Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2020, Vol. 29(4) 962–986

DOI: 10.1177/0962280219856375
journals.sagepub.com/home/smm
SAGE

Latent growth modeling of IRT versus CTT measured longitudinal latent variables

R Gorter,[1] J-P Fox,[2] G Ter Riet,[3] MW Heymans[4] and JWR Twisk[4]

Abstract

Latent growth models are often used to measure individual trajectories representing change over time. The characteristics of the individual trajectories depend on the variability in the longitudinal outcomes. In many medical and epidemiological studies, the individual health outcomes cannot be observed directly and are indirectly observed through indicators (i.e. items of a questionnaire). An item response theory or a classical test theory measurement model is required, but the choice can influence the latent growth estimates. In this study, under various conditions, this influence is directly assessed by estimating latent growth parameters on a common scale for item response theory and classical test theory using a novel plausible value method in combination with Markov chain Monte Carlo. The latent outcomes are considered missing data and plausible values are generated from the corresponding posterior distribution, separately for item response theory and classical test theory. These plausible values are linearly transformed to a common scale. A Markov chain Monte Carlo method was developed to simultaneously estimate the latent growth and measurement model parameters using this plausible value technique. It is shown that estimated individual trajectories using item response theory, compared to classical test theory to measure outcomes, provide a more detailed description of individual change over time, since item response patterns (item response theory) are more informative about the health measurements than sum scores (classical test theory).

Keywords

Longitudinal data, multilevel item response theory, questionnaires, classical test theory, latent growth model, multiple imputation

1 Introduction

The use of questionnaires to measure patient reported outcomes (PROs) in clinical research is widespread when no objective measurement is possible. Questionnaire data is essential in those areas of medical research, to construct convincing evidence and to make inferences. The increasing emphasis on quality-of-life and patient-focused outcomes further stimulates the use and development of questionnaires in clinical research. Furthermore, there is a growing interest in modeling the longitudinal development of these outcomes over time. The analysis of (longitudinal) questionnaire data is, however, not straightforward and no generally accepted gold standard is available.

There are two popular psychometric methods that address the (longitudinal) measurement of persons given their responses to questionnaire items (e.g. Hambleton and Jones[1]). In classical test theory (CTT), test scores are related to (unobservable) true scores, and according to the CTT model an observed (sum) score is linearly linked to

[1]Brain research & Innovation Centre, Ministry of Defence, Utrecht, The Netherlands
[2]Faculty of Behavioural, Management & Social Sciences, Department of Research Methodology, Measurement, and Data Analysis, University of Twente, Enschede, The Netherlands
[3]Department of General Practice, Amsterdam University medical centre, Amsterdam, The Netherlands
[4]Department of Epidemiology & Biostatistics, Amsterdam Public Health Research Institute, Amsterdam University Medical Centre, Amsterdam, The Netherlands

Corresponding author:
R Gorter, Brain research & Innovation Centre, Ministry of Defence, Utrecht 3509, The Netherlands.
Email: rosaliegorter@gmail.com

the sum of a true and an error score. The CTT model is often used because it is a relatively simple way to analyze questionnaire data and the assumptions of the CTT model are easily met. The other psychometric model is referred to as an item response theory (IRT) model which defines a relationship between the patient's response to an item and the construct score (e.g. latent variable). A critical difference is that CTT is based on aggregate data information, where an observed test score is linearly related to the true score. The observed test score is constructed from the categorical item response data, and for this reason it is an aggregate score (Hambleton et al., 1993; Chap.3).[2] The error component represents the deviation between the observed test score and the true score, where the error is assumed to be random across similar parallel test forms. In most medical research with questionnaires, replications of test forms are not available, and a common error variance is assumed across persons (Hambleton et al., 1993; Chap.7).[2] In contrast to CTT, IRT is built on an item-construct relationship, where the latent variable or construct is nonlinearly related to each item response through an item response function. An IRT model is defined at the level of observations and therefore preserves all item-level information.

In the study of Gorter et al.,[3] numerical differences between the two psychometric models were explicitly quantified. They showed that the measurement model choice influences the estimated amount of within-subject and between-subject variance given questionnaire data retrieved under a repeated measurements design. When construct measurements retrieved via sum scores were used as outcomes of a repeated measurements model, the within-subject variance was often overestimated and the between-subject variance underestimated. The estimated CTT scores (sum scores) ignore differences in response patterns between individuals leading to the same sum score and this led to bias in the estimated intra-individual and inter-individual sum score based variances in construct measurements. Under various conditions, construct scores retrieved via IRT did not show bias in the estimated structural variance components. Furthermore, it was shown that CTT scores (sum scores) led to bias in the regression effects under various conditions for data retrieved from a randomized controlled trial (RCT), where effect estimates given IRT scores did not show bias.[4]

In the current study, longitudinal questionnaire data on health-related constructs are considered. This type of data is very common in medical and epidemiological studies. For example, growth trajectories of cognitive abilities,[5] post traumatic stress symptoms[6] or quality of life.[7] In this type of research, latent growth modeling can be used to model trajectories of (measured) constructs.[8] Constructs of interest are measured using a measurement instrument (e.g. a test or questionnaire), and subsequently, a measurement model is required to estimate construct values given multiple-item observations. The purpose of the present paper is to investigate the influence of the measurement model choice on the analyses of growth trajectories of health. Health status is measured across time using a measurement model and questionnaire data. Individual trajectories of health are described by latent growth parameters representing the initial health status, and the linear trend and potentially higher-order trend components to model non-linear trends. Through a simulation study it is shown that the most often used CTT model with sum scores leads to bias in the latent growth parameter estimates, which describe the shape of estimated growth trajectories. The use of sum scores for measuring the longitudinal outcome variables is also shown to lead to biased estimates of relevant predictor variables. Although it is possible in CTT modeling to estimate item parameters and other more advanced techniques, we aim to compare the most popular and frequently used sum score CTT model[2(Ch.7)] with using IRT scores in latent growth models (LGMs) to stress the importance of the measurement model choice and show the influence on the LGM parameter estimates directly. In this simulation study, the latent growth of CTT-based sum scores is directly compared to latent growth of IRT-based latent variables.

The measurement of a repeatedly measured construct (i.e. longitudinal latent variable) is based on the response data but also on the information from the LGM, which implies that the response observations alone are not sufficient information to make inferences about the construct value. The information of the LGM, representing the distribution of the construct across time, is also needed to include for instance the individual trajectory information in the estimation of a time-specific construct measurement. In a Bayesian modeling approach, the measurement model combined with the LGM given the observed response data leads to a posterior distribution for the construct. This posterior distribution depends on the choice of the measurement model, where under IRT a weighted average of the item response data and under CTT a weighted test score is used to construct the posterior. Therefore, in contrast to CTT, under IRT all response information is utilized (even when some items responses are missing), where different response patterns lead to different posterior distributions of the construct.

The health outcomes under both measurement models are not measured on the same scale. As a result, latent trajectory estimates using IRT cannot be compared directly to those using sum scores. When using for instance a mixed logistic regression model with person, time, and item as levels of a three-level design, the scale of the

regression parameter estimates will depend on the scale of the construct measurement. The scale of the regression parameter estimates resulting from a mixed linear model with two levels, person and time, will depend on the scale of the test scores. Therefore, a plausible value procedure is proposed to accommodate scaling differences between measured health constructs, while at the same time dealing with the measurement error for the construct measurements. The re-scaled plausible values under each measurement model (IRT and CTT) serve as outcomes of an LGM to estimate growth parameters. The proposed procedure makes it possible to estimate growth parameters under both measurement models on a common scale.[3,4] In a simulation study, the influence of each measurement model on the estimation of the growth parameters is investigated and the latter are directly compared.

In the present paper, the findings of a simulation study, a longitudinal study on chronic obstructive pulmonary disease (COPD), and a study on coping with back pain are used to show the possibilities of latent growth modeling for longitudinal latent variables measured with IRT, while at the same time the comparison is made with LGMs using sum scores (CTT). The differences between IRT and CTT are stressed in the longitudinal latent variable modeling of health constructs. It is shown that under CTT the magnitude of linear trend effects are underestimated, quadratic trend effects are not always detected, and individual differences in trajectories are underestimated, mainly since CTT does not utilize all response information.

In Section 2, the properties of IRT and CTT are discussed in a brief introduction. In Section 3, latent curve models are discussed for modeling individual trajectories of health measurements. In Section 4, a plausible value method is described to account for scale differences in the IRT-based and CTT-based health measurements. Furthermore, a Markov chain Monte Carlo (MCMC) method is proposed to estimate the parameters of the LGM given the re-scaled plausible values. In a simulation study the implications of using sum-scores as measurements of latent variables in latent growth modeling is shown, while making the comparison with IRT-based plausible values on a common scale. Subsequently, data from two real data studies are used to illustrate the findings of the simulation study. In Section 8, a discussion of the findings and an overall conclusion are presented.

2 Measurement models

There are two important traditions when it comes to measurement models, CTT and IRT. Both measurement models make different assumptions on the relation between the latent variable and the item scores. An important difference is that IRT models describe the relation between the observed item scores and the latent variable using the probability of choosing one over the other answering categories of an item, while taking into account the item characteristics. By using the lower-level item responses, differences in answering patterns across items lead to different scores. This is in contrast to (aggregated) test scores, usually sum scores, used in CTT, where item properties are ignored and different answering patterns can lead to the same construct score. The difference in the calculation of the score leads to more variability in the IRT scores, and thus to a more realistic rendition of the true trait levels. Note that these differences between IRT and CTT become apparent, when concurrently estimating all parameters including those from the LGM. In a one-step (simultaneous) estimation method, the latent variable estimates depend on the response data and the LGM distribution for the latent variable. In that case, the latent variable estimate depends on the item response data (under IRT) or the sum-score (under CTT) and on the LGM. As a result, differences between IRT and CTT become apparent, since different sources of information are used to make inferences about the latent variable estimates.

In IRT modeling, each binary response $Y_{ijk} = 1$ has a success probability that is presumed to be a function of patient i ($i = 1, \ldots, N$) at measurement occasion j ($j = 0, \ldots, n_i$), the parameters of the measurement model for item k ($k = 1, \ldots, K$), and the latent variable θ_{ij}. We use the two parameter normal ogive model[2] for dichotomous items

$$P(Y_{ijk} = 1|\theta_{ij}, a_k, b_k) = \Phi(a_k\theta_{ij} - b_k), \tag{1}$$

where $\Phi(.)$ denotes the standard normal cumulative distribution function. The item difficulty is denoted by b_k and the discrimination parameter by a_k. The latent variables of patient i at occasion j, θ_{ij}, are considered longitudinal latent variables since they are related to observed variables, which are measured at multiple time points. The item parameters are assumed to be invariant across measurement occasions, but item parameters can be occasion-specific, for instance, in case of an incomplete design (i.e. when the questionnaire changes over time and/or when patients do not respond to all questionnaire items).

For ordinal response data, the graded item response model is used (GRM),[9] also called the ordinal probit model,[10,11] which is given by

$$p(Y_{ijk} = c|\theta_{ij}, a_k, \tau_k) = \Phi(a_k\theta_{ij} - \tau_{kc-1}) - \Phi(a_k\theta_{ij} - \tau_{kc}) \quad (2)$$

In the GRM, the probability is modeled that an individual i at occasion j with underlying latent variable θ_{ij}, gives a response into category c ($c = 1, \ldots, C_k$) on item k. The $C_k - 1$ threshold parameters are denoted by τ_{kc}. The probability that the response y_{ijk} falls into category c is the probability of scoring in category c or below minus the probability of scoring in category $c - 1$ or below. The response categories are ordered as $-\infty \leq \tau_{k1} \leq \tau_{k2} \leq \tau_{k3} \leq \infty$.

For CTT, the observed responses are aggregated to a sum score (other test scores are possible), $\bar{Y}_{ij}$, and the measurement of the true score, ϑ_{ij}, of patient i at occasion j is given by

$$\bar{Y}_{ij} = \vartheta_{ij} + E_{ij} \quad (3)$$

where E_{ij} is the (random) measurement error component. The error component represents the difference between the observed score and the true score on a specific test occasion, and independent random errors are assumed for similar (parallel) tests across occasions. In practice, results from parallel tests are not available and a common error variance is assumed across patients. Then, the common error variance, denoted as σ_y^2, can be computed from the patient scores. CTT does not imply a distribution for the errors, but a normal distribution can be assumed for the errors. This distributional assumption facilitates, in a Bayesian modeling approach, the construction of a posterior distribution for the true score. As a result, the errors are independently and normally distributed with a mean of 0 and variance σ_y^2, and the sum scores are considered to be linearly related to the patients' true scores. The IRT latent variable θ_{ij} and the CTT true score ϑ_{ij} are measured on different scales but they represent the same construct (Lord, 1980; p.46),[12] and both are measured using the same items in the test. The (theoretical) construct under each model is similar, although the constructs are measured in different ways. Furthermore, the metric of the scale on which the true score is measured depends on the items in the test, but this also holds for the scale on which the latent variable θ_{ij} is measured.

When considering the CTT (IRT) model as the level-1 part of the model, the LGM describes the change in true (IRT) scores over time and the LGM represents the higher-level part of the model. Inferences about the true (IRT) scores are based on the response data and the LGM using a simultaneous estimation procedure. It is often assumed that a simple random sample of patients is obtained and the latent variable and the true score are modeled as random effects. In that case, patients are randomly selected from a population, and a normal distribution represents the population of patients from which a simple random sample is obtained. Although it is often a priori assumed that the population distribution of the latent variable is normal, the posterior distribution can be asymmetric after updating the prior to the posterior via the likelihood using the response data. In epidemiological studies, a symmetric latent variable (population) distribution is not always present and more advanced parametric measurement techniques are required to describe the response data. For instance, in epidemiological studies, a diagnostic test can be used to identify persons with a deficiency and in the population this concerns a minority of the people. This is represented by an asymmetric population distribution (e.g. Beekman et al.[13] and Reijnen et al.[14]), since people without the deficiency represent the majority.

In the comparison of CTT with IRT, the computation of the posterior distribution of the latent variable (true score) is based on a simultaneous (one-step) estimation method, where all available information is used to make inferences about the latent variable (true score). The reason is that the LGM provides specific information about the time-specific latent variables, which needs to be included in the posterior distribution of the latent variable. Differences between IRT and CTT become apparent when considering the posterior distributions of the model parameters based on the response data and the LGM. Thus, when comparing IRT with CTT in combination with an LGM for longitudinal latent variables, joint inferences will be made by considering a simultaneous estimation method.

Note that if a two-stage estimation method would be used instead of the simultaneous estimation method, the latent variable scores are based solely on the response data per measurement occasion and the LGM parameters are estimated given the latent variable scores. In this two-step estimation approach, the intra-patient correlations across time specified by the LGM are ignored in the computation of the latent variable scores and this will lead to biased LGM parameter estimates.[15] The measurement errors associated with the latent variable measurements are

also ignored, which will also lead to biased LGM estimates and an underestimation of the standard errors of the LGM parameter estimates. When comparing IRT with CTT based on the two-step estimation procedure, and the one-parameter IRT model is considered, the sum score is considered to be a sufficient statistic for the latent variable. As a result, the sum score and the IRT score contain similar information and will not render differences in LGM estimates.

3 Latent growth models for changes in health status

Latent growth models (LGMs) are used to describe differences between individual health trajectories over time.[8,16] A pattern of change is usually described by a random intercept and a random slope, where each patient-specific pair describes a latent trajectory. The random intercept and slope are considered to be latent variables and represent the initial status and the rate of change of a patient's health status. When considering the health status θ_{ij} of patient i measured at occasions $j = 0, \ldots, n_i$, an LGM can be defined by using a latent growth factor $\boldsymbol{\beta}$, which includes a random intercept β_{0i} and a rate of change β_{1i}. For patients $i = 1, \ldots, N$, this LGM is given by

$$\begin{pmatrix} \theta_{i0} \\ \theta_{i1} \\ \vdots \\ \theta_{in_i} \end{pmatrix} = \begin{pmatrix} 1 & t_{i0} \\ 1 & t_{i1} \\ \vdots & \vdots \\ 1 & t_{in_i} \end{pmatrix} \begin{pmatrix} \beta_{0i} \\ \beta_{1i} \end{pmatrix} + \begin{pmatrix} e_{i0} \\ e_{i1} \\ \vdots \\ e_{in_i} \end{pmatrix} \qquad \begin{pmatrix} \beta_{0i} \\ \beta_{1i} \end{pmatrix} = \begin{pmatrix} \gamma_{00} \\ \gamma_{10} \end{pmatrix} + \begin{pmatrix} u_{0i} \\ u_{1i} \end{pmatrix} \tag{4}$$

The error term $\mathbf{e}_i$ is assumed to be normally distributed with a mean zero and variance $\boldsymbol{\Sigma}_e$, where $\boldsymbol{\Sigma}_e$ is a diagonal covariance matrix with diagonal elements σ_j^2. For homoscedastic error variances $\boldsymbol{\Sigma}_e = \sigma^2\mathbf{I}$. The intra-patient correlation across measurements is modeled by the random effects characterizing the latent trajectory. The error term $\mathbf{u}_i$ is multivariate normally distributed with mean zero and covariance matrix $\mathbf{T}$ with diagonal elements τ_0^2 and τ_1^2, and a non-diagonal covariance τ_{01}^2. The LGM for $\boldsymbol{\theta}_i$ can also be represented by

$$\begin{aligned} \boldsymbol{\theta}_i &= \gamma_{00} + u_{0i} + (\gamma_{10} + u_{1i})\mathbf{t}_i + \mathbf{e}_i \\ &= \gamma_{00} + \gamma_{10}\mathbf{t}_i + u_{0i} + u_{1i}\mathbf{t}_i + \mathbf{e}_i \end{aligned} \tag{5}$$

where the mean term is represented by $\gamma_{00} + \gamma_{10}\mathbf{t}_i$. When including explanatory variables, differences at baseline (initial status) can be explained and differences between slopes by including time-variant explanatory variables. When the time of the first measurement, t_{i0}, is coded as zero, then the random intercept, β_{0i}, defines the initial health status and β_{1i} the linear time effect. The other values of t should reflect the spacing between measurement occasions.

The latent growth factor $\boldsymbol{\beta}_i$ represents the characteristics of the trajectory for patient i. The population mean trajectory characteristics are given by a population intercept γ_{00} and a population rate of change γ_{10}. The variation in individual trajectories in the population is described by the covariance matrix $\mathbf{T}$, where τ_0^2 represents the variation in initial values in the population, τ_1^2, the variation in the rate of change in the population, and τ_{01} the covariance between the initial value and the rate of change. The latent growth parameters $\boldsymbol{\beta}_i$ are considered to be second-order latent variables and the $\boldsymbol{\theta}_i$ first-order latent variables.

In Figure 1, the path diagram of the LGM with a random intercept and slope as growth factors is given, according to the latent curve modeling notation,[8,17] where the longitudinal latent variable *health status*, $\boldsymbol{\theta}_i$, is measured by j occasion-specific observations. The model is referred to as M1. The random intercept and random slope are allowed to be correlated and patient-specific explanatory variables can be included to model variation between patients in their initial status and their rate of change.

The modeling of the dynamic changes in longitudinal latent variables can be extended by including higher-order polynomials, leading to nonlinear trajectories. Assuming a linear change over time is often too simplistic. A negative or positive linear trend of patient i can be modified by a quadratic time component to decelerate or to accelerate the trend. The linear and quadratic time components with patient-specific effects can describe a

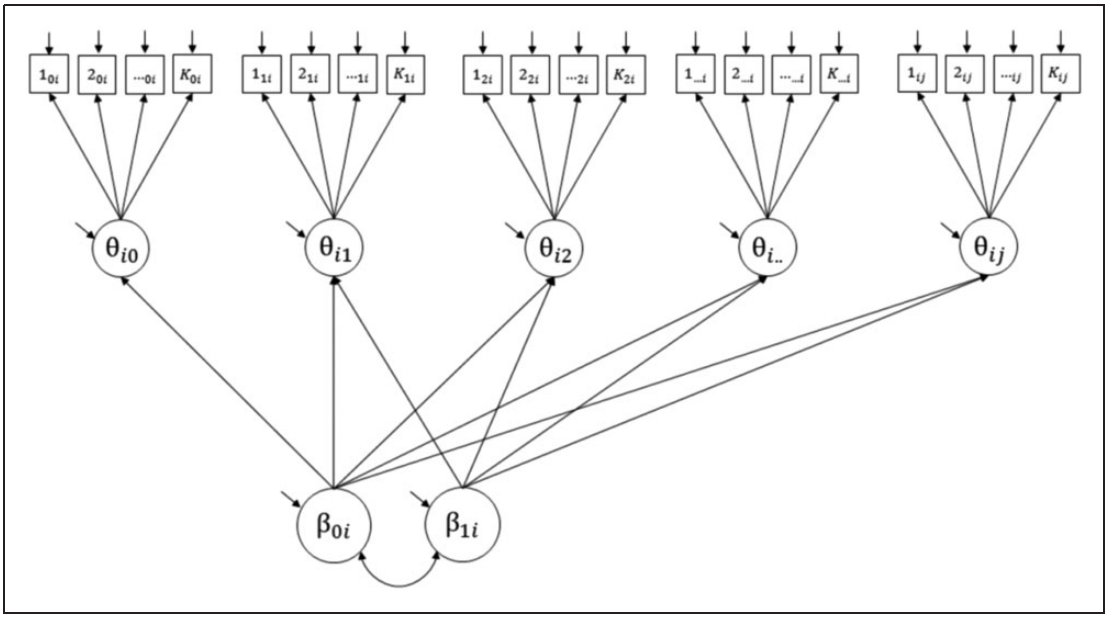

Figure 1. Path diagram of an LGM with a random intercept (initial status) and a linear random slope (linear rate of change).

nonlinear change of the patient's longitudinal latent variable. The quadratic LGM for measurement j of patient i is given by

$$\begin{aligned} \theta_{ij} &= \beta_{0i} + \beta_{1i} t_{ij} + \beta_{2i} t_{ij}^2 + e_{ij} \\ \beta_{0i} &= \gamma_{00} + u_{0i} \\ \beta_{1i} &= \gamma_{10} + u_{1i} \\ \beta_{2i} &= \gamma_{20} + u_{2i} \end{aligned} \tag{6}$$

where the error terms at the patient level $\mathbf{u}_i$, are assumed to be multivariate normally distributed with mean zero and covariance matrix $\mathbf{T}$. The variation in quadratic effects across patients is given by τ_2^2, which represents the variation in the second-order latent variable β_{2i}. In Figure 2, the quadratic LGM of the longitudinal latent variable, θ_{ij}, is given. Although not explicitly mentioned, the mean term (γ_{00}) can be extended with time-varying and time-invariant explanatory variables. The quadratic LGM can also be extended to include even higher-order polynomials such as a cubic trend. For the present study, intensive simulation studies showed that the parameter recovery of random cubic trend models did not show accurate results. Furthermore, the interpretation of cubic and higher-order polynomial trends is complicated and it might be questionable whether such trends can be expected.[18] For many applications, the quadratic LGM defines a good balance between model fit and interpretation.

4 Plausible values to accommodate scale differences

There are several ways to obtain estimates of the LGM parameters, where an (nonlinear) IRT or (linear) CTT model is used to relate the response data to the longitudinal latent variable. Tutz[19(Ch.8)] and McCulloch and Neuhaus,[20(Ch.14)] among others, describe different methods for estimating the parameters. Maximum likelihood (ML) or restricted maximum likelihood (REML) estimates can be obtained using an EM algorithm. It is also possible to obtain more robust parameter estimates through penalized quasi-likelihood estimation or a Laplace approximation. Furthermore, Markov chain Monte Carlo (MCMC) methods can be used to obtain posterior mean estimates. Although there are many ways to estimate the LGM parameters, LGM estimates with IRT as the level-1 component (resulting in a 3-level model with the item responses as the data and items, time and patients as the three levels) cannot be directly compared to those with CTT as the level-1 component (resulting in a two-level

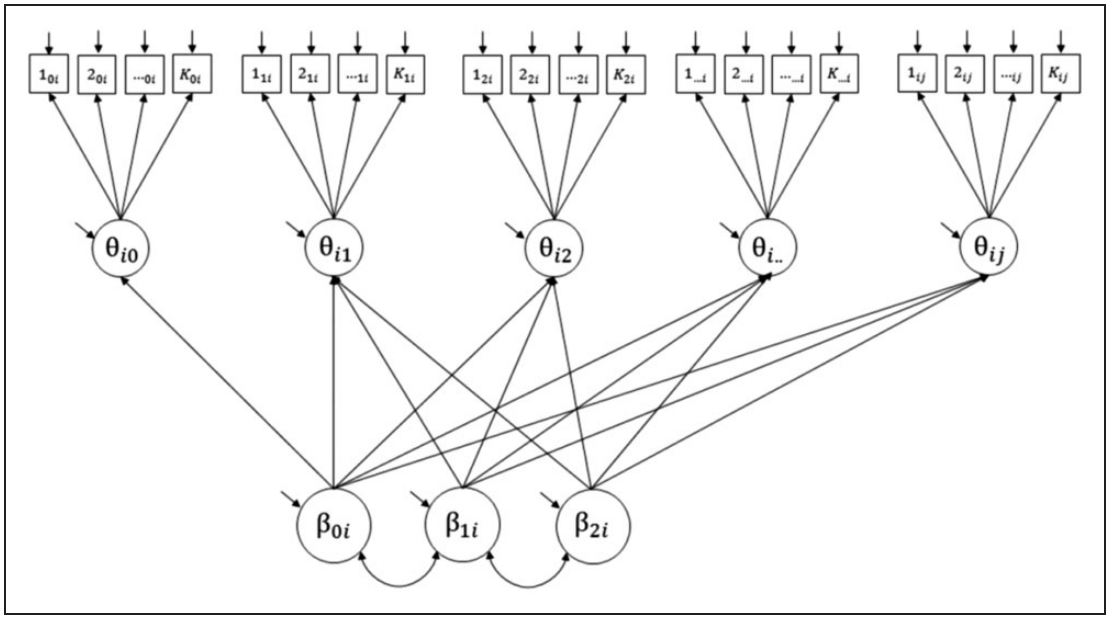

Figure 2. Path diagram of an LGM, Model M2, with a random intercept, linear random slope and a quadratic random slope.

model where the sum scores are the data and time and patients are the levels). The estimates will be defined on different scales, since they are based on different data (i.e. item response data versus a sum score) and it is unknown how to translate the estimates and standard errors from one scale to another. Therefore, a plausible value technique is used, which makes it possible to identify the underlying latent scale and to translate LGM estimates on one scale to another.

Plausible values have been used for the analysis of large-scale surveys, where there is an interest in the parameters of the population distribution of the latent variable. The plausible values represent possible realizations that the latent variable can take given the response data. The plausible values are not point estimates of the latent variable but they are random draws of the posterior distribution of the latent variable. A set of plausible values across patients can be seen as a draw from the population distribution. The theory of plausible values has been developed by Mislevy and colleagues[21,22] and is based on Rubin and Schenkers work on multiple imputations.[23]

There are three arguments to use plausible values in the estimation of population parameters of the latent variable. First, when point estimates of the latent variable are used, bias in the variance estimates of the population parameter estimates can be obtained because of the uncertainty associated with the individual latent variable scores.[22,24] The plausible values can be used to obtain unbiased estimates of the LGM (population) parameters. Second, standard methods (i.e. multilevel models; regression models) can be used when plausible values are available for the latent variable. Third, for complex sampling designs plausible values can be used to get correct standard errors of the parameter estimates. These advantages of plausible values have also been shown in different simulation studies and applications.[21,25–30]

Next to these advantages, plausible values also serve as a perfect tool to handle scale differences in the analysis of latent variables. The procedure is to generate plausible values for the latent variable given the response data for different measurement models. The plausible values cannot be compared across measurement models, since they are generated on different scales. However, the generated plausible values can be transformed afterwards to a common scale by a linear transformation, given that the generated plausible values have a normal distribution. Let θ_{ij}^{PV} denote a plausible value for θ_{ij}, where the vector of plausible values across patients at measurement j has a mean μ_{y_j} and standard deviation σ_{y_j}. The linear transformation of the plausible value θ_{ij}^{PV} to a scale with a mean μ_s and standard deviation σ_s is given by

$$\theta_{ij,s}^{PV} = \mu_s + \frac{\sigma_s}{\sigma_{y_j}}\left(\theta_{ij}^{PV} - \mu_{y_j}\right) \tag{7}$$

where $\theta_{ij,s}^{PV}$ is the linearly transformed latent variable. Different sets of plausible values retrieved with different measurement models can be transformed to a common scale. As a result, estimated population parameters using plausible values defined on different scales can be compared to each other.[3]

When considering the LGM to describe the pattern of change, plausible values are generated for the longitudinal latent variables and are used to estimate the LGM parameters. The plausible values can be generated from the posterior distribution of the latent variable. Let $\mathbf{\Omega}_{ij}$ denote the patient-specific growth parameters and within-patient variance component, $\mathbf{\Omega}_{ij} = (\boldsymbol{\beta}_i, \mathbf{\Sigma}_e)$ for patient i and occasion j. Then, for the IRT model, the posterior distribution of the latent variable θ_{ij} is given by

$$g(\theta_{ij} \mid \mathbf{y}_{ij}, \mathbf{\Omega}_{ij}, \mathbf{a}, \mathbf{b}) = \frac{p(\mathbf{y}_{ij}|\theta_{ij}, \mathbf{a}, \mathbf{b}) f(\theta_{ij} \mid \mathbf{\Omega}_{ij})}{\int p(\mathbf{y}_{ij}|\theta_{ij}, \mathbf{a}, \mathbf{b}) f(\theta_{ij}|\mathbf{\Omega}_{ij}) \mathrm{d}\theta_{ij}} \tag{8}$$

and subsequently, the plausible values are drawn from

$$\theta_{ij}^{PV} \sim g(\theta_{ij}|\mathbf{y}_{ij}, \mathbf{\Omega}_{ij}, \mathbf{a}, \mathbf{b}) \tag{9}$$

This posterior distribution is normal, when conditioning on latent augmented response data (see Fox, 2010; pp.83–85).[31] Then, the plausible values are drawn from a normal posterior distribution given the latent response data and a common error variance. This is shown by introducing the data augmentation step. In the data augmentation step, normally distributed latent response data, Z_{ijk}, are sampled with mean, $a_k\theta_{ij} - b_k$, and variance 1, and $Z_{ijk} > 0$ if $Y_{ijk} = 1$ and $Z_{ijk} \leq 0$ if $Y_{ijk} = 0$. According to equation (8), the posterior distribution of the latent variable θ_{ij} is constructed from the conditional distribution of the item response data given the item parameters and the latent variable, and the conditional distribution of the latent variable given the latent growth parameters (LGM). Let $\mathbf{X}_{ij}$ contain the explanatory variables of the LGM, including the time of measurement. When conditioning on the latent response data, the posterior distribution of the latent variable given the augmented data is given by

$$\begin{aligned} p(\theta_{ij}|\mathbf{Z}_{ij}, \mathbf{a}, \mathbf{b}, \boldsymbol{\beta}_i, \sigma^2) &\propto \exp\left(-\sum_k \left(Z_{ijk} - \left(a_k\theta_{ij} - b_k\right)\right)^2/2\right) \exp\left(-\sigma^{-2}\left(\theta_{ij} - \mathbf{X}_{ij}\boldsymbol{\beta}_i\right)^2/2\right) \\ &\propto \exp\left(-\left(\sum_k a_k^2\left(\hat{\theta}_{ij} - \theta_{ij}\right)^2\right)/2\right) \exp\left(-\sigma^{-2}\left(\theta_{ij} - \mathbf{X}_{ij}\boldsymbol{\beta}_i\right)^2/2\right) \\ &\propto \exp\left(-\left(\sum_k a_k^2 + \sigma^{-2}\right)\left(\theta_{ij}^* - \theta_{ij}\right)^2/2\right) \end{aligned}$$

where

$$\theta_{ij}^* = \frac{\sum_k a_k^2\hat{\theta}_{ij} + \sigma^{-2}\mathbf{X}_{ij}\boldsymbol{\beta}_i}{\sum_k a_k^2 + \sigma^{-2}}$$

and

$$\hat{\theta}_{ij} = \sum_k a_k^{-2}\left(\sum_k a_k\left(Z_{ijk} + b_k\right)\right)$$

It follows that the posterior distribution of the latent variable given augmented data and LGM parameters is normal

$$\theta_{ij}|\mathbf{Z}_{ij}, \mathbf{a}, \mathbf{b}, \boldsymbol{\beta}_i, \sigma^2 \sim N\left(\theta_{ij}^*, \left(\sum_k a_k^2 + \sigma^{-2}\right)^{-1}\right) \tag{10}$$

In the same way, for the CTT model, it follows that the posterior distribution of the latent variable ϑ_{ij} is given by

$$h\left(\vartheta_{ij} \mid \bar{y}_{ij}, \mathbf{\Omega}_{ij}\right) = \frac{p\left(\bar{y}_{ij} \mid \vartheta_{ij}\right) f\left(\vartheta_{ij} \mid \mathbf{\Omega}_{ij}\right)}{\int p\left(\bar{y}_{ij} \mid \vartheta_{ij}\right) f\left(\vartheta_{ij} \mid \mathbf{\Omega}_{ij}\right) d\vartheta_{ij}} \tag{11}$$

and subsequently, the plausible values are drawn from

$$\vartheta_{ij}^{PV} \sim h\left(\vartheta_{ij} \mid \bar{y}_{ij}, \mathbf{\Omega}_{ij}\right) \tag{12}$$

It can be shown that the posterior distribution of the latent variable given the sum score, $\bar{y}_{ij}$, and LGM parameters is normal

$$\vartheta_{ij} \Big| \bar{y}_{ij}, \sigma_y^2, \boldsymbol{\beta}_i, \sigma^2 \sim N\left(\vartheta_{ij}^*, \left(\sigma_y^{-2} + \sigma^{-2}\right)^{-1}\right) \tag{13}$$

where

$$\vartheta_{ij}^* = \frac{\sigma_y^{-2}\bar{y}_{ij} + \sigma^{-2}\mathbf{X}_{ij}\boldsymbol{\beta}_i}{\sigma_y^{-2} + \sigma^{-2}}$$

and σ_y^2 the measurement error variance. Without assuming a distribution for the errors in the CTT model, or assuming $\sigma_y^2 = 0$, it follows from equation (13) that the posterior distribution is also normal

$$\vartheta_{ij} \Big| \bar{y}_{ij}, \boldsymbol{\beta}_i, \sigma^2 \sim N\left(\mathbf{X}_{ij}\boldsymbol{\beta}_i, \sigma^2\right) \tag{14}$$

The conditional distribution $p\left(\bar{y}_{ij} \middle| \vartheta_{ij}\right)$ is only defined, when assuming a distribution for the errors in the CTT model. When assuming a normal distribution, the plausible values under the CTT model are also normally distributed given measurement error variance σ_y^2, as shown in equation (13). Without assuming a distribution for the error scores, $\left(\sigma_y^2 = 0\right)$, the plausible values are normally distributed according to equation (14). In that case, plausible values are drawn from the posterior (predictive) distribution of the latent variable

$$h\left(\vartheta_{ij} \Big| \bar{y}_{ij}\right) = \int p\left(\vartheta_{ij} \middle| \mathbf{\Omega}_{ij}\right) p\left(\mathbf{\Omega}_{ij} \Big| \bar{y}_{ij}\right) d\mathbf{\Omega}_{ij} \tag{15}$$

where the latent variable $\vartheta_{ij} = \bar{y}_{ij}^{rep}$ represents the predictive data (sum scores) under the LGM. In the simulation study and real data study, plausible values will be drawn according to equation (14). This will avoid assuming a specific distribution for the true scores, and will avoid the use of an incorrect measurement error variance σ_y^2, which is usually unknown.

The plausible values generated under CTT and IRT are both normally distributed, when the IRT-generated plausible values are based on the latent response data. It follows that the plausible values generated from the normal distribution in equation (10) (IRT) can be mapped on the scale of the plausible values generated from the normal distribution in equation (14) (CTT) using a linear transformation function. This is done by applying the transformation function defined in equation (7), to obtain plausible values on one common scale. These plausible values are used as outcomes in the LGM to obtain estimation results on a common scale.

Note that the sum score is not linearly related to the latent variable, θ_{ij}, and also not when conditioning on the augmented response data. It follows that, $\sum_k Y_{ijk} = \sum_k I\left(Z_{ijk} > 0 \middle| \theta_{ij}\right)$, which does not imply a linear relationship between the sum score and the latent variable. Through a data augmentation step under the probit (IRT) model, normally distributed plausible values are defined. They can be mapped (through a linear transformation) to the scale of the normally distributed plausible values under the CTT model. When the distribution of the observed data given the latent variable is considered to compute IRT scores, $p\left(\mathbf{y}_{ij} \middle| \theta_{ij}\right)$, then the IRT scores are not normally distributed and the linear scale transformation is not possible.

5 MCMC and the plausible value procedure

MCMC can be used to estimate all model parameters, together with the generation of plausible values. Additionally, given the plausible values, which serve as dependent variables of the LGM, parameters of the LGM can also be estimated using an MCMC algorithm. In this study, both estimation steps will be carried out, since we want to obtain estimation results of IRT-LGM and CTT-LGM on a common scale. A detailed description of the plausible value procedure is given in Section 5.3. In general, MCMC is repeatedly applied to estimate the LGM parameters for each set of plausible values, where the final LGM estimates are computed as the average of estimates across replications.

5.1 Priors

Priors need to be specified for all model parameters. For dichotomous data, a normal prior is specified for the discrimination and difficulty parameter with mean 1 and 0, respectively, both with a high variance (around 10) to specify an uninformative prior. For polytomous data, a diffuse prior is specified for the threshold parameters, which assigns an equal probability for each possible parameter value, while obeying the ordering of the response categories.[31(pp.83–85)] For the LGM, a reference (vague) prior is defined for the population mean trajectory, $p(\gamma) \propto$ constant. For the variance components, an inverse-gamma prior is defined for the variance parameter of the within-patient errors, σ^2, and an inverse-Wishart prior for the covariance matrix $\mathbf{T}$, $\mathbf{T} \sim IW(\nu_T, S_T)$, with ν_T the prior degrees of freedom and S_T the scale parameter. The prior specifications are quite standard, and have been explained by Fox[31] and Lee.[32]

The MCMC algorithm consists of several steps, which describes the sampling from the conditional distributions of the model parameters. The sampling steps are given in Appendix 1. The MCMC algorithm consists of three blocks, which describe the sampling of IRT, CTT, and LGM parameters.

5.2 MCMC algorithm

(1) Sample IRT model parameters. A data augmentation scheme is used to sample latent continuous data, z, which are normally distributed.[31(pp.83–85)] For binary data, item parameters are sampled from normal distributions given the augmented response data. For polytomous data, a Metropolis-Hastings step is used to sample threshold parameters.[31(pp.83–85)] The conditional distribution of the latent variable is given by equation (10), which is a normal distribution, when conditioning on augmented data $\mathbf{z}$ and LGM parameters $\boldsymbol{\beta}$ and σ^2.
(2) The sampling of (CTT) true score values, ϑ_{ij}, given the parameters $\boldsymbol{\beta}$ and σ^2 is described in equation (13). A normal distribution is assumed for the errors in the CTT model, which leads to a normal posterior for ϑ_{ij}. Without assuming a distribution for the errors in the CTT model, true scores are sampled from a normal distribution specified in equation (14).
(3) Sample LGM parameters. Given the latent dependent variable θ_{ij} or ϑ_{ij}, the sampling steps of the LGM parameters $(\gamma, \boldsymbol{\beta}, \sigma^2, \mathbf{T})$ are given in Appendix 1 (see also Klein Entink et al.[33] and Song and Lee[34]). This block contains the sampling steps for the LGM, which can be repeatedly carried out to estimate the LGM parameters given the drawn plausible values in block 1 and 2.

The MCMC algorithm has been implemented in a modified version of the R-Package mlirt.[10] The convergence of the algorithm can be investigated by observing trace plots of the sampled values. At convergence, the sequences of sampled values should mix well and not show any structural patterns. The convergence diagnostics in the R-package Coda[35] can also be used to investigate whether the chains of sampled values has converged.

5.3 Plausible value procedure

The following procedure can be applied to obtain parameter estimates of the LGM model for the IRT (equation (10)) and CTT (equation (14)) generated plausible values. As a result, the final estimates of the LGM parameters given the IRT-generated plausible values, and those given the CTT-generated plausible values are on the same

scale due to the scale transformation of the plausible values.

(1) Generate plausible values for the latent variables θ_{ij} and ϑ_{ij}, according to block 1 and block 2 of the MCMC algorithm, equations (10) and (14), respectively.
(2) Transform each vector of plausible values to a common scale, according to equation (7).
(3) For each set of plausible values, obtain M draws of each LGM parameter, according to block 3 of the MCMC algorithm, and compute the posterior mean and variance of the LGM parameters.
(4) Repeat steps 1–3 multiple times (usually 5).
(5) Pool the LGM estimation results from step 3 for the IRT- and the CTT-generated plausible values.

In the final step (step 5), the results are pooled across the different draws of plausible values to estimate the posterior mean and variance of the model parameters. To describe this procedure, consider the fixed effect parameter γ in the LGM in equation (4). The posterior mean can be estimated by averaging over the MCMC samples for each plausible vector of the latent variable, and then take the mean over the computed averages. Let M and M_{PV} denote the number of MCMC iterations and the number of plausible values, respectively. The posterior mean is estimated by

$$\widehat{\gamma} = \frac{1}{M_{PV}} \frac{1}{M} \sum_{h=1}^{M_{PV}} \sum_{m=1}^{M} \gamma^{(m,h)} = \frac{1}{M_{PV}} \sum_{h=1}^{M_{PV}} \widehat{\gamma}^{(h)}$$

where $\gamma^{(m,h)}$ denotes a sample from the posterior distribution $p(\gamma|\theta^{(h)}, \mathbf{y})$ at MCMC iteration m given a plausible vector $\theta^{(h)}$. The posterior variance of γ is estimated by the sum of the within and between-imputation variance

$$var(\gamma \mid \mathbf{y}) = \frac{1}{M_{PV}} \sum_{h=1}^{M_{PV}} var\left(\gamma \mid \mathbf{y}, \theta^{(h)}\right) + \frac{1}{M_{PV} - 1} \sum_{h=1}^{M_{PV}} \left(\widehat{\gamma}^{(h)} - \widehat{\gamma}\right)\left(\widehat{\gamma}^{(h)} - \widehat{\gamma}\right)^{t} \tag{16}$$

The variance between plausible values (i.e. second term on the right-hand side; between-imputation variance) can be multiplied with $(1 + 1/M_{PV})$ to improve the approximation when M_{PV} is small. The variance term $\text{var}(\gamma|\mathbf{y}, \theta^{(h)})$ is estimated by the sample variance given sampled values and a plausible vector $\theta^{(h)}$. Subsequently, the within-imputation variance is estimated by the mean of the computed sample variances over all plausible vectors.

6 Simulation study

In the simulation study, the plausible value procedure was applied to make a comparison between the LGM estimation results using the CTT-generated latent variable scores with those using the IRT-generated scores.

6.1 Procedure

For each of the $N = 500$ subjects, $J = 10$ (representing a moderate number of follow-ups) and $J = 50$ (representing a high number of follow-ups) repeated measurements were simulated in simulation study 1 and 2, respectively. For each subject i $(i = 1, \ldots, 500)$, latent variables, θ_{ij} $j = 1, \ldots, 10$ in simulation study 1, and $j = 1, \ldots, 50$ in simulation study 2, were simulated from a normal distribution with mean β_{0i} and variance σ_{θ}^2 equal to .50. The true values of the LGM parameters, referred to as model M1, are reported in Tables 1 and 2 under the column labeled ''True''. Data for both conditions were simulated according to model M1 in equation (4). The random intercept and slope, β_{0i} and β_{1i}, were generated from a normal distribution with mean 0 and .4 and variance $\tau_0^2 = .80$ and $\tau_1^2 = .80$, respectively. For the model M2 given in equation (6), a quadratic time effect was simulated from a normal distribution with mean .40 and variance .80.

The item responses were generated using equation (1) combined with the LGM. The item difficulty parameters were sampled from a normal distribution with a mean of 0 and a variance of .25 and the discriminations were all set to 1, which otherwise would disadvantage the CTT model over the IRT model. The generated item parameters and latent variable values were used to simulate item response data. Subsequently, sum scores, $\bar{y}_{ij}$, were computed from the generated response data, which served as observed scores for the CTT model.

Table 1. Simulation study results for latent curve models M1 and M2 over 10 replications, using IRT and CTT-generated plausible values ($N = 500$, $J = 10$, $K = 20$).

				IRT				CTT			
		Par.	True	Mean	SD	BIAS	MSE	Mean	SD	BIAS	MSE
M1	Fixed part	γ_{00}	0	0.00 [−0.09; 0.09]	0.05	0.00	0.00	0.03 [−0.06; 0.11]	0.05	0.03	0.00
	(L-Trend)	γ_{10}	.4	0.40 [0.30; 0.50]	0.06	0.00	0.00	0.36 [0.26; 0.45]	0.07	−0.04	0.00
	Random part	σ^2	.5	0.50 [0.48; 0.52]	0.01	0.00	0.00	0.62 [0.59; 0.64]	0.02	0.12	0.01
	(Intercept)	τ_0^2	.8	0.77 [0.66; 0.89]	0.06	−0.03	0.01	0.75 [0.64; 0.86]	0.06	−0.05	0.01
	(L-Trend)	τ_1^2	.8	0.78 [0.62; 0.94]	0.09	−0.02	0.00	0.52 [0.38; 0.66]	0.10	−0.28	0.08
		τ_{01}	0	−0.01 [−0.10; 0.07]	0.05	−0.01	0.00	−0.01 [−0.09; 0.06]	0.04	−0.01	0.00
M2	Fixed part	γ_{00}	0	−0.03 [−0.14; 0.07]	0.06	−0.03	0.00	−0.05 [−0.16; 0.06]	0.08	−0.05	0.01
	(L-Trend)	γ_{10}	.4	0.52 [0.22; 0.82]	0.19	0.12	0.04	0.85 [0.49; 1.19]	0.30	0.45	0.24
	(Q-Trend)	γ_{20}	.4	0.30 [0.03; 0.57]	0.17	−0.10	0.04	−0.13 [−0.46; 0.18]	0.28	−0.53	0.32
	Random part	σ^2	.5	0.49 [0.47; 0.52]	0.01	−0.01	0.00	0.67 [0.64; 0.70]	0.02	0.17	0.03
	(Intercept)	τ_0^2	.8	0.82 [0.70; 0.94]	0.07	0.02	0.00	0.90 [0.77; 1.03]	0.08	0.10	0.01
	(L-Trend)	τ_1^2	.8	0.86 [0.48; 1.24]	0.24	0.06	0.08	0.49 [0.24; 0.73]	0.20	−0.31	0.12
	(Q-Trend)	τ_2^2	.8	0.71 [0.37; 1.06]	0.23	−0.09	0.09	0.39 [0.18; 0.62]	0.17	−0.41	0.21
		τ_{01}	0	0.00 [−0.10; 0.10]	0.05	0.00	0.00	−0.01 [−0.09; 0.06]	0.04	−0.01	0.00
		τ_{02}	0	0.01 [−0.08; 0.09]	0.05	0.01	0.00	−0.01 [−0.08; 0.05]	0.04	−0.01	0.00
		τ_{12}	0	0.00 [−0.09; 0.09]	0.04	0.00	0.00	0.00 [−0.05; 0.04]	0.02	0.00	0.00

For each replication, the MCMC algorithm was run for 10, 000 iterations to generate plausible values under the IRT model and the CTT model. The convergence of the MCMC chains was checked using an ANOVA test on three groups of 200 from a thinned chain, after a burn-in of 5000 iterations, to investigate mean differences between different parts of the chain. When the test indicated significant mean differences, a new data set was drawn and after estimation checked for convergence. The MCMC chains for the models with only a linear trend (M1) showed convergence. The number of iterations for the MCMC chains was increased for models with a quadratic trend (M2). In simulation condition 1 with the 10 repeated measurements to 100, 000 (burnin = 50, 000), and in condition 2 with 50 repeated measurements to 200, 000 (burnin = 50, 000). After increasing the number of iterations, the chains were inspected for convergence using the potential scale reduction factor.[36] All chains showed adequate convergence. Furthermore, a subset of the chains was additionally inspected for convergence by examining the plots of the sampled parameter values. In Figures 3 and 4, MCMC iterates for the parameters of Model M1 and M2 are plotted, respectively, for one of the 10 replications of simulation condition 1.

The MCMC chains did not show convergence issues. Subsequently, the plausible value procedure was used, where 20,000 iterations were used to estimate the LGM parameters for each vector of plausible values. The LGM estimates were computed for model M1 and M2 using the IRT-generated plausible values (equation (10)) and CTT-generated plausible values (equation (14)) that were rescaled to the (true) scale of the generated latent variables, which were used to simulate the item response data.

Table 2. Simulation study results for latent curve models M1 and M2 over 10 replications, using IRT and CTT-generated plausible values ($N = 500$, $J = 50$, $K = 20$).

				IRT				CTT			
		Par.	True	Mean	SD	BIAS	MSE	Mean	SD	BIAS	MSE
M1	Fixed part	γ_{00}	0	0.01 [−0.07; 0.09]	0.04	0.01	0.00	0.04 [−0.04; 0.12]	0.04	0.04	0.00
	(L-Trend)	γ_{10}	.4	0.38 [0.30; 0.47]	0.04	−0.02	0.00	0.32 [0.25; 0.40]	0.05	−0.08	0.01
	Random part	σ^2	.5	0.50 [0.49; 0.51]	0.01	0.00	0.00	0.62 [0.61; 0.63]	0.01	0.12	0.01
	(Intercept)	τ_0^2	.8	0.80 [0.70; 0.90]	0.05	0.00	0.00	0.81 [0.70; 0.91]	0.06	0.01	0.00
	(L-Trend)	τ_1^2	.8	0.77 [0.67; 0.89]	0.06	−0.03	0.00	0.61 [0.52; 0.70]	0.06	−0.19	0.04
		τ_{01}	0	0.00 [−0.08; 0.07]	0.04	0.00	0.00	−0.07 [−0.14; 0.00]	0.04	−0.07	0.00
M2	Fixed part	γ_{00}	0	−0.01 [−0.09; 0.07]	0.04	−0.01	0.00	0.00 [−0.09; 0.09]	0.05	0.00	0.00
	(L-Trend)	γ_{10}	.4	0.39 [0.25; 0.53]	0.08	−0.01	0.00	0.61 [0.45; 0.76]	0.12	0.21	0.05
	(Q-Trend)	γ_{20}	.4	0.44 [0.29; 0.58]	0.08	0.04	0.00	0.08 [−0.06; 0.22]	0.12	−0.32	0.10
	Random part	σ^2	.5	0.50 [0.49; 0.51]	0.01	0.00	0.00	0.65 [0.64; 0.67]	0.01	0.15	0.02
	(Intercept)	τ_0^2	.8	0.80 [0.70; 0.91]	0.05	0.00	0.00	0.90 [0.78; 1.02]	0.06	0.10	0.01
	(L-Trend)	τ_1^2	.8	0.74 [0.55; 0.94]	0.11	−0.06	0.01	0.77 [0.58; 0.96]	0.14	−0.03	0.02
	(Q-Trend)	τ_2^2	.8	0.86 [0.66; 1.06]	0.12	0.06	0.02	0.35 [0.20; 0.51]	0.13	−0.45	0.21
		τ_{01}	0	0.00 [−0.08; 0.08]	0.04	0.00	0.00	−0.12 [−0.21; −0.03]	0.05	−0.12	0.01
		τ_{02}	0	0.00 [−0.09; 0.09]	0.05	0.00	0.00	−0.06 [−0.12; 0.00]	0.04	−0.06	0.00
		τ_{12}	0	0.01 [−0.09; 0.10]	0.05	0.01	0.00	0.01 [−0.05; 0.07]	0.03	0.01	0.00

6.2 Results

Table 1 shows the results of the simulation study with 10 repeated measurements (condition 1) and Table 2 shows the results for 10 repeated measurements (condition 2). For the IRT model with 10 repeated measurements, the bias for the estimates of the fixed component of model M1 and M2 lies between −.10 and .12 and between −.09 and .06 for the estimates of the random components. For the 50 repeated measurement condition, the bias for the estimates lies between −.02 and .04 for the fixed components and between −.06 and .06 for the random components. The results based on IRT-generated plausible values show accurate results, where the bias and mean squared error (MSE) are around zero. The true parameters were correctly recovered using the MCMC algorithm and the plausible value procedure.

The estimation results based on the CTT-generated plausible values show more bias and higher MSE estimates. The bias for the CTT estimates in the first simulation condition lied between −.53 and .45 for the fixed parts and between −.41 and .17 for the random parts, which is much higher compared to the bias of the estimates based on IRT. The bias in CTT-based estimates in the second simulation condition lied between −.32 and .21 for the fixed parts and between −.45 and .15 for the random parts, which is lower than in the condition with less measurement occasions, however, still much higher compared to the bias in the IRT based estimates. The highest posterior density intervals (HPDs; the 95% shortest credible interval which contains 95% of the most likely values, see for

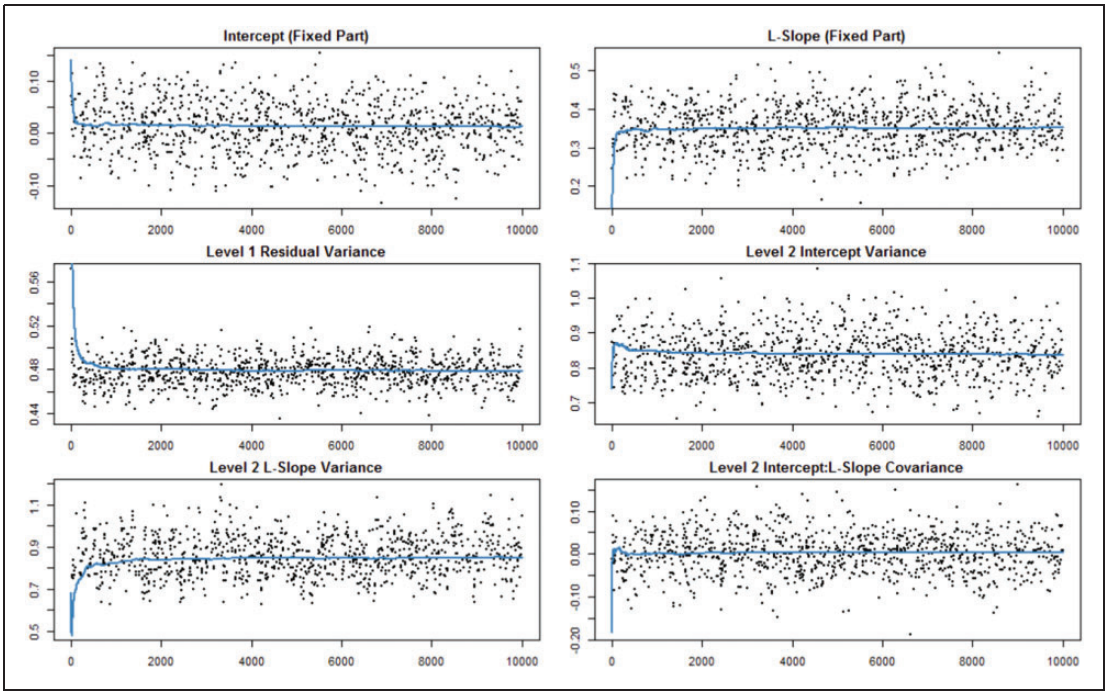

Figure 3. MCMC iterations of the fixed effect and variance parameters of LGM M1 with a linear time effect using IRT-based plausible values.

example Box and Tiao[37]) under the IRT model always contained the true value, whereas for the CTT model, the true values of γ_{02}, σ^2, and τ_1^2 were not located in the HPDs. Most important, when using the CTT model, the mean quadratic time effect was not detected in both conditions. The effect in the first condition (Table 1) was estimated to be $\hat{\gamma}_{20} = -.13(.28)$ $(95\%HPD = [-.46;\ .18])$, and in the second condition (Table 2) $\hat{\gamma}_{20} = .08(.12)$ $(95\%HPD = [-.06;\ .22])$. The true effect was .40 and this quadratic effect accelerated a positive trend of .40, which was not detected when using CTT. For models M1 and M2 in both simulation conditions, the identified bias in the estimates under CTT showed a typical pattern. The (level-1) residual variance σ^2 was overestimated and showed that there was more residual variation detected in latent variable estimates over time (i.e. a positive bias was found). The level-2 variances were underestimated and showed less variation in trends and quadratic effects over subjects (i.e. a negative bias was found).

7 Empirical data examples

7.1 Longitudinal cohort study on COPD

Data were obtained from a longitudinal cohort study on 409 patients with chronic obstructive pulmonary disease (COPD).[38] The questionnaires were administered at several occasions with a maximum of 11 measurements per patient. We examined the longitudinal development of the Chronic Respiratory Disease Questionnaire (CRQ)[39] sub-scale ''Emotion'', which was used to measure COPD complaints over time. The CRQ-emotion sub-scale consists of seven items on the emotional burden of COPD (e.g. ''In the last two weeks, how often did you feel down or discouraged?'', and ''In the last two weeks, how often did you feel embarrassed about your coughing?'', and ''In the last two weeks, how often did you feel restless, agitated, or tense?''). Items contained seven ordinal response categories ranging from ''1: Never'' to ''7: All the time''. One patient was excluded who had not completed the questionnaire on any of the measurement occasions. In order to be able to calculate sum scores, patient measurements with incomplete observations at a certain time point were not taken into account. Furthermore, we excluded 23 patients with only one measurement occasion from our analysis, since we were focused on change in COPD complaints. This led to a total sample size of 385 patients who provided 3236 observations. From this set of included patients, 163 were females and 222 were males. The IRT model that was used for generating the scores was the graded response model, see equation (2). The discrimination parameters were fixed to 1.

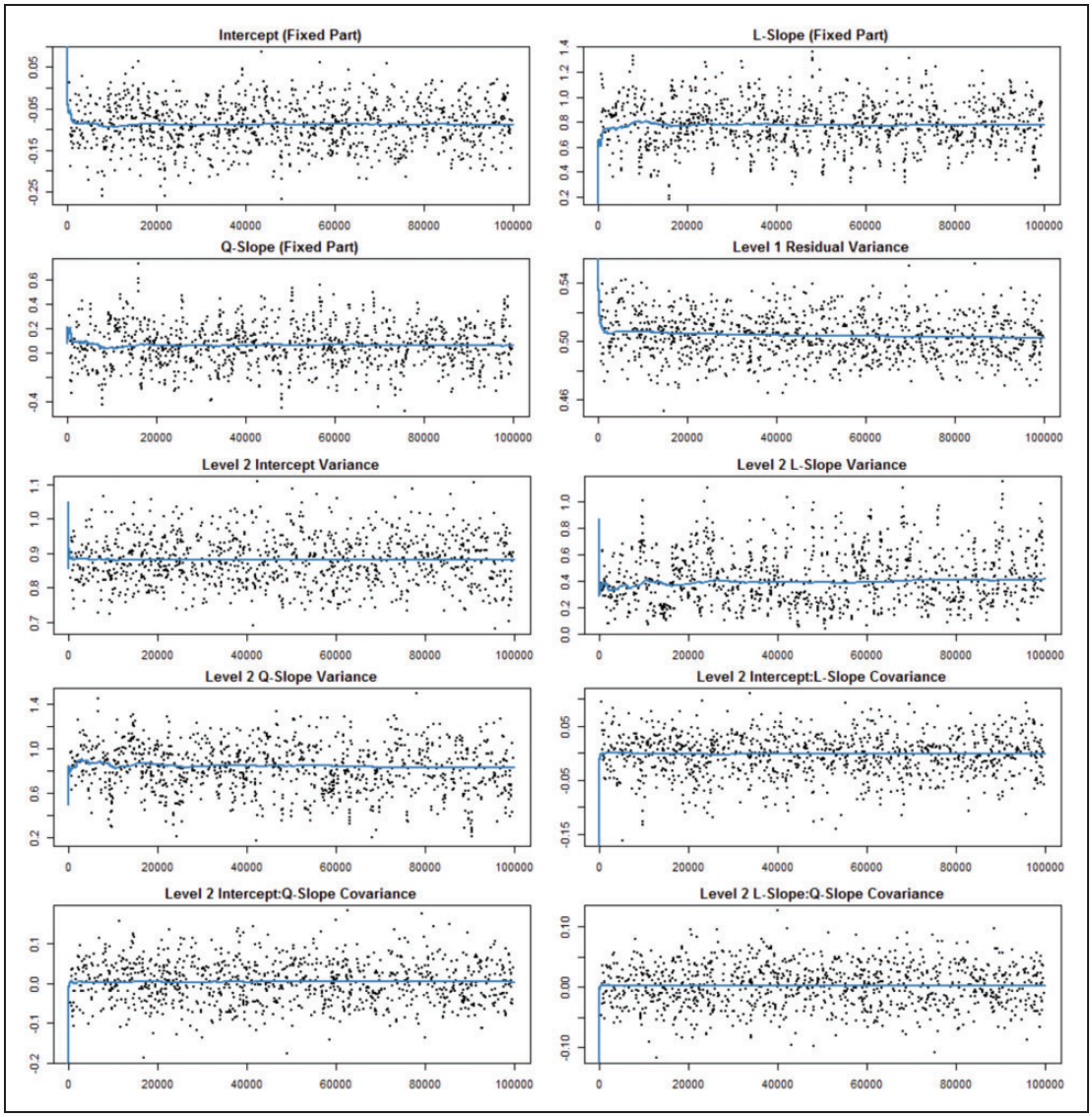

Figure 4. A total of 100,000 MCMC iterations of the fixed effect and variance parameters of LGM M2 with a quadratic time effect using IRT-based plausible values.

The model LGM M1 (equation (4)) was fitted to the data. The random intercept, β_{0i}, and random linear trend, β_{1i}, were assumed to follow a multivariate normal distribution, and were allowed to correlate. Time was defined in weeks from the first observed measurement and was divided by the maximum number of weeks from onset. As a result, the last measurement was taken at time is 1. The first measurement at time 0 represents the level of CRQ emotions at the time of the first measurement in the individual patients. The random intercept variance represents the between-subject variation in CRQ-emotion levels at time 0. A subject-specific linear trend was specified to model the change in CRQ emotions. LGM M2 (equation (6)) was also fitted to the data. In this model, a subject-specific quadratic time effect was added.

The proposed plausible value procedure was applied to examine differences between the use of IRT and CTT scores as outcomes in the LGMs M1 and M2. The plausible values were rescaled to a scale with a mean of zero and a standard deviation of one. Therefore, for each model the CTT- and IRT-based results were directly comparable. The parameter estimates of both LGMs with IRT and CTT-measured outcomes are given in Table 3.

The results for Model 1 in Table 3 show that the IRT level-1 variance estimate is smaller than the CTT level-1 variance estimate, since the corresponding HPDs do not overlap. At the start of the study, there was more variation in scores across patients under IRT (.80) than under CTT (.72). The group of patients showed to be more heterogeneous at time 0 under IRT than under CTT. Furthermore, the linear change in CRQ emotions was

Table 3. Results for LGM M1 and M2 fitted to the (COPD) CRQ-emotional data.

		IRT			CTT		
	Par.	Mean	SD	95% HPD	Mean	SD	95% HPD
Model 1							
Fixed part	γ_{00}	0.02	0.05	[−0.08;0.11]	0.00	0.05	[−0.09;0.09]
(Linear Trend)	γ_{10}	−0.22	0.06	[−0.33; −0.12]	−0.16	0.08	[−0.27; −0.06]
Random part	σ^2	0.20	0.01	[0.18;0.21]	0.26	0.02	[0.25;0.27]
(Intercept)	τ_0^2	0.80	0.06	[0.68;0.92]	0.72	0.07	[0.61;0.84]
(Linear Trend)	τ_1^2	0.51	0.08	[0.38;0.65]	0.41	0.11	[0.29;0.54]
	τ_{01}	−0.09	0.04	[−0.18; −0.01]	−0.06	0.04	[−0.13;0.01]
Model 2							
Fixed part	γ_{00}	0.03	0.05	[−0.08;0.13]	0.00	0.06	[−0.10;0.10]
(Linear Trend)	γ_{10}	−0.27	0.15	[−0.54; −0.02]	−0.19	0.19	[−0.49;0.10]
(Quadratic Trend)	γ_{20}	0.06	0.16	[−0.23;0.32]	0.05	0.19	[−0.26;0.35]
Random part	σ^2	0.20	0.01	[0.19;0.21]	0.26	0.01	[0.24;0.27]
(Intercept)	τ_0^2	0.79	0.06	[0.67;0.91]	0.72	0.06	[0.61;0.83]
(Linear Trend)	τ_1^2	0.37	0.10	[0.20;0.53]	0.30	0.16	[0.14;0.45]
(Quadratric Trend)	τ_2^2	0.20	0.08	[0.07;0.36]	0.22	0.09	[0.08;0.38]
	τ_{01}	−0.05	0.04	[−0.13;0.02]	−0.04	0.04	[−0.11;0.02]
	τ_{02}	−0.03	0.03	[−0.10;0.02]	−0.03	0.03	[−0.09;0.03]
	τ_{12}	0.00	0.02	[−0.04;0.04]	0.00	0.02	[−0.03;0.03]

Note: The estimated LGM parameters of both models are based on five re-scaled plausible values using IRT and using CTT to retrieve estimates on a common scale.

also more negative under IRT (−.22) than under CTT (−.16), indicating that the IRT-based measurements showed on average a steeper decline (i.e. more COPD related emotional problems over the full span of the study). The change in CRQ emotions appeared to be more variable under IRT (.51) than under CTT (.41). Finally, under IRT, slightly more negative covariance was found between the random intercept and the trend, which shows that those with high scores at the intake had on average a more negative decline in CRQ emotions.

When including a quadratic-time effect (Model 2), the estimated average trend was more negative under IRT (−.27) than under CTT (−.19), where the negative trend under CTT was no longer significantly different from 0. The estimated average quadratic-time effect showed a (non-significant) deceleration of the negative trend, which was under IRT around .06 and under CTT around .05. When interpreting the effects, the change in measured CRQ emotions was more negative under IRT than under CTT, but this decline decelerated almost equally under IRT and CTT. The estimated measurement error variance at level-1 was significantly higher under CTT (.26) than under IRT (.20) since the 95% HPD intervals do not overlap. This showed that the IRT-based patient-specific trajectories more accurately describe the change in CRQ emotions than the CTT-based patient-specific trajectories.

The variability in the estimated trajectories across patients under CTT showed less variation in deviations from the average intake score (.72) and from the average negative trend (.30) than under IRT (.79 and .37, respectively). The patient-specific deviations from the population-average trajectory were smaller under CTT than under IRT, and differences between trajectories were less apparent. The estimated variation in the quadratic effects across patients were almost equal under both measurement models. It was concluded that under IRT, the trajectories were better identified due to the enhanced differentiation between persons' CRQ-emotion scores and the reduction in measurement error variance, when comparing them to the estimated trajectories under CTT.

In Figure 5, the two top figures represent the population-average trajectory estimates under CTT and IRT and illustrate the differences in estimated trends for LGM M1 and M2. The time scale is represented in the real study time, starting at week 0 up till week 320. It is apparent that under IRT a more negative trend was estimated than under CTT. When including a quadratic trend component, the negative trend was decelerated in the same way under IRT and CTT.

In a second analysis, a distinction was made between the population-average trajectory of the males and females, by including an indicator variable (i.e. $Male = 0$ and $Female = 1$) in the random effect equations of LGM M1 and LGM M2. The variable Female was included in the random intercept to explain differences between males and females at baseline (time is 0). Furthermore, cross-level interactions between gender and time were investigated, by including the variable Female as a predictor of the random trend, referred to as

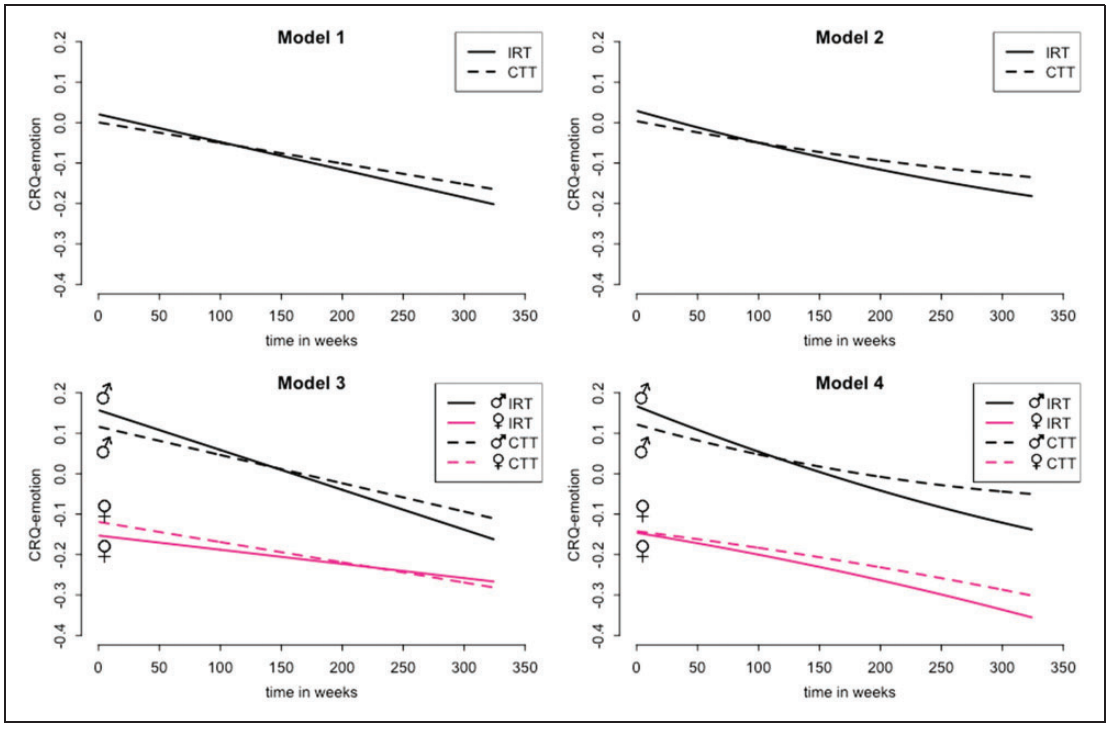

Figure 5. Population-average trajectory estimates of the CRQ-emotion using IRT and CTT for LGMs M1 and M2, and M3 and M4.

model M3, and also as predictor of the random quadratic-time effect, referred to as model M4. The LGM M4 is given by

$$\begin{aligned}\theta_{ij} &= \beta_{0i} + \beta_{1i}t_{ij} + \beta_{2i}t_{ij}^2 + e_{ij}\\ \beta_{0i} &= \gamma_{00} + \gamma_{01}Female_i + u_{0i}\\ \beta_{1i} &= \gamma_{10} + \gamma_{11}Female_i + u_{1i}\\ \beta_{2i} &= \gamma_{20} + \gamma_{21}Female_i + u_{2i}\end{aligned} \tag{17}$$

Differences between females and males in their intercept, trend, and quadratic effect of the trajectory of CRQ-emotion were explored under the CTT and IRT model. The results of LGM M3 and M4 are presented in Table 4.

When using IRT, females scored on average .31 points lower than males at the intake, with a standard deviation of .10. Under CTT, females scored around .24 points lower with a standard deviation of .09. It was concluded that the difference in scores at the intake between males and females was more apparent under IRT. From M3 follows that the estimated trajectory for the females, labeled Female-L-Trend, showed a less steep decline in CRQ-emotion than for the males. The less negative decline for the females was found under IRT (.21) and CTT (.06), where the average difference between the linear trends of the males and females was much more apparent under IRT. When considering the 95% HPD intervals, the estimated linear decline for the females was not significantly different from zero under IRT ($-.11 = -.32 + .21$) and CTT ($-.17 = -.23 + .06$). However, under IRT, with a 90% posterior probability a negative linear trend for the females was identified, $P(\gamma_{10} + \gamma_{11} < 0|\mathbf{Y}) \geq .90$, which was not detected under CTT. Under IRT, the variation in linear trends across patients was around .52, where a variation of around .43 was found under CTT.

From the results of model M4 followed that under IRT, the trajectory of the males showed a more (significantly different from 0) negative trend compared to the trajectory of the females under IRT (males $-.39$ and females $-.16 = -.39 + .23$). Under CTT, this trajectory difference between females and males was smaller (males $-.27$ and females $-.12 = -.27 + .15$), and these linear trend effects were not significantly different from 0 under CTT. Only under IRT, significant negative linear trends were found. It followed that under IRT a (non-significant) positive quadratic effect was estimated, showing a deceleration of the negative trend of around .09, where the females showed a smaller deceleration of the negative trend of $.07 = .09 - .02$.

Table 4. Estimates of LGM M3 and M4, which include gender-specific trend and quadratic trend effects using IRT and CTT-generated plausible values.

		IRT			CTT		
	Par.	Mean	SD	95% HPD	Mean	SD	95% HPD
Model 3							
Fixed part	γ_{00}	0.16	0.06	[0.03;0.28]	0.12	0.07	[−0.01;0.23]
(Linear Trend	γ_{10}	−0.32	0.07	[−0.46; −0.18]	−0.23	0.10	[−0.37; −0.09]
(Female)	γ_{01}	−0.31	0.10	[−0.49; −0.12]	−0.24	0.09	[−0.42; −0.06]
(Female-L-Trend)	γ_{11}	0.21	0.11	[−0.01;0.41]	0.06	0.13	[−0.15;0.27]
Random part	σ^2	0.20	0.01	[0.19;0.21]	0.26	0.01	[0.25;0.28]
(Intercept)	τ^2_{00}	0.78	0.06	[0.66;0.90]	0.70	0.06	[0.60;0.82]
(Linear Trend)	τ^2_{00}	0.52	0.09	[0.39;0.66]	0.43	0.08	[0.31;0.57]
	τ_{01}	−0.09	0.04	[−0.17; −0.01]	−0.06	0.04	[−0.14;0.01]
Model 4							
Fixed part	γ_{00}	0.17	0.08	[0.03;0.30]	0.12	0.07	[−0.01;0.25]
(Linear Trend)	γ_{10}	−0.39	0.23	[−0.77; −0.06]	−0.27	0.25	[−0.63;0.07]
(Quadratic Trend)	γ_{20}	0.09	0.24	[−0.29;0.46]	0.10	0.25	[−0.26;0.46]
(Female)	γ_{01}	−0.31	0.11	[−0.52; −0.11]	−0.26	0.12	[−0.48; −0.07]
(Female-L-Trend)	γ_{11}	0.23	0.32	[−0.31;0.77]	0.15	0.43	[−0.43;0.72]
(Female-Q-Trend)	γ_{21}	−0.02	0.33	[−0.60;0.53]	−0.14	0.51	[−0.74;0.45]
Random part	σ^2	0.20	0.01	[0.19;0.21]	0.27	0.02	[0.25;0.28]
(Intercept)	τ^2_0	0.78	0.06	[0.66;0.90]	0.71	0.06	[0.60;0.82]
(Linear Trend)	τ^2_1	0.34	0.09	[0.17;0.51]	0.22	0.09	[0.09;0.36]
(Quadratic trend)	τ^2_2	0.21	0.09	[0.07;0.37]	0.22	0.08	[0.08;0.38]
	τ_{01}	−0.05	0.04	[−0.13;0.02]	−0.02	0.03	[−0.08;0.03]
	τ_{02}	−0.03	0.03	[−0.10;0.02]	−0.02	0.03	[−0.08;0.03]
	τ_{12}	0.00	0.02	[−0.04;0.04]	0.00	0.02	[−0.03;0.03]

In Figure 5, the two bottom figures represent the population-average trajectories of the males and females under CTT and IRT for the LGM M3 and M4. The lines, marked with a symbol, represent the population-average latent curve for the females and the lines, marked with a symbol, represent the latent curves for the males. The estimated quadratic trend effects were not significant, and therefore attention is focused on differences between the trajectories under LGM M1 and M3. The average trajectories of M1 showed a less conservative linear trend under IRT than under CTT. When including differences between males and females in LGM M3, under IRT it became more apparent, compared to CTT, that the males showed a negative linear trend in CRQ-emotions, where the females showed a less negative trend. Under IRT, the males scored on average much higher at intake than the females, but due to differences in the average negative trends, score differences between males and females were much smaller at the end of the study. Under CTT, the differences across time were less apparent between males and females, since the trends of the males and females were more alike than under IRT. This makes the CTT-based latent curves (dashed lines) less pronounced.

The variation of the random effects in Model M4 shows that there was more variation estimated between patients' intake scores under IRT than under CTT. Also, the variation in patient-specific linear trends was higher under IRT than under CTT. We concluded that the estimated individual trajectories differed much more using IRT than CTT. It was shown that using the CTT scores (based on aggregated scores ignoring differences in response patterns) leads to less individual variation and a more conservative population-average trajectory compared to using all response information (IRT scores).

7.2 RCT on coping with low back pain

The second example concerns an RCT on the effectiveness of two different treatments for non-specific low-back pain.[40,41] The original study included 299 patients, who were randomly assigned to one of two intervention conditions, referred to as intervention A and B, or to the control condition. All patients were measured for four measurement occasions and the sub-scale ''passive coping'' of the Pain Coping Inventory (PCI)[42] was used as outcome variable. The sub-scale contains 21 items with four ordered response categories, which reflects three

cognitive-behavioral strategies, assessing behavioral tendencies to restrict functioning (resting, five items), to avoid environmental stimuli (retreating, seven items) and catastrophic cognitions about the pain (worrying, nine items).[42] A total of three patients reported incomplete questionnaires on the four measurement occasions, and were excluded from the analysis. A total of 42 patients who only filled out the questionnaire on one of the occasions were also excluded. This led to a total of 254 patients who were eligible for latent growth modeling of measured PCI-passive using IRT (equation (2)) and CTT.

Intervention A was given to 88 patients, intervention B to 81 patients and the control condition to 85 patients. A time variable was defined, where time was coded in months from the first wave and divided by 12 months. Time 0 represented the baseline measurement of PCS, and time 1 represented the last wave (at 12 months after baseline).

A linear and quadratic LGM model was fitted, according to LGM M2, which is represented in equation (6). The same typical pattern was found as in the cohort study on COPD, see Table 5 for the results.

The population-average trajectory had a negative linear change, which was decelerated by a positive quadratic trend effect. Furthermore, when comparing the IRT-based population trajectory to the CTT-based population trajectory, it can be seen that the trend in PCI-passive was more negative and the deceleration was more positive and significant under IRT. The CTT-based population trajectory showed a less dynamic change, where the measurement at baseline was lower and the change was more flat over the study period. The estimated variability across patient-specific trajectories in the linear and quadratic trends (1.26 and .65, respectively) was much higher than under CTT (.35 and .32, respectively). The estimated level-1 variance and variation in baseline measurements was higher under CTT than under IRT. This showed that the baseline measurements were more alike, and during the intervention, the trajectories of PCI differed more under IRT than under CTT. Again, under IRT more individualized trajectories were estimated, where changes in PCI were more apparent across patients and time. A strong negative correlation ($-.64$ and $-.23$ for IRT and CTT, respectively) was estimated between the linear and quadratic trend, indicating that large (small) negative trends were decelerated by large (small) positive quadratic trends.

Under IRT, more variance in the outcomes was explained by differences in individual trends over time and less variance was attributed to measurement error variance and to time-invariant score differences across individuals. The estimated individual variation in linear and quadratic trends was much larger under IRT than CTT, while the measurement error variance and the random intercept variance were lower. We concluded that under IRT more of the variance was explained by the random linear and quadratic trend components than under CTT.

In the next step of the latent growth modeling, baseline and cross-level intervention effects were added to the model. The intervention variable was re-coded using the effect coding scheme given in Table 6, where the control

Table 5. Results for LGM M2 for IRT and CTT, given the 21 item PCI data administered on four occasions.

		IRT			CTT		
		Mean	SD	HPD	Mean	SD	HPD
Fixed part	γ_{00}	0.52	0.05	[0.42;0.60]	0.39	0.07	[0.28;0.50]
(Linear Trend)	γ_{10}	−1.76	0.18	[−2.09; −1.45]	−0.98	0.37	[−1.38; −0.58]
(Quadratic Trend)	γ_{20}	0.83	0.16	[0.53;1.12]	0.16	0.34	[−0.22;0.53]
Random part	σ^2	0.18	0.02	[0.15;0.21]	0.37	0.04	[0.32;0.42]
(Intercept)	τ^2_{00}	0.35	0.05	[0.26;0.44]	0.52	0.08	[0.38;0.66]
(Linear Trend)	τ^2_{10}	1.26	0.63	[0.38;2.32]	0.35	0.30	[0.05;0.92]
(Quadratic Trend)	τ^2_{20}	0.65	0.45	[0.11;1.45]	0.32	0.25	[0.05;0.79]
	τ_{01}	0.26	0.11	[0.05;0.45]	−0.05	0.15	[−0.35;0.20]
	τ_{02}	−0.14	0.10	[−0.32;0.04]	0.01	0.15	[−0.23;0.28]
	τ_{12}	−0.64	0.51	[−1.72; −0.06]	−0.23	0.27	[−0.96;0.00]

Table 6. Effect coding schema used to estimate the effects for the treatment conditions.

	Effect A	Effect B
control group	−1	−1
treatment A	1	0
treatment B	0	1

(reference) group was assigned the value −1 and the parameter estimates for both the treatment A and treatment B group were used to estimate the intervention effects of the control group. In Model M5, in equation (18), the baseline and interaction effects are added to detect modifications of the linear and quadratic trend due to the intervention.

$$\begin{aligned}\theta_{ij} &= \beta_{0i} + \beta_{1i}t_{ij} + \beta_{2i}t_{ij}^2 + e_{ij}\\ \beta_{0i} &= \gamma_{00} + \gamma_{01}EffectA + \gamma_{02}EffectB + u_{0i}\\ \beta_{1i} &= \gamma_{10} + \gamma_{11}EffectA + \gamma_{12}EffectB + u_{1i}\\ \beta_{2i} &= \gamma_{20} + \gamma_{21}EffectA + \gamma_{22}EffectB + u_{2i}\end{aligned} \tag{18}$$

The LGM M5 in equation (18) contains the effects for treatment A and treatment B, and random linear and quadratic trend components. The parameter estimates for the control condition were constructed using the MCMC chains of the parameters for Effect A and Effect B, since the sum of the effects were equal to zero. The intercept can be interpreted as a constant value representing the grand mean across all measurements. The effects for the treatment and the control groups represent the overall difference between the grand mean and the group effect. The parameter estimates for both the IRT and CTT model of LGM M5 are represented in Table 7.

The intervention started after the baseline measurement, and there were also no significant differences in scores across the treatment groups measured at the intake. That is, the main effects of the treatments were all not significantly different from 0 under IRT and CTT. This verified the random assignment of individuals to treatment groups, and avoided any pre-existing group differences before the start of the study.

When considering the linear trend, it can be seen that the average linear trend is negative and around −1.79 under IRT and around −.71 under CTT. For intervention group A, the decline was even (significantly) steeper and around $-1.09 = -1.79 - .70$, under IRT. This additional decline for intervention group A was not detected under CTT. Under IRT, a less steep decline was measured for the other treatment groups, around $-1.40 = -1.79 + .39$ for treatment group B, and around $-1.49 = -1.79 + .30$ for the control group, which were not significant with 95% posterior probability. It was concluded that for none of the intervention groups the average linear trend was modified under CTT, where a significant modification of the average linear trend was detected for the treatment

Table 7. Results for Model M5 on the 21 item PCI-passive questionnaire administered on four occasions.

		IRT			CTT		
		Mean	SD	HPD	EAP	SD	HPD
Fixed part	γ_{00}	0.52	0.05	[0.43;0.60]	0.36	0.07	[0.24;0.47]
(Effect A)	γ_{01}	0.06	0.07	[−0.06;0.19]	−0.05	0.09	[−0.21;0.10]
(Effect B)	γ_{02}	0.02	0.07	[−0.11;0.14]	0.07	0.11	[−0.09;0.23]
(Control)		−0.13	0.13	[−0.38;0.11]	0.11	0.16	[−0.21;0.43]
(Linear Trend)	γ_{10}	−1.79	0.17	[−2.1;−1.47]	−0.71	0.28	[−1.11;−0.33]
(Effect A-L-Trend)	γ_{11}	−0.70	0.25	[−1.16;−0.26]	−0.03	0.31	[−0.58;0.52]
(Effect B-L-Trend)	γ_{12}	0.39	0.25	[−0.07;0.84]	−0.01	0.53	[−0.57;0.54]
(Control-L-Trend)		0.30	0.23	[−0.15;0.74]	0.04	0.28	[−0.51;0.58]
(Quadratic Trend)	γ_{20}	0.86	0.16	[0.56;1.14]	−0.08	0.24	[−0.45;0.29]
(Effect A-Q-Trend)	γ_{21}	0.55	0.23	[0.13;0.95]	0.03	0.31	[−0.50;0.54]
(Effect B-Q-Trend)	γ_{22}	−0.36	0.22	[−0.77;0.06]	0.02	0.51	[−0.52;0.53]
(Control-Q-Trend)		−0.19	0.21	[−0.60;0.22]	−0.04	0.26	[−0.56;0.47]
Random part	σ^2	0.18	0.02	[0.15;0.21]	0.33	0.05	[0.29;0.38]
(Intercept)	τ_{00}^2	0.33	0.05	[0.25;0.42]	0.54	0.08	[0.40;0.68]
(Linear Trend)	τ_{10}^2	1.30	0.55	[0.46;2.28]	0.43	0.35	[0.06;1.12]
(Quadratic Trend)	τ_{20}^2	0.64	0.36	[0.14;1.33]	0.37	0.31	[0.06;0.95]
	τ_{01}	0.30	0.12	[0.11;0.48]	0.01	0.16	[−0.29;0.27]
	τ_{02}	−0.18	0.10	[−0.35;-0.01]	−0.05	0.14	[−0.30;0.21]
	τ_{12}	−0.64	0.43	[−1.58;-0.10]	−0.27	0.32	[−1.16;0.00]

Note: Effect coding was used to estimate the baseline treatment effects and the cross-level interaction treatment effects.

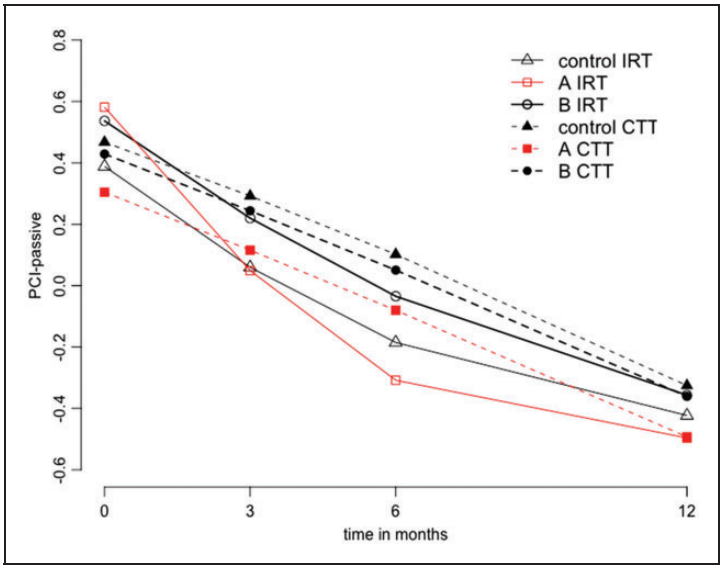

Figure 6. Population-average trajectories of the PCI-passive measurement according to LGM M5 given the parameter estimates presented in Table 7.

group A under IRT. Thus, under IRT a significant cross-level interaction was detected for intervention group A with the time component, which was not found under CTT.

When considering the quadratic trend, a more extreme difference in patterns was found between IRT and CTT. Under IRT, a population-average quadratic trend was estimated of around .86, which decelerated the linear trend. A significant modification of the average quadratic trend was found for the different treatment groups, where intervention group A showed a strong positive (significantly different from 0) deceleration of the linear trend of around .55, and intervention group B and the control group an acceleration of the linear trend of around −.36 and −.19, respectively. The cross-level interaction effect of intervention group B was only identified with 90% posterior probability, and the effect for the control group was not significantly different from 0. Under CTT, the estimated effect of the average quadratic trend component was around −.08 and not significantly different from 0. The estimated cross-level interaction effects did not show a significant modification of the average quadratic trend for any of the treatment groups. It was concluded that under CTT a population-average quadratic trend effect was not found. Finally, when considering the individual variation in trajectories, it can be seen that under IRT a random individual variation in linear and quadratic trends of around 1.30 and .64 was measured, which were smaller and around .43 and .37 under CTT, respectively. Furthermore, under IRT a negative covariance was found between the random intercept and the quadratic trend, and the linear trend and quadratic trend. Under IRT, those with a high score at baseline were more likely to show a more negative linear trend, and those with a more negative linear trend were more likely to show a higher deceleration of this trend. These patterns were not identified under CTT, where only a linear trend was identified which did not correlate with the random intercept.

The predicted population-average trajectories for each treatment group under IRT and CTT are given in Figure 6. It can be seen that under CTT the trajectories of the treatment groups run parallel (i.e. no cross-level interaction between the intervention and the trend) and they show a common small acceleration of the decline at the third measurement occasion. The trajectory patterns are different under IRT, where the intervention groups score on average higher at the intake and have different negative trends. Furthermore, intervention group A shows a positive deceleration of the negative trend at the third measurement occasion, where the intervention group B shows a non-significant acceleration of the negative trend. It is clear that a cross-level interaction between the intervention and the trend is only apparent for the IRT-based trajectories and not for the CTT-based trajectories.

8 Discussion

In the present paper, we compared estimation results using IRT scores and CTT scores when estimating LGMs in several situations. It was shown that IRT performs much better than CTT in all situations. The simulation study showed that the LGM using CTT-based health measurements led to a systematic bias in the growth parameter

estimates, which was not detected using IRT-based measurements. The major difference between IRT and CTT is that CTT is based on an aggregate score, which leads to loss of information, since differences in response patterns leading to the same sum score are ignored. IRT recognizes differences in response patterns. When the items differ in their level of difficulty, each unique response pattern leads to a unique person score. It is shown that the aggregate score under CTT for health measurements does not utilize all response information.

In a simulation study, a comparison was made between the bias in parameter estimates for the non-linear growth model combined with an IRT and a CTT measurement model. The parameter estimates based on IRT modeling are much closer to the true values than the CTT-based estimates over the linear and quadratic growth models. An explanation for this difference is that when the sum-scores (CTT) are used, much of the variance is discarded from the data. When using CTT, systematic bias was found in the estimated growth parameters and an increase in the residual variance in measurements across time. This increase in residual variation across individual measurements led to a reduction in the variation in individual linear trends and diminished the magnitude of the population-average trend.

Not only were the differences between IRT and CTT apparent in the simulation study, in both emperical examples (section 7) similar differences were found. The differences between the IRT- and CTT-based estimates in the empirical examples were comparable to the differences found in the simulation study. The parameters in the random part of the models show the same pattern of differences between the IRT and CTT-based LGM estimates. In the fixed part of the models with only a linear trend, an underestimation of the trend was found in the simulation as well as in both emperical examples. The LGM parameter estimates for the quadratic trend were underestimated and the linear trend were overestimated, when CTT-based scores were used.

These differences were found using a simultaneous estimation method for a hierarchical LGM model, where the measurement model (IRT or CTT) defines the level-1 part of the model and the LGM the higher-level part. The LGM can be viewed as a multilevel model, and subsequently, the posterior mean of the latent variable under IRT is a weighted average of the response information (including the item difficulty parameters) and the LGM information. Under CTT this posterior mean is a weighted average of the sum score and the LGM information, which ignores response pattern differences leading to the same sum score. The mentioned differences between IRT and CTT will not be found in a cross-sectional study, where the data strictly follows the Rasch model[43] and conditional maximum likelihood estimation is used. In that case, the latent variable estimates (construct scores) do not differ for IRT and CTT, and the sum score is considered a sufficient statistic. Furthermore, the two-parameter normal ogive model (equation (1)), equivalent to the two-parameter logistic model, becomes the Rasch (one-parameter) model in the special case that the discrimination parameters are equal to one. As described in section 6.1, discrimination parameters for the simulation study were set to one not to disadvantage the CTT model. In IRT modeling however, we still used the information of the item difficulties. This information is not incorporated in the CTT model (Hambleton et al., 1993; pp.151–172).[2] However, the focus of the presented research was on comparing the influence of the commonly used CTT model with the IRT model. The main objective of the comparison was to investigate the effects of the measurement-model choice on the LGM parameter estimates and not on the IRT and CTT parameter estimates.

The current study is limited to complete response data (i.e. all subjects respond to all items). In practice, it is more common to have incomplete data. To make a comparison between LGM results given CTT- and IRT-based plausible values for incomplete response data, the plausible value technique can be extended with a missing data imputation method to obtain a complete data set. Missing data is more easier handled under IRT than CTT, since IRT can deal with incomplete response data in a more natural way. However, more research is needed to measure and compare effects of missing data on LGM estimates under IRT- and CTT-based plausible values.

Although it is more realistic to assume non-linearity for modeling health over time, a quadratic trend component in the individual trajectories was sometimes not identified under CTT. Furthermore, the estimated effects of predictor variables explaining differences in trajectories were biased, partly due to the increase in residual variation. Across all non-zero parameter values of model M1 and M2, under CTT, the parameters showed an average bias of 46% (condition 1) and 30% (condition 2) compared to an average bias of 8% (condition 1) and 4% (condition 2) under IRT. It can be concluded that IRT utilizes all response information in contrast to CTT, and this led to substantial differences in the latent growth analysis, when comparing it to the CTT-based analysis. Although bias was also expected for more complex models with higher-order polynomials, such as random cubic time effects, the simulation results showed convergence issues for several chains and these results were therefore not used in the comparison.

A diagonal (homogenoeus) covariance matrix was used for the errors in the LGM, assuming conditional independence between successive measurement errors on the same patient. The random effects specifying the

latent growth trajectory define the intra-patient correlation across measurements, and conditionally on these effects the measurement errors were assumed to be independent. In longitudinal research, this assumption may not hold when for instance the correlation between the measurements appear to be stronger between the first measurement occasions and decreases as the time interval increases. Future research could focus on comparing the use of IRT scores and sum scores in a LGM with a more advanced covariance error structure as an autoregressive or a toeplitz covariance pattern model.

A plausible value technique was proposed to obtain health scores and LGM estimates on a common scale, while using different (level-1) measurement models. This straightforward approach utilizes the possibility to rescale plausible values to a specific metric. In the current study, plausible values for health measurements were transformed to obtain estimates of latent growth parameters on a common scale, while using IRT and CTT to measure health. The method can be applied to other models, where health measurements are used as outcomes. This will enable researchers to investigate differences in statistical outcomes caused by using different measurement models. The technique of re-scaling latent variable measurements to obtain a common scale analysis was easily embedded within the plausible value technique. This procedure contributes to the method of plausible values and makes it useful to accommodate scale differences. More research is needed to investigate whether the rescaling of latent variable measurements can also be embedded in other estimation methods as for instance the EM algorithm, where the E-step to render predicted values for the latent variable(s) can be modified.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

R Gorter https://orcid.org/0000-0002-2451-9953
G Ter Riet https://orcid.org/0000-0002-2231-7637

References

1. Hambleton R, Jones R and Rogers J. Influence of item parameter estimation errors in test development. *J Educ Measure* 1993; **30**: 143–155.
2. Lord F, Novick M and Birnbaum A. *Statistical theories of mental test scores*. Boston, MA: Addison-Wesley Publishing Company, Inc., 1968.
3. Gorter R, Fox J-P and Twisk J. Why Item Response Theory should be used for longitudinal questionnaire data analysis in medical research. *BMC Med Res Methodol* 2015; **15**: 1–12.
4. Gorter R, Fox J-P, Apeldoorn A, et al. The influence of measurement model choice for randomized controlled trial results. *J Clin Epidemiol* 2016; **79**: 140–149.
5. Ossenkoppele R, Van Der Flier W, Verfaillie S, et al. Long-term effects of amyloid, hypometabolism, and atrophy on neuropsychological functions. *Neurology* 2014; **82**: 1768–1775.
6. Eekhout I, Reijnen A, Vermetten E, et al. Post-traumatic stress symptoms 5 years after military deployment to Afghanistan: an observational cohort study. *Lancet Psychiatry* 2016; **3**: 58–64.
7. Inge TH, Courcoulas AP, Jenkins TM, et al. Weight loss and health status 3 years after bariatric surgery in adolescents. *New Engl J Med* 2016; **374**: 113–123.
8. Bollen K and Curran P. *Latent curve models: a structural equation approach*. Hoboken, NJ: John Wiley & Sons, 2006.
9. Samejima F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Suppl* 1969; **34**: 100.
10. Fox J-P. Multilevel IRT modeling in practice with the package mlirt. *J Stat Software* 2007; **20**: 1–16.
11. Embretson S and Reise S. *Item response theory for psychologists*. Mawwah, NJ: L.Erbaum Associates, 2000.
12. Lord F. *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates Publishers, 1980.
13. Beekman ATF, Bremmer MA, Deeg DJH, et al. Anxiety disorders in later life: a report from the longitudinal aging study Amsterdam. *Int J Geriatric Psychiatr* 1998; **13**: 717–726.

14. Reijnen A, Rademaker AR, Vermetten E, et al. Prevalence of mental health symptoms in Dutch military personnel returning from deployment to Afghanistan: A 2-year longitudinal analysis. *Eur Psychiatr* 2015; **30**: 341–346.
15. Fox J-P and Glas C. Bayesian modeling of measurement error in predictor variables using item response theory. *Psychometrika* 2003; **68**: 169–191.
16. Meredith W and Tisak J. Latent curve analysis. *Psychometrika* 1990; **55**: 107–122.
17. Muthén B. Latent variable modeling of longitudinal and multilevel data. In: Jordan M, ed. *Learn Graph Models* MIT Press; 1997: 453–480.
18. Hedeker D. An introduction to growth modeling. In: Kaplan D (ed.) *Quantitative methodology for the social sciences*. Thousand Oaks, CA: Sage Publications, 2006.
19. Tutz G. *Regression for categorical data*. vol 34. Cambridge: Cambridge University Press, 2011.
20. McCulloch CE and Neuhaus JM. *Generalized linear mixed models*. Hoboken, NJ: Wiley Online Library; 2001.
21. Mislevy RJ. Randomization-based inference about latent variables from complex samples. *Psychometrika* 1991; **56**: 177–196.
22. Mislevy R, Johnson E and Muraki E. *Chapter 3: Scaling procedures in NAEP*. 1992; **17**: 131–154. DOI: 10.3102/10769986017002131.
23. Rubin D and Schenker N. Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *J Am Stat Assoc* 1986; **81**: 366–374.
24. Little R and Rubin D. On jointly estimating parameters and missing data by maximizing the complete-data likelihood. *Am Stat* 1983; **37**: 218.
25. Asparouhov T and Muthén B. Auxiliary variables predicting missing data. Technical appendix. Los Angeles: Muthèn & Muthèn, http://statmodel.com/download/AuxM2.pdf (2008: 1–4).
26. Glas C, Geerlings H, van de Laar M, et al. Analysis of longitudinal randomized clinical trials using item response models. *Contemporary Clin Trial* 2009; **30**: 158–170.
27. Marsman M, Maris G, Bechger T, et al. What can we learn from plausible values? *Psychometrika* 2016. DOI: 10.1007/s11336-016-9497-x).
28. Thomas N and Gan N. Generating multiple imputations for matrix sampling data analyzed with item response models. *J Educ Behav Stat* 1997; **22**: 425–445.
29. von Davier M, Gonzalez E and Mislevy R. What are plausible values and why are they useful? *IERI monograph series* 2009, pp. 9–36.
30. Wu M. The role of plausible values in large-scale surveys. *Studies Educ Evaluat* 2005; **31**: 114–128.
31. Fox J-P. *Bayesian item response modeling. Theory and applications*. New York, NY: Springer; 2010.
32. Lee P. *Bayesian statistics: an introduction*. 4th ed. New York, NY: Wiley, 2012.
33. Klein Entink R, Fox J-P and van den Hout A. A mixture model for the joint analysis of latent developmental trajectories and survival. *Stat Med* 2011; **30**: 2310–2325.
34. Song X and Lee S. *Basic and advanced Bayesian structural equation modeling*. 1st ed. Hoboken, NJ: Wiley & Sons, 2012.
35. Plummer M, Best N, Cowles K, et al. Coda: Convergence diagnosis and output analysis for MCMC. *R News* 2006; **6**: 7–11.
36. Brooks SP and Gelman A. General methods for monitoring convergence of iterative simulations. *J Computat Graph Stat* 1998; **7**: 434–455.
37. Box G and Tiao G. *Bayesian inference in statistical analysis*. Boston, MA: Addison-Wesley Publishing Company, 1992. DOI: 10.1002/9781118033197
38. Siebeling L, ter Riet G, van der Wal W, et al. ICE COLD ERIC–International collaborative effort on chronic obstructive lung disease: exacerbation risk index cohorts–study protocol for an international COPD cohort study. *BMC Pulmonary Med* 2009; **9**: 15.
39. Wijkstra P, Ten Vergert E, Van Altena R, et al. Reliability and validity of the Chronic Respiratory Disease Questionnaire (CRQ). *Thorax* 1994; **49**: 465–467.
40. Heymans MW, de Vet HCW, Bongers PM, et al. The Effectiveness of high-intensity versus low-intensity back schools in an occupational setting: a pragmatic randomized controlled trial. *Spine* 2006; **31**: 1075–1082.
41. Heymans MW, de Vet HCW, Knol DL, et al. Workers' beliefs and expectations affect return to work over 12 months. *J Occup Rehabil* 2006; **16**: 685–695.
42. Kraaimaat F and Evers A. Pain-coping strategies in chronic pain patients: psychometric characteristics of the pain-coping inventory (PCI). *Int J Behav Med* 2003; **10**: 343–363.
43. Rasch G. *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Nielson an Lydiche (for Danmarks Paedagogiske Institut), 1960, http://eric.ed.gov/?id=ED419814

Appendix 1

The steps of an MCMC algorithm is briefly described for binary response data, and $\mathbf{X}_{ij}$ contain the explanatory variables of the LGM. Three steps are described to discuss (step 1) the sampling of the IRT parameters given the data and LGM parameters, (step 2) the sampling of the true scores given the data and LGM parameters, and

(step 3) the sampling of the LGM parameters given the latent variables $\boldsymbol{\theta}$ (or ϑ). The sampling scheme for the MLIRT model of Fox, 2010 (pp.159–160)[31] can be consulted for specific details of each sampling step.

Step 1a (IRT): Sampling of augmented data $\mathbf{Z}$. Given $\theta_{ij}, \xi_k, y_{ijk}$ the Z_{ijk} is sampled from

$$\begin{aligned} Z_{ijk} &\sim N_-\left(a_k\theta_{ij} - b_k, 1\right) \text{ if } Y_{ijk} = 0 \\ Z_{ijk} &\sim N_+\left(a_k\theta_{ij} - b_k, 1\right) \text{ if } Y_{ijk} = 1 \end{aligned} \tag{19}$$

where $N_+(.,.)$ and $N_-(.,.)$ represents the normal distribution truncated at the left and right by zero, respectively.

Step 1b (IRT): Sampling $\boldsymbol{\theta}$. Given $z_{ijk}, \xi_k, \boldsymbol{\beta}, \sigma^2$, the θ_{ij} is sampled from a normal distribution

$$\theta_{ij} \sim N\left(\theta^*_{ij}, \left(\sum_k a_k^2 + \sigma^{-2}\right)^{-1}\right)$$

where $\theta^*_{ij} = \left(\sum_k a_k^2 + \sigma^{-2}\right)^{-1}\left(\sum_k a_k^2\hat{\theta}_{ij} + \sigma^{-2}\mathbf{X}_{ij}\boldsymbol{\beta}_i\right)$ and $\mathbf{X}_{ij}$ contains the explanatory variables of the LGM.

Step 1c (IRT). Sampling $\xi = (\mathbf{a}, \mathbf{b})$. Let $\mathbf{H} = (\boldsymbol{\theta}, -1)$, item parameters are sampled from a normal distribution

$$\xi_k \sim N\left(\xi^*_k, \left(\boldsymbol{\Sigma}_\xi^{-1} + \mathbf{H}^t\mathbf{H}\right)^{-1}\right)$$

where $\xi^*_k = \left(\boldsymbol{\Sigma}_\xi^{-1} + \mathbf{H}^t\mathbf{H}\right)^{-1}\left(\mathbf{H}^t\mathbf{Z}_k + \boldsymbol{\Sigma}_\xi^{-1}\boldsymbol{\mu}_\xi\right)$, $\boldsymbol{\Sigma}_\xi$ is a diagonal matrix with variance 10, and $\boldsymbol{\mu}_\xi = c(1, 0)$.

Step 2 (CTT). Sampling of true scores (CTT). Assume a normal distribution for the errors in the CTT model, $\bar{y}_{ij} \sim N\left(\vartheta_{ij}, \sigma_y^2\right)$, sample ϑ_{ij} given $\bar{y}_{ij}, \sigma_y^2, \boldsymbol{\beta}_i, \sigma^2$ from a normal distribution

$$\vartheta_{ij} \sim N\left(\vartheta^*_{ij}, \left(\sigma_y^{-2} + \sigma^{-2}\right)^{-1}\right)$$

where $\vartheta^*_{ij} = \left(\sigma_y^{-2} + \sigma^{-2}\right)^{-1}\left(\sigma_y^{-2}\bar{y}_{ij} + \sigma^{-2}\mathbf{X}_{ij}\boldsymbol{\beta}_i\right)$. Without assuming a distribution for the errors, or assuming $\sigma_y^2 = 0$

$$\vartheta_{ij}\left|\boldsymbol{\beta}_i, \sigma^2 \sim N\left(\mathbf{X}_{ij}\boldsymbol{\beta}_i, \sigma^2\right)\right.$$

Step 3a (LGM). Sampling of γ. Given $\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{T}$, sample γ from a normal distribution

$$\gamma \sim N\left(\left(\sum_i \mathbf{w}_i^t\mathbf{T}^{-1}\mathbf{w}_i\right)^{-1}\sum_i \mathbf{w}_i^t\mathbf{T}^{-1}\boldsymbol{\beta}_i, \left(\sum_i \mathbf{w}_i^t\mathbf{T}^{-1}\mathbf{w}_i\right)^{-1}\right)$$

where $\mathbf{w}_i$ represents the person-level variables.

Step 3b (LGM). Sampling of $\boldsymbol{\beta}$. Given $\boldsymbol{\theta}, \gamma, \mathbf{T}, \sigma^2$ sample $\boldsymbol{\beta}$ from a normal distribution

$$\boldsymbol{\beta}_i \sim N\left(\mu_\beta, \Sigma_\beta\right)$$

where $\Sigma_\beta = \left(\boldsymbol{\Sigma}_i^{-1} + \mathbf{T}^{-1}\right)$, $\boldsymbol{\Sigma}_i = \sigma^2\left(\mathbf{x}_i^t\mathbf{x}_i\right)^{-1}$, and $\mu_\beta = \Sigma_\beta^{-1}\left(\Sigma_i^{-1}\hat{\beta}_i + \mathbf{T}^{-1}\mathbf{w}_i\gamma\right)$ and $\hat{\beta}_i = \left(\mathbf{x}_i^t\mathbf{x}_i\right)^{-1}\mathbf{x}_i^t\boldsymbol{\theta}_i$.

Step 3c (LGM). Sampling of σ^2. Given $\boldsymbol{\theta}, \boldsymbol{\beta}$ sample σ^2 from an inverse-gamma distribution with shape parameter $g_1 + \sum_i n_i/2$ and scale parameter

$$\sum_i \left(\boldsymbol{\theta}_i - \mathbf{x}_i\boldsymbol{\beta}_i\right)^t\left(\boldsymbol{\theta}_i - \mathbf{x}_i\boldsymbol{\beta}_i\right)/2 + g2$$

Step 3d (LGM). Sampling of $\mathbf{T}$. Given $\boldsymbol{\beta}, \gamma$ sample $\mathbf{T}$ from an inverse-Wishart distribution with shape parameter $\nu_T + J$ and scale parameter

$$\sum_i \left(\boldsymbol{\beta}_i - \mathbf{w}_i\gamma\right)\left(\boldsymbol{\beta}_i - \mathbf{w}_i\gamma\right)^t + \mathbf{S}_T$$

where an inverse-Wishart prior is used with ν_T degrees of freedom and $\mathbf{S}_T$ the scale parameter.