Quality of Life Research **12**: 189–198, 2003.
© 2003 *Kluwer Academic Publishers. Printed in the Netherlands.*

Unidimensionality and reliability under Mokken scaling of the Dutch language version of the SF-36

P.G.M. van der Heijden[1,3], S. van Buuren[1], M. Fekkes[2], J. Radder[1] & E. Verrips[2]
[1]*Department of Statistics, TNO Prevention and Health, Leiden (E-mail: pgm.vanderheijden@pg.tno.nl);* [2]*Child Health Division, TNO Prevention and Health, Leiden;* [3]*Department of Methodology and Statistics, Utrecht University, The Netherlands*

Accepted in revised form 5 June 2002

Abstract

The sub-scales of the SF-36 in the Dutch National Study are investigated with respect to unidimensionality and reliability. It is argued that these properties deserve separate treatment. For unidimensionality we use a non-parametric model from item response theory, called the Mokken scaling model, and compute the corresponding scalability coefficients. We estimate reliability under the Mokken model, assuming that the items are double homogeneous, and compare it to Cronbach's α. The scalability of the sub-scale general health perceptions is medium ($H = 0.46$), and for the other sub-scales it is strong ($H \geqslant 0.6$). The reliability in terms of α indicates that all sub-scales can be used in basic research ($\alpha > 0.70$), but that only physical functioning can be used for clinical applications of quality of life ($\alpha > 0.90$). The relative merits of our approach are discussed.

Key words: Item response theory, Mokken scaling model, Reliability, Scalability, SF-36

Introduction

Appropriate measurement is fundamental to the field of quality of life (QoL). Many measurement instruments exist, and one of the most popular is the SF-36. In this paper we discuss two important properties of the SF-36, *unidimensionality* and *reliability*, from a psychometric perspective. Our aim is to highlight the role of both concepts in the construction of QoL-instruments.

Why are unidimensionality and reliability important concepts in the construction of QoL-instruments? QoL cannot be measured by devices based on laws of physics, like the sphygmomanometer for measuring blood pressure, but requires the use of one or more questionnaire items. Such items probe (the evaluation of) the health status of the respondent. Using only one item is ideal in the sense that the meaning of the answer will be more or less unambiguous, but at the same time it is also very crude, that is, it has a large measurement error. The measurement can be made more precise (more reliable) by administering more items and by then aggregating the responses to a sum score. The downside of this is that this can only be sensibly done if the items measure the same thing, that is, if they are unidimensional. Reliability without unidimensionality is only of limited value since we do not know *what* we are measuring. Similarly, unidimensionality without reliability is also of limited value because we may be measuring *too crudely*. Thus, unidimensionality and reliability are different concepts. Both play an equally important role in the construction of QoL-instruments. Maximising either one separately may go at the expense of the other, and generally there should be a balance between them. We now discuss the concepts in more detail.

A set of items is called unidimensional if they measure the same trait or property. This property is fundamental to almost any kind of measurement, and forms the key component of content validity. Unidimensionality does not specify the measurement level of the trait. Despite its importance there is no accepted and effective index of unidimensionality. Most authors seem to agree with McDonald that 'a set of items is unidimensional if and only if the set fits a (generally nonlinear) common factor model with just one common factor' [1]. A similar definition was given by Lord and Novick [2], who state that 'an item set is homogeneous if their true scores can be shown to be a monotonic function of some single variable'. Note that this definition uses the term homogeneous. Hattie [3] provides an excellent discussion of the confusion that is being generated by different interpretations of terms like *internal consistency* and *homogeneity*. To circumvent these difficulties, we will use the term unidimensionality throughout.

The reliability of a test is a value that quantifies the amount of measurement error of the total score of a test. Reliability is defined as the ratio of the true score variance to the observed score variance [4]. Under the assumptions that the items fit a linear common one-factor model and that the measurement error variance is equal across all true scores, the popular Cronbach's α gives a sound lower bound to the reliability of a total test score [5]. A well-known limitation of Cronbach's α is the fact that its value depends on the distribution of the true scores of the population. Thus, the coefficient of the reference population can be inadequate for the population to which the test is applied. Latent trait models, by contrast, offer a standard error for each possible test score.

An important problem is that the interpretation of Cronbach's α measure is sometimes extended beyond the use for which it is intended. Some researchers believe that Cronbach's α is a measure that can be used to assess the unidimensionality of a scale. This is fuelled by poorly defined concepts like 'homogeneity' and 'internal consistency', and by popular procedure like SPSS RELIABILITY that support methods to select a set of items that maximises α. It is not always realised that this may only be sensibly done after unidimensionality of the candidate items has been established, presumably by some form of factor analysis. Green et al. [6] demonstrated that α can be high even if the underlying items are clearly multidimensional. Furthermore, it is known that α increases if the number of items in a test is made larger by defining parallel items. If α would really be a measure of unidimensionality, then α should not change by including items that measure the same construct. The existence of some other problems with α leads Hattie to conclude that 'despite its common usage as an index of dimensionality, α is extremely suspect' [3, p. 145].

Other methods than α must be used to study the unidimensionality of a set of items. There are several alternatives that can be classified in three broad groups: linear common factor analysis, parametric item response models and non-parametric item response models.

In common factor analysis, one would be interested in selecting a subset of items that fits the linear common one-factor model. Fit statistics that measure the difference between the observed and the expected covariance matrix can be used to diagnose model fit. Alternatively, one could inspect the off-diagonal elements of residual covariance matrix for aberrant extreme values (see Hambleton et al. [7, p. 56] for other methods). Though easy to apply, the linear model is however not particularly suited to create QoL-instruments as it requires item scores to be measured on an interval scale. In addition, the relation between the latent trait and the observed scores are assumed to be linear. Both assumptions are not necessarily true in QoL research, and if the assumptions are not met, then factor analysis can come up with too many factors differentiating between items that in fact measure the same thing [8].

As an alternative, psychometricians developed a large variety of non-linear parametric one-factor models. Such models are known under names as the Rasch model (RM) the two- and three-parameter models (for binary data), the rating scale model, the graded response model, the partial credit model (for polytomous data). Common to these models is that they specify the probability of a particular category by means of logistic or normal ogive functions of the latent trait. The primary advantage is that responses in individual categories are fitted, thus obviating the need for interval measures. For some models, statistical tests have been developed to test the fit to the data [9, 10]. Such tests can be regarded as tests for unidimensionality, but are not widely used. A practical problem is that those tests are rather stringent. The

RM tends to be rejected for many real QoL data. For example, Raczek et al. [11] conducted a Rasch analysis on the 10 items of the PF sub-scale (PF-10), using data from seven countries. They conclude that 'unidimensionality of the PF scale is upheld' (p. 1201), but their Table 3 indicates that both items 'bending/kneeling' and 'bathing/dressing' do not fit the RM for all seven countries, and that item 'vigorous activities' does not fit in five countries. Other evidence [12] suggests that the PF-10 does not form a perfect hierarchy on a unidimensional scale. If it is so difficult to fit the RM on even the best sub-scale of the SF-36, then more problems can be expected for the other sub-scales. Within the framework of the RM possible solutions to this problem of misfit include: (1) reduce the item set to a more unidimensional subset, or (2) introduce additional flexibility by increasing the number of parameters of the item characteristic curve (ICC), as in the two- and three-parameter models. There is however no guarantee that such solutions work, so the problem may persist.

Consequently, applying less stringent models on QoL data could be useful. A third group of methods, the non-parametric item response model, is such a less stringent model that makes minimal assumptions. This group does not require a parametric functional form for the ICC, but allows for any monotonously increasing function. This approach is known as Mokken scaling [13, 14]. Since the shape of the ICC is much more flexible, Mokken scaling can fit data that will fail to fit parametric models. Mokken scaling can thus be used as a diagnostic tool to discover the shape of the ICC that is dictated by the data, thereby suggesting instances where the more parsimonious parametric models will inappropriately smooth out important characteristics of the data. More importantly, the technique may 'save' items that are not considered unidimensional under a parametric model, but that do fit the Mokken model. This could help to keep the item pool as large as possible, and thus save valuable work. Since the modelling assumptions of Mokken scaling are minimal, we believe that it exemplifies the purest form of unidimensionality.

Both unidimensionality and reliability are important issues for the assessment of usefulness of sub-scales of the SF-36. Therefore the paper has two purposes: (1) to evaluate the unidimensionality of the items in the sub-scales of Dutch language version of the SF-36 using the Mokken scaling model, and (2) to compare the results with the reliability as traditionally estimated by Cronbach's α.

Method

Data

We evaluate the unidimensionality and reliability of the sub-scales of the SF-36 for the so-called Dutch National Study [15]. We chose the SF-36 because this instrument is well established. Our method can be applied to any QoL-instrument, and our results on the SF-36 can act as a benchmark. The National Study is a nation-wide random sample of adults from the Netherlands ($n = 1742$). The data were collected in 1996. Ref. [15] provides descriptive statistics on the item and scale level. Overall the data quality in this study is high if evaluated in terms of percentage of missing data, multi-trait scaling structure and reliability of the sub-scales. Known-group comparisons yielded consistent support for the validity of the SF-36. If individuals have one or more items missing on a sub-scale, they are omitted from this sub-scale (listwise deletion of missing data). The percentage of respondents having one or more items missing was relatively small (ranging between 0 and 4%).

Mokken scaling model

We apply Mokken scaling to each sub-scale of the SF-36 to order the subjects as well as (the levels of) the items in terms of increasing health. Sijtsma [16] discusses many similarities and differences between the Mokken scaling model and the RM. Like the RM, the Mokken scaling model is a probabilistic version of the deterministic Guttman scale. The Guttman model forms the basis of many fit statistics in both the Mokken and Rasch models. In the sequel, we use the Mokken scaling model for polytomous items [17].

We now discuss the assumptions made in Mokken scaling in more detail. Consider a set of dichotomous health items that all aim to measure some quantitative latent trait, for example perceived health. Assume that each health item has two answers, 'yes' (positive about some particular health status) and 'no' (negative). The Mokken model has three assumptions: (1) the latent trait is unidimensional, (2) given the level of the latent trait, the answers on the items are independent, and (3) the ICC of an item is a non-decreasing function.

In Figure 1a a non-decreasing line (the ICC) describes the relation between the probability of a positive answer as a function of the latent trait. Figure 1b shows the ICCs of two items. It shows that lines for different items can be different in form and location. Because of the independence between the answers on the two items given the level of the latent trait, the latent trait explains the observed relation between the items. If these three assumptions hold, a set of items is called a monotonously homogeneous (MH) set. If MH holds, the sum score of the items order the individuals on the latent trait (see Ref. [16]). This has an important practical implication: If the MH assumption holds, the latent trait is a monotone function of the sum score.

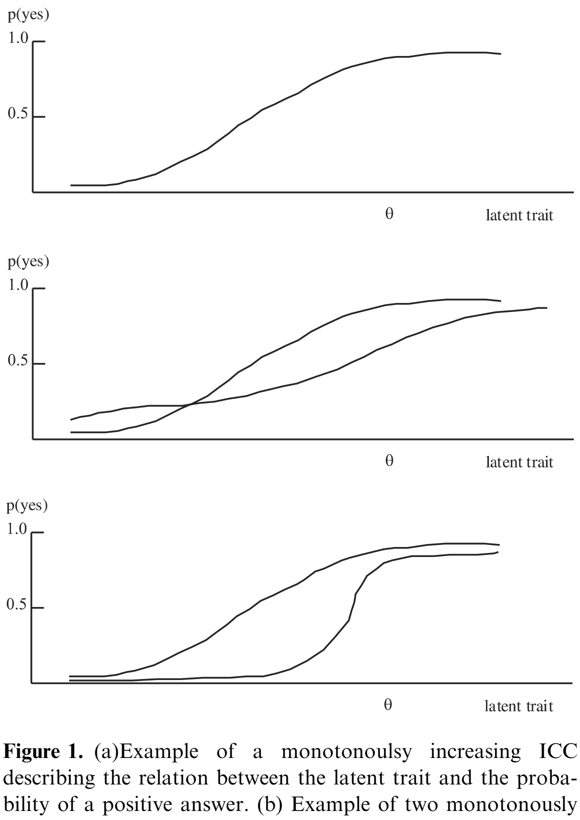

Figure 1. (a)Example of a monotonoulsy increasing ICC describing the relation between the latent trait and the probability of a positive answer. (b) Example of two monotonously increasing ICCs being a MH set. (c) Example of two monotonously increasing ICCs being a DM set.

It is possible to add a fourth assumption leading to a stronger Mokken model: (4) the ICCs may not intersect (see Figure 1c). If this assumption holds, the individuals also order the items, and the proportion of correct responses on a particular sample can be used to estimate the location of the item. We then speak of a doubly monotone (DM) set. The ordering of items is identical across the entire latent scale under the DM model, so the estimated item ordering will be independent of the particular population.

The Mokken scaling model for polytomous items

Thus far we discussed the Mokken scaling model for dichotomous items. However, many items of sub-scales of the SF-36 are polytomous. For polytomous items the Mokken model has been generalised by Molenaar [17]. Assume that each item has $m + 1$ ordered categories. An *item step* is an imaginary threshold between two adjacent categories of an item. Within an item, item steps are dependent, because if, for example, the step from category 1 to category 2 is not taken, then the step from 2 to 3 cannot be taken. By using the item steps it is possible to construct item step characteristic curves (ISCC's) describing the relation between the latent trait and the probability that an item step is taken. By definition ISCC's of one item cannot intersect. The MH assumption of the polytomous Mokken scaling model is that ISCC's are monotone non-decreasing functions. The DM assumption is that ISCC's of different items are not allowed to intersect. Evaluation tools exist to check whether a set of data has MH and DM properties (see Ref. [17, 19]). All analyses are performed with the computer program MSP 3.0 [19].

Check of model assumptions

The Mokken scaling model is a non-parametric model. To check whether the model fits the data, i.e. whether the assumptions underlying the Mokken model hold, certain characteristics of the data should be checked. The Mokken scaling model as it is implemented in MSP 3.0 [18] comes with a fairly elaborate set of diagnostic methods to assess the assumptions of the model. The 'scalability' and the 'single' method evaluate the MH properties, while the 'restscore' and 'pmatrix' methods evaluate DM properties. It is outside the scope of this paper to describe each of these methods in detail, and here we will only give a global indication of how the model properties are checked. We refer to the literature for a thorough description of each of these [13, 18, 19].

The so-called Guttman scalogram model plays a key role in checking the MH model via the scalability. Figure 2 illustrates the Guttman model. There are two items, where the item on the left is item g and the item on the right is item h. For each item one ICC is plotted, and for each of the items the probability of answering ‘yes’ is either zero or one. In this way there are only three admissible response patterns to the item pair, namely, from left to right (no, no), (yes, no) and (yes, yes). An answer (no, yes) is an observed error in terms of the Guttman scalogram model. Thus in the scalogram model some of the response patterns are not allowed. In the Mokken scaling model all response patterns are in principle allowed, but some have a lower probability of occurrence than others. To assess the fit of the Mokken scaling model, we measure the closeness of the solution to the perfect Guttman scale. It is possible to evaluate the scalability of each of the individual items as well as of the whole set of items using a coefficient known as Loevinger’s H [20]. If the ICCs of a set of items are non-decreasing, then $0 \leqslant H \leqslant 1$. It is equal to 1 if the items form a perfect Guttman scale, and it is equal to 0 if and only if all (or all minus one) items are constant functions of the latent trait [11]. Coefficient H can be interpreted as an index for the degree to which subjects can be accurately ordered by means of k items. Mokken [13, 14] speaks of weak scalability of a set of items when $0.3 \leqslant H < 0.4$, of medium scalability when $0.4 \leqslant H < 0.5$, and of strong scalability when $0.5 \leqslant H \leqslant 1.0$. A set of items is unscalable if $H < 0.3$. These guidelines for interpreting H seem coarse, but we could not find other guidelines that allow finer gradations and interpretations of scalability.

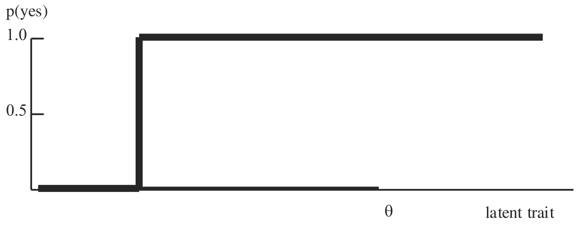

Figure 2. Example of two ICCs for items following a Guttman scale.

The ‘single’ method is the second method to check MH. The starting point is that, if MH holds, then a person falling in a higher level of an item will in general also fall in a higher level on the latent trait [17, 19]. In order to investigate this, certain aspects of the data can be checked. The total number of checks performed is denoted by the MSP-program as #ap. In each check it may turn out that the data do not follow MH. Violations are counted and referred to as #vi. Each violation separately can be tested for significance. However, such tests should be approached reluctantly and with caution, because, even if in a population the MH model would hold, by chance alone we would expect to find some violations, and if the number of pairs compared is large we even expect to find some of them to be significant.

Both methods to check DM compare, each in a different way, data on each of the possible pairs of items with data on the remaining items. Therefore it is only possible to use these methods when the number of items in a sub-scale is larger than two. The so-called ‘rest score’ method is the first method to check DM. It checks whether two ISCC’s of different items do not intersect. Again, in order to investigate this, certain aspects of the data can be checked. The total number of checks performed is denoted by the MSP-program as #rg. In each check it may turn out that the data do not follow DM. Here, violations are also counted and referred to as #vi. Again, each violation separately can be tested for significance, but such tests should be handled with care.

The so-called ‘P matrix’ method is the second method to check DM. It also concerns checking aspects in the data. The number of checks is denoted #ac, and the number of violations by #vi. Again, each violation separately can be tested for significance and interpreted cautiously.

Reliability measures

Cronbach’s α is computed using the computer program SPSS. For dichotomous items a reliability measure based on Mokken’s scaling model has been discussed in Refs. [16, 20] (for a generalization to polytomous items, see Ref. [17]). This reliability measure assumes that the DM assumption holds. This reliability measure is also provided by computer program MSP.

How large should α be? In psychometrics [21] as well as in medical statistics [22] values of 0.70 or 0.80 are considered to be sufficient for basic research, such as comparing groups and calculating correlations. However, for the clinical application (decisions about individuals) α's as high as 0.90 or 0.95 are required [17, 18].

Results

Table 1 summarises the results of the evaluation of the Mokken scaling. We start with a discussion of MH properties, i.e. whether the ISCC's are monotonously increasing. These are investigated by H and the single method. Using the guidelines for the interpretation of H (i.e. $0.4 \leqslant H < 0.5$ is medium scalability and $0.5 \leqslant H < 1.0$ is strong scalability), we find in column 1 that the fit of the MH model of all sub-scales is strong, except for General Health Perceptions where the fit of the MH model is medium.

Another way to check MH properties is by using the 'single' method, which assesses whether the response proportion per category increases with ability level [17, 19]. In the two columns with the heading single, the one with label #ap gives the number of checks made in the data, and the column with the label #vi gives the number of times this a violation of MH is found. It appears that no scales have violations, so there is strong evidence that all scales are monotonously increasing. This means that for each of the scales the persons can be ordered by the items of the scales.

For the DM properties of the data, there are two checks to see whether the ISCC's belonging to different items do not intersect, the 'P matrix' and 'rest score' methods. Since it is only possible to use these methods when the number of items in a sub-scale is larger than two, for the sub-scales Bodily Pain (BP) and Social Functioning (SF) we have no results. In the columns with heading P matrix, the one labelled '#ac' shows the number of checks made in the data, and the next column labelled #vi gives the number of violations of DM found in the data. For Role Limitations due to Physical Health problems (RP) and for Role Limitations due to Emotional Problems (RE) the number of active comparisons made is small, but they do not show any violation of DM. For the other scales some violations are found, but considering the large numbers of active comparisons made, the number of violations found is small (for example, for Physical Functioning (PF) 2880 active comparisons are made and only six show a violation of the model; such a small number could be due to chance, even if the model were true). However, we will study them on item level when we discuss Table 2.

Table 1. Overview of scales

Scale	H	Single		P matrix		Rest score		Reliability	α
		#ap	#vi	#ac	#vi	#rg	#vi		
PF	0.75	47	0	2880	6	396	5	0.95	0.93
RP	0.71	4	0	12	0	6	0	0.88	0.88
BP	0.84	15	0					0.88	0.85
GH	0.46	114	0	1920	22	368	16	0.78	0.77
VT	0.60	90	0	1500	3	325	10	0.83	0.83
SF	0.77	15	0					0.84	0.82
RE	0.71	3	0	3	0	1	1	0.82	0.83
MH	0.61	100	0	3750	12	600	6	0.87	0.86

Scales are Physical Functioning (PF, n = 1675), Role limitations due to Physical Health problems (RP, n = 1703), Bodily Pain (BP, n = 1735), General Health Perceptions (GH, n = 1664), Vitality (VT, n = 1723), Social Functioning (SF, n = 1738), Role Limitations due to Emotional Problems (RE, n = 1708) and General Mental Health (MHe, n = 1705). MH can be checked by scalability H and single method, DM by P matrix and rest score. In single method, #ap is the number of active pairs of items where MH properties of the data are checked, and #vi is the number of violations of MH found in these active pairs. In P matrix method, #ac is the number of active comparisons carried out to check the DM properties of the data, and #vi is the number of violations of DM found in these comparisons. In rest score method #rg is the number of rest group comparisons carried out to check the DM properties of the data, and #vi is the number of violations of DM found in these comparisons.

For the ‘rest score’ method the number of checks made is found in the column labelled #rg, and this column is followed by the column with the number of violations of DM, #vi. Again, RP is doing well. RE is doing a worse, but additional information in the output shows that this violation is only small: the z-value for the violation between the first and third item is 2.22. The other sub-scales also show violations, but considering the number of rest group comparisons carried out, the number of violations in the data is small. We get more insight into these violations when we discuss Table 2.

Table 2 shows the same results at the item level. When we consider MH, we find again that the

Table 2. Mokken scale properties and violations of individual SF-36 items

Scale	Item	*H*	Single		P matrix		Rest score	
			#ap	#vi	#ac	#vi	#rg	#vi
PF	Vigorous activities	0.77	6	0	576	1	72	0
	Moderate activities	0.75	4	0	576	3	80	3
	Lifting/carrying	0.72	4	0	576	2	84	2
	Climbing several	0.73	4	0	576	1	80	1
	Climbing one	0.81	4	0	576	0	76	0
	Bending/kneeling	0.69	12	0	576	0	92	0
	Walking > 1mile	0.73	6	0	576	4	80	4
	Walking blocks	0.79	3	0	576	1	76	0
	Walking block	0.80	1	0	576	0	76	0
	Bathing/dressing	0.72	3	0	576	0	76	0
RP	Amount of time	0.75	1	0	6	0	3	0
	Accomplishes less	0.69	1	0	6	0	3	0
	Kind of work	0.73	1	0	6	0	3	0
	Difficulty	0.70	1	0	6	0	3	0
BP	Had BP	0.84	5	0				
	Did pain interfere	0.84	10	0				
GH	General health is…	0.54	18	0	768	10	160	6
	Sick easier th. oth.	0.44	24	0	768	10	144	4
	As healthy as	0.40	24	0	768	11	160	12
	Expect get worse	0.40	24	0	768	2	144	3
	Is excellent	0.53	24	0	768	11	128	7
VT	Full of pep	0.52	30	0	750	3	175	10
	Lot of Energy	0.62	30	0	750	0	150	3
	Worn out	0.60	15	0	750	1	150	1
	Feel tired	0.63	15	0	750	2	175	6
SF	Normal social activ.	0.77	11	0				
	Social activities	0.77	4	0				
RE	Amount of time	0.73	1	0	2	0	1	1
	Accomplished less	0.75	1	0	2	0	0	
	Not as carefully	0.64	1	0	2	0	1	1
MHe	Nervous person	0.51	13	0	1500	8	225	5
	Nothing cheers up	0.65	28	0	1500	3	225	1
	Calm and peaceful	0.62	30	0	1500	5	250	1
	Downhearted, blue	0.67	14	0	1500	4	275	3
	Happy person	0.60	15	0	1500	4	225	2

MH can be checked by scalability H and single method, DM by P matrix and rest score. In single method, #ap is the number of active pairs of items where MH properties of the data are checked, and #vi is the number of violations of MH found in these active pairs. In P matrix method, #ac is the number of active comparisons carried out to check the DM properties of the data, and #vi is the number of violations of DM found in these comparisons. In rest score method #rg is the number of rest group comparisons carried out to check the DM properties of the data, and #vi is the number of violations of DM found in these comparisons.

sub-scales contain no items that have a much lower *H*-value than the other items, except perhaps for the removal of item 'Did you feel full of pep?' from the sub-scale Vitality and 'Have you been a very nervous person?' from General Mental Health (MHe). But removals are not necessary since *H* for the total sub-scale is larger than 0.5, and this is to be interpreted as good MH properties.

For DM properties this is a bit more complicated, because violations are always related to item step pairs. It follows that many of the violations would disappear if we remove the following items: item 'I am as healthy as anybody I know' from General Health Perceptions (11 out of the 12 rest score violations are significant), item 'Did you feel full of pep?' from Vitality (9 out of the 10 rest score violations are significant), and item 'Have you been a very nervous person?' from MHe (3 out of the 5 rest score violations are significant). A close inspection of the violations of these three items reveals that, for the item 'Did you feel full of pep?' as well as for the item 'Have you been a very nervous person?' most violations are found for the item steps from category 'A good bit of the time' to 'Some of the time' and from category 'Some of the time' to 'A little of the time'. For the item 'I am as healthy as anybody I know' the violations are found in all item steps.

However, it seems better to keep those variables in the sub-scales, because the number of violations is relatively small compared to the very large number of comparisons made. For future research into different populations it deserves attention that these items threaten the DM properties of these sub-scales, and it is of interest to see if the same violations will be found.

We now discuss reliability measures for the sub-scales (see the last two columns of Table 1). Cronbach's α is interpreted as a lower bound for reliability, and it is clear that the lower bounds are quite high, in any case high enough to use each of the scales for basic research purposes in a population comparable to the one we have in this study. The α for PF indicates that this scale can be used for clinical applications, and the α's of the scales RP, BP and MH are close to the range where this usage is allowed. The reliability measure derived under the assumption that the items form a DM set are very close to the estimates of Cronbach's α estimates. Since DM holds, the first reliability measure is preferable because Cronbach's α gives a lower bound and depends on untested linearity assumptions. Note also that the reliability estimates vary independently from the scalability coefficients.

Conclusion

The results make clear that the reliability of the scales is good enough to use all scales in basic research, being around between 0.78 and 0.95, with the lowest reliability for perceived general health. Only PF can be used in clinical applications in populations similar to the one used in this study (α is larger than 0.90), and some of the other scales are very close to this situation.

The scalability of the items is very good: monotone homogeneity is never violated for any of the sub-scales, and *H*-coefficients are almost always larger than 0.7, with again an exception for Perceived General Health, were $H = 0.46$, and Vitality ($H = 0.60$) and MH ($H = 0.61$). Double monotonicity is also good to very good: there is only a very limited number of violations given the enormous amount of comparisons that are made, and it is not unlikely that these violations are due to sample fluctuations. We conclude that the scalability of the items of the sub-scales assessed with the Mokken scaling model is good, both in terms of MH and DM, and no important gains will result by omitting any of the items in the sub-scales.

We conclude that besides good reliability the SF-36 sub-scales also show good scalability/unidimensionality. These results will hold in populations with a health distribution similar to the one in the National study, but might differ somewhat from populations with different health distributions. For example, there appear to be some ceiling effects in the data. In a population where health is worse, such ceiling effects will diminish. Since ceiling effects do not allow for violations of the Mokken scale properties, it might be that the Mokken scale properties of the SF-36 are worse in a population with a worse health. This remains a topic of further study.

Discussion

Unidimensionality and reliability are different concepts. Unidimensionality refers to the question whether items measure the same thing, whereas reliability indicates the variation of the test results if it is repeated under comparable circumstances. In scale construction both aspects need to be assessed, but the limitations of the traditional approach based on common factor analysis and Cronbach's α are well known. Several authors have warned against the inappropriate interpretation of Cronbach's α as a measure for unidimensionality [1, 3, 6]. The present paper investigates a particular alternative, the Mokken scaling model, for assessing unidimensionality and reliability. The Mokken scaling model tries to get the most out the set of the weakest possible assumptions. It is the general item response model for ordering persons according to a sum score. The model can be used to discover the shape of the ICC, and may 'save' items that are not unidimensional according to parametric models. Since the modelling assumptions of Mokken scaling are minimal, it is in some sense the purest form of unidimensionality.

The Mokken approach determines the unidimensionality of a set of items in terms of their scalability. Mokken [13] suggested a number of cut-off values as criteria for scalability. Though such guidance it useful, the exact choice of these values remains of course somewhat arbitrary. Hattie [3, p. 143] argued that scalability is a limited operationalisation of unidimensionality. A serious objection is that these methods can only achieve their upper bounds if the strong assumption of scalability (i.e. a perfect scale) is made. Another criticism is that 'there is nothing in the methods that enables a test of just one trait to be distinguished from a test composed of an equally weighted composite' [3, p. 143]. This criticism is related to the old-standing debate about general vs. specific ability factors. For these and other reasons, Hattie [3] prefers to use the principle of local stochastic independence from a one-factor model as the best way to define unidimensionality. Local independence is one of the key assumptions in Mokken scaling. Some work has been done to actually test the unidimensionality assumption in non-parametric models [23, 24]. This approach has not (yet) been implemented in MSP, but it could potentially be useful in removing some of the arbitrariness of the assessment of unidimensionality.

Similar remarks can be made about the number of violations and the number of active comparison that are needed for testing the DM assumption. For most of the scales, the number of violation is small relative to the number of active comparisons. There is however no statistical test to accompany this statement, and the conclusion remains therefore somewhat arbitrary. More work in this area could be useful.

Within the family of IRT models, the RM is best known. The RM is special in the sense that it possesses statistical properties like sufficiency, separability and specific objectivity [25]. It has been argued that the model can therefore be viewed as a fundamental model of measurement. This implies that if the data fit the model, then the measurements have interval scale properties. Now how does this compare to the non-parametric Mokken model? Both the Mokken and Rasch model are generalisations of the Guttman scaling model, but the Mokken scale makes weaker assumptions than the RM. The Mokken model orders items and persons, and thus produces a scale that has ordinal scale properties at most. Though the Mokken model is more general and will fit the data more often than the RM, this comes at the price that the property of invariant comparison is lost. This is probably not so much of a problem if the test is primarily used for diagnostic purposes, where a specific subsample beyond the cut-off value is referred for further evaluation. The interpretation of difference scores is however complicated by the fact that intervals between different points of an ordinal scale cannot be formally compared. In such a case, the Mokken model can still be useful though to identify the shape of ICC's that are to be fitted within a parametric framework. Last, we also note that the DM property of the data cannot be investigated in the Mokken model, whereas there is no problem to assess the fit of the RM in this case.

If the DM assumption fails, the difficulty order of the items is different for different people. What does that mean for comparing scale scores between people? The most important consequence is that the appropriateness of using the sum score is not theoretically warranted in the polytomous case. If all items are binary, we may validly use the sum score to compare people, as the ranks of the sum score and the latent ability are identical (except for ties). This property holds under both HM and DM models, but it is not true for polytomous items

[26]. In the latter case, we can use the ability estimates provided by the Mokken model as an alternative, but this is less convenient than computing a sum score. Sijtsma and van der Ark conducted a small simulation study about the adequacy of using the sum score as a proxy for the latent ability order. This was done for the polytomous HM model, that is, the model where the largest discrepancies can be expected. The results 'tentatively suggest that, in practice, the use of the sum score does not lead to serious errors when ordering respondents on the latent ability' [26, p. 309]. So, irrespective of the whether the DM property holds, the advice is to use to sum score for ranking respondents on the latent trait.

In conclusion, Cronbach's α is not a measure of unidimensionality, and other methods are needed to assess unidimensionality of the set of items. This seems especially relevant for a maturing field like QoL, where the concepts are still being defined and refined. We hope that the methods used in this paper will stimulate such work.

References

1. McDonald RP. The dimensionality of tests and items. Brit J Math Stat Psychol 1981; 34: 100–117.
2. Lord FM, Novick MR. Statistical Theories of Mental Test Scores. Reading, Massachusetts: Addison-Wesley, 1968.
3. Hattie J. Methodology review: Assessing unidimensionality of tests and items. Appl Psych Meas 1985; 9: 139–164.
4. Novick MR, Lewis C. Coefficient α and the reliability of composite measurements. Psychometrika 1967; 32: 1–13.
5. Cronbach LJ. Coefficient α and the internal structure of tests. Psychometrika 1951; 16: 297–334.
6. Green SB, Lissitz RW, Mulaik SA. Limitations of coefficient α as an index of test unidimensionality. Educ Psychol Meas 1977; 37: 827–838.
7. Hambleton RK, Swaminathan H, Rogers HJ. Fundamental of Item Response Theory. Newbury Park: Sage, 1991.
8. McDonald RP, Ahlawat KS. Difficulty factors in binary data. Brit J Math Stat Psychol 1974; 27: 82–99.
9. Glas CAW, Verhelst ND. Testing the Rasch model. In: Fischer GH, Molenaar, IW (eds), Rasch Models, Berlin: Springer Verlag, 1995; 69–95.
10. Glas CAW, Verhelst ND. Tests of fit for polytomous Rasch models. In: Fischer GH, Molenaar IW (eds), Rasch Models, Berlin: Springer Verlag, 1995; 325–352.
11. Raczek AE, Ware JE, Bjorner JB, et al. 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. J Clin Epidemiol 1998; 51(11): 1203–1214.
12. Jenkinson C, Fitzpatrick R, Garratt A, Peto V, Stewart-Brown S. Can item response theory reduce patient burden when measuring health status in neurological disorders? Results from Rasch analysis of the SF-36 physical functioning scale (PF-10). J Neurol Neurosur Psychiat 2001; 71(2): 220–224.
13. Mokken RJ. Theory and Procedure of Scale Analysis. Den Haag: Mouton; Berlin: De Gruijter, 1971.
14. Mokken RJ. Nonparametric models for dichotomous responses. In: van der Linden WJ, Hambleton RK (eds), Handbook of Modern Item Response Theory, New York: Springer; 1997; 351–367.
15. Aaronson NK, Muller M, Cohen PDA, et al. Translation, validation and norming of the Dutch language version of the SF-36 health survey in community and Chronic disease populations. J Clin Epidemiol 1998; 51: 1055–1068.
16. Sijtsma K. Methodology review: Nonparametric IRT approaches to the analysis of dichotomous item scores. Appl Psychol Meas 1998; 22: 3–31.
17. Molenaar IW. Nonparametric models for polytomous responses. In: van der Linden WJ, Hambleton RK (eds), Handbook of Modern Item Response Theory, New York: Springer, 1997; 369–380.
18. Mokken RJ, Lewis C. A nonparametric approach to the analysis of dichotomous item responses. Appl Psychol Meas 1982; 6: 417–430.
19. Molenaar IW, Debets P, Sijtsma K, Hemker BT. MSP: A Program for Mokken Scale Analysis for Polytomous Items. Version 3.0. Groningen: iecProGAM-MA, 1994.
20. Meijer RR, Sijtsma K, Molenaar IW. Reliability estimation for single dichotomous tems based on Mokken's IRT model. Appl Psychol Meas 1995; 14: 283–298.
21. Nunnally JC, Bernstein IH. Psychometric Theory. 3rd ed. New York: McGraw Hill, 1994.
22. Bland JM, Altman DG. Cronbach's α (Statistics notes). Brit Med J 1997; 314: 572.
23. Stout W. A nonparametric approach for assessing latent trait unidimensionality. Psychometrika 1987; 52: 589–617.
24. Stout W, Goodwin Froelich A, Gao F. Using resampling methods to produce an improved DIMTEST procedure. In: Boomsma A, Van Duijn MAJ, Snijders TAB (eds), Essays of Item Response Theory, Berlin: Springer Verlag, 2001; 357–375.
25. Rost J. The growing family of Rasch models. In: Boomsma A, Van Duijn MAJ, Snijders TAB (eds), Essays of Item Response Theory, Berlin: Springer Verlag, 2001; 25–37.
26. Sijtsma K, van der Ark LA. Progress in NIRT analysis of polytomous item scores: Dilemmas and practical solutions. In: Boomsma A, Van Duijn MAJ, Snijders TAB (eds), Essays of Item Response Theory, Berlin: Springer Verlag, 2001; 297–318.

Address for correspondence: P.G.M. van der Heijden, Department of Statistics, TNO Prevention and Health, Postbus 2215, 2301 CE Leiden, The Netherlands
Phone: +31-71-5181844; Fax: +31-71-5181920
E-mail: pgm.vanderheijden@pg.tno.nl