The Stata Journal (*yyyy*) *vv*, Number *ii*, pp. 1–21

Non parametric Item Response Theory using Stata

Jean-Benoit Hardouin - Faculty of Pharmaceutical Sciences
EA 4275 "Biostatistics, Clinical Research and Subjective Measure in Health Science"
University of Nantes, France
Angélique Bonnaud-Antignac - Faculty of Medicine
ERT A0901 "ERSSCA"
University of Nantes, France
Véronique Sébille - Faculty of Pharmaceutical Sciences
EA 4275 "Biostatistics, Clinical Research and Subjective Measure in Health Science"
University of Nantes, France

Abstract. Item Response Theory (IRT) 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). They are composed of items wich satisfy the three fundamental assumptions of Item Response Theory: unidimensionality, monotonicity and Local Independence. They can be considered as non parametric models in IRT. Traces of the items and Loevinger's H coefficients are particular useful indices for checking whether a set of items constitute a Mokken scale.

However, these indices are not available in general statistical packages. We introduce Stata modules to compute them. The options of these modules are described and exemples of output are shown.

Keywords: Stata, Items traces, 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 non-measurable characteristics of the respondents (latent traits). These models can be applied to measure such a latent variable (measurement models), or to investigate influences of covariates on these latent variables.

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

Most of Item Response Models (IRM) are parametric: they model the probability

to respond at each response category of each item by a function depending on the latent trait, typically considered as a set of fixed effects or as a random variable, and of parameters characterizing the items. The Rasch model and the Birnbaum model for dichotomous items, and the Partial Credit Model and the Rating Scale Model for polytomous items are the most popular IRM and are already described for the Stata software [Hardouin (2007), Zheng and Rabe-Hesketh (2007)].

Mokken (1971) defines a non parametric model to study the properties of a set of items in the framework of IRT. Mokken calls this model the monotonely homogenous model, but it is generally referred to as the Mokken model. This model is implemented on a standalone package MSP [Molenaar et al. (2000)], and codes have been already developped in Stata [Weesie (1999)], SAS [Hardouin (2002)], and R [van der Ark (2007)] langages. We propose modules under Stata in order to study the fit of a set of items to a Mokken model. These modules are more complete than the `mokken` module of Jeroen Weesie that, for example, don't offer the possibility to analyse 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 to parametric IRT models, the Mokken model necessitates few assumption regarding to the relationship between the latent trait and the responses to the items, and so, generally allows keeping a more important number of items. As a consequence, the precision of the individuals ordering is higher [Sijtsma and Molenaar (2002)].

2 The Mokken scales

2.1 Notation

In the following, we use the following notation :

- X_j is the random variable (item) representing the responses to the jth item, $j = 1, ..., J$,
- X_{nj} is the random variable (item) representing the responses to the jth item, $j = 1, ..., J$ 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 on the latent trait and is referred to as a negative response, whereas the m_j non-zero response categories $(1, 2, ..., m_j)$ increase with increasing level on 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 ($j = 1, ..., J$ and $r = 1, ..., m_j$)
- $P(.)$ refer to observed proportions

2.2 Monotonely Homogeneous Model of Mokken

The Mokken Scales are sets of items satisfying a Monotonely Homogeneous Model of Mokken (MMHM) [Mokken (1997) Molenaar (1997), Sijtsma and Molenaar (2002)]. This kind of model is a non parametric IRM defined by the three fundamental assumptions of the Item Response Theory (IRT):

- unidimensionality (the response to the items are explained by a common latent trait)
- local independence (conditionally to the latent trait, the responses to the items are independent)
- monotonicity (the probability to have a response to an item greatear or equal to a given value is a non decreasing 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 to interpret the results. For a questionnaire aiming at measuring several latent traits, such an analysis must be realised 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 with 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 there is not a strong redundance between the items.

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

2.3 Traces of the items

The traces of items can be used to check the monotonicity assumption. We define the score for each individual as the sum of its responses ($S_n = \sum_{j=1}^{J} X_{nj}$). This score is assumed to represent a 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 traces increase. This means 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 easily correct responses to the items. In Health sciences, if we search to measure a dysfunctioning through the presence of symptoms, this means that a patient having a

high level of dysfunctioning will display more symptoms. An exemple of trace is given in figure 1.

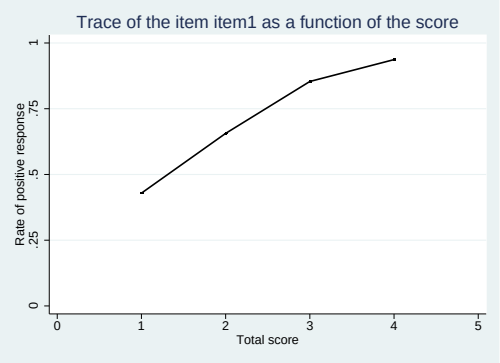

Figure 1: Trace of a dichotomous item as a function of the score.

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. In order to avoid it, the rest-score (computed as the score to 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).

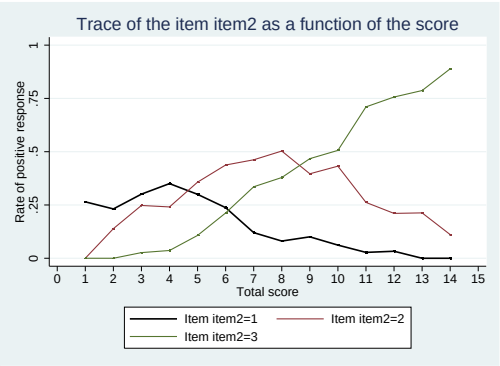

Figure 2: Traces of a polytomous item as a function of the score.

Unfortunately, these traces 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 traces corresponding to each response category does not increase. Cumulative traces 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 respected, these traces increase. An example is given in figure 3.

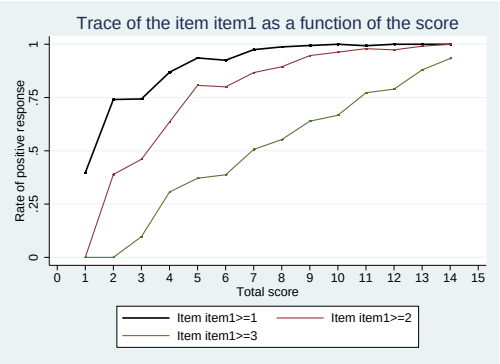

Figure 3: Cumulative traces of a polytomous item as a function of the score.

2.4 The Guttman errors

Dichotomic 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 which have a positive response to the more difficult item and a negative response to the easiest item. In Education sciences, this reprensents the number of individuals that have correctly responded to a given item, but uncorrectly responded to an easier item. In health sciences, this represents the number of individuals which present a given symptom, but which 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 non-missing 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 $\frac{b_{jk}+d_{jk}}{N_{jk}} > \frac{c_{jk}+d_{jk}}{N_{jk}}$ (equivalently, if $b_{jk} > c_{jk}$), and the number of Guttman errors e_{jk} in

this case is $e_{jk} = N_{jk}.P(X_j = 0, X_k = 1) = c_{jk}$. More generally, if we ignore the easier item between j and k:

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

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

$$e_{jk}^{(0)} = N_{jk}. \min \{P(X_j = 0).P(X_k = 1); P(X_j = 1).P(X_k = 0)\} = \frac{(a_{jk} + e_{jk})(e_{jk} + d_{jk})}{N_{jk}} \quad (2)$$

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:

$$\begin{aligned} e_{j(r)k(s)} &= N_{jk}. \min \{P(X_j \geq r, X_k < s); P(X_j < r, X_k \geq s)\} \\ &= N_{jk}. \min \{P(Y_{jr} = 1, Y_{ks} = 0); P(Y_{jr} = 0, Y_{ks} = 1)\} \end{aligned} \quad (3)$$

The number of Guttman errors between the two items is:

$$e_{jk} = \sum_{r=1}^{m_j} \sum_{s=1}^{m_k} e_{j(r)k(s)} \quad (4)$$

Note that if $m_j = m_k = 1$ (dichotomous case), this formula is equivalent to equation 1

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

$$e_{j(r)k(s)}^{(0)} = N_{jk}.P(X_j < r)P(X_k \geq s) = N_{jk}.P(Y_{jr} = 0)P(Y_{ks} = 1) \quad (5)$$

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

$$e_{jk}^{(0)} = \sum_{r=1}^{m_j} \sum_{s=1}^{m_k} e_{j(r)k(s)}^{(0)} \quad (6)$$

2.5 The Loevinger's H coefficients

Loevinger (1948) proposes three indices which can be defined as a function 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:

$$H_{jk} = 1 - \frac{e_{jk}}{e_{jk}^{(0)}} \tag{7}$$

We have $H_{jk} \leq 1$ with $H_{jk} = 1$ if and only if there is no Guttman error between the items j and k. If this coefficient is close of 1, there are few Guttamn errors, and so the two items probably measure the same latent trait. An indice close to 0 signifies that the response to the two items are independent, and therefore that might reveal the fact that the two items probably do not measure the same latent trait. A significantly negative value to this indice is not expected, and can be a flag that an (or several) item(s) has(have) been incorrectly coded, or is(are) incorrectly understood by the respondants.

We can test H_0: $H_{jk} = 0$ (against H_1: $H_{jk} > 0$). Under the null assumption, the statistic

$$Z = \frac{Cov(X_j, X_k)}{\sqrt{\frac{Var(X_j)Var(X_k)}{N_{jk}-1}}} = \rho_{jk}\sqrt{N_{jk}-1} \tag{8}$$

follows a standardized 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 (scale) and j an item which belongs to this scale ($j \in S$). H_j^S is the Loevinger's H coefficient measuring the consistency of the item j with a scale S.

$$H_j^S = 1 - \frac{e_j^S}{e_j^{S(0)}} = 1 - \frac{\sum_{k\in S,\ k\neq j} e_{jk}}{\sum_{k\in S,\ k\neq j} e_{jk}^{(0)}} \tag{9}$$

If the scale S is a good scale (i.e. if it satisfies a MMHM for example), this indice is close to 1 if the item j has a good consistency with the scale S and close to 0 if it has a bad consistency with this scale.

It is possible to test H_0: $H_j^S = 0$ (against H_1: $H_j^S > 0$). Under the null assumption, the statistic

$$Z = \frac{\sum_{k\in S, k\neq j} Cov(X_j, X_k)}{\sqrt{\sum_{k\in S, k\neq j} \frac{Var(X_j)Var(X_k)}{N_{jk}-1}}} \tag{10}$$

follows a standardized 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.

$$H^S = 1 - \frac{e^S}{e^{S(0)}} = 1 - \frac{\sum_{j\in S}\sum_{k\in S,\ k>j} e_{jk}}{\sum_{j\in S}\sum_{k\in S,\ k>j} e^{(0)}_{jk}} \tag{11}$$

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

It is possible to test H_0: $H^S = 0$ (against H_1: $H^S > 0$). Under the null assumption, the statistic

$$Z = \frac{\sum_{j\in S}\sum_{k\in S,k\neq j} Cov(X_j, X_k)}{\sqrt{\sum_{j\in S}\sum_{k\in S,k\neq j} \frac{Var(X_j)Var(X_k)}{N_{jk}-1}}} \tag{12}$$

follows a standardized normal distribution.

We note that in the MSP software [Molenaar et al. (2000)], the Z statistic defined in equation (8), (10) and (12) 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) shows that if a scale S is a a Mokken scale, then $H^S > 0$. But the reciprocity 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 which satisfy a Mokken scale. He suggests that there is a threshold $c > 0.3$ such as, if $H^S > c$, then the scale S satisfies a Mokken scale. This idea is used by Mokken (1971) and adapted by Hemker et al. (1995) to propose the "Mokken Scale Procedure" or "Automated Item Seclection Procedure" (AISP) [Sijtsma and Molenaar (2002)].

More, the fit to the Mokel scale is satisfying if $H^S_j > c$ and $H_{jk} > 0$ whatever j and k two given items of the scale S.

Check of the monotonicity assumption

The monotonicity assumption can be checked by a visual inspection of the traces. Nevertheless, the MSP program Molenaar et al. (2000) propose indexes to evaluate it. The idea of these indexes is to allow the traces to have small decreases.

To check for the monotonicity assumption linked to the jth item ($j = 1...J$), the population is cut in G_j groups (based on the rest-score of the individuals computed as the sum of the items on all the others items). Each group is indexed by $g = 1...G_j$ ($g = 1$ represents the individuals with the lower rest-scores, and $g = G_j$ the individuals with the larger rest-scores).

Let Z_j the random variable representing the groups corresponding to the jth item. It is expected that $\forall j = 1, ..., J$ and $r = 1, ..., m_j$, we have $P(Y_{jr} = 1|Z_j = g) \geq P(Y_{jr} = 1|Z_j = g')$ with $g > g'$. $\frac{G_j(G_j-1)}{2}$ of such comparisons can be realized for the item j (noted $\#ac_j$ for "active comparisons"). In fact, only important violations of the expected results are retained, and a threshold $minvi$ is used to define an important violation $P(Y_{jr} = 1|Z_j = g') - P(Y_{jr} = 1|Z_j = g) > minvi$. Consequently, it is possible for each item to count the number of important violations ($\#vi_j$) and to compute the value of the maximal violation ($maxvi_j$) and the sum of the important violations (sum_j). Last, it is possible to test the null assumptions $H_0 : \ P(Y_{jr} = 1|Z_j = g) \geq P(Y_{jr} = 1|Z_j = g')$ against the alternative assumptions $H_1 : \ P(Y_{jr} = 1|Z_j = g) < P(Y_{jr} = 1|Z_j = g') \ \forall j, r, g, g'$ with $g > g'$.

Let the table

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

Under the null assumption, the statistics

$$z = \frac{2\left(\sqrt{(a+1)(d+1)} - \sqrt{bc}\right)}{\sqrt{a+b+c+d-1}} \tag{13}$$

follows a standardized normal distribution. The maximal value of z for the item j is denoted $zmax_j$ and the number of significant z values is denoted $\#zsig_j$. The criterion used to check the monotonicity assumption linked to the item j is defined by Molenaar et al. (2000) as:

$$\begin{aligned} Crit_j &= 50(0.30 - H_j) + \sqrt{\#vi_j} + 100\frac{\#vi_j}{\#ac_j} + 100maxvi_j + 10\sqrt{sum_j} + 1000\frac{sum_j}{\#ac_j} \\ &\quad +5zmax_j + 10\sqrt{\#zsig_j} + 100\frac{\#zsig_j}{\#ac_j} \end{aligned} \tag{14}$$

It is generally considered that a criterion less than 40 signifies that the violations reported can be ascribed to sampling variation. A criterion exceeding 80 casts serious doubts on the respect of the monotonicity assumption for this item. If the criterion is between 40 and 80, further considerations must be analysed to conclude [Molenaar et al. (2000)].

2.7 The Doubly Monotonely Homogeneous Model of Mokken - DMHMM

The P++ and P−− matrices

The "Doubly" Monotonely Homogeneous Model of Mokken (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 for all the individuals, whatever their level on the latent trait.

The P++ matrix is the matrices $(M \times M)$ where 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.

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

A set of item satisfies the doubly monotone assumption if this set satisfy a MMHM and if the elements of the P++ matrix are increasing in each row, and the elements of the 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 than in the matrices, and on the Y-axis, the probabilities contained in the matrices are represented. The obtained curves must be non decreasing for the P++ matrix, and non increasing for the P−− matrix.

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

Let 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 combinaison of three items. An important violation represents a case where $P(Y_{ks} = 1, Y_{jr} = 1) - P(Y_{lt} = 1, Y_{jr} = 1) > minvi$ with $minvi$ a fixed threshold. For each item j, $j = 1, ...J$, we count the number of comparisons ($\#ac_j$), the number of important violations ($\#vi_j$), the value of maximal important violation ($maxvi_j$), the sum of the important violations ($sumvi_j$). It is pos-

sible to test the null assumption $H_0 : \ P(Y_{ks} = 1, Y_{jr} = 1) \leq P(Y_{lt} = 1, Y_{jr} = 1)$ against the alternative assumption $H_1 : \ P(Y_{ks} = 1, Y_{jr} = 1) > P(Y_{lt} = 1, Y_{jr} = 1)$ 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 the random variable representing the number of individuals in the sample satisfying $Y_{jr} = 1$, $Y_{ks} = 0$ and $Y_{lt} = 1$, or $Y_{jr} = 1$, $Y_{ks} = 1$ and $Y_{lt} = 0$. k and n are the realizations of these two random variables. Molenaar et al. (2000) defines the statistics:

$$z = \sqrt{2k + 2 + b} - \sqrt{2n - 2k + b} \ with \ b = \frac{(2k + 1 - n)^2 - 10n}{12n} \tag{15}$$

Under the null hypothesis, z follows a standardized normal distribution. It is possible and to count the number of significant tests ($\#zsig$) and the maximal value of the statistics z ($zmax$).

A criterion can be computed for each item as the one used in formula 14 using the same thresholds for checking the double monotonicity assumption.

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

From the preeceding formulas, the number of Guttman errors induced by each individual can be computed. Let e_n 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 $e_n^{(0)} = \frac{e^{S(0)}}{N}$. 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 consider as an outlier.

By analogy with the Loevinger coefficient, we can compute the H_n coefficient in the following way $H_n = 1 - \frac{e_n}{e_n^{(0)}}$. 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

$$H^S = \frac{\sum_{n=1}^{N} H_n}{N} \tag{16}$$

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

$$G_{Nn}^p = \frac{e_n}{e_{max,n}} \tag{17}$$

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 full of Guttman errors

The advantages of the G^p_{Nn} indexes are to be bounded between 0 and 1 whatever the number of items and response categories and to be adjusted to the observed score of each individual. Nevertheless, there is no consensual definition of a threshold to identify a no misfit and a misfited individual.

2.9 The Mokken Scale Procedure (MSP) or Automated Item Selection Procedure (AISP)

Algorithm

The Mokken Scale Procedure proposed by Hemker et al. (1995) allows selecting items from a bank of items wich satisfy a Mokken Scale. This procedure puts on the Mokken's definition of a scale [Mokken (1971)]: $H_{jk} > 0$, $H^S_j > c$ and $H^S > c$ whatever j and k two given items of the scale S.

At the initial step, a kernel of items is chosen (at least two items: 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 in the scale $S^{(n-1)}$ the item j if this item satisfies:

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

The MSP is stopped as soon as none item satisfies all these conditions, but it is possible to construct a second scale with the items unselected in the first scale, and so on until there is no more item remaining.

The threshold c is subjectively defined by the user: the authors recommand to fix $c \geq 0.3$: 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 $\frac{J(J-1)}{2}$ such tests. At each following step l, we compare $J^{(l)}$ H_j coefficients to 0 where $J^{(l)}$ is the number of unselected items at the beginning of the 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 et al. (2000)]. At the initial step, we divide α by $\frac{J(J-1)}{2}$ to obtain the level of significance, and at each step l, we divide α by $\frac{J(J-1)}{2} + \sum_{m=1}^{l} J^{(m)}$.

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

Tip to improve the time of computing

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

3 Stata modules

In this section, we presents three Stata modules which allow computing indices and realizing algorithms presented in this paper. These modules have intensively been tested and compared to the output of the MSP software with several datasets. Small (and generally irrelevant) differences with the MSP software can persist, and can be explained by different ways of approximation of the values.

3.1 The `traces` module

Syntax of the `traces` module

The syntax of the `traces` module is (version 3.2 is described here):

`traces` *varlist* [`, score restscore ci test cumulative logistic`
`repfile(`*string*`) scorefile(`*string*`) restscorefile(`*string*`)`
`logisticfile(`*string*`) replace nodraw nodrawcomb onlyone(`*varname*`)`

thrhesholds(*string*)]

Options of the `traces` module

`score` displays the graphical representations of the traces of the items as a function of the total score. This is the default if none of the options `score`, `restscore`, or `logistic` are specified.

`restscore` displays the graphical representations of the traces of the items as a function of the rest-score (total score whithout the item).

`ci` displays the 95% confidence intervals of the traces.

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

`cumulative` displays cumulative traces for polytomous items instead of classical traces.

`logistic` displays the graphical representation of the logitic traces of the items as a function of the score: each trace is the result of a logistic model explaining the response to the item by the score (and a constant). This kind of trace is possible only for dichotomous items. All the logistic traces are represented in the same graph.

`repfile` defines the directory where the files should be saved.

`scorefiles` defines the generic name of the files containing the graphical representations of the traces as a function of the score. The name will be followed by the name of each item and by the .gph extension. If this option is not indicated, the corresponding graphs will not be saved.

`restscorefiles` defines the generic name of the files containing the graphical representations of the traces as a function of the rest-score. The name will be followed by the name of each item and by the .gph extension. If this option is not indicated, the corresponding graphs will not be saved.

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

`nodraw` does not display the graphs by items.

`nodrawcomb` does not display the combined graphs by items.

`replace` replaces graphical files when they already exist.

`onlyone` displays only the trace of a given item.

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

3.2 The `loevH` module

Syntax of the `loevH` module

The syntax of the `loevH` module is (version 7.1 is described here):

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

The Stata module `loevH` requires the modules `traces`, `anaoption` and `gengroup`.

Options of the `loevH` module

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

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

`ppp` displays the P++ matrix (and the associated graph with `graph`).

`pmm` displays the P- - matrix (and the associated graph with `graph`).

`noadjust` allows avoiding the adjustement by $N - 1$ for the test statistics (as in the MSP software).

`generror` defines the prefix of five new variables. The first one (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 independance of the item, the third one (the prefix followed by _H) the quantity 1 minus the ratio between the two preeceding 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, an histogram of the number of Guttman errors by individual is drawn.

`replace` allows replacing 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` displays outputs to check for the monotonicity assumption. It is possible to define in this option the value of `minvi` (only the violations of the monotonicity assumption greater than this value are considered as important, by default the used value is 0.03), the minimal size of each group of patients (`minsize`, 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 a minimum fixed to 50), the significance level (`siglevel` fixed by

default to 0.05). The `details` option allows printing more detailed outputs for polytomous items. It is possible to use all the default values by indicating a $*$ in the `monotonicity` option.

`nipmatrix` display indexes in order to check the non-intersection (Doubly Monotone Mokken Model). It is possible to define in this option the value of `minvi` (only the violations of the non-intersection assumption greater than this value are considered as important, by default the used value is 0.03), and the significance level (`siglevel` fixed by default to 0.05). It is possible to use all the default values by indicating a $*$ in the `nipmatrix` option.

Output of the `loevH` module

Scalars

`r(Obs)`	(matrix) number of individuals used to compute each coefficient H_{jk} (if the `pairwise` option is not used, the number of individuals are the same for each pair of items)	`r(eGuttjk)`	matrix of the numbers of observed Guttman errors associated to each items pair
`r(eGuttjk0)`	matrix of the numbers of theorical Guttman errors associated to each items pair	`r(eGuttj)`	vector of the total numbers of observed Guttman errors associated to the scale
`r(eGuttj0)`	vector of the total numbers of theorical Guttman errors associated to the scale	`r(eGutt)`	total number of observed Guttman errors associated to the scale
`r(eGutt0)`	total number of theorical Guttman errors associated to the scale	`r(loevHjk)`	matrix of the Loevinger's H coefficient for each pair of items
`r(loevHj)`	vector of the Loevinger's H coefficient to measure the integration of each item in the scale	`r(loevH)`	value of the Loevinger's H coefficient of scalability
`r(zHjk)`	matrix of the Z-statistics of the tests concerning the Loevinger's H coefficients for each pair of items	`r(zHj)`	vector of the Z-statistics of the tests concerning the Loevinger's H coefficients of each item in the scale
`r(zH)`	Z-statistics of the tests concerning the Loevinger's H coefficients of scalability	`r(pvalHjk)`	matrix of the p-values of the tests concerning the Loevinger's H coefficients for each pair of items
`r(pvalHj)`	vector of the p-values of the tests concerning the Loevinger's H coefficients of each item in the scale	`r(pvalH)`	p-values of the tests concerning the Loevinger's H coefficients of scalability
`r(P11)`	P++ matrix	`r(P00)`	P-- matrix

3.3 The `msp` module

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

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

The Stata module `msp` requires the module `loevH`.

Options of the `msp` module

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

`kernel` defines the #th first items as the kernel of the first sub-scale. The default is 0.

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

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

`pairwise` uses the `pairwise` option to compute the Loevinger's H coefficients.

`nobon` avoids the Bonferroni's corrections of the level of significance.

`notest` does not test the nullity of the Loevinger's H coefficient.

`nodetails` does not display the detail of the algorithm.

`noadjust` allows avoiding the adjustement by $N - 1$ for the test statistics (as in the MSP software).

Output of the `msp` module

Scalars

`r(dim)`	number of created scales	`r(H#)`	value of the Loevinger's H coefficient of scalability for the #th scale
`r(nbitems#)`	number of selected items in the #th scale	`r(scale#)`	list of the items selected in the #th scale (in the order of selection)
`r(lastitem)`	if only one item is remaining, the name of this item	`r(selection)`	a vector which contains, for each item, the number of the scale where this item is selected

3.4 Output

We present an example of outputs of these programs with items of the french adaptation of the Way of Coping Checklist (WCC) questionnaire Cousson et al. (1996). This questionnaire measures the coping strategies and contains 27 items which compose 3 dimensions: problem-focussed coping, emotional coping and seeking social support. The sample is composed of 100 women with a recent diagnostic of breast cancer.

Output of the module loevH

The **loevH** module allows obtaining the values of the Loevinger's H coefficients. Since the sample was small, it was impossible to obtain several groups of individuals with 50 individuals or more. As a consequence, for the `monotonicity` option, the `minsize`

has been fixed to 30. We studied the emotional dimension composed of 9 items (with 4 response categories of responses per item). The rate of missing data varied from 2% to 15% following the items and only 69 women have a complete pattern of responses, so the `pairwise` option has been prefered to keep a maximum of information.

```
. loevH  item2 item5 item8 item11 item14 item17 item20 item23 item26,pairw
> monot(minsize(30)) nip(*)
```

Item	Obs	Difficulty P(Xj=0)	Observed Guttman errors	Expected Guttman errors	Loevinger H coeff	z-stat.	H0: Hj<=0 p-value	Number of NS 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 correct 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 assumptionis not rejected (no important violation of this assumption, and the criteria are satisfied). This is not the case for the non-intersection of the Pmatrices curves: several criteria are

greater than 80 (items 5, 8 , 11, 14, 26) showing important violation of this assumption. The model followed by this data is therefore more a MHMM than a DMHMM. As the indices suggest that the MMHM holds, the score computed by summing codes associated to the 9 items can be considered as a correct ordinal measure of the studied latent trait (the emotional coping) and the three fundamental assumptions of the IRT (unidimensionality, local independence and monotonicity) can be considered as verified.

Output of the module msp

The **msp** module runs the Mokken Scale Procedure.

```
. msp item2 item5 item8 item11 item14 item17 item20 item23 item26,pairw
Scale: 1
─────────

Significance level: 0.001389
The two first items selected in the scale 1 are item2 and item11 (Hjk=0.6245)
Significance level: 0.001163
The item item17 is selected in the scale 1       Hj=0.5304      H=0.5748
Significance level: 0.001020
The item item5 is selected in the scale 1        Hj=0.5464      H=0.5588
Significance level: 0.000926
The item item8 is selected in the scale 1        Hj=0.4435      H=0.5073
Significance level: 0.000862
The item item20 is selected in the scale 1       Hj=0.3835      H=0.4684
Significance level: 0.000820
None new item can be selected in the scale 1 because all the Hj are lesser than
.3 or none new item has all the related Hjk coefficients significantly greater
than 0.
                      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       265    429.87    0.38354    6.3672    0.00000      0
item8    90     0.3667       286    508.92    0.43802    7.7902    0.00000      1
item5    92     0.3261       225    496.03    0.54640    9.1467    0.00000      0
item17   94     0.7660       168    291.14    0.42296    5.9433    0.00000      1
item2    92     0.2935       261    485.41    0.46231    7.5382    0.00000      0
item11   97     0.5670       249    523.75    0.52458    9.3138    0.00000      0
──────────────────────────────────────────────────────────────────────────────
Scale   100                  727   1367.56    0.46839   13.4151    0.00000

Scale: 2
─────────

Significance level: 0.016667
The two first items selected in the scale 2 are item23 and item26 (Hjk=0.4391)
Significance level: 0.012500
The item item14 is selected in the scale 2       Hj=0.4276      H=0.4313
Significance level: 0.012500
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
```

The AISP procedure creates two groups of items.

On the first hand, the items 2 "Wish that the situation would go away or somehow be over with", 5 "Wish that I can change what is happening or how I feel", 8 "Accept it, since nothing can be done", 11 "Hope a miracle will happen", 17 "I daydream or imagine a better time or place than the one I am in" and 20 "Try to forget the whole thing" concerns items which measures the negation or the wish to forget the reason of the stress. For this set, the scalability coefficient is good (0.47), and there is no problem concerning the monotonicity assumption (maximal criterion per item of -4), nor intersection of the curves (maximal criterion per item of 53). This set seems to satisfy a DMHMM and is composed of 6 of the 11 items composing the "Wishful thinking" and "Detachment" dimensions proposed by Folkman and Lazarus (1985) in an analysis of the WCC among a sample of students.

On the other hand, the items 14 "Realize I brought the problem on myself", 23 "Make a promise to myself that things will be different next time" and 26 "Criticize or lecture myself" concerns items which measures the culpability. 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 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).

We can note that, if we fix the two first items selected in the second set of items as a kernel in the AISP (with the syntax `msp item23 item26 item2 item5 item8 item11 item14 item17 item20,pairw kernel(2)`), all the items are selected in the same set of items. This process is recommanded by Hemker et al. (1995) to confirm or reject the structure found by the AISP. In our case, it is possible to choose between a set of items satisfying a MHMM and two sets of items satisfying each a DMHMM. Since the three sets of items are interpretable (emotional coping for the set of items satisfying MHMM, negation and culpability for the two others sets of items), there is no problem to choose freely following the whished precision of the measured concepts. In the optic of the validation of the questionnaire, it is preferable to chose the set of items containing all the items satisfying the emotional coping, which is closer to the former questionnaire.

4 References

van der Ark, A. 2007. Mokken Scale Analysis in R. *Journal of Statistical Sofware* 20(11): 1–19.

Cousson, F., M. Bruchon-Schweitzer, B. Quintard, J. Nuissier, and N. Rascle. 1996.

Analyse multidimensionnelle d'une chelle de coping : validation française de la Way of Coping Checklist. *Psychologie Franaise* 41(2): 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 LoevH SAS macro-program*. University of Nantes, http://sasloevh.anaqol.org.

———. 2007. Rasch analysis: estimation and tests with the raschtest module. *The Stata Journal* 7(1): 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(4): 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.

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

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*. De Gruyter.

———. 1997. *Nonparametric models for dichotomous responses*, chap. 20, 351–368. New York: Springer Verlag. In W. J. Van der Linden and R. K. Hambleton : Handbook of Modern Item Response Theory.

Mokken, R. J., and C. Lewis. 1982. A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement* 6(4): 417–430.

Molenaar, I. W. 1997. *Nonparametric models for polytomous responses*, chap. 21, 369–380. New York: Springer Verlag. In W. J. Van der Linden and R. K. Hambleton : Handbook of Modern Item Response Theory.

Molenaar, I. W., K. Sijtsma, and P. Boer. 2000. *MSP5 for Windows: A program for Mokken scale analysis for polytomous items - version 5.0*. iec ProGAMMA, Groningen, The Netherlands.

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

Weesie, J. 1999. *MOKKEN: Stata module: Mokken scale analysis*. RePEc EconPapers, 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. *The Stata Journal* 7(3): 313–333.