Paper presented to the Health Economists' Study Group Meeting,
University of Exeter, 9th-11th January 2013

Developing a classification system for an MS-specific preference-based measure

Liz Goodwin, PhD student, Health Economics Group, University of Exeter Medical School

Abstract

Aim/ Background: Research has shown that generic preference-based measures of health fail to capture the full impact of the disease and its treatment when applied to multiple sclerosis (MS). A condition-specific preference based measure (CSPBM) may provide a more relevant, appropriate, sensitive and responsive alternative. This paper describes the first stage in deriving a CSPBM from an existing MS-specific health-related quality of life (HRQoL) measure: reducing the size of the measure to render it suitable for valuation, while minimising the loss of descriptive information.

Methods: Available MS-specific HRQoL instruments were identified using published reviews of HRQoL measurement in MS. The instruments were assessed and compared against criteria developed from the relevant literature. The MS Impact Scale (MSIS-29) was selected to form the basis of a classification system for an MS-specific preference-based measure.

The two-dimensional structure of the MSIS-29 was confirmed using factor analysis. Each of its two subscales, however, included items that represented more than one conceptually distinct dimension of HRQoL. Therefore the relevant literature was analysed to produce a framework of HRQoL in MS, based around a number of conceptual dimensions. The items of the MSIS-29 were then mapped to this framework.

Baseline data (n=529) from a prospective, longitudinal, cohort study of people with MS was subjected to Rasch analysis using the RUMM2030 software and psychometric analysis using SPSS. Rasch analysis was undertaken to identify, then adjust or remove items that exhibited disordered thresholds, differential item functioning or misfit to the Rasch model. The remaining items were assessed against Rasch and psychometric criteria, and the best performing item was selected to represent each conceptual dimension. Further Rasch analysis determined whether respondents were able to distinguish between item levels. These steps were repeated on a validation sample (n=528).

Results: Overall model goodness of fit was achieved for both subscales of the MSIS-29, and the selected items performed well against both Rasch and psychometric criteria. Eight conceptual dimensions of HRQoL in MS were represented in the final MSIS-8D classification system: general physical function, mobility, employment, social function, fatigue, cognition, depression and general emotional wellbeing. Each dimension was represented by a single item. There was no evidence to suggest that item levels should be merged, therefore the four-level structure of the MSIS-29 was retained.

Conclusion: The MSIS-29 can be reduced using Rasch analysis and psychometric criteria to produce a suitable classification system for a CSPBM. The next stages of this research will involve direct valuation of a sample of health states from the MSIS-8D, and statistical modelling to estimate preference weights for all health states.

Introduction

In the economic evaluation of healthcare interventions it is common to employ cost utility analysis, where the QALY is used to compare the relative merits of options for treatment in terms of both length and quality of life. Generic preference-based measures (PBMs) of health are now widely used to estimate the quality of life weights that are required for the calculation of QALYs (Brazier et al, 2012). Indeed, this approach is recommended by the National Institute of Health and Clinical Excellence (NICE), who stipulate that the generic EQ-5D should be used whenever possible (NICE, 2008). PBMs are based around specific dimensions of HRQoL, each with a number of ordinal response levels, enabling respondents to describe their current health state in terms of their level on each dimension. These dimensions and levels provide a health classification system describing a finite number of possible health states, to each of which is assigned a specific quality of life weight. These weights are typically introduced using health state values derived from preference data, elicited from samples of the population (Brazier and Tsuchiya, 2010). Generic PBMs are intended to be suitable for all health conditions, by focusing on the main determinants of health-related quality of life (HRQoL). This broad focus, however, can reduce the sensitivity of generic instruments to differences and changes in aspects of HRQoL that are of particular relevance to some specific health conditions, causing them to underestimate the impact of these diseases and their treatments. In addition, generic PBMs are not always included as outcome measures in clinical trials (Brazier et al, 2012).

Multiple sclerosis (MS) is an inflammatory autoimmune condition, and is the most common cause of neurological disability among young adults in the UK (Bose et al, 2002). Its effects vary widely between individuals, and within the same individual over time. Common symptoms include fatigue, loss of mobility, sensory problems, loss of bowel and bladder function, cognitive impairment and pain (Hemmett et al, 2004). Onset normally occurs between the ages of 20 and 40, the disease course is generally progressive, the impact on life expectancy is only around five to 10 years, and as yet there is no cure; hence people with MS generally experience several decades of increasing neurological disability (Zajicek et al, 2007). Research has shown that people with MS perform significantly worse than the general population on all aspects of health-related quality of life (HRQoL) (Nortvedt and Riise, 2003). The HRQoL decrements associated with mild MS are estimated to be commensurate with those of congestive heart failure or chronic obstructive pulmonary disease (Miller et al, 2010).

Research has shown that generic measures may lack both relevance and sensitivity to changes in HRQoL in MS. The EQ-5D does not directly capture some important factors that influence the HRQoL of people with MS, notably fatigue (Hemmett et al, 2004), and may not

distinguish adequately between mildly and moderately disabled MS patients (Fisk et al, 2005). Both the EQ-5D (Opara et al, 2010) and SF-6D (Fisk et al, 2005) have been found to lack responsiveness, particularly in patients with advanced MS. Most discussion of the relevance and sensitivity of generic measures to MS has centred on the SF-36, upon which the SF-6D is based. Although the SF-36 has been validated for use in MS patients (Nortvedt and Riise, 2003), it has been shown to have limitations. These include significant floor and ceiling effects, particularly in its physical subscales (Gruenewald, 2004), and its mental health summary scores tend to underestimate the negative impacts of MS on emotional and mental wellbeing (Benito-Leon, 2005).

An alternative to generic PBMs is to use a condition-specific preference based measure (CSPBM). These focus on the aspects of health that are most relevant to the condition of interest, potentially rendering them more sensitive to differences and changes in HRQoL (Brazier and Tsuchiya, 2010). Given the limitations of generic measures, an MS-specific CSPBM may provide more accurate estimation of the impacts of the disease and its treatment in a form that is suitable for economic evaluation. Our research aims to develop such a measure. This paper describes the first stage in this process: developing a standardised classification system, based around the dimensions of HRQoL that are most relevant to MS.

Methods

There are two approaches to developing a classification system for a PBM: developing a new classification system from scratch, or deriving one from an existing non-preference based instrument. The latter offers certain advantages. Adapting a well-accepted and frequently used measure enables retrospective economic evaluations to be undertaken using existing trial data, and means that the measure is more likely to be used in future trials, increasing the scope for the new instrument to be used for economic evaluations (Brazier et al, 2007). If the original instrument is accepted by the clinical community, this will add credibility to the results of economic evaluations that use its preference-based off-spring. Therefore we decided to derive a PBM from an existing non-preference based MS-specific measure.

The main stages in developing a PBM have been identified by Brazier et al (2012) as:

1. Establish dimensions
2. Eliminate and select items per dimension
3. Explore item-level reduction
4. Validation
5. Valuation exercise to elicit health-state values for a sample of states
6. Model valuation results to produce utility values for all health states

This paper documents steps 1 to 4. Prior to embarking on this process, however, one additional step is required: the selection of an appropriate instrument to use as the basis for the classification system. NICE guidance recommends using a patient-reported measure of HRQoL: measures describing only symptoms or symptom bother are not deemed suitable (Brazier and Rowen, 2011). We systematically searched Medline and Embase for published reviews of HRQoL measures in MS, and identified 13 reviews, which described 17 MS-specific instruments. We developed a set of criteria (listed in Appendix 1), drawing upon a variety of sources (Brazier et al, 2007; Hobart et al, 2004; Murrell, 1999**;** O'Connor, 2004; Riazi, 2006; Streiner and Norman, 2003), to provide a systematic means of comparing instruments. Application of these criteria produced three candidate instruments: the MS International Quality of Life questionnaire (MusiQoL), MS Impact Scale (MSIS-29) and Functional Assessment of MS (FAMS). Given its limited use in clinical trials to date, which has restricted the availability of evidence to support its responsiveness and acceptability (Bandari et al, 2010), and our lack of access to a dataset containing the measure, we decided to progress no further with the MusiQoL, although we do suggest that further consideration of this instrument could be a productive area for further research. The MSIS-29 and FAMS performed similarly well across all criteria and data is available for both instruments. An exploratory analysis of the FAMS, however, identified a number of problems, which are outlined in Appendix 1. Therefore, we selected the MSIS-29 to form the basis of the classification system for our MS-specific PBM.

The MSIS-29 consists of a physical subscale of 20 items and a psychological subscale of 9 items, and produces an individual score for each subscale, rather than a combined score. Respondents are requested to report the impact of MS on their day-to-day lives. The first three items are preceded by the wording: 'In the *past 2 weeks*, how much has your MS limited your ability to …', and the remaining 26 items by: 'In the *past 2 weeks*, how much have you been bothered by …'. The original version of the instrument had five response levels per item; on the basis of subsequent analysis, these were reduced to four levels per item in Version 2 of the questionnaire: 'not at all', 'a little', 'moderately' and 'extremely'. Items were generated from qualitative analysis of interviews with people with MS, and the subscales were constructed using recognised psychometric scale development techniques (Hobart and Cano, 2009). A number of validation studies have confirmed the acceptability, reliability, validity and responsiveness of the instrument for a range of MS types and clinical settings (Costelloe et al, 2007; Gray et al, 2009; Hoogervorst et al, 2004; McGuigan and Hutchinson, 2004; Ramp, 2009; Riazi et al, 2002). The MSIS-29 is well accepted by clinicians and researchers and has frequently been used in research and clinical trials (Giordano et al, 2009).

Step 1: Establish dimensions

Rather than accepting the published dimensions of the original measure at face value, statistical analysis should be undertaken to determine its dimensional structure. It is important, however, to apply clinical judgement and common sense to ensure that these dimensions make (bio)logical sense (Brazier et al, 2012). Therefore, we undertook an exploratory factor analysis of the MSIS-29, then reviewed the contents of the statistically confirmed factors. We found that each factor included items that represented more than one conceptually distinct dimension of HRQoL; for example, the physical subscale includes items that describe impacts on social activities, as well as on physical functioning. Consequently, selecting items from each factor purely on the basis of statistical criteria could result in important aspects of HRQoL in MS being omitted.

A similar issue has been faced by other CSPBM developers. In deriving a mental health PBM from the CORE-OM, Mavranezouli et al (2011) found that all the emotional aspects of HRQoL were contained within a statistically unidimensional factor. They were able to use the existing published dimensions of the CORE-OM to guide the selection of items from this factor, to ensure that a range of aspects of emotional HRQoL were covered. In our case, however, both published subscales of the MSIS-29 contain conceptually distinct constructs. This left us in need of an external source for the key dimensions of HRQoL in MS. To address this, we analysed the dimensional structures of the other identified MS HRQoL instruments, the papers concerning HRQoL measurement in MS from which those instruments were sourced, and wider literature describing HRQoL both in general and specifically in relation to MS. From this, we constructed a conceptual framework of HRQoL in MS. We then fitted the items of the MSIS-29 to the dimensions of the conceptual framework, enabling items to be selected to cover aspects of HRQoL that are important to people with MS.

Step 2: Eliminate and select items per dimension

When developing a classification system, it is important to consider the cognitive demands on respondents to the valuation exercise (the fifth stage in developing a PBM, as listed on p2). Psychological research has shown that, on average, respondents can deal with five to nine items (Feeny, 2006); typically, however, non-preference based HRQoL instruments far exceed this. The key to deriving a classification system from an existing measure is to reduce the size of the instrument while minimising the loss of descriptive information, and this can be achieved by selecting the most appropriate single item to represent each dimension of HRQoL. This can be achieved through statistical analysis of a dataset that contains the instrument being converted, using a sample that is representative of the target population for the CSPBM (Young et al, 2009).

One recommended approach is to eliminate poorly performing items using Rasch analysis, before selecting those items that perform best against Rasch and traditional psychometric criteria (Brazier et al, 2012). Rasch analysis provides a technique by which ordinal data, such as that generated by HRQoL instruments, can be converted to continuous data. Any unidimensional measure captures 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 on the latent trait. Similarly, item response levels will be located along the same scale according to the level of HRQoL that they represent. Rasch analysis assumes that the probability of a respondent endorsing a particular item response is a logistic function of the relative distance between that individual's position on the latent scale, and the position of the item response on the latent scale (Tennant and Conaghan, 2007). Because Rasch models deal with unidimensional scales, separate models are required for each statistically confirmed dimension of each instrument. A number of tests are available to assess compliance with the key assumptions of Rasch analysis; these explore how well individual items represent the underlying construct, and their ability to distinguish between respondents with differing levels of the underlying construct (Pallant and Tennant, 2007). These tests are therefore ideal for selecting items for a classification system.

Item elimination

Three types of test were employed to identify, then adjust or remove poorly performing items using Rasch analysis: disordered thresholds, differential item functioning, and model and item goodness of fit.

For each individual item, each response level should take it in turns to show the highest probability of endorsement, in order of response level severity, as we progress along the latent scale. This will produce an ordered set of response thresholds for each item, where the threshold between adjacent item response levels is defined as the point at which either response is equally probable. If respondents are unable to discriminate between response levels, they may endorse response levels inconsistently with the level of the latent trait, and this will result in disordered thresholds (Pallant and Tennant, 2007). In order to identify items with disordered thresholds, the item-threshold map for each Rasch model was examined. We then investigated the category probability curves for each item that exhibited disordered thresholds in order to identify options for merging item levels (Ramp et al, 2009). Any items requiring adjustment were retained in the Rasch models, but excluded from consideration as part of the classification system, because they discriminate less effectively across the whole range of condition severity (Young et al, 2009).

The Rasch assumption of unidimensionality requires items to function in the same way for all respondents (Martin, 2007). When responses to an item differ between groups of respondents that exhibit equal levels of the latent characteristic (HRQoL), this is known as differential item functioning, or DIF (Tennant and Conaghan, 2007). Uniform DIF describes a situation where the difference in responses between groups is consistent across the latent scale. With non-uniform DIF, the difference between groups varies along the latent scale. Uniform DIF can be adjusted by splitting the item by the relevant grouping (or 'person factor'), creating two separate items. No such adjustment is available for non-uniform DIF, which requires the affected item to be removed from the analysis (Pallant and Tennant, 2007). The person factors explored for DIF were sex, age group, duration of disease and type of MS. We examined item characteristic curves and DIF summary tables to identify items affected by uniform DIF; we then split the affected items accordingly, one by one, and refitted the model after each adjustment. A final check was then undertaken for any further DIF, including non-uniform DIF. Items exhibiting DIF were excluded from the classification system, although the adjusted versions were retained in the Rasch models, in order to maintain measurement precision and to retain 'the item content contributing to the latent construct' (Hagquist et al, 2009, p385).

Inclusion of respondents whose responses do not fit the expectations of the Rasch model can cause apparent item misfit, therefore these individuals should be removed from the analysis prior to identifying items that misfit the Rasch model (Tennant and Conaghan, 2007). Individual person fit residuals are calculated as the sum of a respondent's deviations from expected values across all items. We removed all respondents with a fit residual > |2.5| from each Rasch analysis, then refitted the models (Pallant and Tennant, 2007).

Lastly, we applied three tests to examine how well the observed data fit the expectations of the Rasch model: item-trait interaction, item and person fit, and the internal consistency of the scale (Ramp et al, 2009). Item-trait interaction is measured by estimating the difference between observed and expected responses across items: if the data fit the Rasch model, the observed and expected responses should be similar, giving a non-significant model χ^2 statistic ($p > 0.01$) (Tennant and Conaghan, 2007). Item and person fit are also assessed according to the difference between observed and expected responses: if the items and respondents fit the Rasch model, the mean item and person fit residuals will be close to zero, with standard deviations close to one (Pallant and Tennant, 2007). The internal consistency of the scale is estimated via the Person Separation Index (PSI), which employs the same formulae as Cronbach's α (Tennant and Conaghan, 2007), and is interpreted in the same way with an acceptable threshold at 0.70 (Ramp et al, 2009).

If these tests indicate poor fit to the Rasch model, overall fit may be improved by identifying and removing individual items that do not fit the model (Hagquist et al, 2009); that is, items with fit residuals > |2.5| and significant χ^2 values ($p < 0.05$). As this process involves multiple tests, we adjusted the threshold χ^2 p values using Bonferroni corrections (Pallant and Tennant, 2007). We removed poorly fitting items one by one, starting with the worst, and refitted the model after removing each item. The process was repeated until only well-fitting items remained, and overall model goodness of fit was achieved. Items that were removed from the Rasch models were not considered for inclusion in the classification system (Brazier et al, 2012).

Item selection

Having eliminated the poorly performing items, we applied further Rasch analysis and psychometric criteria to select the most appropriate item from those that remained 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 range is represented by the latent logit scale. Therefore, items that span a wide range of condition severity will have 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, ie fit residuals close to zero and nonsignificant χ^2, were also taken into account. In addition to the Rasch analysis, we also assessed items against four psychometric criteria: feasibility (percentage of item missing data); internal consistency (Cronbach's α); distribution of responses (floor and ceiling effects); and responsiveness to change over two points in time (standardised response mean). Preference was given to items that spanned the full range of severity and exhibited good fit to the Rasch model (Young et al, 2011).

Step 3: Explore item-level reduction

Research shows that respondents to HRQoL measures can struggle to distinguish between item response levels. Where this is the case, these levels can be merged without losing important descriptive information, and this offers a potential further means of simplifying the classification system (Young et al, 2009). Using Rasch analysis, we examined the threshold probability curves for the items that had been selected for the classification system: curves that cross, or that come close to crossing, may represent levels that could be merged (Brazier et al, 2007).

Step 4: Validation

In order to identify any local dependency between items, we examined the residual correlation matrix using Rasch analysis (Mulhern et al, 2012). We then repeated steps 2 to 4 of the process using a separate validation sample (Brazier et al, 2012).

Dataset

The South West Impact of Multiple Sclerosis (SWIMS) project is a longitudinal cohort study of people aged 18 or over, with a clinical diagnosis of MS or clinically isolated syndrome, living in Devon and Cornwall. Participants complete questionnaire packs, which include a range of generic and condition-specific measures (including the MSIS-29 and FAMS) and collect other clinical and demographic data. The demographic make-up of respondents is consistent with other published UK data and clinical experience; see Zajicek et al (2010) for more detail on this study. In May 2012, we obtained an extract of SWIMS baseline data for 1057 respondents with MS. Initial data cleaning was undertaken in Microsoft Excel, including conversion of age and duration of disease into ordinal variables and reclassification of MS type codes into a smaller number of categories to make the data amenable to Rasch analysis. The total sample was randomly split into a development dataset and a validation dataset; once missing data was taken into account, this provided a suitable sample size for Rasch analysis of around 400-500 respondents. Table 1 reports the descriptive statistics for each dataset. Rasch analysis was undertaken using RUMM2030 software, and psychometric analysis using SPSS.

Table 1: 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%
MS type progressive	20%	24%
MS type RRMS	27%	23%
MS type benign or mild	2%	3%
MS type not recorded	51%	50%

Results

Step 1: Establish dimensions

We undertook an exploratory principle components factor analysis using the full SWIMS dataset. Three factors had Eigenvalues > 1, and a three-factor solution explained 61% of the total variance. The scree plot (Figure 1), however, supported a two factor solution, which explained 57% of total variance and confirmed the original structure of the MSIS-29, with items 1-20 forming the physical subscale and items 21 to 29 forming the psychological subscale. Therefore, we accepted the two factor structure.

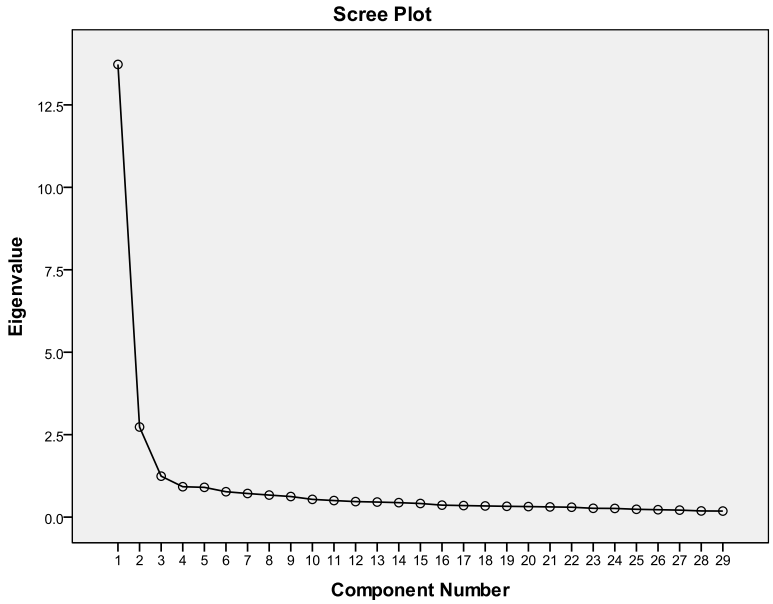

Figure 1: MSIS-29 scree plot

Analysis of the content of other MS-specific HRQoL measures and the HRQoL literature enabled construction of a conceptual framework which included physical, psychological and social impacts of MS on people's HRQoL. The dimensions of the conceptual framework are illustrated in Figure 2.

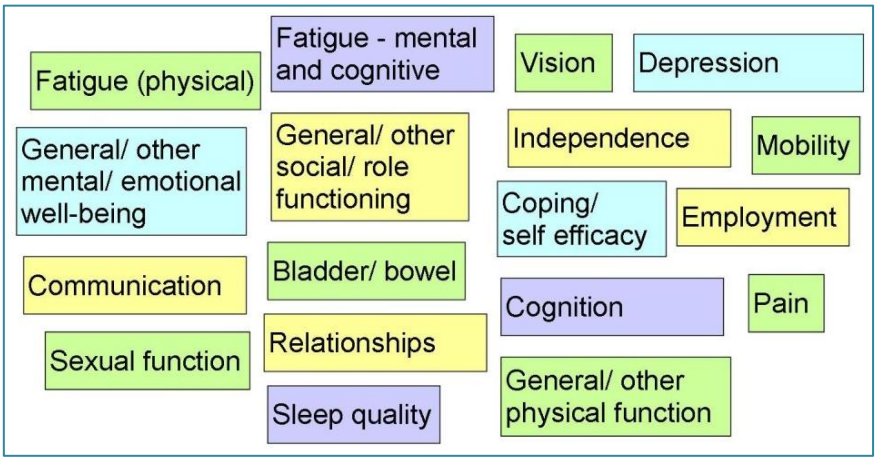

Figure 2: Conceptual framework of HRQoL in MS

We then allocated the items of the MSIS-29 to dimensions of the conceptual framework. 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 mental and emotional wellbeing and the impact of other non-physical symptoms of MS (see Table 2). Not all domains of the conceptual framework were covered; however no measure can realistically include all possible dimensions of HRQoL (Brazier et al, 2007). Three items (IS18: Taking longer to do things; IS19: Difficulty doing things spontaneously; IS21: Feeling unwell) did not fit the conceptual framework. To test the impact of these items, we undertook two versions of the analysis, one including and one excluding them. These items will not be candidates for selection as part of the final classification system, however, because they do not represent a predefined aspect of HRQoL in MS.

Step 2: Eliminate and select items per dimension

We fitted two partial credit polytomous Rasch models, one for the physical and one for the psychological subscale. Table 2 summarises the results of the analysis. Detailed results are available from the author on request.

Item elimination

The item threshold maps exhibited ordered thresholds for all items. Therefore no items were eliminated as a result of item level ordering. Five items from the physical subscale and one item from the psychological subscale required adjustment for uniform DIF. Thirty-five and 22 respondents misfit the Rasch model for the physical and psychological subscales respectively; we removed these respondents from the analyses. 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).

Table 2: MSIS-29 analysis results (development dataset)

MSIS-29 dimension	Conceptual dimension	Code	Item description	Analysis 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?	× Not selected
		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.)?	× Not selected
	Bladder/ bowel	IS20	Needing to go to the toilet urgently?	× Misfit
	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 (e.g. going out on the spur of the moment)?	× Unallocated
Psychological	General/ other mental and emotional wellbeing	IS24	Worries related to your MS?	× Not selected
		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

Overall goodness of fit to Rasch models following item elimination:		**Item fit residual**		**Person fit residual**		**p-value**	**PSI**
		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 selection

The aims of the item selection phase were to confirm whether the items that remained as the sole representative of a dimension were suitable for inclusion, and to decide which items should be selected to represent the General/ other physical functioning, Mobility, and General/ other mental wellbeing dimensions.

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 the highest proportion of missing data of any item.

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 3: Explore item-level reduction

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

Step 4: Validation

The item elimination, item selection and item level reduction processes were repeated using the validation dataset (detailed results available on request). This confirmed the selection of items IS01, IS13, IS14, IS16, IS23, IS26, IS27 and IS29, and the retention of the four item response levels. The only difference between the results of the two datasets was item IS12, representing the Independence dimension, which was included following analysis of the validation dataset, but was eliminated during analysis of the development dataset, where it was affected by DIF according to the duration of disease: 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. In their research into the HRQoL of people with severe MS, Gruenewald et al (2004, p699) found that ‘inappropriate personal assistance’ can have a negative impact. It may be reasonable to hypothesise that people who have experienced MS over many years have had more time to negotiate informal care arrangements that are acceptable to both themselves and their carers, or have adjusted their expectations of personal control and independence. As a plausible explanation for its differential functioning in the development dataset can be provided, we decided to exclude IS12 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 from the outset. This made no difference to the results.

Finally, 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 (Limitations in your social and leisure activities at home) and IS14 (Being stuck at home more than you would like to be). These items represent different dimensions of HRQoL in MS (social function and mobility respectively) and were not correlated in the development dataset, so we did not consider this a major cause for concern.

Based on the results of the analysis of both the development and the validation datasets, the classification system derived from the MSIS-29 is 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 3) describes 65,536 unique health states.

Discussion

This paper describes the first stage in developing a CSPBM: developing a classification system that provides a standardised approach to describing health states experienced by people with the condition. The derivation of classification systems from non-PBMs involves reducing their content, inevitably resulting in a loss of information, which risks over-riding the increased sensitivity of CSPBMs relative to their generic equivalents (Brazier et al, 2007). The challenge of producing a classification system that is both small enough for valuation and comprehensive enough to generate adequate descriptions of health states highlights the importance of adopting robust methods.

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

Figure 3: The MSIS-8D classification system

This study builds on previous research using Rasch analysis and traditional psychometric techniques (Brazier et al, 2012), and these methods have succeeded in producing a classification system from the MSIS-29, which covers important dimensions of HRQoL in MS identified from the relevant literature, with a good balance between its physical, psychological and social aspects.

One issue that has rarely been discussed in the literature describing the derivation of CSPBMs is the selection of the instrument upon which the classification system is to be based. This research has developed a systematic process for comparing available HRQoL measures. Applying a standardised set of criteria enabled the selection of three potential instruments from an initial list of 17 alternative measures. It is important to note, however, that good performance against these criteria does not guarantee suitability, as our experience with the FAMS has shown. Nonetheless, this approach may be of use for future studies where a number of alternative HRQoL instruments are available.

The importance of considering face validity when assessing the dimensional structure of instruments has been highlighted by this research. The statistically confirmed dimensions of the MSIS-29 contained items that represented different dimensions of HRQoL. We developed a novel approach to deal with this scenario: analysing the relevant literature to build a conceptual framework of HRQoL in MS, to which the items of the instrument can be mapped, ensuring that the main conceptually independent dimensions of HRQoL are 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 (Mavranezouli et al, 2011; Young et al, 2010). Although three of the conceptual dimensions originally covered by the MSIS-29 were not included in the final classification system, a reduction in the number of dimensions was necessary to bring the size of the classification system in line with the recommended five to nine items and all the dimensions identified in the literature as major factors influencing the HRQoL of people with MS were present. Mavranezouli et al had a similar experience in deriving their mental health PBM.

There are a number of potential drawbacks to using CSPBMs. Some commentators argue that, in order to compare between patient groups, their health 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 (2012) suggest that, provided the same valuation technique, anchors and type of respondents to the valuation survey are used, comparability can be achieved between different classification systems. Notwithstanding this, some problems with comparability remain, and these arise largely due to the limited coverage of CSPBMs relative to generic measures. CSPBMs are not capable of capturing side effects of interventions that fall outside of the dimensions covered by the classification system, and may not pick up impacts on co-morbidities. They may also be prone to focusing effects, where the impact of the condition may be 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. Preference interactions may occur between dimensions that are included in the classification system, and other aspects of health that are excluded from the classification system but are nonetheless taken into account by respondents to the valuation survey, potentially influencing their preferences between health states and affecting the survey results. Another problem that may arise due to dimensions of health external to the classification system concerns 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 (Brazier et al 2012).

The disadvantages relating to co-morbidities, focusing effects and preference interactions are arguably less important when the condition of interest is the dominant factor in determining HRQoL (Brazier et al, 2007), as is likely to be the case for people with MS. The ability to capture side effects is particularly restricted when the coverage of the classification system is narrow (Brazier et al, 2012), and in this regard the varied impacts of MS on HRQoL may constitute an advantage, as this has resulted in the MSIS-8D classification

system becoming somewhat broader than many other CSPBMs. Therefore, many of the criticisms levelled at CSPBMs, while remaining relevant, may be less problematic for MS than for other conditions. A decision about whether to develop or use a CSPBM will always rest on a trade-off between the advantages and disadvantages of CSPBMs in relation to the condition of interest (Brazier et al, 2012): in the case of MS, the acknowledged limitations of generic measures and lower likelihood of problems arising due to the restricted coverage of the classification system tip the balance in favour of developing a CSPBM.

The next stage of the research will involve direct valuation of a sample of health states from the MSIS-8D, and regression modelling to estimate preference weights for all health states. In order to enhance comparability with the frequently-used EQ-5D, the same valuation technique (time trade-off, using the Measurement and Valuation of Health variant) and respondents (a representative sample of the UK population) will be used to obtain health state utility values for the MSIS-8D. Obtaining a value set that will be comparable with the upcoming EQ-5D-5L could be an area for further research outside the current study. The potential disparity between instrument-specific full health and perfect health will be countered by including a generic upper anchor in the valuation exercise (Brazier et al, 2012).

Classification systems that are comprised of one or more unidimensional scales, such as the MSIS-8D, are likely to be affected by preference interactions between the conceptual dimensions that form each statistically unidimensional scale. This poses a challenge when it comes to selecting a sample of health states for the valuation survey, as traditional methods such as orthogonal arrays run the risk of generating implausible health states. To address this, Mavranezouli et al (2011) developed the Rasch vignette approach, in which those health states that would typically occur at each step along the latent scale are selected for valuation. Because this approach is based on the natural occurrence of health states, it reduces the likelihood of generating implausible states. For the first time, this research will apply the Rasch vignette approach to generating health states from an instrument with two statistical factors, both of which include more than one conceptual dimension.

A further consideration for the valuation survey is whether to specify that the health states being valued are caused by MS. Labelling the condition in this way has been shown to affect respondents' preferences for health states, as they bring their preconceptions about the condition to bear on their valuations. Recent guidance recommends avoiding condition labelling (Brazier et al, 2012); however, this means that the wording of one item (IS01) will differ from the original MSIS-29, so that what respondents are valuing may differ from what patients are reporting (Brazier and Tsuchiya, 2010). The pilot of the valuation survey could be used to explore the effect of condition labelling.

Conclusion

We have demonstrated that a classification system can be derived from the MSIS-29 that covers important dimensions of HRQoL in MS and is suitable for valuation, forming the first stage in developing a CSPBM for MS. Once a valuation survey has been conducted, and health state utility values modelled for all health states described by the classification system, this research will produce a CSPBM that can be used for the economic evaluation of treatments for MS.

Acknowledgement: this research is funded by the Multiple Sclerosis Society.

References

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

Benito-Leon 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(23):1291

Bose U, Ladkani D, Burrell A and Sharief M, 2002: Cost-effectiveness analysis of glatiramer acetate in the treatment of relapsing-remitting multiple sclerosis; J Drug Assessment 5: 67

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

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

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

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)

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

Feeny D, 2006: The multi-attribute utility approach to assessing health-related quality of life; in Jones AM (ed), 2006: The Elgar Companion to Health Economics; Cheltenham: Edward Elgar

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

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

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

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

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

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

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)

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)

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

Martin M, Kosinski M, Bjorner JB, Ware JE, MacLean R and Li T, 2007: Item response theory methods can improve the measurement of physical function by combining the modified health assessment questionnaire and the SF-36 physical function scale; Qual Life Res 16:647

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

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

Miller D, Rudick RA and Hutchinson M, 2010: Patient-centered outcomes: translating clinical efficacy into benefits on health-related quality of life; Neurology 74 (Suppl 3):S24

Mulhern B, Smith SC, Rowen D, Brazier JE, Knapp M, Lamping DL, Loftus V, Young TA, Howard RJ and Banerjee S, 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

Murrell R, 1999: Quality of Life and Neurological Illness: A Review of the Literature; Neuropsychol Rev 9(4):209

NICE, 2008: Guide to the methods of technology appraisal; National Institute for Health and Clinical Excellence, http://www.nice.org.uk/media/B52/A7/TAMethodsGuideUpdatedJune2008.pdf

Nordvedt MW and Riise T, 2003: The use of quality of life measures in multiple sclerosis research; Mult Scler 9:63

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

Opara JA, Jaracz K and Brola W, 2010: Quality of life in multiple sclerosis; J Med Life 3(4):352

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):1

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

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

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

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

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 Rheum 57(8):1358

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

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

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

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

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

Appendix 1: Selecting an appropriate measure of HRQoL

Table A1: Criteria for selection of a HRQoL 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 $\alpha > 0.80$) 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 SE 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

In order to identify available instruments that could be used as the basis for a classification system, we searched Medline and Embase for published reviews of HRQoL measures in MS. The search produced 13 reviews, which described 17 MS-specific instruments. A systematic method was then required by which these instruments could be compared. We developed a standardised set of criteria from the literature on measuring HRQoL, both in general and specifically in relation to MS. In order to accelerate the process, the instruments were initially assessed against five screening criteria (marked with an asterisk in Table A1). The screening criteria were selected because they could be accessed swiftly from the available literature and did not require detailed comparisons between measures. Those measures that passed all screening criteria were then assessed against the remaining criteria, with a view to selecting the best performing measure.

Fourteen measures were rejected on the basis of the screening criteria. Of these, three (NeuroQoL, MSQLI and QOLQ for MS) were not single instruments. Ten (NeuroQoL, MSQLI, RAYS, DIP, HRQOL-MS, MS ADL, MSQoL-54, HAQUAMS, FILMS, QLI-MS) did not involve patients in the generation of items, five did not use recognised scale development techniques (MSQLI, QOLQ for MS, RAYS, HAQUAMS, MSSID), and the techniques used to develop a further two (MSQoL-54, QLI-MS) were not clear. The aspects of HRQoL covered by four of the instruments (MS-ADL, QLI-MS, PRIMUS, LMSQOL) did not adequately reflect our conceptual framework of HRQoL in MS, and two measures (PRIMUS, LMSQOL) aimed to measure overall, rather than health-related, quality of life.

Three measures, the MSIS-29, FAMS and MusiQoL, passed the screening criteria. Next we considered two key practical considerations. Due to its relatively recent development, the use of the MusiQoL in clinical trials is limited, restricting the amount of evidence available to support its psychometric performance. In addition, we had no access to a dataset containing responses to the MusiQoL. Conversely, both the MSIS-29 and the FAMS are well accepted and frequently used (Giordano et al, 2009), and both are included in the SWIMS dataset. Consequently, we rejected the MusiQoL at this stage, and only assessed the MSIS-29 and the FAMS against the remaining criteria. Both instruments performed equally well, therefore we attempted to derive a classification system from each of these two measures.

Potential problems with the FAMS became apparent at the factor analysis stage. The published six-dimensional structure of the FAMS was not supported by the analysis, which instead suggested three alternative versions, with one, three or eight dimensions. Neither the three nor the eight factor version were compatible with the conceptual framework, for example items relating to social and role functioning were spread across more than one factor. A separate Rasch analysis was conducted for each 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.