



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

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



Copy

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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 paper 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,

1
2
3 a unidimensional scale. The aim is to provide a working example of Rasch techniques
4 and in doing so, attempts to demystify the complexities of this psychometric approach.
5
6
7

8 9 **Background**

10
11 Despite the prevalence of CTT being utilised in scale development, it is important to
12 note that CTT has a number of limitations, all of which could have implications upon
13 the scales that have been derived under that methodology. Briefly, CTT is limited in
14 that the scale scores derived are of an ordinal nature, but they are often treated as
15 interval-level data. This means that CTT-based scales may be prone to differential
16 sensitivity at the centre, relative to the extremes, of the score range (DeVellis, 2006).
17 Another limitation is that the evaluations of scales are dependent upon the sample on
18 which they have been tested against, and also, the measurement of people is
19 dependent upon the set of items with which they have been measured (Hobart &
20 Cano, 2009). This means that different samples with different variances will not yield
21 equivalent data or data that can easily be compared across samples (DeVellis, 2006).
22 Hobart and Cano (2009) identified that Rasch analysis represents a logical
23 progression from CTT, as it attempts to improve the scientific quality of the theory
24 underpinning rating scales.
25
26
27
28
29
30
31
32
33
34
35
36
37
38

39 The Rasch model was developed by the Danish mathematician Georg Rasch within
40 the realm of educational psychology (Rasch, 1960). Its primary function is to test how
41 well items within an instrument conform to a unidimensional model. In other words, it
42 checks if the underlying construct being measured has a single dimension on which all
43 of the questionnaire items rely. This is a key concept for instruments where a total
44 summated score is calculated (Hagquist et al., 2009). Historically, in Rasch
45 measurement the position that each item and person occupies on this dimension is
46 termed its item 'difficulty' and person 'ability'. This is because much of the early work
47 was carried out in the field of education using multiple choice exam questions. These
48 terms are also applied, for example, when quantifying physical ability levels in the field
49 of rehabilitation. In symptom measurement, such as breathlessness, 'severity' is a
50 better term and will now be used throughout this paper.
51
52
53
54
55
56
57
58
59
60

1
2
3 There are two key features of the relationship between a person's symptom severity
4 and that expressed by an item in a questionnaire. First, the observed response is
5 dependent on the difference between patient severity and the severity of the item.
6
7 Second, the model is probabilistic as uncertainty (a theory-based probability)
8 surrounds the expected response, consistent with the real life situation (Tesio, et al,
9 2007). The Rasch model assumes that the probability that a person will affirm an item
10 is a logistic function of the *difference* between the person's ability [θ] (in terms of the
11 breathlessness severity level of the person) and the difficulty of the question [β] (again,
12 in terms of the breathlessness severity level represented by the item), and only a
13 function of that difference. For an explanation of the equation, broken down into its
14 component parts, please see Figure 1.
15
16
17
18
19
20
21
22

23
24 Insert figure 1
25

26 The Rasch model tests that items measuring a lower severity are more likely to be
27 endorsed by patients with a higher level of severity (as determined by the responses to
28 all items combined). The converse is true when a person's severity is less than that for
29 the item (Borsboom, 2005). This is a property of scales that is commonly termed
30 Guttman scaling (Guttman, 1944). In Rasch analysis the response patterns obtained
31 are tested against what is expected, so it is a probabilistic form of Guttman scaling
32 (Pallant & Tennant 2007). The resulting severity estimates for items and respondents
33 are reported in 'logits' – the log-odds of responding to an item (Figure 2).
34
35
36
37
38
39
40

41 Insert figure 2
42

43 It also offers an alternative where the limitations of CTT can be overcome; Rasch
44 analysis can provide a transformation of an ordinal score into a linear, interval-level
45 variable, given fit of data to Rasch model expectations (Tennant & Conaghan, 2007).
46 This means that the problem of differential sensitivity can be overcome. Also, the
47 Rasch model has the advantageous property of invariance, meaning that the item and
48 person parameters can be estimated independently of each other (Andrich, 2004). To
49 put this another way; the measurement of people is not dependent upon the sampling
50 distribution of the set of items with which they have been measured, and the difficulty
51 estimates of items on a scale are not dependent upon the distribution of the sample on
52 which they have been derived (Hobart & Cano, 2009). For a more complete account of
53
54
55
56
57
58
59
60

1
2
3 the development of modern test theory in the health sciences, along with its
4 advantages over CTT, the reader is directed towards Hobart & Cano (2009).
5
6
7
8

9
10 Rasch analysis enables an examination of the contribution of items individually as they
11 are added and removed from the item set. This enables the selection of items during
12 questionnaire development phase that provide maximum measurement precision. If
13 data fit model expectations then a fundamental assumption is that each item
14 contributes reliably to the measurement of the single underlying construct. If an item
15 set meets the criteria of invariance and items form a unidimensional scale, then a
16 summated score for the concept being measured can be legitimately obtained.
17
18
19
20
21
22
23
24

25 **The Study**

26 **Aims**

27
28 To describe and critique the application of Rasch analysis to the development of the
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
Dyspnoea-12.

36 **Participants**

37
38 Participants were recruited from out-patient clinics from three NHS Trusts. Participants
39 completed an 81-item list during their clinic visit (n=275, 77%); the remainder
40 completed them at home for return within two weeks. Their baseline characteristics are
41 shown in Table 1. The study received ethical approval from the Local Research Ethics
42 Committee and all participants provided written consent to participate.
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Insert table 1

55 **Dyspnoea-12**

56
57 Details of the overall development of Dyspnoea-12 are described elsewhere (Yorke et
58 al, 2010). Briefly, a pool of 81-items was arranged as a questionnaire that asked
59 patients to respond to each one reflecting their current experience of being breathless.
60

1
2
3 Response options were 'none', 'mild', 'moderate' or 'severe'. Items were removed if
4 more than 50% of the sample responded 'none' or demonstrated bias associated with
5 age. Thirty-four items survived this process and were further reduced and refined
6 using Rasch analysis (Rasch, 1960). The remainder of this paper reflects the
7 application of Rasch analysis to these 34 items which was reduced to a final 12-item
8 set; Dyspnoea-12.
9
10
11
12
13
14
15
16

17 **The Application of Rasch Analysis**

18
19 In this study, analyses were performed using Rasch Unidimensional Measurement
20 Model (RUMM2020) (Andrich et al., 2003). Whilst the steps taken to analyse data and
21 remove/retain items are presented here in a logical step-by-step format, it is important
22 to note that these applications involved an iterative approach. The aim in this instance
23 was to identify the mis-fitting items and examine the effect of their removal on other
24 items and the total item-set.
25
26
27
28
29
30
31
32

33 *Class Intervals*

34
35 In RUMM2020, patients are automatically placed into groups called class intervals
36 (CI). Class intervals are defined by ordering all patients in terms of breathlessness
37 severity (determined by the responses to all items combined) and then splitting them
38 into groups of approximately equivalent size across the sample (Tennant & Conaghan,
39 2007). As demonstrated below, a number of Rasch fit statistics are applied at the CI
40 level. In this study, the 358 patients were grouped into six CI and represented an
41 acceptable dispersion of patient numbers (Table 2). In this sense, the CI's can be seen
42 to represent different, discrete, levels of severity.
43
44
45
46
47
48
49
50

51 *Insert table 2*

52 53 54 55 56 *Ordering of response categories*

57
58 The 34 items each had a polytymous response format ranging from 0 (none) to 3
59 (severe). Patients' responses to these options need to follow a logical sequence. It
60

1
2
3 would be expected that as patient breathlessness severity increases, it is more likely
4 to score a 0, then a 1, then a 2, then a 3 for any particular item. In Rasch analysis
5 terms this would be indicated by an ordered set of response thresholds for each item
6 (Pallant et al., 2006). The term threshold refers to the point between two response
7 categories (e.g. 1 'mild' and 2 'moderate') where the probability of scoring a 1 on the
8 item or scoring a 2 is 50/50 (given that the person will only score a 1 or a 2); in other
9 words, the point where either response is equally probable (Pallant & Tennant, 2007)
10 (Figure 2). Item response patterns that do not follow a logical order are termed
11 disordered, or reversed, thresholds. Disordered thresholds can occur when patients
12 have difficulty consistently discriminating between response options when, for
13 example, there are too many response options, or when the labelling of options is
14 confusing. Correct ordering can often be achieved by combining adjacent response
15 categories and rescoreing the item (Pallant & Tennant, 2007).
16
17
18
19
20
21
22
23
24
25
26
27
28

29 Item thresholds can be assessed graphically using the item category probability curves
30 (Figures 3 and 4). They may also be assessed by looking at the actual numerical
31 threshold estimates. One of the 34 items showed a situation in which patients
32 demonstrate an inconsistent transition between response options ('My breath does not
33 go out all the way') (Figure 4); this item was subsequently removed due to lack of fit to
34 the model.
35
36
37
38
39
40
41
42

43 For an instrument with polytomous items there are two parameterisations of the Rasch
44 model that can be assessed using the RUMM programme: the Rating Scale Model or
45 the Partial Credit Model. These two models differ slightly in their mathematics where
46 the former expects the distances between thresholds to be equal across items
47 (Tennant & Conaghan, 2007). This means that the metric distance between, for
48 example, the thresholds separating categories 1 and 2 is the same across all items,
49 and that the metric distance separating categories 2 and 3 is the same across all
50 items. However, the distance between categories 1 and 2 does not have to be
51 equivalent to the distance between categories 2 and 3. The Rating Scale Model
52 provides a higher degree of specificity; however, it is not always possible to satisfy the
53 assumptions of a Rating Scale Model, in which case the Partial Credit Model should
54
55
56
57
58
59
60

1
2
3
4 be utilised. The likelihood-ratio test provides a test of which version of the model
5 should be utilised by comparing the different parameterisations of the model and
6 providing a Chi-square statistic and probability – if the outcome of the test is not
7 significant ($p > 0.05$), then the Rating Scale Model should be adopted as it is a simpler
8 model. The test for the 34 items used in this study (as is often the case when using
9 real data) was $p < 0.01$; requiring the Partial Credit Model to be used.

14
15
16
17 *Insert Figures 3 and 4*

21 *Tests of Individual Item Fit*

22
23
24 Tests of individual item fit to the Rasch model reflect the differences between the
25 observed responses and that expected by the model (i.e. expected responses given
26 the level of breathlessness severity based on patients' responses to all items
27 combined). This is an important feature of Rasch analysis because it tests the ability of
28 individual items to reliably measure breathlessness at different severity levels (i.e. CI).
29 These tests are presented for each item as a fit residual and as a Chi-Square
30 probability statistic. A residual is a summation of individual item (or person) deviations
31 from model expectations, which are then standardised to form a z-score. Those
32 between ± 2.5 are deemed to generally indicate adequate fit to the model (Pallant &
33 Tennant, 2007). A high negative residual indicates an over-discriminating item, which
34 is also a possible indicator of redundancy. Redundant items offer nothing to the
35 information gained by other items, and removal should improve the fit of those
36 remaining items. In some respects it is analogous with a high item-total correlation
37 used in CTT (Pallant et al, 2006). High positive residuals indicate under-discriminating
38 items, which suggest that these items are not contributing to measure the underlying
39 trait in question. That is, such items do not have discriminant power; the item's
40 responses do not change as much as the underlying severity of the patients change.

41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
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

1
2
3 for multiple testing, indicates good fit to the model (Pallant & Tennant, 2007). The
4 Bonferroni adjustment involves dividing the original probability-level (0.05) by the
5 number of times a statistical test is repeated; this can be done automatically in
6 RUMM2020. If an item demonstrates a significant Chi-Square statistic then it is
7 deemed to misfit model expectations and should be investigated further. If necessary,
8 the mis-fitting item should then be amended or removed. Table 4 illustrates individual
9 item fit including fit-residual and chi-square probability. Individual item fit was also
10 viewed graphically, using the item characteristic curve (ICC). The ICC plots the model
11 fit for each class interval against the expected model for that item (Figures 5 and 6).
12
13
14
15
16
17
18
19
20
21

22 It is important to note that an advantage of Rasch analysis is the ability to gain
23 information about how items are working, both individually and as a scale. There are
24 no set rules as to whether mis-fitting items are retained or removed and is different for
25 each scale. During the iterative process of analysing the 34 items, nine demonstrated
26 a mis-fit and were removed in this instance.
27
28
29
30
31
32
33

34 *Insert Figures 5 and 6*

35 36 37 38 *Differential Item Functioning*

39 Another source of misfit in the data is differential item functioning (DIF). This is a type
40 of bias such as when different patient groups within the sample respond in a different
41 manner to an item despite equally severe levels of breathlessness (Wilson, 2005). It is
42 up to the questionnaire designer to decide which patient variables are entered into
43 Rasch programme to test for DIF. This depends on the construct being measured and
44 the factors thought to potentially impact on patients' responses to the items. Because
45 our aim was to develop a questionnaire that could measure breathlessness across
46 different disease groups we entered diagnosis as a factor. Item bias relating to gender
47 was also assessed. This is a requirement of invariance and is a requirement of scales
48 where summated scores are calculated (Pallant & Tennant, 2007).
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 DIF is tested using analysis of variance (ANOVA). A significant probability $p < 0.05$ (or
4 the Bonferroni adjusted level) indicates significant DIF. There are two forms of DIF.
5 Uniform DIF is where there is uniformity in the difference of severity for an item
6 between groups of patients (Tennant & Conaghan, 2007). For example, where one
7 group displays a consistently higher or lower score on a given item relative to the
8 overall severity judged by the patients' responses to all of the items aggregated
9 together (Figure 7). Non-uniform DIF occurs when there is non-uniformity within the
10 differences between groups (Tennant & Conaghan, 2007) (Figure 8). That is, when the
11 severity differences to confirm an item are inconsistent amongst the groups (i.e. CI).
12 DIF should be accounted for in order to be able to maintain the invariant comparisons
13 between groups. For a summed scale score to remain comparative across groups,
14 items displaying DIF should be removed, unless otherwise justified. In this study,
15 seven items demonstrated DIF associated with gender and 11 associated with
16 diagnosis.
17
18
19
20
21
22
23
24
25
26
27
28
29
30

31 *Insert figures 7 and 8*
32
33
34

35 *Fit of the 12-item set to the Rasch model* 36 37

38 The above tests of individual item fit and DIF were applied and items removed until a
39 set of items that conform to the Rasch model was achieved. This process resulted in
40 12 items being retained and is called the Dyspnoea-12 (Table 3). A number of tests
41 were applied to the 12-item set to examine how well it conformed to the Rasch model.
42 Initially, an estimate of the internal consistency reliability of the scale was tested using
43 the person separation index (PSI). The PSI is analogous to Cronbach's alpha, used in
44 CTT, but uses the logit value as opposed to the raw score (Wilson, 2005). The PSI of
45 the Dyspnea-12 was 0.89 demonstrating good internal reliability.
46
47
48
49
50
51
52
53
54

55 Rasch analysis tested whether the 12-item set behaved in the same way across
56 different levels of severity. This is tested using the Chi-Square statistic and is called
57 item-trait interaction. This test asks the question: "Are all items working as expected at
58 different levels of breathlessness severity?" (Tennant & Conaghan, 2007). This is a
59 formal test of whether the hierarchical severity order of the items remains consistent
60

1
2
3 across different levels of breathlessness severity. This is determined by a non-
4 significant Chi-Square probability value ($p > 0.05$). This value is a summation of all of
5 the individual item chi-square fit statistics, and therefore Bonferroni adjustments may
6 again be applied if necessary. The item-trait interaction statistic for the 12-item set was
7 not significant (chi-square=76.6, df=60, P=0.08).
8
9
10
11

12 13 14 15 *Tests of unidimensionality*

16
17 To classify a construct being measured as unidimensional and justify a total summated
18 score there must be one prominent factor underlying it. Tests to determine scale
19 unidimensionality are still developing. In 2002, Smith reported a *t*-test approach to
20 testing for unidimensionality. This approach has since been amended slightly and is
21 reported in Tennant and Conaghan (2007). This test is done by identifying the two
22 *most different* subsets of items within the scale, and then comparing the person
23 (severity) estimates that are derived using only the items in these subsets. If there is
24 no significant difference between the person estimates generated from the two
25 subsets, then this offers good evidence that the scale is unidimensional. The whole
26 process is internal to the RUMM computer program.
27
28
29
30
31
32
33
34

35
36
37 The subsets of items are identified by testing the factor loadings on the first principal
38 component of the residuals. The highest positive set of correlated items and the
39 highest negative set of correlated items are then selected as the two subsets and
40 individual person estimates are generated from the two item sets. The severity level
41 estimates derived from these subsets is then compared for each person, using a
42 series of *t*-tests, to determine if they significantly differ from each other (Smith 2002). If
43 the person severity estimate is found to significantly differ in more than five percent of
44 patients, this would indicate the presence of multidimensionality (Pallant, et al. 2006).
45 In other words, the two subsets are so different that they measure different, but
46 possibly related, constructs. A confidence interval is then applied and its lower bound
47 should overlap 5% for a non-significant test (Tennant and Conaghan, 2007).
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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.

1
2
3 Some researchers have used a combined approach to questionnaire development
4 with item-total correlations being used to guide the reduction of a large item sets prior
5 to applying Rasch analysis (Jones et al, 2009). Likewise, a principal component
6 analysis may be applied to the data to identify multidimensionality prior to undertaking
7 a Rasch analysis. In fact, this may be a beneficial procedure to carry out as the Rasch
8 model assumes that a unidimensional dataset is being assessed. It is not the function
9 of a Rasch analysis to inform how many dimensions there are in a dataset, and which
10 items are loading onto which dimension.
11
12
13
14
15
16
17
18
19

20 If separate dimensions are identified during an exploratory factor analysis, then the
21 factor loadings could inform item removal. Alternatively, a separate Rasch analysis
22 could be applied to the different components to create separate, but related, scales.
23 There are no set rules regarding this. However, it is important that the questionnaire
24 developer determine these factors early in the process and take into consideration the
25 construct being measured. Our aim was to develop an overall score for breathlessness
26 severity that reflected different aspects of the experience. Therefore, we continued
27 item reduction until all items met model expectations, meaning that an internally robust
28 and unidimensional scale had been obtained.
29
30
31
32
33
34
35
36
37
38

39 The application of Rasch to the development of the Dyspnoea-12 enabled close
40 examination of each item's contribution to the reliability to the overall measure.
41 Whereas CTT will highlight correlated items, it does not signify severity level of
42 individual items or patients; it cannot test the measurement properties of items at
43 different levels of symptom severity (Borsboom, 2005). This study provides insight into
44 information regarding the severity level of breathlessness expressed by different words
45 that patients use to describe the experience. To our knowledge, this study represents
46 the first time that the level of breathlessness severity expressed by different words has
47 been quantified.
48
49
50
51
52
53
54
55
56
57

58 In Rasch methodology, the fit statistics and total item-trait interaction provided a
59 thorough and robust method for testing the effect of removing an item on the reliability
60 of the all items combined. In addition, the ICC's provided a graphical presentation that

1
2
3 enabled detailed assessment of each items performance at different severity levels of
4 breathlessness. These aspects enabled items with the best fit to the model and
5 precision to be retained.
6
7

8
9 A particular advantage of using Rasch in this study was the ability to rigorously
10 scrutinise DIF related diagnosis. This was important to Dyspnoea-12 development
11 because the aim was to produce a questionnaire that was relevant across disease
12 groups. As such, items were required to demonstrate invariance across disease
13 groups. This approach was undertaken because breathlessness is a cardinal symptom
14 of cardiorespiratory disease and many patients have a combination of diseases. This
15 function can also be used to test for bias in relation to country, i.e. it tests that persons
16 from different countries respond to an item in a similar way, given the same severity
17 level of the trait being measured. This has important implications for questionnaires
18 being developed in more than one country (Jones et al, 2009) or the validation of
19 questionnaires into different languages.
20
21
22
23
24
25
26
27
28
29
30

31 An element of Rasch analysis that is still developing is the testing of unidimensionality.
32 We used the method described by Smith (2002) which is often used in health care
33 measures developed with Rasch methods (Gilworth et al, 2007). From the two sets of
34 items that were identified in PCA of the residuals, the person estimates derived from
35 these subsets were not significantly different from one another, thereby supporting the
36 concept that the Dyspnoea-12 provides an overall score for breathlessness severity.
37 The patterning of the items appears to be related to the logit severity associated with
38 each item; items representative of affect tended to have a higher logit value than other
39 items.
40
41
42
43
44
45
46
47
48
49

50 In summary, this paper presents a practical example of Rasch analysis. Whilst no
51 questionnaire is perfect, PROs provide us with a unique reference to the patients'
52 perception of, for example, symptom severity. It is, therefore, vital that these measures
53 are developed and refined using robust psychometric techniques. This paper has
54 demonstrated that Rasch analysis provides a viable option for questionnaire
55 development and refinement. It presents a detailed description of the processes
56
57
58
59
60

1
2
3 involved and provides nurses with a guide to critiquing the robustness of other
4 questionnaires developed using Rasch analysis.
5
6

7 Words = 4,356
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Review Copy

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

FEV₁ forced expired volume in one second

FVC: forced vital capacity

FEV₁: 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

Review Copy

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

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

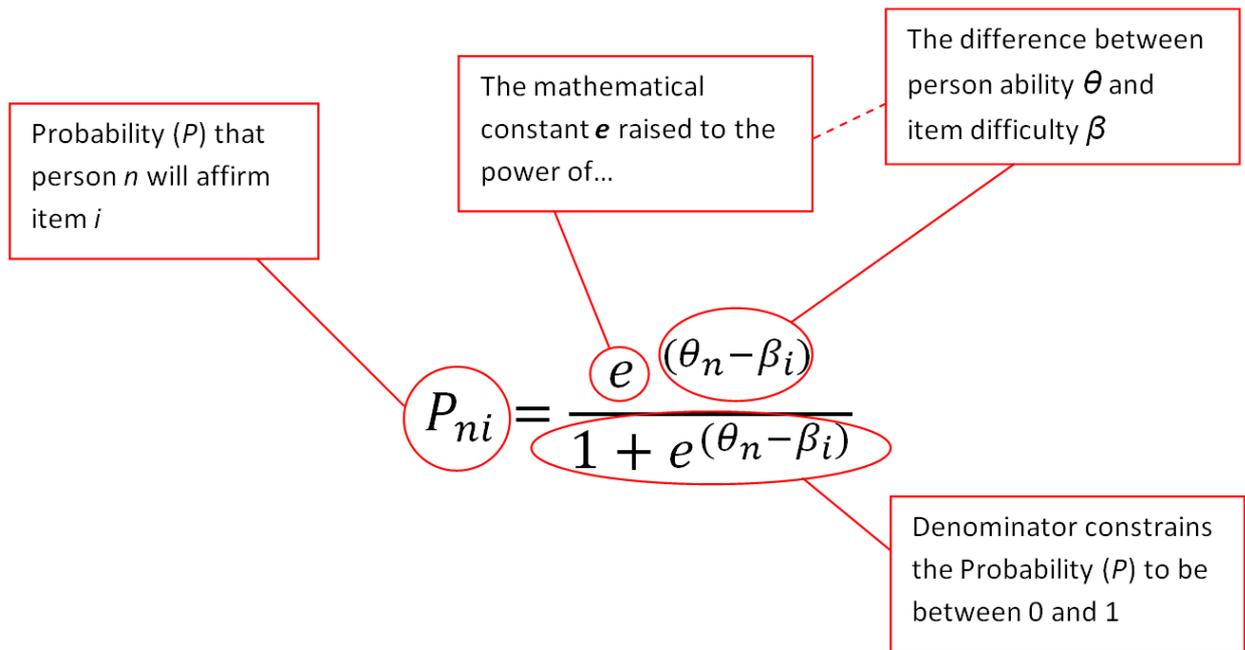


Figure 1: Rasch equation

Review Copy

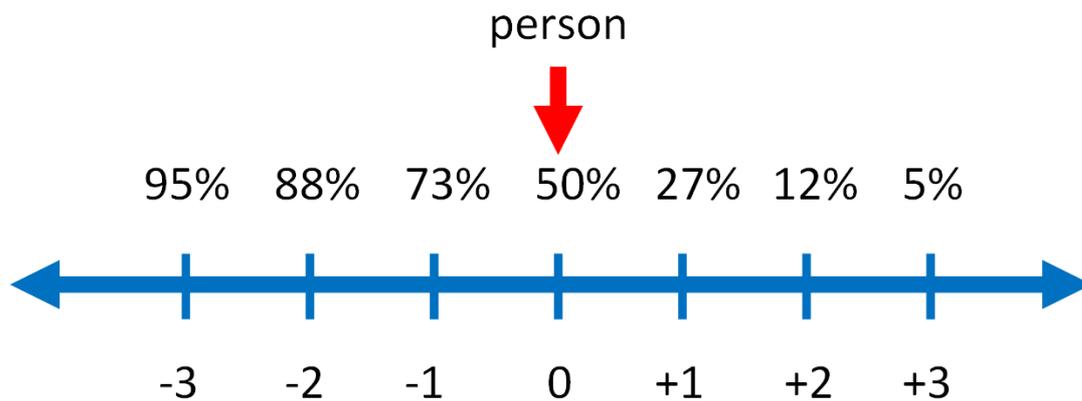


Figure 2: Probability of a person affirming an item.

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%.

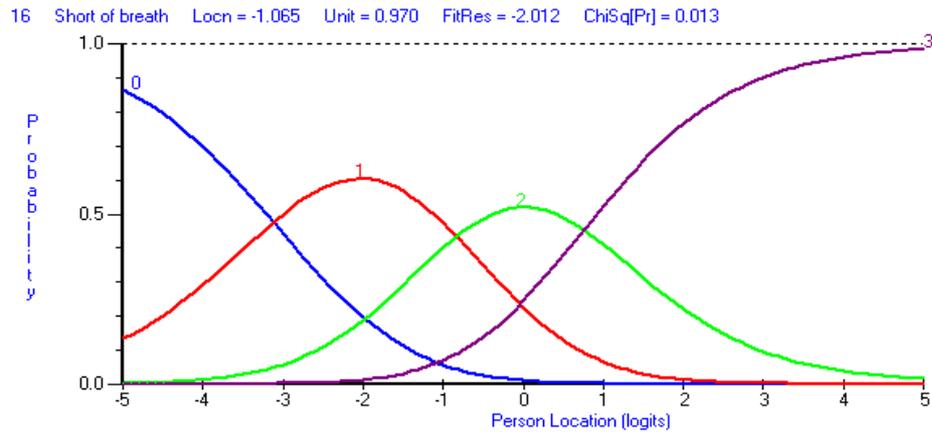


Figure 3: Example of well-ordered transition between categories for the item 'I feel short of breath'. The 'y' axis represents the likelihood of a given response and the 'x' axis represents patient severity. It can be seen that the responses to this item fall in a logical progressive order – category 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.

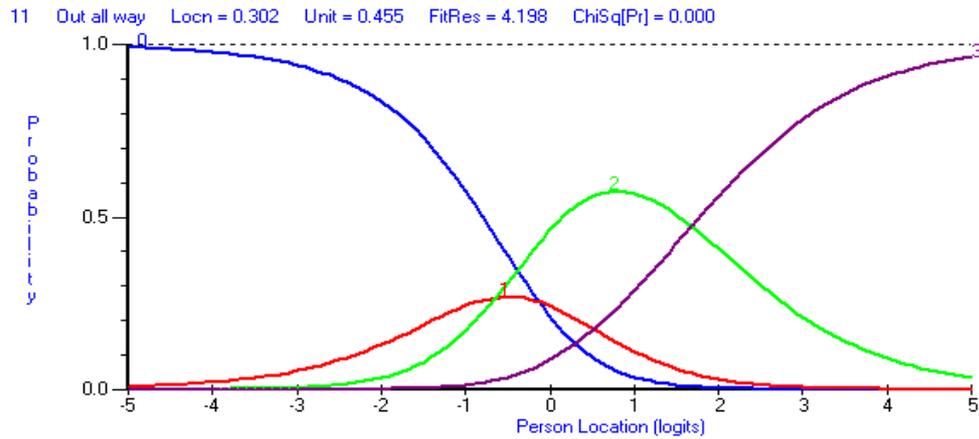


Figure 4: Example of disordered threshold for the item 'My breath does not go out all the way'. At no patient severity is it most likely that category 1 ('mild') will be scored. That is, even at the point where the probability of scoring a 1 is at its peak, it is still more likely that category 0 or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

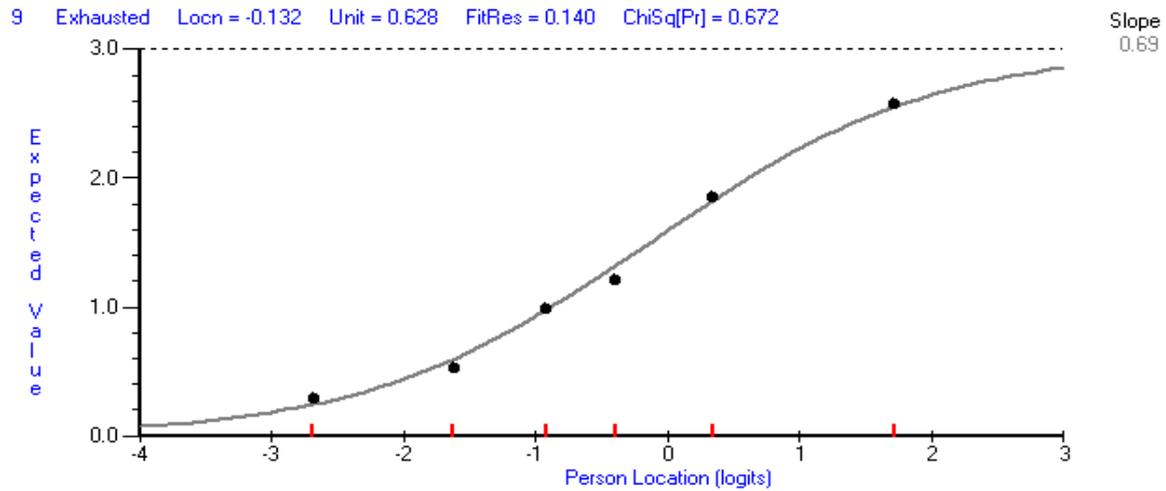


Figure 5: Item Characteristic Curve for a well-fitting item- 'My breathing is exhausting'. The 'y' axis represents the item severity and the 'x' axis represents patient severity in logits. The curved line represents the expected scores for the item, and the dots represent the observed scores for the 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.

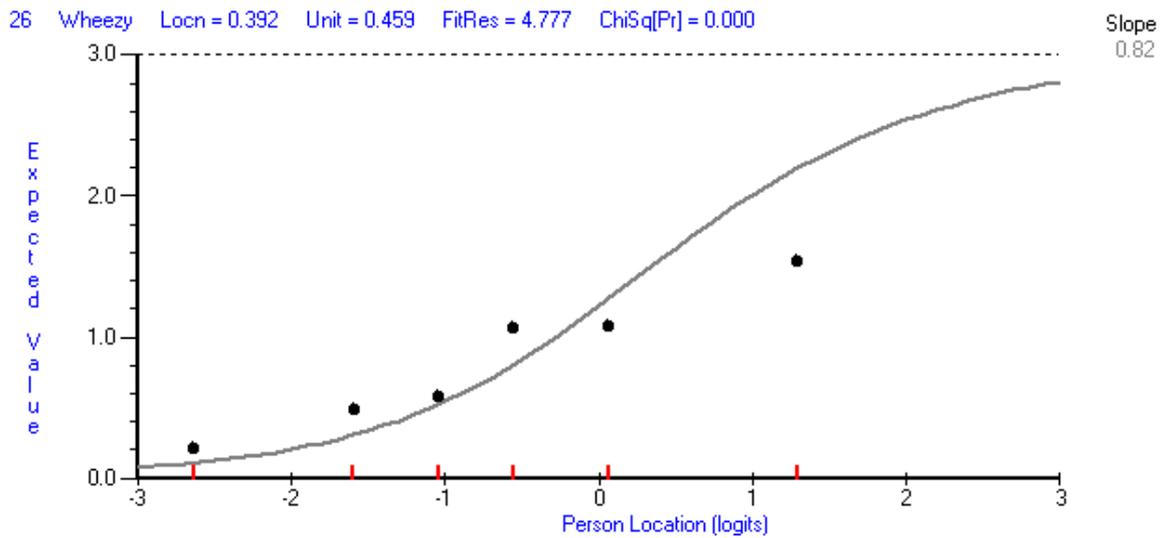


Figure 6: Item Characteristic Curve for a non-fitting item - 'I feel wheezy'. This item is under-discriminating – the observed scores (black dots) form a flatter curve than the expected scores. The fit residual is 4.77 and the Chi-Square is significant ($p < 0.001$). This item was consequently removed.

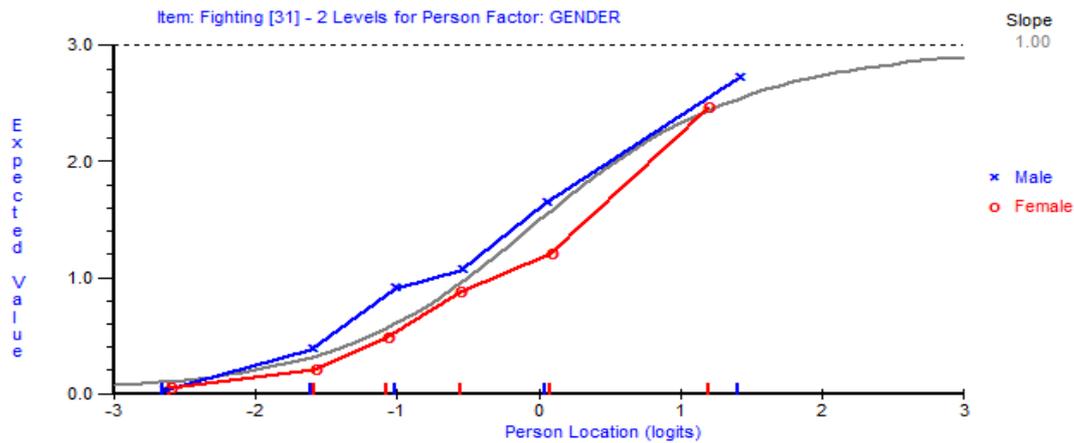


Figure 7: Example of an item – 'I am fighting for breath' demonstrating gender associated uniform DIF. It can be seen that the female group (red line) is slightly but consistently below the male group (blue line). This means that for any level of overall breathlessness severity, females had a lower (i.e.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

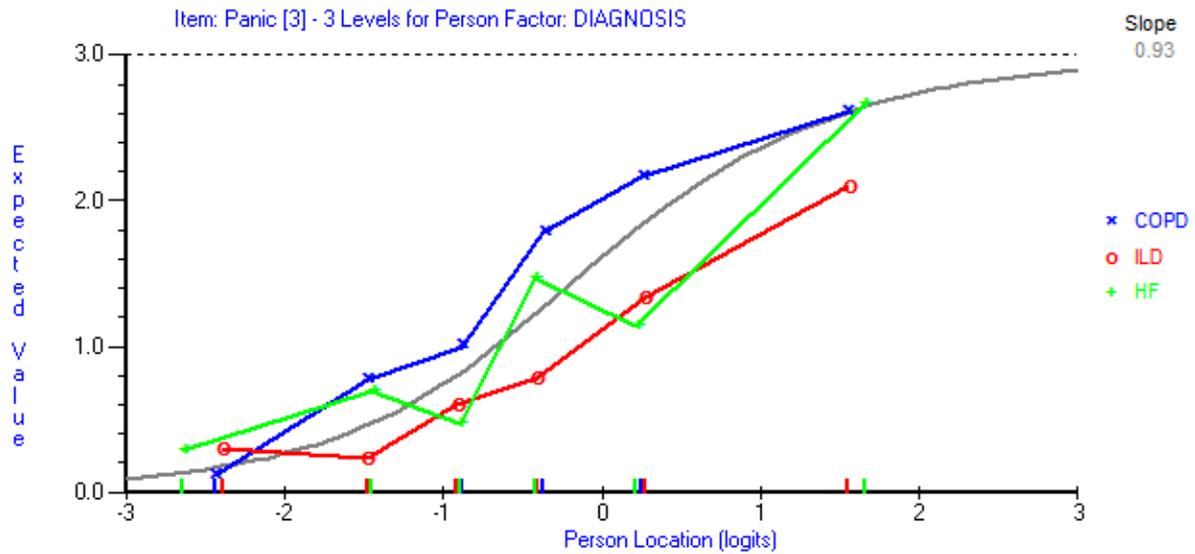


Figure 8: Example of an item – ‘My breathing makes me panic’ demonstrating uniform and non-uniform DIF. At any given level of breathlessness severity, patients with COPD consistently score higher than patients with ILD on this item. Patients with CHF represent the most erratic responses for this item,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

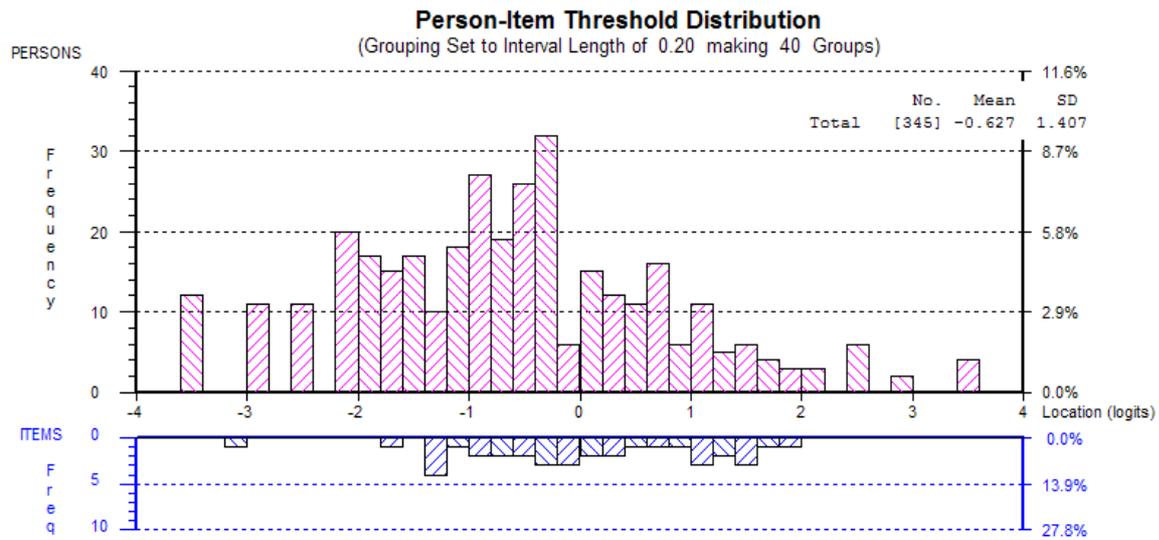


Figure 9: Distribution of patients and item thresholds based on Rasch logit. This figure shows the distributions of patient severity and item severity (locations) along the same linear scale measured in logits. Patients are on the upper part of the graph (pink boxes) and item locations on the lower part (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.

- 1
2
3 Andrich D., Lyne A., Sheridan B., Luo G (2003). RUMM2020. *RUMM Laboratory*,
4 Perth, Australia.
5
6
7 Andrich D. (2004) Controversy and the Rasch model: a characteristic of incompatible
8 paradigms? *Medical Care* 42, I7-I16
9
10
11 Borsboom, D. (2005) *Measuring the Mind: Conceptual issues in contemporary*
12 *psychophysics*. Cambridge, Cambridge University Press.
13
14 DeVellis R.F. (2006) *Scale Development: theory and applications* (2nd Edition), Sage,
15 Thousand Oaks.
16
17
18 Gilworth G., Bhakta B., Eyres S., Carey A., Chamberlain M.A. & Tennant A. (2007).
19 Keeping nurses working: development and psychometric testing of the Nurse-Work
20 Instability Scale (Nurse-WIS). *Journal of Advanced Nursing* 57(5): 543-551.
21
22
23 Guttman L. (1944) A basis for scaling qualitative data. *American Sociological Review*
24 9:139–150
25
26
27 Hagquist C, Bruce M, Gustavsson JP. (2009) Using the Rasch model in nursing
28 research: an introduction and illustrative example. *International Journal of Nursing*
29 *Studies*. 46(3), 380-393.
30
31
32 Hobart J, Cano S. Improving the evaluation of therapeutic interventions in multiple
33 sclerosis: the role of new psychometric methods. *Health Technology Assessment*
34 2009; Vol. 13: No. 12.
35
36
37 Jones P.W., Harding G., Berry P., Wiklund I., Chen W-H., Kline Leidy N. (2009)
38 *Development and first validation of the COPD Assessment Test*. European
39 *Respiratory Journal* 34, 648-654
40
41
42 Nunnally J. (1978). *Psychometric Theory*. New York, Mc Graw-Hill.
43
44
45 Pallant J.F. & Tennant A. (2007). An introduction to the Rasch measurement model:
46 an example using the Hospital Anxiety and Depression Scale (HADS). *British Journal*
47 *of Clinical Psychology* 46(Pt 1): 1-18.
48
49
50 Pallant J.F., Miller R.L., Tennant A. (2006) Evaluation of the Edinburgh Post Natal
51 Depression Scale using Rasch analysis. *BMC Psychiatry* 6-28.
52
53
54 Rasch G. (1960). Probabilistic models for some intelligence and attainment tests.
55 *Danish Institute of Educational Research*.
56
57 Rattray J. & Jones M.C. (2007). Essential elements of questionnaire design and
58 development. *Journal of Clinical Nursing* 16: 234-243.
59
60

1
2
3 Smith E.V. (2002). Detecting and evaluating the impact of multidimensionality using
4 item fit statistics and principal components analysis of residuals. *Journal of Allied*
5 *Measurement* 3: 205-231.
6
7

8 Tennant A. & Conaghan P.G. (2007). The Rasch measurement model in
9 rheumatology: what is it and why use it? When should it be applied, and what should
10 one look for in a Rasch paper? *Arthritis & Rheumatism* 57(8): 1358-62.
11

12 Tesio L, Simone A, Bernardinello M. (2007) Rehabilitation and outcome measurement:
13 where is Rasch analysis-going? *Europa Medicophysphysica* 43(3), 417-26.
14
15

16 van Alphen A., Halfens R., Hasman A., Imbos T. (1994). Likert or Rasch? Nothing is
17 more applicable than good theory. *Journal of Advanced Nursing* 20(1): 196-201.
18
19

20 Watson R. & Thompson D.R. (2006). Use of factor analysis in Journal of Advanced
21 Nursing: literature review. *Journal of Advanced Nursing* 55: 330-341.
22

23 Wilson, M. (2005) *Constructing Measures: An Item Response Modelling Approach*.
24 Routledge Academic, New York.
25
26

27 Wilson R.C. and Jones P.W. (1991) Differentiation between the intensity of
28 breathlessness and the distress it evokes in normal subjects during exercise. *Clinical*
29 *Science* 80(1): 65-70.
30

31 Yorke J. (2008). Dyspnoea and pain: an interesting analogy. *Journal of Clinical*
32 *Nursing* 17(7), 841-842.
33
34

35 Yorke J., Moosavi S.M., Shuldham C., Jones P.W. (2010). Quantification of dyspnoea
36 using descriptors: development and initial testing of the Dyspnoea-12. *Thorax* 65(1):
37 21-26.
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60