A QALY Measure for Multiple Sclerosis: Developing a Patient-Reported Health State Classification System for an MS-Specific Preference-Based Measure

Version 1 (original submission prior to peer review) 14/08/2014

Authors: Elizabeth Goodwin PhD[1], Colin Green PhD[1,2]

[1]Health Economics Group, University of Exeter Medical School, University of Exeter, Exeter, UK

[2]Collaboration for Leadership in Applied Health Research and Care South West Peninsula, University of Exeter Medical School, University of Exeter, Exeter, UK

Contact/Correspondence: Elizabeth Goodwin

Email: e.goodwin@exeter.ac.uk

Telephone: +44 (0) 1392 72 6073

Mailing address: Health Economics Group, Institute of Health Research, Veysey Building, Salmon Pool Lane, Exeter, EX2 4SG

This research is funded by the Multiple Sclerosis Society of Great Britain and Northern Ireland.

This research was supported by the National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care South West Peninsula at the Royal Devon and Exeter NHS Foundation Trust. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

Keywords: multiple sclerosis, health-related quality of life, health states, preference-based measures of health, quality-adjusted life-years, Rasch analysis.

Running title: A QALY Measure for Multiple Sclerosis

A QALY Measure for Multiple Sclerosis: Developing a Patient-Reported Health State Classification System for an MS-Specific Preference-Based Measure

Key points

What is already known about the topic?

- Empirical evidence suggests that generic preference-based measures may lack relevance and sensitivity to differences and changes in the health-related quality of life (HRQoL) of people with multiple sclerosis (MS).
- Condition-specific preference-based measures (CSPBMs) may provide a more relevant and sensitive means of quantifying health outcomes.
- Over recent years, a methodological approach has been developed to derive health state classification systems from existing condition-specific, patient-reported measures of HRQoL.

What does the paper add to existing knowledge?

- The Multiple Sclerosis Impact Scale (MSIS-29) provides a suitable basis for a health state classification system (the MSIS-8D), which represents predefined dimensions of HRQoL of importance to people with MS and is amenable to valuation.
- Where more than one measure of HRQoL is available for the condition of interest, an appropriate measure can be selected using a standardised set of psychometric criteria.
- Where the number of conceptually distinct dimensions of HRQoL represented by the original measure exceeds its statistically independent factors, developing and applying a conceptual framework ensures that the classification system covers a range of dimensions of HRQoL relevant to the condition of interest.

What insights does the paper provide for informing health care-related decision making?

- The next stage of this research will involve estimating utility values for all health states described by the MSIS-8D. This will enable QALY weights to be derived directly from responses to the MSIS-29, for use in cost effectiveness analyses of treatments for MS.

Abstract

Objectives: Increasingly generic preference-based measures of health-related quality of life (HRQoL) are used to estimate quality-adjusted life-years in order to inform resource allocation decisions. Evidence suggests that generic measures may not be appropriate for multiple sclerosis (MS). We report the first stage in the development of an MS-specific preference-based measure to quantify the impact of MS and its treatment: deriving a health state classification system, which is amenable to valuation, from the Multiple Sclerosis Impact Scale (MSIS-29), a widely used patient-reported outcome measure in MS.

Methods: The dimensional structure of the MSIS-29 was determined using factor analysis and a conceptual framework of HRQoL in MS. Item performance was assessed, using Rasch analysis and psychometric criteria, to enable the selection of one item to represent each dimension of HRQoL covered by the MSIS-29. Analysis was undertaken using a sample (n=529) from a longitudinal study of people with MS. Results were validated by repeating the analysis with a second sample (n=528).

Results: Factor analysis confirmed the two subscale structure of the MSIS-29. Both subscales covered several conceptually independent dimensions of HRQoL. Following Rasch and psychometric analysis an eight-dimensional classification system was developed, named the 'MSIS-8D'. Each dimension was represented by one item with four response levels.

Conclusion: Combining factor analysis with conceptual mapping, and Rasch analysis with psychometric criteria, provides a valid method of constructing a classification system for an MS-specific preference-based measure. The next stage is to obtain preference weights so that the measure can be used in studies investigating MS.

Keywords: multiple sclerosis; health-related quality of life; preference-based measures of health; quality-adjusted life-years.

Introduction

Cost utility analysis is a frequently employed technique for evaluating the cost effectiveness of healthcare interventions, in which quality-adjusted life-years (QALYs) are used to compare the relative merits of treatment options in terms of their impact on both length and quality of life. QALYs are calculated by weighting each year of life according to its quality, on a scale from 1 (equivalent to full health) to zero (equivalent to being dead). Increasingly, preference-based measures (PBMs) of health-related quality of life (HRQoL) are used to provide these quality weights. PBMs use a standardised classification system to describe a finite number of possible health states. Each unique health state is assigned a numerical quality weight, typically estimated by eliciting preferences between different health states from a sample of the general population [1]. Cost utility analyses commonly employ generic PBMs, such as the EuroQol EQ-5D [2], Short-Form 6D [3] or Health Utilities Index [4], which are considered applicable for all health conditions. The broad focus of these generic measures has given rise to debate around the extent to which they capture aspects of HRQoL of particular relevance to specific health conditions [5]. The assessment of QALYs in multiple sclerosis is one such case.

Multiple sclerosis (MS) is a neurological condition that affects the central nervous system. It is a complex and progressive condition causing a wide range of symptoms including spasticity, loss of mobility, fatigue, ataxia and loss of vision [6]. The incidence and severity of symptoms differ considerably between individuals and levels of disability increase as the disease progresses [7].

There is empirical evidence to suggest that generic measures may lack the relevance and sensitivity required to capture the many and varied effects of MS on people's HRQoL [6; 8; 9; 10], and that they have limited ability to capture changes in HRQoL across the full range of condition severity [8; 11; 12; 13; 14]. This raises concerns about the content validity of generic PBMs and the interpretability or meaningfulness of their scores when applied to MS [9]. An alternative would be to use a condition-specific preference-based measure (CSPBM).

CSPBMs focus on the aspects of health that are most relevant to the condition of interest, potentially providing greater sensitivity to differences and changes in HRQoL [1]. One approach is to develop a PBM from an existing condition-specific measure. This process has been reported for a range of

conditions, including dementia [15], mental health [16], asthma [17], flushing [18] and overactive bladder [19]. Here we describe the first stage in the development of a CSPBM for MS: deriving a health state classification system from the MS Impact Scale (MSIS-29), a widely used measure of HRQoL in MS with strong psychometric properties [20]. We begin by summarising the basis upon which we selected the MSIS-29, followed by methods for development of the classification system and results.

Measures of HRQoL for MS

Taking as a starting point that only patient-reported measures of HRQoL provide a suitable basis for the development of a classification system for a CSPBM [21], a systematic search of the literature was undertaken to identify existing MS-specific, patient-reported HRQoL instruments. The search identified 13 published reviews of HRQoL measures in MS, from which 17 individual measures were extracted. The existing literature [22; 23; 24**;** 25; 26; 27] was used to develop criteria for assessing the quality of these 17 instruments. These criteria (Table 1) defined our pre-requisites for any potential candidate measure for the CSPBM. A two-stage approach was used, firstly applying five initial criteria to narrow down the selection without need for detailed comparison of measures. Secondly, remaining measures were compared against the remaining selection criteria.

At stage one 14 measures were excluded [28; 29; 30; 31; 32; 33; 13; 34; 35; 36; 37; 38]. Exclusions were primarily due to the development methodology not incorporating qualitative research with patients (NeuroQoL, MSQLI, RAYS, DIP, HRQOL-MS, MS ADL, MSQoL-54, HAQUAMS, FILMS, QLI-MS) and/or recognised scale development techniques (MSQLI, QOLQ for MS, RAYS, HAQUAMS, MSSID).

Three candidate instruments remained after stage one: the MSIS-29 [23], the MS International Quality of Life questionnaire (MusiQoL) [39] and the Functional Assessment of MS (FAMS) [40]. On the basis of practical considerations we decided not to progress further with the MusiQoL: its limited use in clinical trials to date restricted the availability of evidence to support its responsiveness and acceptability [11]. Although we decided not to progress further with the MusiQoL, consideration of this instrument could be a productive area for future research.

At stage two, the MSIS-29 and the FAMS were considered in terms of the remaining criteria. Validation studies have confirmed the acceptability, reliability, validity and responsiveness of the MSIS-29 [41; 42; 43; 44; 45; 46] and the FAMS [47; 48; 40; 49; 50; 51; 52; 53] for a range of MS types and clinical settings. Both instruments are well accepted by clinicians and researchers and have frequently been used in research and clinical trials [54]. Overall, there was more published evidence describing the psychometric properties of the MSIS-29, while validation studies that directly compared the MSIS-29 with corresponding subscales of the FAMS suggest that the former may be superior in terms of acceptability,

internal consistency and responsiveness [54; 55; 56]. Exploratory analyses, assessing both instruments by applying the techniques required to derive a classification system, identified a range of problems with the FAMS (outlined in Appendix 1). Therefore we selected the MSIS-29 to form the basis of the classification system for an MS-specific PBM.

Methods

Typically HRQoL measures include a large number of items and levels. This would result in unreasonable cognitive demands on respondents to the valuation exercise required to estimate quality weights. Therefore, the first stage of deriving a CSPBM involves reducing the size of the existing measure to produce a health state classification system that is amenable to valuation, while minimising the loss of descriptive information [22]. This study adopted a five-stage process [5]:

1. Establish dimensions
2. Eliminate poorly performing items
3. Select one item to represent each dimension
4. Explore item-level reduction
5. Validate the analysis

The Multiple Sclerosis Impact Scale (MSIS-29)

The MSIS-29 consists of a physical subscale of 20 items and a psychological subscale of 9 items. Respondents are requested to report the impact of MS on their day-to-day lives over the preceding two weeks. The amended version, MSIS-29-v2, was used; this has four response levels per item: 'not at all', 'a little', 'moderately' and 'extremely' [20].

Dataset for analysis

The South West Impact of Multiple Sclerosis (SWIMS) project is a longitudinal cohort study of adults with a clinical diagnosis of MS or clinically isolated syndrome, living in Devon and Cornwall. Participants complete questionnaire packs, which include various generic and condition-specific measures, and other clinical and demographic data. The demographic make-up of respondents is consistent with other published UK data and clinical experience [57]. We randomly split an extract of SWIMS baseline data into a development dataset (n=529) and a validation dataset (n=528), providing suitable sample sizes for Rasch analysis. Table 2 reports the descriptive statistics for each dataset.

Analysis

The objective of the analysis was to derive a multi-dimensional, patient-reported health state classification system amenable to valuation. The aim was to reduce the number of items in the MSIS-29 by selecting one item to represent each of the dimensions of HRQoL that were covered by the MSIS-29. Rasch analysis was undertaken using RUMM2030 software, and psychometric analysis using SPSS.

Step 1: Establish dimensions

Exploratory factor analysis was used to investigate the factor structure of the MSIS-29. Each factor included items that represented more than one conceptually distinct dimension of HRQoL. For example, the physical subscale included items that described impacts on social activities, as well as on physical functioning. To address this, a conceptual framework was constructed, based on reviewed literature, to reflect the main dimensions of HRQoL in MS. Particular attention was paid to studies that directly involved people with MS. The items of the MSIS-29 were fitted to this conceptual framework, enabling items to be selected to represent the dimensions of HRQoL that are important to people with MS.

Step 2: Item elimination

Rasch analysis provides a technique by which ordinal data can be converted to continuous data. Unidimensional measures capture an underlying trait (in this case, HRQoL or a particular dimension of HRQoL), which is represented by a latent scale. Individual respondents are located along this scale according to their levels of HRQoL. Similarly, item response levels are located along the same scale according to the level of HRQoL that they represent [58]. Using Rasch methods a number of tests can assess how well individual items represent the underlying construct [59], hence providing a means of assessing the suitability of items for a classification system.

For each subscale of the MSIS-29 a partial credit polytomous Rasch model was fitted and used to assess items in terms of item-level ordering, differential item functioning and goodness of fit to the Rasch model.

Item-level ordering: disordered thresholds

The item-threshold map for each Rasch model was examined to identify items with disordered thresholds. The threshold between adjacent item responses is defined as the point on the latent scale at which either response is equally probable. Ordered thresholds imply that more severe responses have a higher probability of endorsement at lower levels of HRQoL. Disordered thresholds indicate that respondents are unable to distinguish between item response levels [59]. In this case, adjacent response levels are collapsed and, whilst the item should be retained in the Rasch model, it is not suitable for inclusion in the health state classification [19].

Differential item functioning

When responses to an item differ between groups of respondents, this is known as differential item functioning (DIF) [58]. We examined item characteristic curves and DIF summary tables to identify items that exhibited DIF by sex, age group, duration of disease or type of MS. Items exhibiting uniform DIF, where the difference in responses between groups is consistent across the latent scale, should be adjusted by splitting the item by the relevant person factor, creating two separate items [59], which are retained in the Rasch models but not considered for inclusion in the classification system [60].

Model and item goodness of fit

Inclusion of respondents who do not fit the expectations of the Rasch model can cause apparent item misfit, therefore all individuals with a fit residual > |2.5| were removed from the analysis [58, 59].

We applied three tests to examine how well the observed data fit the expectations of the Rasch model:

- Item-trait interaction: non-significant model χ^2 statistic ($p > 0.01$) [58]

- Overall item and person fit: mean item and person fit residuals will be close to zero with standard deviations close to one [59]
- Internal consistency: Person Separation Index (PSI) greater than 0.70 [45].

Overall goodness of fit may be improved by removing individual items that do not fit the model, ie items with fit residuals > |2.5| and significant χ^2 values ($p < 0.05$, adjusted using Bonferroni corrections for multiple tests) [59].

Items were adjusted or removed one at a time, and the analysis was re-run following each change. Items exhibiting disordered thresholds, DIF or misfit to the Rasch model were eliminated from consideration [5].

Step 3: Item selection

The next step was to select the most appropriate item to represent each conceptual dimension of HRQoL. An important feature of a classification system is its ability to span the full range of condition severity. In Rasch analysis, this is represented by a wide spread across the latent space. We judged this using item maps and the spread of response levels at logit zero on each item's threshold probability curves. Individual item goodness of fit statistics (fit residuals close to zero and non-significant χ^2) were also taken into account, as were four psychometric criteria: feasibility (item missing data); internal consistency (Cronbach's α); distribution of responses (floor and ceiling effects); and discriminant validity as a proxy measure of representativeness, using an independent samples t-test to assess the item's ability to distinguish between two sets of respondents, grouped on the basis of their scores on the Expanded Disease Status Scale (EDSS), a clinical measure of disease progression in MS. Preference was given to items that spanned the full range of severity [17].

Step 4: Item-level reduction

Rasch analysis can identify response levels that may be merged without losing descriptive information, offering a further means of simplifying the classification system [19]. Threshold probability curves that cross, or that come close to crossing, represent levels that could be merged [22].

Step 5: Validation

In order to validate the results of the analysis, steps 1 to 4 of the process were repeated using the validation dataset [5]. Two additional tests were undertaken using Rasch analysis. The first employed paired *t*-tests to confirm the unidimensionality of the final models. The percentage of tests significant at the $p < 0.05$ level should not exceed 5% [16]. The second employed residual correlation matrices to identify any local dependency between items. Correlations higher than the average correlation plus 0.2 indicate possible redundancy [61].

Results

Step 1: Establish dimensions

In the exploratory principle components factor analysis of the development dataset, a Varimax rotation produced four factors with Eigenvalues > 1, which explained 66% of the total variance. None of the items loaded most strongly on the fourth component, and there was no conceptual basis for grouping the four items that loaded on the third factor (IS07: Stiffness; IS09: Tremor of arms/legs, IS10: Spasms in limbs; IS22: Problems sleeping) into a separate domain. Allocating these four items to the factor on which they had the second highest loading resulted in a structure that mirrored the original structure of the MSIS-29, with items 1-20 forming one factor (the physical subscale) and items 21 to 29 forming the other (the psychological subscale). This two-factor solution, which explained 50% of the total variance, was supported by the scree plot (presented in Appendix Two).

Figure 1 shows the conceptual framework developed for this analysis, which includes physical, psychological and social impacts of MS on people's HRQoL.

The allocation of MSIS-29 items to these conceptual dimensions is shown in Table 3. The statistically confirmed factors of the MSIS-29 fitted well with the conceptual dimensions: the physical subscale included all items relating to physical and social aspects of HRQoL, and the psychological subscale included all items relating to the impact of non-physical symptoms. Not all domains of the conceptual framework were covered; no measure can realistically include all possible dimensions [22]. Three items (IS18, IS19, IS21) did not fit the conceptual framework, indicating that these items do not represent a predefined aspect of HRQoL in MS and should be excluded from selection.

Step 2: Item elimination

Table 3 summarises the results of the item elimination analysis. More detail is provided in Appendix 2. No items exhibited disordered thresholds. Five items from the physical subscale and one from the psychological subscale exhibited uniform DIF. Thirty-five and 22 respondents were removed from the physical and psychological subscale models respectively due to misfit to the Rasch model. Initial overall fit statistics for both subscales indicated poor fit to the Rasch model. Eight items misfit the model for the

physical subscale, and two misfit the psychological subscale. Removing these items produced good overall fit to both models.

At the end of the item elimination phase, five conceptual dimensions were represented by one item each: General/ other social/ role functioning (IS13); Employment (IS16); Fatigue (IS23); Cognition (IS27); Depression (IS29). A further three dimensions each had two remaining items: General/ other physical functioning (IS01 and IS11); Mobility (IS14 and IS17); General/other mental/ emotional wellbeing (IS24 and IS26). Three dimensions were no longer represented, because their constituent items had been eliminated: Independence (IS12); Bladder/ bowel function (IS20); Sleep quality (IS22).

Step 3: Item selection

The aims of the item selection phase were to confirm the suitability of the items remaining as the sole representative of a dimension, and to decide which items should be selected to represent the General/ other physical functioning, Mobility, and General/ other mental wellbeing dimensions. The results are summarised in Table 4.

All items that remained as the sole representative of a dimension had adequate spread across the latent space and well-spaced threshold probability curves at logit zero. Items IS13 and IS16 performed well across all criteria; IS23 and IS27 failed to meet the threshold for internal consistency but performed well against the other criteria; IS29 struggled against some criteria, but exhibited the strongest internal consistency of any item from the psychological subscale.

General/ other physical functioning: IS01 showed a wider spread across the latent space than IS11, and performed well on all criteria. IS11 had better spaced threshold probability curves, but had a high fit residual and a relatively high proportion of missing data.

Mobility: Although IS14 and IS17 had equivalent spread across the latent space, the thresholds of item IS14 spanned logit zero whereas all thresholds for item IS17 were above logit zero, and the threshold probability curves for item IS14 were more widely spaced. IS14 had a high fit residual whereas IS17 had a large ceiling effect.

General/ other mental wellbeing: Item IS26 showed a wider spread of levels across the latent space, better spaced threshold probability curves, and good performance across all criteria. Item IS24 had a high fit residual, significant p-value and poor internal consistency.

These results supported the selection of items IS01, IS13, IS14, IS16, IS23, IS26, IS27 and IS29 for the classification system.

Step 4: Item-level reduction

Threshold probability curves provided no evidence to suggest that the number of item levels could be reduced.

Step 5: Validation

The analysis was repeated using the validation dataset (detailed results are presented in Appendix 3). The only difference was that item IS12, representing the Independence dimension, passed all item elimination tests during analysis of the validation dataset, but was eliminated during analysis of the development dataset due to DIF: people who had MS for ten or more years reported that they were less bothered by "having to depend on others" than would be expected compared to those in the lower duration groups. For people with severe MS, research suggests that support from others can either increase or decrease their sense of independence [13], providing a possible explanation for the DIF observed in the development dataset. Therefore this item was excluded from the classification system.

In order to test the impact of the unallocated items (IS18, IS19, IS21), we repeated the analysis with these items excluded. This made no difference to the results. Using the Rasch test of unidimensionality in the development dataset, 2.25% of paired t-tests were significant for the physical subscale and 2.51% were significant for the psychological subscale. In the validation dataset, 3.08% were significant for the physical subscale and 2.68% for the psychological subscale. This supported the unidimensionality of all four models. Residual correlations were examined between the items selected for the classification system. In the development dataset, no local dependency was apparent. In the validation dataset, we found a correlation between items IS13 and IS14. These items represent different dimensions of HRQoL

in MS and were not correlated in the development dataset, therefore both were included in the classification system.

The MSIS-8D classification system

Analysis of both datasets produced a classification system comprised of eight items, each of which represents one of the following conceptual dimensions of HRQoL in MS: general physical function, mobility, employment, social function, fatigue, cognition, depression and general emotional wellbeing. Each item has four levels. In total, the MSIS-8D classification system (Figure 2) describes 65,536 health states.

Discussion

We describe the first stage in developing a CSPBM for MS, presenting the MSIS-8D. Building on strong research methodology [5], we have derived the MSIS-8D classification system from an existing HRQoL measure, the MSIS-29. The MSIS-8D covers important dimensions of HRQoL in MS and is suitable for use in a valuation survey. The next stage of the research will involve preference elicitation and related regression-based statistical modelling to derive quality weights for all health states described by the MSIS-8D. This will result in a CSPBM that that is capable of generating health state values for the estimation of QALYs, for use in health policy settings including the economic evaluation of treatments for MS.

We present a strong rationale for the selection of the MSIS-29 as the basis for this MS-specific PBM. All available measures of HRQoL in MS have some limitations, but the MSIS-29 emerged as one of the strongest candidates. Developing a CSPBM from an existing measure of HRQoL offers a number of advantages. Adapting a well-accepted and frequently used measure, such as the MSIS-29, enables retrospective analyses to be undertaken using existing data and increases the likelihood that the measure will be used in future studies [22].

Both subscales of the MSIS-29 contained items that represented different dimensions of HRQoL. We developed a novel approach to deal with this: analysing the relevant literature to build a conceptual framework of HRQoL in MS, to which the items of the instrument were mapped, ensuring that the main conceptually independent dimensions of HRQoL were represented in the classification system. This builds on previous research, where the original dimensional structure of an instrument has been used to guide the selection of items, despite a lack of statistical independence between dimensions [16].

The use of condition-specific measures to inform economic evaluation has generated some debate [1; 5; 22; 62; 63; 64]. Some commentators argue that, in order to compare the results of different economic evaluations, health outcomes must be assessed using the same classification system. This requirement, however, is not found in other areas of economics or in the earlier QALY literature. Brazier et al [5] suggest that, provided the same preference elicitation methods are used to obtain quality weights, comparability can be achieved between different classification systems. This view has informed the

methods used to develop the MSIS-8D. Notwithstanding this, some problems with comparability remain, and these arise largely due to the limited coverage of CSPBMs relative to generic measures. CSPBMs may be incapable of capturing side effects of interventions that fall outside of the dimensions covered by the classification system, or of picking up impacts on co-morbidities. They may be prone to focusing effects, where the impact of the condition is overestimated because respondents to the valuation survey concentrate solely on the dimensions included in the classification system rather than viewing them in a wider context. Respondents may take into account aspects of health that are excluded from the classification system, potentially influencing their preferences between health states and affecting the survey results. Another concern is the relationship between perfect health and the best possible state described by the classification system. It is feasible for a person to attain the best possible health state according to a specific instrument, but to have other health problems not covered by its classification system. The instrument-specific nature of 'best possible' health states makes it difficult to compare results between different PBMs [5].

These disadvantages are arguably less important when the condition of interest is the dominant factor in determining HRQoL [22], as is likely to be the case for people with MS. In addition, the varied impacts of MS on HRQoL have resulted in the MSIS-8D classification system becoming somewhat broader than many other CSPBMs. Deciding whether to develop or use a CSPBM invariably involves a trade-off between the advantages and disadvantages of CSPBMs in relation to the condition of interest [5]. In the case of MS, the potential limitations of existing generic measures, the broad scope of the MSIS-8D classification system and the likely dominant nature of MS in determining HRQoL all support the development and use of a CSPBM.

Research is underway to estimate tariff of quality weights for all MSIS-8D health states. A reliable and valid CSPBM for MS will be a valuable addition to the methods available for the estimation of QALYs for MS health states, to support the assessment of HRQoL and the economic evaluation of treatments for people with MS.

References

1. Brazier J and Tsuchiya A, 2010. Preference-based condition-specific measures of health: what happens to cross programme comparability? Health Econ 19:125

2. Dolan P, 1997. Modeling valuations for EuroQol health states. Med Care 35: 1095-1108

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

4. Horsman J, Furlong W, Feeny D and Torrance G, 2003. The Health Utilities Index (HUI): concepts, measurement properties and applications. Health Qual Life Out 1:54

5. Brazier JE, Rowen D, Mavranezouli I, Tsuchiya A, Young T, Yang Y, Barkham M and Ibbotson R, 2012. Developing and testing methods for deriving preference-based measures of health from condition-specific measures (and other patient-based measures of outcome). Health Technol Assess 16(32)

6. Hemmett L, Holmes J, Barnes M and Russell N, 2004. What drives quality of life in multiple sclerosis? QJM 97:671

7. Zajicek J, Freeman J and Porter B, 2007. Multiple Sclerosis Care: a practical manual. Oxford: Oxford University Press

8. Fisk JD, Brown MG, Sketris IS, Metz LM, Murray TJ and Stadnyk KJ, 2005. A comparison of health utility measures for the evaluation of multiple sclerosis treatments. J Neurol Neurosurg Psychiatry 76:58

9. Kuspinar A, Mayo N. Do generic utility measures capture what is important to the quality of life of people with multiple sclerosis? Health Qual Life Out. 2013;11(71)

10. Orme M, Kerrigan J, Tyas D, Russell N, Nixon R. The effect of disease, functional status, and relapses on the utility of people with multiple sclerosis in the UK. Value Health. 2007;10(1):54-60

11. Bandari, DS, Vollmer TL, Khatri BO and Tyry T, 2010. Assessing Quality of Life in Patients with Multiple Sclerosis. Int J MS Care 12:34

12. Benito- León J, Morales JM, Rivera-Navarro J and Mitchell AJ, 2003. A review about the impact of multiple sclerosis on health-related quality of life. Disabil Rehabil 25:1291

13. Gruenewald DA, Higginson IJ, Vivat B, Edmonds P and Burman RE, 2004. Quality of life measures for the palliative care of people severely affected by multiple sclerosis: a systematic review. Mult Scler 10:690

14. Opara JA, Jaracz K and Brola W, 2010. Quality of life in multiple sclerosis. J Med Life 3:352

15. Mulhern B, Rowen D, Brazier J, Smith S, Romeo R, Tait R, et al. Development of DEMQOL-U and DEMQOL-PROXY-U: Generation of preference-based indices from DEMQOL and DEMQOL-PROXY for use in economic evaluation. Health Technol Assess. 2013;17(5).

16. Mavranezouli I, Brazier JE, Young TA and Barkham M, 2011. Using Rasch analysis to form plausible health states amenable to valuation: the development of CORE-6D from a measure of common mental health problems. Qual Life Res 20:321

17. Young T, Yang Y, Brazier J and Tsuchiya A, 2011. The use of Rasch analysis in reducing a large condition-specific instrument for preference valuation: the case of moving from AQLQ to AQL-5D. Med Decis Making 31:195

18. Young TA, Rowen D, Norquist J and Brazier JE, 2010. Developing preference-based health measures: using Rasch analysis to generate health state values. Qual Life Res 19:907

19. Young T, Yang Y, Brazier J, Tsuchiya A and Coyne K, 2009. The first stage of developing preference-based measures: constructing a health-state classification using Rasch analysis. Qual Life Res 18:253

20. Hobart J and Cano S, 2009. Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. Health Technol Assess 13 (12)

21. Brazier JE and Rowen D, 2011. NICE DSU Technical Support Document 11: Alternatives to EQ-5D for generating health state utility values. http://www.nicedsu.org.uk

22. Brazier J, Ratcliffe J, Salomon JA and Tsuchiya A, 2007. Measuring and Valuing Health Benefits for Economic Evaluation. Oxford: Oxford University Press

23. Hobart JC, Riazi A, Lamping DL, Fitzpatrick R and Thompson AJ, 2004. Improving the evaluation of therapeutic interventions in multiple sclerosis: development of a patient-based measure of outcome. Health Technol Assess 8 (9)

24. Murrell R, 1999. Quality of Life and Neurological Illness: A Review of the Literature. Neuropsychol Rev 9:209

25. O'Connor R, 2004. Measuring Quality of Life in Health. Edinburgh: Churchill Livingstone

26. Riazi A, 2006. Patient-reported Outcome Measures in Multiple Sclerosis. Int MS J 13:92

27. Streiner DL and Norman GR, 2003. Health Measurement Scales: A Practical Guide to their Development and Use. 3rd edition; Oxford: Oxford University Press

28. Doward LC, McKenna SP, Meads DM, Twiss J and Eckert BJ, 2009. The development of patient-reported outcome indices for multiple sclerosis (PRIMUS). Mult Scler 15:1092

29. Ferrans C, Powers M. Quality of Life Index: Development and psychometric properties. ANS Adv Nurs Sci. 1985;8:15-24.

30. Ford HL, Gerry E, Tennant A, Whalley D, Haigh R and Johnson MH, 2001: Developing a disease-specific quality of life measure for people with multiple sclerosis. Clin Rehabil 15:247

31. Freeman JA; Hobart JC and Thompson AJ, 2001. Does adding MS-specific items to a generic measure (the SF-36) improve measurement? Neurology 57:68–74

32. Gold SM, Heesen, C, Schulz H, Guder U, A Mönch A, Gbadamosi J, Buhmann C and Schulz KH, 2001. Disease specific quality of life instruments in multiple sclerosis: Validation of the Hamburg Quality of Life Questionnaire in Multiple Sclerosis (HAQUAMS). Mult Scler 7:119

33. Greenhalgh J, Ford H, Long AF and Hurst K, 2004. The MS Symptom and Impact Diary (MSSID): psychometric evaluation of a new instrument to measure the day to day impact of multiple sclerosis. J Neurol Neurosurg Psychiatry 75:577

34. Neuro-QOL Project, 2010. Measuring Quality of Life in Neurological Disorders: Final Report of the Neuro-QOL Study. Presented to the National Institute of Neurological Disorders and Stroke. http://www.neuroqol.org/Resources/Pages/default.aspx accessed 24/03/2014

35. Ritvo PG, Fischer JS, Miller DM, Andrews H, Paty DW, LaRocca NG, 1997. MSQLI Multiple Sclerosis Quality of Life Inventory: A User's Manual. New York: National Multiple Sclerosis Society

36. Rotstein Z, Barak Y, Noy S and Achiron A, 2000. Quality of life in multiple sclerosis: development and validation of the 'RAYS' scale and comparison with the SF-36. Int J Qual Health Care 12:511

37. Vickrey BG, Hays RD, Genowese BJ, Myers LW, and Ellison GW, 1997: Comparison of a generic to disease-targeted health-related quality-of-life measures for multiple sclerosis. J Clin Epidemiol 50:557

38. Wesson JM, Cooper JA, Jehle LS, Lockhart SN, Draney K and Barber J, 2009: The functional index for living with multiple sclerosis: development and validation of a new quality of life questionnaire. Mult Scler 15:1239

39. Simeoni MC, Auquier P, Fernandez O, et al on behalf of the MusiQoL study group, 2008: Validation of the Multiple Sclerosis International Quality of Life questionnaire. Mult Scler 14:219

40. Cella DF, Dineen K, Amason B, Reder A, Webster KA, Karabatsos G, Chang C, Lloyd S, Mo F, Stewart J, and Stefoski, D, 1996. Validation of the Functional Assessment of Multiple Sclerosis quality of life instrument. Neurology 47: 129

41. Costelloe L, O'Rourke K, Kearney H, McGuigan C, Gribbin L, Duggan M, Daly L, Tubridy N and Hutchinson M, 2007. The patient knows best: significant change in the physical component of the Multiple Sclerosis Impact Scale (MSIS-29 physical). J Neurol Neurosurg Psychiatry 78:841

42. Gray OM, McDonnell GV and Hawkins SA, 2009. Tried and tested: the psychometric properties of the multiple sclerosis impact scale (MSIS-29) in a population-based study. Mult Scler 15:75

43. Hoogervorst ELJ, Zwemmer JNP, Jelles B, Polman CH and Uitdehaag BMJ, 2004. Multiple Sclerosis Impact Scale: relation to established measures of impairment and disability. Mult Scler 10:569

44. McGuigan C and Hutchinson M, 2004. The multiple sclerosis impact scale (MSIS-29) is a reliable and sensitive measure. J Neurol Neurosurg Psychiatry 75:266

45. Ramp M, Khan F, Misajon RA and Pallant J, 2009. Rasch analysis of the Multiple Sclerosis Impact Scale (MSIS-29). Health Qual Life Out 7:58

46. Riazi A, Hobart JC, Lamping DL, Fitzpatrick R and Thompson AJ, 2002. Multiple Sclerosis Impact Scale (MSIS-29): reliability and validity in hospital based samples. J Neurol Neurosurg Psychiatry 73:701

47. Acaster S, Swinburn P, Wang C, Stemper B, Beckmann K, Knappertz V, Pohl C, Sandbrink R, Gondek K, Edan G, Kappos L, Freedman M, Hartung H-P, Arnason B, Comi G, Filippi M, Jeffery D, O'Connor P, Cook S and Lloyd AJ, 2011. Can the functional assessment of multiple sclerosis adapt to changing needs? A psychometric validation in patients with clinically isolated syndrome and early relapsing -remitting multiple sclerosis. Mult Scler 17: 1504

48. Benito-León J, Morales JM and Rivera-Navarro J, 2002. Health-related quality of life and its relationship to cognitive and emotional functioning in multiple sclerosis patients. Eur J Neurol 9: 497

49. Chang C-H, Cella D, Fernández O, Luque G, de Castro P, de Andrés C, Casanova B, Hernández MA, Prieto JM, Fernández VE and de Ramón E on behalf of Grupo Español de Calidad de Vida en Esclerosis Múltiple, 2002. Quality of life in multiple sclerosis patients in Spain. Mult Scler 8: 527

50. Kikuchi H, Mifune N, Niino M, Ohbu S, Kira J, Kohriyama T, Ota K, Tanaka M, Ochi H, Nakane S, Maezawa M and Kikuchi S, 2011. Impact and characteristics of quality of life in Japanese patients with multiple sclerosis. Qual Life Res 20:119

51. Modrego PJ, Pina MA, Simón A and Carmen Azuara M, 2001. The Interrelations Between Disability and Quality of Life in Patients with Multiple Sclerosis in the Area of Bajo Aragon, Spain: A Geographically Based Survey. Neurorehabil Neural Repair 15:69

52. Nicholl CR, Lincoln NB, Francis VM and Stephan TF, 2001. Assessing quality of life in people with multiple sclerosis. Disabil Rehabil 23:597-603

53. Patti F, Russo P, Pappalardo A, Macchia F, Civalleri L and Paolillo A for the FAMS study group, 2007. Predictors of quality of life among patients with multiple sclerosis: An Italian cross-sectional study. J Neurol Sci 252:121

54. Giordano A, Pucci E, Naldi P, Mendozzi L, Milanese C, Tronci F, Leone M, Mascoli N, La Mantia L, Giuliani G and Solari A, 2009. Responsiveness of patient reported outcome measures in multiple sclerosis relapses: the REMS study. J Neurol Neurosurg Psychiatry 80:1023

55. Hobart JC, Riazi A, Lamping DL, Fitzpatrick R and Thompson AJ, 2005. How responsive is the Multiple Sclerosis Impact Scale (MSIS-29)? A comparison with some other self report scales. J Neurol Neurosurg Psychiatry 76:1539

56. Riazi A, Hobart JC, Lamping DL, Fitzpatrick R and Thompson AJ, 2003. Evidence-based measurement in multiple sclerosis: the psychometric properties of the physical and psychological dimensions of three quality of life rating scales. Mult Scler 9:411

57. Zajicek JP, Ingram WM, Vickery J, Creanor S, Wright DE, Hobart JC, 2010. Patient-orientated longitudinal study of multiple sclerosis in south west England (The South West Impact of Multiple Sclerosis Project, SWIMS) protocol and baseline characteristics of cohort. BMC Neurology 10:88

58. Tennant A and Conaghan PG, 2007. The Rasch Measurement Model in Rheumatology: What Is It and Why Use It? When Should It Be Applied, and What Should One Look for in a Rasch Paper? Arthritis & Rheumatol 57:1358

59. Pallant J and Tennant A, 2007. An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). Br J Clin Psychol 46:1

60. Hagquist C, Bruce M and Gustavsson J, 2009. Using the Rasch model in nursing research: an introduction and illustrative example. Int J Nurs Stud 46:380

61. Mulhern B, Smith SC, Rowen D, Brazier JE, Knapp M, Lamping DL, et al, 2012. Improving the measurement of QALYs in dementia: developing patient- and carer-reported health state classification systems using Rasch analysis. Value Health;15(2):323-33

62. Brazier J, Dixon S, 1995. The use of condition specific outcome measures in economic appraisal. Health Econ;4: 255–264

63. Dowie J, 2002. Decision validity should determine whether a generic or condition-specific HRQoL measure is used in health care decisions. Health Econ;11:1-8

64. Gold MR, Siegel JE, Russell LB, Weinstein MS, 1996: Cost Effectiveness in Health and Medicine. New York: Oxford University Press

Table 1: Criteria for selection of a health-related quality of life instrument

Acceptability	*Single instrument, rather than battery of measures Proportion of questionnaires completed Item missing data < 10% High percentage of computable scale scores Floor and ceiling effects < 20% per subscale Does the range of scores span the full scale range? Mean score near scale mid-point
Reliability	Internal consistency (Cronbach's α > 0.80) Test-retest reliability ($r \geq 0.50$)
Construct validity	Convergent validity (correlation $r > 0.70$) Discriminant validity (correlation $r > 0.30$ Group differences validity ($p < 0.05$)
Internal validity	Moderate correlations between subscales ($0.30 < r < 0.70$)
Responsiveness	Effect size: large (>0.80) or moderate (>0.50)
Scale development and scaling assumptions	*Recognised scale development techniques used to devise the instrument Similar mean scores and variances Similar response option frequency distributions Similar and substantial item–total correlations ($r > 0.30$) Item–total exceed item–other correlations by ≥2 standard errors Skewness (–1 to +1)
Content/ face validity	*The underlying concept captured by the instrument is HRQoL *Instrument was constructed on the basis of qualitative work with patients *Extent to which instrument covers domains important for HRQoL in MS
Practical considerations	Acceptability to clinicians/ researchers; use in clinical trials Access to a dataset that includes the measure

* Indicates that this was used as a screening criterion (stage one)

Table 2: Descriptive statistics for development and validation datasets

	Development (n=529)	Validation (n = 528)
Female	73%	74%
Male	27%	26%
Age under 50	47%	48.5%
Age 50 or over	53%	51.5%
Disease duration < 2 yrs	35%	33%
Disease duration 2 to 10yrs	29%	30%
Disease duration > 10 yrs	34%	31%
Diagnosis date not recorded	2%	6%
Progressive MS	20%	24%
Relapsing-remitting MS	27%	23%
Benign or mild MS	2%	3%
MS type not recorded	51%	50%

Table 3: Item elimination results (development dataset)

Subscale	Conceptual dimension	Code	Item description	Results
Physical	General/ other physical functioning	IS01	Do physically demanding tasks	✓
		IS02	Grip things tightly (e.g. turning on taps)	× DIF (gender)
		IS03	Carry things	× DIF (age)
		IS04	Problems with your balance	× DIF (MS type)
		IS06	Being clumsy	× Misfit
		IS07	Stiffness	× Misfit
		IS08	Heavy arms and/or legs	× Misfit
		IS09	Tremor of your arms or legs	× Misfit
		IS10	Spasms in your limbs	× Misfit
		IS11	Your body not doing what you want it to do	✓
		IS15	Difficulties using your hands in everyday tasks	× Misfit
	Mobility	IS05	Difficulties moving about indoors	× DIF (age)
		IS14	Being stuck at home more than you would like to be	✓
		IS17	Problems using transport (e.g. car, bus, train, taxi, etc)	✓
	Bladder/ bowel	IS20	Needing to go to the toilet urgently?	× Misfit

Table 3 continues overleaf

	General/ other social and role functioning	IS13	Limitations in your social and leisure activities at home	✓
	Independence	IS12	Having to depend on others to do things for you	× DIF (duration)
	Employment	IS16	Having to cut down the amount of time you spent on work or other daily activities	✓
	Unallocated items	IS18	Taking longer to do things	× Misfit
		IS19	Difficulty doing things spontaneously (eg going out on the spur of the moment)	× Unallocated
Psychological	General/ other mental and emotional wellbeing	IS24	Worries related to your MS	✓
		IS25	Feeling anxious or tense	× Misfit
		IS26	Feeling irritable, impatient, or short tempered	✓
		IS28	Lack of confidence	× DIF (MS type)
	Depression	IS29	Feeling depressed	✓
	Fatigue	IS23	Feeling mentally fatigued	✓
	Cognition	IS27	Problems concentrating	✓
	Sleep quality	IS22	Problems sleeping	× Misfit
	Unallocated items	IS21	Feeling unwell	× Unallocated

Table 3 continues overleaf

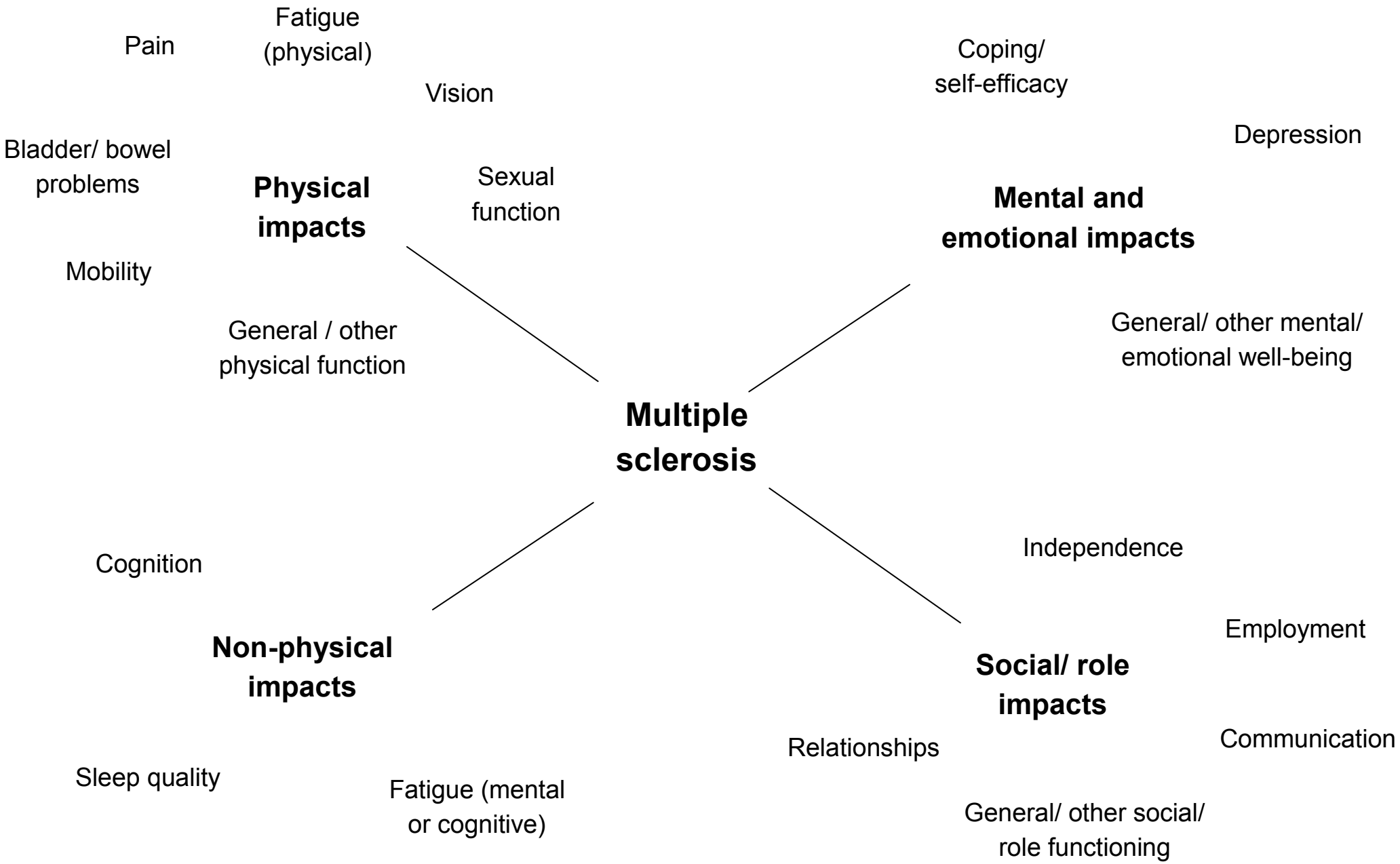
		Item fit residual		Person fit residual		p-value	PSI
Overall goodness of fit to Rasch models following item elimination:		**Mean**	**sd**	**Mean**	**Sd**		
	Physical subscale	-0.159	1.274	-0.265	0.963	0.438	0.892
	Psychological subscale	0.044	0.916	-0.259	0.989	0.069	0.794

✓ = item retained; × = item eliminated; DIF = differential item functioning; sd = standard deviation; PSI = person separation index

Table 4: Summary of Rasch analysis and psychometric criteria for item selection (development dataset)

	Location of item-level threshold on Rasch logit scale			Rasch criteria		Psychometric criteria				
Item (Dimension)	**Level 1-2**	**Level 2-3**	**Level 3-4**	**Fit residual**	***p* -value**	**Missing data %**	**Floor effect %**	**Ceiling effect %**	**Internal consistency**	**Discriminant validity**
IS01 (General physical)	-3.4	-0.8	0.0	-0.279	0.701	0.6	36.6	15.6	0.79	< 0.001
IS11 (General physical)	-1.2	0.4	1.4	2.351	0.281	4.2	16.4	33.4	0.77	< 0.001
IS13 (General social/ role)	-1.4	0.6	1.8	-0.876	0.319	2.8	14.8	33.7	0.80	< 0.001
IS14 (Mobility)	-0.8	0.4	1.0	-1.359	0.163	2.1	22.2	38.5	0.79	< 0.001
IS17 (Mobility)	0.2	1.0	1.6	-0.418	0.372	3.0	14.3	51.4	0.73	< 0.001
IS16 (Employment)	-2.2	0.2	1.0	0.124	0.620	3.4	22.7	25.9	0.78	< 0.001
IS23 (Fatigue)	-2.2	-0.4	0.2	-0.161	0.193	1.9	20.8	19.3	0.68	0.172
IS24 (General/ other EWB)	-1.6	0.6	0.8	1.824	0.045	2.1	12.4	28.0	0.60	0.034
IS26 (General/ other EWB)	-1.8	0.0	1.2	-0.498	0.153	2.5	12.0	27.6	0.70	0.042
IS27 (Cognition)	-1.4	0.2	1.0	0.182	0.514	2.5	11.7	31.1	0.66	0.002
IS29 (Depression)	-0.6	0.6	1.4	-1.339	0.016	3.2	8.0	44.8	0.71	0.005

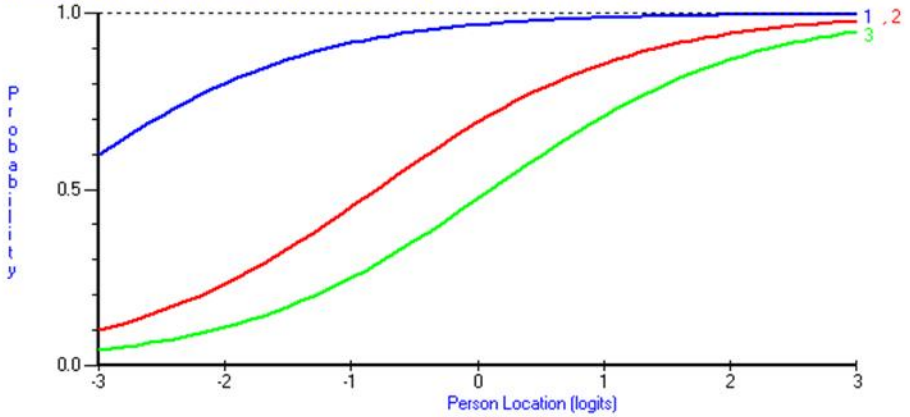
EWB = emotional well-being; internal consistency = corrected item-total (point biserial) correlation


Pain
Fatigue (physical)
Vision
Bladder/ bowel problems
Physical impacts
Sexual function
Mobility
General / other physical function
Coping/ self-efficacy
Depression
Mental and emotional impacts
General/ other mental/ emotional well-being
Multiple sclerosis
Cognition
Non-physical impacts
Sleep quality
Fatigue (mental or cognitive)
Independence
Employment
Social/ role impacts
Relationships
Communication
General/ other social/ role functioning

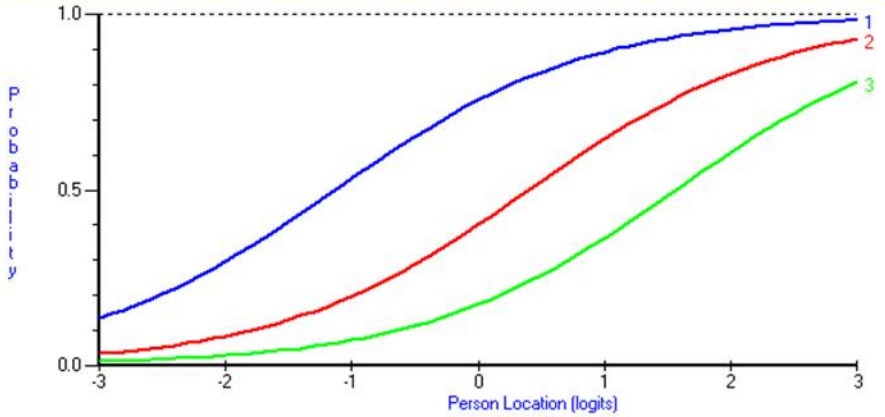
In the *past two weeks*, how much has your MS limited your ability to ...	Not at all	A little	Moderately	Extremely
Do physically demanding tasks?	1	2	3	4

In the *past two weeks*, how much have you been bothered by ...	Not at all	A little	Moderately	Extremely
Limitations in your social and leisure activities at home?	1	2	3	4
Being stuck at home more than you would like to be?	1	2	3	4
Having to cut down the amount of time you spent on work or other daily activities?	1	2	3	4
Feeling mentally fatigued?	1	2	3	4
Feeling irritable, impatient or short-tempered?	1	2	3	4
Problems concentrating?	1	2	3	4
Feeling depressed?	1	2	3	4

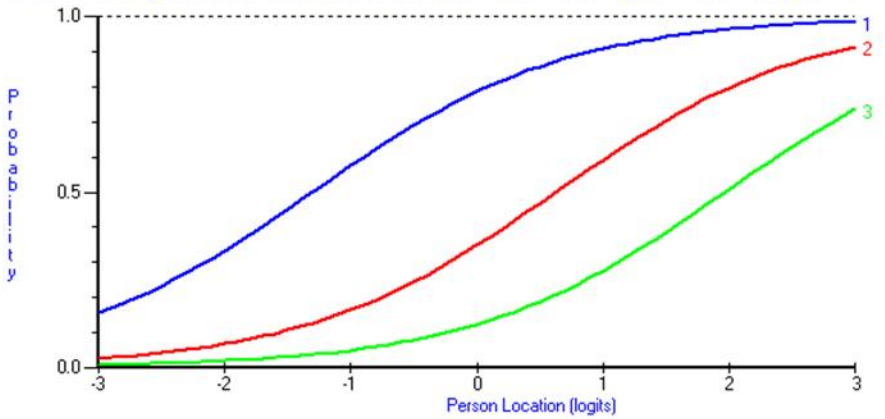
Threshold Probability Curves: IS01 Physically demanding tasks Locn = -1.361 Spread = 0.876 SampleN = 425



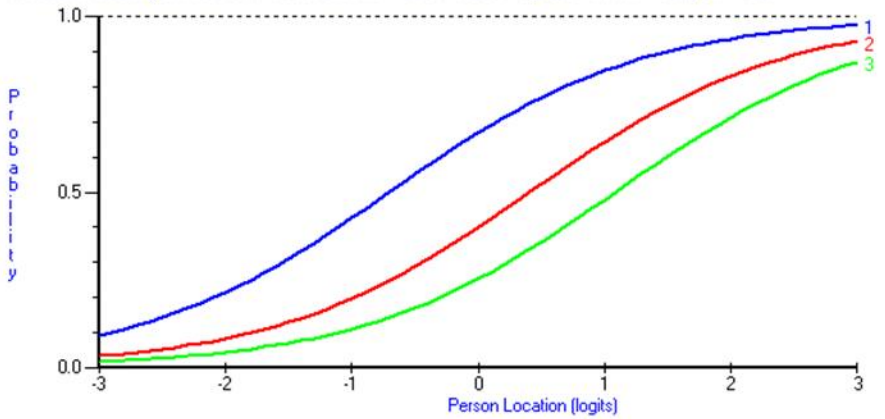
Threshold Probability Curves: IS11 Body not doing what you want i Locn = 0.280 Spread = 0.674 SampleN = 425



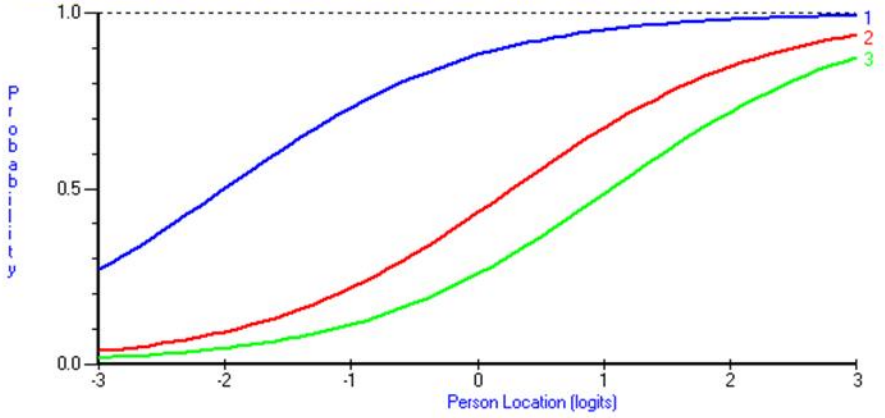
Threshold Probability Curves: IS13 Social and leisure limitations Locn = 0.431 Spread = 0.819 SampleN = 425



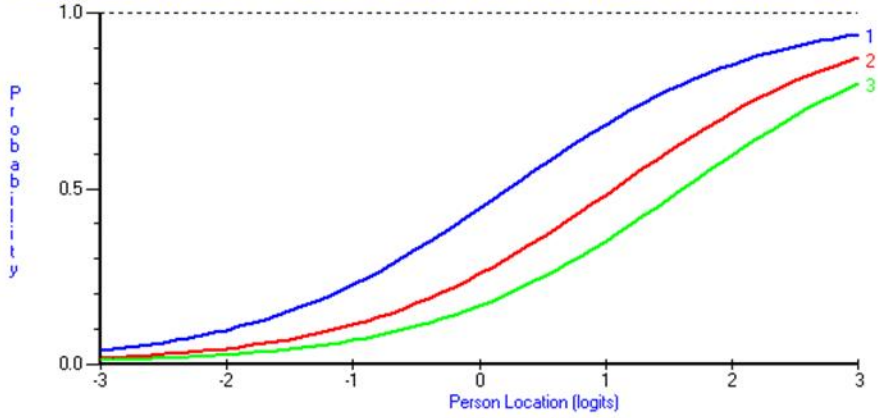
Threshold Probability Curves: IS14 Stuck at home Locn = 0.265 Spread = 0.448 SampleN = 425



Threshold Probability Curves: IS16 Time spent on daily activities Locn = -0.219 Spread = 0.765 SampleN = 425



Threshold Probability Curves: IS17 Problems using transport Locn = 0.973 Spread = 0.346 SampleN = 425



Appendix 1: Factor and Rasch analysis of the FAMS

The published six-dimensional structure of the FAMS was not supported by factor analysis, which instead suggested three alternative versions, with one, three or eight dimensions. Neither the three nor the eight factor version was compatible with the conceptual framework developed for this research. For example items relating to social and role functioning were spread across more than one factor. A separate Rasch analysis was conducted for each factor version of the FAMS. In all three versions, a high proportion of items exhibited disordered thresholds. Respondents had particular difficulty distinguishing between two of the intermediate levels ("somewhat" and "quite a bit"). None of the versions resulted in good overall fit to the Rasch model for all dimensions. In some cases, even where overall model goodness of fit was achieved, no items had survived the item elimination phase unaltered for disordered thresholds or DIF. We concluded, therefore, that the FAMS is unsuitable for use as the basis of a classification system. Details of the analysis of the FAMS can be made available on request.

Appendix Two: Detailed results of analysis using development dataset

Figure 3: MSIS-29 scree plot (development dataset)

Figure 4: Threshold probability curves for physical subscale items (development dataset)

Figure 5: Threshold probability curves for psychological subscale items (development dataset)

Table 5: Detailed results for items eliminated from analysis of the development dataset

MSIS-29 subscale	Item	Conceptual dimension	Items with differential item functioning			Fit statistics for misfitting items		
			Person factor (CIs)	p-value	p-value threshold	Fit residual	p-value	p-value threshold
Physical	IS02	General physical	Gender (3)	0.000022	0.000833			
	IS03	General physical	Age (5)	0.000368	0.000794			
	IS05	Mobility	Age (5)	0.000578	0.000806			
	IS04	General physical	MS type (2)	0.000728	0.000667			
	IS12	Independence	Duration (3)	0.000096	0.000725			
	IS20	Bladder/ bowel				7.241	0.000000	0.001923
	IS18	Unallocated				-5.188	0.000005	0.002000
	IS09	General physical				4.115	0.000000	0.002083
	IS07	General physical				4.165	0.000045	0.002174
	IS10	General physical				4.632	0.000000	0.002273
	IS08	General physical				3.577	0.000721	0.002381
	IS12 2-10yrs					-3.010	0.070653	0.002500
	IS05 younger					-2.763	0.025346	0.002632
	IS02 female					2.125	0.001289	0.002778

	IS02 male					2.099	0.002449	0.002941
	IS06	General physical				2.894	0.016438	0.003333
	IS15	General physical				2.581	0.047698	0.003125
	IS03 older					2.527	0.230317	0.003571
Psychological	IS28	General/ other EWB	MS type (3)	0.000210	0.001852			
	IS22	Sleep quality				5.906	0.000000	0.005000
	IS25	General/ other EWB				-3.542	0.000078	0.005556
CIs = class intervals; threshold p-value = equivalent of $p < 0.05$ after Bonferroni adjustment; EWB = emotional well-being								

Appendix 3: Results of item selection and elimination, using the validation dataset

Table 6: Detailed results for items eliminated from analysis of the validation dataset

MSIS-29 scale	Item	Conceptual dimension	Items with differential item functioning			Fit statistics for misfitting items		
			Person factor (CIs)	p-value	p-value threshold	Fit residual	p-value	p-value threshold
Physical	IS09	General physical	Age (5)	0.000139	0.000833			
	IS20	Bladder/ bowel				9.083	<0.000001	0.002381
	IS18	Unallocated				-5.562	0.000001	0.002500
	IS05	Mobility	MS type (3)	0.000002	0.000794	-4.469	0.000002	0.002632
	IS10	General physical				3.457	0.000004	0.002778
	IS07	General physical				3.864	0.002235	0.002941
	IS08	General physical				3.820	0.000013	0.003125
	IS09 younger					3.695	0.000001	0.003333
	IS09 older					2.828	0.000006	0.003571
	IS04	General physical				3.195	0.013118	0.003846
	IS06	General physical				2.861	0.003464	0.004167
	IS02	General physical				2.897	0.023579	0.004545
	IS15	General physical				2.631	0.000968	0.005000

Psychological	IS22	Sleep quality	Age (5)	0.001639	0.001852			
	IS28	General/ other EWB	MS type (3)		0.001667			
	IS22 older					4.331	<0.000001	0.004545
	IS22 younger					3.009	0.001478	0.005000
	IS25	General/ other EWB				-2.721	0.003305	0.005556

Overall goodness of fit to Rasch models following item elimination:		**Item fit residual**		**Person fit residual**		**p-value**	**PSI**
		Mean	**sd**	**Mean**	**Sd**		
	Physical scale	-0.255	1.404	-0.294	1.096	0.092963	0.88707
	Psychological scale	-0.053	1.257	-0.230	0.949	0.044616	0.78107

CIs = class intervals; threshold p-value = equivalent of $p < 0.05$ after Bonferroni adjustment; EWB = emotional well-being

p-value = item-trait interaction χ^2 p-value; PSI = person separation index

Figure 6: Threshold probability curves for physical subscale items (validation dataset)

Figure 7: Threshold probability curves for psychological subscale items (validation dataset)

Table 7: Results of Rasch analysis and psychometric criteria for item selection, using the validation dataset

	Location of items on Rasch logit scale			Rasch criteria		Psychometric criteria			
Item	**Threshold level 1-2**	**Threshold level 2-3**	**Threshold level 3-4**	**Fit residual**	***p* -value**	**Missing data %**	**Floor effect %**	**Ceiling effect %**	**Internal consistency**
IS01	-3.2	-1.0	0.4	0.004	0.872	1.1	30.5	15.6	0.78
IS03	-2.0	-0.2	1.4	0.908	0.802	1.7	16.9	27.6	0.80
IS11	-1.2	0.4	1.4	1.950	0.662	2.3	15.6	34.8	0.78
IS12	-1.6	0.4	1.2	-1.867	0.069	1.3	17.6	33.1	0.80
IS13	-1.4	0.4	1.8	-1.985	0.113	2.1	13.8	35.3	0.77
IS14	-1.0	0	0.8	-1.445	0.104	0.8	21.2	38.5	0.78
IS16	-1.6	0	0.8	1.258	0.087	1.3	20.9	29.4	0.74
IS17	0	0.6	1.4	-0.559	0.538	2.5	13.1	51.7	0.72
IS23	-2.0	-0.6	0.8	-0.567	0.264	1.0	12.7	35.4	0.66
IS24	-1.4	0.4	0.6	2.023	0.382	0.8	11.3	34.0	0.60
IS26	-1.4	0.2	1.0	0.770	0.039	0.6	10.4	31.5	0.64
IS27	-1.6	0.0	1.0	0.698	0.743	0.6	9.8	30.5	0.59
IS29	-0.4	0.0	1.6	-2.180	0.010	1.1	7.3	47.1	0.71