). Whilst the steps taken to analyse data and remove/retain items are presented here in a logical step-by-step format, it is important to note that these applications involved an iterative approach. The aim in this instance was to identify the mis-fitting items and examine the effect of their removal on other items and the total item-set.

Class Intervals

In RUMM2020, patients are automatically placed into groups called class intervals (CI). Class intervals are defined by ordering all patients in terms of breathlessness severity (determined by the responses to all items combined) and then splitting them into groups of approximately equivalent size across the sample (Tennant & Conaghan, 2007). As demonstrated below, a number of Rasch fit statistics are applied at the CI level. In this study, the 358 patients were grouped into six CI and represented an acceptable dispersion of patient numbers (Table 2). In this sense, the CI’s can be seen to represent different, discrete, levels of seveiry.

Insert table 2

Ordering of response categories

The 34 items each had a polytymous response format ranging from 0 (none) to 3 (severe). Patients’ responses to these options need to follow a logical sequence. It would be expected that as patient breathlessness severity increases, it is more likely to score a 0, then a 1, then a 2, then a 3 for any particular item. In Rasch analysis terms this would be indicated by an ordered set of response thresholds for each item (Pallant et al., 2006). The term threshold refers to the point between two response categories (e.g. 1 'mild' and 2 'moderate') where the probability of scoring a 1 on the item or scoring a 2 is 50/50 (given that the person will only score a 1 or a 2); in other words, the point where either response is equally probable (Pallant & Tennant, 2007) (Figure 2). Item response patterns that do not follow a logical order are termed disordered, or reversed, thresholds. Disordered thresholds can occur when patients have difficulty consistently discriminating between response options when, for example, there are too many response options, or when the labelling of options is confusing. Correct ordering can often be achieved by combining adjacent response categories and rescoring the item (Pallant & Tennant, 2007).

Item thresholds can be assessed graphically using the item category probability curves (Figures 3 and 4). They may also be assessed by looking at the actual numerical threshold estimates. One of the 34 items showed a situation in which patients demonstrate an inconsistent transition between response options ('My breath does not go out all the way') (Figure 4); this item was subsequently removed due to lack of fit to the model.

For an instrument with polytomous items there are two parameterisations of the Rasch model that can be assessed using the RUMM programme: the Rating Scale Model or the Partial Credit Model. These two models differ slightly in their mathematics where the former expects the distances between thresholds to be equal across items (Tennant & Conaghan, 2007). This means that the metric distance between, for example, the thresholds separating categories 1 and 2 is the same across all items, and that the metric distance separating categories 2 and 3 is the same across all items. However, the distance between categories 1 and 2 does not have to be equivalent to the distance between categories 2 and 3. The Rating Scale Model provides a higher degree of specificity; however, it is not always possible to satisfy the assumptions of a Rating Scale Model, in which case the Partial Credit Model should be utilised. The likelihood-ratio test provides a test of which version of the model should be utilised by comparing the different parameterisations of the model and providing a Chi-square statistic and probability – if the outcome of the test is not significant ($p > 0.05$), then the Rating Scale Model should be adopted as it is a simpler model. The test for the 34 items used in this study (as is often the case when using real data) was $p < 0.01$; requiring the Partial Credit Model to be used.

Insert Figures 3 and 4

Tests of Individual Item Fit

Tests of individual item fit to the Rasch model reflect the differences between the observed responses and that expected by the model (i.e. expected responses given the level of breathlessness severity based on patients' responses to all items combined). This is an important feature of Rasch analysis because it tests the ability of individual items to reliably measure breathlessness at different severity levels (i.e. CI). These tests are presented for each item as a fit residual and as a Chi-Square probability statistic. A residual is a summation of individual item (or person) deviations from model expectations, which are then standardised to form a z-score. Those between ±2.5 are deemed to generally indicate adequate fit to the model (Pallant & Tennant, 2007). A high negative residual indicates an over-discriminating item, which is also a possible indicator of redundancy. Redundant items offer nothing to the information gained by other items, and removal should improve the fit of those remaining items. In some respects it is analogous with a high item-total correlation used in CTT (Pallant et al, 2006). High positive residuals indicate under-discriminating items, which suggest that these items are not contributing to measure the underlying trait in question. That is, such items do not have discriminant power; the item's responses do not change as much as the underlying severity of the patients change.

The Chi-Square statistic tests if the difference between the observed values and expected values across the class interval for each item are significant or not. A non-significant Chi-Square statistic less than 0.05, or Bonferroni adjusted value to account for multiple testing, indicates good fit to the model (Pallant & Tennant, 2007). The Bonferroni adjustment involves dividing the original probability-level (0.05) by the number of times a statistical test is repeated; this can be done automatically in RUMM2020. If an item demonstrates a significant Chi-Square statistic then it is deemed to misfit model expectations and should be investigated further. If necessary, the mis-fitting item should then be amended or removed. Table 4 illustrates individual item fit including fit-residual and chi-square probability. Individual item fit was also viewed graphically, using the item characteristic curve (ICC). The ICC plots the model fit for each class interval against the expected model for that item (Figures 5 and 6).

It is important to note that an advantage of Rasch analysis is the ability to gain information about how items are working, both individually and as a scale. There are no set rules as to whether mis-fitting items are retained or removed and is different for each scale. During the iterative process of analysing the 34 items, nine demonstrated a mis-fit and were removed in this instance.

Insert Figures 5 and 6

Differential Item Functioning

Another source of misfit in the data is differential item functioning (DIF). This is a type of bias such as when different patient groups within the sample respond in a different manner to an item despite equally severe levels of breathlessness (Wilson, 2005). It is up to the questionnaire designer to decide which patient variables are entered into Rasch programme to test for DIF. This depends on the construct being measured and the factors thought to potentially impact on patients' responses to the items. Because our aim was to develop a questionnaire that could measure breathlessness across different disease groups we entered diagnosis as a factor. Item bias relating to gender was also assessed. This is a requirement of invariance and is a requirement of scales where summated scores are calculated (Pallant & Tennant, 2007).

DIF is tested using analysis of variance (ANOVA). A significant probability $p < 0.05$ (or the Bonferroni adjusted level) indicates significant DIF. There are two forms of DIF. Uniform DIF is where there is uniformity in the difference of severity for an item between groups of patients (Tennant & Conaghan, 2007). For example, where one group displays a consistently higher or lower score on a given item relative to the overall severity judged by the patients' responses to all of the items aggregated together (Figure 7). Non-uniform DIF occurs when there is non-uniformity within the differences between groups (Tennant & Conaghan, 2007) (Figure 8). That is, when the severity differences to confirm an item are inconsistent amongst the groups (i.e. CI). DIF should be accounted for in order to be able to maintain the invariant comparisons between groups. For a summed scale score to remain comparative across groups, items displaying DIF should be removed, unless otherwise justified. In this study, seven items demonstrated DIF associated with gender and 11 associated with diagnosis.

Insert figures 7 and 8

Fit of the 12-item set to the Rasch model

The above tests of individual item fit and DIF were applied and items removed until a set of items that conform to the Rasch model was achieved. This process resulted in 12 items being retained and is called the Dyspnoea-12 (Table 3). A number of tests were applied to the 12-item set to examine how well it conformed to the Rasch model. Initially, an estimate of the internal consistency reliability of the scale was tested using the person separation index (PSI). The PSI is analogous to Cronbach's alpha, used in CTT, but uses the logit value as opposed to the raw score (Wilson, 2005). The PSI of the Dyspnea-12 was 0.89 demonstrating good internal reliability.

Rasch analysis tested whether the 12-item set behaved in the same way across different levels of severity. This is tested using the Chi-Square statistic and is called item-trait interaction. This test asks the question: "Are all items working as expected at different levels of breathlessness severity?" (Tennant & Conaghan, 2007). This is a formal test of whether the hierarchical severity order of the items remains consistent across different levels of breathlessness severity. This is determined by a non-significant Chi-Square probability value ($p > 0.05$). This value is a summation of all of the individual item chi-square fit statistics, and therefore Bonferroni adjustments may again be applied if necessary. The item-trait interaction statistic for the 12-item set was not significant (chi-square=76.6, df=60, P=0.08).

Tests of unidimensionality

To classify a construct being measured as unidimensional and justify a total summated score there must be one prominent factor underlying it. Tests to determine scale unidimensionality are still developing. In 2002, Smith reported a *t*-test approach to testing for unidimensionality. This approach has since been amended slightly and is reported in Tennant and Conaghan (2007). This test is done by identifying the two *most different* subsets of items within the scale, and then comparing the person (severity) estimates that are derived using only the items in these subsets. If there is no significant difference between the person estimates generated from the two subsets, then this offers good evidence that the scale is unidimensional. The whole process is internal to the RUMM computer program.

The subsets of items are identified by testing the factor loadings on the first principal component of the residuals. The highest positive set of correlated items and the highest negative set of correlated items are then selected as the two subsets and individual person estimates are generated from the two item sets. The severity level estimates derived from these subsets is then compared for each person, using a series of t-tests, to determine if they significantly differ from each other (Smith 2002). If the person severity estimate is found to significantly differ in more than five percent of patients, this would indicate the presence of multidimensionality (Pallant, et al. 2006). In other words, the two subsets are so different that they measure different, but possibly related, constructs. A confidence interval is then applied and its lower bound should overlap 5% for a non-significant test (Tennant and Conaghan, 2007).

Smith's (2002) test of unidimensionality was applied to the 12-items set. The number of significant t-tests was acceptable 6.7% (confidence interval 4.4-9.0%), offering evidence of the scales unidimensionality.

Tests of Item and Patient Severity Level

Since breathlessness severity for patients and items can be placed upon the same logit scale, it is possible to test how well the items are targeted to the population studied (i.e. how well the level of breathlessness severity covered in the items is matched to breathlessness severity of patients). The average breathlessness 'severity' of the patients was -0.63 logits (SD 1.4), and for items was 0 (SD 0.92) (by convention, item severity is centred on zero logits) (Figure 9). This suggests that the items were well matched to the patients, although on average the items were targeted to patients who would be slightly more severe than those recruited in this project.

It is also possible to assess the relative severity level of each item. This is determined by examining the logit score for each individual item; a higher logit indicates an item that expressing more severe breathlessness (Table 4).

Insert figure 9

Insert table 4

Discussion

This paper provides a practical working example of Rasch analysis. We have presented the steps involved in reducing and refining a large item-set by identifying those items which possessed the most reliable measurement properties. We have provided nurse researchers with an alternative to CTT when developing or refining questionnaires that measure PROs. However, it is important to note that to make a true comparison between CTT and Rasch analysis both techniques would need to be applied to the same data set. This paper reports the application of Rasch analysis only.

Some researchers have used a combined approach to questionnaire development with item-total correlations being used to guide the reduction of a large item sets prior to applying Rasch analysis (Jones et al, 2009). Likewise, a principal component analysis may be applied to the data to identify multidimensionality prior to undertaking a Rasch analysis. In fact, this may be a beneficial procedure to carry out as the Rasch model assumes that a unidimensional dataset is being assessed. It is not the function of a Rasch analysis to inform how many dimensions there are in a dataset, and which items are loading onto which dimension.

If separate dimensions are identified during an exploratory factor analysis, then the factor loadings could inform item removal. Alternatively, a separate Rasch analysis could be applied to the different components to create separate, but related, scales. There are no set rules regarding this. However, it is important that the questionnaire developer determine these factors early in the process and take into consideration the construct being measured. Our aim was to develop an overall score for breathlessness severity that reflected different aspects of the experience. Therefore, we continued item reduction until all items met model expectations, meaning that an internally robust and unidimensional scale had been obtained.

The application of Rasch to the development of the Dyspnoea-12 enabled close examination of each item's contribution to the reliability to the overall measure. Whereas CTT will highlight correlated items, it does not signify severity level of individual items or patients; it cannot test the measurement properties of items at different levels of symptom severity (Borsboom, 2005). This study provides insight into information regarding the severity level of breathlessness expressed by different words that patients use to describe the experience. To our knowledge, this study represents the first time that the level of breathlessness severity expressed by different words has been quantified.

In Rasch methodology, the fit statistics and total item-trait interaction provided a thorough and robust method for testing the effect of removing an item on the reliability of the all items combined. In addition, the ICC's provided a graphical presentation that enabled detailed assessment of each items performance at different severity levels of breathlessness. These aspects enabled items with the best fit to the model and precision to be retained.

A particular advantage of using Rasch in this study was the ability to rigorously scrutinise DIF related diagnosis. This was important to Dyspnoea-12 development because the aim was to produce a questionnaire that was relevant across disease groups. As such, items were required to demonstrate invariance across disease groups. This approach was undertaken because breathlessness is a cardinal symptom of cardiorespiratory disease and many patients have a combination of diseases. This function can also be used to test for bias in relation to country, i.e. it tests that persons from different countries respond to an item in a similar way, given the same severity level of the trait being measured. This has important implications for questionnaires being developed in more than one country (Jones et al, 2009) or the validation of questionnaires into different languages.

An element of Rasch analysis that is still developing is the testing of unidimensionality. We used the method described by Smith (2002) which is often used in health care measures developed with Rasch methods (Gilworth et al, 2007). From the two sets of items that were identified in PCA of the residuals, the person estimates derived from these subsets were not significantly different from one another, thereby supporting the concept that the Dyspnoea-12 provides an overall score for breathlessness severity. The patterning of the items appears to be related to the logit severity associated with each item; items representative of affect tended to have a higher logit value than other items.

In summary, this paper presents a practical example of Rasch analysis. Whilst no questionnaire is perfect, PROs provide us with a unique reference to the patients' perception of, for example, symptom severity. It is, therefore, vital that these measures are developed and refined using robust psychometric techniques. This paper has demonstrated that Rasch analysis provides a viable option for questionnaire development and refinement. It presents a detailed description of the processes involved and provides nurses with a guide to critiquing the robustness of other questionnaires developed using Rasch analysis.

Words = 4,356

Table 1: Sample demographics

	COPD n = 123 Mean (SD)	ILD n = 129 Mean (SD)	CHF n = 106 Mean (SD)
Male gender	62 (51%)	47 (36%)	72 (68%)
Age, years	69 (±8)	50 (±12)	68 (±11)
FEV_1%predicted	48 (±16)	69 (±22)	-
FVC % predicted	72 (±19)	69 (±19)	-
FEV_1:FVC % predicted	55 (±12)	83 (±8)	-
Left ventricular ejection fraction	-	-	35 (±15)
MRC Dyspnoea Scale (0-5)	3.4 (±1.1)	3.0 (±1.1)	2.6 (±1.1)

COPD: chronic obstructive pulmonary disease

ILD: interstitial lung disease

CHF: chronic heart failure

FEV_1 forced expired volume in one second

FVC: forced vital capacity

FEV_1: FVC: ratio

MRC: Medical Research Council

Table 2: Class interval patient numbers

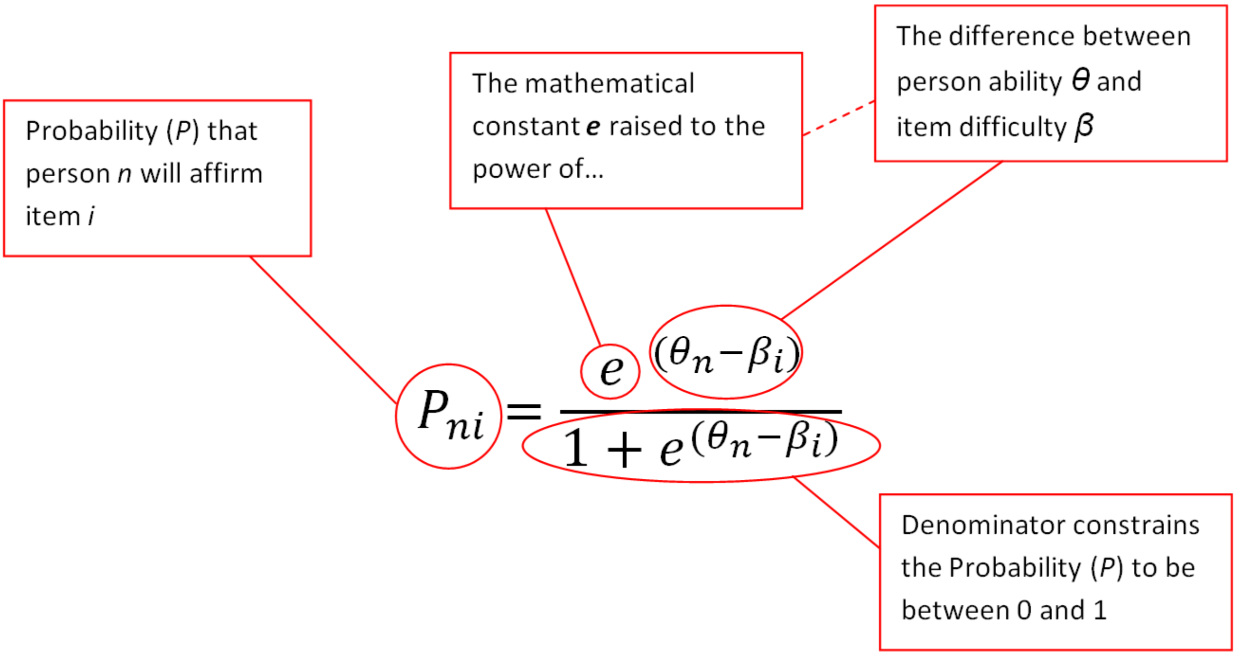
Class Interval	Number of patients
1 (least severe)	59
2	57
3	58
4	57
5	58
6 (most severe)	62

Table 3: Rasch fit statistics for the retained 12 items

Item	Fit residual	Chi-square statistic	Probability (Bonferroni P<0.002)
Irritating	-0.11	2.42	0.79
In all the way	2.17	5.27	0.38
More work	0.08	5.62	0.35
Not get enough air	-0.69	6.68	0.25
Exhausting	0.14	3.18	0.67
Difficulty catching breath	0.99	3.06	0.69
Short of breath	-0.34	12.80	0.03
Uncomfortable	-0.42	11.31	0.05
Agitated	-1.55	12.97	0.02
Distressing	-0.33	4.29	0.51
Depressed	-0.43	4.27	0.51
Miserable	0.21	3.81	0.58

Table 4: Individual item severity expressed in logits

Items	**Logit scores** (lowest to highest severity)
Short of breath	-0.970
Not enough air	-0.462
Exhausting	-0.132
More work	-0.130
Uncomfortable	0.081
Difficulty catching breath	0.168
Depressed	0.202
Not in all way	0.261
Irritating	0.321
Miserable	0.340
Agitated	0.438



$$P_{ni} = \frac{e^{(\theta_n - \beta_i)}}{1 + e^{(\theta_n - \beta_i)}}$$

Figure 1: Rasch equation

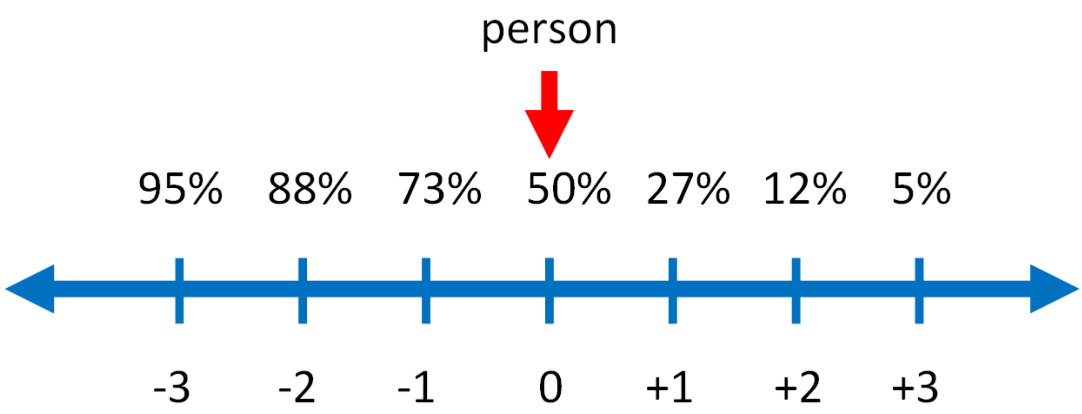


Figure 2: Probability of a person affirming an item.

The bottom of the scale displays the difference in location (in logits) between a person and an item. The top of the scale displays the corresponding probability of the person affirming the item. The probability of a person with ability (severity) of 0 logits affirming an item with a difficulty of 0 logits is 50%. The probability of a person affirming an item with a difficulty that is 2 logits higher than their ability is 12%.

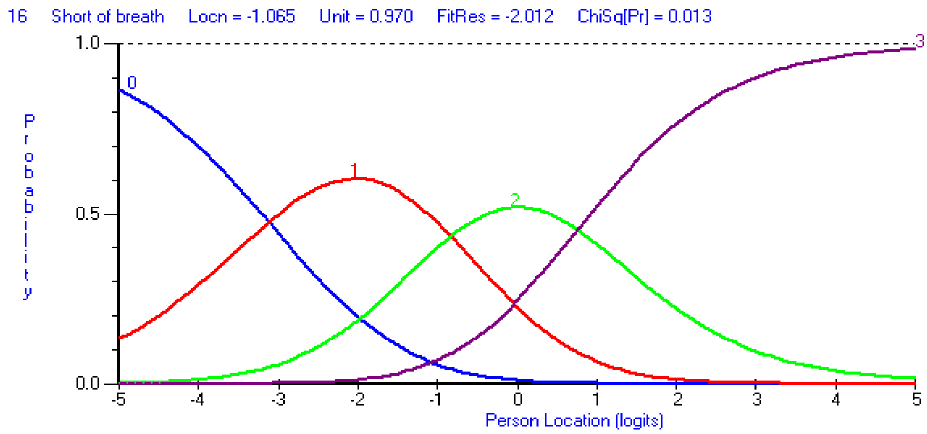


Figure 3: Example of well-ordered transition between categories for the item 'I feel short of breath'. The 'y' axis represents the likelihood of a given response and the 'x' axis represents patient severity. It can be seen that the responses to this item fall in a logical progressive order – category responses represent an increase in breathlessness severity (logits) for each category. For example, at the lowest patient severity (-5 logits) the probability of a score a 0 (i.e. 'none') is most likely, and at the highest patient severity (+5) the probability of scoring a 3 (i.e. 'severe') is most likely.

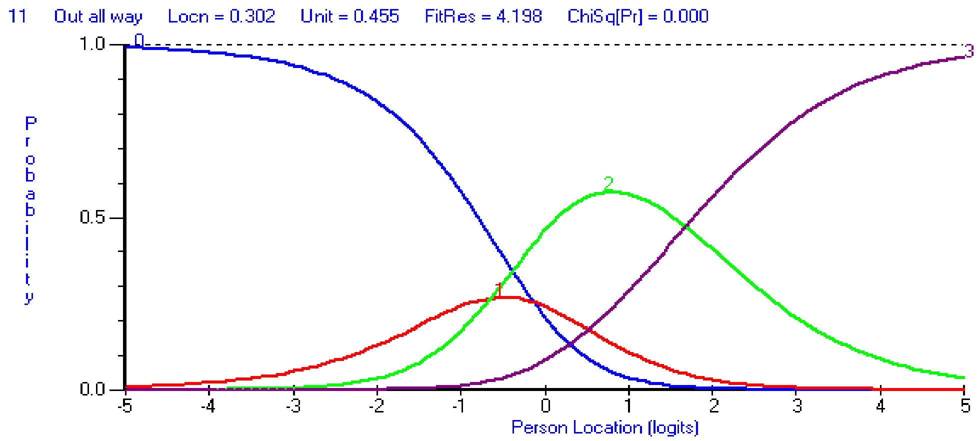


Figure 4: Example of disordered threshold for the item ‘My breath does not go out all the way’. At no patient severity is it most likely that category 1 (‘mild’) will be scored. That is, even at the point where the probability of scoring a 1 is at its peak, it is still more likely that category 0 or category 2 will be scored instead.

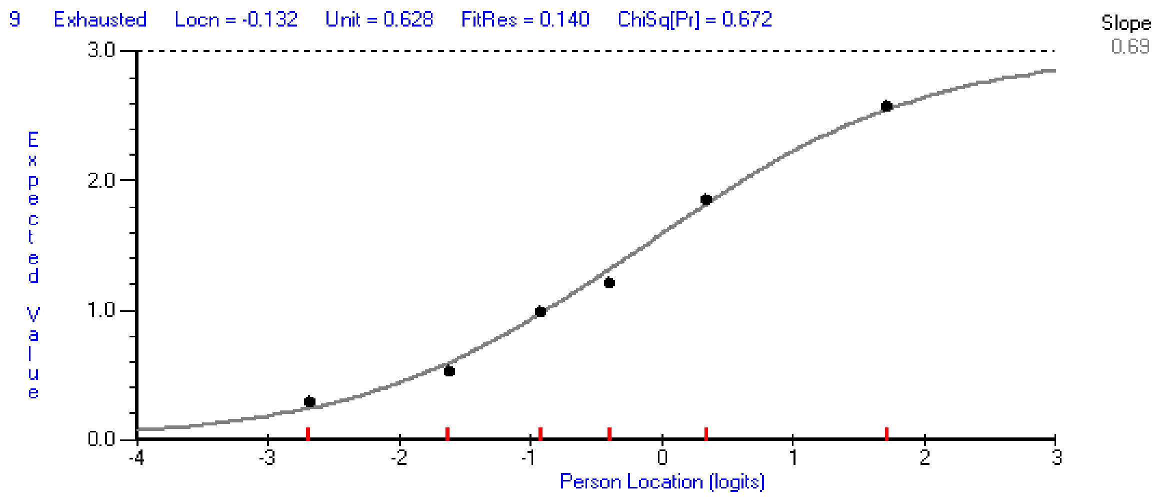


Figure 5: Item Characteristic Curve for a well-fitting item- 'My breathing is exhausting'. The 'y' axis represents the item severity and the 'x' axis represents patient severity in logits. The curved line represents the expected scores for the item, and the dots represent the observed scores for the class intervals at the different severity levels. The fit residual (along the top) is +0.140 and the Chi-Square probability is 0.672, indicating no significant deviation between the expected and observed scores for this item.

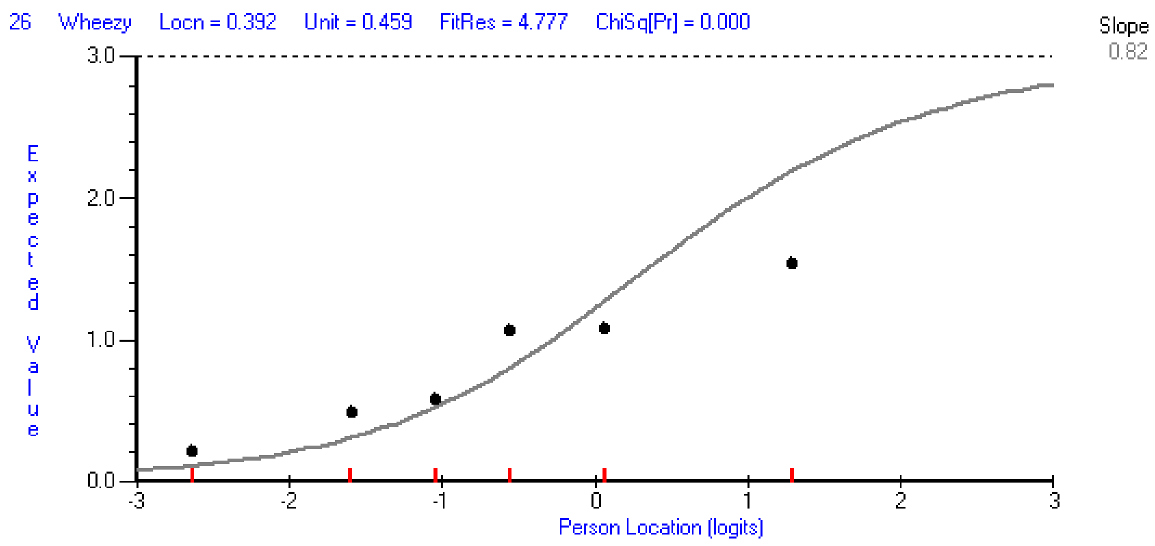


Figure 6: Item Characteristic Curve for a non-fitting item - 'I feel wheezy'. This item is under-discriminating – the observed scores (black dots) form a flatter curve than the expected scores. The fit residual is 4.77 and the Chi-Square is significant ($p < 0.001$). This item was consequently removed.

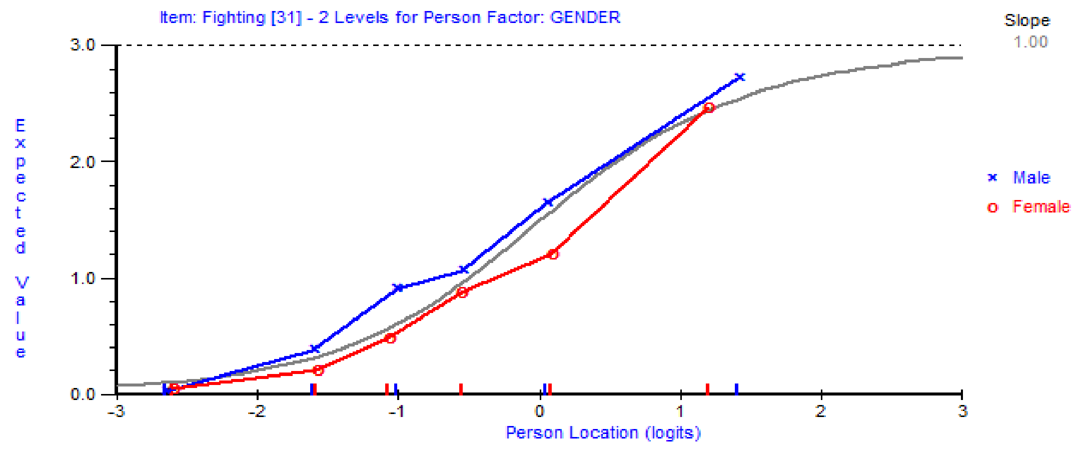


Figure 7: Example of an item – ‘I am fighting for breath’ demonstrating gender associated uniform DIF. It can be seen that the female group (red line) is slightly but consistently below the male group (blue line). This means that for any level of overall breathlessness severity, females had a lower (i.e. less severe) response to this particular item.

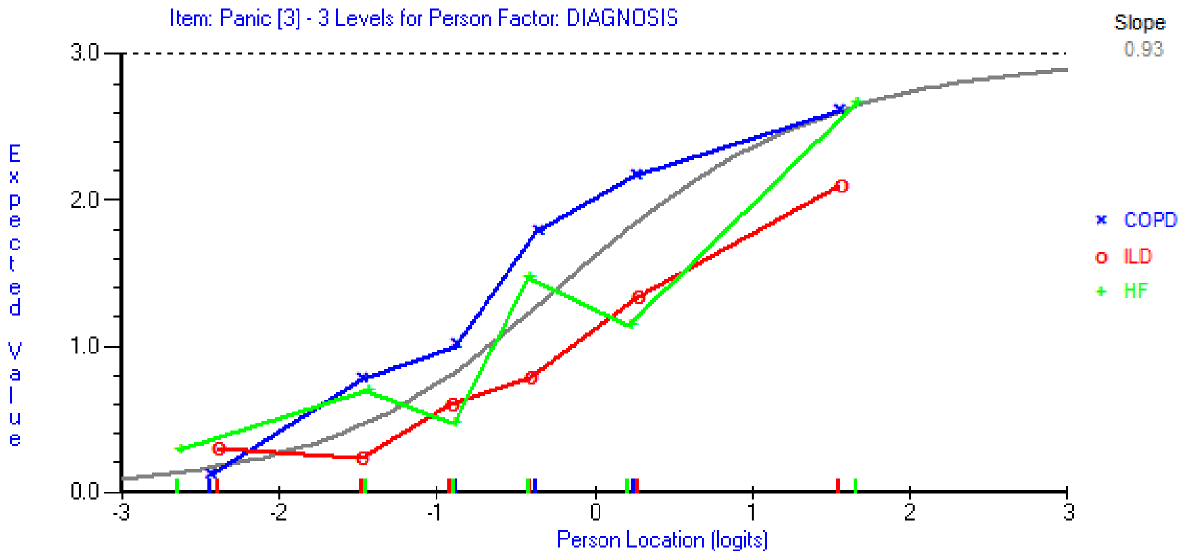


Figure 8: Example of an item – ‘My breathing makes me panic’ demonstrating uniform and non-uniform DIF. At any given level of breathlessness severity, patients with COPD consistently score higher than patients with ILD on this item. Patients with CHF represent the most erratic responses for this item, displaying the lack of consistent difference (non-uniformity) between groups.

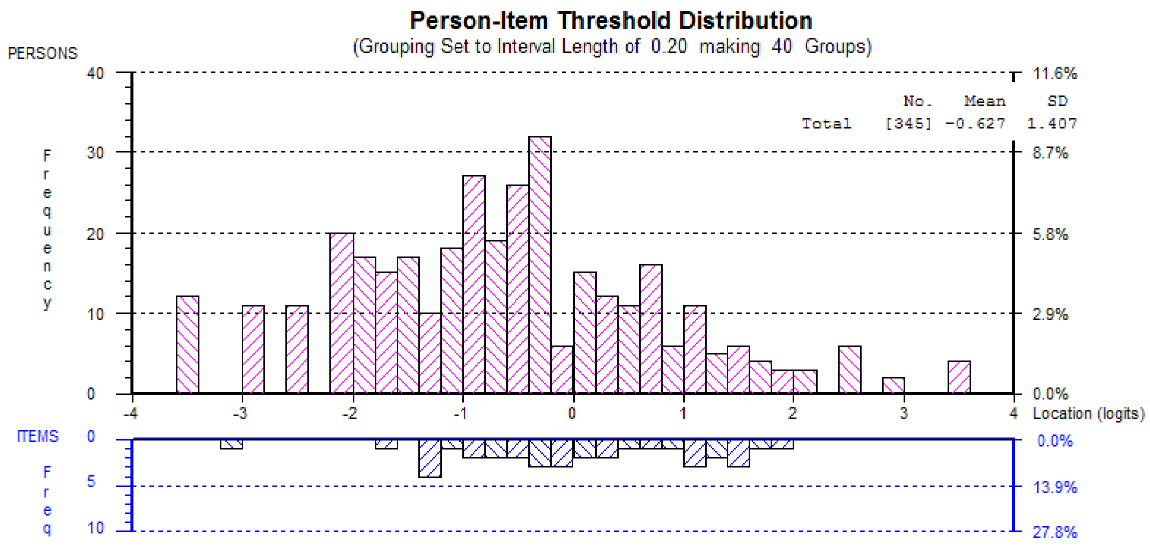


Figure 9: Distribution of patients and item thresholds based on Rasch logit. This figure shows the distributions of patient severity and item severity (locations) along the same linear scale measured in logits. Patients are on the upper part of the graph (pink boxes) and item locations on the lower part (blue boxes). Most of the item thresholds are located between -2 and +2 logits. It can be seen that, on average, patients are at the milder end of the severity scale.

Andrich D., Lyne A., Sheridan B., Luo G (2003). RUMM2020. *RUMM Laboratory*, Perth, Australia.

Andrich D. (2004) Controversy and the Rasch model: a characteristic of incompatible paradigms? *Medical Care* 42, I7-I16

Borsboom, D. (2005) *Measuring the Mind: Conceptual issues in contemporary psychophysics.* Cambridge, Cambridge University Press.

DeVellis R.F. (2006) *Scale Development: theory and applications* (2nd Edition), Sage, Thousand Oaks.

Gilworth G., Bhakta B., Eyres S., Carey A., Chamberlain M.A. & Tennant A. (2007). Keeping nurses working: development and psychometric testing of the Nurse-Work Instability Scale (Nurse-WIS). *Journal of Advanced Nursing* **57**(5): 543-551.

Guttman L. (1944) A basis for scaling qualitative data. *American Sociological Review* 9:139–150

Hagquist C, Bruce M, Gustavsson JP. (2009) Using the Rasch model in nursing research: an introduction and illustrative example. *International Journal of Nursing Studies*. 46(3), 380-393.

Hobart J, Cano S. Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. *Health Technology Assessment* 2009; Vol. 13: No. 12.

Jones P.W., Harding G., Berry P., Wiklund I., Chen W-H., Kline Leidy N. (2009) *Development and first validation of the COPD Assessment Test.* European Respiratory Journal 34, 648-654

Nunnally J. (1978). *Psychometric Theory.* New York, Mc Graw-Hill.

Pallant J.F. & Tennant A. (2007). An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). *British Journal of Clinical Psychology* 46(Pt 1): 1-18.

Pallant J.F., Miller R.L., Tennant A. (2006) Evaluation of the Edinburgh Post Natal Depression Scale using Rasch analysis. *BMC Psychiatry* 6-28.

Rasch G. (1960). Probabilistic models for some intelligence and attainment tests. *Danish Institute of Educational Research*.

Rattray J. & Jones M.C. (2007). Essential elements of questionnaire design and development. *Journal of Clinical Nursing* 16: 234-243.

Smith E.V. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal components analysis of residuals. *Journal of Allied Measurement* 3: 205-231.

Tennant A. & Conaghan P.G. (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 & Rheumatism* 57(8): 1358-62.

Tesio L, Simone A, Bernardinello M. (2007) Rehabilitation and outcome measurement: where is Rasch analysis-going? *Europa Medicophysphysica* 43(3), 417-26.

van Alphen A., Halfens R., Hasman A., Imbos T. (1994). Likert or Rasch? Nothing is more applicable than good theory. *Journal of Advanced Nursing* 20(1): 196-201.

Watson R. & Thompson D.R. (2006). Use of factor analysis in Journal of Advanced Nursing: literature review. *Journal of Advanced Nursing* 55: 330-341.

Wilson, M. (2005) *Constructing Measures: An Item Response Modelling Approach*. Routledge Academic, New York.

Wilson R.C. and Jones P.W. (1991) Differentiation between the intensity of breathlessness and the distress it evokes in normal subjects during exercise. *Clinical Science* 80(1): 65-70.

Yorke J. (2008). Dyspnoea and pain: an interesting analogy. *Journal of Clinical Nursing* 17(7), 841-842.

Yorke J., Moosavi S.M., Shuldham C., Jones P.W. (2010). Quantification of dyspnoea using descriptors: development and initial testing of the Dyspnoea-12. *Thorax* **65**(1): 21-26.