Combining Rasch and cluster analysis: a novel method for developing rheumatoid arthritis states for use in valuation studies

McTaggart-Cowan, H; Brazier, J and Tsuchiya, A
The University of Sheffield

2008

HEDS Discussion Paper 08/15


ScHARR
School of Health and
Related Research

The University Of Sheffield.

ScHARR

Health Economics and Decision Science
Discussion Paper Series

No. 08/15

COMBINING RASCH AND CLUSTER ANALYSIS: A NOVEL METHOD FOR DEVELOPING RHEUMATOID ARTHRITIS STATES FOR USE IN VALUATION STUDIES

Helen McTaggart-Cowan[a,*], John Brazier[a], and Aki Tsuchiya[a,b]

[a]School of Health and Related Research, University of Sheffield, Sheffield, UK.
[b]Department of Economics, University of Sheffield, Sheffield, UK

[*]School of Health and Related Research
University of Sheffield, Regent Court
30 Regent Street
Sheffield, UK
S1 4DA
h.m.cowan@sheffield.ac.uk
Telephone: +44 (0)114 222 0722
Fax: +44 (0)114 272 4095

ABSTRACT

Purpose: Health states that describe an investigated condition are a crucial component of valuation studies. The health states need to be distinct, comprehensible, and data-driven. The objective of this study was to describe a novel application of Rasch and cluster analyses in the development of three rheumatoid arthritis health states.

Methods: The Stanford Health Assessment Questionnaire (HAQ) was subjected to Rasch analysis to select the items that best represent disability. *K*-means cluster analysis produced health states with the levels of the selected items. The pain and discomfort domain from the EuroQol-5D was incorporated at the final stage.

Results: The results demonstrate a methodology for reducing a dataset containing individual disease-specific scores to generate health states. The four selected HAQ items were bending down, climbing steps, lifting a cup to your mouth, and standing up from a chair.

Conclusions: Overall, the combined use of Rasch and cluster analysis has proved to be an effective technique for identifying the most important items and levels for the construction of health states.

Key words: health state, Rasch analysis, cluster analysis, quality of life, rheumatoid arthritis

1. Introduction

Rheumatoid arthritis (RA) afflicts 0.8% of the United Kingdom population (Symmons, 2005). It is a chronic autoimmune inflammatory disorder, which results in upper- and lower-limb disability and discomfort. Although treatments are improving, RA is incurable and can significantly reduce a patient's quality of life (QOL). Physicians often use disease-specific instruments to assess the QOL of their patients; for RA, the Stanford Health Assessment Questionnaire (HAQ) (Fries et al., 1980) is widely used (Bruce and Fries, 2003). The HAQ has been shown to be valid (Marra et al., 2005a) and responsive (Marra et al., 2005b). While scores from the HAQ are meaningful to members of the rheumatology community, the use of disease-specific instruments has limited value in guiding decisions for resource allocation because comparisons across different diseases cannot be drawn.

To ensure that healthcare resources are being utilized efficiently, it is recommended that universal QOL values are used, rather than units specific to the investigated condition (NICE, 2003). While the use of QOL values enables comparisons across diseases, it is important that health states adequately describe the investigated health condition. The descriptions need to portray the symptoms a patient experiences in a manner that is understandable for non-patient respondents when they appraise the states.

Health states can be derived using various approaches, including expert judgements or patient responses. Expert judgements, such as those from physicians and nurses, permit a broad range of patient experiences to be elicited, especially in situations where patients are unable to report their QOL (e.g. severe stroke). However, the experts' opinions are subjected to biases. For example, a physician's viewpoint of a health state may be distorted if his/her patients exaggerate their QOL in an attempt to please their doctor or family members. Patient responses, on the other hand, provide direct information about how the investigated health state impacts their lives. The use of qualitative approaches allow researchers to gain in-depth knowledge from a small number of patients but those who opt to participate in interviews and focus groups might not be representative of most

patients. Responses on disease-specific questionnaires from a large sample of patients may be a better alternative in the construction of health states.

Disease-specific instruments are comprised of items addressing different aspects of QOL. For item responses to be of use in health state descriptions, the number of items to be assessed needs to be minimized to reduce respondent burden. Factor analysis is one statistical technique frequently used to explore the correlation of items in disease-specific instruments (Fayers and Machin, 2000). However, the drawback of this approach is that many disease-specific instruments – including the HAQ – are not necessarily comprised of multi-dimensional items.

One method that can be used to identify the most representative items of unidimensional instruments is Rasch analysis. It is a mathematical technique that converts categorical responses into a continuous latent scale using a logistic model (Rasch, 1960; Tesio, 2003). This method has been employed in the development of QOL instruments (Tennent et al., 2004) and, more recently, in the selection of items for health state classification systems (Young et al., 2007; Young et al., 2008). As Rasch analysis identifies the instrument's most meaningful items, the cognitive and time constraints placed on a respondent are minimized. Furthermore, items can be ranked from easiest to the most difficult; this ensures that health states capture the widest range of severity.

While Rasch analysis reduces the number of items in a large instrument, there is still a need to formulate health states with different combinations of levels of the selected items. A technique that serves this purpose is *k*-means cluster analysis. Sugar et al. (1998) employed *k*-means cluster analysis to identify patterns in the physical and mental health domains of the Medical Outcomes SF-12 questionnaire. These patterns were used to formulate health states. Applying this approach to the Rasch-reduced instrument "allow[s] the data to speak for themselves" (Sugar et al., 1998) in defining the health states. This method assigns different combinations of item level based on the natural groupings of the dataset.

This paper aims to describe the novel combination of Rasch and *k*-means cluster analyses to develop health states in RA. Since the HAQ is commonly used in RA, it was subjected to Rasch analysis to identify the items that best represent disability in RA. Using the reduced HAQ, *k*-means cluster analysis identified three different RA states. The development of three states was deemed appropriate, for our purposes, in providing a range of RA states for respondents to value yet to not overburden them when completing a valuation study.

2. Methods

2.1 The HAQ

While there are other disease-specific instruments that can assess disability in RA (for example, de Jong et al., 1997), the HAQ was chosen because it has been administered in various rheumatic populations for nearly three decades. Despite its popularity, the complete HAQ (which also includes questions on symptoms, medication use, and medical history) is lengthy, posing a burden on respondents. The HAQ also has a floor effect, such that severely disabled individuals are sometimes represented by normal HAQ scores (Wolfe et al., 2004).

The component of the HAQ of relevance to this study contains 20 items and assesses an individual's ability to complete daily tasks in dressing and grooming, arising, eating, walking, personal hygiene, reach, grip, and other activities. Two or three items comprise each domain (Table 1). Each item has four levels: no, some, or much difficulty performing the task, or an inability to perform the task. Respondents can select a score between zero and three, with higher scores implying a greater disability. The score on an individual item is increased by another point when the respondent requires assistive devices or additional help. The greatest item score – the most difficult task – yields the overall score for that domain.

2.2 Dataset

An anonymized dataset containing information from 600 randomly selected RA patients living in the United States was obtained from the National Data Bank for Rheumatic

Diseases (NDB) in Wichita, Kansas. The NDB is a non-profit organization that performs research in RA, osteoarthritis, fibromyalgia, lupus, and other rheumatic diseases (National Data Bank for Rheumatic Diseases, n.d.). The data bank contains longitudinal outcomes research data from patients reporting on all aspects of their illness in detailed semi-annual questionnaires, including the full HAQ; visual analogue scales (VASs) assessing global severity, pain, fatigue, sleep problems, and gastrointestinal symptoms; and the EuroQol-5D (EQ-5D) (Brooks, 1990).

Rasch analysis is sensitive to large sample sizes; using too large a sample generates a greater frequency of statistically significant items (Rasch, 1960), making item reduction difficult. Sample sizes on the order of 400-500 are recommended (Young et al., 2007). For this study an equal number of individuals in three severity ranges (n = 200) ensured that each disability level was well represented. The classification of severity was determined by the patient's total HAQ score (i.e. <1, 1-2, >2). In health state development, it is important to have a good distribution of responses for each item level.

2.3 Initial criteria for reducing the HAQ

The HAQ had to be reduced to a tractable number of items that best represent disability reported by RA patients. Reducing the HAQ to a total of five items was proposed, as previous studies have shown that instruments with five items, such as the EQ-5D and the Asthma Quality of Life Utility Index (Yang et al., 2007), do not overburden respondents. Although five was arbitrarily set, we felt that developing a classification system with more than five items would introduce unnecessary complexity for the respondents of our future valuation study. *A priori* criteria were imposed to ensure that the final health states would contain items that (i) described a combination of upper- and lower-limbed disabilities, (ii) belonged in separate HAQ domains to avoid the potential for collinearity, and (iii) captured the widest range of severity possible.

While Rasch analysis was the main approach in the reduction of HAQ items, other statistical methods were conducted simultaneously to ensure greater strength in the results. The frequency and the internal consistency – the correlations between item and

domain scores – were initially evaluated for the HAQ responses (Young et al., 2007). If some items elicited poor responses (e.g. low frequency) or poor internal consistency (e.g. weak correlation), they were considered to be less representative of disability for the given dataset. As done previously, questions pertaining to the use of assistive devices were excluded from the analysis (Tennant et al., 1996; Wolfe et al., 2004). Although these studies did not discuss their rationale for excluding this information, we felt the relationship between the use of aids and the HAQ items was ambiguous. For example, the use of a walking cane most likely relates to the walking domain (e.g. walking on flat surfaces and climbing up steps) but it is possible that this device may also aid in the arising domain (e.g. standing up from a straight and armless chair). This potential correlation made it difficult to assess which aid corresponded to which item, making it difficult to incorporate such aspects in the modelling procedures.

2.3 Selection of HAQ items that best describe disability

2.3.1 Rasch analysis

Rasch analysis verifies that the scale of the instrument is unidimensional, a fundamental requirement of construct validity. Unidimensionality ensures that the overall score of the instrument is describing what is actually happening and not diluted by items that are insensitive to the underlying construct of the instrument (Streiner and Norman, 1989). Fitting data to the Rasch model allows inferences to be made regarding desirable characteristics of the instrument (Tennant et al., 1996). The Rasch model has claims that:

- The easier the item is, the more likely it will be passed (or affirmed), and
- The more able the respondent, the more likely he/she will pass (or affirm) an item (or do a task) compared to a less able respondent.

Rasch analysis deconstructs each item of the instrument into its component steps: in the HAQ, from zero (i.e. no difficulty with the task) to one (i.e. some difficulty with the task), from one to two (i.e. much difficulty with the task), and from two to three (i.e. unable to perform the task). It then examines the likelihood individuals are in successfully attaining each item level. This gives an estimate of item difficulty, which is then used to assess the ability of the person. As unidimensionality is a pre-requisite for

the summation of any set of items (Streiner and Norman, 1989), the Rasch model assumes that the probability of a given patient passing (or affirming) an item or task is a logistic function of the relative distance between the item location parameter (i.e. the difficulty of the task) and the respondent location parameter (i.e. the ability of the patient):

$$p_i(q) = \frac{e^{(q-b_i)}}{1+e^{(q-b_i)}}, \quad (1)$$

where $p_i(q)$ is the probability that patients with ability q will be able to do item (task) i, and b is the item (task) difficulty parameter.

The Rasch analysis then seeks to combine person ability and item difficulty by taking the difference of these two values. This difference governs the likelihood of what is supposed to happen when a person of given ability uses that ability against a given task (Tennant et al., 1996). The results of the Rasch transformation are reported in logits, the distance along the line of the variable which increases the odds of observing the event (i.e. taking a step on an item) by a factor of 2.718. The relationship between person ability and item difficulty can be best understood by the fact that, for example, a person with a logit score of 2.0 will have an equal probability of passing (or affirming) or not passing an item with a difficulty level of 2.0 logits.

The overall goodness-of-fit test statistic, measured in terms of item-trait interaction, person separation index (PSI), and fit residuals, describes how well the Rasch model fits the original data (Young et al., 2007; Young et al., 2008). Item-trait interaction measures whether the data fits the Rasch model for the given respondent group. PSI calculates the level of agreement between respondents, while fit residuals estimate the degree of divergence between the expected and observed responses for each respondent or item response. Fit residuals are summed over all items for a given person (item fit residuals) or over all persons for a given item (person fit residuals).

To fit the Rasch models, the computer program RUMM2010 (RUMM Laboratory, Duncraig, Australia) was used.

2.3.2 Conducting Rasch analysis

The Rasch analysis framework proposed by Young et al. (2007 and 2008) was used to derive the health states for this study. This process used is summarized in Figure 1 and is described in more detail in the following steps.

Step I: Short-list the HAQ items

After conducting Rasch analysis on the full HAQ, the threshold probability curves for each item were inspected. The threshold probability curves assess the probability of an individual being in each item level across the latent scale, in which the horizontal and vertical axes represent the underlying latent scale and the probability of being in a particular item level, respectively. If any of the levels were disordered, such that a more difficult level was more likely attained than an easier level, the discrepant levels were merged together. Figure 2a illustrates that 'much difficulty' (i.e. a score of 2) is less probable than 'unable to do' (i.e. a score of 3); as a result, the aforementioned levels are merged together. With each merged item, a new base model resulted and the threshold probability curves for each item were re-examined; this step was repeated until all item levels are appropriately ordered (Figure 2b, the ideal graph). Although the merged items were included with each subsequent Rasch modelling, they were not considered in the final health state descriptions; respondents were considered to be indiscriminate towards these item levels.

If, after merging of the disordered item levels, any of the levels for the remaining items were poorly spread (such that the spacing between item levels was not evenly distributed – item level curves were not of approximately equal distance – when inspected visually) or have a low chance of occurring (such that a curve lies close to the bottom horizontal axis), these item levels were merged. An example of poorly spread levels is shown on Figure 2c and an example of a level lying close to 0% probability is shown on Figure 2d. This step requires that the suspect item level be merged with the adjacent level *one item at a time*. This step was conducted independently for each item. Both the overall item-trait fit and the individual item fit test statistics were examined to determine the best

possible model that resulted. If any of the individual items did not contribute to the underlying latent scale ($p < 0.01$), it was excluded from any subsequent modelling. The model with the smallest overall item-fit test statistic (i.e. largest p-value) was chosen to be the best resultant model. This new model was re-fitted and the overall goodness-of-fit statistic examined. The process was repeated until only well-fitting items remained (such as in Figure 2b) and the overall item-trait goodness-of-fit of the model was greater than $p = 0.01$.

Once the model fit was satisfied, the items that were excluded from the construction of the health states were:

- Items that needed merging at the initial Rasch model-fitting stage (i.e. indistinguishable, poorly distributed, and low probability of occurring levels), and
- Items that did not measure the underlying HRQL trait of the HAQ domains (i.e. the items did not fit the Rasch model).

The excluded items coincide with previously published techniques (Young et al., 2007; Young et al., 2008).

Step II: Differential item functioning

Differential item functioning (DIF) analyses provided further information about which HAQ items to include in the final health states. DIF analysis examines whether any items in the questionnaire result in discrepant responses amongst different respondent characteristics. Item characteristic curves were examined to assess whether sex, age (<50 years, 50-65, and >65), duration of RA (<10 years, 10-20 years, and >20 years), and total HAQ score (<1, 1-2, and >2) influenced item responses; these parameters were believed to affect HAQ scores. The levels were selected to ensure approximately equal numbers of individuals in each group. If visual inspection suggested that the likelihood of responses differed significantly between subgroups or between one of the subgroups and the mean response (for example, Figure 3 where there is an apparent difference between sexes), it is possible that this item did not fit the model well. Hence, it should be considered for exclusion from the final health states. This was verified by interpreting the F-test statistic, with the null hypothesis being no difference between subgroups. Any items which

demonstrated DIF were not excluded at this point but were considered as a suspect for removal in the next step.

Step III: Final selection of HAQ items

Using the location scale, the remaining items – after short-listing – were removed one at a time to appraise which eliminated item resulted in a better fitting model. The position of the item on the disability scale indicates the degree of disability represented by the item: a negative value indicates an item of lesser disability and a positive value indicates an item of greater disability. The greater the distance between the maximum and minimum values, the more sensitive the disability scale is. This stage was repeated until the desired number of items remained.

2.4 Forming health states

2.4.1 K-means cluster analysis

Once the reduced set of items from the HAQ was selected for inclusion in the final health states, the next step was to conduct cluster analysis in order to form three RA states. The use of five items (each with four levels) affords 625 possible health states (5^4); cluster analysis provides a statistical approach that reduces these to a smaller set of well-defined health states that, in this case, summarize RA patients' disability. In addition to being distinct, the final health states needed to describe QOL for people with different patterns of health. While there are several types of cluster analysis, including hierarchical, two-step, and expected maximization, the *k*-means algorithm was employed for this work because it allows the formation of asymmetrically-spaced clusters (Sugar et al. 1998).

The *k*-means algorithm aims to group *n* observations into *k* partitions or clusters by finding the centres of natural clusters in the given dataset. The algorithm starts by randomly partitioning the data points into *k* initial sets. Then the mean point, or centre, is calculated for each set. The algorithm then constructs a new partition by associating each point with the closest centre. The centres are re-calculated for the new clusters and the algorithm is repeated until convergence is achieved, such that the data points no longer switch clusters. This approach seeks to identify a set of groups which both minimizes

within-cluster variation and maximizes between-cluster variation in a similar fashion to that of analysis of variance (ANOVA). Figure 4 describes the steps involved in the cluster analysis.

SPSS version 14.0 for Windows (SPSS, Chicago, USA) was used to conduct the *k*-means cluster analysis.

2.4.2 Conducting cluster analysis

Step I: Determine the optimal number of clusters

While the *k*-means algorithm specifies cluster membership once the number of clusters is fixed, the number of clusters needs to be selected in advance (Sugar et al., 1998). The root mean-squared distance to cluster centres – known hereafter as root mean-squared error (RMSE) – was used to determine the number of clusters. As there is a trade-off between having too few or too many clusters – with a single cluster, the RMSE is large and with many clusters, the RMSE approaches zero – a plot of the RMSE versus number of clusters indicates the optimal number of clusters for any given dataset. This plot results in a steadily decreasing curve; at some point, the rate of decrease drops sharply because the data points are genuinely clumped into a fixed number of groups. The part of the curve where the slope changes most abruptly indicates the true optimal number of clusters for the data. While only three clusters were needed for our subsequent valuation study, the ideal number of clusters for the given dataset was still determined. This was achieved by visual inspection of the plot of the RMSE versus number of clusters, and calculation of the slope between adjacent cluster centres.

Step II: Check the stability of the clusters

Three health states, or clusters, needed to be defined for use in our subsequent valuation studies. However, to check the cluster stability, the results from running the algorithm with the optimal number of clusters – as verified by the plot of the RMSE versus cluster numbers – were also evaluated. The *k*-means algorithm was run on the full HAQ and on the Rasch-reduced HAQ. If the disability measures from the reduced HAQ are similar to

the full HAQ then the combinations of item levels after cluster analysis for the reduced HAQ should be similar to that for the full HAQ.

2.6 Pain and discomfort

A common symptom experienced by patients with RA is pain and discomfort. However, in the HAQ, the level of pain experienced is measured continuously on a VAS, ranging from no pain to severe pain. Within the Rasch framework, discrete rather than continuous variables are modeled, and hence pain and discomfort would not have been considered. As a result, we decided to use one of the five items devoted to pain and discomfort, rather than have all items based on the HAQ. The severity of this item was based on the patients' responses to the domain of the same name on the EQ-5D. The allocation of pain level was determined by the proportion of responses in each pre-defined severity group found in the dataset (i.e. a HAQ score <1, 1-2, and >2), such that the most frequent level defined that severity; this was identified for the three defined clusters. As with the reduction of the HAQ items, personal judgements should be employed to ensure that the final health states were comprehensible and plausible to individuals who may not be well informed about RA.

2.7 Additional analysis

The NDB data were characterized, in terms of age, sex, RA duration, HAQ score, EQ-5D score, and EQ-5D VAS score. Continuous variables were presented as means and standard deviations (SD), while categorical variables were presented as the proportion of the sample within each group. Each HAQ score group was characterized by RA duration, EQ-5D score, and EQ-5D VAS score.

The internal consistency of the HAQ responses was tested using Spearman's correlation coefficients; the cut-off value for this criterion was rho $\leq$ 0.7 (Young et al., 2007). With the final health states, one-way ANOVAs evaluated the differences among the respondents' age, RA duration, and instrument scores when stratified by cluster membership.

3. RESULTS

3.1 Study population

Table 2 displays information regarding the demographic variables and the QOL scores of the NDB participants. On average, individuals had been living with diagnosed RA for 16.7 $\pm$ 11.7 years. In terms of EQ-5D, the mean score – derived from converting the individual responses on the classification system using the US societal tariffs (Shaw et al. 2005) – is 0.67 $\pm$ 0.22. The unstandardized mean score from the EQ-5D VAS is 63.16 $\pm$ 21.04. A gradient is observed across HAQ scores, such that people with higher HAQ scores (more severe RA) reported poorer QOL than individuals with mild and moderate RA (Table 3), independent of which valuation process was used.

The results shown in Table 4 demonstrate a gradient between the frequency of responses and the difficulty of the items (items are summarized in Table 1): patients were more likely to report 'no difficulty' than 'some difficulty' and so forth for the items. The most difficult level, 'unable to do', was relatively infrequent amongst most items; however, the items *overhead* and *tubbath* produced a greater frequency of these responses when compared to the adjacent level of 'much difficulty'. Interestingly, the 'unable to do' level for *tubbath* was the most frequently responded level at 40%; this result may be due to the increased number of homes with only showers installed.

The correlation between the item and domain scores was also examined (Table 5). The majority of items were internally consistent (rho $\geq$ 0.7); however *fauceton* (rho = 0.68), *washbody* (rho = 0.59), and *ontoilet* (rho = 0.59) did not fit this criterion. This provided evidence that these items may not be representative of the domain and are therefore inappropriate for inclusion in the final health state descriptions.

3.2 Rasch analysis

Step I: Short-list the HAQ items

The threshold probability curves for all HAQ items were examined. The results indicated that all the item levels were ordered appropriately except for the most severe levels of *tubbath* and *shampoo*; the misfitting nature of these items has been reported elsewhere (Wolfe et al. 2004). As a result, the second (much difficulty with the task) and third (inability to perform the task) levels of these two items were merged together to form a new base model. This new model is identical to the original HAQ, except that *tubbath* and *shampoo* now contain three levels instead of four.

Rasch analysis on the new base model indicated that the *runerand* item did not fit the model ($p \leq 0.01$) and thus, this item was removed from subsequent Rasch modelling. The new base model is similar to the original HAQ, except that both *tubbath* and *shampoo* contain a total of three levels and that *runerand* is not included.

Once again a Rasch model was conducted on the current base model and the threshold probability curves for all items containing four levels were re-evaluated. From these curves, items that were indistinguishable from each other or lay close to the 0% probability line were merged together one at a time until the best model emerged after examining the overall goodness-of-fit and individual item fit. A summary of the items short-listed at this stage is presented in Table 6. At the end of short-listing, nine items remained: *benddown*, *climstep*, *liftcup*, *standup*, *walkflat*, *openjars*, *opencar*, *inoutcar*, and *cutmeat*. These items did not have disordered levels, poorly spread levels, or levels lying close to the 0% probability axis.

Step II: Differential item functioning

DIF analyses were conducted to determine whether or not the remaining HAQ items resulted in differential responses between patient subgroups. Both the item characteristic curves and the F-test statistics indicated which items did not fit the model well. The results indicated that the item *openjars* could potentially be removed from the final health states because the individual curves deviated from the mean item characteristic curve

(see Figure 3 for an example). *Openjars* was deemed eligible for removal; this was verified in Step III. No other item indicated DIF.

Step III: Final selection of the HAQ items

As there were still nine items to be considered for the final health states, each item was removed one at a time, based on the position of the item on the difficulty scale, until the best model with four items was produced. The scale indicated that, for example, most individuals with a mild form of RA would have problems climbing up steps (i.e. represented by a negative value); however, only the most severe cases in this group would be unable to lift a cup to their mouth (i.e. represented by a location value) (Table 7). Of the items under consideration, only *climstep* had a negative location value, representing the mildest form of severity. Thus, it was retained to ensure the final model had the widest range of severity. Using the sequential removal-reassessment process – where one item was removed and individual item test statistics of each model was examined to determine the best model that arose – the following items were removed: *opencar*, *openjars*, and *cutmeat*.

At this stage, *benddown*, *climstep*, *liftcup*, *standup*, *walkflat*, and *inoutcar* remained. To avoid collinearity, the inclusion criteria required that items be from separate domains. Hence, *walkflat* was removed as it was from the same domain as *climstep*, which was previously identified as a desireable item because it represented the mildest form of severity. We also felt that *inoutcar* could result in discrepancies between patients' responses due to the potential for the use of modified cars and potential assistance provided by from travelling companions. As a result, this item was removed from the final description. Therefore, the final Rasch-reduced model (χ^2 = 44.4, p = 0.03) is composed of: *benddown*, *climstep*, *liftcup*, and *standup* (Table 8).

3.3 *K*-means cluster analysis

Step I: Determine the optimal number of clusters

The plot of the RMSE versus number of clusters revealed that the optimal number of clusters appears to be between three and four (Figure 4); this is where the greatest change

in slope was observed (Table 9). As such, three and four cluster centres were examined in greater detail.

Step II: Check the stability of the clusters

Running cluster analyses on both the full and reduced HAQ showed that the combinations of item levels generally remained the same for both models, with only *climstep* moving to the next level of disability (Table 10). Three clusters were needed to represent three RA states for our future valuation study. As the first cluster for all models showed no disability – this state was synonymous to full health – the three-cluster model was insufficient to describe three RA states. Thus, cluster centre two, three, and four of the four-cluster model described three RA states, which seemed to be representative of individuals with a mild, moderate, and severe form of RA, respectively. Cluster one of this model represents individuals with a very mild form of RA.

3.4 Inclusion of pain and discomfort domain

Amongst the three HAQ severity groups, the most frequent response was moderate pain and discomfort (Table 11); this may be due, in part, to the lack of sensitivity between the three levels describing this domain. To ensure that the respondents can differentiate across the three states, pain and discomfort was labelled as mild, moderate, and extreme to describe the health states. This corresponded to the proportion of individuals living in mild, moderate, and severe RA HAQ score groups.

3.5 Respondent characteristics of final health states

Table 12 displays the demographic and QOL information of the patients in the dataset. Although three health state descriptions are needed for subsequent empirical studies, the majority of the patients in this dataset were classified with very mild RA and therefore four states are described. The ANOVA results revealed that there were no differences in age and RA duration across cluster groups ($p = 0.54$ and 0.07, respectively). However, in terms of HAQ, EQ-5D, and EQ-VAS scores, the QOL measures distinguished well across the severity groups ($p \leq 0.001$). In all cases, a monotonic gradient was observed, such that a lower QOL was associated with more severe forms of RA. These results provided

evidence that the health states constructed by cluster analysis had the ability to discriminate between different levels of RA severity.

4. Discussion

The results demonstrate that the combined use of Rasch and cluster analyses can create distinct and plausible health state descriptions; in this example, RA states were developed. To our knowledge, the application of both Rasch and cluster analyses is a novel approach in the construction of health state classification systems.

The construction of health states is based on an analysis of content. Rasch analysis is regarded as a tool to aid in the development of health states. The statistical technique used in assessing the goodness-of-fit of the models provides guidance but do not supersede experienced judgements. The items included in the RA descriptions need to clearly describe the full range of disability associated with this chronic condition. While non-patients may tend to focus on the classic symptoms of RA, such as mobility (e.g. climbing steps), the purpose of the states was to create a descriptive and well-rounded picture of individuals living with RA across three distinct states through a limited number of pre-specified items. Alongside the item describing difficulties with climbing steps, it was important to encompass everyday tasks into the descriptions, such as bending down, lifting a full cup, standing up from a straight and armless chair, and experiencing pain and discomfort.

To date, there have been two other studies that have used Rasch analysis on the HAQ, although their overall objectives were different from this present study. Tennant et al. (1996) investigated the scaling of the HAQ and the fit of the data to the Rasch model. Similar to the findings of the present study, the item *liftcup* adequately represented the upper level of disability, such that those who have difficulty with this task, or find it impossible, have the severest form of RA. While Wolfe et al. (2004) applied Rasch analysis to the HAQ, they opted to do so on a revised version of the HAQ, referred to as the HAQ-II. They reduced the original HAQ down to ten items but their final items were not comparable to the ones selected here. The authors introduced new items when

modifying the HAQ; these items included: waiting in a line for 15 minutes, doing outside work, lifting heavy objects, and moving heavy objects. The remaining six items of the HAQ-II were *ontoilet*, *opencar*, *walkflat*, *overhead*, *standup*, and *climstep*; of which, only *standup* and *climstep* remained in the Rasch-reduced HAQ found in the current work. For the four discrepant items between the two studies, all items except *walkflat* were excluded based on the poorly spread criterion; *walkflat* was eliminated because of concerns of potential collinearity with *climstep* when the latter remained in the RA description.

This is believed to be the first study to combine Rasch and cluster analyses in the development of health states. Rasch analysis demonstrates clear advantages for this type of exercise over the use of factor analysis, which is more useful in examining multi-dimensional instruments, and qualitative approaches, which obtain information from a small number of individuals participating in interviews or focus groups. Although a floor effect has been reported with the HAQ (Wolfe et al. 2004), by fitting the HAQ items into a Rasch model, the floor effect can be reduced by selecting items which contribute to the widest possible severity range on the location scale. The use of *k*-means cluster analysis identified groups of individuals. The descriptions of these individuals differed in terms of severity. Although, this approach is strictly driven by the given dataset, the resulting health states may not be generalizable to other populations.

Four of the five items in the health states were composed of the HAQ. While the HAQ measures pain and discomfort, these responses were not used in the construction of the health states. The continuous representation of pain on the HAQ may be more meaningful to physicians or rheumatologists monitoring their patients over multiple time points. In a cross-sectional study, these ratings may carry less weight when appraised by non-patient respondents. As pain is a common symptom experienced by most patients living with RA, including this domain was critical in providing an appropriate description of the impact of RA on health. Therefore, to ensure that the health states are comprehensible and, more importantly, plausible, the decision was made to include the pain and discomfort domain, but rather than using this domain from the HAQ, to substitute it with

the domain from the EQ-5D data collected from the same patients at the same time as the HAQ data.

In the EQ-5D, the pain and discomfort domain is described by three levels. However, for this study, the current definition of these levels did not provide enough sensitivity to describe the three health states. As such, a fourth level was created to represent "you have mild pain and discomfort"; this was used for the mildest form of the RA state. The three health states were piloted in a small (nine respondents) convenient sample. Overall, they found that the descriptions were clear and plausible.

As with any study, there are limitations; however, none of them should significantly affect the findings. There is a potential for misclassification as the HAQ score was used as a proxy to determine the RA severity of the patients. While this may seem like an adequate approach, the instrument's floor effect implies that the most severe patient could be represented by a low (i.e. decent) HAQ score. However, across the different measures of QOL, which included EQ-5D and EQ-5D VAS, the relationship between RA severity and the QOL measures at the aggregate levels is in the anticipated direction, suggesting that any misclassification of severity by HAQ score is likely minimal. In addition, the nature of how the information was collected may be subject to self-reporting bias. It is possible that individuals may, either consciously or subconsciously, underestimate the individual item score of the HAQ to please their physician and family members for fear that it may alter the current treatment regimen; this result would result in HAQ scores lower (i.e. higher QOL) than they should actually be.

This study was based on data from 600 patients recruited through the NDB in Wichita, Kansas; as such, the results may not be generalizable. Although the sample encompassed all levels of disease severity, the study sample may not be representative of the general RA population in the US. From the NDB website (National Data Bank for Rheumatic Diseases, n.d.), respondents are not offered direct incentives for participation but instead are offered a chance to win one of three $1,000 (US dollar) lotteries if the research questionnaire is returned within two weeks. This financial incentive may provide

differential participation in patients of different social classes. As the dataset did not include information on socio-economic status variables (e.g. annual income, highest attained education level), it is not possible to ascertain whether the socio-economic status of the respondents varied significantly from that of the general RA population. This, however, would be a concern only if the distribution of HAQ scores or the distribution of different clinical manifestations of RA was related to social class.

Despite the potential limitations discussed above, the objective of the first phase of this research was achieved: distinct yet credible RA states were defined. The results presented in this paper demonstrate a methodology for reducing a dataset containing the individual HAQ scores to generate a framework for the RA health states. Overall, the combined use of Rasch and cluster analyses, with the results assessed subjectively based on expert judgement, has proven to be an effective technique for identifying the most important items and levels in the construction of health states to be used in health valuation studies.

5. Acknowledgements

The authors would like to thank Fred Wolfe for generously allowing access to the dataset from the National Data Bank for Rheumatic Diseases and for input during the health state development, and Tracey Young for her invaluable input during the development of the health states and comments during the manuscript preparation stage. Helen McTaggart-Cowan is supported by a doctoral research award from the Canadian Institutes for Health Research.

6. References

Brooks, R. (1996). EuroQol: the current state of play. *Health Policy*, **37**(1), 53-72.

Bruce, B., and Fries, J.F. (2003). The Stanford Health Assessment Questionnaire: a review of its history, issues, progress, and documentation. *Journal of Rheumatology*, **30**, 167-178.

de Jong, Z., van der Heijde, D., Mckenna, S. P., and Whalley, D. (1997). The reliability and construct validity of the RAQoL: a Rheumatoid Arthritis-Specific Quality of Life instrument. *British Journal of Rheumatology*, **36**, 878–883.

Fayers, P.M., and Machin, D. (2000). Factor analysis. In: *Quality of life: Assessment, analysis and interpretation*. New Jersey: John Wiley and Sons, 90-116.

Fries, J.F., Spitz, P., Kraines, G., and Holman, H. (1980). Measurement of patient outcome in arthritis. *Arthritis and Rheumatism*, **23**, 137-145.

Marra, C.A., Woolcott, J.C., Kopec, J.A., Shojania, K., Offer, R., Brazier, J.E., Esdaile, J.M., and Anis, A.H. (2005a). A comparison of generic, indirect utility measures (the HUI2, HUI3, SF-6D, and the EQ-5D) and disease-specific instruments (the RAQoL and the HAQ) in rheumatoid arthritis. *Social Science and Medicine*, **60**(7), 1571-1582.

Marra, C.A., Rashidi, A.A., Guh, D., Kopec, J.A., Abrahamowicz, M., Esdaile, J.M., Brazier, J.E., Fortin, P.R., and Anis, A.H. (2005b). Are indirect utility measures reliable and responsive in rheumatoid arthritis patients? *Quality of Life Research*, **14**(5), 1333-1344.

National Data Bank for Rheumatic Diseases (n.d.). *NDB frequently asked questions* [online]. Available from URL: http://www.arthritis-research.org/ndb_faq.htm [Accessed: 15 November 2007].

National Institute for Clinical Excellence. (2003). *Guide to the methods of technology appraisal*. London: National Institute for Clinical Excellence.

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

Shaw, J.W., Johnson, J.A., and Coons, S.J. (2005). U.S. valuation of the EQ-5D health states. Development and testing of the DI valuation model. *Medical Care*, **43**, 203-220.

Streiner, D., and Norman, G. (1989). *Measurement scales*. Oxford: Oxford University Press.

Sugar, C.A., Sturm, R., Lee, T.L., Sherbourne, C.D., Olshen, R.A., Wells, K.B., and Lenert, L.A. (1998). Empirically defined health states for depression from the SF-12. *Health Services Research*, **33** (4), 911-928.

Symmons, D.P.M. (2005). Looking back: rheumatoid arthritis – aetiology, occurrence and mortality. *Rheumatolgy*, **44**(Suppl. 4): iv14-iv17.

Tennant, A., Hillman, M., Fear, J., Pickering, A., and Chamberlain, M.A. (1996). Are we making the most of the Stanford Health Assessment Questionnaire? *British Journal of Rheumatology* **35**, 574-578.

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

Tesio, L. (2003). Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research. *Journal of Rehabilitation Medicine*, **35**(6), 105-115.

Wolfe, F., Michaud, K., and Pincus, T. (2004). Development and validation of the Health Assessment Questionnaire II. A revised version of the Health Assessment Quesitonnaire. *Arthritis and Rheumatism*, **50**(10), 3296-3305.

Young, T.A., Yang, Y., Brazier, J., and Tsuchiya, A. (2007). The use of Rasch analysis as a tool in the construction of preference-based measure: the case of AQLQ. *Discussion paper 07/01*. Sheffield: Health Economics and Decision Science.

Young, T., Yang, Y., Brazier, J., Tsuchiya, A., and Coyne, K. (2008). Making Rasch decisions: the use of Rasch analysis in the construction of preference based health related quality of life instruments. *Discussion paper 08/05*. Sheffield: Health Economics and Decision Science.

Yang, Y., Tsuchiya, A., Brazier, J., and Young, T.A. (2007). Estimating a preference-based single index from the Asthma Quality of Life Questionnaire (AQLQ). *Discussion paper 07/02*. Sheffield: Health Economics and Decision Science.

7. TABLES

Table 1: Items in the HAQ

HAQ Item	Domain	Description of Item
Dresself	Dressing & Grooming	Dress yourself, including tying shoelaces and doing buttons
Shampoo	Dressing & Grooming	Shampoo your hair
Standup	Arising	Stand up from a straight chair
Inbed	Arising	Get in and out of bed
Cutmeat	Eating	Cut your meat
Liftcup	Eating	Lift a full cup or glass to your mouth
Openmilk	Eating	Open a new milk carton
Walkflat	Walking	Walk outdoors on flat ground
Climstep	Walking	Climb up five steps
Washbody	Hygiene	Wash and dry your body
Tubbath	Hygiene	Take a tub bath
Ontoilet	Hygiene	Get on and off the toilet
Overhead	Reach	Reach and get down a 5-pound object (such as a bag of sugar)
Benddown	Reach	Bend down to pick up clothing from the floor
Opencar	Grip	Open car doors
Openjars	Grip	Open jars which have been previously opened
Faceton	Grip	Turn faucets on and off
Runerand	Activities	Run errands and shop
Inoutcar	Activities	Get in and out of a car
Vacuum	Activities	Do chores such as vacuuming or yard work

Table 2: Characteristics of the study population

	Mean (± SD) or Frequency (%)
Age	61.6 (±12.8)
Duration of RA (years)	16.7 (±11.7)
HAQ	1.41 (±0.87)
EQ-5D	0.67 (±0.22)
EQ-5D VAS	63.16 (±21.04)
Female	474 (79%)

Table 3: Characteristics of each HAQ score group

HAQ Score	Duration in Years (SD)	HAQ Score (SD)	EQ-5D Score (SD)	EQ-5D VAS (SD)
< 1.00	15.1 (10.7)	0.36 (0.32)	0.85 (0.11)	76.28 (17.51)
1.00 to 2.00	16.3 (11.6)	1.51 (0.33)	0.68 (0.17)	60.01 (18.79)
> 2.00	18.7 (12.6)	2.35 (0.22)	0.50 (0.21)	53.20 (19.74)

Table 4: Frequency of level responses for HAQ items

HAQ Item	Number of Responses (%)				
	No Difficulty	Some Difficulty	Much Difficulty	Unable to Do	Missing
Dresself	261 (43.5)	242 (40.3)	68 (11.3)	25 (4.2)	4 (0.7)
Shampoo	329 (54.8)	155 (25.8)	49 (8.2)	46 (7.7)	21 (3.5)
Standup	246 (41.0)	243 (40.5)	91 (15.2)	14 (2.3)	6 (1.0)
Inbed	277 (46.2)	239 (39.8)	63 (10.5)	5 (0.8)	16 (2.7)
Cutmeat	325 (54.2)	169 (28.2)	80 (13.3)	20 (3.3)	6 (1.0)
Liftcup	363 (60.5)	159 (26.5)	59 (9.8)	7 (1.2)	12 (2.0)
Openmilk	224 (37.3)	189 (31.5)	108 (18.0)	63 (10.5)	16 (2.7)
Walkflat	284 (47.3)	211 (35.2)	75 (12.5)	26 (4.3)	4 (0.7)
Climstep	226 (37.7)	196 (32.7)	118 (19.7)	51 (8.5)	9 (1.5)
Washbody	349 (58.2)	196 (32.7)	43 (7.2)	8 (1.3)	4 (0.7)
Tubbath	178 (29.7)	113 (18.8)	58 (9.7)	238 (39.7)	13 (2.2)
Ontoilet	345 (57.5)	215 (35.8)	33 (5.5)	5 (0.8)	2 (0.3)
Overhead	209 (34.8)	177 (29.5)	93 (15.5)	118 (19.7)	3 (0.5)
Benddown	272 (45.3)	212 (35.3)	79 (13.2)	28 (4.7)	9 (1.5)
Opencar	340 (56.7)	185 (30.8)	57 (9.5)	15 (2.5)	3 (0.5)
Openjars	235 (39.2)	252 (42.0)	86 (14.3)	24 (4.0)	3 (0.5)
Fauceton	368 (61.3)	177 (29.5)	46 (7.7)	2 (0.3)	7 (1.2)
Runerand	241 (40.2)	209 (34.8)	96 (16.0)	50 (8.3)	4 (0.7)
Inoutcar	242 (40.3)	269 (44.8)	76 (12.7)	8 (1.3)	5 (0.8)
Vacuum	138 (23.0)	184 (30.7)	128 (21.3)	146 (24.3)	4 (0.7)

Table 5: Correlations between HAQ item and domain scores

HAQ Item	Domain	Spearman's rho
Dresself	Dressing & Grooming	0.911
Shampoo	Dressing & Grooming	0.825
Standup	Arising	0.947
Inbed	Arising	0.839
Cutmeat	Eating	0.813
Liftcup	Eating	0.760
Openmilk	Eating	0.967
Walkflat	Walking	0.841
Climstep	Walking	0.968
Washbody	Hygiene	0.585
Tubbath	Hygiene	0.990
Ontoilet	Hygiene	0.589
Overhead	Reach	0.956
Benddown	Reach	0.757
Opencar	Grip	0.761
Openjars	Grip	0.915
Fauceton	Grip	0.684
Runerand	Activities	0.789
Inoutcar	Activities	0.972
Vacuum	Activities	0.972

Table 6: Summary of items not considered for the final health states after short listing

HAQ Item	Reasons for Exclusion
Dresself	Item levels are poorly distributed
Shampoo	Two most severe levels disordered
Inbed	The most severe level have a low chance of occurring
Openmilk	Item levels are poorly distributed
Washbody	Item levels are poorly distributed
Tubbath	Two most severe levels disordered
Ontoilet	Item levels are poorly distributed
Overhead	Item levels are poorly distributed
Fauceton	The most severe level have a low chance of occurring
Vacuum	Item levels are poorly distributed

Table 7: Individual HAQ item fit for items still under consideration after short listing

HAQ Item*	Location	SE†	Fit Residual	DF†	$\chi^{2‡}$	p-value
Liftcup	1.62	0.09	-0.91	485.1	12.8	0.17
Opencar	1.26	0.08	-1.66	493.6	7.4	0.60
Inoutcar	1.04	0.09	-2.35	491.7	10.5	0.31
Cutmeat	0.89	0.08	-1.06	490.7	5.51	0.79
Standup	0.66	0.08	-0.24	490.7	7.79	0.56
Walkflat	0.58	0.08	0.31	492.6	11.3	0.26
Benddown	0.50	0.08	1.30	487.9	5.20	0.82
Shampoo	0.02	0.09	-1.50	476.6	7.47	0.59
Climstep	-0.24	0.07	0.53	487.9	17.7	0.04

* Runerand excluded from this analysis; bolded items represent items still under consideration for inclusion in final health states.
† DF = degrees of freedom; SE = standard error.
‡ DF for χ^2-test = 9.

Table 8: Individual item fit for the reduced HAQ (Model M11)

HAQ Item	Location	SE*	Fit Residual	DF*	$c^{2†}$	p-value
Liftcup	0.80	0.09	2.99	337.2	23.6	0.003
Inoutcar	0.33	0.09	-2.12	337.2	26.9	0.03
Standup	-0.05	0.09	-0.94	337.2	4.13	0.85
Benddown	-0.20	0.08	0.38	337.2	14.5	0.07
Climstep	-0.90	0.08	-0.93	337.2	4.00	0.86

* DF = degrees of freedom; SE = standard error.
† DF for χ^2-test = 8.

Table 9: Slope of each item to determine the optimal number of clusters

Cluster Numbers	Slope			
	Benddown	Climstep	Liftcup	Standup
3 to 4	-1.93	-2.21	-0.73	-1.70
4 to 5	-0.90	-1.60	-0.31	-1.32
5 to 6	-0.83	-0.92	-0.23	-0.52

Table 10: Results from the *k*-Means cluster analysis

HAQ Item	Cluster Centres*			
	1	**2**	**3**	**4**
Cluster, n = 3				
Dresself	0	1	2	---
Shampoo	0	1	1	---
Standup	**0 (0)**	**1 (1)**	**2 (2)**	---
Inbed	0	1	1	---
Liftcup	**0 (0)**	**0 (1)**	**1 (1)**	---
Openmilk	0	1	2	---
Climstep	**0 (0)**	**1 (1)**	**2 (2)**	---
Washbody	0	1	1	---
Tubbath	0	2	2	---
Ontoilet	0	1	1	---
Overhead	0	1	2	---
Benddown	**0 (0)**	**1 (1)**	**2 (2)**	---
Fauceton	0	0	1	---
Inoutcar	0	1	2	---
Vacuum	0	2	2	---
Cluster, n = 4				
Dresself	0	1	1	2
Shampoo	0	0	1	2
Standup	**0 (0)**	**1 (1)**	**1 (2)**	**2 (2)**
Inbed	0	1	1	2
Liftcup	**0 (0)**	**0 (0)**	**1 (1)**	**2 (2)**
Openmilk	0	1	2	2
Climstep	**0 (0)**	**1 (1)**	**1 (2)**	**2 (3)**
Washbody	0	0	1	1
Tubbath	0	1	2	2
Ontoilet	0	0	1	1
Overhead	0	1	2	2
Benddown	**0 (0)**	**1 (1)**	**1 (1)**	**2 (2)**
Fauceton	0	0	1	1
Inoutcar	0	1	1	2
Vacuum	0	1	2	2

* Items in bold represents the item level for the reduced Rasch model; number in (parenthesis) represents the item level for the reduced Rasch model

Table 11: Frequency of level responses for the pain and discomfort domain of the EQ-5D

	No Pain and Discomfort (%)	Moderate Pain and Discomfort (%)	Extreme Pain and Discomfort (%)
All Respondents	64 (10.7)	418 (69.7)	118 (19.7)
Very Mild RA	58 (23.4)	183 (73.8)	7 (2.8)
Mild RA	2 (1.2)	134 (82.2)	27 (16.6)
Moderate RA	2 (1.9)	60 (55.6)	46 (42.6)
Severe RA	0	26 (52.0)	24 (48.0)
Missing	2 (3.1)	15 (3.6)	14 (11.9)

Table 12: Characteristics of the final health states[*]

	n	Mean	Standard Deviation	Minimum	Maximum
Very Mild Rheumatoid Arthritis					
Age[**]	248	60.7	12.4	17.2	84.0
Duration of RA (years)[†]	248	15.8	10.8	1.2	63.5
HAQ[‡]	248	0.68	0.66	0	2.88
EQ-5D[#]	248	0.83	0.11	0.38	1.00
EQ-5D VAS[δ]	248	74.82	17.70	0	99.00
Female	185 (75%)				
Mild Rheumatoid Arthritis					
Age[**]	163	61.9	12.6	16.8	88.3
Duration of RA (years)[†]	163	16.1	12.3	1.3	61.1
HAQ[‡]	163	1.59	0.58	0.13	3.00
EQ-5D[#]	163	0.66	0.16	0.17	1.00
EQ-5D VAS[δ]	163	59.58	18.10	15.00	98.00
Female	127 (78%)				
Moderate Rheumatoid Arthritis					
Age[**]	108	61.2	14.1	23.3	89.6
Duration of RA (years)[†]	108	17.3	10.5	0.7	44.2
HAQ[‡]	108	2.17	0.29	0.88	2.75
EQ-5D[#]	108	0.52	0.18	0.20	0.85
EQ-5D VAS[δ]	108	52.25	18.32	13.00	99.00
Female	92 (85%)				
Severe Rheumatoid Arthritis					
Age[**]	50	63.1	12.4	31.4	89.9
Duration of RA (years)[†]	50	20.4	15.0	3.0	75.8
HAQ[‡]	50	2.51	0.25	2.00	3.00
EQ-5D[#]	50	0.39	0.20	-0.04	0.75
EQ-5D VAS[δ]	50	45.06	21.55	5.00	92.00
Female	43 (86%)				

[*] EQ-5D = EuroQol 5D; EQ-5D VAS = EuroQol 5D visual analogue scale; HAQ = Health Assessment Questionnaire; RA = rheumatoid arthritis.

[**] one-way ANOVA results: $F = 0.72$, $p = 0.54$

[†] one-way ANOVA results: $F = 2.35$, $p = 0.07$

[‡] one-way ANOVA results: $F = 285.03$, $p \leq 0.001$

[#] one-way ANOVA results: $F = 188.32$, $p \leq 0.001$

[δ] one-way ANOVA results: $F = 64.76$, $p \leq 0.001$

8. FIGURES

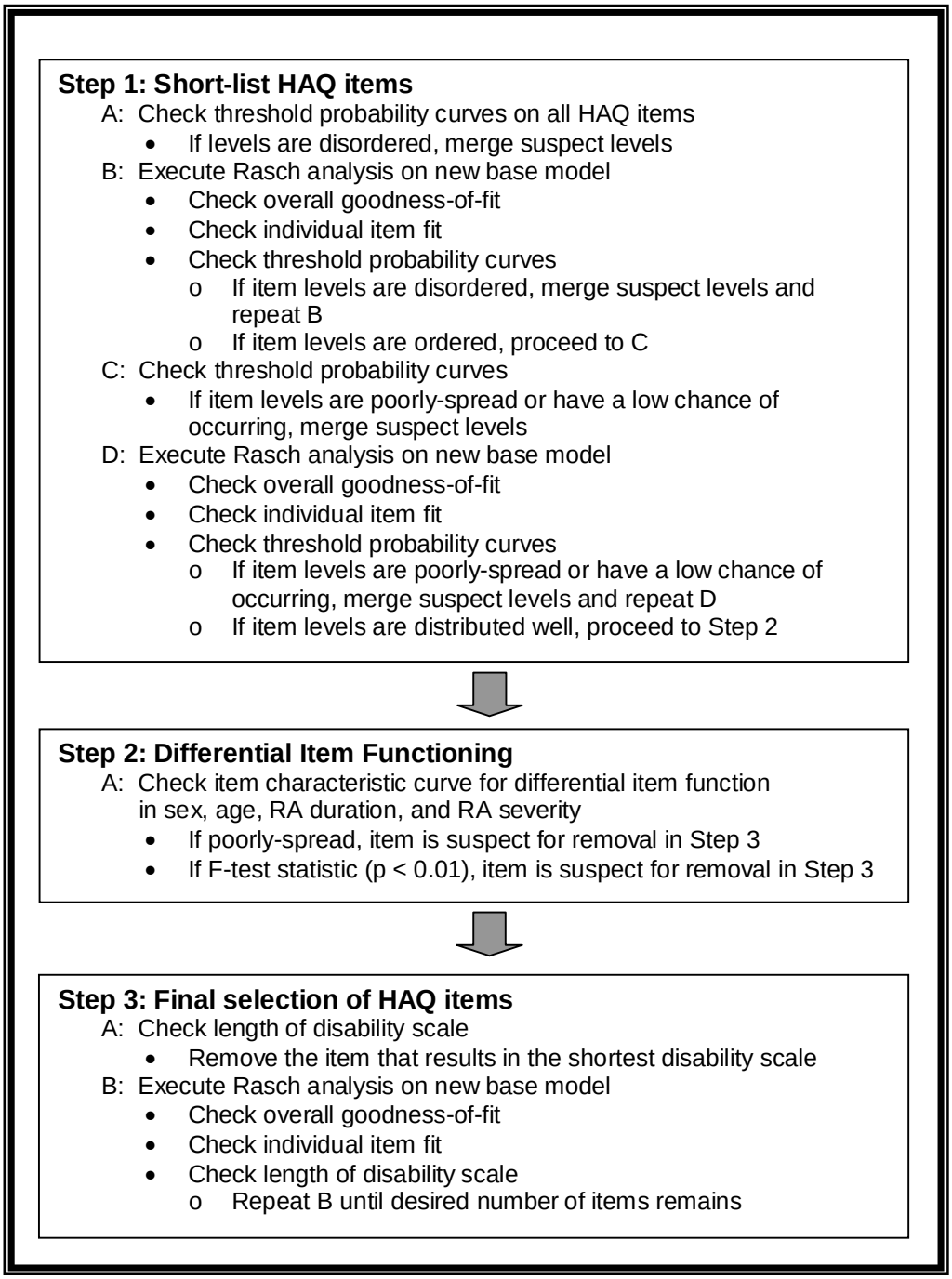

Figure 1: A schematic overview of Rasch analysis

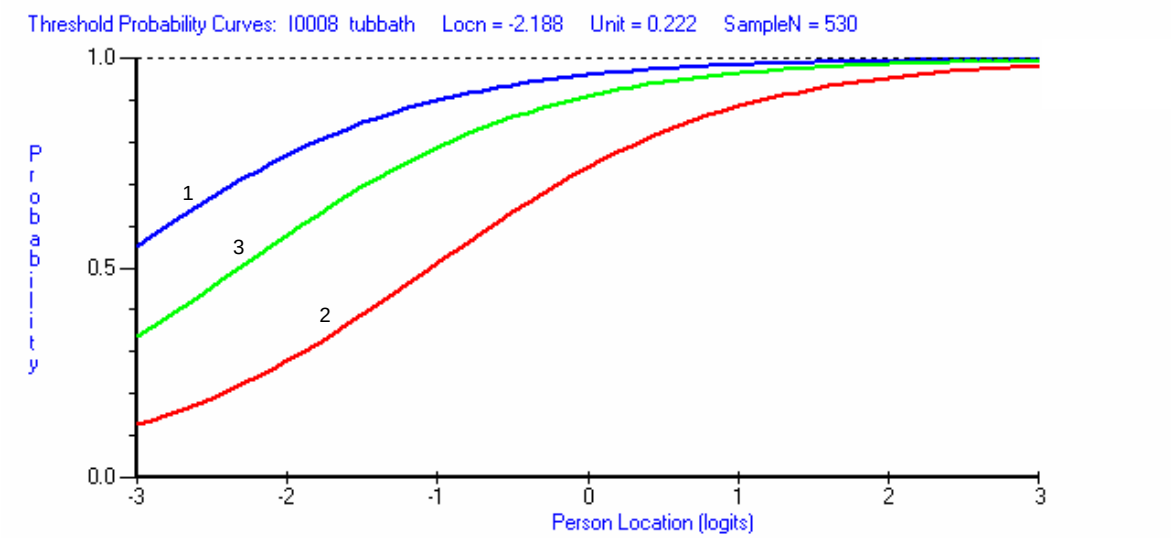

Figure 2a: An example of disordered item levels

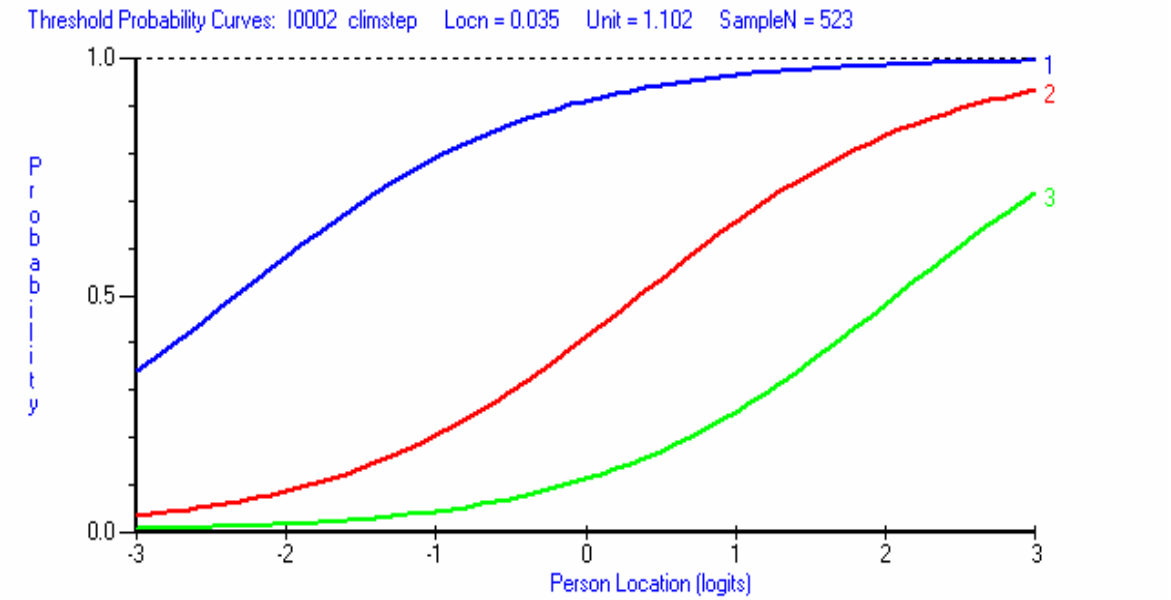

Figure 2b: An example of appropriately ordered item levels

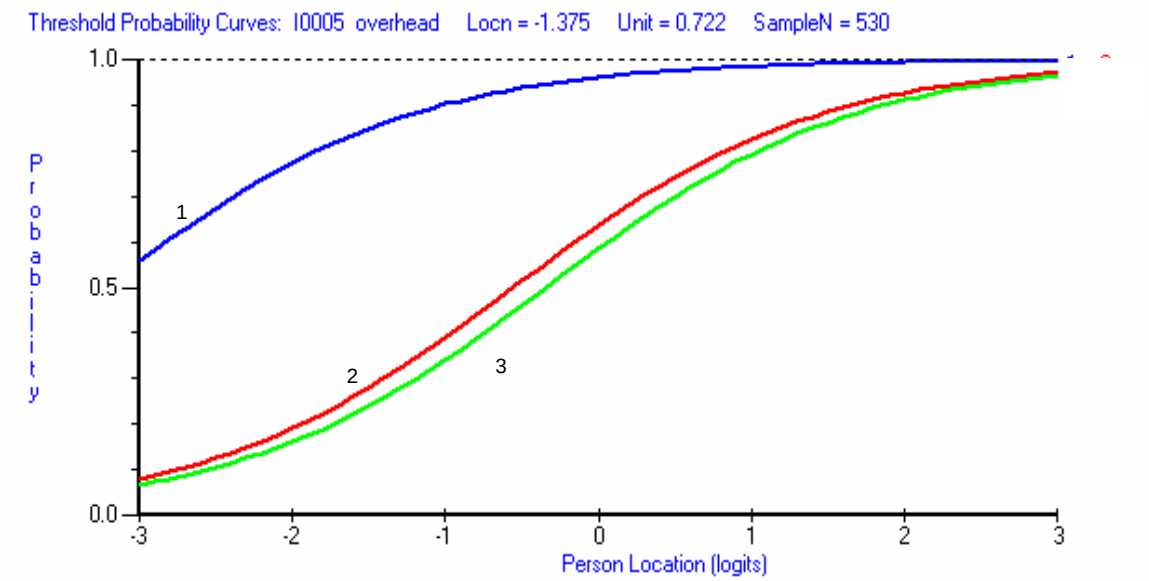

Figure 2c: An example of poorly spread item levels

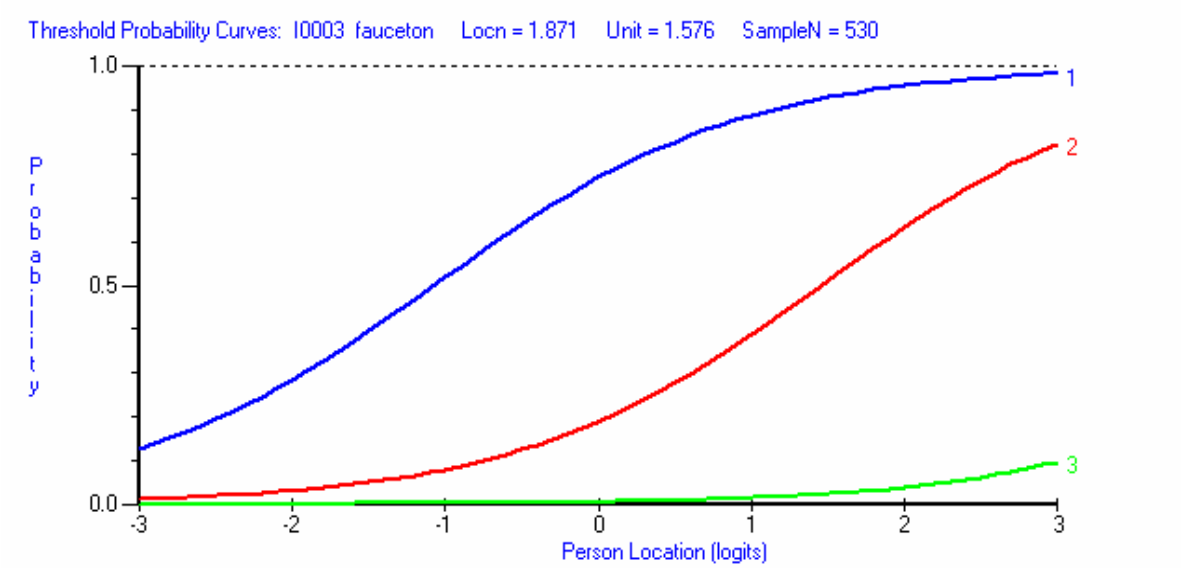

Figure 2d: An example of a level lying close to 0% probability

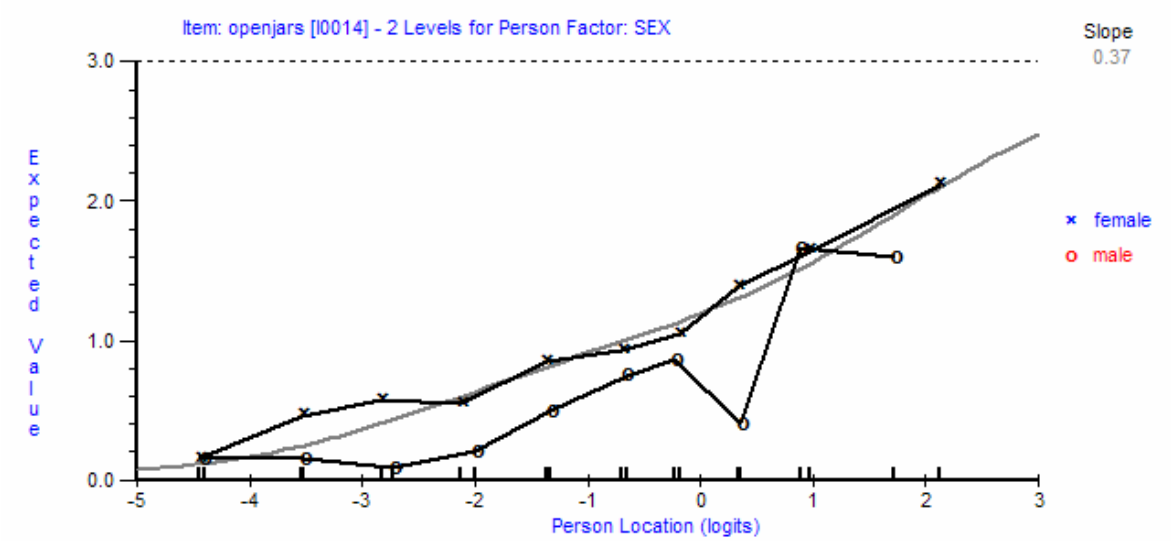

Figure 3: Item characteristic curve for *openjar* item stratified by sex

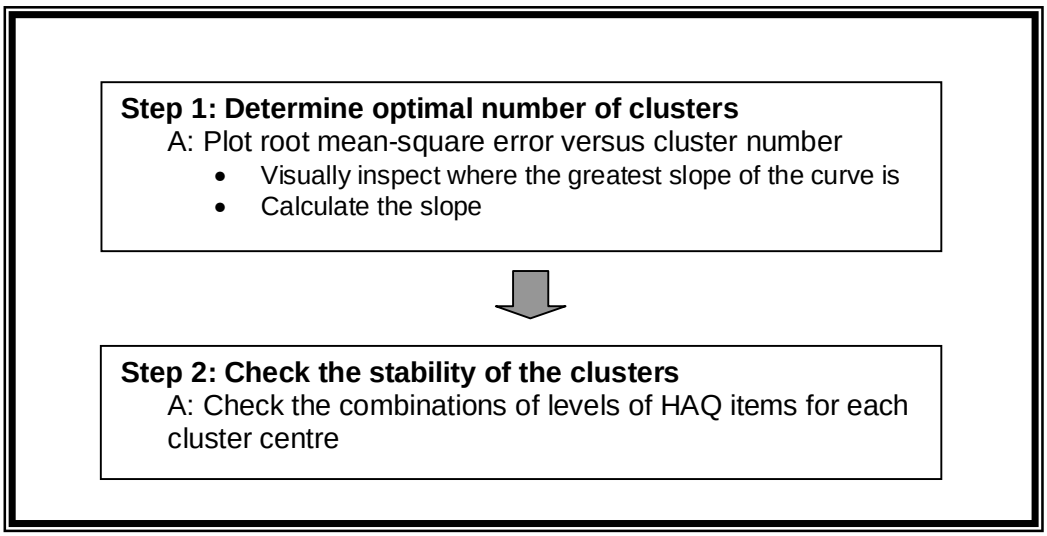

Figure 4: A schematic overview of *k*-means cluster analysis

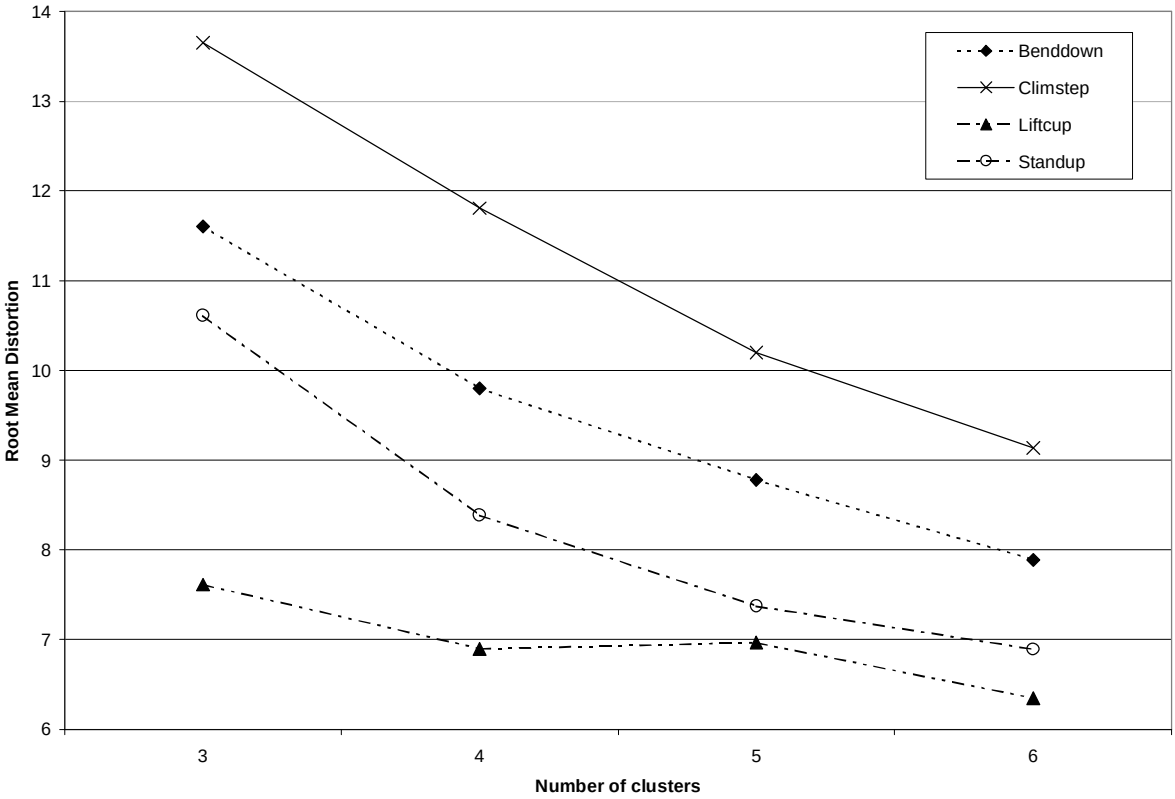

Figure 5: Root mean distortion versus number of clusters for the *k*-means method

Health State 1	Health State 2	Health State 3
You have **some** difficulty bending down to pick up clothes from the floor.	You have **some** difficulty bending down to pick up clothes from the floor.	You have **much** difficulty bending down to pick up clothes from the floor.
You have **some** difficulty climbing up 5 steps.	You have **much** difficulty climbing up 5 steps.	You are **unable** to climb up 5 steps.
You have **no** difficulty lifting a full cup or glass to your mouth.	You have **some** difficulty lifting a full cup or glass to your mouth.	You have **much** difficulty lifting a full cup or glass to your mouth.
You have **some** difficulty standing up from a straight and armless chair.	You have **much** difficulty standing up from a straight and armless chair.	You have **much** difficulty standing up from a straight and armless chair.
You have **mild** pain and discomfort.	You have **moderate** pain and discomfort.	You have **extreme** pain and discomfort.

Figure 6: Final health state descriptions