Qual Life Res (2008) 17:1183–1193
DOI 10.1007/s11136-008-9395-5

Deriving utility scores from the SF-36 health instrument using Rasch analysis

Graeme Hawthorne · Konstancja Densley · Julie F. Pallant · Duncan Mortimer · Leonie Segal

Accepted: 4 September 2008 / Published online: 30 September 2008

Abstract
Background Utility scores for use in cost-utility analysis may be imputed from the SF-36 health instrument using various techniques, typically regression analysis. This paper explored imputation using partial credit Rasch analysis.
Method Data from the Assessment of Quality of Life (AQoL) instrument validation study were re-analysed ($n = 996$ inpatients, outpatients and a community sample). For each AQoL item, factor analysis identified those SF-36 items forming a unidimensional scale. Rasch analysis located scale logit scores for these SF-36 items. The logit scores were used to assign AQoL item scores. The standard AQoL scoring algorithm was then applied to obtain the utility scores.
Results Many SF-36 items were limited predictors of AQoL items; some items from both instruments obtained disordered thresholds. All imputed scores were consistent with the AQoL model and fell within AQoL score boundaries. The explained variance between imputed and true AQoL scores was 61%.
Discussion Rasch-imputed mapping, unlike many regression-based algorithms, produced results consistent with the axioms of utility measurement, while the proportion of explained variance was similar to regression-based modelling. Item properties on both instruments implied that some items should be revised using Rasch analysis. The methods and results may be used by researchers needing to impute utility scores from SF-36 health scores.

Keywords AQoL · Cost-utility analysis · Health services research · SF-36 · Item response theory · Quality of life · Rasch analysis

Abbreviations

AQoL	Assessment of Quality of Life multi-attribute utility instrument
CUA	Cost-utility analysis
DIF	Differential item functioning
EFA	Exploratory factor analysis
ICC	Intra-class correlation
MAU	Multi-attribute utility
QALY	Quality-adjusted life year
SF-36	Short Form-36 health survey

Introduction

Decisions concerning the allocation of available health care resources require judgements which, ideally, are informed

G. Hawthorne (✉)
Department of Psychiatry, Level 1 North, Main Building, Royal Melbourne Hospital, Grattan Street, Parkville, Victoria 3050, Australia
e-mail: graemeeh@unimelb.edu.au

G. Hawthorne · K. Densley
Department of Psychiatry, Faculty of Medicine, Dentistry and Health Sciences, The University of Melbourne, Melbourne, Australia

J. F. Pallant
School of Rural Health, Faculty of Medicine, Dentistry and Health Sciences, The University of Melbourne, Shepparton, Australia

D. Mortimer
Centre for Health Economics, Faculty of Business and Economics, Monash University, Melbourne, Australia

L. Segal
Division of Health Sciences, University of South Australia, Adelaide, Australia

by evidence concerning the comparative performance of alternative service or intervention options. Often, however, findings from economic evaluations may quantify health gains in very different ways, making comparisons difficult. Cost-utility analysis (CUA) is increasingly used as a way of overcoming this difficulty because it provides a common outcome metric, the quality-adjusted life year (QALY), which enables the calculation of cost-per-QALY ratios for use in economic evaluation. The most common method of obtaining utilities is through the administration of a multi-attribute utility (MAU) instrument [1]. Where repeated observations are available, the utilities (preferences for health states) obtained from MAUs can be used to compute QALYs.

Many studies, however, have not collected utilities but have quantified health gains using descriptive measures of health status (typically the Short Form-36 [SF-36] [2]), thus restricting the availability of QALYs for comparative purposes. Consequently there have been several attempts at mapping health status scores to utility scores. Generally researchers have used one of two techniques. Most have used regression to model SF-36 to utility scores [3–8]. In contrast Brazier et al. developed a direct algorithm for a utility measure based on selected SF-36 items [9, 10]. Mappings or cross-walks allow utilities to be recovered where only descriptive measures of health status have been collected; cost-effectiveness to be expressed in cost per QALY terms rather than as a cost per point improvement on the main clinical outcome. The Rasch partial credit model considered here provides a third alternative for imputing utilities or QALY weights from descriptive measures of health status.

A substantial literature has now been accumulated concerning the methods, application and validity of techniques for imputing or predicting QALY weights from descriptive measures [11]. Recently, this literature has started to address some of the shortcomings of the relatively simple regression-based approaches that have dominated the literature and the capacity of such methods to deal with discontinuity and non-normality in the data. For example, the derivation of regression-based methods using subscale- or scale-level data on a descriptive measure such as the SF-36 has the effect of imposing restrictions on the relationship between the information contained in the descriptive system measure and utility scores. Regression using item scores will entail fewer restrictions. The results reported by Mortimer et al. [6] did not, however, find that algorithms for converting SF-36 item scores into AQoL utilities resulted in a significantly lower magnitude of error than scale- or subscale-based algorithms when predicting between-group differences.

However, even the use of item scores may have the effect of imposing restrictions on functional form and obscuring discontinuities in the data that might arise from disordered item responses (a disordered response is one that has a higher probability of being selected by respondents than is warranted by their underlying health state, e.g. if persons in 'good' health endorse a 'fair' health state, then the categories 'good' and 'fair' health may be disordered). For this reason, Lundberg et al. [12] and Brazier et al. [13] derived regression-based algorithms using response-level data, with each level of each item entered as a categorical variable, thereby, avoiding the inappropriate imposition of ordinal or interval properties. Similarly, Gray et al. [14] employed multinomial logistic regression to directly map to response categories on the preference-based target measure. The use of response-level rather than scale- or subscale-level data produced only modest improvements in the predictive power [13]. Quantifying any loss of predictive validity associated with the sort of restrictions on function form that arise from the use of scale-, subscale- or item-level data is of particular interest because scale- or subscale-based mappings generally have a wider application than item-based mappings (due to the fact that scale- or subscale-level data [but not item-level data] are commonly available from published studies).

Rasch modelling may provide an alternative means of imputing utility from health scores with alternative restrictions on function form. The model is a one-parameter model in the sense that it meets Thurstone's scaling requirements (that the measurement scale used must be independent of the object of measurement) [15, 16]. The Rasch model uses probabilistic models consistent with a probabilistic interpretation of the axioms of Guttman scaling [17]. Guttman scales are particularly useful in health research because they place people in order on a unidimensional scale. Diagnostic and screening tests require standard rules that enable the classification of people into the correct population with minimum levels of misclassification; e.g. healthy/diseased, benign/malignant, excellent health/good health/fair health/poor health, smoker/occasional smoker/non-smoker [18].

The original Rasch dichotomous model [19] specifies that the probability of item endorsement is a function of two different parameters, which are the underlying ability of the respondent and their expected item response. Masters generalized this to the partial credit Rasch model (hereafter, Rasch analysis) for use with items with multiple response categories [20], thus:

$$\pi_{xni} = \frac{\exp \sum_{j=0}^{x} \left(\beta_n - \delta_{ij}\right)}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^{k} \left(\beta_n - \delta_{ij}\right)} \quad (1)$$

where π_{xni} is the probability of a respondent, n, endorsing an item, i, at a given level, x, as a function of the respondent's ability, β_n, and the difficulty of the item m-steps, δ,

where m represents successive thresholds between the item response categories and j is the number of response levels. Thus, δ_{ij} is the difficulty of the probability threshold between item i response categories $j - 1$, where there are $mi + 1$ categories. δ_{i0} is 0 by convention. As reported by Masters [20], the equation permits the elimination of the respondent's ability from item estimations, thus, creating two internal parameters representing the respondent's ability and the item difficulty.

The relationship between these parameters enables the researcher to judge how well an item is performing. Figure 1 illustrates this with respect to item #4 from the AQoL instrument [21]. Along the bottom (x-axis) is a person's parameter or ability to respond to the item, where this refers to the latent trait of interest within the individual—in the current context a person's intrinsic health status. Up the side (y-axis) is the expected response value, based on the available probability thresholds (if there are four response levels, then there are three available thresholds between levels 1/2 and 2/3 and 3/4). The resulting empirical curve showing the relationship between the two parameters is the item characteristic curve, which plots the probability of a person with a logit ability selecting an expected response level.

For example, when a person selects an item response category the probability of endorsing that category in preference to the next category can be calculated. Response category thresholds refer to the point where the probability of endorsing one response category is equal to the probability of endorsing the next category (in Fig. 1, the horizontal lines represent equi-interval expected or probability thresholds). Good fitting models are where there is a graded monotonic relationship between the person ability and the item response categories such that persons with low abilities (e.g. being unable to wash or toilet themselves) endorse low response categories (e.g. unable to carry out personal care tasks). Two important function form assumptions implicit in Rasch analysis are those of invariance (that item properties exist independently of the respondent) and local independence (that where the abilities influencing responses are held constant the responses to items are assumed to be independent of each other).

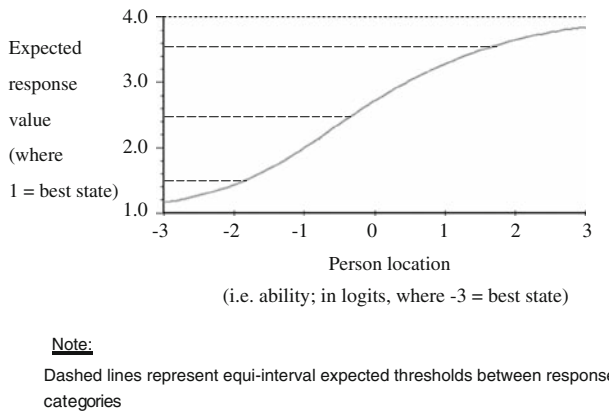

Fig. 1 Item characteristic curve for the Assessment of Quality of Life (AQoL) item #4, "Personal care"

Rasch analysis, then, may provide a method of controlling for variation in respondent characteristics, such as in a biassed unhealthy sample (where it might be expected that most people will report poor health). Importantly, it is suited to the situation where there is ordinal data or where the data are non-normal [20, 22, 23]. The SF-36 instrument and all utility measures use item response ordinal scales producing non-normal data distributions.

In the current context, Rasch analysis offers three methods of assessing the properties of items: (a) the ability to examine items' response scale performance, thus making sure that the response levels within an item are discriminating as expected (threshold analysis); (b) the opportunity to observe if known groups differ in their interpretation of an item (differential item functioning [DIF]); and (c) the capacity to identify the location of items on a common logit (logistic) scale thus enabling test linking [24]. It is this last property that is of particular interest for imputing utility from health.

At this present time, no studies have been published using Rasch analysis (or any other item response theory model) to examine the relationship between health function and scores suitable for use in cost-utility analysis. Of the several possible reasons for this situation, three may be particularly important. In general, utility instruments were developed by health economists whose interests were not in measurement theory but in the problem of valuing health outcomes for inclusion in economic evaluation. The second is that scales on which scores are represented have different meanings, thus, a score of, say, 0.5 on a utility scale does not have the same meaning as a score of 50, which is a percentage score, on one of the SF-36 scales. The third is that Rasch analysis demands that the items contributing to a scale are unidimensional, whereas MAU instruments attempt to cover the whole of life. By definition, they are heterogeneous (the requirement for unidimensionality, however, may be overcome by examining unidimensional sub-scales within MAU instruments).

The use of Rasch analysis may therefore hold the promise of providing an alternative way of mapping health states to utility scores. This has not been previously reported, although Rasch analysis has been used to examine the response levels within items [25], to select items for inclusion in a descriptive system [26], to examine the performance of a descriptive system cross-culturally [27] or of different response levels within a MAU instrument [28].

Somewhat differently, this study investigated the use of Rasch analysis to derive an algorithm enabling the computation of AQoL utility scores from the SF-36 health status measure. We expected that a Rasch analysis would provide imputed AQoL item scores which could be used in the standard AQoL scoring algorithm.

Methods

Data from the Victorian validation study of the AQoL were re-analysed [1, 29].

Participants

Participants were a stratified sample of Victorian residents, selected to cover a very broad range of health conditions, from those who were healthy through to those who were terminally ill. The strata were: (a) randomly selected community members weighted by socio-economic status to achieve representativeness of the Australian population; (b) outpatients attending two of Melbourne's largest public hospitals (the method used was random sampling within selected time frames); and (c) inpatients from three Melbourne hospitals (purposive sampling was used within wards based on the severity of condition).

Measures

The SF-36 version 1 [2] was administered to participants. This is a health status instrument comprising 36 items covering physical functioning, role physical, bodily pain, general health, vitality, social functioning, role emotion and mental health. For this study, only individual items were used. Scale scores were not computed or used.

The Assessment of Quality of Life (AQoL) instrument is an MAU instrument [1, 21]. There are 12 items forming four dimensions: independent living, social relationships, physical senses and psychological well-being. The standard utility algorithm for scoring it is as follows. Item responses from an individual are replaced with community preference values, where those values were obtained from a representative sample of the population using time trade-off (TTO). A multiplicative model is then used to combine these new scores into the four-dimension scores, again weighted by community preferences obtained through TTO. The resulting four-dimension scores are then combined into a single score, which is re-weighted (again from a community sample based on TTOs) and presented as a utility score on a life–death scale, where the endpoints are −0.04 (worse than death HRQoL equivalent state), 0.00 (death equivalent HRQoL state) to 1.00 (best HRQoL).

Statistical analyses

A three-part analysis procedure was followed.

The first part used exploratory factor analysis (EFA) to identify for each AQoL item the associated SF-36 items. The purpose was to identify, for each AQoL item, those SF-36 items forming a unidimensional scale with it being suitable for Rasch analysis [30, 31]. The reason for the use of EFA to do this was the requirement of Rasch analysis that items form a unidimensional scale. The robustness of factor analysis under non-normality has been previously demonstrated [32]. Prior to analysis, the data were checked to test for adequate sample size, the presence of outliers, inter-item correlation, the number of variables and sampling adequacy [33–36]. The procedure itself involved an iterative series of principal component and rotated analyses aimed at extracting the maximum variance from the dataset with the minimum number of unrelated components, assuming that for each item all of the variance could be ascribed to a common underlying factor. Following the principal component analysis, the factors were rotated using oblimin (with delta = 0.00) and then varimax rotations. To determine the number of factors to be retained, only factors with an eigenvalue of 1.0 or more were retained for further investigation. Iterative forced analyses [37] of each AQoL item with 35 SF-36 items showed that the best models, overall, were those with three-factor solutions. Each AQoL item was entered and SF-36 items iteratively entered and withdrawn until those SF-36 items sharing the same vector as the AQoL item of interest were identified.

Following completion of the EFA, Rasch analyses were constructed where each model was based on sampling those with variance among the items of interest (i.e. all those with ceiling scores among the items of interest for each particular model were excluded). Generally, this provided sample sizes of between 800 and 970 cases. There was one instance where the sample size was $n = 250$, which was half of the sample size reported by Stone and Yumoto [38] for infit and outfit error convergence, but which was above the minimum sample size requirements for Rasch analysis with restricted targetting [39], which was the case in this study. A separate model was constructed for each AQoL item and its associated SF-36 items; thus, 12 iterative models were constructed—one for each AQoL item. Item probability thresholds were then examined and, where necessary, items were recoded to remove disordered thresholds. The procedure for combining categories was to accept the recoding option which provided the greatest differentiation in the expected threshold probability curves between item levels. Figure 2 illustrates this for the AQoL item #4, "Personal care." As shown in Fig. 2a, response category 2 (needs occasional

help with person care) is dominated by response category 1 (needs no help); after combining these two categories to remove the disordered threshold the probability curves presented in Fig. 2b were obtained.

Poor or redundant items from each resulting scale were then iteratively removed using Rasch analysis. Items were removed if the fit residual exceeded 2.50 [40] (which approximates $P < 0.01$, whereas a value of 2.00 is an approximation of $P < 0.05$, which is a less stringent unidimensionality requirement) and removal improved the overall unidimensional model fit statistics. The reason for accepting this was that fit statistics are a function of sample size and test length, as well as unidimensionality; with small samples minor discrepancies will result in items being (by chance) classified as misfitting and, therefore, excluded even though they may function perfectly well [38, 41]. We wished to avoid this situation through being over-inclusive rather than exclusive. The process was iteratively continued until the item and person model fit statistics deteriorated following the removal of an item.

The models were then re-run, excluding each AQoL item, to extract the logit scale scores for the SF-36 items associated with the AQoL item of interest. These logit scale scores were then plotted against the actual responses to the AQoL item and the cut-points on each logit scale defined. The mean logit scale score for each AQoL item response level was observed, as were the standard deviations around that logit mean. The standard deviations for each pair of consecutive AQoL item levels were then compared and, for each pair of standard deviations, the mean logit score computed; this then became the cutpoint for assigning the imputed AQoL item scores. For example, for the AQoL item #4, "Personal care," the final SF-36 items used to impute scores were items SF3A, SF3F, SF3G, SF3I and SF3J. Simple summation of these items provided a scale with scores from 5 (worst possible health) to 15 (best possible health). Based on the logic scale, the cutpoints for AQoL item #4 (value in brackets) were ≤5 (4), 6 (3), 7–10 (2) and ≥11 (1).

The standard AQoL scoring algorithm was then applied to the data for scoring the AQoL. The results were compared with directly obtained AQoL utility scores.

Data analysis was conducted using SPSS version 13.0 software for Windows [42] and RUMM2020 4.0 [43].

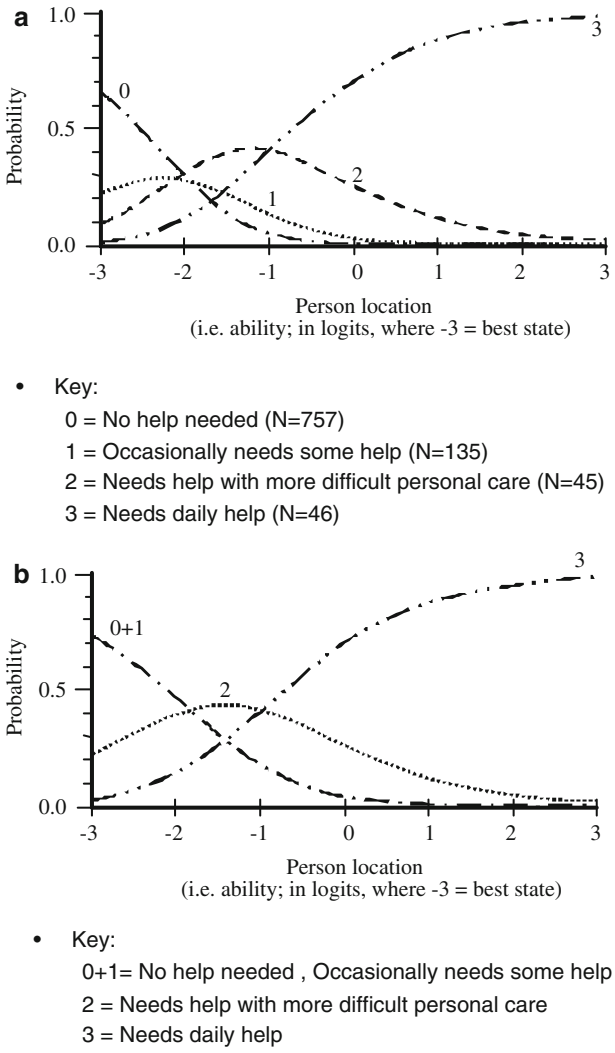

Fig. 2 **a** Unrecoded probability curves for AQoL item #4, "Personal care." **b** Recoded probability curves for AQoL item #4, "Personal care"

Results

The sample comprised 996 adults. The response rates were 58% ($n = 396$) for the community sample, 43% ($n = 334$) for outpatients and 68% ($n = 266$) for inpatients. The community sample comprised 46% of the study population, outpatients 38% and inpatients 16%. Fifty percent of respondents were male, the mean age was 52 (standard deviation [SD] = 18) years, 75% were Australian-born and 89% reported that English was their first language. Sixty percent of the sample were partnered (married, de facto), 18% were single, 11% were separated or divorced and 12% widowed. For education attainment, 64% had completed either primary or high school, 13% held a trade certificate and 23% a university degree. Forty-seven percent were in the labour force (working or studying), 34% were retired, 10% were homemakers and 9% were unemployed or reported other non-working activities.

The results of the EFA exploring the relationship between each AQoL item and its associated SF-36 items loading on the principal component are shown in Table 1. The table shows the percentage of explained variance for each model, the factor loading of each AQoL item and the associated SF-36 items. As mentioned in the Methods

section, a variety of forced models were explored and three-factor models were found to provide the best fit to the data. For example, the initial solution for the AQoL item #5 yielded five components explaining 67% of the variance. The fifth factor, however, was poorly represented, so a forced four-factor solution was examined. This explained 64% of the variance, but not all of the communalities met the criteria of >0.30, nor did all variables load substantially on just one component (>0.30). When a three-factor solution was tested, the proportion of explained variance was 61% and all communalities were >0.30, the pattern coefficient was >0.40 and all variables loaded substantially on only one component (hence meeting the need for unidimensionality). The other 11 AQoL items contributing to the AQoL utility score were subjected to similar analysis.

The loadings roughly followed item content. The AQoL items on the "Independent living" dimension (#A4, #A5 and #A6 in Table 2) were associated with the SF-36 physical items and the AQoL "Social relationships" items (#A7, #A8 and #A9) were associated with the SF-36 items assessing energy to do things or restrictions on doing things. The "Physical sense" AQoL items (#A10, #A11 and #A12) varied more. The AQoL items assessing vision (#A10) and communication (#A12) were associated with SF-36 items concerned with the energy to do things or restrictions on doing things, whereas the AQoL item covering hearing (#A11) was associated with the SF-36 physical items. It is possible that this relationship reflects hearing loss in older adults who may also suffer increased physical restrictions. The AQoL "Psychological" dimension items (#A13, #A14 and #A15) were associated with SF-36 items concerned with the energy to do things or restrictions on doing things. #A15 (pain) was also associated with the SF-36 items covering pain.

A feature of the table is the limited factor loadings (<0.30) of five of the AQoL items: items assessing family role (#A9), vision (#A10), hearing (#A11), communication (#A12) and sleeping (#A13). The reason for this is that there are no SF-36 items assessing these aspects of people's lives.

After recoding of the item response scales to remove disordered thresholds where they occurred, each of the EFA models (the columns in Table 1) were then examined using Rasch analyses. Table 2 provides an example of the modelling for the AQoL item #4 ("Do I need any help looking after myself?") and shows the various item fit statistics which were obtained. As shown, several of the SF-36 items which loaded on the EFA in Table 1 on the principal vector as the AQoL item #4 were discarded (iteratively) because they failed to meet the Rasch analysis criteria outlined in the Methods section, including SF-36 items #1, #3b, #3c, #3d, #3e and #3 h. The final model shown here was satisfactory, as indicated by the various summary statistics in the table. That the SF-36 items in Table 2 have negative fit residuals is indicative of local dependency; i.e. that these items are dependent upon one another (e.g. consider SF-36 items SF3g and SF3i. These two items are dependent upon each other—if a person cannot walk 100 metres, he/she certainly cannot walk 1 km. This lack of independence means these items 'overfit' (i.e. are more predictable) rather than misfit (as would be indicated by a positive fit residual)).

Table 3 summarizes the results for all AQoL items. Of the 36 items comprising the SF-36 instrument, 25 were used across the different models. The number of items in the models varied from 4 to 10. The most commonly used items from the SF-36 were items #9D, #9E and #9I assessing having sufficient energy, being calm and peaceful and being tired, respectively. The fit statistics were all satisfactory in that for all models the item and person (ability) fit statistics fell within the acceptable range. Similarly, there were no significant interactions between item and person (ability) fit, indicating good separation of these.

As described in the Methods section, each model was then re-run after excluding the AQoL item, the SF-36 items scale score estimator extracted and plotted against the relevant AQoL item, cut-points on the estimator established as described and the AQoL item responses imputed. Table 3 shows the cutpoints on the summed scaled SF-36 items for each AQoL item.

The standard AQoL scoring algorithm was then applied to the imputed item data for scoring the AQoL. Following imputation, the relationship between the SF-36 logit estimator imputed AQoL utility scores and the original AQoL utility scores was examined. The Spearman correlation was $r_s = 0.77$ ($n = 867$, $P < 0.01$) and the linear $r^2 = 0.61$. Figure 3 shows a scatterplot of the two AQoL scores with a linear trendline and 95% confidence intervals.

Discussion

Because utility data have often not been collected in studies, researchers have sought to model utility scores on MAU instruments using either regression or the direct revaluation of SF-36 health states [11]. Mappings obtained from direct revaluation of the descriptive measure typically condense the descriptive measure before valuation. Predicted utilities obtained from direct revaluation can therefore only ever be a subset of the set of health states covered by the original descriptive measure, and this might manifest as floor or ceiling effects. In this context, it is worth noting that the direct revaluation of the SF-36 to produce the SF-6D produces a floor effect at around 0.30. SF-6D utilities are therefore unsuitable for estimating QALY weights in severely ill populations or for severe

Table 1 Factor loadings and the percentage of variance explained by three-factor models with AQoL items

	AQoL items											
	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15
Percentage of variance explained (%)	58	59	59	58	58	59	58	58	58	58	59	50
AQoL item loading	−0.536	0.649	−0.638	0.428	0.621	0.267	0.263	0.257	0.251	0.359	0.813	−0.367
SF 1	−0.367	0.391	−0.373	0.438	0.400	0.452	0.450	0.397	0.447	0.465	0.394	
SF 3a	−0.556	0.579	−0.549					0.595				
SF 3b	−0.767	0.777	−0.760					0.774				
SF 3c	−0.804	0.802	−0.798					0.795				
SF 3d	−0.806	0.812	−0.805					0.821				
SF 3e	−0.890	0.884	−0.890					0.880				
SF 3f	−0.789	0.784	−0.783					0.788				
SF 3g	−0.867	0.873	−0.864					0.874				
SF 3h	−0.926	0.916	−0.924					0.912				
SF 3i	−0.878	0.855	−0.873					0.845				
SF 3j	−0.703	0.663	−0.681					0.643				
SF 4a												−0.873
SF 4b												−0.853
SF 4c												−0.810
SF 4d												−0.777
SF 5a				0.308	0.334		0.301		0.318		0.385	−0.527
SF 5b				0.330	0.354		0.320		0.337	0.304	0.406	−0.534
SF 5c				0.358	0.371		0.350		0.369	0.337	0.415	−0.381
SF 6												−0.516
SF 7												−0.601
SF 8												−0.645
SF 9a				0.501	0.479	0.508	0.507		0.515	0.516	0.495	−0.349
SF 9b				0.758	0.773	0.759	0.757		0.761	0.756	0.795	
SF 9c				0.753	0.766	0.753	0.751		0.764	0.750	0.798	
SF 9d				0.719	0.696	0.724	0.720		0.721	0.729	0.723	
SF 9e				0.461	0.432	0.475	0.472		0.474	0.487	0.434	
SF 9f				0.798	0.804	0.805	0.797		0.806	0.803	0.831	
SF 9g				0.615	0.608	0.628	0.621		0.624	0.631	0.618	
SF 9h				0.777	0.763	0.780	0.770		0.775	0.779	0.781	
SF 9i				0.586	0.565	0.596	0.593		0.596	0.599	0.570	
SF 10				0.303	0.302		0.303		0.314	0.300	0.333	−0.438
SF 11a				0.538	0.499	0.552	0.547		0.545	0.557	0.505	
SF 11b				0.435	0.392	0.454	0.448		0.441	0.464	0.383	
SF 11c				0.400	0.364	0.410	0.409		0.398	0.422	0.348	
SF 11d				0.478	0.436	0.499	0.496		0.489	0.510	0.438	

Factor loadings <0.30 are suppressed from the table

health states [9, 10]. By way of comparison, the imputed utilities reported here cover the entire −0.04 to 1.00 range of the AQoL scale.

Imputed utilities obtained from Rasch analysis might also offer advantages over many regression-based mappings. Regression-based models published to date have explained between 51 and 70% of the variance in observed utility scores [3, 5–8, 11]. Generally, these results should be regarded as being very good, given that they may be sample-dependent and that there is sufficient evidence suggesting that quality of life is determined by salutogenesis rather than pathogenesis, thus giving rise to adaptation. As health declines other aspects of life, such as relationships, may become more important.

Table 2 Final partial credit Rasch analysis of the recoded AQoL item #4 with SF-36 items: individual item fit statistics

		Location (a)	SE (b)	Fit residual (c)	df (d)	χ^2 (e)	df (d)	*P*-value (f)
A4	Need help looking after self	−0.95	0.13	1.41	202.08	8.15	7	0.32
SF3A	Vigorous activities	2.56	0.12	−0.05	201.25	4.67	7	0.70
SF3F	Bending, kneeling or stooping	0.52	0.12	−1.10	201.25	11.37	7	0.12
SF3G	Walking >1 km	0.65	0.12	−1.45	201.25	9.11	7	0.24
SF3I	Walking 100 m	−1.16	0.14	−2.35	202.08	9.43	7	0.22
SF3J	Bathing or dressing self	−1.62	0.15	−0.67	202.08	6.57	7	0.47

Summary statistics: mean item fit residual: −0.70 (SD = 1.29); mean person fit residual: −0.47 (SD = 1.07); person separation index (i.e. Rasch test reliability analysis): 0.86; $\chi^2 = 49.32$, df = 42, $P = 0.20$

a = in logits

b = standard error

c = item fit residual (i.e. the relationship of the observed score with the expected score for each person–item relationship)

d = degrees of freedom

e = Chi-square

f = probability value

* = misfitting items, fit residuals >|2.50|, equivalent to $P < 0.01$

Rasch analysis may assist with this problem because it separates persons from their scores and, as used in this study, uses probabilistic models of Guttman scaling, enabling the imputation of item values between instruments. Rasch analysis also imposes a different set of assumptions on function form than those implicit in regression-based imputations, as discussed in the Methods section. Thus, in deciding how to impute utility scores in any particular study, there may be a trade-off between the assumptions of regression against those of Rasch analysis. The assumptions are that the probability of a response is a function of the respondent's underlying state, that items possess local independence and that items form a unidimensional scale. The findings presented in the tables suggest that these assumptions were met: the factor analyses ensured unidimensionality, which in turn is an accepted test for local independence [44].

In this study, many of the SF-36 items were poor predictors of AQoL items (Table 1). The reason for this may be that the AQoL is concerned with assessing a person's quality of life from the perspective of handicap arising from an intrinsic health condition [21]. In contrast, the SF-36 measures health function. Indeed Ware et al. in the SF-36 Version 1 Manual and Interpretation Guide deny that the SF-36 measures quality of life at all [2].

This was an exploratory study which used some of the techniques from Rasch analysis test-equating [24, 44–46], but in the health-related quality of life field. To our knowledge it is the first study to attempt to do this. Possible reasons for this are that predicting utility scores from health scores does not meet the requirement for test-equating, that using Rasch analysis for this is more resource-intensive than regression-modelling and because it places higher demands on available datasets because it requires the use of individual data.

The theoretical consistency of imputed AQoL utility scores with the properties of the AQoL scale was superior to that of the predicted AQoL utility scores reported in our previous paper, where we used sophisticated regression modelling with the same dataset [6]. In contrast to predicted AQoL scores from our regression models, all imputed AQoL scores from the Rasch analyses fell within the bounds of the AQoL scale and relate to health states that exist in the AQoL descriptive system. Where a researcher has a database with raw SF-36 item scores, Table 3 can be used to impute AQoL data through summing the scores of the relevant SF-36 items and then imputing each AQoL item score using the cutpoints provided. The resulting data can then be used with the normal AQoL utility scoring algorithm.

The predictive validity of the Rasch analysis imputation model was, however, modest. As shown in Fig. 3, the explained variation between the imputed AQoL utility scores and the observed AQoL scores was 61%, which was consistent with the range reported in the literature for regression-based models imputing utility scores from the SF-36 [3, 5, 7, 8]. However, it was considerably lower than that reported for our regression models using the same dataset [6]. There are several reasons which may help to explain this modest finding.

As already noted, there is no necessary reason that SF-36 items should predict AQoL items because these two instruments are concerned with different aspects of people's lives. It should also be borne in mind that neither the SF-36 nor AQoL were designed to be used at the individual level. The implication is that the accuracy of items will be

Table 3 Rasch analysis models, statistics and scoring for all AQoL items

AQoL item		SF-36 items*	Rasch analysis model statistics						
No.	Content		Mean item fit residual (SD)	Mean person fit residual (SD)	PSI (a)	Item–person interaction			Cutpoints on summed scales from SF-36 items (e)
						χ^2 (b)	df (c)	*P*-value (d)	
A4	Self-care	3A, 3F, 3G, 3I, 3J	−0.70 (1.29)	−0.47 (1.07)	0.86	49.32	42	0.20	≤5 (4)/6 (3)/7–10 (2)/≥11 (1)
A5	Daily tasks	3A, 3F, 3I, 3J	−0.83 (0.52)	−0.46 (0.67)	0.84	41.42	30	0.08	≤5 (4)/6–7 (3)/8–9 (2)/≥10 (1)
A6	Mobility	3B, 3C, 3G, 3I	0.18 (0.31)	−0.26 (0.80)	0.91	45.75	35	0.11	≤4 (4)/5 (3)/6 (2)/≥7 (1)
A7	Intimacy	5B, 9C, 9G, 9I, 1*, 9E*, 9D*, 9F	−0.11 (1.22)	−0.32 (1.11)	0.88	76.25	81	0.63	≤19 (4)/20–22 (3)/23–31 (2)/≥32 (1)
A8	Friendship	5A, 5B, 9C, 9I, 11A, 9E*, 9D*, 9H*, 11D*	−0.09 (1.26)	−0.28 (1.06)	0.85	99.59	90	0.23	≤18 (4)/19–23 (3)/24–32 (2)/≥33 (1)
A9	Family role	9C, 9F, 9G, 9I, 9A*, 9E*, 11D*	−0.14 (0.65)	−0.38 (1.14)	0.88	82.06	72	0.20	≤10 (4)/11–15 (3)/16–27 (2)/≥28 (1)
A10	Vision	5B, 5C, 9C, 9I, 9D*, 9H*, 11D*	−0.48 (1.39)	−0.35 (1.01)	0.79	89.17	72	0.08	≤6 (4)/7–10 (3)/11–21 (2)/≥22 (1)
A11	Hearing	9C, 9F, 9G, 9I, 1*, 9E*, 9D*, 9H*, 11D*	0.04 (1.26)	−0.37 (1.22)	0.87	100.61	90	0.21	≤7 (4)/8–9 (3)/10–19 (2)/≥20 (1)
A12	Communication	5A, 5B, 9C, 9F, 9G, 9I, 9A*, 9E*, 9D*, 11D*	−0.52 (1.01)	−0.35 (1.15)	0.90	106.24	108	0.53	≤9 (4)/10–16 (3)/17–27 (2)/≥28 (1)
A13	Sleeping	9C, 9F, 9I, 1*, 9E*, 9D*	0.19 (1.55)	−0.40 (1.31)	0.83	75.49	63	0.13	≤17 (4)/18–24 (3)/25–28 (2)/≥29 (1)
A14	Anxiety	5A, 5C, 9C, 9F, 9A*, 9E*, 9D*, 9H*	0.07 (0.82)	−0.40 (1.12)	0.90	90.35	73	0.08	≤16 (4)/17–23 (3)/24–31 (2)/≥32 (1)
A15	Pain	4A, 5A, 5B, 5C, 6*, 7*	−0.62 (0.99)	−0.35 (0.90)	0.83	75.81	58	0.06	≤7 (4)/8–9 (3)/10–15 (2)/≥16 (1)

* Reversed item before summation

a = person separation index (i.e. Rasch test reliability analysis)

b = Chi-square

c = degrees of freedom

d = probability

e = cutpoints on summed scales of SF-36 items, AQoL item values in parentheses

limited. This is shown by the item and instrument score standard deviations, which are typically 20% of the scale range. The consequence is that there will be substantial misclassification at the item imputation level. A related reason may be that selected items from both instruments suffered poor response distributions (in that there was considerable sparse data) and disordered thresholds. For these reasons, many of the items were recoded prior to analysis; the analyses should be regarded, therefore, as rather imprecise estimates. A third reason may be the rather arbitrary nature of the cut-points we used, which were based on the standard deviations around item response levels, although we note that no better procedure has been suggested in the literature [40].

In summary, there is good reason to suspect that the findings from this study are a function of the limitations of the different purposes of the two instruments, the items themselves and the procedure used. The general conclusion is that, although we have demonstrated that the Rasch analysis procedure can be used to map health status to utility and overcome some of the limitations of regression-based models (which systematically ignore issues of meaning and item measurement), its use exposes the psychometric properties and meanings of items to a high level of scrutiny. There may be a trade-off; regression-based imputations assume relationships that have the advantage of obscuring or minimising item difficulties, whereas Rasch analysis requires instruments with similar meanings and with items that have better properties than those currently available, at least in the SF-36 and AQoL. This conclusion is consistent with Teresi's argument that the mathematics of measurement are now well ahead of the quality of manifest items [46]. It is also possible that the procedures outlined by Reise et al. [47] in relation to

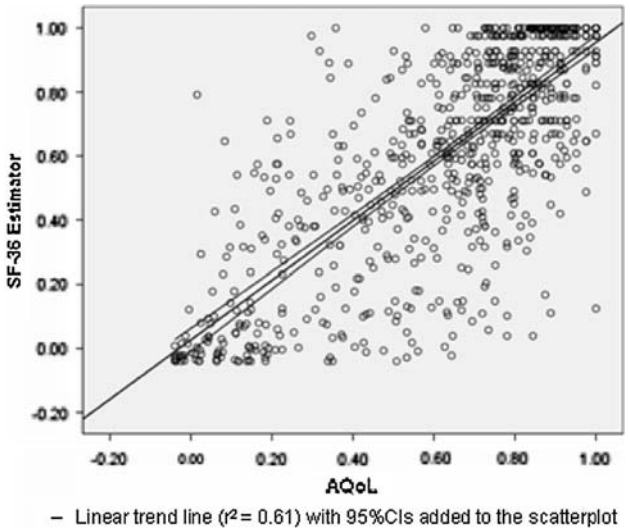

Fig. 3 Scatter plot of AQoL with SF-36 estimator imputed AQoL scores

bifactor models may yield better results than the unidimensional models used in this study.

The findings suggest that further refinements using Rasch analysis may produce better results. The study needs to be replicated in other samples and with other instruments in order to investigate whether the limitations outlined above hold true across different utility instruments or if they are a function of the Rasch analysis itself in this context. Importantly, as discussed above, mapping from health to utility as currently practised does not fully meet the conventional requirements for test-equating.

If the findings of this study are a function of item properties, then there is a prima facie case for the revision of health status and HRQoL instrument items using modern test theory. In the meantime, our findings suggest that partial credit Rasch-imputed mapping from the SF-36 to the AQoL produces results that are, generally, as good as those reported in the literature from regression-based modelling. The methods and results from this study may be used by researchers wherever SF-36 items scores are available and there is a need to impute utility scores.

Acknowledgements The research reported in this paper was supported by an Australian National Health and Medical Research Council (NHMRC) Project Grant, the Department of Psychiatry in the Faculty of Medicine, Dentistry and Health Sciences at The University of Melbourne and the Centre for Health Economics at Monash University. Much of this work was completed while Professor Leonie Segal was working at the Centre for Health Economics at Monash University prior to taking up her current position as the Professor of Health Economics, Division of Health Sciences, University of South Australia. The views expressed herein are the sole responsibility of the authors. The authors have no competing interests to declare.

References

1. Hawthorne, G., Richardson, J., & Day, N. A. (2001). A comparison of the Assessment of Quality of Life (AQoL) with four other generic utility instruments. *Annals of Medicine*, *33*, 358–370. doi:10.3109/07853890109002090.
2. Ware, J. E., Snow, K. K., Kosinski, M., & Gandek, B. (1993). *SF-36 health survey: manual and interpretation guide*. Boston, MA: The Health Institute, New England Medical Center.
3. Franks, P., Lubetkin, E. I., Gold, M. R., & Tancredi, D. J. (2003). Mapping the SF-12 to preference-based instruments: convergent validity in a low-income, minority population. *Medical Care*, *41*, 1277–1283. doi:10.1097/01.MLR.0000093480.58308.D8.
4. Franks, P., Lubetkin, E. I., Gold, M. R., Tancredi, D. J., & Jia, H. (2004). Mapping the SF-12 to the EuroQol EQ-5D Index in a national US sample. *Medical Decision Making*, *24*, 247–254. doi:10.1177/0272989X04265477.
5. Fryback, D. G., Lawrence, W. F., Martin, P. A., Klein, R., & Klein, B. E. (1997). Predicting Quality of Well-being scores from the SF-36: results from the Beaver Dam Health Outcomes Study. *Medical Decision Making*, *17*, 1–9. doi:10.1177/0272989X9701700101.
6. Mortimer, D., Segal, L., Hawthorne, G., & Harris, A. (2007). Item-based versus subscale-based mappings from the SF-36 to a preference-based quality of life measure. *Value in Health*, *10*, 398–407. doi:10.1111/j.1524-4733.2007.00194.x.
7. Nichol, M. B., Sengupta, N., & Globe, D. R. (2001). Evaluating quality-adjusted life years: estimation of the health utility index (HUI2) from the SF-36. *Medical Decision Making*, *21*, 105–112. doi:10.1177/02729890122062352.
8. Segal, L., Day, S. E., Chapman, A. B., & Osborne, R. H. (2004). Can we reduce disease burden from osteoarthritis? An evidence-based priority setting model. *The Medical Journal of Australia*, *180*, S11–S17.
9. Brazier, J., Roberts, J., & Deverill, M. (2002). The estimation of a preference-based measure of health from the SF-36. *Journal of Health Economics*, *21*, 271–292. doi:10.1016/S0167-6296(01)00130-8.
10. Brazier, J., Usherwood, T., Harper, R., & Thomas, K. (1998). Deriving a preference-based single index from the UK SF-36 Health Survey. *Journal of Clinical Epidemiology*, *51*, 1115–1128. doi:10.1016/S0895-4356(98)00103-6.
11. Mortimer, D., & Segal, L. (2008). Comparing the incomparable? A systematic review of competing techniques for converting descriptive measures of health status into QALY-weights. *Medical Decision Making*, *28*, 66–89. doi:10.1177/0272989X07309642.
12. Lundberg, L., Johannesson, M., Isacson, D. G., & Borgquist, L. (1999). The relationship between health-state utilities and the SF-12 in a general population. *Medical Decision Making*, *19*, 128–140. doi:10.1177/0272989X9901900203.
13. Brazier, J. E., Kolotkin, R. L., Crosby, R. D., & Williams, G. R. (2004). Estimating a preference-based single index for the Impact of Weight on Quality of Life-Lite (IWQOL-Lite) instrument from the SF-6D. *Value in Health*, *7*, 490–498. doi:10.1111/j.1524-4733.2004.74012.x.
14. Gray, A. M., Rivero-Arias, O., & Clarke, P. M. (2006). Estimating the association between SF-12 responses and EQ-5D utility values by response mapping. *Medical Decision Making*, *26*, 18–29. doi:10.1177/0272989X05284108.
15. Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, *33*, 529–554. doi:10.1086/214483.
16. Wright, B. D. (1989). Rasch model from Thurstone's scaling requirements. *Rasch Measurement Transactions*, *2*, 13–14.

17. Andrich, D. (1985). An elaboration of Guttman scaling with Rasch models for measurement. In N. Brandon-Tuma (Ed.), *Sociological methodology* (pp. 33–80). San Francisco, CA: Jossey-Bass.
18. Coste, J., & Pouchot, J. (2003). A grey zone for quantitative diagnostic and screening tests. *International Journal of Epidemiology*, *32*, 304–313. doi:10.1093/ije/dyg054.
19. Rasch, G. (1960). In B. Rasc (Ed.), *Probabilistic models for some intelligence and attainment tests* (p. 184). Copenhagen, Denmark: Danmarks Paedagogiske Institut.
20. Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174. doi:10.1007/BF02296272.
21. Hawthorne, G., Richardson, J., & Osborne, R. (1999). The Assessment of Quality of Life (AQoL) instrument: a psychometric measure of health-related quality of life. *Quality of Life Research*, *8*, 209–224. doi:10.1023/A:1008815005736.
22. Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561–573. doi:10.1007/BF02293814.
23. Mokken, R. J., & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement*, *6*, 417–430. doi:10.1177/014662168200600404.
24. Zhu, W. (1998). Test equating: what, why, how? *Research Quarterly for Exercise and Sport*, *69*, 11–23.
25. Dobrez, D., Cella, D., Pickard, A. S., Lai, J. S., & Nickolov, A. (2007). Estimation of patient preference-based utility weights from the functional assessment of cancer therapy—general. *Value in Health*, *10*, 266–272. doi:10.1111/j.1524-4733.2007.00181.x.
26. Brazier, J. E., & Roberts, J. (2004). The estimation of a preference-based measure of health from the SF-12. *Medical Care*, *42*, 851–859. doi:10.1097/01.mlr.0000135827.18610.0d.
27. Raczek, A. E., Ware, J. E., Bjorner, J. B., Gandek, B., Haley, S. M., Aaronson, N. K., et al. (1998). Comparison of Rasch and summated rating scales constructed from SF-36 physical functioning items in seven countries: results from the IQOLA Project. International Quality of Life Assessment. *Journal of Clinical Epidemiology*, *51*, 1203–1214. doi:10.1016/S0895-4356(98)00112-7.
28. Pickard, A. S., Kohlmann, T., Janssen, M. F., Bonsel, G., Rosenbloom, S., & Cella, D. (2007). Evaluating equivalency between response systems: application of the Rasch model to a 3-level and 5-level EQ-5D. *Medical Care*, *45*, 812–819. doi:10.1097/MLR.0b013e31805371aa.
29. Hawthorne, G., & Richardson, J. (2001). Measuring the value of program outcomes: a review of multiattribute utility measures. *Expert Review of Pharmacoeconomics and Outcomes Research*, *1*, 215–228.
30. Chan, K.-Y., Drasgow, F., & Sawin, L. L. (1999). What is the shelf life of a test? The effect of time on the psychometrics of a cognitive ability test battery. *The Journal of Applied Psychology*, *84*, 610–619. doi:10.1037/0021-9010.84.4.610.
31. Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, *16*(Suppl 1), 5–18. doi:10.1007/s11136-007-9198-0.
32. Rummel, R. J. (1970). *Applied factor analysis* (pp. 24–27). Evanston, IL: Northwestern University Press.
33. Conway, J. M., & Huffcutt, A. I. (2003). A review and evaluation of exploratory factor analysis practices in organizational research. *Organizational Research Methods*, *6*, 147–168. doi:10.1177/1094428103251541.
34. Ford, J. K., MacCallum, R. C., & Tait, M. (1986). The application of exploratory factor analysis in applied psychology: a critical review and analysis. *Personnel Psychology*, *39*, 291–314. doi:10.1111/j.1744-6570.1986.tb00583.x.
35. Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed., p. 966). New York: Harper Collins.
36. Reise, S. P., Waller, N. G., & Comrey, A. L. (2000). Factor analysis and scale revision. *Psychological Assessment*, *12*, 287–297. doi:10.1037/1040-3590.12.3.287.
37. Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate data analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall.
38. Stone, M., & Yumoto, F. (2004). The effect of sample size for estimating Rasch/IRT parameters with dichotomous items. *Journal of Applied Measurement*, *5*, 48–61.
39. Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions*, *7*, 328–331.
40. Pallant, J. F., & Tennant, A. (2007). An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). *The British Journal of Clinical Psychology*, *46*, 1–18. doi:10.1348/014466506X96931.
41. Hambleton, R. K., Jones, R. W., & Rogers, H. J. (1993). Influence of item parameter estimation errors in test development. *Journal of Educational Measurement*, *30*, 143–155. doi:10.1111/j.1745-3984.1993.tb01071.x.
42. SPSS Inc. (2004). SPSS for Windows, version 13.0. Chicago, IL: SPSS Inc.
43. Andrich, D., Sheridan, B., & Luo, G. (2005). *RUMM2020*. Perth, Australia: RUMM Laboratory Pty Ltd.
44. Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park: Sage Publications.
45. Dorans, N. J. (2007). Linking scores from multiple health outcome instruments. *Quality of Life Research*, *16*(Suppl 1), 85–94. doi:10.1007/s11136-006-9155-3.
46. Teresi, J. A. (2006). Overview of quantitative measurement methods. Equivalence, invariance, and differential item functioning in health applications. *Medical Care*, *44*, S39–S49. doi:10.1097/01.mlr.0000245452.48613.45.
47. Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research*, *16*(Suppl 1), 19–31. doi:10.1007/s11136-007-9183-7.