R-package LNIRT for joint modeling of response accuracy and times

Jean-Paul Fox[1], Konrad Klotzke[1] and Ahmet Salih Simsek[2]

[1] Faculty of Behavioral, Management, and Social Sciences, University of Twente, Enschede, Netherlands
[2] Department of Measurement and Evaluation in Education, University of Kirsehir Ahi Evran, Kirsehir, Turkey

ABSTRACT

In computer-based testing it has become standard to collect response accuracy (RA) and response times (RTs) for each test item. IRT models are used to measure a latent variable (*e.g.*, ability, intelligence) using the RA observations. The information in the RTs can help to improve routine operations in (educational) testing, and provide information about speed of working. In modern applications, the joint models are needed to integrate RT information in a test analysis. The R-package LNIRT supports fitting joint models through a user-friendly setup which only requires specifying RA, RT data, and the total number of Gibbs sampling iterations. More detailed specifications of the analysis are optional. The main results can be reported through the summary functions, but output can also be analysed with Markov chain Monte Carlo (MCMC) output tools (*i.e.*, coda, mcmcse). The main functionality of the LNIRT package is illustrated with two real data applications.

INTRODUCTION

In computer (adaptive) testing next to response accuracy (RA) response times (RTs) can be automatically recorded. The information from RTs can be used to improve the assessment of the test-takers (*e.g.*, latent ability estimation), to improve the design of a test (*e.g.*, adaptive item selection), and to learn more about a test-taker's response process (*e.g.*, test cheating, item exposure). The traditional item response theory (IRT) models to measure ability and item characteristics (*e.g.*, one- and two-parameter IRT models) offer no options to include this important information from RTs. These IRT models were constructed for paper-and-pencil tests and assume the RA data (response-accuracy observations to test items) represents the complete-data information. Therefore, to include the RT data (response-time observations to test items) in the psychometric-measurement analysis the traditional IRT models need to be modified.

IRT models have been adjusted to include RT information through a new parameter (*e.g.*, a time parameter) (see, *e.g.*, *Roskam, 1997*; *Thissen, 1983*; *Verhelst, Verstraalen & Jansen, 1997*). In another approach, the RT data is modeled separately using an RT-specific measurement model (*Maris, 1993*, *Scheiblechner, 1979*). Both approaches are not popular, mainly because the analysis of RT and RA data requires a joint modeling approach, which

Corresponding author
Jean-Paul Fox, j.p.fox@utwente.nl

includes the modeling of (underlying) associations between RA and RT data. This motivation has lead to an hierarchical modeling approach in which the RTs and the RA are each conditionally independently modeled (*van der Linden, 2007*). According to IRT measurement principles, a latent variable, referred to as accuracy (*i.e.*, ability), is measured with the RA data. In the same way, it is assumed that a latent variable speed is measured with the RT data. Therefore, in the joint modeling approach of RA and RT data, the latent variables speed and accuracy (*i.e.*, ability) are assumed to be measured by the RT and RA data, respectively. Response observations are assumed to be conditionally independently distributed given the latent variables speed and accuracy. At a higher modeling level (*i.e.*, the level of test-takers), the association between speed and ability is modeled. A speed-accuracy trade-off can be defined, where the level of accuracy is expected to increase when the level of speed is decreased. The characteristics of each test item represent the properties of measuring accuracy (*e.g.*, item difficulty) and of measuring speed (*e.g.*, item time-intensity), which are expected to be related to each other.

The hierarchical model for speed and accuracy can be viewed as two-component joint model, where an IRT model is defined for the RA and an RT model for the RTs. The continuous nature of the RT data supports the use of a normal distribution to model the heterogeneity in RTs. However, RTs are restricted to be positive, since they have a lower bound at zero. Therefore, a log-transformation is applied on the RTs, and the log-RTs are assumed to be normally distributed. The fit of the log-normal distribution for RTs from psychometric tests have been examined in different studies and was considered to be good (*Thissen, 1983*, *Schnipke & Scrams, 1997*, *van der Linden, Scrams & Schnipke, 1999*). Furthermore, the log-normal model for RTs has a specification which closely resembles IRT models for continuous RA (see, *e.g.*, *Samejima, 1973*; *Shi & Lee, 1998*).

Summarized, the joint model for RTs and RA represents a two-component joint model in which the log-normal model describes the RTs and a two-parameter IRT (*i.e.*, normal ogive two-parameter IRT model) describes the RA. The latent variables speed and accuracy are assumed to be measured by the RT and RA data, respectively. Item parameters represent the properties of an item for measuring accuracy and for measuring speed. This joint model is referred to as the log-normal item response theory (LNIRT) model and has become a popular model for analysing jointly RT and RA data.

The R-package LNIRT (*Fox, Klotzke & Klein Entink, 2019*) supports this (Bayesian) joint modeling of RA and RTs and comprehends several Gibbs samplers for parameter estimation. The R-package is available on CRAN: https://cran.r-project.org/web/packages/LNIRT/index.html. The R-package LNIRT is the successor of the cirt package of *Fox, Klein Entink & van der Linden (2007)*. The cirt program was implemented in Visual Pro FORTRAN, which led to problems in maintaining and updating the software. The program LNIRT has been developed in R, which makes it open source and makes the maintenance of the program much easier. The Gibbs sampler for the hierarchical joint model implemented in the cirt package has been integrated in the LNIRT package, and several important extensions have been added to LNIRT (see points 2–5 below). The

LNIRT package offers several important contributions for jointly analysing item patterns of RA and RTs:

1. Statistical computations are done through a powerful Gibbs sampling method that allows for fast MCMC convergence and efficient MCMC sampling.
2. Extensive residual analysis tools are available for the evaluation of item- and person fit, and outlier detection (*Fox & Marianti, 2017*).
3. Different parameterizations of the joint model can be applied. For instance, the log-normal RT model can be fitted with the parameterization of *van der Linden (2007)* or of *Klein Entink, Fox & van der Linden (2009)*.
4. Explanatory variables can be included at the item and person level, and missing item response data can be treated as missing at random (MAR) or missing by design (*i.e.*, incomplete test design) (*van der Linden & Fox, 2016*).
5. It offers a generalized measurement model for RTs in which working speed can be modeled through a latent growth component to allow for differential working speed across items (*Fox & Marianti, 2016*).

The remainder of this article discusses the main contributions of the package. Two detailed data examples are presented which show how the package can be used. In Section 2 the joint model is described, which includes the different measurement model parameterizations, the inclusion of explanatory variables at the item and person level, and (hyper) prior distributions. In "Model Fit", statistical tools are discussed to evaluate the fit of the model. This includes person-fit and item-fit tools to flag extreme persons and items under the model, respectively. "Applications" describes an illustration of the package by analysing the Credential data set collected from 1,636 candidates who took the licensure exam (*Cizek & Wollack, 2016*). Furthermore, the joint model with variable working speed is discussed using the Amsterdam chess data (*Van der Maas & Wagenmakers, 2005*). Then, in "Summary and Discussion", the conclusions are given.

THE JOINT MODEL

The LNIRT (joint) model is represented by the description of the measurement models for speed and accuracy at level 1. At a second (hierarchical) level, the distribution is described for the latent variables speed and accuracy, which operate as level-1 parameters to describe the RA and RTs. Furthermore, the distribution for the item parameters, which also operate as level-1 parameters, is also described at level 2. The joint model for RA and RTs was introduced by *van der Linden (2007)* and generalized by among others *van der Linden & Fox (2016)* and *Fox & Marianti (2016)*. The current description of the joint model follows the presentation given by *Fox, Klein Entink & van der Linden (2007)*, who described the (Bayesian) joint model with multivariate normal priors for the person and item parameters.

Level 1: measurement models

RA model

The matrix of RA-observations $\mathbf{Y}$ contains the responses to $k = 1, \ldots, K$ items of the $i = 1, \ldots, N$ test takers. Each RA is described by item characteristics and a test-taker's accuracy. In the two-parameter IRT model, the probability of a correct response is defined given the accuracy of a test taker i, θ_i, and item parameters (see, *e.g.*, *Lord & Novick, 1968*). Accordingly, the probability of a correct response to item k is defined as:

$$P(Y_{ik} = 1 \mid \theta_i, a_k, b_k) = \Phi(a_k\theta_i - b_k), \tag{1}$$

where a_k and b_k are generally known as the discrimination parameter and difficulty parameter of item k, respectively, and Φ denotes the normal cumulative distribution function. To define the item difficulties on the same scale as the ability scale, additional brackets need to be placed in the mean component. Then, the probability of a correct response to item k is given by,

$$P(Y_{ik} = 1 \mid \theta_i, a_k, b_k) = \Phi\left(a_k\left(\theta_i - \tilde{b}_k\right)\right), \tag{2}$$

where θ_i and $\tilde{b}_k$ are defined on the same scale. In LNIRT, both parameterizations are implemented. The item difficulty parameters in Eqs. (1) and (2) are not directly comparable, and are defined on different scales. For the three-parameter model, a guessing parameter c_k is introduced, representing the probability of guessing item k correctly, and this leads to the following measurement model for the success probability:

$$P(Y_{ik} = 1 \mid \theta_i, a_k, b_k, c_k) = c_k + (1 - c_k)\Phi(a_k\theta_i - b_k). \tag{3}$$

RT model

The matrix of RT-observations $\mathbf{RT}$ contains the log-RTs of the N test takers to the K items. When assuming a constant working speed, each test taker works with a constant speed represented by ζ_i. The time needed to complete an item also depends on item characteristic parameters. They are denoted as ϕ_k and λ_k, and can be seen as a time-discrimination and time-intensity parameter, respectively. The log-RTs, RT_{ik}, are assumed to be normally distributed given the speed and item parameters,

$$\begin{aligned} RT_{ik} &= \lambda_k - \phi_k\zeta_i + \epsilon_{ik} \\ \epsilon_{ik} &\sim N\left(0, \sigma^2_{\epsilon_k}\right). \end{aligned} \tag{4}$$

The time-intensity parameter λ_k represents the average time needed to complete the item (on a logarithmic scale). The speed parameter, ζ_i, represents the working speed of test taker i, and the time-discrimination parameter, ϕ_k, the item-specific effect of working speed on the RT. When increasing the time-intensity, λ_k, the average time needed to complete the item will increase. When increasing speed ζ_i the average time needed of person i to complete the item will decrease. In the same way as for the IRT model in

Eq. (2), the time intensities can be defined on the same scale as the speed parameter. By including extra brackets in the mean term

$$RT_{ik} = \phi_k\left(\tilde{\lambda}_k - \zeta_i\right) + \epsilon_{ik}, \tag{5}$$

the time-discrimination parameter operates on the term $\tilde{\lambda}_k - \zeta_i$. *Fox, Klein Entink & van der Linden (2007)* and *Klein Entink, Fox & van der Linden (2009)* introduced this time-discrimination parameter to improve the modelling of the association between speed and RTs. This item-specific discrimination effect of speed on the RT represents the item-specific change in RT, when increasing or decreasing speed. For a constant time-discrimination across items, a change in speed leads to item-invariant changes in RTs. For a non-constant item-specific time-discrimination, the change in speed leads to item-specific changes in RTs. The latter is more natural, since the effect of a change in speed is likely to depend on an item-specific processes instead of a test-specific process.

Fox & Marianti (2016) showed that the time-discrimination parameters in Eq. (4) contributes to the modeling of the covariance structure of the RTs. Furthermore, the item-specific measurement error variance in Eq. (4) represents unexplained variance in RTs due to stochastic behavior of a test taker. When test takers operate with different speed values, take small pauses during the test, or change their time management, the RTs might show more systematic variation than explained by the structural mean term. The item-specific error component might accommodate for these differences and avoids bias in the parameter estimates.

The time-discrimination parameter defined in LNIRT differs from the one defined by *van der Linden (2007)*. In his approach, the item-specific measurement-error precision, defined as the reciprocal of the standard deviation of the measurement error, is referred to as the time discrimination parameter. This parameter operates on the *squared* difference between observed RT and speed (and time intensity), instead of operating directly on the difference between speed and time intensity. Thus, the parameter does not represent the average change in RT due to a change in speed, but represents an effect of a change in the unexplained heterogeneity between RTs and speed. This parameterization also has the consequence that a measurement-error variance is not included in the RT model.

It is assumed that a test taker operates with a constant speed throughout the test conditioning on the time intensities. Each test taker's speed level does not change whatever the test conditions. In Section 3.4, the differential working-speed model is discussed to model changes in working speed, for example, due to fatigue or the adoption of a new strategy during the test, which is also implemented in the R-package LNIRT.

Level 2 and 3: structural person and item parameter models
Persons
A bivariate normal (population) distribution is defined for the ability and speed parameter,

$$(\theta_i, \zeta_i) \sim \mathcal{N}_2(\mu_P, \Sigma_P)$$
$$\mu_P = (\mu_\theta, \mu_\zeta)$$
$$\Sigma_P = \begin{pmatrix} \sigma_\theta^2 & \rho \\ \rho & \sigma_\zeta^2 \end{pmatrix}.$$

(6)

The covariance between the person parameters is represented by ρ. The level-2 model for speed and ability can be considered to represent a population distribution for the test takers. The test takers are defined to be exchangeable, and the distribution represents the prior for the person parameters. Without an identification restriction(s) on the variance parameters, the hyperprior for the covariance matrix Σ_P is the inverse-Wishart distribution with degrees of freedom v_P and scale parameter V_P.

Items

A multivariate normal distribution is specified for the item parameters,

$$(a_k, b_k, \phi_k, \lambda_k) \sim \mathcal{N}_4(\mu_I, \Sigma_I),$$

$$\Sigma_I = \begin{pmatrix} \Sigma_{a,b} & \Sigma_{(a,b),(\phi,\lambda)} \\ \Sigma_{(\phi,\lambda),(a,b)} & \Sigma_{\phi,\lambda} \end{pmatrix}$$

$$= \begin{pmatrix} \sigma_a^2 & \sigma_{a,b} & \sigma_{a,\phi} & \sigma_{a,\lambda} \\ \sigma_{b,a} & \sigma_b^2 & \sigma_{b,\phi} & \sigma_{b,\lambda} \\ \sigma_{\phi,a} & \sigma_{\phi,b} & \sigma_\phi^2 & \sigma_{\phi,\lambda} \\ \sigma_{\lambda,a} & \sigma_{\lambda,b} & \sigma_{\lambda,\phi} & \sigma_\lambda^2 \end{pmatrix}.$$

(7)

Parameters a_k and ϕ_k are restricted to be positive. The covariance matrix for the item parameters allows for correlation between item parameters. The hyperprior for (μ_I, Σ_I) is a normal-inverse-Wishart distribution:

$$\Sigma_I \sim IW_{v_I}(V_I^{-1})$$

(8)

$$\mu_I \mid \Sigma_I \sim \mathcal{N}(\mu_0, \Sigma_I/\kappa),$$

(9)

where v_I and V_I are the degrees of freedom and scale matrix of the inverse Wishart distribution, μ_0 is the prior mean and κ the number of prior measurements, which is given a default value of one to specify a vaguely informative prior. A Beta prior is specified for the guessing parameter, c_k, where the default hyper-parameter values are 20 and 80. This leads to a prior proportion of guessing of 1/5 with a standard deviation of 0.04.

Explanatory variables

The multivariate models for persons and item parameters can be extended to include explanatory variables. Let $\mathbf{X}_\theta$ denote the predictors for the ability parameter and $\mathbf{X}_\zeta$ for the speed parameter. The mean component for the person parameters can be expressed as

$$\mu_\theta = \mathbf{X}_\theta \beta_\theta, \mu_\zeta = \mathbf{X}_\zeta \beta_\zeta.$$

For the mean component of the difficulty and time-intensity parameters a similar extension is defined,

$$\mu_b = \mathbf{X}_b \beta_b, \ \mu_\lambda = \mathbf{X}_\lambda \beta_\lambda.$$

Explanatory information can be included to explain differences between persons and item characteristics. Noninformative normal priors are defined for the regression parameters with a mean of zero and a large variance. In the LNIRT package, the option to include predictors for discrimination and time discrimination are not implemented, since the variance in parameter values is (very) small. For categorical predictor variables, dummy coding is required.

Parameter estimation and model identification

A summarized description of the estimation method and model identification rules is presented. A more detailed description of the parameter estimation can be found in *Fox, Klein Entink & van der Linden (2007)*, *Klein Entink, Fox & van der Linden (2009)* and *Fox (2010)*. IRT models are usually identified by fixing the mean and variance of the latent variable to zero and one, respectively. Typically, this can be done directly by restricting the prior mean μ_θ and variance σ_θ^2, or by putting restrictions on the item parameters. The joint model can be identified in the same way. However, for the identification rules in the package LNIRT restricting the variance of a person parameter has been avoided. When restricting the variance of a (random) person parameter, the covariance matrix in Eq. (6) is also restricted, and the inverse-Wishart distribution does not apply to a restricted covariance matrix. For this scenario, *Klein Entink, Fox & van der Linden (2009)* redefined the prior for the person parameters, where ρ becomes a regression parameter in the regression of working speed on ability with the working speed variance included in the error variance. However, this identification procedure would also complicate other modeling features (*e.g.*, model-fit tools, variable working speed).

The LNIRT package provides two options to identify the model. For both options, the variance of the latent scales are identified by restricting the product of discriminations and of the time discriminations to one, $\prod_k a_k = 1$ and $\prod_k \phi_k = 1$, respectively. For option one (referred to as (R-code) ident=1), the mean of the scales are identified by restricting the sum of the difficulty and of the time-intensity parameters to zero, $\sum_k b_k = 0$ and $\sum_k \lambda_k = 0$, respectively. For option two (referred to as (R-code) ident=2), the mean of the scales are identified by fixing the mean of the ability parameter to zero, $\mu_\theta = 0$, and of the speed parameter to zero, $\mu_\zeta = 0$.

Model parameters are estimated through Gibbs sampling from their joint posterior distribution. The procedure involves the division of all unknown parameters into blocks, with iterative sampling of the conditional posterior distributions of the parameters in each block given the preceding draws for the parameters in all other blocks (*Fox, 2010*). In various simulation studies and real data analysis, the Gibbs samplers in the LNIRT package showed good convergence properties and generally returned efficient MCMC samples (*e.g.*, *Fox & Marianti, 2016*, *2017*; *Klein Entink, Fox & van der Linden, 2009*).

Logistic *vs* probit model item parameter estimates

The LNIRT package uses the normal-ogive IRT model (Probit model) to model the probability of a correct/positive response. The item parameter estimates under the normal ogive (Probit) model can be transformed to a Logistic scale and vice versa. Therefore, the logistic scale factor is used (1.7) to transform parameters on a Logistic scale to those on a Probit scale. Let μ_L, σ_L and μ_P, σ_P denote the mean and variance of the latent scale under the Logistic and Probit model, respectively. Then, the item parameters a_k and b_k are invariant under both scales, when applying the logistic scale factor

$$\underbrace{a_k \frac{(\theta_i - \mu_P)}{\sigma_P} - b_k}_{Probit\,Model} = \underbrace{1.7 \left(a_k \frac{(\theta_i - \mu_L)}{\sigma_L} - b_k \right)}_{Logistic\,Model}. \tag{10}$$

Assume that data were generated under the Logistic IRT model, and item parameter estimates were obtained under the Probit model. The Probit model estimates can be transformed to the Logistic model estimates;

$$\frac{\hat{a}_k}{\sigma_P} = \frac{1.7 a_k}{\sigma_L}$$

and, it follows that

$$\frac{\left(\frac{\hat{a}_k}{\sigma_P} \right)}{1.7} = \frac{a_k}{\sigma_L}.$$

In the same way, the difficulty parameters under the Probit model can be transformed. When assuming that μ_P and μ_L are zero, the transformation is

$$\hat{b}_k = b_k$$
$$1.7 \hat{b}_k = b_k^*,$$

since $b_k^* = 1.7 b_k$ is the difficulty parameter under the Logistic model.

MODEL FIT

Tools for evaluating the fit of the joint model have been introduced by *Marianti et al. (2014)* and *Fox & Marianti (2017)*. The current description follows their approach and shows the main features of their methods as implemented in the R-package LNIRT.

Person fit

The general idea is (1) to define a person-fit statistic for an RA and RT pattern, (2) to define classification variables, as a function of the statistics, to represent a non-aberrant and an aberrant state, and (3) to compute the posterior probability of the aberrant state. A person-fit test for RA and RT patterns is discussed. For each test, a dichotomous classification variable is introduced, which states whether a pattern is considered extreme. Then, the person-fit test for the joint model is constructed from the dichotomous classification variables for RA and RT patterns.

Person-fit statistic for RA pattern

Drasgow, Levine & Williams (1985) proposed a standardized version of Levine-Rubin statistic, which is shown to have statistical power to detect aberrant response patterns (*Karabatsos, 2003*). The log-likelihood is used to evaluate the fit of an RA pattern. Following a two-parameter IRT model, the person-fit statistic, denoted as l_0, is given by,

$$l^y(\theta_i, \mathbf{a}, \mathbf{b}; \mathbf{y}_i) = -\ln p(\mathbf{y}_i \mid \theta_i, \mathbf{a}, \mathbf{b})$$
$$= -\sum_{k=1}^{K} \left[\mathbf{y}_{ik} \ln p(\mathbf{y}_{ik}) + (1 - \mathbf{y}_{ik}) \ln(1 - p(\mathbf{y}_{ik})) \right] \tag{11}$$

where $p(\mathbf{y}_{ik}) \equiv p(\mathbf{y}_{ik} = 1 \mid \theta_i, a_k, b_k)$. The person-fit statistic can be standardized and this standardized version, denoted as l_s^y, is approximately standard normally distributed. Under the 3PL model, a dichotomous classification variable S_{ik} is introduced that classifies a correct response to be either a correct random guess with probability c_i ($S_{ik} = 0$) or a correct response according to the two-parameter IRT model ($S_{ik} = 1$). Then, l_s^y is defined conditionally on $S_{ik} = 1$ to evaluate the extremeness of non-guessed responses in the RA pattern. The guessed responses are ignored in the evaluation of the extremeness of an RA pattern.

To quantify the extremeness of an RA pattern, the posterior probability is computed that the person-fit statistic is greater than a threshold C, which can be defined according to its standard normal distribution. Note that the logarithm-of-response-probabilities are negative, thus increasing values of the negative log-likelihood correspond to misfit. The statistic is integrated over the prior distributions of all parameters and can therefore be interpreted as a prior predictive test. The (marginal) posterior probability of the statistic being greater than the threshold C is expressed as;

$$P(l_s^y(\mathbf{y}_i) > C) = \int \cdots \int P(l_s^y(\theta_i, \mathbf{a}, \mathbf{b}; \mathbf{y}_i) > C) p(\mathbf{a}, \mathbf{b}) d\theta_i d\mathbf{a} d\mathbf{b}$$
$$= \int \cdots \int \varphi(l_s^y(\theta_i, \mathbf{a}, \mathbf{b}; \mathbf{y}_i) > C) p(\mathbf{a}, \mathbf{b}) d\theta_i d\mathbf{a} d\mathbf{b}, \tag{12}$$

where φ denotes the normal density function. This approach corresponds to the prior predictive approach to hypothesis testing advocated by *Box (1980)*, since the posterior probability of an extreme RA pattern is computed by integrating over the prior distributions of the model parameters.

By introducing a classification variable, the posterior probability of an aberrant RA pattern for a given significance level can be computed. Let F_i^y denote a random variable that takes the value one when an observed pattern $\mathbf{y}_i$ is marked as extreme and zero otherwise,

$$F_i^y = \begin{cases} 1 & \text{if} \quad P(l_s^y(\theta_i, \mathbf{a}, \mathbf{b}; \mathbf{y}_i) > C), \\ 0 & \text{if} \quad P(l_s^y(\theta_i, \mathbf{a}, \mathbf{b}; \mathbf{y}_i) \leq C). \end{cases} \tag{13}$$

The posterior probability of an extreme pattern (F_i^y equals one) is computed by integrating over the model parameters. Then, for a fixed critical level C, the posterior probability represents how likely it is that the RA pattern is extreme under the model.

Person-fit statistic for RT pattern

Analogously, the log-likelihood of the RT pattern of a test taker i is used to define a person-fit statistic to quantify the extremeness of the pattern. The person-fit statistic, based on the log-likelihood of an RT pattern, is represented by

$$l_i^t(\zeta_i, \boldsymbol{\lambda}, \phi, \sigma^2; \mathbf{rt}_i) = \sum_{k=1}^{K} (\mathbf{rt}_{ik} - (\lambda_k - \phi_k \zeta_i))^2 / \sigma_k^2. \tag{14}$$

The sum of standardized errors is an increasing function of the negative log-likelihood, and an unusually large person-fit value corresponds to a misfit. The $l_i^t(\zeta_i, \boldsymbol{\lambda}, \phi, \sigma^2; \mathbf{rt}_i)$ given the model parameters is chi-squared distributed with K degrees of freedom. For a threshold C, representing the boundary of a critical region, the posterior probability of an extreme RT pattern is expressed as

$$P(l_i^t(\zeta_i, \boldsymbol{\lambda}, \phi, \sigma^2; \mathbf{rt}_i) > C) = P(\chi_K^2 > C) = p_{l^t}. \tag{15}$$

A classification variable can be defined to quantify the posterior probability of an extreme RT pattern given a threshold value C. Let F_i^t denote the random variable which equals one when the RT pattern is flagged as extreme and zero otherwise;

$$F_i^t = \begin{cases} 1 & \text{if} \quad P(l_i^t(\zeta_i, \boldsymbol{\lambda}, \phi, \boldsymbol{\sigma}^2; \mathbf{rt}_i) > C), \\ 0 & \text{if} \quad P(l_i^t(\zeta_i, \boldsymbol{\lambda}, \phi, \boldsymbol{\sigma}^2; \mathbf{rt}_i) \leq C). \end{cases} \tag{16}$$

The posterior probability of an extreme RT pattern is computed for each pattern with MCMC.

Person-fit statistic for RA and RT pattern

An observed RT pattern $\mathbf{rt}_i$ is reported as extreme when the F_i^t equals one with a least 0.95 posterior probability. To identify a joint pattern of RA and RT to be extreme, another classification variable is defined. Let $F_i^{t,y}$ equal one, when both F_i^y and F_i^t are equal to one, and equal zero otherwise. This joint classification variable represents the situation that both patterns of a test taker are extreme or not. The classification variable for the joint pattern is defined as

$$F_i^{t,y} = \begin{cases} 1 & \text{if} \quad P(l_i^t(\zeta_i, \boldsymbol{\lambda}, \phi, \sigma^2; \mathbf{rt}_i) > C, l_s^y(\theta_i, \mathbf{a}, \mathbf{b}; \mathbf{y}_i) > C), \\ 0 & \text{if} \quad 1 - P(l_i^t(\zeta_i, \boldsymbol{\lambda}, \phi, \sigma^2; \mathbf{rt}_i) > C, l_s^y(\theta_i, \mathbf{a}, \mathbf{b}; \mathbf{y}_i) > C). \end{cases} \tag{17}$$

The posterior probabilities of the classification variables in Eqs. (13), (16), and (17) are Bayesian significance probabilities. They represent the posterior probability of an extreme person-fit statistic given the data. They are computed using MCMC taking into account dependencies in the joint model parameters.

Item fit

Without reproducing the equations for item-fit statistics, the person-fit tests for RA and RT patterns can be modified to examine the fit of RA and RT item patterns. Therefore, the log-likelihood of RA and of RTs of an item is considered to define an item-fit statistic for RA and RTs, respectively. The estimated posterior probability of each item-fit statistic represents the probability that the statistic is extreme under the model, which means that the pattern of responses to the item is extreme.

Residual analysis

Bayesian residual computation has been considered by *Albert & Chib (1993)*, *Johnson & Albert (2006)*, and *Fox (2010*, Chapter 5) to evaluate the fit of an IRT model. This approach is extended to the joint model, and latent residuals e_{ik} are defined, which represent the difference between a latent continuous RA and the mean component. The latent continuous RA is defined through a data augmentation method to facilitate a Gibbs sampling algorithm (*Fox, 2010*, Chapter 3 and 4). Conditional expectation of a latent residual is derived by integrating out the augmented latent variable to obtain a Rao-Blackwell estimator for the latent residuals. The estimated latent residuals are used to quantify the total percentage of outliers per item and per test taker.

For the RA, the conditional expected latent residual is given by

$$E(e_{ik} \mid a_k, b_k, \theta_i) = \begin{cases} \frac{-\varphi(b_k - a_k\theta_i)}{\Phi(b_k - a_k\theta_i)} & y_{ik} = 0 \\ \frac{\varphi(b_k - a_k\theta_i)}{\Phi(b_k - a_k\theta_i)} & y_{ik} = 1 \end{cases} \tag{18}$$

where φ and Φ are the normal density function and the cumulative distribution function, respectively. Subsequently, the posterior probability that a latent residual is greater than a threshold C is given by

$$P(|e_{ik}| > C \mid a_k, b_k, \theta_i) = \begin{cases} \frac{\Phi(C)}{1 - \Phi(a_k\theta_i - b_k)} & y_{ik} = 0 \\ \frac{\Phi(-C)}{\Phi(a_k\theta_i - b_k)} & y_{ik} = 1 \end{cases} \tag{19}$$

The residuals for the RTs, referred to as ϵ_{ik}, can be estimated directly as the difference between the RT_{ik} and the mean, $\epsilon_{ik} = (rt_{ik} - (\lambda_k - \phi_k\zeta_i))$. The extremeness of an RT residual is expressed as the posterior probability that the residual is greater than a threshold C. This posterior probability can be expressed as

$$P(|\epsilon_{ik}| > C \mid \zeta_i, \lambda_k, \phi_k, rt_{ik}) = \Phi\left(-C - \frac{\epsilon_{ik}}{\sigma_k}\right) + 1 - \Phi\left(C - \frac{\epsilon_{ik}}{\sigma_k}\right). \tag{20}$$

Distribution RT residuals

The distribution of the RT residuals is evaluated using the Kolmogorov-Smirnov (KS) test. The empirical distribution of the residuals is compared to the assumed normal distribution. The KS test is a goodness-of-fit test and the posterior probability is computed that RT residuals of an item are non-normally distributed. The empirical distribution is given by

$$F_N(\epsilon) = \frac{1}{N} \sum_{i=1}^{N} I(\epsilon_{ik} < \epsilon)(\epsilon_{ik}).$$

The implemented KS test represents the difference between the cumulative empirical distribution and the normal cumulative distribution function:

$$D_N = \sup_{\epsilon} |F_N(\epsilon) - \Phi(\epsilon)|. \tag{21}$$

The distribution of D_N is the Kolmogorov distribution, which is used to compute the probability that D_N is greater than a threshold C:

$$p_{KS} = P(D_N > C \mid \mathbf{rt}_k, \lambda_k, \phi_k, \boldsymbol{\zeta}). \tag{22}$$

The marginal posterior probability is computed using MCMC, and the estimated significance probability represents the probability that the RT residuals of item k are non-normally distributed.

Differential Working Speed

The generalization of the joint model to allow differential working speed has been introduced by *Fox & Marianti (2016)*. A summarized version of their procedure is given, which shows how the joint model can be modified to include differential working speed.

Differential working speed is defined as a change in working speed of a test taker during the test. This relaxes a basic assumption of the joint model to work with a constant speed throughout the test. The change in working speed is modeled using a latent growth model. A random intercept, a linear trend component, and a quadratic time component are considered to model the speed trajectory of each test taker. The random intercept, ζ_{i0}, represents the initial value of working speed. The linear trend component, ζ_{i1}, is used to model a linear change in speed. Test takers can start slowly (fast) to increase (decrease) their speed later on. The quadratic time component, ζ_{i2}, is used to decelerate or accelerate the linear trend. For instance, a positive linear trend in working speed can be decelerated by a negative quadratic component.

The log-normal differential speed model with a random trend and quadratic time variable is represented by

$$
\begin{aligned}
RT_{ik} &= \lambda_k - \phi_k\big(\zeta_{i0} + \zeta_{1i}X_{ik} + \zeta_{2i}X_{ik}^2\big) + \epsilon_{ik} \\
\zeta_{i0} &\sim N\big(0, \sigma_{\zeta_0}^2\big) \\
\zeta_{i1} &\sim N\big(0, \sigma_{\zeta_1}^2\big) \\
\zeta_{i2} &\sim N\big(0, \sigma_{\zeta_2}^2\big),
\end{aligned}
\tag{23}
$$

where the time variable X_{ik} represents the order in which the test items are solved. The beginning of the test is represented by the time-point $X_{i1} = 0$ for each test taker i. This does not mean that each test taker should start the test at the same time. It is simply a

reference point for the speed process of each test taker. The time scale is defined on an equidistant scale to reduce the MCMC computations. Let $\mathbf{X}_{(i)} = X_{(i1)}, X_{(i2)}, \ldots, X_{(iK)}$ represent the order in which items are solved by test taker i. Then, a convenient time scale is defined by $X_{ik} = (X_{(ik)} - 1)/K$—the times are defined on a scale from 0 to 1, with 1, the upper bound, representing an infinite number of items. The time scale addresses the order in which the items are made and assumes an equidistant property of the speed measurements.

The average of the time intensities defines the average time to complete the test, since the random effects have a population mean of zero. The random intercept represents the speed measured at the first item. The population-average speed trajectory is constant, and shows no change in speed. Test takers can work faster (slower) than this population-average level, which corresponds to a positive (negative) initial speed level. A negative (positive) growth rate shows a decrease (increase) in speed, which can be decelerated (accelerated) by the quadratic time component. The random component variances represent the variance in growth parameters of the working speed trajectories. The covariance between random speed effects is modeled through the time-discrimination parameters. When the time-discrimination parameters are all fixed to one, the full covariance matrix of the random speed components is freely estimated.

Ability and differential working speed

A multivariate normal distribution is assumed for the random component ability and speed to allow for relationships between ability and the different speed components. This multivariate model for the random person parameters, $(\theta_i, \boldsymbol{\zeta}_i) = (\theta_i, \zeta_{0i}, \zeta_{1i}, \zeta_{2i})$, is given by

$$\begin{pmatrix} \theta_i \\ \boldsymbol{\zeta}_i \end{pmatrix} = N_4 \left(\begin{pmatrix} \mu_\theta \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \sigma_\theta^2 & \Sigma_{\theta,\zeta} \\ \Sigma_{\zeta,\theta} & \Sigma_\zeta \end{pmatrix} \right). \tag{24}$$

The relationship between ability and the speed components is defined by the covariance components $\Sigma_{\theta,\zeta}$. The growth components defining the speed trajectory influence ability. When test takers do not vary speed, ability is only influenced by the random intercept speed. When test takers vary their speed, the trend and quadratic time component will also influence ability, which is a specific feature of this generalization of the constant speed model.

Whether a change in speed improves the accuracy of the responses depends on the application. The differential speed joint model can be used to examine different test strategies across test takers. For instance, it is possible to estimate the speed trajectories of test takers with different levels of ability. The speed trajectories of high-ability students can differ from the low-ability students. The speed trajectories of test takers may also differ across tests. The model can be used to explore effects of time limits on test takers' changes in working speed. Benefits of exploring heterogenous speed trajectories in relation to ability will depend on the application.

SOFTWARE

The main function of the package LNIRT is the *LNIRT* function to fit the joint model for RA and RTs. It checks the input, arranges the input for the MCMC algorithm, and constructs the data output. An object of class **LNIRT** is generated, and a summary function can be used to get a summarized view of the estimation results. In the most simple case, the user passes the data to the *LNIRT* function and uses the package's summary function to get an overview of the results. The complete functionality of the package can be accessed by making further input specifications.

Input

The following arguments are mandatory for the *LNIRT* function:

- RT: A matrix RT containing the *log-normally* transformed response times in wide format, (N) persons in rows and (K) items in columns.
- Y: A matrix Y containing the (binary) RA data in wide format, (N) persons in rows and (K) items in columns. The binary RA data is coded as zero (incorrect) and one (correct) (missing values as NA).
- data (optional if RT and Y are given): A list or a *simLNIRT* object (output object of the function *simLNIRT*) containing the RT and RA matrices and optionally predictors for the item and person parameters. If a *simLNIRT* object is provided, in the summary the simulated item and time parameters are shown alongside of the estimates. If the required variables cannot be found in the list, or if no data object is given, then the variables are taken from the environment from which *LNIRT* is called.
- XG: The integer number of MCMC iterations (the default is 1,000), this includes the burn-in period.

The remaining arguments are optional for the *LNIRT* function:

- burnin: The percentage of the total number of MCMC iterations (XG) which will serve as the burn-in period of the chains. The default is 10%.
- ident: Identification rule, (ident=1) restrict sum of difficulties and sum of time intensities to zero and (ident=2) restrict mean person parameters to zero. The default is (ident=2). The product of (time) discriminations is restricted to one.
- guess: A logical variable with the default FALSE, where TRUE (FALSE) represents (not) a guessing parameter in the IRT model.
- par1: A logical variable with the default FALSE, where TRUE represents the bracket notation for the mean term of the IRT component as in Eq. (2), and FALSE the non-bracket notation as in Eq. (1). In general, the MCMC performance is better for the non-bracket parameterization.
- XGresid: the number of MCMC iterations (default is 1,000) to be done before starting the residual computation.
- residual: A logical variable with the default FALSE. When TRUE, a complete residual analysis is done together with the estimation of the joint model parameters, as discussed

in "Model Fit". The residual computations are started after XGresid MCMC draws. Therefore, XG should be greater than XGresid, and a sufficient number of MCMC iterations should be made after XGresid MCMC iterations to obtain accurate residual estimates. Preferably, at least 5,000 MCMC iterations are made, when doing a residual analysis.

- td: A logical variable with the default TRUE. When TRUE, the time-discrimination parameter is estimated. When FALSE, the time discrimination is restricted to one.
- WL: A logical variable with the default FALSE. When TRUE, the time-discrimination parameter represents the inverse of the measurement error variance parameter according to the parameterization of *van der Linden (2007)*.
- alpha: An optional vector of length K of pre-defined item discrimination parameters.
- beta: An optional vector of length K of pre-defined item difficulty parameters.
- phi: An optional vector of length K of pre-defined time-discrimination parameters.
- lambda: An optional vector of length K of pre-defined time-intensity parameters.
- XPA: An optional matrix of predictor variables for the ability parameters, where the columns represent the predictor variables. Categorical predictor variables need to be dummy coded.
- XPT: An optional matrix of predictor variables for the speed parameters, where the columns represent the predictor variables. Categorical predictor variables need to be dummy coded.
- XIA: An optional matrix of predictor variables for the item difficulty parameters, where the columns represent the predictor variables. Categorical predictor variables need to be dummy coded.
- XIT: An optional matrix of predictor variables for the time-intensity parameters, where the columns represent the predictor variables. Categorical predictor variables need to be dummy coded.
- MBDY: An optional missing-by-design indicator matrix—of the same size as Y—for missing values (coded NA) in the Y matrix due to the test design (0 is missing by design, 1 is not missing by design). Multiple imputations are simulated for missing values (missing at random) in the Y matrix which are not assigned in MBDY as missing-by-design.
- MBDT: An optional missing-by-design indicator matrix—of the same size as RT—for missing values (coded NA) in the RT matrix due to the test design (0 is missing by design, 1 is not missing by design). Multiple imputations are simulated for missing values (missing at random) in the RT matrix which are not assigned in MBDT as missing-by-design.

Output

The *LNIRT* function creates an object of class **LNIRT**, which stores the MCMC output, posterior draws and posterior mean estimates. The output of the function *LNIRT* is described in an itemized way, without including a residual computation.

- data: If available a data object from the function *simLNIRT*.
- burnin: Percentage from XG representing the burn-in MCMC iterations.
- ident: Same as the input variable ident.
- guess: Same as the input variable guess.
- MAB: The MCMC sampled values for the item parameters (discrimination, difficulty, time discrimination, time intensity), object is an array of dimension XG (number of MCMC iterations) by K (number of items) by 4 (number of item parameters).
- MCMC.Samples: This object contains the sampled MCMC values from *LNIRT*, where the following objects are stored:

  – Cov.Person.Ability.Speed: Samples of covariance of ability and speed.
  – CovMat.Item: Array of dimension XG (MCMC iterations) by K (items) by 2 (time-discrimination and time-intensity parameters), it contains the sampled variance of time discrimination and time intensity as well as their sampled covariance.
  – Item.Dificulty: Samples of difficulty parameters (XG by K).
  – Item.Discrimination: Samples of discrimination parameters (XG by K).
  – Item.Guessing: Samples of guessing parameters (XG by K).
  – Mu.Item.Difficulty: Sampled values of mean item difficulty parameter.
  – Mu.Item.Discrimination: Sampled values of mean item discrimination parameter.
  – Mu.Person.Ability: Sampled values of mean ability parameter.
  – Mu.Person.Speed: Sampled values of mean speed parameter.
  – Mu.Time.Discrimination: Sampled values of mean time-discrimination parameter.
  – Mu.Time.Intensity: Sampled values of mean time-intensity parameter.
  – Person.Ability: Sampled values of person's ability parameter (XG by N).
  – Person.Speed: Sampled values of person's speed parameter (XG by N).
  – Sigma2: Sampled values of measurement error variance parameters (XG by K).
  – Time.Discrimination: Sampled time-discrimination parameters (XG by K).
  – Time.Intensity: Sampled time-intensity parameters (XG by K).
  – Var.Person.Ability: Sampled variance ability parameters.
  – Var.Person.Speed: Sampled variance speed parameters.

- Mguess: Sampled values for the guessing parameter (under the default Beta prior B (20,80)).
- MmuI: Sampled values of the mean discrimination, difficulty, time discrimination, and mean time intensity (in that order).
- MmuP: Sampled values of the mean ability and mean speed parameters.
- MSI: Sampled values of the covariance matrix of item parameters. Array of dimension XG by 4 by 4 (discrimination, difficulty, time discrimination, time intensity), see also CovMat.Item.
- Msigma2: Sampled values of measurement error variance parameters.

- MSP: Sampled values of the covariance matrix of person parameters (ability and speed), in an array of dimension XG by 2 by 2.
- Mtheta: Posterior mean estimates of ability and speed of dimension N by 2.
- MTSD: Posterior standard deviation of ability and speed of dimension N by 2.
- par1: Same as the input variable par1.
- Post.Means: Posterior mean estimates of the following parameters:

    – Cov.Person.Ability.Speed: Covariance ability and speed.
    – CovMat.Item: Covariance matrix item parameters.
    – Item.Difficulty: Item difficulty estimates.
    – Item.Discrimination: Item discrimination estimates.
    – Mu.Item.Difficulty: Mean item difficulty estimate.
    – Mu.Item.Discrimination: Mean item discrimination estimate.
    – Mu.Person.Ability: Mean ability estimate.
    – Mu.Person.Speed: Mean speed parameter.
    – Mu.Time.Discrimination: Mean time discrimination.
    – Mu.Time.Intensity: Mean time intensity.
    – Person.Ability: Ability parameters.
    – Person.Speed: Speed parameters.
    – Sigma2: Measurement error variance.
    – Time.Discrimination: Time-discrimination parameters.
    – Time.Intensity: Time-intensity parameters.
    – Var.Person.Ability: Variance ability parameter.
    – Var.Person.Speed: Variance speed parameter.

- RT: Logarithm of the RT data.
- td: Same as the input variable td.
- WL: Same as the input variable WL.
- XIA: Same as the input variable XIA.
- XIT: Same as the input variable XIT.
- XPA: Same as the input variable XPA.
- XPT: Same as the input variable XPT.
- Y: Same as the input variable Y.

When the residual computation is included (residual=TRUE), then the **LNIRT** object includes the following additional output variables:

- EAPCP1: Posterior probability that the RT pattern is flagged as aberrant according to Eq. (16), using the posterior probability that the person-fit statistic l^t is extreme as defined in Eq. (15) with a significance level of 0.05.

- EAPCP2: Posterior probability that the RA pattern is flagged as aberrant according to Eq. (13), using the posterior probability that the person-fit statistic, l_s^{y}, is significant (with a significance level of 0.05).
- EAPCP3: Posterior probability that both patterns (RA and RT) are flagged as aberrant according to Eq. (17).
- EAPKS: Posterior probability of an extreme KS-test result according to Eq. (22), which indicates that the RT residuals are not normally distributed.
- EAPKSA: A significance probability of an extreme KS-test that the latent residuals of RA items are not normally distributed. This significance test has no power.
- EAPresid: Posterior probability of an extreme (standardized) residual (RT data), which is greater than plus or minus two in absolute value.
- EAPresidA: Posterior probability of an extreme (standardized) latent residual (RA data), which is greater than plus or minus two in absolute value.
- IFl: The (negative) log-likelihood statistic for item fit under the IRT model (high values represent misfit).
- IFlp: The posterior significance probability of an extreme item fit under the IRT model.
- lZI: Item-fit statistic representing the posterior probability that the squared sum of item residuals are extreme under the model.
- EAPl0: The log-likelihood contribution of each RA observation, where a low value represents a misfit.
- PFl: The standardized (negative) log-likelihood contribution of each RA pattern, where a high value represents a misfit.
- PFlp: Posterior (significance) probability of observing a more extreme person-fit statistic for the RA pattern than the observed one.
- lZPA: Posterior significance probability of an extreme person-fit test for RA pattern based on latent residuals. This significance test has no power (under construction).
- lZPT: The (unstandardized) estimated person-fit statistic according to Eq. (14).
- lZP: Posterior (significance) probability of observing a more extreme person-fit statistic for the RT pattern than the observed one, according to Eq. (15).

MCMC output analysis

The posterior draws in the output of LNIRT are obtained through a Gibbs sampler. MCMC convergence tests are necessary to make sure that the MCMC chains converged before making statistical inferences. The coda package (*Plummer et al., 2006*) can be used to do an MCMC convergence analysis. The LNIRT function returns a single MCMC chain for each model parameter, which can be directly translated to an **MCMC** object with the function *as.mcmc*. The single-chain convergence diagnostics, for instance Geweke, and Heidelberg-Welch, can be used to examine if the chains did not converge. The multiple-chain convergence diagnostics, for instance Gelman and Rubin's convergence statistic, requires multiple calls to LNIRT to create multiple MCMC chains with different starting values—LNIRT uses random starting values when parameter values are not pre-specified.

Table 1 Credential data (Form 1): the test design.

	Item set			
		Pretest		
Group	Test (1–170)	K1 (171–180)	K2 (181–190)	K3 (191–200)
G1	X	X	–	–
G2	X	–	X	–
G3	X	–	–	X

With the summary function of the package LNIRT, posterior means and standard deviations are computed and reported. Although the LNIRT Gibbs samplers produce efficient MCMC samples with low autocorrelation, it is possible that for a specific dataset some of the chains have a medium to high autocorrelation. Then, the reported standard deviation estimates may underestimate the standard deviation, since autocorrelation in the chains are ignored. The coda and the mcmcse package can be used to compute the autocorrelation and the Monte Carlo standard error. Note that the autocorrelation has no effect on the posterior mean estimate. The effective sample size—the sample size of independently distributed values with the same variance as the autocorrelated MCMC sample—can also be computed to examine if the run was long enough to make accurate and reliable inferences. A reasonable rule of thumb is to have an effective sample size of 400. Then, the Monte Carlo standard error is less than 5% of the overall uncertainty of the posterior mean.

The default burn-in period is 10% of the total number of MCMC iterations. This burn-in period is used in the computation of the posterior estimates, which are also reported with the summary function. Extensive simulation studies showed that the burn-in period is usually below 100 MCMC iterations, but this could be higher for a specific data set. Furthermore, the MCMC properties are better for the RA model in Eq. (1) than for the RA model with the additional brackets given in Eq. (2).

APPLICATIONS

The credential (Form 1) data

The credential data set of *Cizek & Wollack (2016)* is analyzed to illustrate the functionality of the package LNIRT. The credential data set concerns 1,636 test takers who applied for licensure. Form 1 of the test was administered, which consisted of 170 licensure exam items, and 30 pretest items. A total of 10 background variables of the test takers was stored —this includes the country where the candidate received his/her educational training, the state in which the test taker applied for licensure, and the center where the candidate took the exam. The RA and RT data were also stored. The collected data followed from a year of testing using a computer-based program that tests continuously.

Each candidate completed one of the three pretest forms, each consisting of 10 items. In Table 1, the test design is given, which shows the incomplete test design of in total 200 items for three groups.

The RA and RT data were extracted from the data to be used in the LNIRT() function. The RA data consisted of 1,636 test takers (in rows) and 170 exam items (in columns).

```
R> Y<- as.matrix(data[c(which(colnames(data)=="iraw.1")
+:which(colnames(Cdata)=="iraw.170"))])
R> head(Y[,1:5],3)
```

```
        iraw.1  iraw.2  iraw.3  iraw.4  iraw.5
[1,]         1       1       1       0       1
[2,]         1       0       1       0       0
[3,]         0       0       0       0       1
```

The RT data also consisted of 1,636 test-takers and 170 exam items. The RTs were transformed to a logarithmic scale. A total of 105 RTs were equal to zero. Possibly the item was skipped and a default incorrect response was recorded, since the corresponding RA were all incorrect. The zero RTs were converted into NAs, since it was unknown why a zero RT was recorded. The minimum RT was 2 s, and the 100 highest RTs ranged from 6 to 12 min.

```
R> RT<-as.matrix(data[c(which(colnames(data)=="idur.1"):
+ which(colnames(data)=="idur.170"))])
R> RT[RT==0]<-NA
R> RT<-log(RT)
R> head(RT[,1:5],3)
```

```
         idur.1    idur.2    idur.3    idur.4    idur.5
[1,]   4.094345  3.555348  3.555348  3.465736  3.295837
[2,]   4.007333  4.060443  4.343805  3.850148  4.110874
[3,]   4.234107  3.761200  4.007333  3.178054  4.127134
```

In the first analysis, the (default) joint model was fitted to the data to explore the item parameter estimates and the item and person covariance matrix. The model was identified by restricting the population means of ability and speed to zero and by restricting the product of time discriminations and discriminations to one. A total of 5,000 MCMC iterations were computed, and the burn-in period was 500 (*i.e.*, 10% of the total number of MCMC iterations).

```
R> library(LNIRT)
R> out0 <- LNIRT(RT=RT, Y=Y, XG=5000,ident=2,burnin=10)
R> summary(out0)
```

MCMC convergence

Multiple MCMC chains were checked for convergence and run length by computing the Geweke and Heidelberger statistics, the effective sample sizes, and the MCMC standard errors. The convergence statistics did not show any problems of non-convergence for the examined chains. The lowest effective sample size was 427 for the discrimination parameter of item 155 (MCMC standard error of 0.007), and the highest 4,000 for the speed parameter of test taker 1 (MCMC standard error less than 0.000). The naive

```
> summary(out0)

Log-Normal RT-IRT Modeling, 2013, J.-P. Fox

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
MCMC iterations:          5000
Burn-in period:           10%
Identification:           mu_theta = mu_zeta = 0
                          prod(alpha) = prod(phi) = 1
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

Summary of results

       Item Discrimination parameter    Item Difficulty parameter
       item    EAP    SD                item    EAP    SD

       1       1.235  0.164             1       -1.287  0.047
       2       0.881  0.130             2       -0.863  0.036
       3       1.992  0.163             3       -1.076  0.042
       4       0.544  0.123             4       -0.513  0.033
       5       0.860  0.124             5       -0.391  0.032

       Time Discrimination    Time Intensity          Measurement Error Variance
       item    EAP    SD       item    EAP    SD       EAP    SD

       1       1.237  0.059    1       3.899  0.010    0.174  0.006
       2       0.678  0.080    2       4.310  0.015    0.361  0.012
       3       1.347  0.076    3       3.841  0.014    0.321  0.011
       4       0.471  0.080    4       4.062  0.014    0.345  0.012
       5       1.096  0.068    5       3.945  0.012    0.232  0.008
```

Figure 1 Screenshot of the LNIRT summary output: item parameter estimates of the first five items.

posterior standard deviation—when ignoring autocorrelation in the chain—was often just smaller than the times-series standard error estimate using the coda package. It was concluded that the burn-in period and the total length of the chains were sufficiently long to compute accurate parameter estimates.

Output analysis

The general summary function of LNIRT shows the item number (which corresponds to the column number of Y and of RT), and the posterior mean (labeled EAP) and standard deviation (labeled SD) estimates of the item and of the prior parameters of the items and persons. In Fig. 1, the estimation results for the first five item parameters are given.

In Fig. 2, the item and person prior parameter estimates are given. The mean of the items, μ_I, is labeled in the output as mu_a (mean discrimination), mu_b (mean difficulty), mu_phi (mean time discrimination) and mu_lam (mean time intensity). The posterior mean estimate of the covariance matrix of the items, Σ_I, is given under the label Sigma_I. The estimated mean item discrimination is around 1.19 with a variance of around 0.32. Around 10 items have an estimated item discrimination below 0.40. The item difficulty estimates ranges from −1.84 to.75, with a mean of −0.70 and a variance of 0.27. The estimated average time-discrimination and time-intensity is around 1.03 and 3.96, with a variance of 0.05 and 0.11, respectively. The time discrimination ranged from 0.47 to 1.42,

```
Mean and Covariance matrix Items (mu_a,mu_b,mu_phi,mu_lambda)

--- Population Mean Item ---
mu_a    SD      mu_b    SD      mu_phi  SD      mu_lam  SD
1.191   0.052   -0.697  0.038   1.025   0.016   3.958   0.020

--- Covariance matrix Items (a,b,phi,lambda)---
      SigmaI                    SD SigmaI                  SigmaI (Correlation)
 0.322 -0.128  0.060 -0.002     0.044  0.027  0.012  0.015     1.000 -0.433  0.488 -0.011
-0.128  0.272 -0.049  0.081     0.027  0.030  0.010  0.015    -0.433  1.000 -0.433  0.464
 0.060 -0.049  0.047 -0.029     0.012  0.010  0.006  0.006     0.488 -0.433  1.000 -0.400
-0.002  0.081 -0.029  0.112     0.015  0.015  0.006  0.012    -0.011  0.464 -0.400  1.000

Person Effects and Covariance matrix Persons (ability,speed)

--- Person Effects (Ability - Speed)---
      Ability Predictor Effects
      EAP     SD
X 1   0.000   0.007
      Speed Predictor Effects
      EAP     SD
X 2   0.000   0.004

SigmaP          SD SigmaP       SigmaP (Correlation)
0.074  0.018    0.003  0.001    1.000  0.395
0.018  0.028    0.001  0.001    0.395  1.000
```

Figure 2 Screenshot of the LNIRT summary output: prior parameter estimates of persons and items.

and it appears that the items show better discriminating performance in speed than in ability. However, the estimated working-speed variance is very small and around 0.03. The variance in RTs is mostly explained by the time-intensity parameters and hardly by differences in working speed across test takers. Thus, the time discriminations are somewhat higher than the item discriminations, but the time discriminations have a minor contribution in explaining variance in RTs.

The covariance matrix of the item parameters shows that the less difficult items are more discriminating (correlation of −0.43), and the less time-intensive items are also more time-discriminating (correlation of −0.40). Differences in ability and speed are better measured with less difficult and less time-intensive items. The more difficult items are also more time-intensive (correlation of 0.46). The time-discriminating items are positively correlated with the item discriminations (correlation of 0.49).

The covariance matrix of the person parameters (ability and speed), Σ_P, is given under the label Sigma_P. The variance in ability and speed across test takers are both small. There is a positive correlation between ability and speed of around 0.40, which states that more able test takers also work faster than the less able ones.

Explanatory variables

Dummy coded variables were created for the pretest groups, where effect coding is used such that the general intercept can be restricted to zero to identify the joint model. Differences in ability and speed were examined across pretest groups (stored in objects XA and XT for ability and speed), and the test-taker's total test time was used to explain the correlation between ability and speed (stored in object XA for ability). The observed total test times contained information on top of the speed values, since the latter were shrunk

towards the population-average test time depending on the variance of the speed parameters. The joint model parameters—with the explanatory variables for ability and speed—are estimated. The main estimation results are again computed and displayed using the (R-code) summary function (using defaults burnin=10 and ident=2):

```
R>  XFT$Pgroup[XFT$Pretest==6,1]  <-  -1
R>  XFT$Pgroup[XFT$Pretest==6,2]  <-  -1
R>  XFT$Pgroup[XFT$Pretest==7,1]  <-  1
R>  XFT$Pgroup[XFT$Pretest==8,2]  <-  1
R>  out1  <-  LNIRT(RT=RT,  Y=Y,  XG=5000,  XPA=XA,  XPT=XT)
R>  summary(out1)
```

The effect of the total test time on ability was significant and around −0.07. Those who finished earlier scored on average higher than those who finished the test later. When accounting for the total test-time differences in measuring ability, the correlation between ability and speed was around 0.1. So, the total test-time explained around 75% of the correlation. There were no significant speed differences between the pretest groups, and the variance of speed was also small and around 0.03. There were estimated differences in ability between the pretest groups: group G_1 scored on average −0.13 lower, group G_2 0.07 higher, and group G_3 0.05 higher than the general average, but they were not significant.

Planned missing by design

In the planned missing data design, simultaneous parameter estimation for all examinees is possible using (R-code) LNIRT() function. It requires defining an indicator matrix for the planned missing data. In the matrix MBDM, a zero is a designed missing and a one a designed observation. The planned missing data matrix can be defined separately for the RA and RT data, and arguments MBDY and MBDT represent the design matrix for the RA and RT data, respectively. For instance, it is possible to include RA data in the analysis for which RT data was only partly collected. When including the pretest items in the measurement of ability and speed, the test design contains planned missing data. Then, the indicator matrix MBDM for the planned missing data is defined and included in the call to LNIRT:

```
R>  MBDM<-matrix(rep(0,1636*200),nrow=1636,ncol=200)
R>  MBDM[XFT$Pretest==6,171:180]<-1
R>  MBDM[XFT$Pretest==7,181:190]<-1
R>  MBDM[XFT$Pretest==8,191:200]<-1
R>  MBDM[,1:170]<-1
R>  outmbdm  <-  LNIRT(RT=RTt,Y=Yt,XG=5000,alpha=alpha1,
MBDY=MBDM,MBDT=MBDM)
```

In the call to LNIRT, the arguments RTt and Yt contain the RT and RA data for all 200 items. Pre-specified item discriminations, alpha1, are used, since the model with free item discrimination parameters would not fit. The output is not discussed for reasons of brevity.

Model fit

The joint model-fit tools are illustrated by re-running the analysis with explanatory variables for the 170 items and activating the residual computation (R-code residual=TRUE).

```
R> out1 <- LNIRT(RT=RT,Y=Y,XG=5000,XPA=XA,XPT=XT,
residual=TRUE)
R> summary(out1)
```

The summary report contains an overview of the extreme residuals under the header *Residual Analysis*. A total of 0.07% of RTs residuals were considered extreme with at least 95% posterior probability (Eq. (20)). This concerns RTs that were small and around 2–6 s or much higher than the item-average RTs. For the RA residuals, around 0.02% was estimated to be extreme with 95% posterior probability (Eq. (19)). This concerns incorrect responses from test takers with an above-average ability. For 62 items (36.5%) the assumption of log-normally distributed residuals was violated for which a significant probability of the KS test was computed (Eq. (22)). The variance in working speed and time intensities are small, and the estimated residual variance is around 0.26. Therefore, RT outliers more easily affect the fit of the log-normal distribution. The item-fit statistics (RA and RT data) did not identify a significant misfit of an item. Despite the outliers, log-likelihoods of item patterns were not significant under the joint model.

The EAPCP1 is reported and around 18.34%, representing the percentage of RT patterns which are considered extreme with 95% posterior probability. The percentage of significant extreme RT patterns is around 19.5%, when using a significance level of 0.05. The reported EAPCP2 is around 1.59% and the significant RA patterns is around 1.65%, which shows that there are only a few RA patterns identified as extreme. Finally, around 0.31% of the joint patterns (RA and RT) are extreme (object EAPCP3).

The heterogeneity in person-fit statistics for RA an RT patterns is examined with a linear regression of the statistics on the number of test attempts, country, and pretest group. The country variable was (dummy) recoded (US (Cgroup1=1),Philippines (Cgroup2=1),India (Cgroup3=1),Others (intercept)). The pretest groups were already represented by two dummy coded variables.

```
R> summary(lm(out1$PFl ~ as.factor(XFT$Attempt)+(XFT
$Cgroup)+(XFT$Pgroup)))
Coefficients:
```

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-0.204245	0.045153	-4.523	6.53e-06	***
as.factor(XFT$Attempt)2	-0.004323	0.060106	-0.072	0.94267	
as.factor(XFT$Attempt)3	-0.102401	0.079525	-1.288	0.19805	
as.factor(XFT$Attempt)4	-0.253421	0.114568	-2.212	0.02711	*
as.factor(XFT$Attempt)4+	0.017670	0.080984	0.218	0.82731	
XFT$Cgroup1	0.051825	0.046087	1.124	0.26097	
XFT$Cgroup2	-0.215907	0.054834	-3.937	8.58e-05	***

```
XFT$Cgroup3              -0.109458    0.053455   -2.048   0.04075 *
XFT$Pgroup1               0.011785    0.034349    0.343   0.73157
XFT$Pgroup2               0.095390    0.029290    3.257   0.00115 **
```

It follows that the RA patterns of test takers with more test attempts are less extreme than those with a fewer attempts. Test takers from the Philippines and India are less likely to have an aberrant RA pattern. Furthermore, test takers from the third pretest group (variable Pgroup2) have higher person-fit statistic scores. They correspond to RA patterns that are less likely to be observed under the model than RA patterns with lower person-fit statistics. For the person-fit test for RA patterns, for a significance level of 0.05 the critical value is 1.645. The average-statistic scores (number of attempts, country, pretest group) are much lower than the critical value, and test takers with an aberrant RA pattern are also outliers in their groups.

The heterogeneity in person-fit statistic scores for the RT patterns is also explored through a linear regression with the same explanatory variables.

```
R>  summary(lm(out1$lZPT ~ as.factor(XFT$Attempt)+(XFT$Cgroup)
+(XFT$Pgroup)))
Coefficients:
                          Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)               169.9231      3.0855   55.071   < 2e-16 ***
as.factor(XFT$Attempt)2     1.1506      4.1073    0.280   0.7794
as.factor(XFT$Attempt)3    10.2410      5.4343    1.885   0.0597.
as.factor(XFT$Attempt)4     2.6948      7.8290    0.344   0.7307
as.factor(XFT$Attempt)4     2.3246      5.5341    0.420   0.6745
XFT$Cgroup1                -7.7563      3.1494   -2.463   0.0139 *
XFT$Cgroup2                 3.0138      3.7471    0.804   0.4213
XFT$Cgroup3                15.4076      3.6528    4.218   2.6e-05 ***
XFT$Pgroup1                 2.5152      2.3473    1.072   0.2841
XFT$Pgroup2                 0.2681      2.0016    0.134   0.8935
```

For the l_i^t statistic, the critical value is 201.4 when the significance level is 0.05 (l_i^t is chi-square distributed with 170 degrees of freedom). The intercept corresponds to test takers from pretest group 1 from citizens outside the US, India, and the Philippines, who took the test for the first time. For test takers with more attempts the person-fit statistic is on average higher and closer to the critical value. Those test takers are more likely to give a very fast or much slower response. Test takers from the US have on average much lower person-fit scores, and test takers from India have a much higher person-fit score.

In Fig. 3, the person-fit scores for RA patterns are plotted against those of RT patterns to provide a more comprehensive overview of the aberrant and non-aberrant patterns per country. For this example, the R-code is given below

```
set1  <-  which(XFT$Country=="USA")
set2  <-  which(XFT$Country=="India")
set3  <-  which(XFT$Country=="Philippines")
plot(out1$PFl,out1$lZPT,xlab="Person-fit  Statistic  RA",
  ylab="Person-fit Statistic RT",col="black",cex=.5,bty="l",xlim=c(-3,3),
```
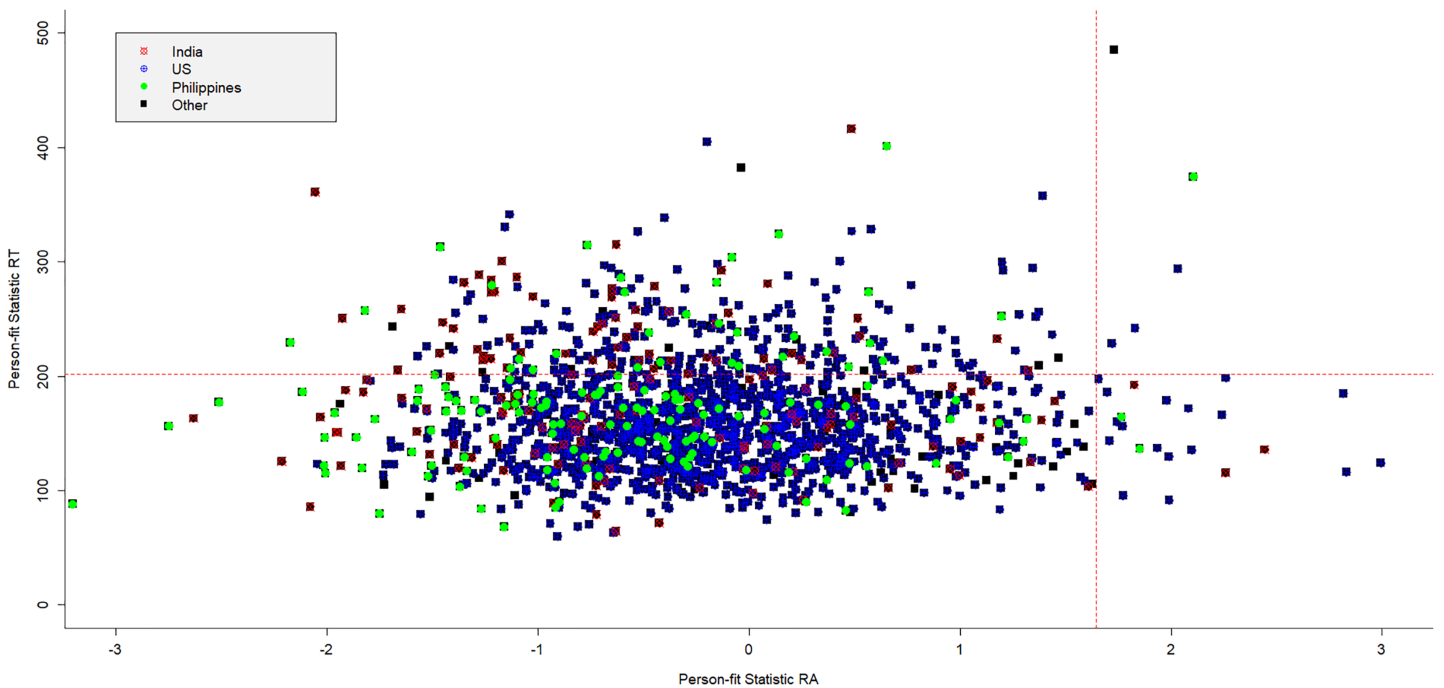
Figure 3 For different countries the person-fit statistic of RA patterns plotted against those of RT patterns.

```
        ylim=c(0,500),cex.main=.8,cex.axis=.7,cex.lab=.8,pch=15)
points(out1$PFl[set1],out1$lZPT[set1],col="blue",pch=10,cex=.5)
points(out1$PFl[set2],out1$lZPT[set2],col="red",pch=13,cex=.5)
points(out1$PFl[set3],out1$lZPT[set3],col="green",pch=16,cex=.5)
abline(h = qchisq(.95, df= 170),lty = 2,col="red")
abline(v = qnorm(.95),lty = 2,col="red")
legend(-3,500,c("India","US","Philippines","Other"), col=c("red","blue",
    "green","black"),pch = c(13,10,16,15),  bg = "gray95",cex=.7)
```

For both person-fit statistics, the threshold value of the significant area is marked with a dotted red line. With respect to aberrant RT patterns, a serious number of test takers are marked as aberrant, since their value is above the threshold of 201.4. Test takers from India have relatively more aberrant persons with respect to RT patterns, and the US test takers relatively less, although many of the aberrant test takers are US citizens. A few test takers are marked as aberrant with respect to their RA pattern, since their statistic value is above 1.645. Only five persons have been marked as aberrant in terms of the RA and RT patterns. None of the four quadrants seems to be dominated by one specific country.

The amsterdam chess data

The Amsterdam Chess data was collected by *Van der Maas & Wagenmakers (2005)* and is used to demonstrate a differential working speed analysis with the package LNIRT. The data study of *Fox & Marianti (2016)* is followed who also used this data set to illustrate

their differential working speed model. In the Amsterdam Chess study, RA and RT patterns are collected of 259 test takers who responded to 40 chess tasks. The chess items have three subdimensions; tactical skill (20 items), positional skill (10 items), and end-game skill (10 items). Each item represents a chess board situation, and the problem-solving task was to select the best possible move. Both RAs and RTs were recorded, where RA was coded as 1 (correct) and 0 (incorrect). In this analysis, the differential working speed model in Eq. (23) is fitted to explore heterogeneity in speed trajectories and to examine the between-person relationship between ability and random speed components.

The LNIRTQ function is used to run the joint model with a differential working speed model with a random trend and quadratic time component. The RA and RT data were extracted from the Amsterdam Chess data to be used in the LNIRTQ function. The RA data consisted of 259 test takers (in rows) and 40 chess items (in columns).

```
R> data(AmsterdamChess)
R> N <- nrow(AmsterdamChess)
R> Y <- as.matrix(AmsterdamChess[c(which(colnames
(AmsterdamChess)=="Y1"):
              which(colnames(AmsterdamChess)=="Y40"))])
R> K <- ncol(Y)
R> Y[Y==9] <- NA
R> RT <- as.matrix(AmsterdamChess[c(which(colnames
(AmsterdamChess)=="RT1"):
              which(colnames(AmsterdamChess)=="RT40"))])
R> RT[RT==10000.000] <- NA
R> RT <- log(RT) #logarithm of RTs
```

The missing RA and RT values coded as 9 and 10,000 are replaced by NAs. There are three records with all missing values (rows 147,201, and 209). In LNIRTQ, imputations are generated under the model for the missing data even if complete records are missing. A time scale is defined by starting at zero and taking steps of $1/K$ to end in $(K-1)/K$. The LNIRTQ model is identified by restricting the difficulty and discrimination parameters to zero and one, respectively. Furthermore, the mean of each random speed parameter (intercept, trend, quadratic) is restricted to zero, and the product of time discriminations is restricted to one. Finally, the covariance of the speed components is restricted to zero, since the covariance among the speed components is modeled by the time discriminations. The LNIRTQ function is ran for 10,000 MCMC iterations, with a default burnin-period of 10%.

```
R> X <- 1:K
R> X <- (X - 1)/K
R> outchess <- LNIRTQ(Y=Y,RT=RT,X=X,XG=10000)
R> summary(outchess)
```

MCMC chains can be checked in the same way for convergence and run length by computing the Geweke and Heidelberger statistics, the effective sample sizes, the MCMC standard errors. The output of LNIRTQ contains MCMC chains of item parameters and hyper prior parameters for items and persons. The examined chains did not have any non-convergence problems. The effective sample size of the chains ranged from 466 (time

discrimination item 1) to 9,000 (time intensity item 1). The naive posterior standard
deviation were often close to the time-series standard error. The 10,000 MCMC iterations
were sufficiently long and the chains showed convergence after 1,000 iterations.

The summary function of LNIRTQ provides a similar output as the LNIRT function.
The posterior mean estimates and (naive) standard deviations of the item parameters are
given, in addition to the covariance matrix of the item and person parameters. For reasons
of brevity, the correlation matrix of the items and the covariance matrix of the persons are
discussed.

The correlation matrix of the items is again given in the order of a, b, ϕ, and λ. The
correlation between discrimination and difficulty is around 0.37, and between time
discrimination and intensity around −0.53. The difficult items tend to discriminate better
in ability than the less difficult items. The high time-intensive items do not discriminate
well between speed levels. RTs were hardly affected by increasing working speed for the
high time-intensive items. For low time-intensive items this effect on the RTs is much
higher. There is a strong correlation between item difficulty and time intensity of 0.79,
which states that the difficult items also took more time to be completed.

```
        ---  Covariance  matrix  Items  ---
     Item  Matrix  Correlation
         [,1]     [,2]     [,3]     [,4]
   [1,]  1.000    0.371    0.050    0.406
   [2,]  0.371    1.000   -0.602    0.785
   [3,]  0.050   -0.605    1.000   -0.534
   [4,]  0.406    0.785   -0.534    1.000
        ---  Covariance  matrix  Person  ---
     Estimated  Value
     Covariance  Matrix
             Theta Intercept Slope1 Slope2
   Theta      0.423     0.114 -0.004 -0.014
   Intercept 0.114     0.060  0.000  0.000
   Slope1   -0.004     0.000  0.114  0.000
   Slope2   -0.014     0.000  0.000  0.063
```

The covariance matrix of the random person parameters are given under the label *Theta*
(ability), *Intercept* (speed intercept), *Slope1* (speed trend), and *Slope2* (speed quadratic
component). The variance in working speed intercepts (*i.e.*, the variance in starting speed
of persons) is small (0.06). The speed trajectories of the persons show a variance of 0.11 in
trends. The variance in deceleration/acceleration in working speed is around 0.06. The
positive covariance between ability and the speed intercept shows that high-ability persons
are more likely to start with a higher speed. The negative covariance between ability and
the speed trend shows that the speed trajectories of high-ability persons has a negative
trend (decrease in speed), in comparison to low-ability persons. The covariance between
ability and the quadratic speed component is around −0.014, which means that high-
ability persons are more likely to show an acceleration in the negative trend in speed.

For the records with missing values, the imputed data leads to population-average estimates of the random person effects. The ability estimate is around the population average of −0.162, and the speed components are around zero. For instance, for the record of 147 the estimated random person effects are (ability, intercept, trend, quadratic):

```
R>  outchess$Mtheta[147,]
 [1]  -0.228  -0.018  -0.004  -0.004
```

SUMMARY AND DISCUSSION

Computer-based testing has made it possible to collect more information from respondents to improve our understanding and interpretation of test performances and of the behaviors of respondents. RT, the amount of time a test taker spends on answering an item, has shown to be a very useful source of information. RTs have been used to measure working speed, item time-intensity, or the relationship between speed and accuracy (*i.e.*, speed-accuracy trade-off). The modeling of RTs and its integration in the measurement of ability has led to the development of joint models for RA and RTs. However, there is a lack of software tools that supports a joint model data analysis, which includes recent joint modeling extensions.

The R-package LNIRT has been developed with the purpose to provide MCMC estimation methods for (hierarchical) joint models for RA and RT data, while including also new developments in this area. The IRT-based measurement models for RA and RT data have been implemented under different parameterizations to make the program suitable for different users. Imputation methods have been integrated to deal with missing data, while the program can also deal with planned missing by design data. Furthermore, Bayesian significance tests are included to evaluate person and item fit for RT and RA patterns. The developed person-fit statistics can be used to identify aberrant test takers with respect to their RT pattern, RA pattern or both patterns. Explanatory variables can be used to explain differences between persons and items.

The LNIRT package has been designed to estimate the parameters of a joint model with multiple link functions (linear, probit) and with non-exponential family distributions. The cross-classified nature of the (random) effects (item and person parameters) further complicates the use of standard (multilevel) software for parameter estimation. Furthermore, Bayesian simulation methods are preferred to handle the required high-dimensional numeric integration for parameter estimation. However, black-box MCMC methods (*e.g.*, JAGS) (1) can be very slow for medium to large data sets and for high-dimensional models (2) cannot integrate identification restrictions in the simulation procedure—for instance re-scaling latent variables to an identified scale in each MCMC iteration—(3) necessary identifying restrictions on parameters can lead to complex priors for the model parameters, and (4) the computation of the model-fit tools is complex, since the tools need to be integrated in the model description. The MCMC sampler in the LNIRT package has been designed to collect posterior samples with low autocorrelation and a high effective sample size. This leads to a faster and more efficient algorithm than a black-box MCMC method, which is not designed to optimize the information content of the posterior samples.

The LNIRT has some limitations which we hope to address in next releases. Currently, the RA data is limited to binary observations. This is a matter of integrating MCMC schemes for ordinal RA data, which have been discussed by *Fox (2010)*. A more elegant way is to integrate a Gibbs sampler for mixed response types that can deal with RA data with different levels of response types. The joint model is limited to two levels of hierarchy, but extensions to more levels have been discussed by *Fox (2010)* and *Klein Entink, Fox & van der Linden (2009)*.

Furthermore, in the differential working speed model test takers can change their speed and the speed components can influence the level of ability. However, the ability component represents a summary measure of accuracy and cannot capture within-individual changes in accuracy. Change in accuracy can be modeled directly with a latent growth model and jointly with a latent growth model for speed. However, binary outcomes contain less information than continuous RTs. This limits the number of growth components that can be estimated and limits the flexibility in describing a pattern of change in accuracy (*Gorter et al., 2020*). The change in accuracy can also be modeled by a hidden markov model (*Molenaar et al., 2016*), where dynamic response behavior is modeled by different item-level states. However, the model complexity increases rapidly when increasing the number of states. Furthermore, the flexibility in modeling change depends on the number of specified states. A more efficient way to model change in accuracy is to adjust a person-specific (or group-specific) discrimination parameter by the level of working speed. Then, to model change in accuracy the contribution of the ability component is adjusted by a discrimination parameter, which level is moderated by working speed. The inclusion of item-specific person-level and person-specific item-level variables to allow the speed-accuracy trade-off to vary between items has been considered by *Goldhammer, Naumann & Greiff (2015)* and *Klotzke & Fox (2019)*. More research is needed to integrate such an approach in LNIRT.

Funding
Ahmet Salih Simsek was supported financially by The Scientific and Technological Research Council of Turkey (TUBITAK) through project number 1059B191800628. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures
The following grant information was disclosed by the authors:
The Scientific and Technological Research Council of Turkey (TUBITAK): 1059B191800628.

Competing Interests
The authors declare that they have no competing interests.

Author Contributions

- Jean-Paul Fox conceived and designed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, development of the computer software, and approved the final draft.
- Konrad Klotzke performed the computation work, prepared figures and/or tables, development of the computer software, and approved the final draft.
- Ahmet Salih Simsek conceived and designed the experiments, analyzed the data, prepared figures and/or tables, and approved the final draft.

Data Availability

The software R-package is available at CRAN: https://cran.r-project.org/web/packages/LNIRT/index.html.

The data and software is available after installing the R-package in R or Rstudio.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.1232#supplemental-information.

REFERENCES

Albert JH, Chib S. 1993. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88(422)**:669–679 DOI 10.1080/01621459.1993.10476321.

Box GE. 1980. Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society: Series A (General)* **143(4)**:383–404 DOI 10.2307/2982063.

Cizek GJ, Wollack JAE. 2016. *Handbook of quantitative methods for detecting cheating on tests.* New York, NY, USA: Taylor & Francis.

Drasgow F, Levine MV, Williams EA. 1985. Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology* **38(1)**:67–86 DOI 10.1111/j.2044-8317.1985.tb00817.x.

Fox JP. 2010. *Bayesian item response modeling: theory and applications.* Berlin: Springer Science & Business Media.

Fox JP, Klein Entink R, van der Linden WJ. 2007. Modeling of responses and response times with the Package CIRT. *Journal of Statistical Software* **20(7)**:1–14 DOI 10.18637/jss.v020.i07.

Fox JP, Klotzke K, Klein Entink R. 2019. *LNIRT: log-normal response time item response theory models.* R package version 0.5.1. *Available at* https://cran.r-project.org/package=LNIRT.

Fox JP, Marianti S. 2016. Joint modeling of ability and differential speed using responses and response times. *Multivariate Behavioral Research* **51(4)**:540–553 DOI 10.1080/00273171.2016.1171128.

Fox JP, Marianti S. 2017. Person-fit statistics for joint models for accuracy and speed. *Journal of Educational Measurement* **54(2)**:243–262 DOI 10.1111/jedm.12143.

Goldhammer F, Naumann J, Greiff S. 2015. More is not always better: the relation between item response and item response time in raven's matrices. *Journal of Intelligence* **3(1)**:21–40 DOI 10.3390/jintelligence3010021.

Gorter R, Fox J-P, Riet GT, Heymans M, Twisk J. 2020. Latent growth modeling of IRT versus CTT measured longitudinal latent variables. *Statistical Methods in Medical Research* **29(4)**:962–986 DOI 10.1177/0962280219856375.

Johnson VE, Albert JH. 2006. *Ordinal data modeling.* Berlin: Springer Science & Business Media.

Karabatsos G. 2003. Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education* **16(4)**:277–298 DOI 10.1207/S15324818AME1604_2.

Klein Entink RH, Fox JP, van der Linden WJ. 2009. A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika* **74(1)**:21–48 DOI 10.1007/s11336-008-9075-y.

Klotzke K, Fox JP. 2019. Bayesian covariance structure modeling of responses and process data. *Frontiers in Psychology* **10**:1675 DOI 10.3389/fpsyg.2019.01675.

Lord FM, Novick MR. 1968. *Statistical theories of mental test scores.* Boston, MA, USA: Addison-Wesley, Reading.

Marianti S, Fox JP, Avetisyan M, Veldkamp BP, Tijmstra J. 2014. Testing for aberrant behavior in response time modeling. *Journal of Educational and Behavioral Statistics* **39(6)**:426–451 DOI 10.3102/1076998614559412.

Maris E. 1993. Adaptive and multiplicative models for gamma distributed variables, and their application as psychometric models for response times. *Psychometrika* **58(3)**:445–469 DOI 10.1007/BF02294651.

Molenaar D, Oberski D, Vermunt J, Boeck PD. 2016. Hidden markov item response theory models for responses and response times. *Multivariate Behavioral Research* **51(5)**:606–626 DOI 10.1080/00273171.2016.1192983.

Plummer M, Best N, Cowles K, Vines K. 2006. CODA: convergence diagnosis and output analysis for MCMC. *R News* **6(1)**:7–11.

Roskam EE. 1997. Models for speed and time-limit tests. In: van der Linden WJ, Hambleton RK, eds. *Handbook of Modern Item Response Theory*. New York, NY, USA: Springer, 187–208.

Samejima F. 1973. Homogeneous case of the continuous response level. *Psychometrika* **38(2)**:203–219 DOI 10.1007/BF02291114.

Scheiblechner H. 1979. Specifically objective stochastic latency mechanisms. *Journal of Mathematical Psychology* **19(1)**:18–38 DOI 10.1016/0022-2496(79)90003-8.

Schnipke DL, Scrams DJ. 1997. Representing response time information in item banks. LSAC Computerized Testing Report 97-09, Law School Admission Council, Newton, PA.

Shi J-Q, Lee S-Y. 1998. Bayesian sampling-based approach for factor analysis models with continuous and polytomous data. *British Journal of Mathematical and Statistical Psychology* **51(2)**:233–252 DOI 10.1111/j.2044-8317.1998.tb00679.x.

Thissen D. 1983. Timed testing: an approach using item response theory. In: Weiss DJ, ed. *Latent Trait Test Theory and Computerized Adaptive Testing*. New York, NY, USA: Academic Press, 179–203.

van der Linden WJ. 2007. A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika* **72(3)**:287–308 DOI 10.1007/s11336-006-1478-z.

van der Linden WJ, Scrams DJ, Schnipke DL. 1999. Using response-time constraints to control for speededness in computerized adaptive testing. *Applied Psychological Measurement* **23(3)**:195–210 DOI 10.1177/01466219922031329.

van der Linden WJ, Fox JP. 2016. Joint hierarchical modeling of responses and response times. In: *Handbook of Item Response Theory, Volume One: Models*. Boca Raton, FL, USA: CRC Press, 481–500.

Van der Maas HL, Wagenmakers EJ. 2005. A psychometric analysis of chess expertise. *The American Journal of Psychology* **118(1)**:29–60 DOI 10.2307/30039042.

Verhelst ND, Verstraalen HHFM, Jansen MG. 1997. A logistic model for time limit tests. In: van der Linden WJ, Hambleton RK, eds. *Handbook of Modern Item Response Theory*. New York, NY, USA: Springer, 169–185.