Using Rasch analysis to aid the construction of preference based measures from existing quality of life instruments.

Tracey Young, Yaling Yang, John Brazier, Aki Tsuchiya,
HEDS, ScHARR, University of Sheffield, S1 4DA

Health Economics Study Group Meeting
29th June to 1st July 2005
University of Newcastle

Corresponding author:
Tracey Young
Health Economics and Decision Science (HEDS)
ScHARR, University of Sheffield
Regent Court, 30 Regent Street,
Sheffield S1 4DA
Tel: 0114 2220837
E-mail: **T.A.Young@sheffield.ac.uk**

Acknowledgements

This study is funded by Novartis. The “Trial” dataset was provided by Novartis. The “Observation” dataset was collected by the COGENT project at the University of Newcastle, funded by the NHS programme ‘Methods to promote the uptake of research findings. The usual disclaimer applies.

Introduction

The Asthma Quality of Life Questionnaire (AQLQ) has been designed to assess health related quality of life in patients with asthma. It has been shown to be reliable and valid for use in this population [Juniper et al, 1993, 1999], but it cannot be used in cost utility analyses because the scores are not preference-based.

To derive a preference-based single index measure from the AQLQ for use in calculating QALYs, we planned to apply the hypothetical health state valuation method, which has been successfully used on the SF-36 by Brazier et al [2002]. The first stage was to derive a health state classification system from the AQLQ that is amenable to valuation using a preference elicitation technique. The second stage is a valuation survey of a selection of states defined by the new classification system by a sample of the UK general population. The third stage is to estimate a range of econometric models for predicting the health state values for all states defined by the new classification system, which in turn will enable the calculation of QALYs based on AQLQ data.

This paper sets out a structured approach of the process of selecting a small number of AQLQ items for establishing a classification system to use in an evaluation survey using Rasch analysis.

Rasch analysis is a mathematical modelling technique [Rasch G, 1960], commonly used in education in the development and validation of examination papers [Willmott & Fowless, 1974]. It is increasingly being used in health related quality of life studies in the development of new quality of life questionnaires [see for example Duncan *et al*, 2003; Gilworth *et al*, 2004; Pesudovs *et al*, 2004] and in the validation of existing questionnaires [e.g. Raczeck *et* al, 1998; Valders *et* al, 2004; Wiren *et* al, 2000]. However, to date Rasch analysis has not been used in the development of a preference based utility index.

Therefore, this paper sets out to use Rasch analysis alongside conventional psychometric methods firstly to select items for a preference based utility index and secondly to reduce the existing number of levels of the AQLQ from seven to a more manageable number for the valuation survey to be conduced in stage two. The use of Rasch analysis, as set out in this paper, in the development of a preference based measure, is not a definitive one, but a solution set out by the authors in an attempt to formalise the selection of items for a preference based measure.

In what follows, we will first give a brief description of the AQLQ instrument, Rasch analysis, and the psychometric tests used. This will be followed by a description of the data used, the results of the analysis, and a discussion.

The AQLQ

The AQLQ consists of 32 items that ask a series of questions across four domains: symptoms (12 items), activity limitations (11 items), emotional function (5 items) and environmental stimuli (4 items). Table 1 gives a description of each of the 32 items in the AQLQ and which domain each question belongs to. A score can be derived for each of the four domains or a combined quality of life score can be obtained across all 32 items. For each item the respondent is asked to choose from a series of seven levels ranging from extreme problems (score 1) to no problems (score 7). Scores are summed and averaged over items to obtain a domain or overall score – where a higher score indicates better quality of life, but neither scoring system is preference based.

Two versions of the AQLQ exist which are virtually identical, with the exception of the first five items: the "individualised" version of the AQLQ asks respondents to chose five from a list of 27 activities and then state how each of these chosen activities has been limited by asthma within the previous 2 weeks, whereas the "standardised" version of the questionnaire asks how asthma has limited strenuous, moderate, social, work related activities or sleeping.

Rasch analysis

This section sets out the concept behind Rasch analysis and the formal process of checking model goodness of fit. Rasch analysis is a mathematical technique that converts qualitative (categorical) responses to a continuous (unmeasured) latent scale using a logit model, and can be conceptualised as "a statistical approach to the measure of human performance, attitudes and perceptions" [Tesio, 2003]. In quality of life each categorical item can be mapped to the continuous latent scale, where the scale is conceived to be a continuous measure of quality of life.

In education a person's position on the underlying latent scale in Rasch analysis is measured from the responder's ability to answer questions and the difficulty of the items set within an examination or test. For quality of life each responders position on the latent scale accounts for a persons disease severity and item difficulty in terms of how the item relates to severity. For example, a responder indicating that they are frequently short of breath (item 8) is more likely to have severe asthma symptoms than someone indicating little problems for this items but some difficulty with distress or discomfort due to coughing (item 12).

The Rasch model used in this paper is known as the Rasch rating scale model and is fitted to allow for multi level responses to all items, as is the case with the AQLQ. The Rasch model assumes that item and person responses are independent variables that may be estimated separately. The mathematic formula for the model is set out below:

$$\ln\left(\frac{p_{nij}}{1-p_{nij}}\right) = f(q_n, d_i, t_j)$$

where p_{nij} is the probability of person n with asthma severity q_n responding to item i with item difficulty d_i and at level j with level difficulty t_j.

The overall goodness of fit of the Rasch model is measured in terms of item-trait interaction, the person separation index and the person and item fit residuals.

The item-trait interaction measures whether data fit the Rasch model for discrete groups of responders. The groups are selected by dividing the responders into a series of subgroups based on where each responder lies on the latent scale of the Rasch model. Thus responders who tend to have similar quality of life for an AQLQ domain will be grouped together. Observed and expected responses are compared across items and traits and can be summarised using the chi-squared test statistic. A good fitting Rasch model has a p-value for the chi-squared goodness of fit for n degrees of freedom which is greater than 0.01, indicating that there is no significant deviation between the observed and expected responses.

The ***person separation index*** lies on a scale of 0 to 1 and is a measure of how each of the AQLQ domains discriminates between respondents. A person separation index of 0.8 or more indicates a well fitting Rasch model.

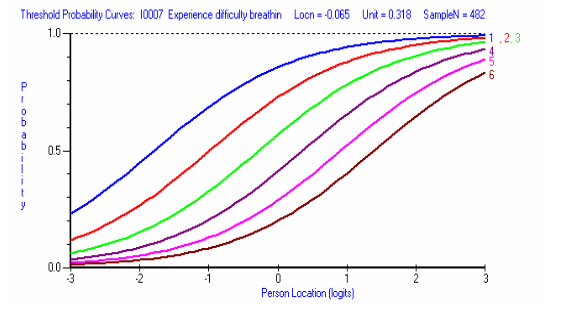
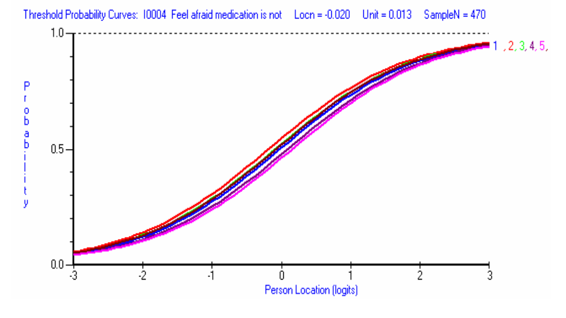
Fit residuals estimate the amount of divergence between the expected and observed responses for each person-item response, summed over all items (item fit residual) or summed over all persons (person fit residuals). The residuals are standardised to approximate the Z-score and therefore the mean item or person fit residual should be approximately zero with standard deviation approximately equal to 1.

In this paper we assume that the underlying latent scale can be used both as a measure of quality of life and as a preference based measure of utilities and use Rasch analysis at three stages in the item selection process: 1) Eliminate items which do not fit the Rasch model or do not meet selection criteria for a set of five psychometric tests, 2) to select items for a preference based measure, 3) to collapse the number of item levels from seven to a more manageable number.

1. Eliminating AQLQ items

The first step in the Rasch analysis is to fit the Rasch model to the AQLQ domains and make sure that responders to the AQLQ are able to distinguish between different level choices for items in the questionnaire. As an illustrative example, if the latent scale for each level of an item is ordered then respondents are able to distinguish between levels (Figure 1a), however if the latent scale is unordered then responders have difficultly distinguishing between item levels (Figure 1b). In Figure 1 the x-axis depicts the underlying latent (logit) scale and the y-axis the probability of being in a particular item level and thus each curve shows the probability of being in a particular level across the latent scale. If curves are unordered adjacent item levels should be merged and the Rasch model refitted using the merged levels.

Figure 1: Illustration of the ordering of threshold curves for two AQLQ items

1a) Item 7: Feel concerned about having asthma	1b) Item 21: Feel afraid of not having your asthma medication available
	

Ordered curves: colour sequence highest to lowest – blue, red, green, purple, pink, brown (levels 1, 2, 3, 4, 5, 6 for item 7) Unordered curve: colour sequence highest to lowest – Red, green, brown, blue, purple, pink, (levels 2, 3, 6, 1, 4, 5 for item 21) NB: Only six curves are shown as level 7 – no problems/limitations acts as a comparator (the probability of being in level 1 to 6 in comparison with level 7)

The selection of the appropriate levels to merge is left to the analyst, a summary of the frequency of responses per level can aid this choice as can eyeballing the threshold probability curves (a plot of the probability of being in each item level across the latent scale) and merging levels where curves lie close together.

Levels may be collapsed generically, across all items in a questionnaire, regardless of the ordering across all items, by questionnaire domain, or individually item by item. The choice is left to the analyst and may be dependent on the number of items being examined (if a questionnaire contained a large number of items it may be preferable to collapse generically across the whole questionnaire). In this paper items are collapsed at the individual level.

After ordering has been achieved across all items the goodness of fit statistics for the Rasch model are examined. If the overall item-trait fit of the model, measured using the chi-squared statistic, had a p-value of < 0.01 the fit of each of the individual items included in the Rasch models were examined. As with the overall model fit, items with a chi-squared probability of < 0.01 are said to not fit the Rasch model. The item with the highest chi-squared value (p-value closest to zero) is removed and the model refitted and the overall goodness of fit statistics examined for the new model. The process repeated until only well fitting items remain and the overall item-trait goodness of fit of the model is greater than 0.01.

Once the model fit is satisfied the analyst is left with a sub-sample of items for each of the AQLQ domains and the process of selecting items for a preference based utility measure can begin.

Psychometric methods

Alongside the Rasch analysis the performance of the AQLQ was tested across all 32 items using five conventional psychometric tests, where the tests were used as a method for identifying and eliminating items that did not meet specific criteria: missing data, correlation between scores at different time points, distribution of responses, responsiveness between baseline and follow-up visits and regression between general health and item responses. These five criteria were chosen to represent conventional psychometric criteria for assessing instruments and the cut-off level chosen for each criterion is arbitrary, unless otherwise stated. The five methods are described below:

1. The prevalence of missing data

High prevalence of missing data reduces the usefulness of an item. Therefore items where the percentage of missing data was greater than 1% were excluded.

2. The correlation of an item score and its domain score

Within each domain, we tested the relationship between individual items and the domain score. The assumption was that: if certain items within a domain do not have good correlation with the domain score, these items cannot be said to be representative of the domain. 0.65 was set as the threshold correlation coefficient.

3. Distribution of responses across the 7 response levels

Since the reduced instrument for valuation will be covering the spectrum of asthma with a smaller number of items, ideal items are those that utilise the whole range of 7 response levels rather than just a few of them. Thus, for each of the 32 AQLQ items, the distribution of respondents' answers was examined across the seven levels of response. If the proportion of responders in the extreme levels (i.e. level 1 or level 7) was over 20%, indicating that the scale might suffer from floor or ceiling effects, then the item was rejected.

4. The responsiveness at two time points

Responsiveness is generally defined as the ability of an instrument to detect accurately a change when it has occurred, and thus, it is a test to be used for time series data. Of the several possible ways to calculate this, effect size (ES) was used, defined as:

$$ES = \text{Average } (x_f - x_b) / \text{Standard Deviation } (x_f - x_b)$$

where x represents the item score and subscripts f and b represent follow-up and baseline data respectively. Where, on average, respondent's scores improved between baseline and follow-up. There is no gold standard for good or acceptable levels of responsiveness, but ES > 0.2 was used for an acceptable level since it was below the criterion for a "small" ES established by Cohen [1978].

5. Regression coefficients between a general health indicator and the item

Items used in the new instrument will on the one hand present specific domains of asthma related quality of life, and on the other hand represent a component of a more general concept of health related quality of life. In other words, an ideal item will be correlated with some measure of overall health. This was pursued by regression analysis between the general health question of the SF-36 (as the dependent variable) and the individual AQLQ items (as categorical explanatory variables). Technically, and strictly speaking, the use of an SF-36 item as a continuous dependent variable is problematic, but this was intended as an exploration of the importance of each item with respect to the patients' overall health. A very low R^2 would suggest that the item has little to do with the patient's perception of their health. The criterion was set at $R^2 = 0.15$.

2. Item Selection

Any item which did not meet the above five criteria and had not been deselected within the Rasch analysis was excluded from further consideration at this stage. This left a number of items for each domain that were suitable for selection in a utility index. The aim was to select one item from each of the four domains of the AQLQ and, given that the AQLQ also asked four questions about sleep (Items 5 (standardised version only), 20, 24 and 29), an additional item was selected for a fifth domain asking about how asthma affected sleep.

Items were then selected for the preference based measure by domain based predominantly on the spread of levels per item across the latent scale (obtained from the Rasch analysis), their position on the latent scale in relation to the other chosen items (obtained from the Rasch analysis) and the five psychometric criteria described above.

Given that the logit scale in Rasch analysis is centred around zero (see the horizontal axis in the figures above) the threshold probability curves for the different levels should be spread evenly across the threshold space – where the greater the spread the more likely the respondent was to distinguish between item levels.

3. Collapsing levels

Having selecting five items for the preference based measure the authors felt that it would be possible to collapse the number of response levels, per selected item, from seven to a smaller number and felt that five would be a more manageable choice of levels for responders to choose from. Threshold probability curves were examined and item levels which were felt to be closer together in comparison with other levels were merged.

RUMM2020© was used to fit the Rasch models [RUMM Laboratory Pty Ltd, 2004].

Data sets

Two data sets of asthma responders to the AQLQ were used in this study. The “Trial” data comes from a multinational trial, and included patients with severe persistent asthma. Patients filled in the individualized version of the AQLQ on several occasions throughout the trial period. For this paper, patients from the treatment and the placebo arms are not distinguished, and unless otherwise stated, analyses are based on 482 “baseline” observations of this Trial data. However, there were three exceptions.

Firstly, since the responsiveness analysis (psychometric analysis no. 4 above) needs more than one observation, the baseline and follow-up information from the Trial data were used. Furthermore, in order to carry out the regression analysis (psychometric analysis no. 5), information from the first round of the Observation data was used, since it included the SF-36 questionnaire alongside the AQLQ questionnaire.

Lastly, three Rasch analyses were conducted for the symptoms, environmental function and environmental stimuli domains, but regarding the activity domain, given that the first five questions of the individualised version of the questionnaire depend on the responder's choice of activities, it was felt to be inappropriate to include the responses to these questions in the item selection process. Therefore, in order to supplement for this, the "Observation" dataset was introduced. This comes from a UK trial of computerised decision support (and thus observational in the therapeutic sense), with 3000 patients covering a wide range of asthma severities (for further details, see Eccles et al, 2000). The standardised version of the AQLQ was used at the second round of observations, consisting of 2,119 cases. To date no paper has been published on sample size calculations for Rasch analysis [Personal communication Alan Tennant, University of Leeds, 2005], but the current consensus among Rasch experts is a sample size of 300 to 400 cases is adequate where anything greater could results in over fitting of the Rasch model. Therefore a total of 413 cases were randomly selected from the second round of the Observation dataset to carry out the Rasch analysis for the activity domain.

Results

Using Rasch analysis, examination of items on the latent scale across the seven levels showed that respondents were unable to distinguish between levels for items: 30 (symptoms), 4 (activity), 11 (activity), 25 (activity), 28 (activity), 21 (emotion), 9 (environment) and 17 (environment). Items were collapsed individually and ordering of levels was achieved by collapsing the two mildest levels (none and very little problems) the two most severe levels (problems all or the majority of the time) and the three middle levels across all unordered items, leaving just three levels for each of these items.

To achieve overall Rasch model goodness of fit (measured from the item-trait c^2 statistic p-value > 0.01) the following poorly fitting items (item c^2 statistics p-value < 0.01) were excluded from the Rasch models: items 12, 16 and 30 from the symptoms domain, 1, 5, 11, 19 and 28 from the activity domain, and item 17 from the environment domain. It was unnecessary to exclude items from the emotion domain. Table 2 summarises the over Rasch model statistics for each domain. The person separation index for the environment domain does not reach the recommended 0.8 level, although all other fit statistics are met for the environment domain and for the remaining three domains.

Conventional psychometric tests were then performed in order to exclude further AQLQ items and the results of these tests are presented in Table 3, where items not meeting the test criteria are highlighted in bold. This resulted in a further six items being removed from the selection process: items 3, 4, 9, 20, 21 and 26 and left a total of 17 items for possible inclusion in a preference based questionnaire (8 items from the symptoms domain; 4 from the activity limitations domain; 4 from the emotional functioning domain; 1 from the environmental stimuli domain, where two of the items in the symptoms domain asked about sleep).

Table 4 summarises the statistics used in the selection process of remaining items for a preference based measure. The statistics were taken from the four Rasch analyses performed on the AQLQ domains. Given that only one item remained in the environment domain this item was automatically selected for the preference based measure (Item 23). Two items remained that asked about sleep, item 29 had the better fitting c^2 statistic (lower c^2 value) and had a slightly higher spread of probabilities across levels at logit 0 and was therefore selected. Item 2 in the activity domain was removed as this item differed for the two different versions of the AQLQ questionnaire. Item 32 was selected as the best fitting item with the greatest spread of probabilities across levels at logit 0. Item 8 was selected from the remaining non-sleep related symptoms items due to having the greatest spread of probabilities across levels at logit 0. Finally, item 7 was chosen from the emotional functioning domain because it had the largest spread at logit 0.

A summary of the selected items is presented below:
Item 7: (Emotional Function) In general, how much of the time during the last 2 weeks did you: Feel concerned about having asthma?

Item 8: (Symptoms) In general, how much of the time during the last 2 weeks did you: Feel short of breath as a result of your asthma?

Item 23: (Environmental Stimuli) In general, how much of the time during the last 2 weeks did you: Experience asthma symptoms as a result of the weather or air pollution outside?

Item 29: (Symptoms) In general, how much of the time during the last 2 weeks did you: Has your asthma interfered with getting a good night's sleep?

Item 32: (Activity Limitations) Overall, among all the activities you have done during the last 2 weeks, how limited have you been by your asthma?

It was felt that the existing number of levels for each chosen item (7) was too many for the valuation stage of deriving the preference based index and that a more realistic number of levels would be five. An attempt was made to use Rasch analysis to reduce the number of levels from seven to five by studying threshold probability curves for each of the five chosen items (Figure 2). The levels relating to curves which were close together were collapsed. However various scenarios and combinations of levels failed to produce five level items that fitted within the Rasch assumptions (levels were no longer ordered when collapsed or threshold plots suggested that levels should be collapsed further). Therefore consensus was reached partially based on Rasch results and through agreement amongst the authors which levels should be collapsed: for items 7, 8, 23 and 29 which asked about how asthma effected the responders in terms of time, levels were collapsed as follows: none of the time, a little or hardly any of the time, some of the time, most of the time, and all of the time. The wording for levels item 32 related to limitations and the five levels were chosen as follows: not at all limited, a little limitation, moderate or some limitation, extremely or very limited, and totally limited.

Discussion

This study has shown how Rasch analysis can be used alongside conventional psychometric methods in the selection process of items for a preference based measure. After the initial stage, where poorly performing items based upon Rasch model or psychometric selection criteria were excluded, a total of 17 possible items representing all four AQLQ domains and including two questions on sleep were available from which to choose items for a preference based measure. Rasch analysis was then used in the second stage to select what were felt to be the best performing items per domain and finally, the number of item levels was reduced from seven to five, although this process was based in part on author opinion rather than Rasch analysis results.

The first stage excluded items that did not meet either Rasch or psychometric criteria and suggests that Rasch analysis can be used alongside psychometric methods as neither method alone failed to select all of the fifteen items which were initially rejected as part of the selection process of items. Seven items were identified by both Rasch and psychometric techniques and only two of the items (1 and 2) identified in the Rasch analysis were not identified by conventional methods. There are no obvious reasons why conventional methods failed to identify these two items, though this could be due to the arbitrary exclusion criteria chosen. The proportion of common items (47%) excluded by the two approaches could be unique to this data set.

We wished to select the best performing items from each domain for inclusion in a preference based measure and at the item selection stage there was sometimes little to distinguish between items for a particular domain. In this analysis items were selected based primarily upon the spread of item levels at logit 0 in the threshold probability space, where respondents were thought to be making use of the full range of possible responses. At the time of the analysis we felt that this was the most appropriate approach to use, however, other approaches could be equally applicable. For example, choosing the best fitting items according to the item c^2 statistic, had this approach been chosen a slightly different set of item would have been chosen: 15, 18, 23, 29, and 32 – excluding items 1 to 5 from this process. No guidelines currently exist on this process and it remains the choice of the researchers how they choose to select items and we have used Rasch analysis in an attempt to formalise this selection process.

The role of Rasch analysis in the collapsing of the number of levels to a preconceived target of 5 levels proved unsuccessful. However, this might have been because when the five items were chosen they were based on the spread of levels across the latent space at logit 0, and the items chosen tended to have evenly spread levels. Therefore, any attempt to reduce these levels resulted in violation of Rasch assumptions, as set out previously, and poor fitting Rasch models.

The choice of Rasch analysis and the selection of cases to be included in the Rasch analysis could also be varied. A separate Rasch analysis for sleep related items was not included here, but instead these items were selected from the Rasch analysis on the two domains (symptoms and activities from which sleep related items were included). Additionally, we could have selected a mixture of baseline and follow-up data from the Trial data set and could also have run a separate Rasch analysis on a selected number of cases from the Observation data. Performing a number of Rasch analyses could aid in validating the selection process and in future analysis should be built into the item selection process.

In addition to its role in the selection of items for a preference based measure Rasch analysis has inadvertently been used as a validation tool for the AQLQ by checking the assumption that the items selected for each domain are appropriately selected as a measure of quality of life for a particular domain in addition to checking the appropriateness of using seven level to discriminate between states (for 8 items, 3 levels were found to be a more appropriate choice). Analysis not presented here showed that if all items were collapsed generically ordering of levels was achieved across all items using three levels (merging scores 1&2, 3, 4&5 and 6&7).

We have made the assumption here that the underlying latent scale in the Rasch models has the ability to measure both preference based utilities and quality of life on the latent scale. We would be interested in the HESG audiences thoughts on this merits of this assumption.

Conclusions

Rasch analysis has proved to be a useful in assisting in the initial process of selecting items from an existing quality of life instrument in the construction of a preference based measure. This method is recommended for use alongside traditional psychometric testing to aid the selection of items in the development of preference based measures.

References

Brazier,J., Roberts,J., Deverill,M. (2002), The estimation of a preference-based measure of health from the SF-36, Journal of Health Economics, 21:271-292

Cohen J (1978) Statistical power analysis for the behavioural sciences. New York Academic Press

Duncan P.W. Bode R.K. Min Lai S *et al* (2003) Rasch analysis of a new stroke-specific outcome scale: the Stroke Impact Scale. Archives of Physical Medicine and Rehabilitation 84(7): 950-63

Eccles M, Grimshaw J, Steen N, Parkin D, Purves I, McColl E, et al. (2000) The design and analysis of a randomized controlled trial to evaluate computerized decision support in primary care: the COGENT study. Family Practice 17(2):180-186

Gilworth G. Chamberlain M.A. Bhakta B. *et al* (2004) Development of the BD-QoL: a quality of life measure specific to Behcet's disease. Journal of Rheumatology 31(5): 931-7

Juniper EF, Guyatt GH, Ferrie PJ, Griffith LE. Measuring quality of life in asthma. American Review of Respiratory Disease 1993;147:832-838

Juniper EF, Buist AS, Cox FM, Ferrie PJ, King DR. Validation of a standardized version of the Asthma Quality of Life Questionnaire. Chest 1999;115(5):1265-1270

Pesudovs K. Garamendi E Elliott D.B. (2004) The quality of life impact of refractive correction (QIRC) questionnaire: development and validation. Optometry and Vision Science 81(10): 769-77

Raczek A.E. Ware J.E. Bjorner J.B. *et al* (1998) Comparison of Rasch and summated rating scales constructed from SF-36 physical functioning items in seven countries: Results from the IQOLA project. Journal of Clinical Epidemiology 15(11): 1203-14

Rasch G. (1960) Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press (Reprinted 1980)

Rasch Unidimensional Measurement Models (RUMM) 2020 ©. RUMM Laboratory Pty Ltd 1997-2004

Tesio L. (2003) Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research. Journal of Rehabilitation Medicine 35: 105-15

Willmott A.S. Fowles D.E. (1974) The objective interpretation of test performance: The Rasch model applied. Windsor NFER

Wiren L. Whalley D. McKenna S. Wilhelmsen L (2000) Application of a disease-specific, quality-of-life measure (QoL-AGHDA) in growth hormone-deficient adults and a random population sample in Sweden: validation of the measure by Rasch analysis. Clinical Endocrinology 52(2): 143-52

Valderas J.M. Alonso J. Prieto L. (2004) Content-based interpretation aids for health-related quality of life measures in clinical practice. An example for the visual function index (VF-14). Quality of Life Research 13(1): 35-44

Table 1: AQLQ items (Standardized)

Item N^{o}.	Question (during the last 2 weeks) as a result of your asthma	Domain	Wording
1	Limited strenuous activities	Activity	Limitations
2	Limited moderate activities	Activity	Limitations
3	Limited social activities	Activity	Limitations
4	Limited work-related activities	Activity	Limitations
5	Limited sleeping	Activity	Limitations
6	How much discomfort or distress as a result of chest tightness	Symptoms	Quantity
7	Feel concerned about having asthma	Emotional	Time
8	Feel short of breath as a result of your asthma	Symptoms	Time
9	Experience asthma symptoms as a result of being exposed to cigarette smoke	Environment	Time
10	Experience a wheeze in your chest	Symptoms	Time
11	Feel you had to avoid a situation or environment because of cigarette smoke	Activity	Time
12	How much discomfort or distress have you felt as a result of coughing	Symptoms	Quantity
13	Feel frustrated as a result of your asthma	Emotional	Time
14	Experience a feeling of chest heaviness	Symptoms	Time
15	Feel concerned about the need to use medication for your asthma	Emotional	Time
16	Feel the need to clear your throat	Symptoms	Time
17	Experience asthma symptoms as a result of being exposed to dust	Environment	Time
18	Experience difficulty breathing out as a result of your asthma	Symptoms	Time
19	Feel you had to avoid a situation or environment because of dust	Activity	Time
20	Wake up in the morning with asthma symptoms	Symptoms	Time
21	Feel afraid of not having your asthma medication available	Emotional	Time
22	Feel bothered by heavy breathing	Symptoms	Time
23	Experience asthma symptoms as a result of the weather or air pollution outside	Environment	Time
24	Were you woken at night by your asthma	Symptoms	Time
25	Avoid or limit going outside because of the weather or air pollution	Activity	Time
26	Experience asthma symptoms as a result of being exposed to strong smells or perfume	Environment	Time
27	Feel afraid of getting out of breath	Emotional	Time
28	Feel you had to avoid a situation of environment because of strong smells or perfume	Activity	Time
29	Has your asthma interfered with a good night's sleep	Symptoms	Time
30	Have a feeling of fighting for air	Symptoms	Time
31	How much has your range of activities you would like to have done been limited by your asthma	Activity	Limitations
32	Among all the activities you have done how limited have you been by your asthma	Activity	Limitations

Table 2: Summary of Rasch goodness of fit statistics for the four AQLQ domains

	Items in model	Overall χ^2	DF*	P-value	Mean item fit (SD)	Mean person fit (SD)	Person separation index
Symptoms	6, 8, 10, 14, 18, ***20,*** 22, ***24, 29***	87.04	63	0.024	0.12 (1.59)	-0.53 (1.48)	0.923
Activity	2, 3, 4$_{(3)}$, 25$_{(3)}$, 31, 32	38.11	36	0.374	-0.55 (1.87)	-0.43 (1.09)	0.945
Emotion	7, 13, 15, 21$_{(3)}$, 27	37.12	35	0.372	0.16 (1.51)	-0.48 (1.31)	0.845
Environment	9$_{(3)}$, 23, 26	28.41	21	0.129	0.60 (1.33)	-0.38 (1.11)	0.739

* DF = Degrees of Freedom for χ^2 test $_{(3)}$ = Denotes items where number of levels was collapse to three to ensure ordering of levels Items in ***bold italics*** ask questions about sleep

Table 3: Overall summary of five psychometric tests (failed items are in bold)

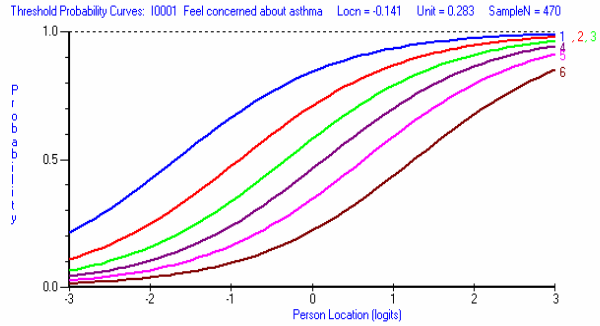
test	Proportion at level 1	Proportion at level 7	Effect Size	Missing data	Correlation with domain score	Regression with general health
threshold	(>=20%)	(>=20%)	(<=0.2)	(>=1.0%)	(<=0.6)	(R^2<=0.15)
item 1	6.0	3.5	0.5	0.4	0.73	0.22
item 2	7.3	3.6	0.5	0.8	0.74	0.20
item 3	6.9	2.5	0.5	**1.2**	0.71	0.21
item 4	6.7	3.6	0.5	**1.0**	0.74	0.23
item 5	5.5	4.2	0.4	**1.9**	0.71	0.25
item 6	6.2	5.2	0.4	0.2	0.74	0.25
item 7	10.0	7.3	0.4	0.2	0.74	0.22
item 8	6.0	2.3	0.5	0.2	0.77	0.28
item 9	18.6	14.6	0.3	0.6	0.70	**0.08**
item 10	6.5	5.4	0.4	0.4	0.76	0.23
item 11	**26.8**	12.5	0.4	0.2	**0.60**	**0.06**
item 12	7.1	8.1	0.3	0.2	0.70	0.21
item 13	11.2	10.0	0.4	0.0	0.79	0.24
item 14	4.0	9.2	0.4	0.4	0.77	0.23
item 15	11.8	16.8	0.3	0.0	0.75	0.17
item 16	10.6	5.0	**0.2**	0.6	**0.59**	0.17
item 17	17.0	6.6	0.4	0.0	0.77	**0.09**
item 18	5.4	5.8	0.4	0.2	0.74	0.22
item 19	14.3	10.6	**0.2**	0.0	0.67	**0.12**
item 20	**21.2**	9.1	0.4	0.0	0.77	0.20
item 21	**21.8**	**21.4**	0.3	0.2	0.77	**0.13**
item 22	7.5	9.1	0.5	0.0	0.76	0.22
item 23	13.7	5.4	0.5	0.2	0.69	0.18
item 24	8.8	15.2	0.4	0.4	0.75	0.17
item 25	7.7	**20.4**	0.3	0.4	**0.65**	0.24
item 26	10.8	16.0	0.3	0.2	0.74	**0.15**
item 27	11.0	15.8	0.4	0.2	0.76	0.24
item 28	11.7	20.0	**0.2**	0.4	**0.55**	0.16
item 29	10.0	14.3	0.4	0.2	0.76	0.21
item 30	5.4	18.1	0.4	0.2	0.74	0.23
item 31	9.4	4.2	0.5	0.4	0.71	0.30
item 32	1.3	3.5	0.4	0.4	0.78	0.31

Table 4: Summary of item statistics for 18 remaining AQLQ items by domain

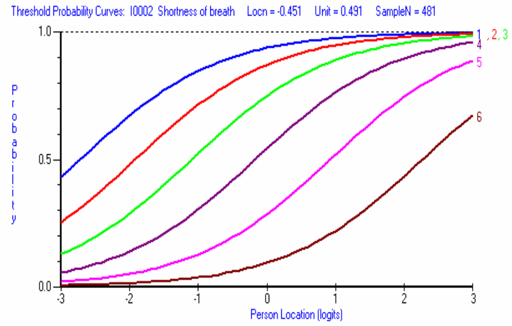
Item	Domain	c^2	DF	P-value	Item difficulty (logit)	Spread of levels: Average
6	Symptoms	8.18	7	0.317	0.011	0.73
8	Symptoms	17.31	7	0.016	-0.446	0.88
10	Symptoms	9.27	7	0.234	-0.016	0.78
14	Symptoms	5.94	7	0.547	0.319	0.72
18	Symptoms	5.03	7	0.657	0.082	0.78
22	Symptoms	10.47	7	0.163	0.074	0.64
24 (Sleep)	Symptoms	13.09	7	0.070	0.212	0.56
29 (Sleep)	Symptoms	9.22	7	0.237	0.127	0.57
2	Activity	2.12	6	0.909	-0.456	0.91
25(3)	Activity	7.13	6	0.309	0.594	0.64
31	Activity	7.02	6	0.319	-0.863	0.75
32	Activity	5.17	6	0.522	-0.282	0.93
7	Emotion	9.40	7	0.225	-0.155	0.68
13	Emotion	9.95	7	0.192	-0.043	0.53
15	Emotion	4.34	7	0.740	0.103	0.45
27	Emotion	5.25	7	0.630	0.124	0.46
23	Environment	3.17	7	0.869	-0.311	0.62

Figure 2: Initial probability threshold curves for items 7, 8, 23, 29, and 32 before collapsing

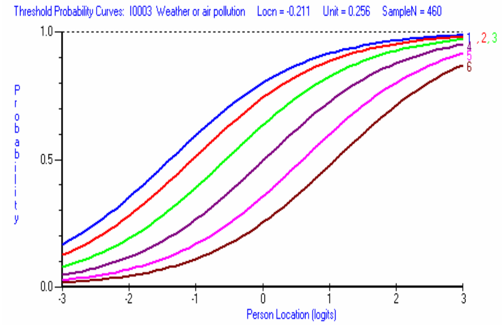
Item 7



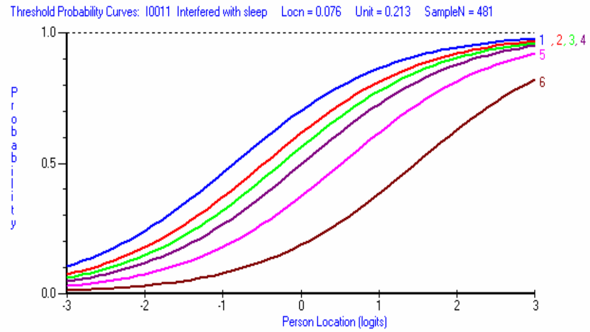
Item 8



Item 23



Item 29



Item 32