

RespOnse Shift ALgorithm in Item response theory (ROSALI) for response shift detection with missing data in longitudinal patient-reported outcome studies

Alice Guilleux · Myriam Blanchin · Antoine Vanier · Francis Guillemin ·
Bruno Falissard · Carolyn E. Schwartz · Jean-Benoit Hardouin · Véronique Sébille

Accepted: 20 November 2014
© Springer International Publishing Switzerland 2014

Abstract

Purpose Some IRT models have the advantage of being robust to missing data and thus can be used with complete data as well as different patterns of missing data (informative or not). The purpose of this paper was to develop an algorithm for response shift (RS) detection using IRT models allowing for non-uniform and uniform recalibration, reprioritization RS recognition and true change estimation with these forms of RS taken into consideration if appropriate.

Methods The algorithm is described, and its implementation is shown and compared to Oort's structural equation modeling (SEM) procedure using data from a clinical study assessing health-related quality of life in 669 hospitalized patients with chronic conditions.

Results The results were quite different for the two methods. Both showed that some items of the SF-36 General Health subscale were affected by response shift,

but those items usually differed between IRT and SEM. The IRT algorithm found evidence of small recalibration and reprioritization effects, whereas SEM mostly found evidence of small recalibration effects.

Conclusion An algorithm has been developed for response shift analyses using IRT models and allows the investigation of non-uniform and uniform recalibration as well as reprioritization. Differences in RS detection between IRT and SEM may be due to differences between the two methods in handling missing data. However, one cannot conclude on the differences between IRT and SEM based on a single application on a dataset since the underlying truth is unknown. A next step would be to implement a simulation study to investigate those differences.

Keywords Item response theory · Response shift · Missing data · Attrition · Bias · Quality of life

A. Guilleux · M. Blanchin · A. Vanier · J.-B. Hardouin ·
V. Sébille (✉)
Biostatistics, Pharmacoepidemiology and Subjective Measures
in Health Sciences, EA 4275, University of Nantes, Nantes,
France
e-mail: veronique.sebille@univ-nantes.fr

A. Vanier
Department of Biostatistics, UPMC, Univ. Paris 06, Paris,
France

A. Vanier
Department of Biostatistics Public Health and Medical
Informatics, AP-HP, University Hospitals Pitié-Salpêtrière
Charles-Foix, Paris, France

F. Guillemin
EA 4360 Apemac, Lorraine University, Paris Descartes
University, Nancy, France

B. Falissard
INSERM 669, Université Paris-Sud, Université Paris Descartes,
Paris, France

B. Falissard
Département de santé publique, AP-HP, Hôpital Paul Brousse,
Villejuif, France

C. E. Schwartz
DeltaQuest Foundation, Inc., 31 Mitchell Road, Concord,
MA 01742, USA

C. E. Schwartz
Department of Medicine and Orthopedic Surgery, Tufts
University School of Medicine, Boston, MA, USA

J.-B. Hardouin · V. Sébille
Clinical Research Unit, Department of Methodology and
Biostatistics, University Hospital of Nantes, Nantes, France

Introduction

Response shift (RS) is an important issue in clinical research, and missing data remain a challenge when assessing longitudinal patient-reported outcome (PRO) data and possibly associated response shift (RS) phenomenon. Indeed, most proposed analytical statistical strategies, such as structural equation modeling (SEM) [1] allowing for the detection of all forms of RS (recalibration, reprioritization and reconceptualization), assume that data are either missing completely at random (MCAR) or missing at random (MAR). In these cases, the missingness probability either depends on the observed data (MAR) or is independent of all previous, current and future assessments (MCAR). However, it is very likely that most data are not missing at random (MNAR) [2] and that the probability of missingness depends on unobserved data (e.g., patients might be too tired to fill in the PRO on fatigue). This type of data is truly problematic since it can lead to very poor and biased estimates of RS and true change in PRO data [3]. The issues regarding the choice of the most appropriate methodological approach for both the identification of RS occurrence and its appropriate adjustment in the analyses of longitudinal PRO data with possibly informative missing data (MNAR data) are challenging and remain debated.

Item response theory (IRT), and in particular Rasch family models [4], could be an interesting alternative for response shift detection, with some advantages compared to SEM. Indeed, with IRT it is possible to estimate a latent trait with interval scale property, unlike raw scores or their linear transformations [5]. That is, a unit difference characterizes the same amount when measured from different initial levels on the latent trait scale. Furthermore regarding the management of missing data, Rasch-based IRT models possess a very interesting property of *specific objectivity*, which allows one to obtain consistent estimates of the parameters associated with the latent trait (quality of life for instance) whether or not an item is observed [6, 7]. Consequently, unbiased estimates of the latent trait can be obtained even when some items are missing, in a framework that can be ignorable (MCAR or MAR data) or not (MNAR data) [8–11]. Rasch family models could therefore provide a valid methodological approach for RS identification in longitudinal studies with potentially informative missing data.

The purpose of this paper was to develop an algorithm for RS detection using IRT models allowing for non-uniform and uniform recalibration, reprioritization RS recognition and true change estimation with these forms of RS taken into consideration if appropriate. The algorithm follows the sequence of the algorithm of Oort's procedure for SEM and relies on longitudinal polytomous IRT models including the partial credit model (PCM) of the Rasch family model and

the generalized partial credit model (GPCM). The algorithm is described, and its implementation is shown and compared to Oort's SEM procedure using data from a clinical study assessing health-related quality of life in hospitalized patients with chronic conditions.

Methods

Algorithm for response shift detection using IRT models

We propose a new algorithm for response shift analyses using polytomous IRT models following the sequence of the algorithm of Oort's procedure for SEM [1]: the RespOnse Shift ALgorithm in Item response theory (ROSALI). This algorithm makes it possible to detect two types of RS using data from two measurement occasions: non-uniform and uniform recalibration and reprioritization. After this detection, the true change (the mean change of the latent trait between the two times) can be estimated taking RS into account. Unlike the Oort's SEM method, ROSALI is currently based on unidimensional IRT models and does not yet include the possibility of reconceptualization detection which would require multidimensional IRT modeling.

Finally, the observed change can be separated into two components: change due to response shift and true change in the level and variability of the latent construct the PRO is intended to measure. The different steps of the algorithm are the following: a preliminary step for estimating the item difficulties at the first measurement occasion (step 0), establishing a measurement model (step 1), fitting a model with no RS and overall evaluation of RS (step 2), RS detection (step 3) and true change estimation (step 4). Figure 1 summarizes each step for both methods, IRT and SEM.

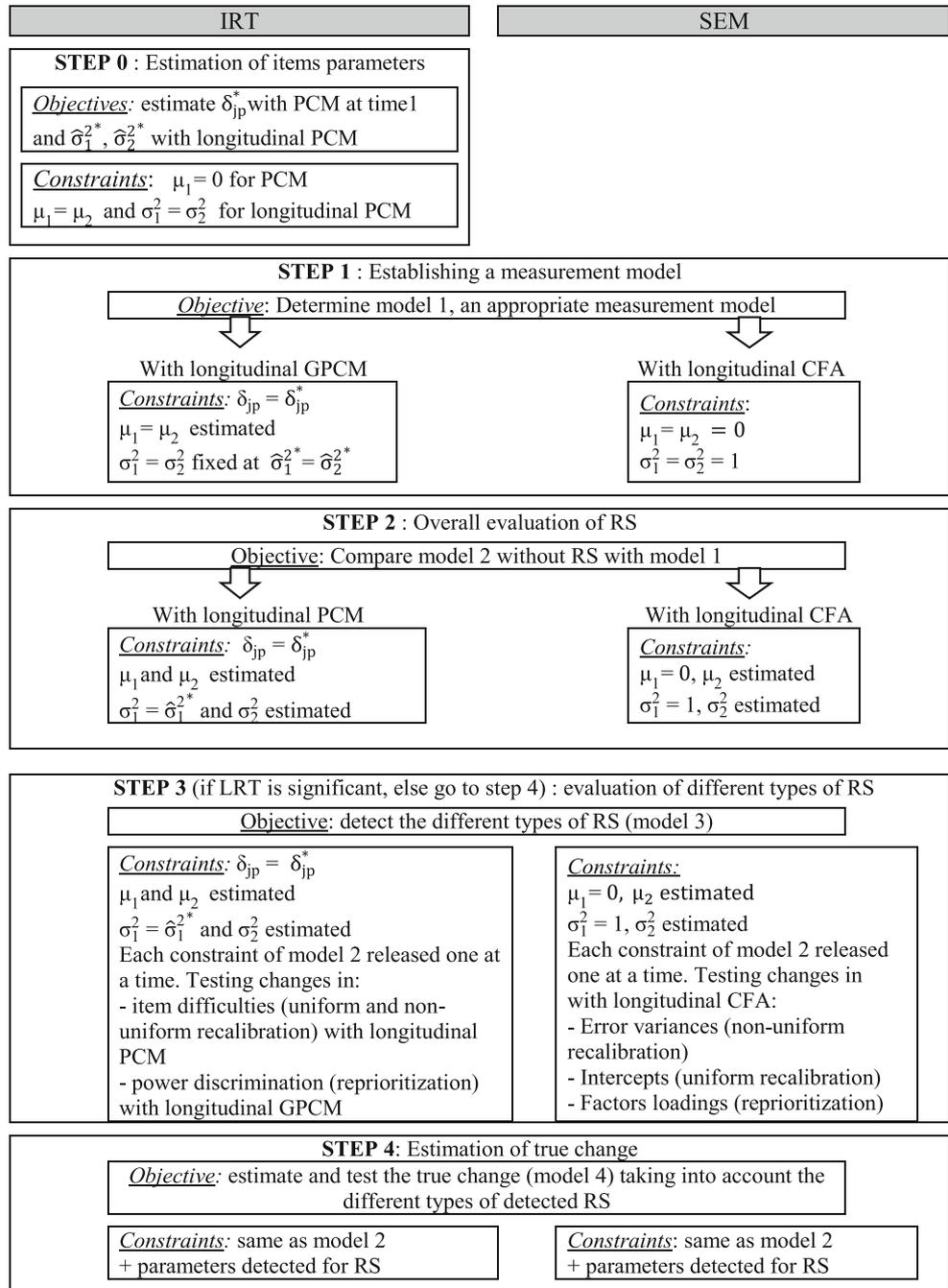
Item response theory models

Suppose patients' quality of life is measured using a questionnaire which includes polytomous items on two measurement occasions t ($t = 1, 2$). The presentation of the IRT algorithm will focus on one dimension of this questionnaire assumed to be composed of J items.

The items' responses of the patients on the two measurement occasions can be modeled using a longitudinal GPCM as follows:

$$P\left(X_{ij}^{(t)} = h|\theta_i^{(t)}, \delta_{j1}, \dots, \delta_{jm_j}, \eta_{j1}^{(t)}, \dots, \eta_{jm_j}^{(t)}, \alpha_1^{(t)}, \dots, \alpha_j^{(t)}\right) \\ = \frac{\exp\left(\alpha_j^{(t)}\left(h\theta_i^{(t)} - \left[\sum_{p=1}^h \delta_{jp} + \eta_{jp}^{(t)}\right]\right)\right)}{\sum_{l=0}^{m_j} \exp\left(\alpha_j^{(t)}\left(l\theta_i^{(t)} - \left[\sum_{p=1}^l \delta_{jp} + \eta_{jp}^{(t)}\right]\right)\right)} \quad (1)$$

Fig. 1 Steps and constraints for each method: SEM and IRT. *PCM* partial credit model, *GPCM* generalized partial credit model, *CFA* confirmatory factor analysis, *IRT* item response theory, *SEM* structural equation modeling, *RS* response shift, *LRT* likelihood ratio test



$$\begin{bmatrix} \theta^{(1)} \\ \theta^{(2)} \end{bmatrix} \sim N \left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}; \Sigma \right)$$

with $X_{ij}^{(t)}$ the response of patient i ($i = 1, \dots, N$) to item j ($j = 1, \dots, J$) at time t ($t = 1, 2$), $\theta_i^{(t)}$ a realization of the random variable θ , that is the latent trait of patient i at time t . Σ is the covariance matrix $\begin{bmatrix} \sigma_1^2 & \sigma_{2,1} \\ \sigma_{1,2} & \sigma_2^2 \end{bmatrix}$ where σ_1^2 and σ_2^2 are the variances of the latent trait at times $t = 1$ and $t = 2$, respectively, and $\sigma_{1,2} = \sigma_{2,1}$ is the covariance

between the latent traits at times $t = 1$ and $t = 2$. The parameter δ_{jp} corresponds to the item difficulty for each positive category p of item j , and the number of positive response categories for this item is equal to m_j . $\eta_{jp}^{(t)}$ is the change in item difficulties between time $t = 1$ and time $t = 2$ with $\eta_{jp}^{(1)} = 0 \forall j, p$. $\alpha_j^{(t)}$ are the discriminating powers at time $t = 1$ and $t = 2$, respectively, for each item $j = 1, \dots, J$. In order to obtain a PCM at time $t = 1$, the following constraints are used: $\alpha_j^{(1)} = 1 \forall j = 1, \dots, J$.

If no RS is assumed, $\alpha_j^{(t)} = 1$ for all values of t ($t = 1, 2$) and j ($j = 1, \dots, J$), and the former longitudinal GPCM becomes a longitudinal PCM. In this model, only characteristics of the latent trait (level of the individuals on the measured concept) can vary over time, unlike the item parameters which are held fixed.

Response shift can be considered as a modification of the perception of the questionnaire over time by the patients. If RS is assumed, the most flexible model allowing for response shift is used to model the data: the longitudinal GPCM (Eq. 1) which includes parameters allowing for change in item difficulties and in discriminating powers between time $t = 2$ and time $t = 1$ for each item $j = 1, \dots, J$ to allow for modeling possible non-uniform and/or uniform recalibration and/or reprioritization response shift.

Modeling and interpretation of different sources of response shifts

Recalibration Recalibration is considered as change in the respondent's internal standards of measurement [12]. For non-uniform recalibration (Fig. 2), it is assessed by a change of item difficulties (parameter $\eta_{jp}^{(t)}$ with $\eta_{jp}^{(1)} = 0 \forall j, p$) that can occur in various directions and magnitude toward lower or greater difficulties. For uniform recalibration (Fig. 3), this phenomenon is assessed by a change of all difficulties of a given item in the same direction and to the same extent. Figures 2 and 3 show the two types of recalibration (non-uniform and uniform) when IRT was applied on the clinical data from the illustrative example. For both, recalibration involves a change of item difficulties between the two times, $\eta_{jp}^{(2)}$. For example, if we observe a non-uniform recalibration on item j , each curve on the graph can be shifted in different directions and/or magnitude. If we observe a uniform recalibration on item j , each curve on the graph will be shifted at time $t = 2$ by the same amount.

Reprioritization Reprioritization corresponds to a change in the respondent's values, that is to say a shift in the importance of items constituting the target construct. The change in discriminating power for each item between the two times $\alpha_j^{(2)}$ can be used to assess this type of RS. If there is reprioritization on an item, the value of its discriminating power, which is equal to 1 at time $t = 1$, is different from 1 at time $t = 2$ (Fig. 4 illustrates the application of IRT on the clinical data from the illustrative example). In this case, this item has become more or less discriminating at time $t = 2$ than at time $t = 1$. For example, for item j , if its discriminating power is higher than 1 at time $t = 2$, item j is more predictive of the latent trait level, and conversely, if it is lower than 1 at time $t = 2$, item j is less predictive of the latent trait level.

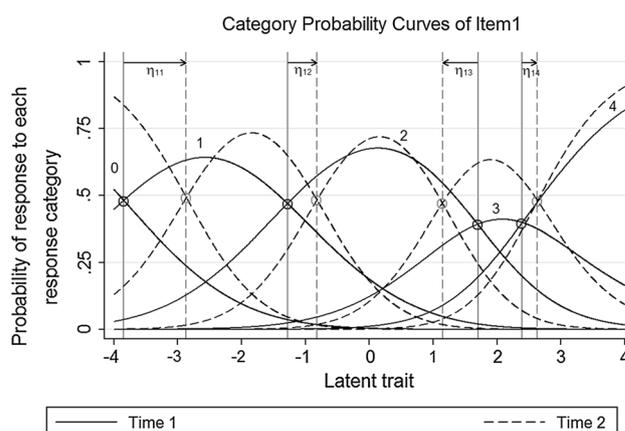


Fig. 2 Category probability curves for non-uniform recalibration on item 1 for the SatisQoL data. Item 1 is one of the items of the General Health (GH) subscale of the SF-36: “In general, would you say your health is” with responses: Excellent/Very good/Good/Fair/Poor. The parameter η_{1p} is the change in item difficulties between time $t = 2$ and time $t = 1$ for item 1 with $\eta_{1p} = 0$ at time $t = 1$ for all p ($p = 1, \dots, 4$)

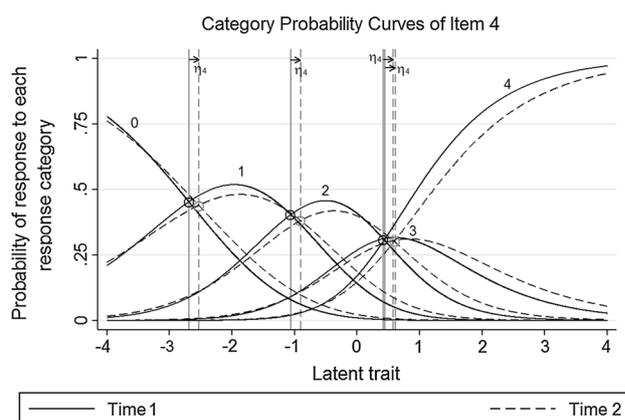


Fig. 3 Category probability curves for uniform recalibration on item 4 for the SatisQoL data. Item 4 is one of the items of the General Health (GH) subscale of the SF-36: “I expect my health to get worse” with responses: Definitely true/Mostly true/Don't know/Mostly false/Definitely false. The parameter η_4 is the change in item difficulties between time $t = 2$ and time $t = 1$ for item 4 with $\eta_4 = 0$ at time $t = 1$ for all p ($p = 1, \dots, 4$)

Procedure

The procedure was developed with the SAS software to obtain an algorithm for RS detection using IRT models. It is composed of 5 steps and 4 models. A preliminary step (step 0) was added to estimate item parameters at time $t = 1$.

Step 0: estimation of item parameters

This preliminary step allows estimating the item difficulties at time $t = 1$ δ_{jp} by fitting a PCM whose fit is assessed

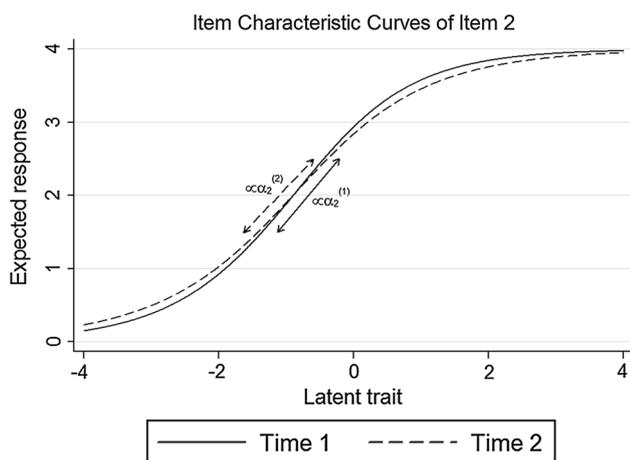


Fig. 4 Item characteristic curves on item 2 for reprioritization for the SatisQoL data. Item 2 is one of the items of the General Health (GH) subscale of the SF-36: “I seem to get sick a little easier than other people” with responses: Definitely true/Mostly true/Don’t know/Mostly false/Definitely false. The parameter $\alpha_2^{(t)}$ is the discriminating power at time t ($t = 1, 2$) for item 2

using the MIRT software [13, 14]. For all the next steps of the algorithm, the item difficulties are fixed to their estimated values at step 0.

Step 1: establishing a measurement model (model 1)

Model 1 is a GPCM (Eq. 1) which takes into account the two possible types of RS: recalibration (non-uniform or uniform) and reprioritization. The discriminating powers are all equal to 1 at time $t = 1$, and their values are estimated at time $t = 2$ as well as the change in item difficulties between the two times $\eta_{jp}^{(2)}$.

Step 2: overall evaluation of response shift (model 2)

Model 2 is a model assuming no RS. It is a longitudinal PCM because we impose invariance constraints such that we assume no non-uniform or uniform recalibration and no reprioritization. To evaluate the presence of overall RS, we compare model 1 and model 2 using a likelihood ratio test. If the test is significant, we go to step 3 to improve model 2 and detect which type of RS occurs on which item. If the test is not significant, we may assume that there is no RS, and we skip step 3.

Step 3: response shift detection (model 3)

Step 3 is an iterative step in which model 3 is updated constantly according to the results. The first model used at this step corresponds to model 2, and we remove the constraints of model 2 one by one to identify RS. The algorithm first tests the presence of recalibration on each item j . For

each j , a test to detect whether a change in item difficulties, estimated by $\eta_{jp}^{(2)}$, has occurred is performed using the likelihood ratio test to compare the new model with the previous one. If the test is significant, we consider the presence of recalibration on item j . Among the items for which significant tests are found, we choose the item associated with the model with the maximum value for the likelihood. Then, we determine the type of recalibration on this item by testing whether the item difficulties change in the same direction and to the same magnitude. If this test is significant, non-uniform recalibration is suspected, else uniform recalibration is assumed. Finally, the model is updated to take into account recalibration if appropriate. This first part of step 3 may be repeated on the updated model until there are no more items displaying recalibration remaining, and we obtain a model taking into account all non-uniform and uniform recalibrations that have been detected.

After detecting recalibration, the algorithm determines whether there is reprioritization by looking at the change in discriminating powers between the two times. To begin, a global test is performed to determine whether there is reprioritization on some items by testing whether all discriminating powers on all items are equal to 1 at time $t = 2$ using a likelihood ratio test. If the test is significant, we consider that reprioritization has occurred on at least one item. To locate the involved item(s), a likelihood ratio test is performed on the discriminating power of each item to determine whether there is reprioritization. Among the significant tests, we choose the item associated with the model which maximizes the likelihood. Finally, the model is updated to take into account the detected reprioritization. This part of step 3 may be repeated until there are no more items displaying reprioritization remaining.

Step 4: true change assessment (model 4)

During the last step, the true change is estimated and tested. The true change is evaluated using model 4 that contains the significant parameters found in step 3 for RS adjustment. True change is evidenced if the difference of the means of the latent variable between the two times of the study is significantly different from 0. The test of true change is performed with a Wald test.

Response shift detection using SEM

Detection of RS using SEM was performed using Oort’s procedure [1]. This procedure has four steps: (1) establishing an appropriate measurement model, (2) fitting a no RS model, (3) RS detection, and (4) assessment of true change. Each of these steps is associated with a particular longitudinal confirmatory factor analysis model. SEM models were fitted

using robust maximum-likelihood estimator with a Satorra–Bentler correction [15]. SEM analyses were conducted using lavaan 0.5-13 package [16] for R software 3.0.1 [17].

Step 1: establishing a measurement model (model 1)

The appropriate measurement model (model 1) is established following the structure of the chosen questionnaire. No across time constraints are imposed. A root mean square error of approximation (RMSEA) close to 0.05 (p of close fit >0.10), comparative fit index (CFI) ≥ 0.95 and standardized root mean residual (SRMR) ≤ 0.05 are used as indicators of good fit [18].

Step 2: overall evaluation of response shift (model 2)

Model 2 is a model assuming no RS in which all RS parameters are constrained to be equal across times. Model 1 and model 2 are compared using a likelihood ratio test. If the test is significant, we go to step 3 to improve model 2 and detect which type of RS occurs on which item. If the test is not significant, we may assume that there is no RS, and we skip step 3.

Step 3: response shift detection (model 3)

Untenable constraints on RS parameters are released one at a time, starting from model 2. Each modification is tested by likelihood ratio tests [19]. Specification search is guided using modification indices. Releasing constraints on error variances (non-uniform recalibration) is tested first, followed by intercepts (uniform recalibration) and factor loadings (reprioritization).

Step 4: true change assessment (model 4)

A final model is estimated, in which differences in factor means are indicative of “true change” after accounting for RS.

Handling of missing data

Missing data are not handled in the same way if IRT models or SEM are used. With IRT models, there is no need for imputation, which is usually associated with some unverifiable assumptions, and it is possible to use these models on all available data even if some items responses are missing. This “available case” approach was chosen for IRT models since Rasch family models were shown to provide unbiased estimations and good power for MCAR but also MNAR data [7, 10]. With SEM, several strategies can be used: complete case analysis (observations with any missing values are discarded) [20], assuming MCAR data, full information

maximum likelihood assuming MCAR or MAR data as well as normality [21], or some data imputation technique under ignorable missing data conditions (MCAR or MAR data) [20]. For the latter, robust maximum-likelihood estimator with a Satorra–Bentler correction can be used when data are not assumed to be normally distributed [15]. This last approach was used since SEM analyses were performed at the item level to be comparable with IRT analyses.

Application of the IRT- and SEM-based procedures for RS detection on clinical data—an illustrative example

The algorithm that was developed for RS detection using IRT models was applied on a clinical dataset and compared to Oort’s procedure for SEM.

Study sample and data collection procedures

The application of both procedures based on IRT or SEM was performed on a subsample of the SatisQoL study. The SatisQoL study is a French multicenter (3 centers) cohort study designed to assess the relationships between satisfaction with care and health-related quality of life (HRQL) after being hospitalized in a university hospital for a medical or surgical intervention related to a chronic disease [22].

The exclusion criteria were the following: patients under 18 or above 75 years old, patients not suffering from a chronic disease for less than 6 months at initial admission, no medical or surgical intervention during hospitalization. Patients were asked to fill in a variety of questionnaires (including HRQL measurement) shortly after admission, and at 6 months after discharge. In this study, we focused on patients who underwent surgery which was believed to initiate a response shift (catalyst).

Main outcome

HRQL was assessed at baseline and 6 months after discharge using the SF-36 version 1.3 in French [23, 24]. As it was necessary to work on a unidimensional psychometric construct to allow comparisons of RS detection techniques between SEM and IRT, it was decided to restrict the application of both procedures to the general health (GH) dimension of the SF-36.

Missing data

Handling of missing data in this study is depicted on Fig. 5. For both methods (IRT and SEM), the 91 patients (13.6 %) who did not respond to any of the 5 items of the GH subscale on one or two of the measurement occasions were excluded from the analyses. Missing data were subsequently handled in a

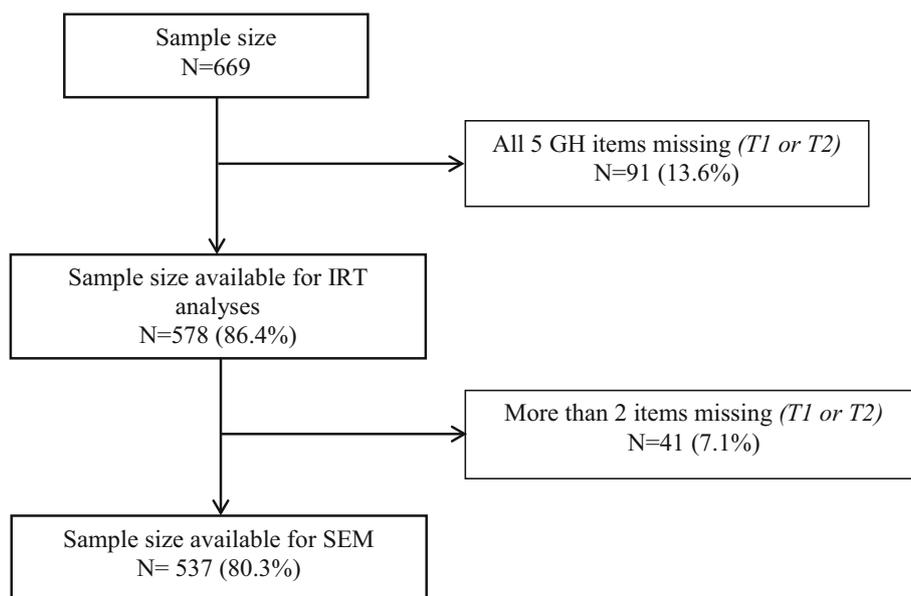
different way according to whether IRT or SEM was used. For IRT analyses, no imputation was done, and all available remaining data were used: The IRT analyses were thus performed on 578 patients, corresponding to 86.4 % of the initial sample size. For SEM analyses, missing data were handled according to guidelines of the SF-36 manual [25]. As such, if no more than 2 items were missing at one time of measurement for a patient, missing items were imputed by the mean of other responses at the corresponding time of measurement. Therefore, if there were more than 2 items missing at any time of measurement for a patient, he/she was excluded from SEM analyses. Hence, 537 patients (80 % of the initial sample size) remained for SEM analyses since 41 patients had more than two items missing at the first or second time of measurement. For SEM, in accordance with the guidelines of the SF-36 manual, imputation by the mean was performed on 38 patients ($38/537 = 7.1\%$), as responses to one or two items were missing to at least one time of measurement for these 38 patients. Table 1 presents the characteristics of participants with and without missing data.

Results

Sample characteristics

Table 2 summarizes characteristics of the 669 patients included in the SatisQoL study. The average age was 55 years, and 356 (53.2 %) were men. These 669 patients went through various surgical procedures belonging to 11 medical areas. The average observed GH score was 58.5 at baseline, almost equal to the average GH score at 6 months after discharge (58.1).

Fig. 5 Flow chart depicting missing data patterns for the General Health (GH) subscale of the SF-36



Detection of response shift using IRT and SEM

Step 1—model 1

For IRT, a measurement model (longitudinal GPCM, model 1) was established for the 5 items of the GH subscale dimension. The fit of model 1 was correct (p value = 0.34). For SEM, a longitudinal measurement model with the 5 items of the GH subscale loading on one dimension at each time point led to RMSEA (0.054, p of close fit = 0.32), CFI (0.982) and SRMR (0.038) being below the desired cutoff.

Step 2—model 2 (no RS model)

For IRT and SEM, the test of overall RS was significant ($p < 10^{-4}$ for IRT and $p < 0.016$ for SEM) which was considered as overall evidence for RS for both procedures. Consequently, step 3 was realized for IRT and SEM to determine the type of RS.

Step 3—model 3

For IRT, non-uniform recalibration on item 1 “In general, would you say your health is ...” (Fig. 2) and uniform recalibration on item 4 “I expect my health to get worse” (Fig. 3) of the GH subscale were detected. Moreover, all items seemed affected by reprioritization (example given for item 2 “I seem to get sick a little easier than other people” in Fig. 4). For SEM, non-uniform recalibration was detected in two different items of the GH subscale as compared to IRT: items 2 and 4; uniform recalibration was

Table 1 Comparison of patient characteristics with and without missing data at baseline and at 6 months

Characteristic, measure	Sample without missing data (n = 499)	Sample with missing data (n = 170)
Age (years), mean \pm SD	54.3 \pm 13.5	57.5 \pm 13.4
Gender (M/F), n (%)	265/234 (53.1/46.9)	91/79 (53.5/46.5)
<i>Medical diagnosis n (%)</i>		
ENT–Ophthalmology	103 (20.6)	36 (21.1)
Gastrointestinal	90 (18.0)	30 (17.6)
Rheumatology	92 (18.4)	27 (15.8)
Circulatory system	60 (12.0)	40 (23.5)
Urology–Nephrology	61 (12.0)	12 (7.1)
Others	93 (18.6)	24 (14.1)
<i>SF-36 General Health (GH) score (/100), mean \pm SD</i>		
Baseline	59.5 \pm 20.9	55.2 \pm 22.6
6 months after	58.8 \pm 22.1	51.5 \pm 26.2

Table 2 Patient characteristics at baseline and SF-36 scores on the General Health (GH) subscale at baseline and at 6 months

Characteristic, measure	Sample (n = 669)
Age (years), mean \pm SD	55.1 \pm 13.5
Gender (M/F), n (%)	356/313 (53.2/46.8)
<i>Medical diagnosis n (%)</i>	
ENT–Ophthalmology	139 (20.8)
Gastrointestinal	120 (17.9)
Rheumatology	119 (17.8)
Circulatory system	100 (14.9)
Urology–Nephrology	73 (10.9)
Others	118 (17.7)
<i>SF-36 General Health (GH) score (/100), mean \pm SD</i>	
Baseline	58.5 \pm 21.3
6 months after	58.1 \pm 22.6

similarly found on item 4 but also on item 3 “I am as healthy as anybody I know” of the GH subscale. Moreover, reprioritization was only found on item 5 “My health is excellent” for SEM.

Step 4—model 4

For IRT, the estimation of true change in model 4 was nearly the same as in model 2 with no RS (0.027 vs. -0.014). RS seemed to have a very small impact on the true change estimation. Both true changes estimated in the GH subscale dimension for model 2 (with no RS) and for model 4 (RS accounted for) were not significantly different from 0 at $\alpha = 5\%$. For SEM, change in factor means over time in model 4, as indicative of true change in GH after

taking RS into account, was slightly larger in absolute value than in model 2 (-0.034 vs. -0.025). Both of them were non-significantly different from 0 at $\alpha = 5\%$.

The parameter estimates for IRT and SEM applied on the SatisQoL data appear in Table 3.

Discussion

We proposed an algorithm, the RespOnse Shift ALgorithm in Item response theory (ROSALI), following the sequence of the algorithm proposed by Oort for SEM, to detect response shift. IRT and, in particular, Rasch family models are robust to missing data and thus can be used with complete data as well as all patterns of missing data [7, 10]. This procedure allows detecting two types of RS: non-uniform and uniform recalibration as well as reprioritization. The true change can be subsequently estimated after taking RS into consideration, if appropriate. The technical feasibility of the IRT algorithm was assessed through its application on a clinical dataset, and the results were compared to the ones obtained using Oort’s procedure for SEM on the same data. We may note that usually, the methods used to detect RS are applied at the dimension level and that one of the consequences of the use of IRT models is that we worked at the item level; both applications presented here, comparing IRT and SEM, were thus performed at the item level.

The whole procedure of the IRT algorithm could be performed and applied on the clinical data as well as Oort’s procedure for SEM. The results were quite different for the two methods (Table 4) and showed that some items of the GH subscale of the SF-36 were affected by response shift and that they usually differed according to the chosen approach (IRT or SEM). Non-uniform recalibration was found on item 1 for IRT, whereas it was found on item 2 and item 4 for SEM. Uniform recalibration was found on item 4 for IRT and SEM, but it was also detected on item 3 using SEM. Finally, reprioritization was evidenced on all items of the GH subscale for IRT: items 1, 3 and 5 became more predictive of the latent trait level at 6 months, while items 2 and 4 became less predictive of the latent trait level. However, in SEM, reprioritization was only detected on item 5, and it went in an opposite direction as compared to IRT. Indeed, this item became less predictive of the latent trait level at 6 months. We cannot rule out, for IRT, the possibility that change in item difficulties’ parameters with time (recalibration) can affect the slope of the item characteristic curves and that it could have an influence on the discrimination parameters values. Consequently, recalibration could possibly create an artificial detection of reprioritization. However, true change estimations were close for both methods, and both were not significantly different from 0.

Table 3 Parameter estimates for item response theory (IRT) and structural equation modeling (SEM) applied on the SatisQoL data

Item	Response category	IRT					SEM					
		Time 1		Time 2			Time 1			Time 2		
		δ_{jp}	α_j	$\delta_{jp} + \eta_{jp}^a$	$\delta_{jp} + \eta_{jp}^b$	α_j^c	Intercepts ^a	Error ^b variances	Factor ^c loadings	Intercepts	Error variances	Factor loadings
Item 1			1			1.67	3.37	0.32	0.62	3.37	0.32	0.62
	1	-3.85		$\delta_{jp} + 0$	$\delta_{jp} + 0.98$							
	2	-1.28		$\delta_{jp} + 0$	$\delta_{jp} + 0.46$							
	3	1.7		$\delta_{jp} + 0$	$\delta_{jp} - 0.55$							
	4	2.39		$\delta_{jp} + 0$	$\delta_{jp} + 0.24$							
Item 2			1			0.78	3.72	1.09	0.62	3.72	0.94	0.62
	1	-2.11		$\delta_{jp} + 0$	$\delta_{jp} + 0$							
	2	-1.25		$\delta_{jp} + 0$	$\delta_{jp} + 0$							
	3	-0.47		$\delta_{jp} + 0$	$\delta_{jp} + 0$							
	4	0.12		$\delta_{jp} + 0$	$\delta_{jp} + 0$							
Item 3			1			1.38	3.23	0.53	0.85	3.33	0.53	0.85
	1	-2.24		$\delta_{jp} + 0$	$\delta_{jp} + 0$							
	2	-0.5		$\delta_{jp} + 0$	$\delta_{jp} + 0$							
	3	-0.38		$\delta_{jp} + 0$	$\delta_{jp} + 0$							
	4	1.91		$\delta_{jp} + 0$	$\delta_{jp} + 0$							
Item 4			1			0.84	3.43	1.03	0.67	3.33	0.87	0.67
	1	-2.69		$\delta_{jp} + 0.16$	$\delta_{jp} + 0$							
	2	-1.06		$\delta_{jp} + 0.16$	$\delta_{jp} + 0$							
	3	0.42		$\delta_{jp} + 0.16$	$\delta_{jp} + 0$							
	4	0.45		$\delta_{jp} + 0.16$	$\delta_{jp} + 0$							
Item 5			1			1.78	2.98	0.36	1.04	2.98	0.36	0.96
	1	-1.48		$\delta_{jp} + 0$	$\delta_{jp} + 0$							
	2	-0.25		$\delta_{jp} + 0$	$\delta_{jp} + 0$							
	3	-0.15		$\delta_{jp} + 0$	$\delta_{jp} + 0$							
	4	2.43		$\delta_{jp} + 0$	$\delta_{jp} + 0$							
Mean of latent trait		0.004					0			-0.034		
				0.017								
Variance of latent trait		1.239					1			1.193		
				1.329								

δ_{jp} : item difficulty for each positive category p of item j ; η_{jp} : change in item difficulties between time $t = 2$ and time $t = 1$ for item j with $\eta_{jp} = 0$ at time $t = 1$ for all p ; α_j : discriminating power at time t ($t = 1, 2$) for item j with $\alpha_j = 1$ at time $t = 1$ for all j . Item 1–5 are the items of the General Health (GH) subscale of the SF-36

Bold indicate the values are significantly different between time 1 and time 2

^a Uniform recalibration

^b Non-uniform recalibration

^c Reprioritization

Several reasons might explain some of the differences obtained between the two methods. Firstly, these methods do not handle missing data in the same way. For SEM, missing data were imputed when it was possible, whereas, in IRT, imputation was not realized. Therefore, these two methods were not performed on the same number of patients. In the framework of IRT, they were $N = 578$ patients available for analysis (patients who have responded to at least one item among the 5 items of the GH

dimension at the two times of measurement). However, fewer patients were available for analysis for SEM ($N = 537$) corresponding to the available sample for IRT minus the patients who had more than 2 missing items to at least one time of measurement ($N = 41$). Among those 537 patients, the data of 38 patients were imputed at least once. For imputation, we assume that the missing data are MAR, hence ignorable, but it is not possible to know whether missing data in the SatisQoL study come from an ignorable

Table 4 Uniform, non-uniform recalibration and reprioritization response shift detection using item response theory (IRT) and structural equation modeling (SEM)

		Non-uniform recalibration		Uniform recalibration		Reprioritization	
		IRT	SEM	IRT	SEM	IRT	SEM
Item 1 ^a	In general, would you say your health is	X	–	–	–	X	–
Item 2 ^b	I seem to get sick a little easier than other people	–	X	–	–	X	–
Item 3 ^b	I am as healthy as anybody I know	–	–	–	X	X	–
Item 4 ^b	I expect my health to get worse	–	X	X	X	X	–
Item 5 ^b	My health is excellent	–	–	–	–	X	X

X: response shift detected; –: no response shift detected

^a Answers: Excellent/Very good/Good/Fair/Poor

^b Answers: Definitely true/Mostly true/Don't know/Mostly false/Definitely false

(MCAR or MAR) or a non-ignorable mechanism (MNAR). Some assumptions can only be made looking at the data and the results: It seems that patients with missing data at one or two measurement occasions ($N = 170$) were more likely to have a deteriorated quality of life than patients without missing data ($N = 499$). At baseline, the average observed GH score was 59.5 ± 20.9 for patients without missing data compared to 55.2 ± 22.6 ($p = 0.04$) for patients with at least one missing item; at 6 months after discharge, it was 58.9 ± 22.1 (without missing data) versus 51.5 ± 26.2 (patients with at least one missing item, $p = 0.03$). This might suggest that the missing data could be MNAR. Moreover, if the missing data were MCAR, we could have expected that the results regarding RS detection would have been similar between SEM and IRT which was not the case, most items displaying RS being different for the two methods. However, this is very speculative, and testing the hypothesis of ignorable missing data relies on strong and unverifiable assumptions because the data that would allow testing for ignorability (MCAR or MAR versus MNAR data) are actually missing [26]. Hence, uncertainty caused by missing data is high and seriously challenges appropriate inference from the data.

We have chosen to work with a method based on IRT models because they have interesting performances in the presence of missing data. They indeed allow estimating parameters of the latent trait without making any assumptions regarding missing data. In particular, imputation is not required for these models [9, 27]. Furthermore, using Rasch family models, we benefit from the specific objectivity property. This property implies that the estimation of the latent trait is independent of the set of items used for the measurement. Consequently, unbiased estimation of the latent trait can be obtained even when some items responses are missing. Previous simulation studies have shown that Rasch family models provided unbiased estimations and were more powerful than classical test theory-based analyses in the framework of incomplete datasets

and especially in the presence of MNAR data [7, 10, 11]. We thus can hypothesize that the IRT-based algorithm may provide an interesting tool for RS analysis in case of missing data and in particular MNAR data as compared to SEM which makes the assumption that missing data are either MCAR or MAR. For SEM, simple imputation has been made following the SF-36 manual; multiple imputation could also be of value for SEM and may provide interesting results, but it is usually performed assuming MCAR or MAR data as well. Finally, among the methodological choices that were made, the use of the longitudinal GPCM for reprioritization can be discussed. Indeed, this model does not possess the specific objectivity property of the Rasch family models, and hence, it implies that, unlike the PCM (used for recalibration detection), we do not know whether we are adequately dealing with missing data regarding bias and power in the presence of reprioritization with possible MNAR data.

Conclusion

An algorithm has been developed for response shift analyses using IRT models and allows the investigation of non-uniform and uniform recalibration as well as reprioritization. This IRT-based procedure has been applied and compared to the procedure proposed by Oort's for SEM on a clinical dataset. This gives some clues regarding the technical feasibility of the proposed procedure for IRT but does not provide any formal conclusion on whether IRT and SEM differ or not regarding RS analyses with missing data or whether they both appropriately deal with missing data at the item level. In fact, the underlying "truth" is never known with certainty using a single dataset: Was RS truly present or not? Which items were really affected by RS? What was the population true change magnitude? Were the missing data MCAR, MAR or MNAR? etc. Moreover, we have hypothesized that the IRT-based

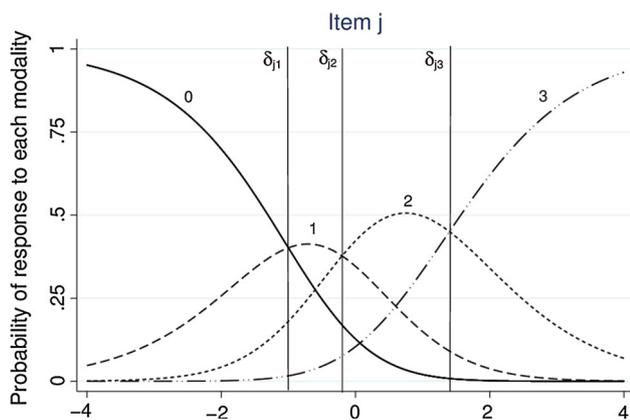


Fig. 6 Category probability curves of item j . The parameter δ_{jp} corresponds to the item difficulty for each positive category p ($p = 1, \dots, 3$) of item j

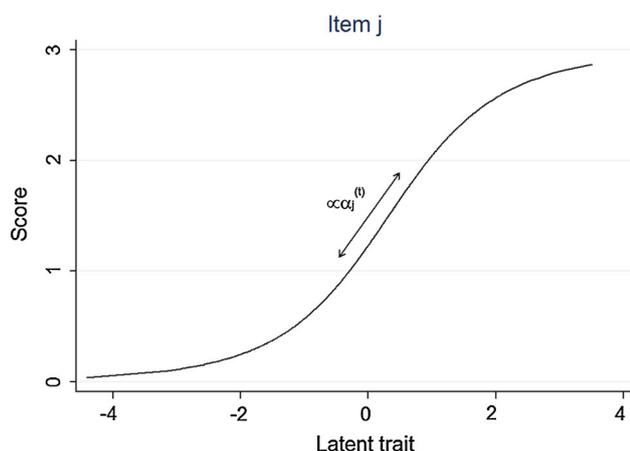


Fig. 7 Item characteristic curves for item j at time t . The parameter $\alpha_j^{(t)}$ is the discriminating power at time t ($t = 1, 2$) for item j

method for RS detection should be reliable in the presence of missing data, but its performances have to be explored more thoroughly according to the type and amount of missing data.

The next step is to implement a simulation study to be able to assess and compare the suitability of the different statistical models as well as the bias in the parameter estimates in relation to a known simulated truth. Such a study would allow validating both IRT- and SEM-based procedures and ensure that the different steps can detect the correct form of RS, on the appropriate items, and that true change estimation is unbiased. A simulation study can help in investigating the complex relationship between missing data and response shift detection by controlling for the amount of missing data and missing data mechanisms as well as the presence and type of response shift.

Acknowledgments This study was supported by the Institut National du Cancer, under reference “INCA_6931.” The SatisQoL cohort project (investigators: P. Auquier, F. Guillemin (PI), M. Mercier) was supported by an IRESP (Institut de recherche en santé publique) Grant from Inserm, and a PHRC (Programme Hospitalier de Recherche Clinique) National Grant from French Ministry of Health, France.

Appendix

The category probability curves (Fig. 6) represent the responses’ probabilities for an individual to endorse each response category for an item as a function of his latent trait level. δ_{j1} can be interpreted as the value of the latent trait for which the probability to respond negatively to item j is equal to the probability to answer positively to the first positive category (coded 1) for item j . The item characteristic curves (Fig. 7) are a representation of the expected score to an item as a function of the latent trait level. For example, for item j at time t , the slope of the curve is linked to the value of the discriminating power $\alpha_j^{(t)}$.

References

- Oort, F. J. (2005). Using structural equation modeling to detect response shifts and true change. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, 14(3), 587–598.
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. Hoboken: Wiley.
- Swartz, R. J., Schwartz, C., Basch, E., Cai, L., Fairclough, D. L., McLeod, L., et al. (2011). The king’s foot of patient-reported outcomes: Current practices and new developments for the measurement of change. *Quality of Life Research*, 20(8), 1159–1167.
- Fischer, G. H., & Molenaar, I. W. (1995). *Rasch models: Foundations, recent developments, and applications*. New York: Springer.
- Wang, W.-C., & Chyi-In, W. (2004). Gain score in item response theory as an effect size measure. *Educational and Psychological Measurement*, 64(5), 758–780.
- Andrich, D. (2011). Rating scales and rasch measurement. *Expert Review of Pharmacoeconomics & Outcomes Research*, 11, 571–585.
- De Bock, E., Hardouin, J. -B., Blanchin, M., Le Neel, T., Kubis, G., Bonnaud-Antignac, A., et al. (2013). Rasch-family models are more valuable than score-based approaches for analysing longitudinal patient-reported outcomes with missing data. *Statistical Methods in Medical Research* (in press).
- Sébille, V., Hardouin, J.-B., & Mesbah, M. (2007). Sequential analysis of latent variables using mixed-effect latent variable models: Impact of non-informative and informative missing data. *Statistics in Medicine*, 26(27), 4889–4904.
- Hardouin, J.-B., Conroy, R., & Sébille, V. (2011). Imputation by the mean score should be avoided when validating a Patient Reported Outcomes questionnaire by a Rasch model in presence of informative missing data. *BMC Medical Research Methodology*, 11(1), 105.

10. De Bock, É. de, Hardouin, J. -B., Blanchin, M., Neel, T. L., Kubis, G., & Sébille, V. (2014). Assessment of score- and Rasch-based methods for group comparison of longitudinal patient-reported outcomes with intermittent missing data (informative and non-informative). *Quality of Life Research*, 1–11.
11. Hamel J. F., Sébille V., Le Neel T., Kubis G., & Hardouin J. B. (2012) Study of different methods for comparing groups by analysis of patients reported outcomes: Item response theory based methods seem more efficient than classical test theory based methods when data is missing. Under review.
12. Schwartz, C. E., & Sprangers, M. A. (1999). Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. *Social Science & Medicine*, 48(11), 1531–1548.
13. Glas, C. A. W. (1988). The derivation of some tests for the Rasch model from the multinomial distribution. *Psychometrika*, 53, 525–546.
14. Glas, C. A. W. (2010). http://www.utwente.nl/gw/omd/Medewerkers/temp_test/mirt-manual.pdf
15. Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399–419). Thousand Oaks, CA, US: Sage Publications, Inc.
16. Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1–36.
17. R Development Core Team. (n.d.). R Development Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
18. Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research-Online*, 8, 23–74.
19. Bryant, F. B., & Satorra, A. (2012). Principles and practice of scaled difference Chi Square testing. *Structural Equation Modeling: A Multidisciplinary Journal*, 19(3), 372–398.
20. Enders, C. K. (2013). Analyzing structural equation models with missing data. In Structural Equation (Ed.), *Modeling : a second course* (pp. 493–519). Charlotte, NC: IAP, Information Age Publ.
21. Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling*, 8(3), 430–457.
22. Kepka, S., Baumann, C., Anota, A., Buron, G., Spitz, E., Auquier, P., Guillemin, F., Mercier, M. (2013). The relationship between traits optimism and anxiety and health-related quality of life in patients hospitalized for chronic diseases: data from the SATIS-QOL study. *Health and Quality of Life Outcomes*, 11(1), 134.
23. Ware, J. E., & Sherbourne, C. D. (1992). The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Medical Care*, 30(6), 473–483.
24. Leplège, A., Ecosse, E., Verdier, A., & Perneger, T. V. (1998). The French SF-36 Health Survey: translation, cultural adaptation and preliminary psychometric evaluation. *Journal of Clinical Epidemiology*, 51(11), 1013–1023.
25. Leplège, A. (2001). *Le questionnaire MOS SF-36: manuel de l'utilisateur et guide d'interprétation des scores*. Paris: Editions ESTEM.
26. Fairclough, D. L. (2002). *Design and analysis of quality of life studies in clinical trials: Interdisciplinary statistics*. London: Chapman & Hall/Crc.
27. Hamel, J.-F., Hardouin, J.-B., Le Neel, T., Kubis, G., Roquelaure, Y., & Sébille, V. (2012). Biases and power for groups comparison on subjective health measurements. *PLoS ONE*, 7(10), e44695.