HEDS Discussion Paper 08/05

Published paper

Young Y, Yang Y, Brazier J, Tsuchiya A, Coyne K. (2009) The first stage of developing preference-based measures: constructing a health-state classification using Rasch analysis. Quality of Life Research 2009;18(2):253-65.


ScHARR
School of Health and
Related Research

The University Of Sheffield.

ScHARR

Health Economics and Decision Science Discussion Paper Series

No. 08/05

Making Rasch Decisions: The use of Rasch Analysis in the Construction of Preference Based Health Related Quality of Life Instruments

Tracey Young[1, 2], Yaling Yang[1], John Brazier[1], Aki Tsuchiya[1, 3], Karin Coyne[4]

1. School of Health and Related Research, University of Sheffield, Sheffield, S1 4DA UK
2. Trent Research Development and Support Unit (RDSU), University of Sheffield, Sheffield, S1 4DA UK
3. Department of Economics, University of Sheffield, Sheffield, S1 4DT UK
4. United BioSource Corporation Center for Health Outcomes Research, Maryland, US

Corresponding author:
Dr. Tracey Young
HEDS, School of Health and Related Research
University of Sheffield
Regent Court
30 Regent Street
Sheffield S1 4DA
Telephone 44 (0)114 2220837
Fax: 44 (0)114 2724095
E-mail: t.a.young@sheffield.ac.uk

ABSTRACT

Objective: To set out the methodological process for using Rasch analysis alongside traditional psychometric methods in the development of a health state classification that is amenable to valuation.

Methods: The overactive bladder questionnaire is used to illustrate a four step process for deriving a reduced health state classification from an existing non-preference based health related quality of life instrument. Step I excludes items that do not meet the initial validation process and step II uses criteria based on Rasch analysis and psychometric testing to select the final items for the health state classification. In step III, item levels are examined and Rasch analysis is used to explore the possibility of reducing the number of item levels. Step IV repeats steps I to III on alternative data sets in order to validate the selection of items for the health state classification.

Conclusions: The techniques described enable the construction of a health state classification amenable for valuation exercises that will allow the derivation of preference weights. Thus, the health related quality of life of patients with conditions, like overactive bladder, can be valued and quality adjustment weights such as quality adjusted life years derived.

Key Words:

Rasch analysis; health related quality of life; condition specific measure; preference-based measures; overactive bladder syndrome

Abbreviations

DIF	Differential item functioning
HRQL	Health related quality of life
OAB	Overactive bladder syndrome
OAB-q	Overactive bladder questionnaire
PSI	Person separation index
QALY	Quality adjusted life years
SRM	Standardised response mean

INTRODUCTION

Economic evaluations are performed when investigating health care technologies to inform the allocation of resources between competing health care interventions. The usual methodology is cost-effectiveness analysis, where interventions are compared in terms of their cost per quality adjusted life years (QALYs) gained. The QALY captures health benefits in terms of the length of life (in the form of 'years') and the quality of life (in the form of 'health state values') into a single summary measure, based on people's preferences.

A common way to obtain health state values is to use a 'generic' preference-based instrument, such as the EQ-5D [1] HUI3 [2], or SF-6D [3], where it is assumed that the generic measure is relevant to all health care interventions and patient groups. This assumption has been supported in some interventions and disease groups, for example Marra *et al* [4] demonstrated the discriminative ability of four generic measures across severity levels for patient with rheumatoid arthritis. However, generic measures of health have been found to be inappropriate or insensitive for some medical conditions [5]. For this reason many clinicians and researchers prefer condition specific measures. However, most condition specific measures are not preference-based and cannot be used to derive the 'quality adjustment weight' for use in QALYs.

One approach has been to try to 'map' from the condition specific measures onto the generic preference-based measures by judgement based methods, for example, using panels of experts [6], or empirically, using a separate data set containing the non-preference based measure and the generic preference-based measure [7-8]. Several studies have mapped condition specific non-preference based measures onto preference-based measures using regression based approaches (e.g. Tsuchiya *et al,* [9]). However, a review of mapping functions from 28 studies found that their

performance, in terms of model fit and predictions, varied considerably [10]. More concerning is the tendency of these mapping functions to over predict the preference-based index at the lower, more severe end of the health related quality of life (HRQL) scale and under predict at the higher end (e.g. Gray *et al,* [11]). Further, mapping methods are limited by the degree of overlap in terms of coverage between the descriptive systems (questionnaire items) of measures, where important dimensions of one instrument are not necessarily covered by the other instrument.

These issues regarding mapping functions lead to the exploration of a more direct methodological approach of deriving preference-based quality adjusted weights for the condition specific measure. In order to do this, it is usually necessary to reduce the length of the original instrument, as existing measures are typically too large to value. The approach described here develops the methodology pioneered for the SF-36, where the 36-item instrument was reduced to a six-dimension health state classification that was amenable to valuation by respondents [3, 12].

This paper sets out to use Rasch analysis as a tool alongside conventional psychometric methods to help derive a health state classification system amenable to valuation using the Overactive Bladder Questionnaire (OAB-q), a condition specific measure, as an example. Rasch analysis is a mathematical modelling technique [13] that, in relation to HRQL questionnaires, converts each categorical (qualitative) item response to a continuous (unmeasured) latent scale using a logit model, where the scale is conceived to be a continuous measure of HRQL.

The process is the first stage of a three part process in creating a preference-based measure that can be obtained directly from the original instrument [14]. In the first stage Rasch analysis is used, firstly to select items for from the original instrument and secondly to reduce the existing number of item levels to a more manageable

number for valuation purposes. In this paper we do not argue that Rasch analysis is offering a single formulaic solution to the problem of developing a health state classification small enough to be subjected to a valuation survey from an existing HRQL instrument. What we attempt to show is how Rasch analysis can help ensure that the process of creating a preference-based measure makes the best use of the richness and sensitivity of the original instrument.

The second stage is to undertake a survey which asks a representative sample of the general population to value a selection of states defined by the reduced classification system. The third stage is modelling the sample health states values and uses these econometric models to predict health state values for all possible states defined by the new classification system. The resulting scoring algorithm or population value set will provide a preference-based single index measure for different condition specific health states, contributing towards the calculation of QALYs. This paper reports on the first stage, and the second and third stages are addressed in a separate paper [Yang *et al (unpublished)*].

THE OVERACTIVE BLADDER QUESTIONNAIRE (OAB-q)

Overactive bladder (OAB) is characterized by urinary urgency, with or without urgency incontinence, and is often accompanied by urinary frequency and nocturia [15]. It has been demonstrated that generic quality of life measures such as the SF-36 and the EQ-5D may not be sensitive enough to detect change in HRQL over time for patients' with specific symptoms [16], therefore, a condition specific measure may be more appropriate (e.g. OAB-q).

The OAB-q is a 33-item condition specific questionnaire that was developed to assess the symptom bother and impact of OAB on patients. The OAB-q has been well validated in patients with continent and incontinent OAB and has demonstrated

good internal consistency, test-retest reliability, concurrent validity, discriminate validity and responsiveness to treatment-related change [17-19]. The OAB-q consists of an eight-item Symptom Bother scale and a 25-item HRQL scale with four subscales: Coping (eight items), Sleep (five items), Concern (seven items) and Social Interaction (five items). Responses to the OAB-q are based upon a six-point Likert scale with one denoting no impact and six high impact. Results are reported in terms of domain scores or overall scores, where neither scoring system is preference based.

Two versions of OAB-q exist – the 33-item version described above and the 19-item short-form OAB-q comprising of six items on symptom bother and a further 13 items on HRQL. The 19 items in the short-form OAB-q are identical in wording and item response options to those in the longer version of the instrument.

OAB-q Datasets and Rasch Modelling

Rasch rating scale models [13] are fitted to responses to the OAB-q from patients with OAB, with or without urge incontinence, from two datasets, 'Trial I' and 'Trial II'. Both Trial I and Trial II are multicentre randomised controlled trials where participants with OAB are asked to complete the OAB-q at baseline, 4 weeks and 12 weeks [20]. A sub-sample of 391 patients randomly selected from the baseline analysis of Trial I are used to select items for the reduced health state classification. Data from the remaining patients in Trial I and Trial II are then used to validate the item choices.

RASCH ANALYSIS IN THE CONSTRUCTION OF A HEALTH STATE CLASSIFICATION

Prior to the item selection process, a decision should be made as to whether the reduced health state classification can be derived from all versions of the existing condition specific measure when multiple versions of the condition specific measure

exist, (e.g. the 33-item OAB-q and the 19-item short-form OAB-q). Consideration should be given to any potential differences in wording for those items common across versions or to versions that combine two or more items from previous versions. These differences are important at the valuation stage of the health state classification where any change in wording could lead to responders placing different values on the health states under consideration.

For OAB, the research team decided that the health state classification should be derived from both versions of the OAB-q. Thus, only the 19 questions common to both instruments are considered for selection in the reduced health state classification (Table 1).

The subsequent process of selecting items can be divided into four steps; the first step excludes items that do not meet the initial validation process and the second step uses a set of criteria based on Rasch analysis and psychometric testing to select the final items for the preference based measure. In the third step, item levels are examined and Rasch analysis is used to explore the possibility of reducing the number of item levels. The final step repeats steps one to three on alternative data sets in order to validate the selection of items for the health state classification.

Step I – Using Rasch Analysis to Eliminate Items

Rasch analysis [13] is typically used in HRQL studies in the development of new HRQL questionnaires [e.g. 21-23], and in the validation of existing HRQL questionnaires [e.g. 24-26]. In the above situations Rasch analysis is used as a tool to validate *item level ordering*, establish *differential item functioning (DIF),* and identify items that do not *fit* the underlying Rasch models. This same process can be used as a first stage in the construction of a health state classification where items failing item level ordering, DIF, and model fit can be eliminated.

Item level ordering

The first step is to examine item threshold maps for item level ordering. For an ordered item, the thresholds between item levels are the points at which each item level is equally likely to occur. Unordered items highlight the inability of responders to distinguish between item levels. Ordering can be achieved by uniformly merging adjacent item levels across all items or using an item by item approach: either approach is valid. When creating a reduced health state classification, one of the objectives is to select items that respond to the full range of severity across the condition being measured. Therefore, any item where it is necessary to collapse item levels fails to respond to the full range of condition specific severity and is excluded from further consideration.

A total of five items, common to both the 33-item OAB-q and the 19 item short-form OAB-q, (items 4, 11, 12, 16, and 19) were unordered in the initial Rasch models (Table 1). For four of the seven items “A good bit of the time” and “most of the time” and “all of the time” were collapsed into one item level.

DIF

DIF is examined to establish whether responses to the HRQL questionnaire differ across patient characteristics. For example younger patients might select less severe item response options for an item asking about physical abilities than older patients. Items where it is necessary to adjust for systematic DIF across groups of responders are of limited value for making cross population comparisons and therefore are excluded from further consideration.

Gender (male/female), age (less than 65 years/greater than 65 years) and OAB severity (mild/moderate/severe) were examined for DIF. Adjustments were made to

Rasch models where DIF occurred. However, any item where it was necessary to adjust for DIF was excluded from further consideration. Three items were split for DIF, item 3 due to different responses to OAB-q by symptom severity and items 9 and 19 due to different responses by age (Table 1).

Rasch model goodness of fit

The final validation step is to achieve overall model goodness of fit by examining item-trait interactions, the person separation index (PSI) and person and item fit residuals. The fit of each of the individual items included in the Rasch model is examined and poor fitting items (items that do not meet the unidimensionality of the underlying latent scale) are removed until overall Rasch model goodness of fit is achieved. Any item that is removed is excluded from further consideration.

No items were removed from the symptoms, coping, or social domains, item 10 was removed from the sleep domain, and item 17 from the concerns domain to achieve model goodness of fit. Table 2 presents the overall model goodness of fit statistics after removing poor fitting items from the Rasch models. All models achieved high PSI, had reasonable item and person fit residuals and meet the (chi-squared) model goodness of fit requirements ($p > 0.01$).

Factor Analysis and Cluster Analysis

Factor analysis attempts to identify underlying factors that explain patterns of correlation within a set of observed variables [27] and can therefore be used alongside Rasch analysis and psychometric analysis to identify items that are poorly correlated (not loading) within a domain. Any items that are strongly correlated (high loadings) are candidates for inclusion in the final health state classification.

Additionally, cluster analysis of items can be used to validate the HRQL domain structure and identify items that do not fit naturally into the instrument domain structure (candidates for deletion) or which belong to an alternative domain structure. In the later case, the alternative structure can be explored using Rasch analysis and the results taken into consideration in the item selection process.

Results from factor analysis of the symptom domain indicated that items 4 and 5, which asked about night time urination, loaded differently to the remaining items in the domain; cluster analysis of all 33 items confirmed that items 4 and 5 loaded with the sleep domain rather than the symptom domain. Further examination of the symptoms domain using cluster analysis indicated that items 3 and 6, related to urine loss, grouped differently to items 1 and 2 which asked about frequency and urge (results available from the authors on request). Therefore, it was decided to select one item each from four OAB-q HRQL domains (coping, sleep, concern and social) and two items from the symptoms domain, one related to urine loss and the other to urge, in order to represent both of these symptoms in the health state classification system.

Step II – Using Rasch Analysis to Select Items

Step I results in a series of items from which to select the most 'representative' items from the full condition specific measure. A total of ten OAB-q items remained for consideration at this stage: Items 1, 2 and 6 from the symptoms bother domain, items 7, 13 and 14 from the coping domain, item 8 from the sleep domain, item 18 from the social domain and item 15 from the concern domain. The aim of the selection process for the reduced health state classification is to select items that span the full range of condition severity. The criteria used are described below.

Rasch analysis

Item selection is predominantly based upon the spread of item levels across the latent space, as selected items should span the full range of condition severity, where the wider the spread the better the item. Threshold probability curves are used to examine the spread of item levels at a point on the latent scale, typically at the central logit zero. However the spread at the mean logit for each item is also an acceptable approach to use.

Item goodness of fit statistics (χ^2) are also taken into consideration when selecting items, where the better the goodness of fit (high χ^2 value and non-significant p-value) the better the item represents the underlying unidimensional latent scale. Finally, overall item logit 'score' is considered to ensure a spread of logit scores across latent space in order to represent items that span different levels of condition severity.

Psychometric analysis

It is recognised that Rasch analysis should be used alongside conventional psychometric methods in the construction of HRQL measures [28]. Therefore, the performance of items across conventional psychometric criteria should be taken into consideration when selecting items, such as: feasibility (rate of missing data), internal consistency (correlation between item and domain scores), distribution of responses (e.g. the absence of ceiling or floor effects) and responsiveness (between baseline and follow-up visits). The responsiveness of OAB-q was measured with the standardised response mean (SRM) [29] using data from Trial I at baseline and 12 weeks follow-up.

The overall performance of items across Rasch and psychometric tests was examined with preference given to items with reasonable Rasch model goodness of

fit and spread of item levels. Rasch goodness of fit statistics and item level spread criteria were obtained from Rasch models fitted to the five OAB-q domains. These Rasch models included items that were not under consideration for inclusion in the reduced health state classification, due to the exclusion criteria applied in step I, as these items contribute to the original OAB-q overall domain measure of HRQL.

Item specific results from Rasch models and psychometric tests are presented in Table 3 for the 10 items under consideration in the reduced health state classification. Item 18, the only item under consideration from the social domain, had a very high floor effect, where 61.8% of responders felt that OAB never "affected your relationship with family and friends". Additionally, item 18 had poor responsiveness (SRM = 0.43) compared with other items. Therefore, the team decided not to include an item from the social domain in the preference-based measure.

Of the three items in the symptoms domain item 6 was the remaining item that asked about loss of urine and was therefore included in the reduced health state classification. Of the remaining two items, item 2 had a poor Rasch model goodness of fit and was excluded. Therefore items 1 and 6 were selected.

All three items in the coping domain performed well in psychometric tests. Item 13 was selected from the coping domain as it had the best goodness of fit in the Rasch model and largest item spread at logit 0. For both the sleep and concern domains one item remained for possible selection (item 8 – sleep and item 15 – concern) and these items were also included.

Thus the five items selected for the reduced health state classification, to be known as OAB-5D, were:

Item 1 – an uncomfortable urge to urinate

Item 6 – urine loss associated with a strong desire to urinate

Item 8 – bladder symptoms interfered with your ability to get a good night's sleep

Item 13 – bladder symptoms caused you to plan escape routes

Item 15 – bladder symptoms caused you embarrassment

Step III: Using Rasch Analysis to Explore Collapsing Item Levels

In the construction of a reduced health state classification it is important to ensure that information relating to items and item levels is not redundant. Previous work with health state classifications has found that respondents sometimes have problems distinguishing between item levels. Rasch analysis, alongside psychometric criteria, can be used to reduce the number of item levels and examine whether respondents are able to distinguish between item levels and whether ordering makes sense. Item level collapsing should be carried out at the generic level, for items sharing a common wording of item levels, for example "I have problems with X none of the time / some of the time / all of the time". A generic collapsing of item levels will aid responders in the valuation survey in distinguishing between health states.

Threshold probability curves, which show the distribution of item levels across latent space, can be examined, where item levels that are closer together, in comparison with other levels, being candidates for item level collapsing. The distribution of responses across an item should also be taken into account when merging item levels, where adjacent item levels with low response frequencies are candidates for collapsing. More than one possible solution to item merging may be available and each solution should be tested using Rasch analysis and the results compared, using threshold maps and item goodness of fit statistics, before deciding on the most appropriate solution.

For the OAB-5D the aim was to explore the possibility of having either four or five item levels rather than the original six item levels. Items 1 and 6 present item choices in terms of severity, ranging from “not at all” to “a very great deal”. The remaining three items ask about frequencies and range from “none of the time” to “all of the time”.

An examination of threshold curves (Figure 1) for item 6 suggested that levels “a little bit” and “somewhat” were closest together, this pattern was also true across other items in the symptoms domain not selected for OAB-5D (results not shown). However, item responses (Table 4) suggested collapsing “not at all” with “a little bit” or “a great deal” with a very great deal” as these response categories had the fewest responders. Each of the three options listed above were tested in Rasch analysis on the validated symptoms domain and “a little bit” and “somewhat” gave the best Rasch model goodness of fit statistics and thus was used in OAB-5D.

It was difficult to identify potential item response categories for collapsing from the proportion of responses at each level as these were fairly evenly distributed across all item levels for items 8, 13 and 15 (Table 4). The threshold probability curve for item 15 suggested that “most of the time” and “all of the time” could be collapsed; item 13 implied “a good bit of the time” and “most of the time” could be collapsed and item 8 suggested “a good bit of the time” and “most of the time” and “all of the time.” Each option was tested in Rasch analysis: when two levels, “most of the time” and “all of the time”, or three levels, “a good bit of the time” and “most of the time” and “all of the time” were collapsed, item 13 (coping) was no longer ordered in Rasch models, suggesting further level collapsing. However, all three items performed well when “a good bit of the time” and “most of the time” were collapsed together. Figure 2 presents the final instrument for OAB-5D.

Step IV: Validation of Item Selection for the Preference-Based Instrument

The item selection process, described above, should be validated by repeating the item selection and item level reduction process on a random sample from the original data set, or a second independent sample of patients. Further, if the preference based instrument is to be amenable to multiple versions of the condition specific measure, then item selection and item level reduction should be validated for all versions of the instrument. This validation check will inevitably strengthen the justification for items selected for the final preference-based instrument.

Rasch models were validated on the remaining 746 patients from Trial I. Item selection was also validated on a second dataset, Trial II, (N = 793). All Rasch models were first fitted to the 33-item OAB-q and further validated on the short-form OAB-q. All analysis confirmed the selection of items 1, 6, 8, 13 and 15 in OAB-5D. (Results are available from the authors on request.)

DISCUSSION AND CONCLUSIONS

The first stage of deriving a preference-based single index measure for use in calculating QALYs is to derive a health state classification system that is amenable to valuation using a preference elicitation technique. This paper has presented a solution to this problem using Rasch analysis, alongside psychometric criteria. The advantage of this, is that the process of item selection becomes more systematic than methods used previously. Using the methods described in this paper, the developers of preference-based instruments can define a set of criteria, based on mathematical models, to be met when selecting items.

Some of the methods suggested here are dependent on data availability, for example patient and disease characteristics for DIF analysis and additional response data for psychometric criteria and validation. It is not necessary to use all of these criteria

when deriving a preference based measure. However, if the information is available then using it in the derivation of the reduced health state classification will strengthen the selection process and increase the validity of the final item selection.

The process described here should be used as a guide in the first stage of the development of health state classification and developers of such instruments set their own criteria based on the aims and objectives of the instrument they are deriving. For example, the developers may wish to represent all domains of the original instrument in the new instrument, as was the case for OAB-q. However other developers may, *a priori,* believe that it is not necessary to represent all the domains of the instrument or to include more than one item from each dimension. We advise that conventional methods such as factor analysis and cluster analysis are used in the justification of these decisions.

In addition, consideration may be given to the wording of the items. It is important to ensure that the wording of the health states derived from the items are comprehensible and amenable to valuation. The item wording should be reasonable in length to be conceivably used in a preference-base measure. Additionally, the item should not be linked or paired in anyway to other items in the original non-preference based questionnaire as the item may be valued differently within the full instrument to how it is valued within the reduced health state classification.

Similarly, the methods used to achieve a reduced number of item levels may vary by developers, with some developers deciding not to reduce the number of item levels. We have found that respondents to valuation exercises sometimes struggle in distinguishing between item levels and it is worthwhile exploring the possibility of collapsing redundant levels. However, one of the problems of using Rasch analysis in the reduction process is that the initial selection criteria is the spread of item levels

– where the wider the item spread the better, and the chosen items typically had evenly spread item levels. Therefore, any attempt to reduce the number of levels may result in violation of Rasch model assumptions; thus the results from Rasch analysis should be considered alongside the spread of item responses when making a judgement on how to reduce item levels.

A further consideration when constructing a health state classification is the choice of sample. Developers should ensure that the datasets they use to construct and validate the health state classification are representative of the population they wish to use the final instrument on. Failure to use a representative sample could result in a measure that is not sensitive to changes on HRQL for some subgroups of the population.

This paper has described the first step in developing a preference-based measure, illustrated using OAB-5D. A valuation survey has been undertaken for OAB-5D on the general population and a single index has been derived based on these results [Yang *et al (u*npublished)].

The techniques described here mean that the HRQL of patients with some conditions, like OAB, that might not be measurable using generic preference-based measures, can be valued and quality adjustment weights like QALYs derived, thus increasing the use of cost-effectiveness analysis with QALYs in the evaluation of health care technologies.

Acknowledgements

This study is funded by Pfizer Inc. John Brazier is funded by the Medical Research Council Health Service Research Collaboration. Zoe Kopp provided advice throughout the study, the Trial I and Trial II datasets were provided by Pfizer Inc. The usual disclaimer applies.

References

1. Brooks R. (1996) EQ-5D, the current state of play. Health Policy; 37: 53-72

2. Feeny D, Furlong W, Torrance GW *et al.* (2002) Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. Med Care; 40: 113-128

3. 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(2):271-92

4. Marra CA. Woolcott JC. Kopec JA, e*t al.* (2005) A comparison of generic, indirect utility measures (the HUI2, HUI3, SF-6D, and the EQ-5D) and disease-specific instruments (the RAQoL and the HAQ) in rheumatoid arthritis. Social Science & Medicine; 60(7):1571-82

5. Brazier, JE, Deverill M., Harper R., Booth A. (1999) A review of the use of Health Status measures in economic evaluation*. Health Technol Assess* 3(9)

6. Coast J. (1992) Reprocessing data to form QALYs. BMJ; 305(6845):87-90

7. Fryback DG, Dasbach EJ, Klein R *et al.* (1993) The Bever Dam Health Outcomes Study: initial catalogue of health-state quality factors. Med Decis Making; 13: 89-102

8. Nichol MB, Sengupta N, Globe DR. (2001) Evaluating quality adjusted life-years: estimation of the Health Utility Index (HUI2) from the SF-36. Med Decis Making; 21: 105-112

9. Tsuchyia A, Brazier J, McColl E, Parkin D. (2002) Deriving preference-based condition-specific instruments: converting AQLQ into EQ-5D indices. Health Economics and Decision Science Discussion Paper Series No. 02/01, Accessed from: http://www.shef.ac.uk/content/1/c6/01/87/47/DP0201.pdf

10. Brazier J, Yang Y, Tsuchiya A (2007) Review of methods for mapping between measures of health related quality of life onto generic preference-based measures: a road to nowhere? (Paper presented at the Health Economics Study Group Meeting, Brunel University, Uxbridge)

11. Gray A, Rivero-Arias O, Clarke PM. (2006) Estimating the association between SF-12 responses and EQ-5D utility calues by response mapping. Med Decis Making; 26: 18-29

12. Brazier J, Usherwood T, Harper R, Thomas K. (1998) Deriving a preference-based single index from the UK SF-36 Health Survey. J Clin Epidemiol; 51(11):1115-28

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

14. Brazier J, Roberts J, (2005) Estimation of a preference based index measure of health for the SF-12 and comparison to the SF-36. Medical Care; 42(9):851-859

15. Abrams P, Cardozo L, Fall M et a (2002) The standardization of terminology of lower urinary tract function: report from the Standardization Sub-committee of the International Continence Society, Neurourol Urodyn; 21: 167–178

16. Kobelt G, Kirchberger I, Malone-Lee J. (1999) Quality-of-life aspects of the overactive bladder and the effect of treatment with tolterodine. BJU International; 83(6):583-90

17. Coyne KS, Matza LS, Thompson CL. (2005) The responsiveness of the overactive bladder questionnaire (OAB-q). Qual Life Res; 14(3):849-55

18. Coyne K, Revicki D, Hunt T, *et al.* (2002) Psychometric validation of an overactive bladder symptom and health-related quality of life questionnaire: the OAB-q. Qual Life Res; 11(6):563-74

19. Matza LS, Thompson CL, Krasnow J, Brewster-Jordan J, Zyczynski T, Coyne KS. (2005) Test-retest reliability of four questionnaires for patients with overactive bladder: the overactive bladder questionnaire (OAB-q), patient perception of bladder condition (PPBC), urgency questionnaire (UQ), and the primary OAB symptom questionnaire (POSQ). Neurourology & Urodynamics; 24(3):215-25

20. Siami P Seidman LS, Lama D. (2002) A multicentre, prospective, open-label study of tolterodine extended-release 4mg for overactive bladder: the speed of onset of therapeutic assessment trial (STAT). Clin Ther; 24(4): 616-628

21. Duncan PW, Bode RK, Min Lai S, Perera S. (2003) Glycine Antagonist in Neuroprotection Americans Investigators. Rasch analysis of a new stroke-specific outcome scale: the Stroke Impact Scale. Arch Phys Med Rehabil; 84(7): 950-963

22. Gilworth G, Chamberlain MA, Bhakta B, Haskard D, Silman A, Tennant A. (2004) Development of the BD-HRQL: a quality of life measure specific to Behcet's disease. J. Rheumatol; 31(5): 931-937

23. Pesudovs K, Garamendi E, Elliott DB. (2004) The quality of life impact of refractive correction (QIRC) questionnaire: development and validation. Optom Vis Sci; 81(10): 769-777

24. Raczek AE, Ware JE, Bjorner JB, Gandek B, Haley SM, Aaronson NK, *et al.* (1998) Comparison of Rasch and summated rating scales constructed from SF-36 physical functioning items in seven countries: Results from the IHRQLA project. J Clin Epidemiol; 15(11): 1203-1214

25. White LJ, Velozo CA. (2002) The use of Rasch measurement to improve the Oswestry classification scheme. Arch Phys Med Rehabil; 83(6): 822-831

26. Valderas JM, 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). Qual Life Res; 13(1): 35-44

27. Chatfield C, Collins AJ. (1980) Introduction to Multivariate Analysis. Chapman and Hall; University Press, Cambridge

28. Tennant A, McKenna S.P. Hagell P. (2004) Application of Rasch analysis in the development and application of quality of life instruments. Value Health; 7(Supplement 1): S22-S26

29. Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR. (2002) Clinical Significance Consensus Meeting Group. Methods to explain the clinical significance of health status measures. Mayo Clin Proc; 77(4): 371–383

Figure 1: Probability Threshold Curves for Items 1, 6, 8, 13 and 15 of the OAB-5D Prior to Item Level Collapsing

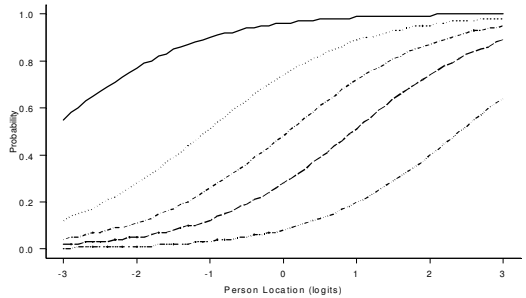
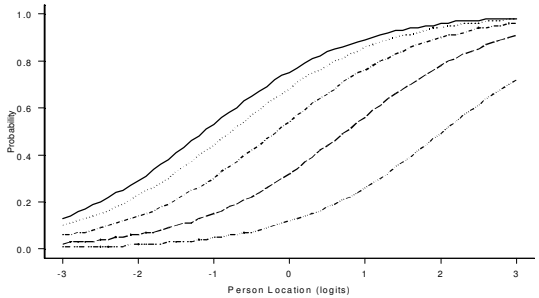
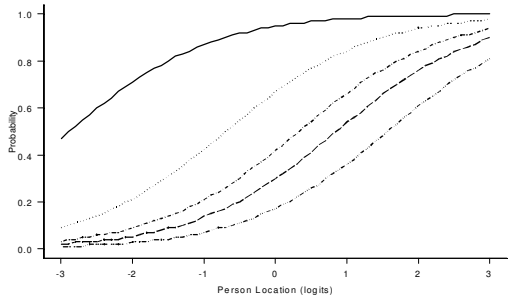
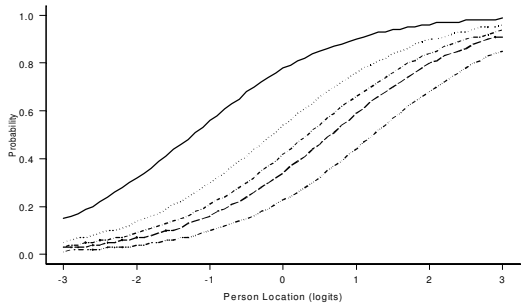
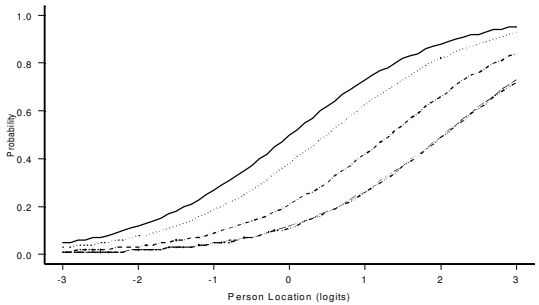
<table>
<tr><td>Item 1</td><td>Item 6</td><td>Item 8</td></tr>
<tr><td> 
</td><td> 
</td><td> 
</td></tr>
<tr><td> 
</td><td> 
</td><td>Threshold between:
Item levels 1 and 2
Item levels 2 and 3
Item levels 3 and 4
Item levels 4 and 5
Item levels 5 and 6</td></tr>
<tr><td>Item 13</td><td>Item 15</td><td></td></tr>
</table>

Figure 2 OAB-5D Classification System

URGE
1 Not at all bothered by an uncomfortable urge to urinate
2 Bothered by an uncomfortable urge to urinate a little bit or somewhat
3 Bothered by an uncomfortable urge to urinate quite a bit
4 Bothered by an uncomfortable urge to urinate a great deal
5 Bothered by an uncomfortable urge to urinate a very great deal
URINE LOSS
1 Not at all bothered by urine loss associated with a strong desire to urinate
2 Bothered by urine loss associated with a strong desire to urinate a little bit or somewhat
3 Bothered by urine loss associated with a strong desire to urinate quite a bit
4 Bothered by urine loss associated with a strong desire to urinate a great deal
5 Bothered by urine loss associated with a strong desire to urinate a very great deal
SLEEP
1 Bladder symptoms interfered with your ability to get a good night's rest none of the time
2 Bladder symptoms interfered with your ability to get a good night's rest a little of the time
3 Bladder symptoms interfered with your ability to get a good night's rest some of the time
4 Bladder symptoms interfered with your ability to get a good night's rest a good bit or most of the time
5 Bladder symptoms interfered with your ability to get a good night's rest all of the time
COPING
1 Bladder symptoms caused you to plan 'escape routes' to restrooms in public places none of the time
2 Bladder symptoms caused you to plan 'escape routes' to restrooms in public places a little of the time
3 Bladder symptoms caused you to plan 'escape routes' to restrooms in public places some of the time
4 Bladder symptoms caused you to plan 'escape routes' to restrooms in public places a good bit or most of the time
5 Bladder symptoms interfered with your ability to get a good night's rest all of the time
CONCERN
1 Bladder symptoms caused you embarrassment none of the time
2 Bladder symptoms caused you embarrassment a little of the time
3 Bladder symptoms caused you embarrassment some of the time
4 Bladder symptoms caused you embarrassment a good bit or most of the time
5 Bladder symptoms caused you embarrassment all of the time

Table 1: Overall Summary of the 19 items common to the OAB-q and short form OAB-q and initial Rasch Validation Criteria

Item	Question	Domain	In short-form OAB-q	Item levels collapsed	DIF - Characteristic and split	Excluded from Rasch validation
1	An uncomfortable urge to urinate	Symptom bother	Yes			
2	A sudden urge to urinate with little or no warning	Symptom bother	Yes			
3	Accidental loss of small amounts of urine	Symptom bother	Yes		Symptom severity Mild V Moderate/Severe	
4	Nighttime urination	Symptom bother	Yes	Not at all AND A little bit		
5	Waking up at night because you had to urinate	Symptom bother	Yes			
6	Urine loss associated with strong desire to urinate	Symptom bother	Yes			
7	Made you uncomfortable while travelling with others because of needing to stop for a restroom	Coping	Yes			
8	Interfered with your ability to get a good night's rest	Sleep	Yes			
9	Awakened you during sleep	Sleep	Yes		Symptom severity Mild V Moderate/Severe	
10	Interfered with getting the amount of sleep you needed	Sleep	Yes			Yes
11	Made you avoid activities away from restrooms (i.e. walks, running, hiking)	Coping	Yes	A good bit of the time AND Most of the time AND All of the time		
12	Caused you to decrease your physical activities (exercising, sport etc.)	Coping	Yes	None of the time AND A little of the time Most of the time AND All of the time Some of the time AND A good bit of the time		
13	Caused you to plan "escape routes" to restrooms in public places	Coping	Yes			
14	Caused you to locate the closest restroom as soon as you arrive at a place you have never been	Coping	Yes			
15	Caused you embarrassment	Concern	Yes			
16	Made you feel like there is something wrong with you	Concern	Yes	A good bit of the time AND Most of the time AND All of the time		
17	Made you frustrated or annoyed about the amount of time you spend in the restroom	Concern	Yes		Age Less than 65 V greater than 65	Yes
18	Affected your relationship with family and friends	Social	Yes			
19	Caused you to have problems with your partner or spouse	Social	Yes	A good bit of the time AND Most of the time AND All of the time		

Table 2: Rasch Model Goodness if Fit for the five OAB-q Domains

Domain	χ^2	Degrees of Freedom	P-value	Person Separation Index	Item Fit (SD)	Person Fit (SD)
Symptom bother	50.6	45	0.262	0.89	0.51 (1.06)	-0.66 (1.84)
Coping	53.4	45	0.183	0.93	0.03 (0.96)	-0.36 (1.24)
Sleep	42.8	25	0.015	0.93	0.05 (1.63)	-0.43 (1.02)
Social	31.9	20	0.035	0.88	0.04 (2.03)	-0.46 (1.08)
Concerns	38.1	25	0.045	0.84	-0.15 (1.05)	-0.29 (0.81)

Table 3: Rasch and Psychometric Results for the 10 items not excluded from Rasch model validation common to OAB-q and short-form OAB-q

		Rasch Criteria				Psychometric Criteria				
Item	Domain	Location	Residual	χ^2_5 (P-value)	Spread at logit 0	% response at floor (level 1)	% response at ceiling (level 6)	SRM	% missing data	Correlation with domain score
1	Symptoms	-0.17	-0.11	5.28 (0.382)	0.88	2.0	7.9	0.79	6.5	0.67
2	Symptoms	-0.21	-0.38	14.98 (0.010)	0.81	3.1	9.6	0.82	7.8	0.74
5	Symptoms	-0.09	1.99	11.90 (0.036)	0.53	7.4	15.0	0.80	3.3	0.65
6	Symptoms	0.14	1.22	0.60 (0.988)	0.63	6.6	8.7	0.72	18.4	0.69
7	Coping	0.23	1.73	9.37 (0.095)	0.43	27.3	11.4	0.72	6.7	0.85
13	Coping	0.16	0.10	2.42 (0.788)	0.55	25.8	10.2	0.62	2.8	0.85
14	Coping	-0.51	-0.46	6.28 (0.280)	0.54	16.4	19.1	0.77	1.6	0.84
8	Sleep	-0.17	-0.47	18.51 (0.002)	0.78	11.0	28.4	0.84	1.3	0.91
18	Social	0.40	-1.86	6.46 (0.264)	0.63	61.8	4.0	0.43	3.4	0.88
15	Concern	1.18	-0.13	4.83 (0.437)	0.39	32.2	9.2	0.65	1.5	0.82

Table 4: Distribution of responses across item levels for five items selected for OAB-5D

Item	***Not at all***	***A little bit***	***Somewhat***	***Quite a bit***	***A great deal***	***A very great deal***
1	8 (2.0%)	60 (15.3%)	100 (25.6%)	104 (26.6%)	64 (16.4%)	31 (7.9%)
6	26 (6.6%)	45 (11.5%)	61 (15.6%)	90 (23.0%)	58 (14.8%)	34 (8.7%)
	None of the time	***A little of the time***	***Some of the time***	***A good bit of the time***	***Most of the time***	***All of the time***
8	43 (11.0%)	57 (14.6%)	69 (17.6%)	47 (12.0%)	59 (15.1%)	111 (28.4%)
13	101 (25.8%)	78 (19.9%)	71 (18.2%)	39 (10.0%)	47 (12.0%)	40 (10.2%)
15	126 (32.2%)	80 (20.5%)	74 (18.9%)	39 (10.0%)	29 (7.4%)	36 (9.2%)