Research Article

Statistics in Medicine

Received 26 October 2009, Accepted 1 November 2010 Published online 28 December 2010 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4153

Comparison of CTT and Rasch-based approaches for the analysis of longitudinal Patient Reported Outcomes

Myriam Blanchin,[a]*† Jean-Benoit Hardouin,[a] Tanguy Le Neel,[a] Gildas Kubis,[a] Claire Blanchard,[b] Eric Mirallié[b] and Véronique Sébille[a]

Health sciences frequently deal with Patient Reported Outcomes (PRO) data for the evaluation of concepts, in particular health-related quality of life, which cannot be directly measured and are often called latent variables. Two approaches are commonly used for the analysis of such data: Classical Test Theory (CTT) and Item Response Theory (IRT). Longitudinal data are often collected to analyze the evolution of an outcome over time. The most adequate strategy to analyze longitudinal latent variables, which can be either based on CTT or IRT models, remains to be identified. This strategy must take into account the latent characteristic of what PROs are intended to measure as well as the specificity of longitudinal designs. A simple and widely used IRT model is the Rasch model. The purpose of our study was to compare CTT and Rasch-based approaches to analyze longitudinal PRO data regarding type I error, power, and time effect estimation bias. Four methods were compared: the Score and Mixed models (SM) method based on the CTT approach, the Rasch and Mixed models (RM), the Plausible Values (PV), and the Longitudinal Rasch model (LRM) methods all based on the Rasch model. All methods have shown comparable results in terms of type I error, all close to 5 per cent. LRM and SM methods presented comparable power and unbiased time effect estimations, whereas RM and PV methods showed low power and biased time effect estimations. This suggests that RM and PV methods should be avoided to analyze longitudinal latent variables. Copyright © 2010 John Wiley & Sons, Ltd.

Keywords: Item Response Theory; Classical Test Theory; Patient Reported Outcomes; longitudinal data; simulation study

1. Introduction

Patient Reported Outcomes (PRO) data are widely used in health sciences to evaluate concepts, such as health-related quality of life (HRQoL), pain, fatigue, or anxiety [1], which are often referred to as latent variables because they cannot be directly observed from patients. PRO data are evaluated using the answers of patients to items often grouped into several dimensions in a questionnaire. Two approaches are commonly used for the analysis of such data: Classical Test Theory (CTT) and Item Response Theory (IRT). The CTT is an approach based on the computation of a score usually computed as the sum of the item responses. This score is an estimation of a 'true' score assumed to represent the evaluated outcome (e.g. HRQoL). The observed and true scores are assumed to be linked by a linear relation. In IRT, item responses have a central role. The probability to answer to an item is a function (not necessary linear) of the latent variable which represents the evaluated outcome. IRT models are a large family

[a] *EA 4275 'Biostatistics, Clinical Research and Subjective Measures in Health Sciences', Faculty of Pharmaceutical Sciences, University of Nantes, Nantes, France*

[b] *Department of Digestive and Endocrine Surgery/Institut des Maladies de l'Appareil Digestif, CHU Nantes, Faculty of Medicine, University of Nantes, France*

**Correspondence to: Myriam Blanchin, EA 4275 'Biostatistics, Clinical Research and Subjective Measures in Health Sciences', Faculté de Pharmacie—Université de Nantes, 1, rue Gaston Veil—44035 Nantes Cedex 1, France.*

† *E-mail: myriam.blanchin@univ-nantes.fr*

of models relying generally on the same three assumptions: unidimensionality, monotonicity, and local independence. A simple and widely used IRT model is the Rasch model [2] or one-parameter logistic model (1-PLM). The Rasch model is adapted to the analysis of dichotomous items and models the probability of a response to an item through a person parameter (person ability) and an item parameter (item difficulty). Nowadays, a large proportion of scales are developed and validated using models of the Rasch family due to their interesting psychometrics properties including the exhaustivity of the score on the latent trait and the specific objectivity.

Longitudinal data are often collected in order to analyze the evolution of an outcome over time. In this case, the correlation between measurements from each patient over time has to be taken into account. Linear mixed models [3] are commonly used to analyze such data.

In the case of longitudinal PRO data, the latent characteristic of what PRO are intended to measure as well as the specificity of longitudinal designs with repeated measurements should probably both be taken into account to provide reliable analysis. To date, the choice of a statistical strategy for the analysis of such data is usually based on CTT rather than on IRT and seems to more likely rely on the researcher's practice and familiarity with CTT than on scientific grounds. Hence, the most adequate strategy to analyze longitudinal latent variables, which can be either based on the CTT or IRT approach, remains to be identified. The purpose of our study was to compare a CTT- and three Rasch-based approaches to analyze longitudinal PRO data regarding type I error, power and time effect estimation bias. Data from the evaluation of quality of life of patients with primary hyperparathyroidism were used to illustrate simulation results.

2. Methods

2.1. Statistical models

2.1.1. The Rasch model. The Rasch model [2, 4] came from psychometrics and was developed for achievement tests. Later it came to be used in health sciences, in particular for construction, validation, and reduction of questionnaires [5, 6]. In this framework, the responses to the items of a questionnaire are assumed to be the manifestation of a latent variable which cannot be directly observed. The Rasch model proposes to model the relationship between responses to dichotomous items and the latent variable, denoted θ. Let Y_{ij} be the dichotomous variable representing the response of person i $(i=1\dots N)$ to an item j $(j=1\dots J)$.
For a questionnaire containing J dichotomous items, the model can be written as follows:

$$P(Y_{ij}=y|\theta_i;\delta_j)=\frac{\exp(y(\theta_i-\delta_j))}{1+\exp(\theta_i-\delta_j)} \tag{1}$$

where $y=0$ for a negative response (the most pejorative response) and $y=1$ for a positive response. δ_j is called the difficulty or item parameter and is associated with item j . The personal parameter θ_i is the individual value of the latent trait for patient i and represents the ability of the patient (e.g. HRQoL).

The Rasch model relies on three hypotheses:

- *Unidimensionality*: A unique latent variable explains the response to the items.
- *Monotonicity*: The probability of a positive response to an item is a non-decreasing function of the latent variable.
- *Local independence*: Given an individual, the item responses are independent of one another.

A Rasch model, where the latent trait θ is considered as a random variable and usually has a normal distribution $\mathrm{N}(\mu,\sigma^2)$, is a mixed-effects logistic model [7]. The parameters to be estimated in the model are μ and σ^2 to characterize the distribution of the latent trait θ and the difficulty parameters δ_j $(j=1\dots J)$. They can be jointly estimated using marginal maximum likelihood (MML). The marginal likelihood is expressed as

$$L(\delta_1,\dots,\delta_J,\mu,\sigma^2|\mathbf{y})=\prod_{i=1}^{N}\int\prod_{j=1}^{J}\frac{\exp(y_{ij}(\theta-\delta_j))}{1+\exp(\theta-\delta_j)}G(\theta/\mu,\sigma^2)\,\mathrm{d}\theta \tag{2}$$

with $G(./\mu,\sigma^2)$ the normal distribution function with mean μ and variance σ^2.

In order to ensure the identifiability of the model, one constraint has to be adopted. In general, the mean of the latent trait or the sum of difficulty parameters is assumed to be equal to 0.

2.1.2. Longitudinal mixed Rasch model. For repeated measures data, a longitudinal form of the Rasch model has been developed [8] in the field of educational and psychological testing. The interest of researchers using learning tests was to measure learning ability as well as its evolution. A longitudinal mixed Rasch model for modeling of quality of life evolution was derived from the modeling of learning and change, considering the latent variable θ as a random variable rather than as fixed parameters.

For a questionnaire containing J dichotomous items and measures repeated T times for each person i, the probability of a response to an item j at time t can be written as follows:

$$P(Y_{ij}^{(t)}=y^{(t)}|\theta_i^{(t)};\delta_j)=\frac{\exp(y^{(t)}(\theta_i^{(t)}-\delta_j))}{1+\exp(\theta_i^{(t)}-\delta_j)} \tag{3}$$

Item parameters $\delta_j, j=1\dots J$ are assumed to be constant with time meaning that the characteristics of the questionnaire are assumed not to vary through time.

The marginal likelihood is expressed as

$$L(\delta_1,\dots,\delta_J,\boldsymbol{\mu},\boldsymbol{\Sigma}|\mathbf{y})=\prod_{i=1}^{N}\int_{\mathbb{R}^T}\prod_{t=1}^{T}\prod_{j=1}^{J}\frac{\exp(y_{ij}^{(t)}(\theta^{(t)}-\delta_j))}{1+\exp(\theta^{(t)}-\delta_j)}G(\boldsymbol{\theta}/\boldsymbol{\mu},\boldsymbol{\Sigma})d\boldsymbol{\theta} \tag{4}$$

with $G(./\boldsymbol{\mu},\boldsymbol{\Sigma})$ the multivariate normal distribution function with mean vector $\boldsymbol{\mu}=(\mu_1\dots\mu_T)'$ and covariance matrix $\boldsymbol{\Sigma}$.

A constraint is needed to ensure identifiability of the longitudinal mixed Rasch model. The usual constraint made on the parameters is $\mu_1=0$.

2.1.3. Covariance pattern models. A simple way to analyze repeated measures data where the focus is on the mean responses and their evolution with time is to use a covariance pattern model [9, 10]. In linear mixed models, random effects are used to model individual variation around the mean trajectory when the primary interest is in individual trajectories. Covariance pattern models contain only fixed effects that characterize the mean behavior of the population which is our primary interest. Moreover, this type of mixed model allows to specify a pattern for the correlation between measurements from the same patient.
Let

- n_i be the number of observations on patient i, $i=1\dots N$
- p be the number of parameters
- $\boldsymbol{Y}_i$ be the $(n_i\times 1)$ vector containing the responses for the patient i
- $\boldsymbol{\beta}$ be the $(p\times 1)$ vector of fixed effects parameters
- $\boldsymbol{X}_i$ be the $(n_i\times p)$ design matrix
- $\boldsymbol{e}_i$ be the $(n_i\times 1)$ vector of error terms, characterizing the overall variation and measurement error
- $\boldsymbol{\Sigma}_i$ be the $(n_i\times n_i)$ covariance matrix of error terms
- $M=\sum_{i=1}^{N}n_i$ be the total number of observations.

A covariance pattern model can be written as follows:

$$\begin{aligned}\boldsymbol{Y}_i&=\boldsymbol{X}_i\boldsymbol{\beta}+\boldsymbol{e}_i\\ \mathrm{var}(\boldsymbol{e}_i)&=\boldsymbol{\Sigma}_i\\ \boldsymbol{Y}_i&\sim N_{n_i}(\boldsymbol{X}_i\boldsymbol{\beta},\boldsymbol{\Sigma}_i)\end{aligned} \tag{5}$$

The parameters to be estimated in the model are $\boldsymbol{\beta}$ that characterizes the mean and $\boldsymbol{\omega}$ that characterizes $\boldsymbol{\Sigma}_i$. For example, $\boldsymbol{\omega}=(\sigma^2,\rho)$ for an AR(1) structure of $\boldsymbol{\Sigma}_i$. Two methods based on likelihood maximization are used to estimate unknown parameters: Maximum Likelihood (ML) or REstricted Maximum Likelihood (REML) estimation methods. The use of ML estimation leads to unbiased estimate for $\boldsymbol{\beta}$ but $\boldsymbol{\omega}$ is known to be biased when N is not too large. In REML method, the likelihood is modified to include an extra term for correction of the bias on $\boldsymbol{\omega}$ parameter.

Let $\boldsymbol{Y}$ be the $(M\times 1)$ vector summarizing the vectors $\boldsymbol{Y}_i (i=1,\ldots,N)$ into one vector. The joint density of $\boldsymbol{Y}$ is expressed as

$$f(\boldsymbol{y})=\prod_{i=1}^{N}(2\pi)^{-n_i/2}|\boldsymbol{\Sigma}_i|^{-1/2}|\boldsymbol{X}_i'\boldsymbol{\Sigma}_i^{-1}\boldsymbol{X}_i|^{-1/2}\exp\{-(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta})'\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta})/2\} \tag{6}$$

The estimator of $\boldsymbol{\omega}$ resulting from REML estimation is known to be less biased than the estimation based on ML [3].

2.2. Comparison of methods

Four methods to analyze longitudinal PRO data have been compared: (i) Score and Mixed models (SM), (ii) Rasch and Mixed models (RM), (iii) Plausible Values (PV), and (iv) Longitudinal Rasch model (LRM).

Score and Mixed models. The SM method, corresponding to the CTT approach, consisted in calculating a score by summing the item responses for each patient. A linear mixed model was then used to explain the evolution of score with time.

Four covariance structures $\boldsymbol{\Sigma}$ are often used with longitudinal data: UN (unstructured), ARH(1) (heterogeneous first-order autoregressive), AR(1) (first-order autoregressive), or CSH (heterogeneous compound symmetry). The unstructured matrix is the most general possible structure. It is used when no hypotheses can be made on the structure of the covariance matrix but lead to estimate an important number of parameters. The ARH(1) structure takes into account the correlation between measures in time. Correlations are assumed to decrease when measures get further apart from each other in time. The use of the AR(1) structure makes an extra hypothesis compared to the ARH(1) structure: the variances are assumed to be equal. On the contrary, the choice of a CSH structure assumes that the variances are not equal but the correlation is constant over time.

The simulated datasets contained only one group and balanced data measured on three different occasions. In the presence of a single group, $n_i=3\ \forall i$ and $\boldsymbol{\Sigma}_i=\boldsymbol{\Sigma}$ could be assumed to be the same for all patients.

The model can be written as

$$\begin{aligned}\boldsymbol{S}_i &= \boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{e}_i\\ \mathrm{var}(\boldsymbol{e}_i) &= \boldsymbol{\Sigma}\\ \boldsymbol{S}_i &\sim N_{n_i}(\boldsymbol{X}\boldsymbol{\beta},\boldsymbol{\Sigma})\end{aligned} \tag{7}$$

where $S_i^{(t)}=\sum_j y_{ij}^{(t)}$ for $t=(1,2,3)$ and

$$\boldsymbol{\Sigma}=\begin{pmatrix}\sigma_1^2 & \sigma_{12} & \sigma_{13}\\ \sigma_{12} & \sigma_2^2 & \sigma_{23}\\ \sigma_{13} & \sigma_{23} & \sigma_3^2\end{pmatrix}\ \text{for UN}\quad \boldsymbol{\Sigma}=\begin{pmatrix}\sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho^2\\ \sigma_1\sigma_2\rho & \sigma_2^2 & \sigma_2\sigma_3\rho\\ \sigma_1\sigma_3\rho^2 & \sigma_2\sigma_3\rho & \sigma_3^2\end{pmatrix}\ \text{for ARH(1)}$$

$$\boldsymbol{\Sigma}=\sigma^2\begin{pmatrix}1 & \rho & \rho^2\\ \rho & 1 & \rho\\ \rho^2 & \rho & 1\end{pmatrix}\ \text{for AR(1)}\quad \boldsymbol{\Sigma}=\begin{pmatrix}\sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho\\ \sigma_1\sigma_2\rho & \sigma_2^2 & \sigma_2\sigma_3\rho\\ \sigma_1\sigma_3\rho & \sigma_2\sigma_3\rho & \sigma_3^2\end{pmatrix}\ \text{for CSH}$$

Mean parameters $\boldsymbol{\beta}$ and covariance parameters $\boldsymbol{\omega}$ were estimated using the REML method in order to reduce the bias on covariance parameters. As would be performed on real data, AIC for each covariance matrix structure were compared to choose the adequate structure.

An estimate of $\boldsymbol{\mu}$ could be given by $\hat{\boldsymbol{\mu}}=(\hat{\mu}_1\quad \hat{\mu}_2\quad \hat{\mu}_3)'=\boldsymbol{X}\hat{\boldsymbol{\beta}}$. The time effect between two consecutive measures $(t=(1,2))$ was $d_{t,t+1}=\mu_{t+1}-\mu_t$. The time effect between time 1 and time 2 was estimated as $\hat{d}_{12}=\hat{\mu}_2-\hat{\mu}_1$.

The test of a time effect used an approximate F-test:

$$H_0:\mu_1=\mu_2=\mu_3=\mu\Leftrightarrow\beta_1=\beta_2=\beta_3\Leftrightarrow L\boldsymbol{\beta}=0$$

$$H_1:\exists i|\mu_i\neq\mu\Leftrightarrow L\boldsymbol{\beta}\neq 0$$

Define

$$\boldsymbol{L}=\begin{pmatrix}1 & -1 & 0\\ 1 & 0 & -1\end{pmatrix}$$

Under H_0, $F_L=(\boldsymbol{L}\hat{\boldsymbol{\beta}})'(\boldsymbol{L}\hat{\boldsymbol{V}}_\beta\boldsymbol{L}')^{-1}\boldsymbol{L}\hat{\boldsymbol{\beta}}/\text{rank}(\boldsymbol{L})$ has approximately an $F_{r,\text{df}}$ distribution where the numerator degrees of freedom, r, equals the rank of $\boldsymbol{L}$, df is the appropriate denominator degrees of freedom and $\hat{\boldsymbol{V}}_\beta=(\sum_{i=1}^{N}\boldsymbol{X}'\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X})^{-1}$ is the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$.

Rasch and mixed models and plausible values. The Rasch and Mixed models (RM) and the Plausible Values (PV) methods were performed in two steps. In the first step, a mixed Rasch model was applied on the overall sample and an individual value of the latent trait was estimated for each patient at each time point. A linear mixed model was then fitted to explain the evolution of the estimated latent traits with time. A covariance pattern model was used as it was for the score.

First step of RM and PV methods. The parameters μ and σ^2, which characterize the distribution of θ, and δ_j, $j=1\dots J$ were estimated using MML estimation from a mixed Rasch model [4]. Item parameters δ_j were constant with time.

RM and PV differ in the method used to estimate individual values of the latent trait. The *Bayes Expected A Posteriori* estimator (EAP), used in the first step of the RM method, is a point estimate defined as the mean of the posterior distribution [11]:

$$\text{EAP}(\theta_i)=E(\theta|\boldsymbol{Y},\boldsymbol{\delta},\mu,\sigma^2)=\frac{\int\theta B(\theta)\exp(\theta s_i)G(\theta|\mu,\sigma^2)\,\text{d}\theta}{\int B(\theta)\exp(\theta s_i)G(\theta|\mu,\sigma^2)\,\text{d}\theta}$$

where $\boldsymbol{\delta}$ was the vector of difficulty parameters, s_i the observed raw score for patient i ($s_i=\sum_j y_{ij}$), $B(\theta)=\prod_j[1+\exp(\theta-\delta_j)]^{-1}$, and $G(./\mu,\sigma^2)$ the normal distribution function with mean μ and variance σ^2. Owing to the use of EAP estimates in RM method, all patients with the same total score and hence the same posterior distribution will have the same estimated value of the latent trait $\hat{\theta}$.

The PV method consisted in first estimating a value for the latent variable by plausible value imputation. The plausible value imputation is based on the multiple imputation theory of Rubin [12] and is used in large-scale educational surveys, such as PISA and NAEP [13, 14]. Instead of using the mean of the posterior distribution, a plausible value ($\hat{\theta}$) is randomly drawn from the posterior distribution. This method allows patients with the same total score (and hence the same posterior distribution) to have different plausible values. Usually, several draws of plausible values are used. The same analysis is made on each draw and results of all the analyses are pooled to obtain an estimate of the parameter of interest and an estimate of its variance as in multiple imputation. The multiple draws allow to obtain an estimate of the uncertainty due to the estimate of θ. If this uncertainty has to be taken into account in the analysis but has not to be explicitly estimated, one draw of plausible values is sufficient [15]. Furthermore, Wu [16] has shown that one plausible value could be sufficient to adequately recover population parameters. In health sciences, studies are focused on the evolution of the population and not on the individual trajectories.

Given the item response pattern $\boldsymbol{y}$ and the latent variable θ, $f(\boldsymbol{y}|\theta)$ is the item response probability of the Rasch model also called item response model. Assuming that θ comes from a normal distribution, $g(\theta)\sim\text{N}(\mu,\sigma^2)$ is called the population model. The posterior distribution $h(\theta|\boldsymbol{y})$ of an individual with item response pattern $\boldsymbol{y}$ is defined as

$$h(\theta|\boldsymbol{y})=\frac{f(\boldsymbol{y}|\theta)g(\theta)}{\int f(\boldsymbol{y}|\theta)g(\theta)\,\text{d}\theta} \tag{8}$$

Plausible values are randomly drawn from the posterior distribution with density $h(\theta|\boldsymbol{y})$. As the EAP estimator is defined as the mean of the posterior distribution, we use the EAP estimates and its standard errors to draw the plausible values $\hat{\theta}_i^{(t)}$ for each person i at each time point t from a normal distribution with mean equal to the EAP estimate of θ of the person i and standard error equal to the corresponding estimated standard error.

Second step of RM and PV methods. For both methods, the linear mixed model was expressed as

$$\begin{aligned}\hat{\theta}_i &= X\beta + e_i \\ \mathrm{var}(\boldsymbol{e}_i) &= \boldsymbol{\Sigma} \\ \hat{\theta}_i &\sim N_{n_i}(\boldsymbol{X\beta}, \boldsymbol{\Sigma})\end{aligned} \tag{9}$$

Mean parameters $\boldsymbol{\beta}$ and covariance parameters $\boldsymbol{\omega}$ were estimated using the REML method. As for the SM method, four structures were investigated for $\boldsymbol{\Sigma}$: UN, ARH(1), AR(1), and CSH.

An estimate of $\boldsymbol{\mu}$ could be given by $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1 \ \ \hat{\mu}_2 \ \ \hat{\mu}_3)' = \boldsymbol{X}\hat{\boldsymbol{\beta}}$. The time effect between time 1 and time 2 was estimated by $\hat{d}_{12} = \hat{\mu}_2 - \hat{\mu}_1$.

The test of a time effect used an approximate F-test such as for the SM method.

Longitudinal mixed Rasch model. The method LRM was based on a longitudinal Rasch model that estimated the time effect, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ of the latent trait in the same step.

The longitudinal mixed Rasch model, used to estimate $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and δ_j, $j = 1 \dots J$, was expressed as

$$P(Y^{(t)} = y^{(t)} | \theta^{(t)}; \delta_j) = \frac{\exp(y^{(t)}(\theta^{(t)} - \delta_j))}{1 + \exp(\theta^{(t)} - \delta_j)} \tag{10}$$

The estimation was based on MML where the marginal likelihood was expressed as

$$L(\delta_1, \dots, \delta_J, \boldsymbol{\mu}, \boldsymbol{\Sigma} | \mathbf{y}) = \prod_{i=1}^{N} \int_{\mathbb{R}^T} \prod_{t=1}^{T} \prod_{j=1}^{J} \frac{\exp(y_{ij}^{(t)}(\theta^{(t)} - \delta_j))}{1 + \exp(\theta^{(t)} - \delta_j)} G(\boldsymbol{\theta} / \boldsymbol{\mu}, \boldsymbol{\Sigma}) \mathrm{d}\boldsymbol{\theta} \tag{11}$$

with $G(./\boldsymbol{\mu}, \boldsymbol{\Sigma})$ the multivariate normal distribution function with mean vector $\boldsymbol{\mu} = (\mu_1 \cdots \mu_T)'$ and covariance matrix $\boldsymbol{\Sigma}$ of unstructured type.

Constraint of the nullity of latent variable mean at time 1 was used to ensure identifiability. Owing to this constraint, $\hat{\mu}_2$ and $\hat{\mu}_3$ represented, respectively, $\hat{d}_{12}$ and $\hat{d}_{13}$.

The test of a time effect used an approximate Wald test:

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = \mu_3 = \mu \Leftrightarrow L\boldsymbol{\mu} = 0 \\ H_1 &: \exists i | \mu_i \neq \mu \Leftrightarrow L\boldsymbol{\mu} \neq 0 \end{aligned}$$

Define

$$\boldsymbol{L} = \begin{pmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}$$

Under H_0, $T_L = (\boldsymbol{L}\hat{\boldsymbol{\mu}})'(\boldsymbol{L}\hat{\boldsymbol{V}}\boldsymbol{L}')^{-1}\boldsymbol{L}\hat{\boldsymbol{\mu}}$ has approximately a χ_r^2 distribution, where $r =$ rank of $\boldsymbol{L}$ and $\hat{\boldsymbol{V}}$ is the estimated covariance matrix.

2.3. Simulation of data

Responses of patients to dichotomous items in a repeated measures setting were simulated with a longitudinal mixed Rasch model including three times of assessment according to equation (3). The latent trait vector $\theta = (\theta^{(1)} \ \ \theta^{(2)} \ \ \theta^{(3)})'$ had a multivariate normal distribution with

$$\boldsymbol{\mu} = (\mu_1 \ \ \mu_2 \ \ \mu_3)' \quad \text{and} \quad \boldsymbol{\Sigma} = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{pmatrix}$$

corresponding to a first-order autoregressive structure for the covariance matrix. The correlation between measures decreased as measures got farther apart from each other in time. The latent trait was considered as having the same variance at each time point ($\sigma^2 = 1$). Three different values for the correlation coefficient of the latent trait between two consecutive times ρ were used to simulate data: $\rho = 0.4$ (small correlation), $\rho = 0.7$, and $\rho = 0.9$ (high correlation).

The time effect between two consecutive measures ($t=(1,2)$) was $d_{t,t+1}=\mu_{t+1}-\mu_t$ and $ES_{t,t+1}$ was the effect size between two consecutive measures. $ES_{t,t+1}=d_{t,t+1}/\sigma\ \forall t=(1,2)$. Under H_0, $ES_{t,t+1}=0$. Under H_1, $ES_{t,t+1}=0.2$.

The data were assumed to come from a 4-item scale or a 7-item scale with dichotomous items. The values of difficulty parameters were $\delta_1=-1$, $\delta_2=-0.5$, $\delta_3=0.5$, $\delta_4=1$ for a 4-item scale and $\delta_1=-1.5$, $\delta_2=-1$, $\delta_3=-0.5$, $\delta_4=0$, $\delta_5=0.5$, $\delta_6=1$, $\delta_7=1.5$ for a 7-item scale. The sample size could be of 100 or 200 individuals. The different values of the number of items, the number of individuals, and the correlation led to consider 24 different cases. Five hundred simulated datasets were generated and analyzed for each case.

Studied criteria. In order to compare the methods to analyze longitudinal PRO data, three criteria were studied: the type I error, the power and the bias of the time effect estimation.

The type I error of the tests was classically computed as the proportion of rejection of H_0 under the null hypothesis. Rejection of H_0 was based on a test of simultaneous equality of mean estimations, i.e. the absence of time effect.

The power calculation used the same tests but calculated the proportion of rejection of H_0 under the alternative hypothesis.

RM, PV, and LRM methods were based on IRT models, hence the calculation of time effect estimation bias was possible. The estimated value of time effect $\hat{d}_{t,t+1}$ was compared with the fixed value $d_{t,t+1}$ used for data simulation. As SM method was based on the classical approach (CTT) and IRT was used to simulate data, the true value of the time effect on the score scale was not known. But since under H_0, no time effect was assumed on the latent variable, no time effect was expected on the score scale under H_0 as well and hence the time effect bias could be calculated under H_0. Under H_1, the true value of time effect was not known on the score scale hence the time effect bias could not be assessed.

The mean of time effect estimation for each case was compared to the true value using a t-test.

Simulations and analyses were performed using SAS 9.1 [17] and Stata 10 [18, 19].

3. Simulation results

For each of the 24 cases, the AIC for the four structures of covariance matrix for mixed models were compared for RM and SM methods.

For the SM, when $\rho=0.4$ or $\rho=0.7$, the AIC was more often minimized by the choice of an AR(1) structure for the covariance matrix. When $\rho=0.9$, the CSH structure more often minimized the AIC of the models than the other structures. The results further presented are from covariance pattern models estimated through the REML method with an AR(1) structure for the covariance matrix.

For the RM method, the AIC was also more often minimized by the choice of an AR(1) structure for the covariance matrix in most of the cases under H_0. The results presented further for the RM method come from analyses with an AR(1) structure.

The LRM method used an unstructured covariance matrix.

3.1. *Type I error of the tests*

Table I and Figure 1 show the type I error for each method for different values of the parameters: sample size, number of items, and latent variable correlation. All type I errors were close to 5 per cent. All methods showed comparable results whatever the value of the parameters. No ρ, J, or N effects were observed on the type I error values.

All 95 per cent confidence intervals included the target value of 5 per cent. Type I error ranged from 3.6 to 6.0 per cent for the SM method, from 3.4 to 5.8 per cent for the RM method, from 3.8 to 6.2 per cent for the PV method, and from 4.4 to 6.7 per cent for the LRM method.

3.2. *Power of the tests*

Table I and Figure 2 show the power for each method for different values of the parameters: sample size, number of items, and latent variable correlation. Whatever the value of these parameters, the powers of the LRM method were comparable to those of the SM method. We denoted that the LRM method presented a power a little higher than the SM method whatever the value of correlation coefficient ρ. Moreover, LRM and SM methods achieved much larger power than the RM and PV

Table I. Type I error and power of the tests for Score Mixed model (SM), Rasch Mixed model (RM), Plausible Values (PV), and Longitudinal Rasch Mixed model (LRM) methods for different values of sample size (N), number of items (J), and latent variable correlation (ρ). Results from analyses with an AR(1) structure for the covariance matrix in RM, PV, and SM methods and an unstructured covariance matrix for LRM method.

			SM		RM		PV		LRM	
N	J	ρ	Type I error	Power	Type I error	Power	Type I error	Power	Type I error	Power
100	4	0.4	0.054	0.384	0.052	0.146	0.062	0.166	0.054	0.388
		0.7	0.040	0.372	0.040	0.120	0.042	0.142	0.044	0.423
		0.9	0.038	0.424	0.042	0.156	0.056	0.156	0.048	0.508
	7	0.4	0.036	0.428	0.034	0.108	0.052	0.162	0.044	0.470
		0.7	0.046	0.522	0.048	0.096	0.050	0.150	0.058	0.568
		0.9	0.044	0.564	0.044	0.070	0.054	0.160	0.046	0.688
200	4	0.4	0.060	0.634	0.058	0.280	0.058	0.288	0.052	0.654
		0.7	0.048	0.678	0.048	0.276	0.060	0.304	0.056	0.721
		0.9	0.060	0.756	0.060	0.250	0.052	0.330	0.067	0.826
	7	0.4	0.050	0.810	0.048	0.184	0.042	0.304	0.050	0.822
		0.7	0.036	0.880	0.038	0.154	0.040	0.314	0.048	0.916
		0.9	0.048	0.918	0.048	0.148	0.038	0.306	0.058	0.955

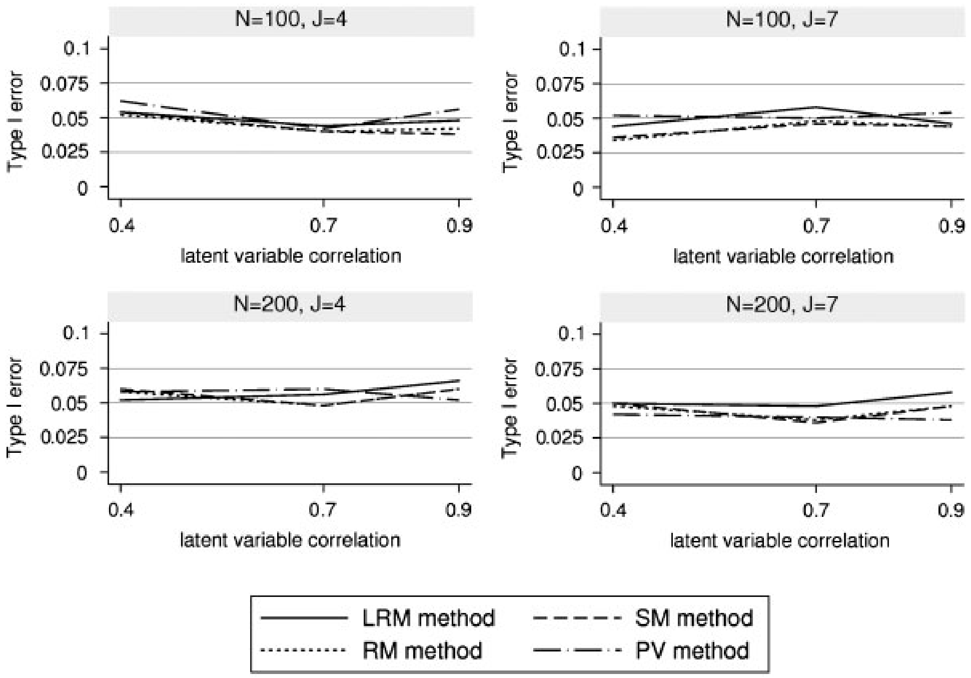

Figure 1. Type I error of the tests for Score Mixed model (SM), Rasch Mixed model (RM), Plausible Values (PV), and Longitudinal Rasch Mixed model (LRM) methods for different values of sample size (N), number of items (J), and latent variable correlation (ρ). Results from analyses with an AR(1) structure for the covariance matrix in RM, PV, and SM methods and an unstructured covariance matrix for LRM method.

methods. Differences in power were the highest when $N = 200$ and $J = 7$. In this case, LRM and SM powers were all higher than 80 per cent, whereas RM power ranged from 15 to 18 per cent and PV power were close to 30 per cent.

For LRM and SM methods, the power increased with the rise of the correlation between two measures, the number of items or the number of individuals. For example, power of the SM and LRM methods were close to 38 per cent when $N = 100$, $J = 4$, and $\rho = 0.4$. Power was much higher, greater than 90 per cent, when $N = 200$, $J = 7$, and $\rho = 0.9$.

Power from the RM and PV methods was quite stable whatever the values of the parameters. We could note a slight effect of J on the RM method: when the number of items increased, the power decreased

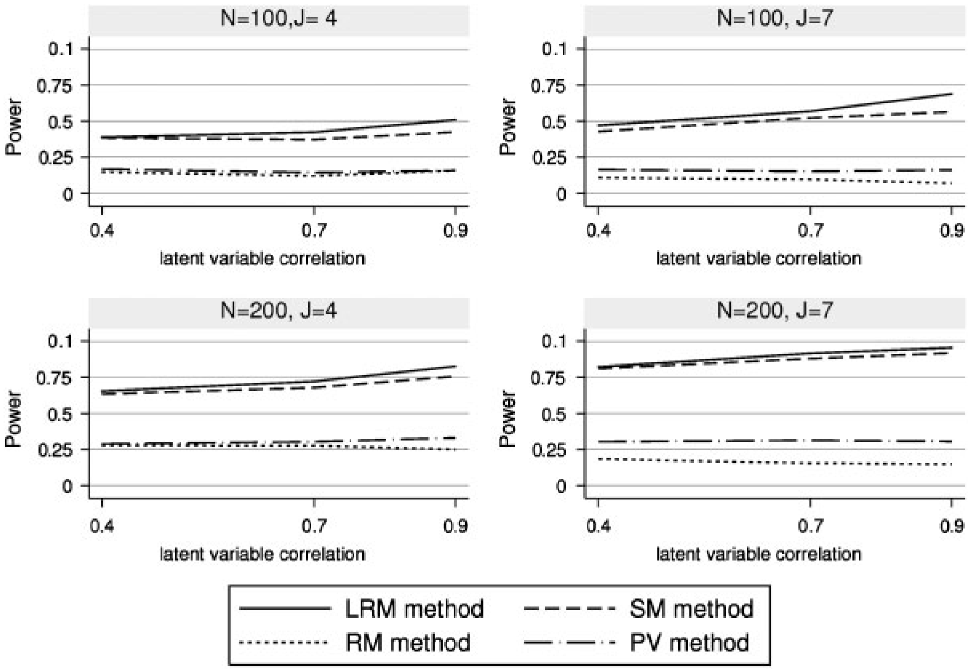

Figure 2. Power of the tests for Score and Mixed models (SM), Rasch and Mixed models (RM), Plausible Values (PV), and Longitudinal Rasch Mixed model (LRM) methods for different values of sample size (N), number of items (J), and latent variable correlation (ρ). Results from analyses with an AR(1) structure for the covariance matrix in RM, PV, and SM methods and an unstructured covariance matrix for LRM method.

by contrast with the LRM and SM methods. A N effect was also shown: the power increased with the sample size.

3.3. Bias on the time effect estimation

Table II shows the time effect estimation between the first and second time of measurement for each method for different values of the parameters: sample size, number of items, and latent variable correlation. For all methods, all cases presented unbiased estimation of time effect between time 1 and time 2 under H_0.

For RM and PV methods, the time effect estimation under H_1 was always biased. For both methods, d_{12} was underestimated in all cases. Table II showed that the estimated time effect was 10 to 20 times less than the true value for RM and 2 to 3 times less than the true value for PV. A J effect could be observed for RM. As the number of items increased, the estimation of time effect decreased and was more biased.

On the contrary, estimations from the LRM method were always unbiased. Remember that the true value of time effect on score was not known. Thus, the bias on time effect estimation could not be assessed for the SM method. The results on time effect estimation between time 2 and time 3 are comparable to the results on time effect estimation between time 1 and time 2 (results not shown).

4. Illustrative example

An analysis was performed on a longitudinal study aimed at evaluating the evolution of nonspecific symptoms and quality of life in primary hyperparathyroidism before and after surgery [20]. The study was multicentric and took place in six academic departments of Endocrine Surgery in France. Patients with primary hyperparathyroidism scheduled for parathyroidectomy were asked to fill out a questionnaire about nonspecific symptoms and a quality of life questionnaire (SF-36). Patients were evaluated during the preoperative period and at 3 and 6 months after surgery.

The SF-36 is a 36-item generic scale made of 8 dimensions: physical functioning, social functioning, bodily pain, general health perceptions, vitality, role limitations due to emotional problems

Table II. Time effect estimation between time 2 and time 1 ($\hat{d}_{12}$) under H_0 and H_1 for Score Mixed model (SM), Rasch Mixed model (RM), Plausible Values (PV), and Longitudinal Rasch Mixed model (LRM) methods for different values of sample size (N), number of items (J), and latent variable correlation (ρ). Results from analyses with an AR(1) structure for the covariance matrix in SM, PV, and RM methods and an unstructured covariance matrix for LRM method. Mean values of $\hat{d}_{12}$ and standard errors (s.e.).

					SM		RM		PV		LRM	
	N	J	ρ	d_{12}	$\hat{d}_{12}$	s.e.	$\hat{d}_{12}$	s.e.	$\hat{d}_{12}$	s.e.	$\hat{d}_{12}$	s.e.
H_0	100	4	0.4	0	0.010	0.007	0.006	0.004	0.008	0.006	0.013	0.009
			0.7	0	−0.003	0.006	−0.002	0.004	0.001	0.006	−0.004	0.008
			0.9	0	−0.007	0.006	−0.005	0.003	−0.003	0.006	−0.010	0.008
		7	0.4	0	0.005	0.009	0.002	0.004	0.000	0.006	0.005	0.007
			0.7	0	0.008	0.009	0.004	0.004	0.008	0.006	0.007	0.007
			0.9	0	0.011	0.007	0.005	0.003	0.005	0.006	0.009	0.006
H_0	200	4	0.4	0	−0.003	0.005	−0.001	0.003	0.001	0.004	−0.003	0.006
			0.7	0	0.007	0.005	0.004	0.003	0.007	0.004	0.009	0.006
			0.9	0	−0.007	0.004	−0.004	0.002	−0.004	0.004	−0.009	0.006
		7	0.4	0	−0.003	0.007	−0.001	0.003	−0.002	0.004	−0.002	0.005
			0.7	0	0.002	0.006	0.001	0.003	0.003	0.004	0.003	0.005
			0.9	0	−0.005	0.005	−0.002	0.002	−0.003	0.004	−0.004	0.004
H_1	100	4	0.4	0.2	0.139	0.007	0.023*	0.002	0.082*	0.006	0.184	0.009
			0.7	0.2	0.138	0.006	0.022*	0.002	0.073*	0.006	0.186	0.008
			0.9	0.2	0.156	0.006	0.022*	0.002	0.084*	0.006	0.211	0.008
		7	0.4	0.2	0.237	0.009	0.009*	0.001	0.080*	0.006	0.189	0.007
			0.7	0.2	0.246	0.008	0.010*	0.001	0.091*	0.006	0.197	0.007
			0.9	0.2	0.250	0.007	0.007*	0.001	0.084*	0.006	0.202	0.006
H_1	200	4	0.4	0.2	0.148	0.005	0.023*	0.001	0.093*	0.004	0.197	0.006
			0.7	0.2	0.155	0.004	0.023*	0.001	0.086*	0.004	0.207	0.005
			0.9	0.2	0.150	0.004	0.022*	0.001	0.087*	0.004	0.202	0.006
		7	0.4	0.2	0.255	0.007	0.011*	0.001	0.089*	0.004	0.203	0.005
			0.7	0.2	0.243	0.006	0.008*	0.001	0.082*	0.004	0.194	0.005
			0.9	0.2	0.254	0.005	0.008*	0.001	0.088*	0.004	0.202	0.004

*The t-test of d_{12} ($H_0: d_{12}=0$ or $d_{12}=0.2$) is significant at 5 per cent.
Under H_1, the time effect bias on the score scale could not be assessed.

(role emotional), role limitations due to physical health problems (role physical), and mental health. The role physical dimension (RP) includes four dichotomous items. The data of the 57 patients from the study were analyzed using the three methods to estimate time effect, variances, and correlations. Item responses were summed to obtain a score on a scale of 0 (lowest symptom level) to 4 (highest symptom level).

The results of the analysis of the RP dimension of the SF-36 with SM, RM, PV, and LRM methods are presented in Table III. SM, RM, and PV methods used a covariance matrix of AR(1) type. LRM method used an unstructured covariance matrix. All the methods have rejected the hypothesis of equality of the means at $\alpha=5$ per cent level. The estimated values of time effect claim for an improvement of quality of life on role physical dimension at 3 months after surgery and a stability between the third and sixth month after surgery. On simulation data, time effect was underestimated for the RM and PV methods under H_1 whereas the time effect was unbiased for the LRM method. Effect sizes estimated on SF-36 data are large (from 0.25 to 1.13), much larger than effects sizes used for simulation data ($d_{t,t+1}=0.2$), especially between the first and second times. The fact that all methods conclude to a time effect, even for the RM and PV methods for which power is low, might be due to the large effect size. A large effect size can be more often detected than a medium one. The correlation coefficients between two measurements were estimated around 0.6 for each method except for the PV method where the correlation coefficient was estimated around 0.3. The case of simulation data where $N=100$, $J=4$, and $\rho=0.7$ is the closest to the SF-36 data. For this simulation case, the power was estimated to 37.2, 12, 14.2, and 42.3 per cent for the SM, RM, PV, and LRM methods, respectively. An effect of the sample size on power was shown. We might argue that power for SF-36 data might be lower than power found in the simulation study that was evaluated for a larger sample size.

Table III. Results of the analysis of the dimension 'Role Physical' of the SF-36 for Score and Mixed models (SM), Rasch and Mixed models (RM), Plausible Values (PV), and Longitudinal Rasch Mixed model (LRM). Estimations of time effect between time t and t' ($\hat{d}_{tt'}$), variance at time t ($\hat{\sigma}_t^2$), and correlation between time t and t' ($\hat{\rho}_{tt'}$) for $t=1,2,3$ and $t \neq t'$. Test statistic and p-value of the test of equality of the means (Fischer test for RM and SM, Wald test for LRM).

	SM	RM	PV	LRM
$\hat{d}_{12}$	1.32	2.23	2.71	3.10
$\hat{d}_{23}$	−0.45	−0.76	−0.78	−0.94
$\hat{d}_{13}$	0.87	1.47	1.93	2.16
$\hat{\sigma}_1^2$	2.41	6.87	9.80	9.46
$\hat{\sigma}_2^2$	2.41	6.87	9.80	5.61
$\hat{\sigma}_3^2$	2.41	6.87	9.80	18.86
$\hat{ES}_{12}$	0.85	0.85	0.86	1.13
$\hat{ES}_{23}$	−0.29	−0.29	−0.25	−0.27
$\hat{ES}_{13}$	0.56	0.56	0.62	0.57
$\hat{\rho}_{12}$	0.56	0.56	0.31	0.54
$\hat{\rho}_{23}$	0.56	0.56	0.31	0.61
$\hat{\rho}_{13}$	0.31	0.31	0.10	0.65
Test statistic	16.53	16.62	11.79	12.4
p-value	<0.0001	<0.0001	<0.0001	0.002

5. Discussion

PRO data are widely used in health sciences, in particular for the evaluation of HRQoL. Such data are often measured several times on the same patients in order to study the evolution of the outcome with time. Four methods to analyze longitudinal PRO data were compared: the SM method based on the CTT approach, the RM, PV, and LRM methods all based on the Rasch model. The four methods have shown comparable results in terms of type I error with type I error rates close to 5 per cent. The LRM and SM methods presented comparable power and unbiased time effect estimations. It has been shown that the rise of the sample size, the questionnaire length or the correlation between measures increased the values of power. This expected rise with these parameters is concordant with the results in Glas *et al.* [15]. The impact of sample size and the number of items on power leads to take with caution the results from studies of longitudinal PRO data with small sample sizes and short questionnaires.

The RM and PV methods presented much lower power as compared with SM and LRM ones. Moreover, it has been shown that RM and PV gave biased time effect estimations under H_1. The large underestimation of time effect under H_1 explains the important loss of power of these methods as compared to the two others that were unbiased. The RM and PV methods seem to be inadequate to analyze longitudinal patient reported outcomes data and should be avoided.

An explanation of the poor performance of both methods might come from the first step of estimation with Bayes *Expected A Posteriori* or plausible values. In the RM method, individual latent traits are estimated based on the EAP estimate. The EAP estimate is a point estimate defined as the mean of the posterior distribution. The mean of the EAP estimates is known to be an unbiased estimate of the population mean. Nevertheless, the variance of the EAPs is an underestimate of the variance population [21]. Other point estimates exist that produce estimates by maximizing the likelihood of observed item responses. For example, the maximum likelihood estimate (MLE) is obtained from joint maximum likelihood estimation. The mean of the MLE estimates is also an unbiased estimate of the population mean and the variance of the MLEs is an overestimate of the population variance. The bias in variance of MLE and EAP is not reduced when the sample size increases but it goes down when the number of items increases [16]. A correction based on the reliability index can be used to eliminate the bias in the EAP case [22]. EAP estimates present the problem of shrinkage toward the mean of prior distribution.

It has been shown that the bias due to shrinkage is minimal with over than 20 items [23]. In this study, the number of items is too small to avoid the bias due to shrinkage. As a consequence of estimating the individual latent trait on the overall sample without time factor, the EAPs are shrunk to zero. This leads to an underestimation of the time effect for the RM method.

In the PV method, plausible values are randomly drawn from the posterior distribution of each individual. In contrast to point estimates, plausible values allow patients with the same total score (and so the same posterior distribution) to have different plausible values. The main difference between the RM method and plausible value imputation is that the latter considers the variability of the estimated value for the latent trait in time effect estimation, whereas the RM method does not take into account the uncertainty related to the latent variable estimation. The underestimation of the time effect is reduced by the use of plausible values instead of EAP estimates but the observed bias is still important and the power is much lower than for the SM and LRM methods. As expected with the PV method, the variance is no longer biased but the correlation coefficient is still underestimated as with the RM method (results not shown). This bias probably affects the estimated covariance matrix of $\hat{\beta}$ in the mixed model step and may explain the poor performance in terms of power because the F-test used for testing time effect and the computation of type I error and power uses this estimated covariance matrix. The poor performance of the 2-step Rasch-based methods (RM and PV) against 1-step Rasch-based method (LRM) pleads for the use of a multivariate form of the Rasch model to account for the particular structure of repeated measures.

The plausible value imputation is widely used in large-scale educational surveys where the number of items and sample size are much larger than in health sciences. In health sciences, Glas *et al.* [15] have shown that, in most of the cases that were studied, a longitudinal IRT model and plausible value imputation methods lead to comparable results in terms of type I error rate and power in the context of two groups with two time points. Nonetheless, the longitudinal IRT model performs better than plausible value imputation method when the number of items is small ($J=5$ or 10) which is the case in our simulation study ($J=4$ or 7). Furthermore, no comparisons were made with a method based on the CTT, which is widely used in practice, and no more than two time points were studied by the authors.

A linear time effect was assumed in this study. This assumption can be inadequate in some cases. For example, in a QOL study where three assessments take place before treatment, during treatment and after treatment, the treatment can have a deleterious effect on quality of life level. We can assume, for instance, that the quality of life decreases between the first and second assessment and increases between the second and third assessment hence leading to a quadratic evolution with time.

Covariance structures such as AR(1) are adequate in a context of equally spaced time, but in practice time of assessments they can be unequally spaced due to the design of the study or problems of recruitment and followup. In the illustrative example, LRM method has shown different variances over time. This indicates that the assumption of constant variances made in this simulation study can be inappropriate for some other studies. Moreover, covariance matrix in the example did not seem to have an AR(1) structure as the correlation between time 1 and time 3 was not close to the square of correlation between consecutive times. Correlation seemed to be constant whatever the time points chosen.

All results from covariance pattern models were presented for an AR(1) structure for the covariance matrix. Regarding the value of AIC for different structures of covariance matrix, the CSH structure most often minimized the AIC of the models for $\rho=0.9$. Investigating the impact on results of misspecification of covariance matrix structure by performing analyses with a CSH structure led to comparable results than the analyses using AR(1) structure in terms of type I error, power and time effect bias (results not shown).

The repeated use of a questionnaire can cause a problem in terms of response-shift. For instance, the assessment of quality of life over time is based on the assumption that the perception that patients have of their own quality of life will not change over time. But patients are faced with a disease and its treatment that may change their perception leading to the phenomenon called response-shift. As patients are adapting to the adverse effects of the disease and its treatments, the repeated measures of quality of life become difficult to compare due to the response-shifts. An observed evolution of a patient's quality of life may confound a true change in the quality of life and the change of patient's perception. As defined by Barclay *et al.* in a recent review of the subject [24], response-shift involves a change in the meaning of an individual's self evaluation of HRQoL as a result of a change in their internal standards, values and/or concepts of HRQoL. Three components of response-shift are identified: recalibration (a change in the respondent's internal standard of measurement), reprioritization (a change in the importance of component domains constituting the target construct), and reconceptualization (a redefinition of the target construct) [25]. Specific designs like the then-test have been developed to detect response-shift. They have been first used in the area of educational training interventions and then in the area of quality of life, in particular for cancer patients. Treatments of cancer patients can be harmful for quality of life and it has been shown that these patients succeed in adapting to the adverse effects of the disease and its treatments [26]. Since then, the impact of health state changes on an individual's quality of life has gained increased attention in social and medical clinical research. The response-shift that may occur is now considered in studies on evolution of quality of life but the debate on which method to use to detect the response-shift still continues. Some methods are addressing the problem of the response-shift at the design stage of the study, such as the then-test and the individualized methods. Other methods are statistical methods to address response shift, such as factor analysis, growth curve analysis, and Rasch analysis. Among all these possibilities, the then-test is the most commonly used method to measure response-shift. Different components of the response-shift are detected from a method to another. The major point of development remains the quantification of the response-shift. Whereas each method allows to detect it, only the then-test and the factor analysis give a value of change of quality of life adjusted for response-shift effect. Although this simulation study assumed no response-shift, this subject has gained major concern in longitudinal PRO studies and it will be of interest to study the behavior of CTT and Rasch-based methods when response-shift is present.

Finally, many longitudinal studies are faced with the problem of missing observations. In this case, different approaches are often adopted: complete-case analysis, available-data analysis, and imputation. Because the SM method is based on the score computed by summing item responses, in the presence of missing data, the analysis can only be performed through complete-case analysis or imputation approach. In the LRM method, based on item responses, the analysis can also be performed with available-data approach.

Little and Rubin [27] made distinctions between missing value processes. A missing data process is said to be missing completely at random (MCAR) if the missingness is independent of both unobserved and observed data. Data are missing at random (MAR) if the missingness is independent of the unobserved measurements, conditional on the observed data. Otherwise, the missing data process is missing not at random (MNAR). Likelihood-based analyses that ignore the missing data mechanism lead to valid analyses when the missingness is ignorable (MCAR or MAR) [28]. Selection models and pattern-mixture models [29] were proposed to model nonignorable nonresponse. They are an interesting way of dealing with MNAR missing data process by modeling explicitly the missing data mechanism. These models have to be used with caution because untestable assumptions have to be made on the missing data process for selection models and untestable identifying restrictions are used in pattern-mixture models.

Each approach for handling missing data leads to different results and possible bias. It seems important to study the impact of missing data on the performances of methods to analyze longitudinal latent variables. We suspect that there will be a more important loss of information using a CTT-based method than a Rasch-based method because of the necessity to impute for missing data or to use only complete cases in SM method. In the presence of missing data, we expect that the LRM method will present better results than the SM method as it has been shown in the context of sequential analysis of latent variables [30].

This simulation study is based on the assumption that the data follow a Rasch model. The different results on the performance of the methods will probably be affected to different extents if the Rasch model does not correctly fit the data. In this case, we can expect that the CTT approach will perform better than methods based on the Rasch model.

In conclusion, it has been shown that using either the SM or LRM method give comparable and satisfying results. These two methods are adequate for the analysis of longitudinal PRO data following a Rasch model without missing data.

Acknowledgements

This work was supported by the Ligue Nationale Contre le Cancer.

References

1. Gotay CC, Kawamoto CT, Bottomley A, Efficace F. The prognostic significance of patient-reported outcomes in cancer clinical trials. *Journal of Clinical Oncology* 2008; **26**(8):1355–1363.
2. Rasch G. *Probabilistic Models for Some Intelligence and Attainment Tests* (expanded edn). University of Chicago Press: Chicago, 1980.
3. Verbeke G, Molenberghs G. *Linear Mixed Models for Longitudinal Data*. Springer: New York, 2001.
4. Fischer GH, Molenaar IW. *Rasch Models*, *Foundations*, *Recent Developments*, *and Applications*. Springer: New York, 1997.
5. Bjorner JB, Petersen MA, Groenvold M, Aaronson N, Ahlner-Elmqvist M, Arraras JI, Brédart A, Fayers P, Jordhoy M, Sprangers M, Watson M, Young T. Use of item response theory to develop a shortened version of the EORTC QLQ-C30 emotional functioning scale. *Quality of Life Research* 2004; **13**(10):1683–1697.
6. Garcia SF, Cella D, Clauser SB, Flynn KE, Lad T, Lai JS, Reeve BB, Smith AS, Stone AA, Weinfurt K. Standardizing patient-reported outcomes assessment in cancer clinical trials: a patient-reported outcomes measurement information system initiative. *Journal of Clinical Oncology* 2007; **25**(32):5106–5112.
7. Rijmen F, Tuerlinckx F, De Boeck P, Kuppens P. A nonlinear mixed model framework for Item Response Theory. *Psychological Methods* 2003; **8**(2):185–205.
8. Embretson S. A multidimensional latent trait model for measuring learning and change. *Psychometrika* 1991; **56**(3):495–515.
9. Fitzmaurice G, Laird N, Ware SJ. *Applied Longitudinal Analysis*. Wiley: Hoboken, 2004.
10. Fitzmaurice G, Davidian M, Verbeke G, Molenberghs G. *Longitudinal Data Analysis*: *A Handbook of Modern Statistical Methods*. Chapman & Hall/CRC: London, 2008.
11. Hoijtink H, Boomsma A. On person parameter estimation in the dichotomous Rasch model. In *Rasch Models*, Fischer GH, Molenaar IW (eds). Springer: New York, 1997; 53–68.
12. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. Wiley-IEEE: New York, 2004.
13. Wu M, Adams RJ. *PISA 2000 Technical Report*. OECD Publications: Paris, 2002.
14. Thomas N. Assessing model sensitivity of the imputation methods used in the national assessment of educational progress. *Journal of Educational and Behavioral Statistics* 2000; **25**(2):351–371.
15. Glas CAW, Geerlings H, van de Laar MAFJ, Taal E. Analysis of longitudinal randomized clinical trials using item response models. *Contemporary Clinical Trials* 2009; **30**(2):158–170.
16. Wu M. The role of plausible values in large-scale surveys. *Studies in Educational Evaluation* 2005; **31**(2–3):114–128.
17. Littell RC, Milliken GA, Stroup WW, Wolfinger R. *SAS System for Mixed Models*. SAS Insitute Inc: Cary, NC, 1996.
18. Hardouin J. Rasch analysis: estimation and tests with Rasch test. *Stata Journal* 2007; **7**(1):22–44.
19. Zheng X, Rabe-Hesketh S. Estimating parameters of dichotomous and ordinal item response models with gllamm. *Stata Journal* 2007; **7**(3):313–333.
20. Caillard C, Sebag F, Mathonnet M, Gibelin H, Brunaud L, Loudot C, Kraimps JL, Hamy A, Bresler L, Charbonel B, Leborgne J, Henry JF, Nguyen JM, Mirallié E. Prospective evaluation of quality of life (SF-36v2) and nonspecific symptoms before and after cure of primary hyperparathyroidism (1-year follow-up). *Surgery* 2007; **141**(2):153–160.
21. Mislevy RJ, Beaton AE, Kaplan B, Sheehan KM. Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement* 1992; **29**:133–161.
22. Adams RJ. Reliability as a measurement design effect. *Studies in Educational Evaluation* 2005; **31**(2–3):162–172.
23. Wainer H, Thissen D. Estimating ability with the wrong model. *Journal of Educational and Behavioral Statistics* 1987; **12**(4):339–368.
24. Barclay-Goddard R, Epstein JD, Mayo NE. Response shift: a brief overview and proposed research priorities. *Quality of Life Research* 2009; **18**(3):335–346.
25. Schwartz CE, Sprangers MAG. Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. *Social Science and Medicine* 1999; **48**(11):1531–1548.
26. Sprangers MAG. Response-shift bias: a challenge to the assessment of patients' quality of life in cancer clinical trials. *Cancer Treatment Reviews* 1996; **22**(Supplement 1):55–62.
27. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. Wiley: New York, 1987.
28. Troxel AB, Fairclough DL, Curran D, Hahn EA. Statistical analysis of quality of life with missing data in cancer clinical trials. *Statistics in Medicine* 1998; **17**:653–666.
29. Little RJA. Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association* 1995; **90**(431):1112–1121.
30. Sébille V, Hardouin J, Mesbah M. Sequential analysis of latent variables using mixed-effect latent variable models: impact of non-informative and informative missing data. *Statistics in Medicine* 2007; **26**(27):4889–4904.