Qual Life Res (2008) 17:275–290
DOI 10.1007/s11136-007-9281-6

Nonparametric IRT analysis of Quality-of-Life Scales and its application to the World Health Organization Quality-of-Life Scale (WHOQOL-Bref)

Klaas Sijtsma · Wilco H. M. Emons · Samantha Bouwmeester · Ivan Nyklíček · Leo D. Roorda

Accepted: 6 November 2007 / Published online: 2 February 2008

Abstract
Background This study investigates the usefulness of the nonparametric monotone homogeneity model for evaluating and constructing Health-Related Quality-of-Life Scales consisting of polytomous items, and compares it to the often-used parametric graded response model.
Methods The nonparametric monotone homogeneity model is a general model of which all known parametric models for polytomous items are special cases. Merits, drawbacks, and possibilities of nonparametric and parametric models and available software are discussed. Particular attention is given to the monotone homogeneity model (also known as the Mokken model), and the often-used parametric graded response model.
Results Data from the WHOQOL-Bref were analyzed using both the monotone homogeneity model and the graded response model. The monotone homogeneity model analysis yielded unidimensional scales for each content domain. Scalability coefficients further showed that some items have limited scalability with respect to the other items in the same scale. The parametric IRT analyses lead to the rejection of some of the items.
Conclusions The nonparametric monotone homogeneity model is highly suited for data analysis in a health-related quality-of-life context, and the parametric graded response model may add interesting features to measurement provided the model fits the data well.

Keywords Health-related quality-of-life measurement · Item response theory · Nonparametric monotone homogeneity model · Parametric graded response model

Acronyms

GRM	Graded response model
HRQoL	Health-related quality-of-life
IRF	Item response function
IRT	Item response theory
ISF	Item score function
ISRF	Item-step response function
LI	Local independence
M	Monotonicity
MHM	Monotone homogeneity model
UD	Unidimensionality
WHOQOL-Bref	World Health Organization Quality-of-Life Scale

Introduction

Questionnaires for health-related quality-of-life (HRQoL) measurement are important for several reasons. First, they may be used to compare the mean level of different patient groups with respect to physical, mental and social health. A researcher may want to find out whether these patient groups have different needs with respect to, for example, therapy or medication or whether different adaptations of their environment are in order so as to improve their

K. Sijtsma (✉)
Department of Methodology and Statistics FSW, Tilburg University, PO Box 90153, Tilburg 5000 LE, The Netherlands
e-mail: k.sijtsma@uvt.nl

W. H. M. Emons · I. Nyklíček
Tilburg University, Tilburg, The Netherlands

S. Bouwmeester
Erasmus University, Rotterdam, The Netherlands

L. D. Roorda
Jan van Breemen Institute and VU University Medical Center, Amsterdam, The Netherlands

conditions of living. Second, HRQoL questionnaires are also important for the measurement of mean change—either progress or deterioration—of such groups due to, for example, therapy. The researcher's interest then lies in the effectiveness of therapy with respect to HRQoL. Third, the total score a patient obtains on an HRQoL questionnaire may be used to diagnose this patient's general level of physical health and psychological well-being, for example, so as to be able to estimate the budget needed for his/her treatment during a particular period.

To effectively measure HRQoL, we argue that an instrument must meet two requirements. The first requirement is that it is clear what the instrument measures: one overall dimension of HRQoL or several dimensions reflecting different aspects of HRQoL. If the instrument measures one dimension, one can use the total score on all items to obtain an impression of the overall level of HRQoL. If the instrument measures multiple dimensions, it may be recommendable to determine total scores on subsets of items (e.g., domain scores), each reflecting a particular aspect of HRQoL (e.g., HRQoL with respect to physical, psychological, and social limitations) and then assess individuals or compare groups on a profile of scores. These two cases may be characterized as unidimensional and multidimensional measurement.

The second requirement is that the psychometric properties of the items are known and found sufficient. One important psychometric item property is the item's location on the scale that quantifies the HRQoL aspect of interest. For example, patients are likely to experience fewer problems when engaging in activities like bathing and dressing than in more demanding activities such as shopping and travelling. The items concerning bathing and dressing require a lower level of physical functioning than the other two items. Thus, bathing and dressing are located further to the left (at a lower level of the scale) than shopping and travelling. A good diagnostic HRQoL instrument contains items of which the locations are widely spread along the scale. Such a scale allows for measurement at varying levels of physical functioning and may be used, for example, for assessing mean differences between groups, mean change due to therapy, and individual patients' levels of physical functioning.

Another important psychometric item property is the item's discrimination power. This is the degree to which the item distinguishes patients with relatively low psychological well-being levels from patients with relatively high psychological well-being levels. The higher the discrimination power the higher the item's contribution to reliable measurement ([1], pp. 101–124). A good diagnostic instrument has items with high discrimination power that each contributes effectively to reliable measurement of patients at different locations along the scale. Such an instrument picks up differences between groups, effects of therapy, and individual levels of activity limitation.

Item response theory (IRT) models [1] are becoming more popular as statistical tools for scale construction in the HRQoL context. IRT can be used effectively to investigate the dimensionality of an instrument and the psychometric properties of its constituent items. The goal of this study is to discuss one particular class of IRT models known as *nonparametric* IRT models [2–5], and to argue that this class in particular provides a general and flexible data analysis framework for studying the dimensionality of a set of polytomously scored items (with dichotomously scored items as special cases) and ascertaining ordinal scales for the measurement of HRQoL aspects which contain items that have varying locations and sufficient discrimination power. Over the past few years, nonparametric IRT models already have been used occasionally for constructing HRQoL scales; see [6–9].

Our point of view is that, given that the researcher has formulated desirable measurement properties, (s)he should construct his/her scale by means of an IRT model that is as general as possible while satisfying the desired measurement properties. Examples of such properties are that the items measure the same dimension, that the measurement level is at least ordinal, and that measurement values are reliable. An HRQoL researcher who has constructed and pre-tested a questionnaire consisting of, say, 40 items is not served well when his/her data are analyzed by means of an IRT model that is unnecessarily restrictive, the result of which is that, say, half of the items are discarded. We will argue that the most general IRT model that serves one's purposes well, often (but not always) is a nonparametric IRT model.

Many questionnaires are used for assessing differences in HRQoL between groups, change due to therapy, and individual patients' scale levels, and a general nonparametric IRT model then is the perfect choice for analyzing one's data. Nonparametric IRT models have several advantages over more-restrictive *parametric* IRT models [1]: Nonparametric IRT models (1) are based on less-restrictive assumptions, thus they allow more items into the scale while maintaining desirable measurement properties; (2) offer diverse tools for HRQoL analysis that give ample information about the dimensionality of the data and the properties of the items; and (3) provide patient measurement values and item location and discrimination values, which have an interpretation that is close to intuition and therefore easy to interpret for users of HRQoL scales. For computerized adaptive HRQoL testing, more-restrictive parametric IRT models such as the Rasch [10] model (dichotomous items) and the generalized partial credit model [11] and the rating scale model [12] (polytomous items) are more appropriate than nonparametric models.

Parametric IRT models have been used more than nonparametric IRT models, especially in psychological and educational measurement, and also in HRQoL research (e.g., [13–15]). A reason for this may be that nonparametric IRT models were developed later than parametric IRT models. See [16–18] for reviews of nonparametric IRT.

This paper is organized as follows. First, we explain assumptions of IRT and compare parametric IRT and nonparametric IRT. Second, we discuss methods and software from nonparametric IRT that can be used for analyzing the polytomous item scores obtained from HRQoL questionnaires. Third, we use this software to analyze data from the World Health Organization Quality-of-Life Scale (WHOQOL-Bref) [19]. The results are compared to those obtained by means of a parametric IRT model. Finally, we provide recommendations for HRQoL researchers on how to use nonparametric IRT methods for analyzing their data.

Definitions and assumptions

IRT models are suited for the analysis of multi-item questionnaire data that typically result from HRQoL questionnaires. The data are discrete scores representing the responses of N respondents to J items (items are indexed j; $j = 1, ..., J$). Many HRQoL questionnaires use items that have three or more ordered answer categories represented by three or more ordered scores, also called polytomous item scores. For simplicity, we assume that all items have the same number of ordered answer categories; this number is denoted by $M + 1$. Let X_j be the random variable representing the discrete score on item j, and let the items be scored $X_j = 0,...,M$. For example, for an item asking whether one is satisfied with one's sleep, score 0 may represent '*very much dissatisfied*' and score M may represent '*very much satisfied*', and the intermediate scores represent intermediate levels of satisfaction. With dichotomous item scoring, $X_j = 0,1$, and possible intermediate satisfaction levels are not quantified separately but collapsed into the score categories 0 and 1. Usually, researchers summarize the J item scores for each patient by the total score (sometimes referred to as the sum score), which is formally defined as $X_+ = \sum_{j=1}^{J} X_j$. Total score X_+ is an estimate of a patient's true score T; this is the expectation of X_+ across independent replications of the measurement procedure ([20], pp. 29–30).

IRT models distinguish observable or manifest variables such as item score X_j and total score X_+ from latent variables. These latent variables play the role of summaries of the behavior that is described by the responses to the items. Sometimes, latent variables are interpreted as if they were causal agents driving responses to items and individual differences between patients. We will also use the distinction between latent and manifest variables in our examples. Thus, we assume that patients are characterized by either one latent HRQoL attribute (meaning that measurement is unidimensional) or different HRQoL attributes (meaning that measurement is multidimensional) which together represent the patient's latent physical, mental or social health. For example, assume that measurement is unidimensional and that the latent variable is mental health or psychological well-being. Psychological well-being then is an unobservable state in each patient, and inferences about it are made on the basis of the manifest responses reported by patients in reaction to the items in an HRQoL questionnaire. Latent variables are denoted by notation θ. If measurement is unidimensional, the IRT model contains one latent variable θ, and if it is multidimensional multiple latent variables are needed. We only consider unidimensional IRT models here; see [21] for a discussion of multidimensional IRT models.

Several families of IRT models for polytomous item scores have been proposed (e.g., [22]). The family of graded response models (GRMs; [23]) is suitable for analyzing ordered item scores collected by means of polytomous response scales [22, 24, 25]. Suppose the item 'How much do you enjoy life?' has four ordered answer categories running from 'Not at all' (score 0) to 'Very much' (score 3); thus $X_j = 0,...,3$ and $M = 3$. For this item, GRMs conceptualize the response process by means of four conditional response probabilities, called item-step response functions (ISRFs) and denoted by $P(X_j \geq m|\theta)=P_{jm}(\theta)$ $(m = 0,...,M)$. It may be noted that for $m = 0$ we have that $P_{j0}(\theta) = 1$ by definition, irrespective of the latent variable level; thus, this response probability is uninformative about the response process and may be ignored. The ISRF describes the relationship between expressing at least a particular minimum level of enjoying life (i.e., having at least a score of m on the example item) and the latent variable of psychological well-being (θ).

Figure 1a shows an example of the ISRFs of two items having four answer categories each (i.e., $M = 3$). The solid lines denote the ISRFs for one item and the dashed lines denote the ISRFs for the other item. Two things are noteworthy. First, the ISRFs of different items may intersect but based on the cumulative character of the definition of the ISRFs, the ISRFs of the same item cannot intersect. Second, the ISRFs have been drawn as monotone curves with rather irregular shapes. Such shapes are typical of a nonparametric GRM [24] and found not to prevent a set of items from having favourable measurement properties, as we will see shortly. Figure 2b also shows monotone ISRFs but now having smooth S-shapes, typical of a parametric GRM. This parametric GRM nearly has the same measurement properties as the nonparametric GRM, but

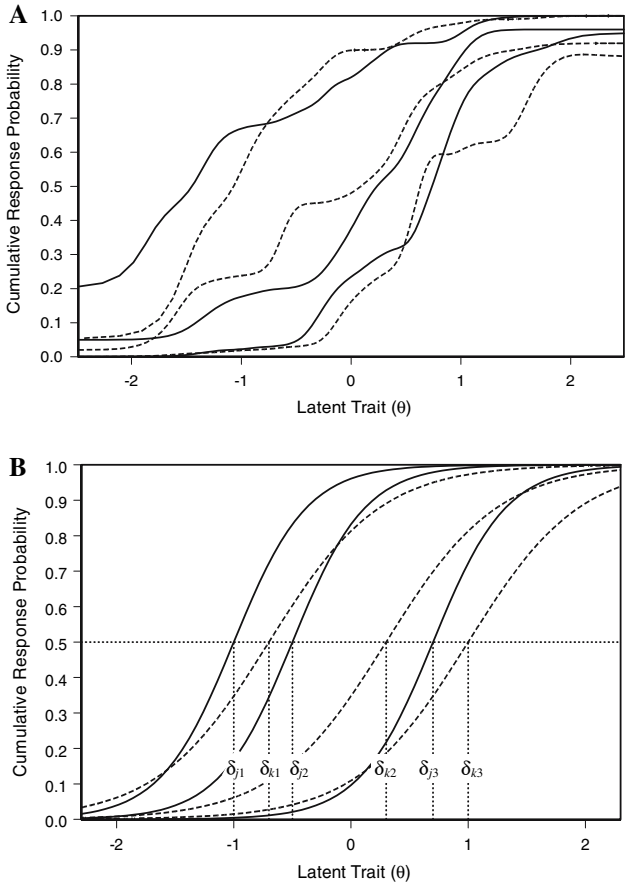

Fig. 1 Examples of item step response functions of (**a**) a nonparametric item response model and (**b**) a parametric item response model

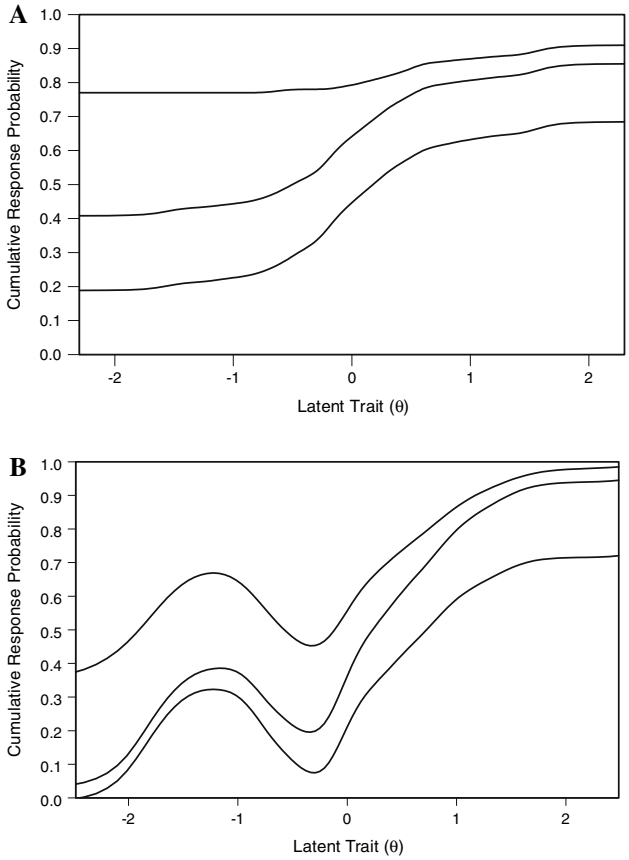

Fig. 2 Examples of (**a**) flat ISRFs and (**b**) non-monotone ISRFs

because it assumes smooth ISRFs (Fig. 1b) instead of irregular ISRFs (Fig. 1a), it is more restrictive and often leads to the rejection of more items (and thus shorter scales) than its nonparametric counterpart.

GRMs have the next three assumptions as a point of departure. The first assumption is unidimensionality (UD); that is, each model assumes that one latent variable θ summarizes the variation in the J item scores in the questionnaire. Assumption UD implies that respondents can be ordered meaningfully by means of a single number. The second assumption is local independence (LI); that is, if we condition on θ, the J item scores are statistically independent. An implication of LI is that in a subgroup of patients who have the same θ value, all covariances between item scores are 0. The third assumption is monotonicity (M); that is, the ISRFs are each assumed to be monotone increasing functions in θ (see Figs. 1a and b). Applied to the latent variable of psychological well-being (θ) and the items from the WHOQOL-Bref, assuming UD, LI and M means that we hypothesize that (1) only psychological well-being drives responses to items and has a systematic effect on individual differences in item scores and total scores (UD); (2) given a fixed level of psychological well-being, relationships between concrete aspects of psychological well-being such as represented by the items 'How much do you enjoy life?' and 'Are you able to concentrate?' are explained completely (i.e., the covariance between these items, conditional on θ, equals 0) (LI); and (3) the higher the level of psychological well-being, the higher the probability that one enjoys life and is able to concentrate.

Parametric and nonparametric graded response models

Parametric Graded Response Model. The ISRFs of the parametric GRM [26] are defined by logistic functions that have the following parameters:

- δ_{jm}: the location parameter of the mth ISRF of item j (i.e., $P_{jm}(\theta)$) on the scale of θ;
- α_j: the slope parameter or 'discrimination power' of item j.

The meaning of these parameters is explained after the ISRF of the parametric GRM is introduced. This ISRF is defined as

$$P_{jm}(\theta) = \frac{\exp[\alpha_j(\theta - \delta_{jm})]}{1 + \exp[\alpha_j(\theta - \delta_{jm})]}, \quad j = 1, ..., J; \quad m = 1, ..., M.$$

Figure 1b shows the logistic ISRFs of two items ($M = 3$). The two items are denoted by j (solid ISRFs) and k (dashed ISRFs). For each item, the three location parameters are also shown. For item j, by definition we have that $\delta_{j1} < \delta_{j2} < \delta_{j3}$, and for item k by definition $\delta_{k1} < \delta_{k2} < \delta_{k3}$. An ISRF's location parameter is the value of θ for which the probability of having an item score of at least m equals .5: that is, $P_{jm}(\theta) = P_{km}(\theta) = .5$, $m = 1,..., M$.

Figure 1b also shows that the slopes of the ISRFs of the same item are equal (mathematically, they must be equal or the ISRFs would intersect; this is impossible given the cumulative definition of the ISRFs), but also that the ISRFs of item j are steeper than the ISRFs of item k. Steepness of slopes is evaluated as follows. For ISRF m of item j, consider the point with coordinates (δ_{jm}, .5). This is the point in which the slope of a logistic ISRF is steepest, and this steepest slope is taken to be typical for the whole ISRF. Parameter α_j expresses this maximum steepness (but is not exactly equal to it). In the example in Fig. 1b, we have that $\alpha_j > \alpha_k$.

Nonparametric Graded Response Model. Instead of choosing a parametric function, nonparametric IRT models typically define order restrictions on the ISRFs. The nonparametric GRM, better known as the monotone homogeneity model (MHM) for polytomous items [27, 4], assumes UD, LI, and M: that is, for any pair of θs, say, θ_a and θ_b, the MHM assumes that

$$P_{jm}(\theta_a) \leq P_{jm}(\theta_b), \quad \text{whenever } \theta_a < \theta_b.$$

Thus, the ISRF is monotone non-decreasing in θ; see Fig. 1a for examples of ISRFs that are monotone but not logistic. This assumption says that a higher level of psychological well-being induces a higher probability of obtaining at least an item score of m (i.e., a higher item score). The ISRFs of different items can have any monotone form and be very different. Requiring monotone ISRFs only is less restrictive than requiring monotone logistic ISRFs; thus, the MHM is a more general model for describing the data than the GRM (henceforth, we call the nonparametric GRM by its better known name (in fact, acronym) MHM, and the parametric GRM simply the GRM).

Unlike the GRM, the MHM does not provide numerical estimates of the latent variable θ . Instead, the MHM allows that total score X_+ orders patients stochastically on latent variable θ in almost all practical measurement situations [28]. This means that, for two total scores X_+ denoted v and w,

$$E(\theta|X_+ = v) \leq E(\theta|X_+ = w), \quad \text{for } 0 \leq v < w \leq J,$$

[4]. This inequality says that as the total score increases the mean θ also increases (or stays the same). Thus, groups of patients that have higher total scores, on average also have higher latent variable values. This result may not seem spectacular at first sight, but it (1) ascertains an ordinal scale for patient measurement (2) using only observable total scores (without requiring the actual estimation of θ). For the psychological well-being example, if the MHM fits the data, ordering patients by means of the total score by implication orders them on the latent variable θ.

Also, the MHM does not provide numerical estimates of the item parameters δ and α. Instead, a distinction can be made between drawing information about item functioning from estimates of the complete ISRFs and item parameters typical of the MHM. Estimates of the complete ISRFs provide much information about the exact relationship between the item scores and the latent variable [16, 29]. ISRFs that are relatively flat or fail to be monotone can be studied in much detail so as to reveal why they dysfunction.

Figure 2a shows three relatively flat ISRFs of a hypothetical item ($M = 3$). This item does not distinguish low θ and high θ patients well. It may be noted that the ISRFs do not all need to be flat simultaneously, but given that they cannot intersect if one ISRF is flat others are likely to be relatively flat as well and the item as a whole contributes little to the reliable ordering of patients on θ.

Figure 2b shows three non-monotone ISRFs of another hypothetical item. Each shows relative good distinction between low θ values and between above-average θ values but bad distinction just below the middle of the scale. Again, non-intersection of the ISRFs of the same item implies that often several ISRFs simultaneously show such disturbing non-monotonicities. For the example item one may conclude that when the questionnaire contains few items, which are effective in the high θ area, this item may be retained to cover this area even though this would be at the expense of measurement quality just below the middle of the scale.

For each item, the MHM framework provides M location parameters and a scalability coefficient, which provides information about item discrimination. The location parameters are the proportions of the population of interest, which have at least a score m on item j, and which are denoted by π_{jm}, $m = 1, ..., M$. For the same item, due to non-intersection of ISRFs we have that $\pi_{j1} \geq \ ... \geq \pi_{jM}$, whereas in the GRM item location parameters are ordered oppositely, $\delta_{j1} \leq ... \leq \delta_{jM}$. The item scalability coefficient H_j (e.g., [2], pp. 148–153; [18], chap. 4) summarizes the discrimination power of an item across its M ISRFs.

The numerical H_j value is determined by the interplay of the slope and the location of the ISRFs of all J items and the distribution of the latent variable θ [30], and it expresses how well item j separates patients given the ISRFs of item j relative to the other items' ISRFs and the distribution of θ. Mathematically, holding constant this distribution and the location of all items' ISRFs, coefficient H_j is higher when the slopes of ISRFs of item j are steeper [18].

Nonparametric versus parametric graded response models

Relationships Among Models. The MHM shares assumptions UD, LI and M with the GRM, but the MHM is less restrictive with respect to the shape of the ISRFs. Thus, the MHM is more general than the GRM or, equivalently, the GRM is a special case of the MHM. This hierarchy implies that, if the GRM fits the data, by implication the more general MHM also fits but if the MHM fits the data, this does not imply that the GRM also fits. Fit of the GRM then needs to be investigated separately. Because of the hierarchical relationship, for any data set the MHM fits as least as many items as the GRM (Table 1).

Patient and Item Parameters. The MHM and the GRM provide the following patient and item parameters (also, see Table 1):

- For patient measurement, the MHM uses total score X_+ to order patients on latent variable θ. Because total score X_+ has an easy interpretation and, moreover, in many IRT models X_+ and θ tend to correlate extremely high suggesting a strong linear relationship [31], total score X_+ may be preferred in practice. Total score X_+ is the sum of the rating scale scores on the J items, whereas estimates of θ are expressed on a logit scale, which does not have a straightforward interpretation for users of HRQoL scales. In general, the ordinal relationship of θ with total score X_+ (which can be approximated well by a linear relation) enables users to switch between scales, and use the one that suits their goals best.
- Because item location δ_{jm} is expressed on the same scale as latent variable θ, it also has an interpretation in logits. For many users, proportion π_{jm}, the proportion of patients who have at least an item score of m, has an easier interpretation.
- Item discrimination α_j gives the maximum slope of the logistic ISRF irrespective of the locations of the other ISRFs of item j and the other items in the questionnaire, and irrespective of the distribution of θ. Thus, information on ISRF slopes is absolute in the sense that a particular α_j value does not provide information on the item's suitability for measurement in a particular group (characterized by a particular distribution of θ) by means of a set of J items (characterized by particular location and slope parameters). On the other hand, item scalability coefficient H_j depends explicitly on the interplay of the distribution of θ, the spread of locations of the ISRFs, and the slopes of the ISRFs. In particular, keeping two of these factors fixed, H_j tends to increase in the third. This dependence on the distribution of θ and the item properties informs the researcher precisely how well item j separates patients with low and high θ values in the particular group of patients under consideration using the particular set of items. The difference between absolute slope information (α_j) and relative slope information (H_j) is illustrated as follows.

Two data sets of size $N = 5{,}000$ and five items ($J = 5$) ($M = 3$ for each item) were generated using the item parameters in Table 2. The first data set came from a

Table 1 Comparison of monotone homogeneity model (MHM) and graded response model (GRM)

	Nonparametric IRT (MHM)	Parametric IRT (GRM)	
Restrictiveness of models	Low; many items admitted to the scale	High; fewer items admitted to the scale	
Interpretation of parameters	Intuitively appealing	More-complicated	
Parameters			(*typical range*)
Person level	T, X_+	θ	$(-3 \leq \theta \leq 3)$
ISRF location	π_{jm}	δ_{jm}	$(-3 \leq \delta_{jm} \leq 3)$
ISRF discrimination	H_j	α_j	$(0.5 \leq \alpha_j \leq 2.5)$
Data analysis	Exploratory, data as point of departure	Confirmatory, model as null hypothesis	
Applications	Comparing groups	Comparing groups	
	Measuring change	Measuring change	
	Diagnosing patients	Diagnosing patients	
		Constructing item banks	
		Adaptive testing	

Table 2 Example of item parameters of the graded response model (GRM), and Item H Values in two different populations with normally distributed latent traits

Item j	Item parameters GRM				H_j	
	α_j	δ_{j1}	δ_{j2}	δ_{j3}	$\theta \sim$ N(–2,1)	$\theta \sim$ N(1,1)
1	1.4	–2.4	–2.2	–1.0	0.36	0.26
2	1.4	–4.0	–2.0	1.0	0.37	0.39
3	1.4	–1.0	2.0	2.5	0.39	0.42
4	1.4	–3.0	–2.0	–1.0	0.36	0.25
5	1.4	–2.5	–2.0	–1.5	0.36	0.25

Note: H_j values are based on simulated data for sample size $N = 5{,}000$

hypothetical clinical population with $\theta \sim$ N(−2,1) (i.e., low psychological well-being level), and the second data set came from a hypothetical healthy population with $\theta \sim$ N(1,1) (i.e., high psychological well-being level). It may be noted that for each of the five items, $\alpha_j = 1.4$ by definition, irrespective of ISRF location parameters and the θ distribution. The H_j values in the clinical group were computed and found to range from .36 to .39. In the healthy group, the H_j values were found to be smaller for items 1, 4, and 5 ($H_1 = .26$, $H_4 = H_5 = .25$). In the clinical group, the ISRFs of these three items were located more closely to the middle of the θ distribution (Fig. 3 shows this for Item 4) such that higher H_j values resulted, but in the healthy group the ISRFs were located further in the lower tail of the θ distribution (Fig. 3) resulting in lower H_j values (.3 is considered minimally acceptable; [2], chap. 5). Thus, the items 1, 4, and 5 are well suited for measurement in the clinical group but not in the healthy population, despite their overall discrimination power expressed by $\alpha_j = 1.4$, for $J = 1, ..., 5$. For item 3, H_3 was a little higher in the healthy group because the second and third ISRFs discriminate particularly well at higher ranges of θ. For item 2, the location parameters of the ISRFs were widely spread across the θ distribution, resulting in good discrimination both at lower and higher levels of θ in both distributions.

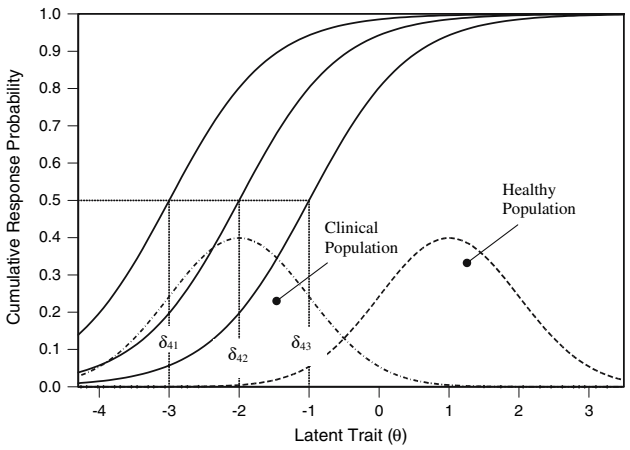

Fig. 3 Three ISRFs of the same item (Item 4) relative to two different distributions of the latent variable

We conclude that item scalability coefficient H_j has the advantage that it takes the item (and not the individual ISRF) as a unit and depends simultaneously on the θ distribution, the slopes of the ISRFs, and the spread of the locations of the ISRFs. Thus, H_j informs the researcher whether item j discriminates well in the group under consideration using the particular set of items, whereas item discrimination α_j provides information about the discrimination power irrespective of the patient group under consideration and the item properties of item j and the other items in the scale.

Confirmatory and exploratory data analysis. In general, before IRT models are accepted as reasonable descriptions of the data their goodness-of-fit to these data must be investigated and assessed. In general, goodness-of-fit research is different for parametric and nonparametric IRT (with the GRM and the MHM as special cases, respectively). In a parametric IRT analysis the model often serves as null-hypothesis and it is tested whether this null-hypothesis must be rejected or may be supported by the data. Nonparametric IRT analysis in general takes the data as point of departure and (1) instead of positing a unidimensional or multidimensional latent variable structure analyzes the data to find its true dimensionality, and (2) instead of positing a logistic or other functional shape estimates the ISRFs from the data so as to diagnose the items' functioning [16, 18]. This research strategy renders nonparametric IRT a more flexible data-analysis tool than parametric IRT. One could also characterize this distinction as confirmatory (parametric IRT) versus exploratory (nonparametric IRT) (Table 1).

Application of parametric and nonparametric IRT. If a nonparametric IRT model such as the MHM fits the data, the result is a scale on which patients can be ordered by means of the total score X_+. This total score has a strong linear correlation with latent variable θ. Such a scale suffices in many applications. Examples are the comparison of groups, the measurement of change due to therapy, and the establishment of the patient's psychological well-being level as low, medium, or high (Table 1). The practical

advantage of nonparametric IRT models over parametric IRT models is that the scales they produce contain more items thus reducing the risk of wasting items that have non-logistic but monotone ISRFs that discriminate well in (part of) the group under consideration (e.g., Fig. 2b). Such items contribute well to reliable measurement. In addition, rejection of such items may also harm the coverage of the latent attribute.

If a fitting parametric IRT model is obtained for a set or a subset of the items, one has a parsimonious description of the item characteristics, and one can use the estimated item parameters to scale the items, and the estimated θs as interval level measures to locate patients on this scale. If a large set of items, also known as an item bank [32], is available, and if a parametric IRT model fits the item bank, parametric IRT models have the advantage that the patient's θ can be assessed using different sets of items from the item bank. This may be useful for the measurement of change when change is so large that the set of items that was used initially no longer captures the higher or lower θ levels needed for the second measurement, thus necessitating the use of other items. Another application of parametric IRT is computerized adaptive testing (CAT), which selects items that match the patient's θ level well from a huge item bank so as to optimize accuracy of θ measurement. In principle, CAT requires different item sets for different θ values (Table 1).

Nonparametric IRT analysis in practice

Software for nonparametric IRT analysis. Several programs are available for data analysis using nonparametric IRT but not each can handle polytomous item scores. We briefly discuss the programs MSP ([33]; also see [4]) and TestGraf98 [34]. Both programs are used regularly and, together, they provide the researcher with a clear and informative picture of (1) the dimensionality of the data, (2) the (lack of) monotonicity of the ISRFs, and (3) estimates of item locations and item discrimination.

Program *MSP* uses the MHM as the main analysis model but another nonparametric model not discussed here is also included in the program. Basically the program consists of two parts. The first part of MSP is an algorithm for exploring the dimensionality of the data ([2], chap. 5; [4], chap. 5). The algorithm uses item scalability coefficient H_j to select items. Items that are related to the same latent variable θ are selected one by one on the basis of their H_j value. Suppose that one latent variable drives the responses to one subset of the items, another latent variable drives the responses to another subset of items, and so on. Then, the algorithm selects mutually exclusive clusters of items each of which is driven by a different latent variable.

The second part of MSP provides several statistical tools for exploring the shape of the ISRFs [29]. This is most useful after the dimensionality of the data has been ascertained. For example, due to their strong positive tendency the non-monotone ISRFs in Fig. 2b may have an H_j value that is high enough for the items to be selected in a unidimensional cluster, but the non-monotonicity also may distort parts of the ordinal scale defined by the items in the cluster. MSP estimates ISRFs by means of a number of discrete points that are connected to form a jagged 'curve'. Figure 4a shows four discrete ISRFs of the same item (i.e., $M = 4$) each estimated by means of eleven points. The researcher can manipulate the number of estimated points. If more points are estimated from the same data more details of the ISRF become visible (i.e., bias is reduced) but because for each point fewer data are available, accuracy decreases. In statistics, this is known as the bias-accuracy trade-off, and it is advisable to try several options to reach

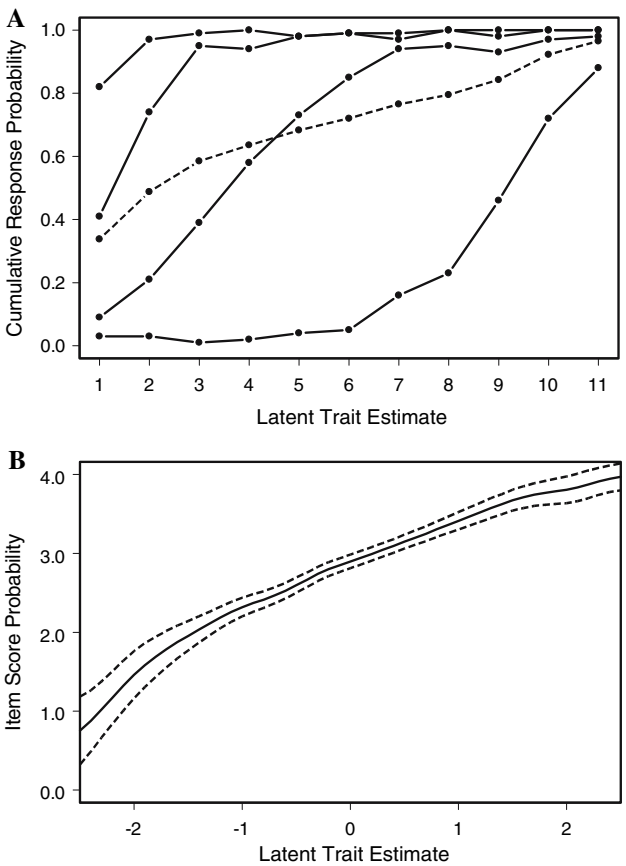

Fig. 4 (**a**) Discrete estimates of four ISRFs (solid curves; from MSP) of Item 16 ('Satisfied doing daily activities?', from Physical Health and Well Being domain), and the corresponding discrete item-score function (dashed curve), (**b**) also estimated as a continuous curve (from Testgraf98)

a good decision. MSP tests observable deviations from monotonicity for significance. Figure 4a also shows a 'mean' ISRF (dashed curve), which we may call the item score function (ISF) and which is not standard output of MSP. This function must be monotone nondecreasing.

Program *TestGraf98* [34] can be used for studying the shape of the ISF. Unlike MSP, TestGraf98 produces continuous estimates of response functions ([35]; here, only the ISFs); and like MSP, TestGraf98 shows graphical displays of these estimates that can be manipulated with respect to bias and accuracy, and also here it is advisable to try several options. The quality of the decision can be improved by using the confidence envelopes for the estimated ISFs for statistical testing. Figure 4b shows an example of an estimated ISF (solid curve) and its confidence envelopes (dashed curves), which were estimated by means of TestGraf98.

Research strategies for nonparametric IRT analysis. MSP provides a method for investigating the dimensionality of the data, and MSP and TestGraf98 both can be used to investigate assumption M. The investigation of dimensionality and monotonicity serves to identify the items that together constitute an ordinal patient scale for the same latent variable.

For investigating dimensionality, MSP offers the researcher the possibility to set a positive lower bound c on H_j. Under the MHM, the lowest admissible value is $c = .0$; MSP's default is $c = .3$ ([2], chap. 5; [4, 33]). This default value ascertains a lower bound on the overall discrimination power of the items (but researchers are free to choose a higher value) and, as a result, item clusters consist only of sufficiently discriminating items that measure the same latent variable. Thus, MSP aims to produce unidimensional scales that allow accurate patient measurement.

TestGraf98 estimates the ISFs (e.g., Fig. 4b) by means of the nonparametric regression method known as kernel smoothing (e.g., [36], chap. 2; [35]). The availability of confidence envelopes for the continuous ISF estimates provides detailed information of (lack of) monotonicity for each item. TestGraf98 provides these estimates irrespective of the dimensionality of the data. Thus, a good research strategy is to first investigate item-set dimensionality by means of MSP and then use MSP and TestGraf98 to study the ISRFs and the ISFs in dimensionally distinct clusters. See [37] for another method for assessing the shape of these curves.

A real-data example: The World Health Organization Quality-of-Life Scale

The WHOQOL-Bref was developed for assessing individuals' perception and feelings of their daily life. The questionnaire starts with two items, which ask for global estimates of one's quality of life, and then continues with 24 items covering four domains: (a) physical health and well-being (seven items); (b) psychological health and well-being (six items); (c) social relations (three items); and (d) environment (eight items). The two general items were left out of the analysis. In agreement with their numbering in the WHOQOL-Bref, the other 24 items were numbered from 3 to 26. Examples of items are:

- Do you have enough energy for daily life? (physical domain)
- How much do you enjoy life? (psychological domain)
- How satisfied are you with your personal relationships? (social domain)
- How safe do you feel in your daily life? (environmental domain)

Each item uses a five-point rating scale (i.e., $X_j = 0, ..., 4$); the higher the item score, the better one's quality of life on the specific domain covered by the item.

The data were collected by undergraduate psychology students of Tilburg University as part of a course Research Practical in the academic year 2005–2006. Students were instructed to strive for a sample of participants equally distributed across both sexes and the following age categories: 30–39, 40–49, 50–59, and more than 60 years. The final sample consisted of $N = 589$ respondents from the Dutch population. Of these respondents, 55% were women, mean age was 55.2 years (SD = 14.6), 32% had completed community college or university, 36% had completed vocational school, 20% had high school at most, and 12% had only elementary school or less.

$N = 55$ cases had missing item scores. Missing values were estimated using two-way imputation. Comparable to an analysis-of-variance layout, this method uses both a person effect and an item effect for estimating a missing score (for details, see [38]). MSP and TestGraf98 were used to analyze these data and construct one or more scales, thus illustrating the possibilities of the MHM. For the sake of comparison, we ran a principal component analysis and a GRM scale analysis on the data.

Results

Sample statistics of item and scale scores

Table 3 shows that the mean item scores ranged from 2.58 (Item 20: 'Satisfied with sex life?') to 3.46 (Item 25: 'Moving around well?'). The mean X_+ scores were 21.04 (physical domain), 17.01 (psychological domain), 8.52 (social domain), and 24.27 (environmental domain). Correlations between the domain scores ranged from .37

Table 3 Results from MSP item selection procedure (Item clusters, item H_j values, and total H), and Item H_j values and total H for each content domain

			MSP item selection procedure							H_j per content domain
			$c = .3$		$c = .4$					
j		Mean	1	2	1	2	3	4	5	
Physical health and well-being										
3	Distraction due to pain[a]	3.04		.59		.59				.40
10	Experiencing energy[a]	2.98	.43		.53					.46
15	Satisfied with sleep	2.66	.22	–	–	–	–	–	–	.28
25	Moving around well	3.46	.36		.43					.41
16	Satisfied doing daily activities	2.84	.41		.56					.52
4	Need medical treatment for daily functioning[a]	3.17		.59		.59				.43
17	Satisfied work capacity	2.89	.40		.57					.52
	Scale value	21.04								.43
Psychological health and well-being										
5	Enjoying life	2.66	.34		.42					.37
7	Being able to concentrate	2.80	.32		–	–	–	–	–	.29
18	Satisfied with yourself	2.95	.41		.48					.45
11	Acceptance physical appearance	3.23	.33		–	–	–	–	–	.35
26	Experiencing negative feelings[a]	2.72	.30		–	–	–	–	–	.34
6	Life meaningful	2.66	.30		–	–	–	–	–	.37
	Scale value	17.01								.36
Social relations										
19	Satisfied relationship with other people	3.06	.34						.50	.50
20	Satisfied with sex life	2.58	.30						.42	.42
21	Satisfied support from others	2.88	.28	–					.40	.40
	Scale value	8.52								.44
Environment										
8	Feeling safe in daily life	3.08	.30		–	–	–	–	–	.33
22	Satisfied living conditions	3.13	.43		–	–	–	–	–	.43
12	Enough financial resources	3.08	.33				.53			.42
23	Satisfied getting adequate health care	2.87	.29	–				.52		.36
13	Availability information needed in daily life	3.06	.34				.53			.40
14	Opportunities leisure	2.90	.34				.49			.39
9	Healthy environment	2.88	.29	–	–	–	–	–	–	.31
24	Satisfied with transport in daily life	3.28	.34					.52		.40
	Scale value	24.27								.38

[a] Reversely scored items

(between physical and social domains) to .51 (between physical and environmental domains).

Dimensionality analysis

Principal components analysis. Dimensionality was explored by means of a principal components analysis using polychoric correlations. The ratio of the first to the second eigenvalue of the polychoric correlation matrix was 8.428/1.986 = 4.24. A ratio of 4:1 is taken as evidence of considerable strength of the first dimension (e.g., [39]). The first factor explained 32.4% of the variance. A confirmatory factor analysis of the four a priori domain scales of the WHOQOL-bref improved fit over the one-factor model ($P \leq 0.001$). However, the factors correlated from .50 (physical and social domain) to .79 (psychological and social domain). The explorative factor analysis in conjunction with the confirmative factor analysis justifies the assumption of a general HRQoL dimension underlying each scale.

Monotone homogeneity model analysis. Next, MSP was used treating all 24 items as a fixed scale. The MHM does

not allow negatively correlating items in one scale. Item 4 ('Need medical treatment for daily functioning?') and Item 20 ('Satisfied with sex life?') correlated negatively but not significantly ($P > .05$); thus all 24 items were used for analysis. The item H_j values (not tabulated) ranged from .21 (Item 15: 'Satisfied with sleep?') to .40 (Item 22: 'Satisfied with living conditions?' and Item 10: 'Enough energy for everyday life?'). The total-scale H coefficient was equal to .30. The results suggest that the items tend to cover one latent HRQoL aspect, which, however, induces only weak general association between the items**.** Thus, in addition to this common aspect it seems reasonable to also look for more-specific HRQoL aspects that are covered by subsets of items.

Dimensionality was investigated by means of the MSP search algorithm using several c values, starting with .3 (default) and then increasing c with steps of .05 in each next analysis round. We only report results for $c = .3$ and $c = .4$ (other values did not reveal interesting results). For $c = .3$, one scale consisting of 18 items ($H = .35$) and one scale consisting of 2 items ($H = .59$) were found (Table 3). The four remaining items were not selected because their H_j values were under .3 (i.e., .22 and .28, .29, and .29). The 18-item scale had a rather heterogeneous content. The 2-item scale asked about distraction due to pain (Item 3) and the need for medical treatment for daily functioning? (Item 4). Thus, their high scalability may be explained by the use of palliative medicines.

For $c = .4$, five scales were found consisting of 6, 2, 3, 2, and 3 items, respectively (Table 3). The first scale consisted of items from the physical domain and the psychological domain. The other scales consisted of items from one domain. Scale 2 again covered Item 3 and Item 4. Scales 3 and 4 covered environmental-domain aspects. Scale 5 contained all social-domain items.

Thus, for default $c = .3$, 18 of the 24 items were selected in one scale. The pattern of item selection for higher c values such as $c = .4$ showed that the item set progressively crumbled into many smaller scales while other items remained unselected. Sijtsma and Molenaar ([4], pp. 80–86; see also [7] [40]) argued that this typical pattern of results gives evidence that the 18-item set constitutes a unidimensional scale. The total-scale H equaled .35, giving evidence of weak scalability ([2], p. 185). Most of the item H_j values were between .3 and .4, also suggesting a weak relationship with the latent variable ([2], p. 185).

Finally, it was investigated whether the four a priori identified item domains could be considered as separate scales. Table 3 (last column) shows that the total-scale H values ranged from .36 (environmental domain) to .44 (social domain). Thus, based on Mokken's classification of scales [2] the four a priori item domains constituted weak to medium scales. Two items had H values just smaller than $c = .3$ (i.e., Item 15 ($H_{15} = .28$): 'Satisfied with sleep?' and Item 18 ($H_{18} = .29$): 'Satisfied with yourself?'). The content domains may be considered as unidimensional clusters of items measuring distinct aspects of HRQoL, each of which are related to a more general underlying HRQoL construct. Because of their conceptual clarity, the remaining analyses were done on the a priori defined item domains.

Monotonicity assessment in each item domain

For each item domain, MSP was used to assess the ISRFs' shapes. First, ISRFs were estimated accurately (i.e., many cases were used to estimate separate points of the ISRFs) but at the expense of possible bias (i.e., only few points were estimated). Second, ISRFs were estimated with little bias (i.e., many points were estimated) but at the expense of accuracy (i.e., few cases were used to estimate each point).

For the physical domain, the first analysis (high accuracy, more risk of bias) revealed four items of which one or more ISRFs showed minor violations of monotonicity, but none of these violations were significant (5% level, one-tailed test, because only sample decreases are tested as violations; increases support monotonicity). The second analysis (more inaccuracy, less bias) revealed that for all seven items one or more ISRFs showed one or more local decreases, but none them were significant. Figure 5a shows the local, nonsignificant decreases in the ISRFs for Item 3 ('Distraction due to pain?').

For the psychological domain, for both analyses (i.e., high accuracy versus little bias) two items were found which had ISRFs showing significant local decreases. For example, Fig. 6a shows for Item 7 ('Being able to concentrate?') that the estimate of ISRF $P_{72}(\theta) = P(X_7 \geq 2|\theta)$ (the second curve from the top) has several local decreases, the largest of which was significant (5% significance level, $P = 0.019$). The estimate of ISRF $P_{73}(\theta) = P(X_7 \geq 3|\theta)$ shows two small decreases; they were not significant. For the social domain and the environmental domain, no significant violations of the monotonicity assumption were found. It can be concluded that assumption M holds for each of the four scales.

Next, TestGraf98 was used to investigate assumption M for the ISFs. Several sample sizes were used for estimating curve fragments of the ISFs and balancing the bias-accuracy trade-off. Figure 5b shows the estimated ISF of Item 3 ('Distraction due to pain?'). The confidence envelopes show that the local decrease of the estimated ISF can be ignored safely. For the estimated ISF of Item 7 ('Being

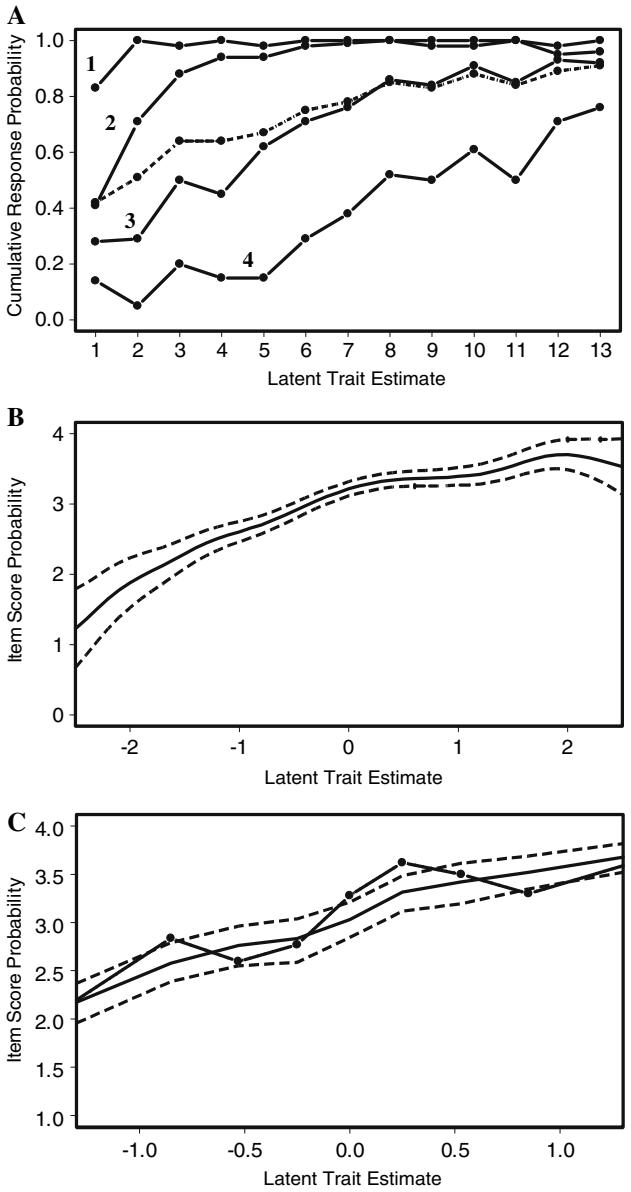

Fig. 5 Four ISRFs of Item 3 ('Distraction due to pain?', from Physical Health and Well Being domain) showing nonsignificant violations of assumption M and rejected by the GRM: (**a**) Results from MSP (including the ISF); (**b**) results from Testgraf98 (ISF and confidence envelopes); (**c**) and results from Multilog7.0 (GRM)

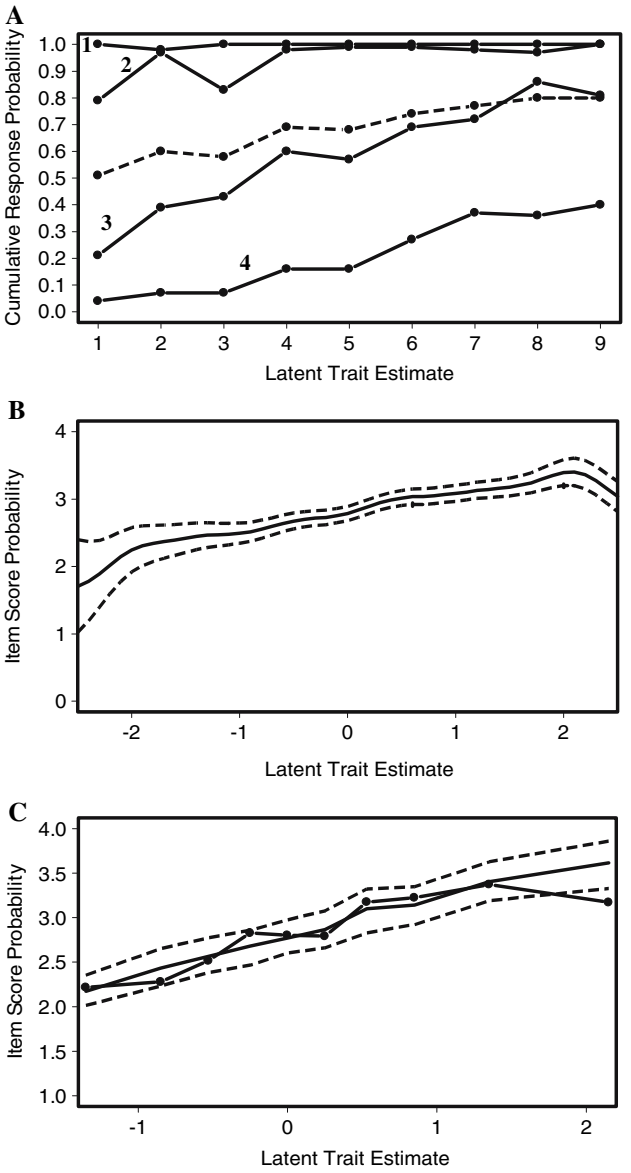

Fig. 6 Four ISRFs of Item 7 ('Being able to concentrate?') showing significant violations of assumption M and rejected by the GRM: (**a**) Results from MSP (including the ISF); (**b**) results from Testgraf98 (ISF and confidence envelopes); and (**c**) results from Multilog7.0 (GRM)

able to concentrate?'), Fig. 6b suggests a violation of assumption M for high latent variable levels (i.e., $\theta > 2$). With MSP the user specifies the minimum number of observations used for estimating each point of an ISRF, but Testgraf98 controls the bias-accuracy trade-off by means of a bandwidth parameter, also to be specified by the user but without being able to control the number of observations for estimating separate curve fragments. The effect may be that, in particular at the lower end and higher end of the scale, the ISF is estimated very inaccurately. Combining the results from MSP (no significant decreases of the ISRFs at the higher end of the θ scale) and Testgraf98 (a smooth monotone increasing ISF in the middle of the θ scale), we may conclude that assumption M holds for item scores. Thus, the expected item score monotonically increases in the latent variable.

Comparison of the MHM with the GRM

The GRM was fitted using Multilog7.0 [41]. Estimation problems occurred for item scoring 0–4 because some

score categories were (almost) empty. This was resolved by combining scores 0 and 1 into score 0, and re-scoring the items 0–3. Table 4 provides the estimated slope parameters (α_j) and the three location parameters (δ_{jm}, $j = 1, 2, 3$), and also the estimated item H_j coefficients (also in Table 3, last column) and the nonparametric location parameters (π_{jm}, $j = 1, 2, 3$). The location parameters indicate that the items are relatively popular (highly endorsed).

Several methods are available for assessing the fit of the GRM, but many are problematic and no generally accepted standard goodness-of-fit method for the GRM is presently available ([42] pp. 85–89). We investigated goodness-of-fit of the GRM by means of posterior predictive assessment [43], which can provide graphical and numerical evidence about model fit. Model fit was investigated at the level of items. The most interesting results were found for items from the physical and psychological domains. Results are only discussed for these domains. For the physical domain, Item 3 ('Distraction due to pain?') and Item 7 ('Being able to concentrate') showed significant misfit ($P < .01$). In Fig. 5c, the curve made up by dots connected by straight line pieces represents the estimated nonparametric ISF of Item 3, the solid curve represents the expected ISF of Item 3 under the GRM, and the dotted curves represent the 95% confidence envelopes. The GRM rejects this item. Thus, modeling the jagged pattern of the estimated ISF by means

Table 4 Results of monotone homogeneity model (MHM) scale analysis and estimated item parameters from the graded response model (GRM)

		MHM					GRM [b]			
j		H_j	π_{j1}	π_{j2}	π_{j3}	π_{j4}	α_j	δ_{j1}	δ_{j2}	δ_{j3}
Physical health and well-being										
3	Distraction due to pain[b]	.40	.99	.93	.72	.40	1.14	−2.71	−0.99	0.49
10	Experiencing energy[a]	.46	.98	.95	.79	.45	1.97	−2.24	−0.63	0.56
15	Satisfied with sleep	.28	.99	.95	.70	.34	0.85	−2.09	−0.79	1.78
25	Moving around well	.41	.99	.96	.91	.61	1.23	−3.08	−2.24	−0.41
16	Satisfied doing daily activities	.52	.98	.93	.72	.21	3.96	−1.59	−0.58	0.87
4	Need medical treatment for daily functioning[a]	.43	.98	.92	.75	.23	1.23	−2.82	−1.29	0.24
17	Satisfied work capacity	.52	.99	.98	.77	.20	3.58	−1.56	−0.69	0.81
	Scale value	.43								
Psychological health and well-being										
5	Enjoying life	.37	1.00	.98	.61	.07	1.93	−2.79	−0.38	1.97
7	Being able to concentrate	.29	.99	.98	.77	.20	0.82	−4.27	0.68	1.69
18	Satisfied with yourself	.45	1.00	.93	.64	.15	1.99	−2.85	−0.97	1.12
11	Acceptance physical appearance	.35	.99	.98	.80	.45	1.08	−3.93	−1.56	0.22
26	Experiencing negative feelings[a]	.34	1.00	.97	.61	.08	1.11	−2.84	−0.60	1.90
6	Life meaningful	.37	1.00	.96	.62	.23	1.90	−2.71	−0.35	1.94
	Scale value	.36								
Social relations										
19	Satisfied relationship with other people	.50	.99	.96	.82	.29	2.52	−2.21	−1.10	0.97
20	Satisfied with sex life	.40	.97	.88	.58	.15	1.32	−3.19	−0.90	1.38
21	Satisfied support from others	.42	.99	.97	.72	.20	1.38	−1.88	−0.34	1.62
	Scale value	.44								
Environment										
8	Feeling safe in daily life	.33	1.00	.98	.78	.32	1.12	−4.12	−1.37	0.84
22	Satisfied living conditions	.43	.98	.90	.68	.34	1.96	−2.71	−1.27	0.63
12	Enough financial resources	.42	.99	.95	.70	.44	1.84	−2.22	−0.73	0.21
23	Satisfied getting adequate health care	.36	.99	.97	.68	.23	1.32	−2.84	−0.93	1.33
13	Availability information needed in daily life	.40	.99	.95	.72	.21	1.64	−3.21	−0.98	0.60
14	Opportunities leisure	.39	.99	.98	.75	.34	1.61	−1.94	−0.69	0.60
9	Healthy environment	.31	.99	.98	.83	.32	1.02	−3.95	−0.89	1.43
24	Satisfied with transport in daily life	.40	.99	.97	.86	.45	1.66	−2.77	−1.55	0.17
	Scale value	.38								

[a] Reversely scored items; [b] For the GRM analysis, items were recoded by collapsing item scores 0 and 1 into item score 1

of logistic ISRFs having the same slopes would do injustice to the data. However, it is noteworthy that Item 3 has good measurement properties (Table 4: e.g., $H_3 = .40$) under the more general MHM, and from this model's perspective it might be retained in the scale.

Item 7 ('Being able to concentrate?') was a popular item (Table 4; $\pi_{j1} = .99$, $\pi_{j2} = .98$; also, see Fig. 6a, upper two ISRFs). As a result, the GRM could not be estimated accurately; item parameters were estimated very inaccurately (standard errors >.25). Figure 6c shows evidence of misfit at $\theta > 2$, for which the observed ISF fell outside the 95% confidence interval. This means that the GRM gives biased results for $\theta > 2$. The nonparametric estimates of the ISRFs were monotone. This provides evidence that the MHM adequately fitted Item 7. However, $H_7 = .29$, which is rather low. A reason to keep this item in the scale is that it may help measuring differences at the lower and middle ranges of the θ scale, which are the most relevant ranges for measuring HRQoL.

Summary of the scale properties

The WHOQOL-Bref is most often used in scientific research (e.g., epidemiological studies and clinical trails) and by health professionals (e.g., to assess treatment efficacy) [19]. The nonparametric MHM analyses revealed that the scales have adequate properties for comparing groups on the underlying HRQoL aspects. Each of the four domains of the WHOQOL-Bref constitutes a unidimensional scale, each scale measuring a different aspect of HRQoL in addition to a weak common HRQoL attribute. This justifies reporting both separate domain scores and possibly an overall HRQoL score. The scalability results showed that the domain scales are weak to moderate, with scalability coefficients H ranging from .36 to .44. The test-score reliabilities of the four domain scores were .82, .76, .66, and .81, respectively. The rank correlations between sum score X_+ and the estimated θ from the GRM varied from .91 ('physical health') to .96 ('environment') (Pearson correlations ranged from .94 ('physical health') to .99 ('environment')). Thus, X_+ and estimated θ carry nearly the same rank order (and numerical) information. This interesting result further justifies the use of the nonparametric MHM for scale analysis, and the use of X_+ for (at least) ordinal measurement of persons.

Because the item-score distributions were severely skewed to the left, the lower response categories 0 and 1 were ineffective for HRQoL measurement in the general population. The locations of the ISRFs for the higher response categories 2, 3, and 4 were well spread along the θ scale. The ISRFs' discrimination power as reflected by the H_j values often was in the weak to medium range. Thus, the higher response categories are modestly informative across a wide range on the θ scale.

The relatively short WHOQOL-Bref may also be considered for use as a tool for assessing HRQoL at the individual level in clinical and medical settings. For example, the WHOQOL-Bref may be used to evaluate whether a patient's HRQoL has improved after taking medication. An interesting feature of a fitting IRT model is that psychometric properties can be evaluated conditionally on the latent variable. For example, the measurement error of X_+ can be evaluated at different values of the latent variable. TestGraf98 provides graphical information about the standard error of measurement based on the MHM model. For example, for the physical-health domain Testgraf98 estimated a standard error of measurement ranging from 2.8 for $X_+ \leq 20$ to 1.8 for $X_+ \geq 28$. Thus, to be significant at the 5% significance level differences between two observed X_+ scores have to be larger than $2.8 \times 1.96 \times \sqrt{2} \approx 8$ for $X_+ \leq 20$, and larger than $1.8 \times 1.96 \times \sqrt{2} \approx 5$ for $X_+ \geq 28$ (e.g., see [44], p. 209). This is a substantial standard error of measurement relative to the length of scale. This relatively large measurement error appears to be consistent with the observed H_j values, which indicate weak to moderate scalability. For the other three content domains, the standard error of measurement was also substantial. Thus, caution has to be exercised when drawing conclusions about differences and changes in individual levels of HRQoL based on observed X_+ scores from the WHOQOL-Bref and any other HRQOL measure—see [45].

Discussion

This study explained how the nonparametric monotone homogeneity model contributes to the construction of scales for the measurement of HRQoL. The MHM is more general than parametric IRT models [24], such as the much-used parametric graded response model [26] but also the partial credit model [13–15] and the generalized partial credit model [11]. Hemker et al. [46] showed that all known parametric IRT models for polytomous items are special cases of the nonparametric MHM. This means that any item set satisfying the requirements of a parametric IRT model for polytomous items also satisfies the requirements of the nonparametric MHM. Given the greater generality and flexibility of the nonparametric MHM, which results in longer scales, and because X_+ and estimated θ carry the same rank order information (based on the approximate stochastic ordering property of θ given X_+), the nonparametric MHM is highly suited for person measurement.

In an HRQoL context, often little is known about the psychometric properties of new questionnaires. A typical nonparametric MHM analysis explores the dimensionality of the data by capitalizing on model assumptions such as monotonicity (MSP), and studies the shapes of the ISRFs and the ISFs in order to learn more about the (mal-)functioning of individual items (MSP and TestGraf98). This results in scales on which groups can be compared and changes monitored without making unduly restrictive assumptions about the data.

The properties of any IRT model only hold for the application at hand when the model fits the data. In case of misfit, the structure of the model does not match the structure of the data. One cause of misfit is that the data are multidimensional while the model assumes unidimensionality. Another cause of misfit is that the real ISRFs may not be monotone or that they are monotone but fail to have the logistic shape assumed by many parametric models. Other causes of misfit, such as a multiple-group structure as in differential item functioning (e.g., [47]) or person misfit [48] were not considered here.

When the MHM fits the data, the researcher may decide to also investigate goodness of fit of the GRM or other parametric IRT models for polytomous items. The choice of a parametric model may be based on the flexibility of the model. For example, the partial credit model only has item location parameters but assumes the slopes of the response functions to be the same within and between items, whereas the generalized partial credit model also allows for varying slope parameters between items, just as the GRM. If one pursues a parametric IRT analysis, misfit may be a good reason to resort to a nonparametric IRT model and still have an ordinal patient scale. If CAT is pursued, one of the parametric models is a better option provided the model fits the data well. In an HRQoL context, CAT indeed could prove to be successful because patients have a definitive interest in providing truthful answers (in the educational context, in which CAT originated, CAT requires that items be kept secret. This requires item banks often containing hundreds of calibrated items). As a result, in HRQoL measurement CAT presently meets with a growing interest (e.g., [49–54]).

References

1. Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory. Principles and applications*. Boston: Kluwer Nijhoff.
2. Mokken, R. J. (1971). *A theory and procedure of scale analysis*. Berlin: De Gruyter.
3. Mokken, R. J. (1997). Nonparametric models for dichotomous responses. In W. J. v. d. Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 351–367). New York: Springer.
4. Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage.
5. Petersen, M. A. (2004). Book review: Introduction to nonparametric iterm response theory. *Quality of Life Research, 14*, 1201–1202.
6. Ringdal, K., Ringdal, G. I., Kaasa, S., Bjordal, K., Wisløff, F., Sundstrøm, S., & Hermstad, M.J. (1999). Assessing the consistency of psychometric properties of the HRQOL scales within the EORTC QLC-C30 across populations by means of the Mokken scaling model. *Quality of Life Research, 8*, 25–41.
7. Moorer, P., Suurmeijer, Th. P. B. M., Foets, M., & Molenaar, I. W. (2001). Psychometric properties of the RAND-36 among three chronic diseases (multiple sclerosis, rheumatic diseases and COPD) in the Netherlands. *Quality of Life Research, 10*, 637–645.
8. Van der Heijden, P. G. M., Van Buuren, S., Fekkes, M., Radder, J., & Verrips, E. (2003). Unidimensionality and reliability under Mokken scaling of the dutch language version of the SF-36. *Quality of Life Research, 12*, 189–198.
9. Roorda, L. D., Roebroeck, M. E., Van Tilburg, T., Molenaar, I. W., Lankhorst, G. J., Bouter, L.M., & the Measuring Mobility Studying Group (2005). Measuring activity limitations in walking: Development of a hierarchical scale for patients with lower-extremity disorders who live at home. *Archives of Physical Medicine and Rehabilitation, 86*, 2277–2283.
10. Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Nielsen & Lydische.
11. Muraki, E. (1997). A generalized partial credit model. In W. J. v. d. Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 153–164). New York: Springer.
12. Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561–573.
13. Barley, E. A., & Jones, P. W. (2006). Repeatability of a Rasch model of the AQ20 over five assessments. *Quality of Life Research, 15*, 801–809.
14. Fitzpatrick, R., Norquist, J. M., Jenkinson, C., Reeves, B. C., Morris, R. W., Murray, D. W., & Gregg, P. J. (2004). A comparison of Rasch with likert scoring to discriminate between patients' evaluations of total hip replacement surgery. *Quality of Life Research, 13*, 331–338.
15. Kosinski, M., Bjorner, J. B., Ware, J. E. Jr, Batenhorst, A., & Cady, R. K. (2003). The responsiveness of headache impact scales scored using 'classical' and 'modern' psychometric methods: A re-analysis of three clinical trials. *Quality of Life Research, 12*, 903–912.
16. Junker, B. W., & Sijtsma, K. (2001). Nonparametric item response theory in action: An overview of the special issue. *Applied Psychological Measurement, 25*, 211–220.
17. Stout, W. F. (2002). Psychometrics: From practice to theory and back. *Psychometrika, 67*, 485–518.
18. Sijtsma, K., & Meijer, R. R. (2007). Nonparametric item response theory and related topics. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, vol. 26: Psychometrics* (pp. 719–746). Amsterdam: Elsevier.
19. The WHOQoL Group (1998). Development of the World Health Organisation WHOQOL-Bref QoL assessment. *Psychological Medicine, 28*, 551–559.
20. Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
21. Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. J. v. d. Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271–286). New York: Springer.

22. Mellenbergh, G. J. (1995). Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement, 19*, 91–100.
23. Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, Monograph supplement No. 17.
24. Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika, 62*, 331–347.
25. Van Engelenburg, G. (1997). *On psychometric models for polytomous items with ordered categories within the framework of item response theory*. Ph.D. Thesis, Amsterdam, The Netherlands: University of Amsterdam.
26. Samejima, F. (1997). Graded response model. In W. J. v. d. Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). New York: Springer.
27. Molenaar, I. W. (1997). Nonparametric models for polytomous responses. In W. J. v. d. Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory.* (pp. 369–380). New York: Springer.
28. Van der Ark, L. A. (2005). Stochastic ordering of the latent trait by the sum score under various polytomous IRT models. *Psychometrika, 70*, 283–304.
29. Junker, B. W., & Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement, 24*, 65–81.
30. Mokken, R. J., Lewis, C., & Sijtsma, K. (1986). Rejoinder to 'The Mokken scale: A critical discussion'. *Applied Psychological Measurement, 10*, 279–285.
31. Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*, 357–382.
32. Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.
33. Molenaar, I. W., & Sijtsma, K. (2000). *User's manual MSP5 for Windows*. Groningen, the Netherlands: iecPROGAMMA.
34. Ramsay, J. O. (2000). *Testgraf. A program for the analysis of multiple choice test and questionnaire data*. Montreal, Canada: Department of Psychology, McGill University.
35. Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56*, 611–630.
36. Fox, J. (1997). *Applied regression analysis, linear models, and related methods*. Thousand Oaks, CA: Sage.
37. Rossi, N., Wang, X., & Ramsay, J. O. (2002). Nonparametric item response function estimates with the EM algorititm. *Journal of Educational and Behavioral Statistics, 27*, 291–317.
38. Van Ginkel, J. R., & Van der Ark, L. A. (2005). SPSS syntax for missing value imputation in test and questionnaire data. *Applied Psychological Measurement, 29*, 152–153.
39. Reise, S. P., & Waller, N. G. (1990). Fitting the two-parameter model to personality data. *Applied Psychological Measurement, 14*, 45–58.
40. Hemker, B. T., Sijtsma, K., & Molenaar, I. W. (1995). Selection of unidimensional scales from a multidimensional itembank in the polytomous IRT model. *Applied Psychological Measurement, 19*, 337–352.
41. Thissen, D., Chen, W.-H., & Bock, R. D. (2003). *Multilog (version 7) [computer sotware]*. Lincolnwood, IL: Scientific Software International.
42. Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models*. Thousand Oaks, CA: Sage.
43. Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement, 30*, 298–321.
44. Allen, M.J., & Yen, W. M. (1979). *Introduction to measurement theory*. Belmont, CA: Wadsworth.
45. Hays, R.D., Brodsky, M., Johnston, M. F., Spritzer, K. L., & Hui, K. (2005). Evaluating the statistical significance of health-related quality of life change in individual patients. *Evaluation and the Health Professions, 28*, 160–171.
46. Hemker, B. T., Van der Ark, L. A., & Sijtsma, K. (2001). On measurement properties of continuation ratio models. *Psychometrika, 66*, 487–506.
47. Crane, P. K., Gibbons, L. E., Narasimhalu, K., Lai, J.-S., & Cella, D. (2007). Rapid detection of differential item functioning in assessments of health-related quality of life: The functional assessment of cancer therapy. *Quality of Life Research, 16*, 101–114.
48. Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2005). Global, local, and graphical person-fit analysis using person response functions. *Psychological Methods, 10*, 101–119.
49. Bjorner, J. B., Kosinski, M., & Ware, J. E. jr. (2003). Calibration of an item pool for assessing the burden of headaches: An application of item response theory to the Headache Impact Test (HIT^{TM}). *Quality of Life Research*, *12*, 913–933.
50. Bjorner, J. B., Kosinski, M., & Ware, J. E. jr. (2003). The feasibility of applying item response theory to measures of migraine impact: A re-analysis of three clinical studies. *Quality of Life Research, 12*, 887–902.
51. Fliege, H., Becker, J., Walter, O. B., Bjorner, J. B., Klapp, B. F., & Rose, M. (2005). Development of a computer-adaptive test for depression (d-cat). *Quality of Life Research, 14*, 2277–2291.
52. Lai, J.-S., Cella, D., Chang, C.-H., Bode, R. K., & Heinemann, A. W. (2003). Item banking to improve, shorten and computerize self-reported fatigue: An illustration of steps to create a core item bank from the facit-fatigue scale. *Quality of Life Research, 12*, 485–501.
53. Petersen, M. A., Groenvold, M., Aaronson, N., Fayers, P., Sprangers, M., Bjorner, J. B., et al. (2006). Multidimensional computerized adaptive testing of the EORTC QLQ-C30: Basic developments and evaluations. *Quality of Life Research, 15*, 315–329.
54. Schwartz, C., Welch, G., Santiago-Kelley, P., Bode, R., & Sun, X. (2006). Computerized adaptive testing of diabetes impact: A feasibility study of hispanics and non-hispanics in an active clinic population. *Quality of Life Research, 15*, 1503–1518.