The Stata Journal (2011)
11, Number 1, pp. 30–51

Nonparametric item response theory using Stata

Jean-Benoit Hardouin
University of Nantes
Faculty of Pharmaceutical Sciences
Biostatistics, Clinical Research, and Subjective Measures in Health Sciences
Nantes, France
jean-benoit.hardouin@univ-nantes.fr

Angélique Bonnaud-Antignac
University of Nantes
Faculty of Medicine
ERT A0901 ERSSCA
Nantes, France

Véronique Sébille
University of Nantes
Faculty of Pharmaceutical Sciences
Biostatistics, Clinical Research, and Subjective Measures in Health Sciences
Nantes, France

Abstract. Item response theory is a set of models and methods allowing for the analysis of binary or ordinal variables (items) that are influenced by a latent variable or latent trait—that is, a variable that cannot be measured directly. The theory was originally developed in educational assessment but has many other applications in clinical research, ecology, psychiatry, and economics.

The Mokken scales have been described by Mokken (1971, *A Theory and Procedure of Scale Analysis* [De Gruyter]). They are composed of items that satisfy the three fundamental assumptions of item response theory: unidimensionality, monotonicity, and local independence. They can be considered nonparametric models in item response theory. Traces of the items and Loevinger's H coefficients are particularly useful indexes for checking whether a set of items constitutes a Mokken scale.

However, these indexes are not available in general statistical packages. We introduce Stata commands to compute them. We also describe the options available and provide examples of output.

Keywords: st0216, tracelines, loevh, gengroup, msp, items trace lines, Mokken scales, item response theory, Loevinger coefficients, Guttman errors

1 Introduction

Item response theory (IRT) (Van der Linden and Hambleton 1997) concerns models and methods where the responses to the items (binary or ordinal variables) of a questionnaire are assumed to depend on nonmeasurable characteristics (latent traits) of the respondents. These models can be applied to measures such as a latent variable (in measurement models) or to investigate influences of covariates on these latent variables.

Examples of latent traits include health status; quality of life; ability or content knowledge in a specific field of study; and psychological traits such as anxiety, impulsivity, and depression.

Most item response models (IRMs) are parametric: they model the probability of response at each category of each item by a function, depending on the latent trait, which is typically considered as a set of fixed effects or as a random variable, and they model the probability of parameters characterizing the items. The most popular IRMs for dichotomous items are the Rasch model and the Birnbaum model, and the most popular IRMs for polytomous items are the partial credit model and the rating scale model. These IRMs are already described for the Stata software (Hardouin 2007; Zheng and Rabe-Hesketh 2007).

Mokken (1971) defines a nonparametric model for studying the properties of a set of items in the framework of IRT. Mokken calls this model the monotonely homogeneous model, but it is generally referred to as the Mokken model. This model is implemented in a stand-alone package for the Mokken scale procedure (MSP) (Molenaar, Sijtsma, and Boer 2000), and codes already have been developed in Stata (Weesie 1999), SAS (Hardouin 2002), and R (Van der Ark 2007) languages. We propose commands under Stata to study the fit of a set of items to a Mokken model. These commands are more complete than the `mokken` command of Jeroen Weesie, for example, which does not offer the possibility of analyzing polytomous items.

The main purpose of the Mokken model is to validate an ordinal measure of a latent variable: for items that satisfy the criteria of the Mokken model, the sum of the responses across items can be used to rank respondents on the latent trait (Hemker et al. 1997; Sijtsma and Molenaar 2002). Compared with parametric IRT models, the Mokken model requires few assumptions regarding the relationship between the latent trait and the responses to the items; thus it generally allows keeping more important items. As a consequence, the ordering of individuals is more precise (Sijtsma and Molenaar 2002).

2 The Mokken scales

2.1 Notation

In the following text, we use the following notation:

- X_j is the random variable (item) representing the responses to the jth item, $j = 1, \ldots, J$.

- X_{nj} is the random variable (item) representing the responses to the jth item, $j = 1, \ldots, J$, for the nth individual, and x_{nj} is the realization of this variable.
- $m_j + 1$ is the number of response categories of the jth item.
- The response category 0 implies the smallest level of the latent trait and is referred to as a negative response, whereas the m_j nonzero response categories $(1, 2, \ldots, m_j)$ increase with increasing levels of the latent trait and are referred to as positive responses.
- M is the total number of possible positive responses across all items: $M = \sum_{j=1}^{J} m_j$
- Y_{jr} is the random-threshold dichotomous item taking the value 1 if $x_{nj} \geq r$ and 0 otherwise. There are M such items ($j = 1, \ldots, J$ and $r = 1, \ldots, m_j$).
- $P(.)$ refers to observed proportions.

2.2 Monotonely homogeneous model of Mokken (MHMM)

The Mokken scales are sets of items satisfying an MHMM (Mokken 1997; Molenaar 1997; Sijtsma and Molenaar 2002). This kind of model is a nonparametric IRM defined by the three fundamental assumptions of IRT:

- unidimensionality (responses to items are explained by a common latent trait)
- local independence (conditional on the latent trait, responses to items are independent)
- monotonicity (the probability of an item response greater than or equal to any fixed value is a nondecreasing function of the latent trait)

Unidimensionality implies that the responses to all the items are governed by a scalar latent trait. A practical advantage of this assumption is the easiness of interpreting the results. For a questionnaire aiming at measuring several latent traits, such an analysis must be realized for each unidimensional latent trait.

Local independence implies that all the relationships between the items are explained by the latent trait (Sijtsma and Molenaar 2002). This assumption is strongly related to the unidimensionality assumption, even if unidimensionality and local independence do not imply one another (Sijtsma and Molenaar 2002). As a consequence, local independence implies that a strong redundancy among the items does not exist.

Monotonicity is notably a fundamental assumption that allows validating the score as an ordinal measure of the latent trait.

2.3 Traces of the items

Traces of items can be used to check the monotonicity assumption. We define the score for each individual as the sum of the individual's responses ($S_n = \sum_{j=1}^{J} X_{nj}$). This score is assumed to represent an ordinal measure of the latent trait. The trace of a dichotomous item represents the proportion of positive responses $\{P(X_j = 1)\}$ as a function of the score. If the monotonicity assumption is satisfied, the trace lines increase. This means that the higher the latent trait, the more frequent the positive responses. In education sciences, if we wish to measure a given ability, this means that a good student will have more correct responses to the items. In health sciences, if we seek to measure a dysfunction through the presence of symptoms, this means that a patient having a high level of dysfunction will display more symptoms. An example trace is given in figure 1.

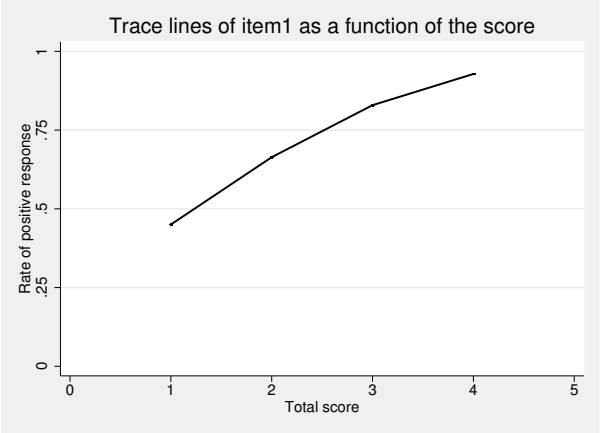

Figure 1. Trace of a dichotomous item as a function of the score

The score and the proportion of positive responses to each item are generally positively correlated, because the score is a function of all the items. This phenomenon can be strong, notably if there are few items in the questionnaire. To avoid the phenomenon, the rest-score (computed as the score of all the other items) is more generally used.

For polytomous items, we represent the proportion of responses to each response category $\{P(X_j = r)\}$ as a function of the score or of the rest-score (an example is given in figure 2).

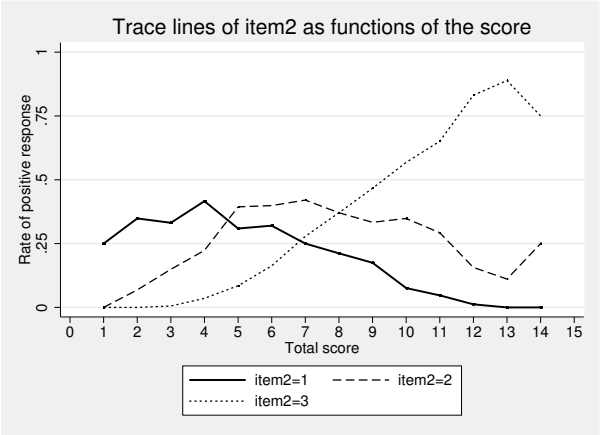

Figure 2. Traces of a polytomous item as functions of the score

Unfortunately, these trace lines are difficult to interpret, because an individual with a moderate score will preferably respond to medium response categories, and an individual with high scores will respond to high response categories, so the trace lines corresponding to each response category do not increase. Cumulative trace lines represent the proportions $P(Y_{jr} = 1) = P(X_j \geq r)$ as a function of the score or of the rest-score. If the monotonicity assumption is satisfied, these trace lines increase. An example is given in figure 3.

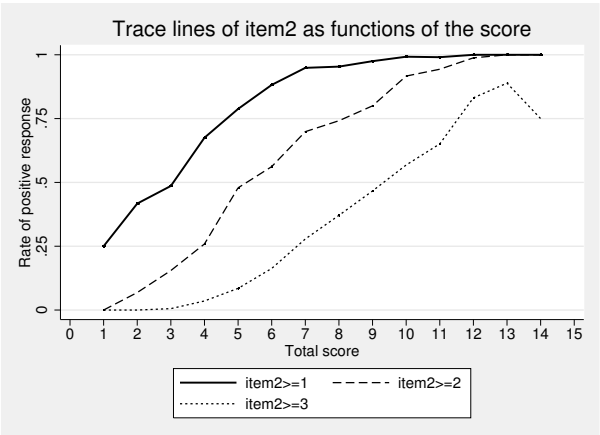

Figure 3. Cumulative trace lines of a polytomous item as functions of the score

2.4 The Guttman errors

Dichotomous case

The difficulty of an item can be defined as its proportion of negative responses. The Guttman errors (Guttman 1944) for a pair of dichotomous items are the number of individuals having a positive response to the more difficult item and a negative response to the easiest item. In education sciences, this represents the number of individuals who correctly responded to a given item but incorrectly responded to an easier item. In health sciences, this represents the number of individuals who present a given symptom but do not present a more common symptom.

We define the two-way tables of frequency counts between the items j and k as

		Item j		
		0	1	
Item k	0	a_{jk}	b_{jk}	$a_{jk}+b_{jk}$
	1	c_{jk}	d_{jk}	$c_{jk}+d_{jk}$
		$a_{jk}+c_{jk}$	$b_{jk}+d_{jk}$	N_{jk}

N_{jk} is the number of individuals with nonmissing responses to the items j and k.

An item j is easier than the item k if $P(X_j = 1) > P(X_k = 1)$—that is to say, if $(b_{jk} + d_{jk}/N_{jk}) > (c_{jk} + d_{jk}/N_{jk})$ (equivalently, if $b_{jk} > c_{jk}$), and the number of Guttman errors e_{jk} in this case is $e_{jk} = N_{jk} \times P(X_j = 0, X_k = 1) = c_{jk}$. More generally, if we ignore the easier item between j and k,

$$e_{jk} = N_{jk} \times \min\left\{P(X_j = 0, X_k = 1), P(X_j = 1, X_k = 0)\right\} = \min\left(b_{jk}, c_{jk}\right) \quad (1)$$

$e_{jk}^{(0)}$ is the number of Guttman errors under the assumption of independence of the responses to the two items:

$$\begin{aligned} e_{jk}^{(0)} =& N_{jk} \times \min\left\{P(X_j = 0) \times P(X_k = 1), P(X_j = 1) \times P(X_k = 0)\right\} \\ =& \frac{\left(a_{jk} + e_{jk}\right)\left(e_{jk} + d_{jk}\right)}{N_{jk}} \end{aligned}$$

Polytomous case

The Guttman errors between two given response categories r and s of the pair of polytomous items j and k are defined as

$$\begin{aligned} e_{j(r)k(s)} &= N_{jk} \times \min\left\{P(X_j \geq r, X_k < s), P(X_j < r, X_k \geq s)\right\} \\ &= N_{jk} \times \min\left\{P(Y_{jr} = 1, Y_{ks} = 0), P(Y_{jr} = 0, Y_{ks} = 1)\right\} \end{aligned}$$

The number of Guttman errors between the two items is

$$e_{jk} = \sum_{r=1}^{m_j} \sum_{s=1}^{m_k} e_{j(r)k(s)}$$

If $m_j = m_k = 1$ (the dichotomous case), this formula is equivalent to (1).

Under the assumption of independence between the responses to these two items, we have

$$e^{(0)}_{j(r)k(s)} = N_{jk} \times P(X_j < r)P(X_k \geq s) = N_{jk} \times P(Y_{jr} = 0)P(Y_{ks} = 1)$$

if $P(X_j \geq r) > P(X_k \geq s)$ and

$$e^{(0)}_{jk} = \sum_{r=1}^{m_j} \sum_{s=1}^{m_k} e^{(0)}_{j(r)k(s)}$$

2.5 The Loevinger's H coefficients

Loevinger (1948) proposed three indexes that can be defined as functions of the Guttman errors between the items.

The Loevinger's H coefficient between two items

H_{jk} is the Loevinger's H coefficient between the items j and k:

$$H_{jk} = 1 - \frac{e_{jk}}{e^{(0)}_{jk}}$$

We have $H_{jk} \leq 1$ with $H_{jk} = 1$ only if there is no Guttman error between the items j and k. If this coefficient is close to 1, there are few Guttman errors, and so the two items probably measure the same latent trait. An index close to 0 signifies that the responses to the two items are independent, and therefore reveals that the two items probably do not measure the same latent trait. A significantly negative value to this index is not expected, and it can be a flag that one or more items have been incorrectly coded or are incorrectly understood by the respondents.

We can test H_0: $H_{jk} = 0$ (against H_a: $H_{jk} > 0$). Under the null hypothesis, the statistic

$$Z = \frac{\mathrm{Cov}(X_j, X_k)}{\sqrt{\frac{\mathrm{Var}(X_j)\mathrm{Var}(X_k)}{N_{jk}-1}}} = \rho_{jk}\sqrt{N_{jk} - 1} \tag{2}$$

follows a standard normal distribution, where ρ_{jk} is the correlation coefficient between items j and k.

The Loevinger's H coefficient measuring the consistency of an item within a scale

Let S be a set of items (a scale), and let j be an item that belongs to this scale ($j \in S$). H_j^S is the Loevinger's H coefficient that measures the consistency of the item j within a scale S.

$$H_j^S = 1 - \frac{e_j^S}{e_j^{S(0)}} = 1 - \frac{\sum_{k \in S,\ k \neq j} e_{jk}}{\sum_{k \in S,\ k \neq j} e_{jk}^{(0)}}$$

If the scale S is a good scale (that is, if it satisfies an MHMM, for example), this index is close to 1 if the item j has a good consistency within the scale S, and this index is close to 0 if it has a bad consistency within this scale.

It is possible to test H_0: $H_j^S = 0$ (against H_a: $H_j^S > 0$). Under the null hypothesis, the statistic

$$Z = \frac{\sum_{k \in S, k \neq j} \mathrm{Cov}(X_j, X_k)}{\sqrt{\sum_{k \in S, k \neq j} \frac{\mathrm{Var}(X_j)\mathrm{Var}(X_k)}{N_{jk}-1}}} \tag{3}$$

follows a standard normal distribution.

The Loevinger's H coefficient of scalability

If S is a set of items, we can compute the Loevinger's H coefficient of scalability of this scale.

$$H^S = 1 - \frac{e^S}{e^{S(0)}} = 1 - \frac{\sum_{j \in S} \sum_{k \in S,\ k > j} e_{jk}}{\sum_{j \in S} \sum_{k \in S,\ k > j} e_{jk}^{(0)}}$$

We have $H^S \geq \min_{j \in S} H_j^S$. If H^S is near 1, then the scale S has good scale properties; if H^S is near 0, then it has bad scale properties.

It is possible to test H_0: $H^S = 0$ (against H_a: $H^S > 0$). Under the null hypothesis, the statistic

$$Z = \frac{\sum_{j \in S} \sum_{k \in S, k \neq j} \mathrm{Cov}(X_j, X_k)}{\sqrt{\sum_{j \in S} \sum_{k \in S, k \neq j} \frac{\mathrm{Var}(X_j)\mathrm{Var}(X_k)}{N_{jk}-1}}} \tag{4}$$

follows a standard normal distribution.

In the MSP software (Molenaar, Sijtsma, and Boer 2000), the z statistics defined in (2), (3), and (4) are approximated by dividing the variances by N_{jk} instead of by $N_{jk} - 1$.

2.6 The fit of a Mokken scale to a dataset

Link between the Loevinger's H coefficient and the Mokken scales

Mokken (1971) showed that if a scale S is a Mokken scale, then $H^S > 0$, but the converse is not true. He proposes the following classification:

- If $H^S < 0.3$, the scale S has poor scalability properties.
- If $0.3 \leq H^S < 0.4$, the scale S is "weak".
- If $0.4 \leq H^S < 0.5$, the scale S is "medium".
- If $0.5 \leq H^S$, the scale S is "strong".

So Mokken (1971) suggests using the Loevinger's H coefficient to build scales that satisfy a Mokken scale. He suggests that there is a threshold $c > 0.3$ such that if $H^S > c$, then the scale S satisfies a Mokken scale. This idea is used by Mokken (1971) and is adapted by Hemker, Sijtsma, and Molenaar (1995) to propose the MSP or automated item selection procedure (AISP) (Sijtsma and Molenaar 2002).

Moreover, the fit of the Mokken scale is satisfactory if $H_j^S > c$ and $H_{jk} > 0$ for all pairs of items j and k from the scale S.

Check of the monotonicity assumption

The monotonicity assumption can be checked by a visual inspection of the trace lines. Nevertheless, the MSP program that Molenaar, Sijtsma, and Boer (2000) proposed calculates indexes to evaluate the monotonicity assumption. The idea of these indexes is to allow the trace lines to have small decreases.

To check for the monotonicity assumption linked to the jth item ($j = 1, \ldots, J$), the population is cut into G_j groups (based on the individual's rest-score for item j as the sum of the individual's responses to the other items). Each group is indexed by $g = 1, \ldots, G_j$ ($g = 1$ represents the individuals with the lower rest-scores, and $g = G_j$ represents the individuals with the larger rest-scores).

Let Z_j be the random variable representing the groups corresponding to the jth item. It is expected that $\forall j = 1, \ldots, J$ and $r = 1, \ldots, m_j$. We have $P(Y_{jr} = 1|Z_j = g) \geq P(Y_{jr} = 1|Z_j = g')$ with $g > g'$. $G_j(G_j - 1)/2$ of such comparisons can be realized for the item j (denoted as $\#\text{ac}_j$ for active comparisons). In fact, only important violations of the expected results are retained, and a threshold minimum violation (minvi) is used to define an important violation $P(Y_{jr} = 1|Z_j = g') - P(Y_{jr} = 1|Z_j = g) > \text{minvi}$. Consequently, it is possible for each item to count the number of important violations ($\#\text{vi}_j$) and to compute the value of the maximum violation (maxvi_j) and the sum of the important violations (sum_j). Lastly, it is possible to test the null hypothesis $H_0 : P(Y_{jr} = 1|Z_j = g) \geq P(Y_{jr} = 1|Z_j = g')$ against the alternative hypothesis $H_a : P(Y_{jr} = 1|Z_j = g) < P(Y_{jr} = 1|Z_j = g')$ $\forall j, r, g, g'$ with $g > g'$.

Consider the table

		Item	Y_{jr}
		0	1
Group	g'	a	b
	g	c	d

Under the null hypothesis, the statistic

$$z = \frac{2\left\{\sqrt{(a+1)(d+1)} - \sqrt{bc}\right\}}{\sqrt{a+b+c+d-1}}$$

follows a standard normal distribution. The maximal value of z for the item j is denoted $z\text{max}_j$, and the number of significant z values is denoted $\#z\text{sig}_j$. The criterion used to check the monotonicity assumption linked to the item j is defined by Molenaar, Sijtsma, and Boer (2000) as

$$\begin{aligned}\text{Crit}_j &= 50(0.30 - H_j) + \sqrt{\#\text{vi}_j} + 100\frac{\#\text{vi}_j}{\#\text{ac}_j} + 100\text{maxvi}_j + 10\sqrt{\text{sum}_j} + 1000\frac{\text{sum}_j}{\#\text{ac}_j} \\ &\quad +5z\text{max}_j + 10\sqrt{\#z\text{sig}_j} + 100\frac{\#z\text{sig}_j}{\#\text{ac}_j} \qquad (5)\end{aligned}$$

It is generally considered that a criterion less than 40 signifies that the reported violations can be ascribed to sampling variation. A criterion exceeding 80 casts serious doubts on the monotonicity assumption for this item. If the criterion is between 40 and 80, further analysis must be considered to draw a conclusion (Molenaar, Sijtsma, and Boer 2000).

2.7 The doubly monotonely homogeneous model of Mokken (DMHMM)

The P++ and P−− matrices

The DMHMM is a model where the probabilities $P(X_j \geq l)\ \forall j, l$ produce the same ranking of items for all persons (Mokken and Lewis 1982). In practice, this means that the questionnaire is interpreted similarly by all the individuals, whatever their level of the latent trait.

$\mathbf{P}++$ is an $\mathbf{M} \times \mathbf{M}$ matrix in which each element corresponds to the probability $P(X_j \geq r, X_k \geq s) = P(Y_{jr} = 1, Y_{ks} = 1)$. The rows and the columns of this matrix are ordered from the most difficult threshold item $Y_{jr}\ \forall j, r$ to the easiest one.

$\mathbf{P}--$ is an $\mathbf{M} \times \mathbf{M}$ matrix in which each element corresponds to the probability $P(Y_{jr} = 0, Y_{ks} = 0)$. The rows and the columns of this matrix are ordered from the most difficult threshold item $Y_{jr}\ \forall j, r$ to the easiest one.

A set of items satisfies the doubly monotone assumption if the set satisfies an MHMM, and if the elements of the $\mathbf{P}++$ matrix are increasing in each row and the elements of the $\mathbf{P}--$ matrix are decreasing in each row.

We can represent each column of these matrices in a graph. On the x axis, the response categories are ordered in the same order as in the matrices; and on the y axis, the probabilities contained in the matrices are represented. The obtained curves must be nondecreasing for the $\mathbf{P}++$ matrix and must be nonincreasing for the $\mathbf{P}--$ matrix.

Check of the double monotonicity assumption via the analysis of the P matrices

Consider three threshold items Y_{jr}, Y_{ks}, and Y_{lt} with $j \neq k \neq l$. Under the DMHMM, if $P(Y_{ks} = 1) < P(Y_{lt} = 1)$, then it is expected that $P(Y_{ks} = 1, Y_{jr} = 1) < P(Y_{lt} = 1, Y_{jr} = 1)$. In the set of possible threshold items, we count the number of important violations of this principle among all the possible combinations of three items. An important violation represents a case where $P(Y_{ks} = 1, Y_{jr} = 1) - P(Y_{lt} = 1, Y_{jr} = 1) >$ minvi, where minvi is a fixed threshold. For each item j, $j = 1, \ldots, J$, we count the number of comparisons ($\#\text{ac}_j$), the number of important violations ($\#\text{vi}_j$), the value of maximal important violation (maxvi_j), and the sum of the important violations (sumvi_j). It is possible to test the null hypothesis H_0: $P(Y_{ks} = 1, Y_{jr} = 1) \leq P(Y_{lt} = 1, Y_{jr} = 1)$ against the alternative hypothesis H_a: $P(Y_{ks} = 1, Y_{jr} = 1) > P(Y_{lt} = 1, Y_{jr} = 1)$ with a McNemar test.

Let K be the random variable representing the number of individuals in the sample who satisfy $Y_{jr} = 1$, $Y_{ks} = 0$, and $Y_{lt} = 1$. Let N be the random variable representing the number of individuals in the sample who satisfy $Y_{jr} = 1$, $Y_{ks} = 0$, and $Y_{lt} = 1$, or who satisfy $Y_{jr} = 1$, $Y_{ks} = 1$, and $Y_{lt} = 0$. k and n are the realizations of these two random variables. Molenaar, Sijtsma, and Boer (2000) define the statistic:

$$z = \sqrt{2k + 2 + b} - \sqrt{2n - 2k + b} \quad \text{with} \quad b = \frac{(2k + 1 - n)^2 - 10n}{12n}$$

Under the null hypothesis, z follows a standard normal distribution. It is possible to count the number of significant tests ($\#z\text{sig}$) and the maximal value of the z statistics ($z\text{max}$).

A criterion can be computed for each item as the one used in (5), using the same thresholds for checking the double monotonicity assumption.

2.8 Contribution of each individual to the Guttman errors, H coefficients, and person-fit

From the preceding formulas, the number of Guttman errors induced by each individual can be computed. Let e_n be this number for the nth individual. The number of expected Guttman errors under the assumption of independence of the responses to the item is equal to $e_n^{(0)} = e^{S(0)}/N$. An individual with $e_n > e_n^{(0)}$ is very likely to be an individual whose responses are not influenced by the latent variable, and if e_n is very high, the individual can be considered an outlier.

By analogy with the Loevinger coefficient, we can compute the H_n coefficient in the following way: $H_n = 1 - (e_n/e_n^{(0)})$. A large negative value indicates an outlier, and a positive value is expected (note that $H_n \leq 1$).

It is interesting to note that when there is no missing value,

$$H^S = \frac{\sum_{n=1}^{N} H_n}{N}$$

Emons (2008) defines the normalized number of Guttman errors for polytomous items (G^p_N) as

$$G^p_{Nn} = \frac{e_n}{e_{\max,n}}$$

where $e_{\max,n}$ is the maximal number of Guttman errors obtained with a score equal to S_n. This index can be interpreted as

- $0 \leq G^p_{Nn} \leq 1$
- if G^p_{Nn} is close to 0, the individual n has few Guttman errors
- if G^p_{Nn} is close to 1, the individual n has many Guttman errors

The advantages of the G^p_{Nn} indexes are that they always lie between 0 and 1, inclusive, regardless of the number of items and response categories and that dividing by $e_{\max,n}$ adjusts the index to the observed score S_n. However, there is no threshold standard to use to judge the closeness of the index to 0 or 1.

2.9 MSP or AISP

Algorithm

The MSP proposed by Hemker, Sijtsma, and Molenaar (1995) allows selecting items from a bank of items that satisfy a Mokken scale. This procedure uses Mokken's definition of a scale (Mokken 1971): $H_{jk} > 0$, $H^S_j > c$, and $H^S > c$, for all pairs of items j and k from the scale S.

At the initial step, a kernel of at least two items is chosen (we can select, for example, the pair of items having the maximal significant H_{jk} coefficient). This kernel corresponds to the scale S^0.

At each step $n \geq 1$, we integrate into the scale $S^{(n-1)}$ the item j if that item satisfies these conditions:

- $j \notin S^{(n-1)}$
- $S^{(n)} \equiv S^{(n-1)} \cup j$
- $j = \arg\max_{k \notin S^{(n-1)}} H^{S^{*(n)}}$ with $S^{*(n)} \equiv S^{(n-1)} \cup k$
- $H^{S^{(n)}} \geq c$
- $H^{S^{(n)}}_j \geq c$
- $H^{S^{(n)}}_j$ is significantly positive
- H_{jk} is significantly positive $\forall k \in S^{(n-1)}$

The MSP is stopped as soon as no item satisfies all these conditions, but it is possible to construct a second scale with the items not selected in the first scale, and so on, until there are no more items remaining.

The threshold c is subjectively defined by the user: the authors of this article recommend fixing $c \geq 0.3$. As c gets larger, the obtained scale will become stronger, but it will be more difficult to include an item in a scale.

The Bonferroni corrections

At the initial step, in the general case, we compare all the possible H_{jk} coefficients to 0 using a test: there are $J(J-1)/2$ such tests. At each following step l, we compare $J^{(l)}$ H_j coefficients with 0, where $J^{(l)}$ is the number of unselected items at the beginning of step l.

Bonferroni corrections are used to take into account this number of tests and to keep a global level of significance equal to α (Molenaar, Sijtsma, and Boer 2000). At the initial step, we divide α by $J(J-1)/2$ to obtain the level of significance; and at each step l, we divide α by $\{J(J-1)/2\} + \sum_{m=1}^{l} J^{(m)}$.

When the initial kernel is composed of only one item, only $J-1$ tests are realized at the first step, and the coefficient $J(J-1)/2$ is replaced by $J-1$. When the initial kernel is composed of at least two items, this coefficient is replaced by 1.

Tip for improving the speed of computing

At each step, the items k (unselected in the current scale) that satisfy $H_{jk} < 0$ with an item j already selected in the current scale are automatically excluded.

3 Stata commands

In this section, we present three Stata commands for calculating the indexes and algorithms presented in this article. These commands have been intensively tested and compared with the output of the MSP software with several datasets. Small (and generally irrelevant) differences from the MSP software can persist and can be explained by different ways of approximating the values.

3.1 The tracelines command

Syntax

The syntax of the `tracelines` command (version 3.2 is described here) is

`tracelines` *varlist* [`, score restscore ci test cumulative logistic`
`repfiles(`*directory*`) scorefiles(`*string*`) restscorefiles(`*string*`)`
`logisticfile(`*string*`) nodraw nodrawcomb replace onlyone(`*varname*`)`
`thresholds(`*string*`)`]

Options

`score` displays graphical representations of trace lines of items as functions of the total score. This is the default if neither `restscore` nor `logistic` is specified.

`restscore` displays graphical representations of trace lines of items as functions of the rest-score (total score without the item).

`ci` displays the confidence interval at 95% of the trace lines.

`test` tests the null hypothesis that the slope of a linear model for the trace line is zero.

`cumulative` displays cumulative trace lines for polytomous items instead of classical trace lines.

`logistic` displays graphical representations of logistic trace lines of items as functions of the score: each trace comes from a logistic regression of the item response on the score. This kind of trace is possible only for dichotomous items. All the logistic trace lines are represented in the same graph.

`repfiles(`*directory*`)` specifies the directory where the files should be saved.

`scorefiles(`*string*`)` defines the generic name of files containing graphical representations of trace lines as functions of the score. The name will be followed by the name of each item and by the `.gph` extension. If this option is not specified, the corresponding graphs will not be saved.

`restscorefiles(`*string*`)` defines the generic name of files containing graphical representations of trace lines as functions of the rest-scores. The name will be followed by the name of each item and by the `.gph` extension. If this option is not specified, the corresponding graphs will not be saved.

`logisticfile(`*string*`)` defines the name of the file containing graphical representations of logistic trace lines. This name will be followed by the `.gph` extension. If this option is not specified, the corresponding graph will not be saved.

`nodraw` suppresses the display of graphs for individual items.

`nodrawcomb` suppresses the display of combined graphs but not of individual items.

`replace` replaces graphical files that already exist.

`onlyone(`*varname*`)` displays only the trace of a given item.

`thresholds(`*string*`)` groups individuals as a function of the score or the rest-score. The *string* contains the maximal values of the score or the rest-score in each group.

3.2 The loevh command

Syntax

The syntax of the `loevh` command (version 7.1 is described here) is

`loevh` *varlist* [`, pairwise pair ppp pmm noadjust generror(`*newvar*`) replace graph monotonicity(`*string*`) nipmatrix(`*string*`)`]

`loevh` requires that the commands `tracelines`, `anaoption`, `gengroup`, `guttmax`, and `genscore` be installed.

Options

`pairwise` omits, for each pair of items, only the individuals with a missing value on these two items. By default, `loevh` omits all individuals with at least one missing value in the items of the scale.

`pair` displays the values of the Loevinger's H coefficients and the associated statistics for each pair of items.

`ppp` displays the $\mathbf{P}++$ matrix (and the associated graph with `graph`).

`pmm` displays the $\mathbf{P}--$ matrix (and the associated graph with `graph`).

`noadjust` uses N_{jk} as the denominator instead of the default, $N_{jk}-1$, when calculating test statistics. The MSP software also uses N_{jk}.

`generror(`*newvar*`)` defines the prefix of five new variables. The first new variable (only the prefix) will contain the number of Guttman errors attached to each individual; the second one (the prefix followed by `_0`), the number of Guttman errors attached to each individual under the assumption of independence of the items; the third one (the prefix followed by `_H`), the quantity 1 minus the ratio between the two preceding values; the fourth one (the prefix followed by `_max`), the maximal possible Guttman errors corresponding to the score of the individual; and the last one (the prefix followed by `_GPN`), the normalized number of Guttman errors. With the `graph` option, a histogram of the number of Guttman errors by individual is drawn.

`replace` replaces the variables defined by the `generror()` option.

`graph` displays graphs with the `ppp`, `pmm`, and `generror()` options. This option is automatically disabled if the number of possible scores is greater than 20.

`monotonicity(`*string*`)` displays indexes to check monotonicity of the data (MHMM). This option produces output similar to that of the MSP software. The *string* contains the following suboptions: `minvi()`, `minsize()`, `siglevel()`, and `details`. If you want to use all the default values, type `monotonicity(*)`.

`minvi(`#`)` defines the minimal size of a violation of monotonicity. The default is `monotonicity(minvi(0.03))`.

`minsize(`#`)` defines the minimal size of groups of patients to check the monotonicity (by default, this value is equal to $N/10$ if $N > 500$, to $N/5$ if $250 < N \leq 500$, and to $N/3$ if $N \leq 250$ with the minimal group size fixed at 50).

`siglevel(`#`)` defines the significance level for the tests. The default is `monotonicity(siglevel(0.05))`.

`details` displays more details with polytomous items.

`nipmatrix(`*string*`)` displays indexes to check the nonintersection (DMHMM). This option produces output similar to that of the MSP software. The *string* contains two suboptions: `minvi()` and `siglevel()`. If you want to use all the default values, type `nipmatrix(*)`.

`minvi(`#`)` defines the minimal size of a violation of nonintersection. The default is `nipmatrix(minvi(0.03))`.

`siglevel(`#`)` defines the significance level for the tests. The default is `nipmatrix(siglevel(0.05))`.

Saved results

`loevh` saves the following in `r()`:

Scalars

`r(pvalH)`	p-value for Loevinger's H coefficient of scalability
`r(zH)`	z statistic for Loevinger's H coefficient of scalability
`r(eGutt0)`	total number of theoretical Guttman errors associated with the scale
`r(eGutt)`	total number of observed Guttman errors associated with the scale
`r(loevh)`	Loevinger's H coefficient of scalability

Matrices

`r(Obs)`	(matrix) number of individuals used to compute each coefficient H_{jk} (if the `pairwise` option is not used, the number of individuals is the same for each pair of items)
`r(pvalHj)`	p-values for consistency of each item with the scale
`r(pvalHjk)`	p-values for pairs of items
`r(zHj)`	z statistics for consistency of each item with the scale
`r(zHjk)`	z statistics for pairs of items
`r(P11)`	$\mathbf{P}++$ matrix
`r(P00)`	$\mathbf{P}--$ matrix
`r(eGuttjk0)`	theoretical Guttman errors associated with each item pair
`r(eGuttj0)`	theoretical Guttman errors associated with the scale
`r(eGuttjk)`	observed Guttman errors associated with each item pair
`r(eGuttj)`	observed Guttman errors associated with the scale
`r(loevHjk)`	Loevinger's H coefficients for pairs of items
`r(loevHj)`	Loevinger's H coefficients for consistency of each item with the scale

3.3 The msp command

Syntax

The syntax of the `msp` command (version 6.6 is described here) is

`msp` *varlist* [`, c(#) kernel(#) p(#) minvalue(#) pairwise nobon notest nodetails noadjust`]

`msp` requires that the `loevh` command be installed.

Options

`c(#)` defines the value of the threshold c. The default is `c(0.3)`.

`kernel(#)` defines the first # items as the kernel of the first subscale. The default is `kernel(0)`.

`p(#)` defines the level of significance of the tests. The default is `p(0.05)`.

`minvalue(#)` defines the minimum value of an H_{jk} coefficient between two items j and k on a same scale. The default is `minvalue(0)`.

`pairwise` omits, for each pair of items, only the individuals with a missing value on these two items. By default, `msp` omits all individuals with at least one missing value in the items of the scale.

`nobon` suppresses the Bonferroni corrections of the levels of significance.

`notest` suppresses testing of the nullity of the Loevinger's H coefficient.

`nodetails` suppresses display of the details of the algorithm.

`noadjust` uses N_{jk} as the denominator instead of the default, $N_{jk}-1$, when calculating test statistics. The MSP software also uses N_{jk}.

Saved results

`msp` saves the following in `r()`:

Scalars

`r(dim)`	number of created scales
`r(nbitems#)`	number of selected items in the #th scale
`r(H#)`	value of the Loevinger's H coefficient of scalability for the #th scale

Macros

`r(lastitem)`	when only one item is remaining, the name of that item
`r(scale#)`	list of the items selected in the #th scale (in the order of selection)

Matrices

`r(selection)`	a vector that contains, for each item, the scale where it is selected (or `0` if the item is unselected)

3.4 Output

We present an example of output of these programs with items of the French adaptation of the Ways of Coping Checklist questionnaire (Cousson et al. 1996). This questionnaire measures coping strategies and includes 27 items that compose three dimensions: problem-focused coping, emotional coping, and seeking social support. The sample is composed of 100 women, each with a recent diagnosis of breast cancer.

Output of the loevh command

The `loevh` command allows researchers to obtain the values of the Loevinger's H coefficients. Because the sample was small, it was impossible to obtain several groups of 50 individuals or more. As a consequence, for the `monotonicity()` option, the `minsize()` has been fixed at 30. We studied the emotional dimension composed of nine items (with four response categories per item). The rate of missing data varied from 2% to 15% per item. Only 69 women have a complete pattern of responses, so the `pairwise` option was employed to retain a maximum of information.

```
. use wccemo
. loevh item2 item5 item8 item11 item14 item17 item20 item23 item26, pairwise
> monotonicity(minsize(30)) nipmatrix(*)
                        Observed  Expected                                  Number
           Difficulty   Guttman   Guttman  Loevinger           H0: Hj<=0    of NS
Item   Obs    P(Xj=0)    errors    errors    H coeff   z-stat.   p-value     Hjk
-------------------------------------------------------------------------------
item2   92     0.2935       453    732.03    0.38117   7.4874    0.00000       1
item5   92     0.3261       395    751.61    0.47446   9.5492    0.00000       1
item8   90     0.3667       515    788.65    0.34699   7.6200    0.00000       4
item11  97     0.5670       519    862.50    0.39826   9.2705    0.00000       1
item14  98     0.6327       532    752.63    0.29314   6.8306    0.00000       3
item17  94     0.7660       299    487.40    0.38653   7.4598    0.00000       1
item20  95     0.6632       494    711.53    0.30573   6.7867    0.00000       1
item23  85     0.5412       525    729.72    0.28054   6.1752    0.00000       2
item26  89     0.6517       502    710.59    0.29355   6.3643    0.00000       2
-------------------------------------------------------------------------------
Scale  100                 2117   3263.33    0.35128  15.9008    0.00000

        Summary per item for check of monotonicity
        Minvi=0.030    Minsize=   30  Alpha=0.050

Items  #ac  #vi  #vi/#ac   maxvi     sum  sum/#ac    zmax #zsig  Crit
-------------------------------------------------------------------------------
item2    3    0                                                  -4     graph
item5    3    0                                                  -9     graph
item8    3    0                                                  -2     graph
item11   3    0                                                  -5     graph
item14   3    0                                                   0     graph
item17   2    0                                                  -4     graph
item20   3    0                                                  -0     graph
item23   3    0                                                   1     graph
item26   3    0                                                   0     graph
-------------------------------------------------------------------------------
Total   52    0   0.0000  0.0000  0.0000   0.0000  0.0000           0
-------------------------------------------------------------------------------
```

```
        Summary per item for check of non-Intersection via Pmatrix
        Minvi=0.030    Alpha=0.050
```

Items	#ac	#vi	#vi/#ac	maxvi	sum	sum/#ac	zmax	#zsig	Crit
item2	1512	49	0.0324	0.0990	2.2005	0.0015	1.6844	1	51
item5	1512	85	0.0562	0.1239	4.1743	0.0028	2.9280	6	81
item8	1512	90	0.0595	0.1105	4.2927	0.0028	2.5221	4	81
item11	1512	120	0.0794	0.1105	5.4429	0.0036	2.5221	6	89
item14	1512	88	0.0582	0.1081	4.1701	0.0028	2.3015	7	88
item17	1512	52	0.0344	0.0865	2.4122	0.0016	2.0662	2	57
item20	1512	52	0.0344	0.0830	2.2127	0.0015	2.3015	1	57
item23	1512	90	0.0595	0.0990	4.2123	0.0028	1.8742	3	77
item26	1512	94	0.0622	0.1239	4.3258	0.0029	2.9280	4	87

This scale has a satisfactory scalability ($H^S = 0.35$). Three items (14, 23, 26) display a borderline value for the H_j^S coefficient (0.28 or 0.29). The monotonicity assumption is not rejected because no important violation of this assumption occurred and the criteria are satisfied. This is not the case for the nonintersection of the Pmatrix curves: several criteria are greater than 80 (items 5, 8, 11, 14, 23, 26), showing an important violation of this assumption. The model followed by these data is therefore more an MHMM than a DMHMM. Because the indexes suggest that the MHMM is appropriate, the score computed by summing codes associated with the nine items can be considered a correct ordinal measure of the studied latent trait (the emotional coping), and the three fundamental assumptions of IRT (unidimensionality, local independence, and monotonicity) can be considered verified.

Output of the msp command

The `msp` command runs the Mokken scale procedure.

```
. msp item2 item5 item8 item11 item14 item17 item20 item23 item26, pairwise
Scale: 1
─────────
Significance level: 0.001389
The two first items selected in the scale 1 are item2 and item11 (Hjk=0.6245)
The following items are excluded at this step: item14 item23
Significance level: 0.001220
The item item17 is selected in the scale 1       Hj=0.5304      H=0.5748
The following items are excluded at this step: item8
Significance level: 0.001136
The item item5 is selected in the scale 1        Hj=0.5464      H=0.5588
The following items are excluded at this step: item26
Significance level: 0.001111
The item item20 is selected in the scale 1       Hj=0.3758      H=0.4864
Significance level: 0.001111
There is no more items remaining.
```

```
                        Observed  Expected                                 Number
            Difficulty   Guttman   Guttman   Loevinger          H0: Hj<=0   of NS
Item   Obs    P(Xj=0)    errors    errors     H coeff  z-stat.   p-value     Hjk
-------------------------------------------------------------------------------
item20  95     0.6632       212    339.64     0.37582   5.5460   0.00000       0
item5   92     0.3261       179    376.71     0.52484   7.5735   0.00000       0
item17  94     0.7660       124    233.07     0.46797   5.8889   0.00000       0
item2   92     0.2935       186    367.75     0.49422   6.9525   0.00000       0
item11  97     0.5670       181    400.10     0.54761   8.4434   0.00000       0
-------------------------------------------------------------------------------
Scale  100                  441    858.64     0.48640  10.9364   0.00000
Scale: 2
-------

Significance level: 0.008333
The two first items selected in the scale 2 are item23 and item26 (Hjk=0.4391)
The following items are excluded at this step: item8
Significance level: 0.007143
The item item14 is selected in the scale 2       Hj=0.4276      H=0.4313
Significance level: 0.007143
There is no more items remaining.

                        Observed  Expected                                 Number
            Difficulty   Guttman   Guttman  Loevinger          H0: Hj<=0   of NS
Item   Obs    P(Xj=0)    errors    errors    H coeff   z-stat.   p-value     Hjk
-------------------------------------------------------------------------------
item14  98     0.6327       115    200.89    0.42756    5.4739   0.00000       0
item23  85     0.5412       109    193.44    0.43651    5.2885   0.00000       0
item26  89     0.6517       114    200.00    0.43000    5.4109   0.00000       0
-------------------------------------------------------------------------------
Scale  100                  169    297.17    0.43129    6.5985   0.00000

There is only one item remaining (item8).
```

The AISP creates two groups of items.

On the one hand, five items measure negation or the wish to forget the reason for the stress: `item2`, "Wish that the situation would go away or somehow be over with"; `item5`, "Wish that I can change what is happening or how I feel"; `item11`, "Hope a miracle will happen"; `item17`, "I daydream or imagine a better time or place than the one I am in"; and `item20`, "Try to forget the whole thing". For this set, the scalability coefficient is good (0.49), and there is no problem concerning the monotonicity assumption (maximal criterion per item of -4), nor is there a problem concerning the intersection of the curves (maximal criterion per item of 38). This set seems to satisfy a DMHMM and is composed of 5 of the 11 items composing the "wishful thinking" and "detachment" dimensions proposed by Folkman and Lazarus (1985) in an analysis of the Ways of Coping Checklist questionnaire among a sample of students.

On the other hand, three items measure culpability: `item14`, "Realize I brought the problem on myself"; `item23`, "Make a promise to myself that things will be different next time"; and `item26`, "Criticize or lecture myself". For this set, the scalability coefficient is good (0.43), and there is no problem concerning the monotonicity assumption (maximal criterion per item of -6), nor concerning the intersection of the curves (maximal criterion per item of -6). This set seems to satisfy a DMHMM and is composed of the three items of the "self blame" dimension proposed by Folkman and Lazarus (1985).

In our case, it is possible to choose between a set of items that satisfy an MHMM and two sets of items that each satisfy a DMHMM. Because the three sets of items are interpretable (emotional coping for the set of items satisfying MHMM; negation and culpability for the two other sets of items), there is no problem to choose freely from the available types of measured concepts. Concerning the validation of the questionnaire, it is preferable to choose the set of items containing all items satisfying the emotional coping, which is closer to the output returned by the `loevh` command.

4 References

Cousson, F., M. Bruchon-Schweitzer, B. Quintard, J. Nuissier, and N. Rascle. 1996. Analyse multidimensionnelle d'une échelle de coping: validation française de la W.C.C. (way of coping checklist). *Psychologie Française* 41: 155–164.

Emons, W. H. 2008. Nonparametric person-fit analysis of polytomous item scores. *Applied Psychological Measurement* 32: 224–247.

Folkman, S., and R. S. Lazarus. 1985. If it changes it must be a process: Study of emotion and coping during three stages of a college examination. *Journal of Personality and Social Psychology* 48: 150–170.

Guttman, L. 1944. A basis for scaling qualitative data. *American Sociological Review* 9: 139–150.

Hardouin, J.-B. 2002. *The SAS Macro-program "%LOEVH"*. University of Nantes, http://sasloevh.anaqol.org.

———. 2007. Rasch analysis: Estimation and tests with raschtest. *Stata Journal* 7: 22–44.

Hemker, B. T., K. Sijtsma, and I. W. Molenaar. 1995. Selection of unidimensional scales from a multidimensional item bank in the polytomous Mokken IRT model. *Applied Psychological Measurement* 19: 337–352.

Hemker, B. T., K. Sijtsma, I. W. Molenaar, and B. W. Junker. 1997. Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika* 62: 331–347.

Loevinger, J. 1948. The technic of homogeneous tests compared with some aspects of scale analysis and factor analysis. *Psychological Bulletin* 45: 507–529.

Mokken, R. J. 1971. *A Theory and Procedure of Scale Analysis: With Applications in Political Research*. Berlin: De Gruyter.

———. 1997. Nonparametric models for dichotomous responses. In *Handbook of Modern Item Response Theory*, ed. W. J. van der Linden and R. K. Hambleton, 351–368. New York: Springer.

Mokken, R. J., and C. Lewis. 1982. A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement* 6: 417–430.

Molenaar, I. W. 1997. Nonparametric models for polytomous responses. In *Handbook of Modern Item Response Theory*, ed. W. J. van der Linden and R. K. Hambleton, 369–380. New York: Springer.

Molenaar, I. W., K. Sijtsma, and P. Boer. 2000. *User's Manual for MSP5 for Windows: A Program for Mokken Scale Analysis for Polytomous Items (Version 5.0)*. University of Groningen, Groningen, The Netherlands.

Sijtsma, K., and I. W. Molenaar. 2002. *Introduction to Nonparametric Item Response Theory*. Thousand Oaks, CA: Sage.

van der Ark, L. A. 2007. Mokken scale analysis in R. *Journal of Statistical Sofware* 20: 1–19.

van der Linden, W. J., and R. K. Hambleton, ed. 1997. *Handbook of Modern Item Response Theory*. New York: Springer.

Weesie, J. 1999. mokken: Stata module: Mokken scale analysis. Statistical Software Components, Department of Economics, Boston College. http://econpapers.repec.org/software/bocbocode/sjw31.htm.

Zheng, X., and S. Rabe-Hesketh. 2007. Estimating parameters of dichotomous and ordinal item response models with gllamm. *Stata Journal* 7: 313–333.

About the authors

Jean-Benoit Hardouin and Véronique Sébille are, respectively, attached professor and full professor in biostatistics at the Faculty of Pharmaceutical Sciences of the University of Nantes. Their research applies item–response theory in clinical research. Angélique Bonnaud-Antignac is an attached professor in clinical psychology at the Faculty of Medicine of the University of Nantes. Her research deals with the evaluation of quality of life in oncology.