Person-Fit Statistics for Joint Models for Accuracy and Speed

Jean-Paul Fox University of Twente

Sukaesi Marianti University of Brawijaya

Abstract

Response accuracy and response time data can be analyzed with a joint model to measure ability and speed of working, while accounting for relationships between item and person characteristics. Person-fit statistics are proposed for joint models to detect aberrant response accuracy and/or response time patterns. The person-fit tests take the correlation between ability and speed into account, as well as the correlation between item characteristics. They are posited as Bayesian significance tests, which have the advantage that the extremeness of a test statistic value is quantified by a posterior probability. The person-fit tests can be computed as by-products of an MCMC algorithm. Simulation studies were conducted in order to evaluate their performance. For all person-fit tests, the simulation study shows good detection rates in identifying aberrant patterns. A real data example is given to illustrate the person-fit statistics for the evaluation of the joint model.

Keywords: Bayesian procedures, Joint models, Person fit, Hierarchical models

Introduction

In computer-based testing, response accuracy (RA) and response time (RT) data can be collected. The response times (RTs) used to respond to items provide information about working speed, where RA data provide information about ability. RTs are collected in order to estimate speed and item time-intensity (i.e., population-average amount of time needed to complete an item), to investigate relationships with speed components and accuracy, but also to investigate issues in educational testing. For instance, the impact of time limits on test takers' performances and to what extent time limits initiate rapid-guessing behavior has been investigated using RTs (e.g., Chang, Tsai & Hsu, 2014; Schnipke & Scrams, 1997). RTs have also been useful in identifying test takers with aberrant response behavior (e.g., cheating, guessing) (Marianti, Fox, Avetisyan, Veldkamp, & Tijmstra, 2014).

This has raised an increased interest in modeling and analyzing RTs. Goldhammer (2015), Lee and Chen (2011), and van der Linden (2006, 2007) give an overview of recent developments in RT modeling, where joint models for speed and ability are extensively discussed. In this joint modeling approach, ability is considered a latent variable, where item responses serve as indicators to measure ability. Speed is also considered a latent variable, since it cannot be observed directly and is always measured in relation to time intensities (e.g., van der Linden, 2011).

Despite the methodological advances in the joint modeling of speed and ability, and the many applications in educational assessment, there has not been much attention in testing the underlying assumptions of a joint model for speed and ability. Model comparison criteria have been proposed for RT models (e.g., Klein Entink, Fox, et al., 2009). However, model-fit statistics for the evaluation of the joint model have been sparsely developed and evaluated. Bolsinova and Maris (2016) developed a test for conditional independence between RA and RT data given an exponential family model for the response accuracy. Van der Linden and Guo (2008) developed Bayesian posterior predictive checks to identify aberrant response time patterns. Some model-fit tests for item response theory (IRT) model can been used for the joint model but their performances for the joint model have not been fully tested. Tests to evaluate the fit of RT models (e.g., Marianti et al., 2014) have also not been implemented and evaluated for the joint model.

In this paper, person-fit tests are developed for the joint model to distinguish test takers with aberrant RA patterns and/or RT patterns from test takers with normal item response patterns and/or RT patterns. The object is to identify aberrant response behavior, where the extremeness of an RA pattern and RT pattern needs to be evaluated simultaneously, since speed and ability affects the response behavior. This relates to van der Linden (2009) and Goldhammer (2015), who argued that in the measurement of ability the influence of speed on the response behavior needs to be controlled. In the same way, it can be argued that in the evaluation of RA patterns the RT patterns needs to be taken into account.

The proposed person-fit statistics can be classified as Bayesian significance tests. In this Bayesian test procedure, the extremeness of each person-fit value is quantified by computing the posterior probability (p-value) that the value is greater than a certain threshold given the data. This Bayesian p-value is directly interpretable and represents the posterior probability of the extremeness of a person-fit value given the data.

The person-fit tests for the joint model can be used to identify aberrant response behavior (e.g., cheating, guessing, random responding) manifested in the RA pattern and/or manifested in the RT pattern. The developed person-fit tests are based on the log-likelihood of an RA pattern (Drasgow, Levine, & Williams, 1985) and of an RT pattern (Marianti et al., 2014). The joint distribution of ability and speed is included in the computation of the person-fit tests to account for the relation between speed and ability in evaluating the extremeness of the RA and RT pattern. The performance of the person-fit statistics is evaluated using simulation studies.

After introducing the joint model for ability and speed, person-fit statistics are defined under the log-normal RT model and IRT model. It will be shown that given all information, RA and RT patterns can be identified as aberrant with a specific posterior probability, according to the Bayesian significance test procedure. In a simulation study, the power to detect the aberrancies is investigated by simulating various types of aberrant RA and RT data. A real data study is used to illustrate the use of person-fit statistics for different joint models. Several directions for future research are presented.

The Joint Modeling Framework

RA and RT data can be jointly modeled using a hierarchical latent variable model. Working speed and ability, defined at the level of persons, are assumed to underlie the RT and RA data, defined at the level of observations, respectively. Fox, Klein Entink, and van der Linden (2007), Klein Entink, Fox, et al. (2009), and van der Linden (2007) have developed a Bayesian modeling framework where a lognormal RT model and an IRT model are used to model the level-1 observations. At level 2, population distributions are defined for the latent variables speed and ability, and for the item parameters in the RT model and the IRT model.

Although this hierarchical model for speed and ability consists of distinct distributions for RA and RT data, a relationship between speed and ability is specified at the level of persons. This joint model (e.g., van der Linden, 2007; Klein Entink, Fox, & van der Linden, 2009) has been used to explain a cognitive structure (Klein Entink, Kuhn, Hornke, & Fox, 2009), to test hypothesis about relationships between speed and ability, to calibrate test items, and to investigate test design characteristics, speededness, and cheating.

Another joint modeling approach considered the generalized linear model for RA and RT data (Molenaar, Tuerlinckx, & van der Maas, 2015). This generalized linear approach restricts items to have equal discriminations. However, it is more realistic to assume that the effect of an increase in ability (working speed) can have a different effect across items on the accuracy in responses (time to respond). A joint IRT modeling approach for categorical RTs has also been considered (e.g., DeBoeck & Partchev, 2012; Partchev & DeBoeck, 2012; Ranger & Kuhn, 2012). Treating continuous time observations as categorical RTs will always reduce the available amount of information, where the measurement precision can be increased by modeling the continuous RTs. A general overview of different modeling approaches is given by van der Linden (2009).

In the present joint modeling approach, a two-parameter IRT model for binary responses is considered, and the log-normal model for RTs including time-discriminations and item-specific error variances. For the RT model, let RT_{ik} denote the RT of person *i* $\left(i=1,\ldots,N\right)$ on item *k* $\left(k=1,\ldots,K\right)$. A lognormal RT distribution is considered to account for the positively skewed characteristic of RT distributions. The lognormal distribution for the RT is given by

$$\ln RT_{ik} = \lambda_k - \phi_k \zeta_i + \varepsilon_{ik},\ \varepsilon_{ik} \sim N\left(0, \sigma^2_{\varepsilon_k}\right), \tag{1}$$

where 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. Fox et al. (2007) and Klein Entink, Fox, et al. (2009) introduced the time-discrimination parameter as a slope parameter for speed, which models the sensitivity of the item for different speed-levels of the test takers.

This specification of the time discrimination parameter differs from the time discrimination parameter defined by van der Linden (2009). In his approach, the reciprocal of the standard deviation of the measurement error is defined to be the time discrimination. This also allows for item-specific variances. However, the time discriminations in Equation (1) also model covariances between RTs. When considering responses to item *k* and *l,* the covariance between RTs of test taker *i* is given by,

$$\operatorname{cov}\left(RT_{ik}, RT_{il}\right) = \operatorname{cov}\left(\lambda_k - \phi_k \zeta_i + \varepsilon_{ik}, \lambda_l - \phi_l \zeta_i + \varepsilon_{il}\right) = \operatorname{cov}\left(-\phi_k \zeta_i, -\phi_l \zeta_i\right) = \phi_k \operatorname{var}\left(\zeta_i\right) \phi_l, \quad (2)$$

when assuming independent errors and time intensities. So, the covariance between the two RTs is influenced by both time discriminations. Furthermore, the additional error term in Equation (1) can model variations in RTs due to stochastic behavior of the 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 avoid bias in the parameter estimates.

Besides observing the RT for item *k*, let Y_{ik} denote the RA of person *i* on item *k* (coded 1 if a correct answer is given and 0 if a wrong answer is given). A two-parameter IRT model is considered to describe the RA. The probability of a correct response is given by

$$P\left(Y_{ik} = 1 \middle| \theta_i, a_k, b_k\right) = \Phi\left(a_k \theta_i - b_k\right) \quad (3)$$

where Φ denotes the normal cumulative distribution function, a_k the discrimination parameter, and b_k the difficulty parameter. A latent response variable Z_{ik} can be defined, which is normally distributed with mean $a_k\theta_i - b_k$ and variance 1. So, the probability of a correct response is equal to the probability that the latent response variable is greater than 0. The latent response variable Z_{ik} is often used to represent a latent continuous response given an observed categorical response (Fox, 2010), and is also referred to as an auxiliary response. The latent response variable is useful in a MCMC simulation procedure to estimate the model parameters (e.g., Albert & Chib, 1993; Fox, 2010; Klein Entink, Fox et al., 2009; Meng, Tao, & Chang, 2015). For the two-parameter model, the response Y_{ik} is the indicator of the latent response variable Z_{ik} being positively truncated.

The joint model for the latent continuous responses and RTs is given by,

$$\begin{aligned} Z_{ik} &= a_k\theta_i - b_k + e_{ik},\ e_{ik} \sim N\left(0,1\right) \\ \ln RT_{ik} &= \lambda_k - \phi_k\zeta_i + \varepsilon_{ik},\ \varepsilon_{ik} \sim N\left(0,\sigma^2_{\varepsilon_k}\right). \end{aligned} \tag{4}$$

The test takers are assumed to be randomly selected from a population and the ability and speed variable are assumed to have a multivariate normal population distribution

$$\begin{pmatrix} \theta_i \\ \zeta_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_\theta \\ \mu_\zeta \end{pmatrix}, \begin{pmatrix} \sigma^2_\theta & \rho_{\theta\zeta} \\ \rho_{\theta\zeta} & \sigma^2_\zeta \end{pmatrix}\right), \tag{5}$$

where the population means μ_θ and μ_ζ represent the population average level of ability and working speed, respectively. The population variances σ^2_θ and σ^2_ζ represent the variance in ability and working speed in the population, where $\rho_{\theta\zeta}$ represents the common covariance between ability and speed.

The population distribution of the item characteristics is a multivariate normal, which is given by,

$$\begin{pmatrix} a_k \\ b_k \\ \phi_k \\ \lambda_k \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_a \\ \mu_b \\ \mu_\phi \\ \mu_\lambda \end{pmatrix}, \begin{pmatrix} \sigma_a^2 & \rho_{ab} & \rho_{a\phi} & \rho_{a\lambda} \\ \rho_{ab} & \sigma_b^2 & \rho_{b\phi} & \rho_{b\lambda} \\ \rho_{a\phi} & \rho_{b\phi} & \sigma_\phi^2 & \rho_{\phi\lambda} \\ \rho_{a\lambda} & \rho_{b\lambda} & \rho_{\phi\lambda} & \sigma_\lambda^2 \end{pmatrix} \right). \tag{6}$$

The mean of the multivariate distribution represents the test-average item characteristics. The covariance matrix represents the variation in item characteristics and the covariance between them.

Person Fit for Speed and Accuracy

When a joint model is used to measure ability and speed, person-fit statistics are required to distinguish test takers with aberrant RA patterns and/or RT patterns from test takers with normal RA patterns and/or RT patterns. Scores on a test can be spuriously high or low due to cheating, guessing or random responding (e.g., Meijer, 1996; van der Linden & Lewis, 2015), which can be directly detected from the RA pattern. However, it is also possible that the aberrant response behavior is manifested in the RT pattern. Subsequently, an inaccurate working speed measurement can bias an estimated relation between speed and ability. It is also possible that both the RA pattern and the RT pattern indicate aberrant response behavior. In that case the RT simply contributes to the evidence to identify aberrant-responding test takers.

Before discussing the person-fit test for the joint model, a person-fit test for RA patterns and one for RT patterns is discussed. For each person-fit 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 Patterns

Meijer and Sijtsma (2001), Karabatsos (2003), and Rupp (2013) give an overview of the large number of person-fit statistics to detect aberrant RA patterns. The parametric tests are based on the principle that the fit of responses to a set of items are evaluated under the IRT model. The popular person-fit statistic of Drasgow, Levine, and Williams (1985) and the standardized version of Levine and Rubin (1979) is based on the log-likelihood of the RA pattern. This person-fit statistic is often used in educational measurement and is shown to have power to identify aberrant RA patterns (Karabatsos, 2003; Dimitrov and Smith, 2006).

The log-likelihood of the responses is used to evaluate the fit of an RA pattern, and is referred to as a person-fit statistic, denoted as l^y. Given the log-likelihood according to a two-parameter IRT model, the person-fit statistic is given by,

$$\begin{aligned} l^y\left(\theta_i,\mathbf{a},\mathbf{b};\mathbf{y}_i\right) &= \log p\left(\mathbf{y}_i \,|\, \theta_i,\mathbf{a},\mathbf{b}\right) \\ &= \sum_{k=1}^{K} \log p\left(y_{ik} \,|\, \theta_i,a_k,b_k\right) \\ &= \sum_{k=1}^{K} \left[y_{ik} \log p\left(y_{ik} \,|\, \theta_i,a_k,b_k\right) + \left(1-y_{ik}\right)\log\left(1-p\left(y_{ik} \,|\, \theta_i,a_k,b_k\right)\right)\right]. \end{aligned} \tag{7}$$

The expected value and variance are given by

$$\begin{aligned} E\left(l^y\left(\theta_i,\mathbf{a},\mathbf{b};\mathbf{y}_i\right)\right) &= \sum_{k=1}^{K} p\left(y_{ik}=1 \,|\, \theta_i,a_k,b_k\right)\log p\left(y_{ik}=1 \,|\, \theta_i,a_k,b_k\right)+ \\ &\quad \left(1-p\left(y_{ik}=1 \,|\, \theta_i,a_k,b_k\right)\right)\log\left(1-p\left(y_{ik}=1 \,|\, \theta_i,a_k,b_k\right)\right), \end{aligned} \tag{8}$$

and

$$\begin{aligned} Var\left(l^y\left(\theta_i,\mathbf{a},\mathbf{b};\mathbf{y}_i\right)\right) &= \sum_{k=1}^{K} \Bigg[p\left(y_{ik}=1 \,|\, \theta_i,a_k,b_k\right)\left(1-p\left(y_{ik}=1 \,|\, \theta_i,a_k,b_k\right)\right) \\ &\quad \log\left(\frac{p\left(y_{ip}=1 \,|\, \theta_i,a_k,b_k\right)}{1-p\left(y_{ip}=1 \,|\, \theta_i,a_k,b_k\right)}\right)^2 \Bigg], \end{aligned} \tag{9}$$

respectively. Subsequently, the standardized version of this person-fit statistic is given by

$$l_s^y\left(\theta_i,\mathbf{a},\mathbf{b};\mathbf{y}_i\right)=\frac{l^y\left(\theta_i,\mathbf{a},\mathbf{b};\mathbf{y}_i\right)-E\left(l^y\left(\theta_i,\mathbf{a},\mathbf{b};\mathbf{y}_i\right)\right)}{\left(Var\left(l^y\left(\theta_i,\mathbf{a},\mathbf{b};\mathbf{y}_i\right)\right)\right)^{\frac{1}{2}}}, \tag{10}$$

which is assumed to be approximately standard normally distributed.

A Bayesian significance test can be defined to compute the extremeness of each RA pattern. From Equation (7) follows that the person-fit statistic is an increasing function of the likelihood of the responses. The logarithm-of-response-probabilities are negative values, and the logarithm of more likely response probabilities is higher than the logarithm of less likely response probabilities. Therefore, increasing values of the negative person-fit statistic correspond to misfit.

The negative person-fit statistic is considered in the computation of a critical region. As a result, for increasing l_s^y values the posterior probability of misfit is increasing. The posterior probability that the person-fit statistic is greater than a certain threshold value is given by,

$$P\left(l_s^y\left(\theta_i,\mathbf{a},\mathbf{b};\mathbf{y}_i\right)>C\right)=\Phi\left(l_s^y\left(\theta_i,\mathbf{a},\mathbf{b};\mathbf{y}_i\right)>C\right)=p_{l^y}. \tag{11}$$

For the joint model, the computation of the person-fit statistic is more complex due to relationships between model parameters such as speed and ability. These relationships need to be taken into account in the computation of the person-fit statistic. The MCMC algorithm developed for the joint model (e.g., Klein Entink, Fox et al., 2009) supports the computation of expected a posteriori estimates of the model parameters, while accounting for the full covariance structure between model parameters. This MCMC algorithm can also be used to compute the person-fit statistic in Equation (11) , taking into account the full covariance structure. Therefore, a dichotomous classification variable F_i^y is defined, which equals 1 when the RA pattern of test taker *i* is marked aberrant and 0 otherwise, given the population and item parameters,

$$F_i^y = \begin{cases} 1 & \text{if} \quad P\left(l_s^y\left(\theta_i, \mathbf{a}, \mathbf{b}; \mathbf{y}_i\right) > C\right) \\ 0 & \text{if} \quad P\left(l_s^y\left(\theta_i, \mathbf{a}, \mathbf{b}; \mathbf{y}_i\right) \leq C\right). \end{cases} \tag{12}$$

The status of F_i^y can be computed in each MCMC iteration, and the average over MCMC iterations is used as an estimate of the posterior probability of an aberrant RA pattern. This method makes it possible to identify aberrant RA patterns, for which the person-fit statistic value is greater than the threshold value, with a specific posterior probability. That is, the extremeness of each RA pattern can be quantified with a posterior probability given the threshold *C*.

Person-fit Statistic for RT Patterns

In Marianti et al. (2014), a person-fit statistic for RTs is defined, which is based on the likelihood of RT patterns. Let $RT_{ik}^* = \ln\left(RT_{ik}\right)$ denote the logarithm of the RT of test taker i on item k. Given the model specification in Equation (1), the likelihood of an RT pattern is represented by,

$$\begin{aligned} -2\log p\left(\mathbf{rt}_i^* \middle| \zeta_i, \boldsymbol{\lambda}, \boldsymbol{\phi}, \boldsymbol{\sigma}^2\right) &= -2\sum\nolimits_{k=1}^{K} \log p\left(rt_{ik}^* \middle| \zeta_i, \lambda_k, \phi_k, \sigma_k^2\right) \\ &= \sum\nolimits_{k=1}^{K} \left(\left(\frac{rt_{ik}^* - \mu_{ik}}{\sigma_k} \right)^2 + \log\left(2\pi\sigma_k^2\right) \right) \\ &= \sum\nolimits_{k=1}^{K} \left(Z_{ik}^2 + \log\left(2\pi\sigma_k^2\right) \right), \end{aligned} \tag{13}$$

where Z_{ik} is standard normally distributed, since it represents the standardized error of the normally distributed logarithm of RT. The sum of standardized errors is an increasing function of the negative log-likelihood of RTs. This error function is used as the likelihood-based person-fit statistic for RTs,

$$l_i^t\left(\zeta_i, \boldsymbol{\lambda}, \boldsymbol{\phi}, \boldsymbol{\sigma}^2; \mathbf{rt}_i^*\right) = \sum_{k=1}^{K} Z_{ik}^2 \tag{14}$$

where unusually large statistic values indicate a misfit. The statistic represents a departure of the RTs from expected RTs under the model. The posterior distribution of the statistic can be used to examine whether a pattern of observed RTs is extreme under the model.

The distribution of the test statistic is chi-square with K degrees of freedom given the model parameters. Let threshold C define the boundary of a critical region. This critical region contains the set of values for the observed statistic value for which the null hypothesis is rejected. This critical value C can be determined from the chi-square distribution,

$$P\left(l_i^t\left(\mathbf{rt}_i^*\right) > C\right) = P\left(\chi_K^2 > C\right) = p_{l^t}. \tag{15}$$

The p_{l^t} is the posterior probability that the observed statistic value is larger than *C* given the RT pattern, and the chi-square distribution of the test statistic can be used to evaluate the extremeness of an RT pattern.

The person-fit statistic, l^t, depends on the parameter values of the RT model, which are also related to the parameters of the IRT model. To compute the posterior probability in Equation (15), again a dichotomous classification variable can be used. Let variable F_i^t equal 1 for test taker *i* when his/her RT pattern is marked aberrant, and *0* otherwise; that is,

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

The status of F_i^t can be computed in each MCMC iteration, given draws of the parameters. The average over MCMC iterations is used as an estimate of the posterior probability of an aberrant RT pattern for test taker *i*.

Person-fit Statistic for RA and RT Patterns

A third classification variable is defined to identify test takers with aberrant RT and RA patterns. This third classification variable equals 1 when the other classification variables are both equal to 1, $F_i^t = 1$ and $F_i^y = 1$, and equals 0 otherwise. For test taker *i*, the classification variable to identify an extreme pattern for responses and for RTs is defined by,

$$F_i^{t,y} = \begin{cases} 1 & \text{if} \quad P\left(l^t\left(\zeta_i, \phi, \boldsymbol{\lambda}; \mathbf{rt}_i\right) > C,\ l_s^y\left(\theta_i, \boldsymbol{\alpha}, \boldsymbol{\beta}; \mathbf{y}_i\right) > C\right) \\ 0 & \text{if} \quad 1 - P\left(l^t\left(\zeta_i, \phi, \boldsymbol{\lambda}; \mathbf{rt}_i\right) > C,\ l_s^y\left(\theta_i, \boldsymbol{\alpha}, \boldsymbol{\beta}; \mathbf{y}_i\right) > C\right). \end{cases} \tag{17}$$

The computation of the posterior probability of extreme RA and RT patterns can be done using MCMC. In each MCMC iteration, the status of the classification variable is evaluated and the average over MCMC iterations is an estimate of the marginal posterior probability of an aberrant test-taker concerning the RA and RT pattern.

The posterior probabilities based on the classification variables defined in Equations (12), (16), and (17), are significance probabilities that express the extremeness of statistic values given the data. They can be used to identify aberrant response behavior with respect to RA and/or RT patterns. MCMC is used to compute the significance probabilities, where dependencies between the joint model parameters are also taken into account.

Three Parameter IRT Model

Guessing behavior can lead to aberrant RA patterns under the two-parameter IRT model. However, this type of response behavior can also be controlled for by the three-parameter IRT model. Then, the probability of a correct response is a random guess with probability c_k. With probability $1 - c_k$ the response is not guessed but a response is given according to a two-parameter IRT model. This three-parameter IRT model is given by

$$P\left(Y_{ik} = 1 \middle| \theta_i, a_k, b_k, c_k\right) = c_k + \left(1 - c_k\right)\Phi\left(a_k\theta_i - b_k\right). \tag{18}$$

Although the three-parameter model accounts for guessing behaviour, other types of aberrant response behaviour for the non-guessed responses can still be investigated. The person-fit statistic defined in Equation (11) can be defined for the non-guessed responses. Therefore, a latent variable S_{ik} is introduced, which is equal to one, when the test taker *i* knows the correct response to item *k*, and is equal to zero otherwise.

Béguin and Glas (2001) and Glas and Meijer (2003) defined such a latent variable and showed methods to simulate $\mathbf{S}_i$ values under the model given binary responses.

The person-fit statistic, defined in Equation (11), is used to evaluate the fit of the responses for which the test taker *i* knows the response $(S_{ik}=1)$. Then, a Bayesian significance probability can be defined, which is the probability of an extreme statistic value for the non-guessed responses integrated over the possible non-guessed responses of test taker *i*,

$$p_{l^y} = \int \Phi\left(l_i^y\left(\theta_i, \mathbf{a}, \mathbf{b}; \mathbf{s}_i\right) > C\right) p\left(\mathbf{s}_i \,\middle|\, \mathbf{y}_i, \theta_i, \mathbf{a}, \mathbf{b}\right) d\mathbf{s}. \qquad (19)$$

An MCMC method is used to sample latent variable values $\mathbf{S}_i$, and to compute the significance probability.

MCMC Estimation

A Markov chain Monte Carlo (MCMC) algorithm was implemented to estimate the parameters of the joint model. The MCMC procedure is based on the algorithm developed by Fox et al., (2007) and Klein Entink, Fox, et al., (2009). The MCMC algorithm was extended to include the estimation of the three-parameter normal-ogive model parameters. This extension was based on the simulation of auxiliary data **S**, and sampling of guessing parameters c_k, where a Beta prior distribution was used for the guessing parameter. The other priors were similar to those of Fox et al., (2007) and

Klein Entink, Fox, et al., (2009). The R-code of the estimation method can be found in the R-package LNIRT[1].

Simulated Data Analysis

Parameter recovery of the joint model with guessing

The performance of the MCMC algorithm to estimate the joint model with a three-parameter IRT model for the RA data was evaluated. Therefore, a total of 20 data sets were simulated. Each data set consisted of simulated RA and RT data to K=20 items of N=500 persons according to the joint model. Furthermore, for each simulated data set, the joint model parameters were simulated from their prior distributions. The item parameters were simulated from the multivariate prior with a mean vector of $\mu_a = 1$, $\mu_b = 0$, $\mu_\phi = 1$, and $\mu_\lambda = 0$, and with a diagonal covariance matrix with elements; $\sigma_a^2 = .05$, $\sigma_b^2 = 1$, $\sigma_\phi^2 = .05$, and $\sigma_\lambda^2 = 1$. The simulated guessing parameters were equal to .10 to simulate a moderate level of guessing. As a result, the performance of the MCMC algorithm was investigated without having overwhelming data evidence about guessing behaviour.

Speed and ability were simulated from a multivariate normal distribution with means equal to zero and variances to one. The covariance between speed and ability was equal to .75 to simulate a high correlation between the person parameters. A high correlation was simulated to show that the estimation method can also support the computation of the person-fit test for the joint model (Equation (17)) for a more extreme situation. The person-fit test exploits the correlation between speed and ability, and a higher correlation will induce more correlation between RA and RT patterns. The

[1] The R-package LNIRT (https://cran.r-project.org/package=LNIRT) contains the MCMC algorithm for the joint model, the person-fit statistics and several tools for residual analysis, which are described in the web-supplement.

measurement error variances were simulated from a lognormal distribution with a mean of zero and standard deviation of .20.

For each replicated data set, the MCMC algorithm was ran for 10,000 iterations and a burn-in period of 1,000 iterations was used. The MCMC algorithm showed rapid convergence without using informative starting values. The R-Coda package (Plummer, Best, Cowles & Vines, 2006) was used to investigate the convergence of the MCMC chains. The commonly used convergence diagnostics (e.g., Geweke, Heidelberger and Welch) did not show any issues.

Figure 1 shows the estimated bias (y-axis) across 20 data replications, as the deviation between estimated and true value of each item parameter (x-axis). The biases are small, generally ranging from -0.077 to 0.045, for all item parameters. The estimated posterior standard deviations, averaged across 20 data replications, ranged from 0.022 to 0.127, and represent the uncertainty of the estimates. It was concluded that there exists a close agreement between estimated values and true values. It shows that the MCMC method, for estimating the joint model with a three-parameter IRT model, was able to recover the true item parameters.

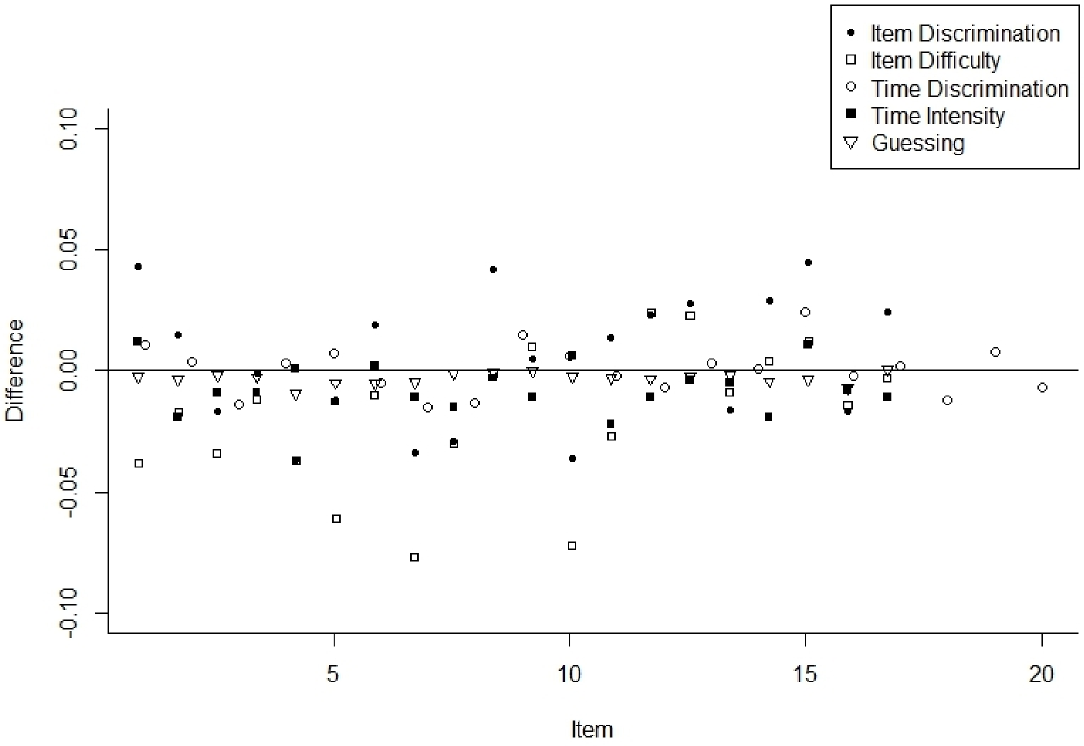

Figure 1: For each item, the difference between simulated and estimated values, averaged across 20 datasets, of the item parameters are plotted.

Evaluate Performance Person-Fit Tests

Two types of aberrant patterns were simulated. One type represented cheating, where aberrant RA data were simulated by letting low-ability persons correctly answer difficult items (Meijer, 1996; St-Onge et al., 2011). More specifically, in this simulation study, low-ability persons had a probability of .75 to answer the 10 most difficult items correctly.

The other type represented random RT behavior. For this type of behavior, random RTs were generated for persons who copied answers or used a cheat sheet and did not require time to complete the items. Their RT patterns contained extreme RTs, although their total time to finish the test was not extreme. To simulate random RT behavior, aberrant RTs were generated from a log-normal distribution with the mean equal to the average RTs, and the standard deviation equal to three times the average standard deviation of RTs. As a result, the average test time for simulated aberrant RT patterns was similar to the average test time of non-aberrant RT patterns. However, for the same item randomly selected RTs of aberrant patterns deviated significantly from average RTs of non-aberrant patterns.

Three conditions were considered. In condition 1, aberrant RA and RT patterns were simulated according to cheating and random RT behavior. In condition 2, aberrant RT patterns were simulated according to random RT behavior, and in condition 3, aberrant RA patterns were simulated according to cheating. Furthermore, for each condition a total of 1,000 response and RT patterns for a 20-item test were simulated, where 100 (10% of the sample) and 200 (20% of the sample) patterns were simulated to be aberrant.

For 20 replicated data sets, the MCMC algorithm showed accurate estimation results, and they showed sufficient variation in parameter estimates given the estimated standard deviations. Therefore, in this simulation study, 50 data replications were expected to provide accurate information about the detection rates of the person-fit statistics. In Table 1, the estimated detection rates of both person-fit statistics are given, which are based on 50 replicated data sets.

A significance level of .05 was used. In Table 1, the header “Aberrant” refers to all simulated persons with aberrant patterns, and header “Total” represents all persons (aberrant and non-aberrant) in the sample. The Type-I errors are given under the header “Model Fit”: 4.2% and 3.1% of persons in the sample (under the label “Total”) were detected as aberrant according to l_t and l_s^y, respectively. That is, in the condition that no aberrant patterns were simulated, the detection rates slightly underestimated the true Type-I errors, but will improve when increasing the number of simulated patterns.

TABLE 1

Detection rates of person-fit tests l_t and l_s^y*, to identify aberrant RA and/or RT patterns for N=1,000 and K=20.*

		l_t statistic		l_s^y statistic	
		Aberrant	**Total**	**Aberrant**	**Total**
Model Fit *Significance level 0.05*		-	4.2	-	3.1
10%	*Condition 1*	9.4	9.7	10	13.4
	Condition 2	9.4	9.7	-	3.2
	Condition 3	-	3.9	10	13.1
20%	*Condition 1*	15.6	15.8	20	23.3
	Condition 2	15.6	15.8	-	3.2
	Condition 3	-	4.1	19.9	23.2

Note: Condition 1, random RT behavior and cheating; condition 2, random RT behavior; condition 3, cheating.

In the first condition (*Random RT Behavior and Cheating*), both types of aberrant behavior were simulated. The results show that with 10% simulated aberrant patterns the l_t and l_s^y statistics were able to detect 9.4% and 10% of them, respectively. When 20% of the persons were simulated with aberrant responses, the l_t and l_s^y were able to detect 15.6% and 20% of them, respectively.

The performance of each person-fit statistic was also evaluated under each condition separately. Both the l_t and l_s^y showed less performance, when increasing the percentage of aberrant patterns from *10*% to *20*%. This corresponds to the finding of Karabatsos (2003), who reported that generally, detection rates decrease as the percentage of aberrant patterns increases. Both statistics show good detection rates in identifying aberrant patterns and the performance is similar to the results of the condition with both violations. In the condition with only aberrant patterns due to cheating, the l_t statistic showed detection rates close to the detection rates obtained under the Model Fit condition. In the condition with only aberrant patterns due to random RT behavior, the l_s^y statistic showed detection rates close to the Type-I error.

Real Data Analysis

Cizek and Wollack (2016) discuss a real data set (referred to as the credentialing data) concerning 1,636 test takers who applied for licensure. The candidates made Form 1 of the test, which consisted of 170 items, and their RA and RT data were stored. The collected data followed from a year of testing using a computer-based program that tests continuously. Besides the RA information, background information of each candidate was available, for instance, 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 test takers were pretested using three different item sets. The average scores varied significantly across the differently pretested groups.

In this study, RT and RA patterns of 723 test takers, who were pretested with the same item set, were analyzed using the joint model. The person-fit tests were used to detect aberrant response behavior, without using any background information. The LNIRT program was used to estimate all model parameters and to compute the person-fit statistics.

The joint 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.

The MCMC convergence diagnostics were used to evaluate the convergence of the chains. According to the diagnostics, a burn-in period of 1,000 iterations and a total of 5,000 MCMC iterations were made to estimate the model parameters (R-Coda package; Plummer et al., 2006).

In Figure 2, the estimated item parameters are shown. The estimated item difficulties and time intensities were rescaled to have a mean of zero to present them in the same plot on the same scale. The estimated mean of the item difficulties is -.56 and the item difficulties ranges from -1.71 to .79. The estimated mean of the time intensities is around 4.00 and the time intensities ranges from 2.89 to 4.85. So, the average RT to complete each item ranges from $\exp(2.89) \approx 18s$ to $\exp(4.85) \approx 128s$.

It can be seen that the range in item difficulties is relatively large, which gives support to accurate estimation of test-takers' ability on the entire range of the scale. Some of the item discriminations are relatively small (slightly above .30). The highly discriminating items are also the most difficult items. The variety in time discriminations is not very high, which ranges from .4 to 1.6 on a logarithmic scale. The average population level of speed was fixed to zero to identify the scale.

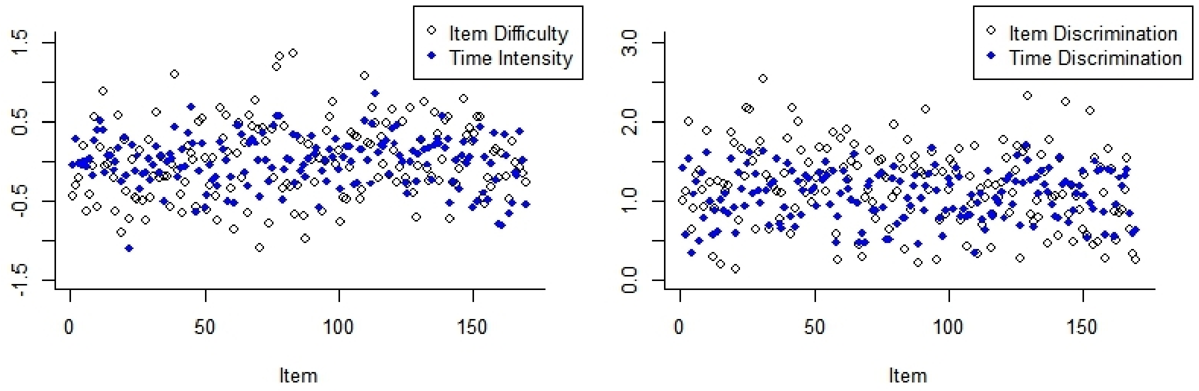

Figure 2. The estimated item parameters; difficulty and time intensity in the left subplot, and item discrimination and time-discrimination in the right subplot.

The person-fit statistics to detect aberrant response behavior were computed. In Figure 3, the estimated person-fit statistic values, l^t, are plotted against the posterior probability of significance. The statistic values are chi-square distributed with 170 degrees of freedom, under the joint model. Subsequently, the critical statistic value is 201.4, when the level of significance equals .05. Estimated statistic values higher than

201.4 are located in the critical region. Given this significance level, the estimated number of aberrant patterns is 124, which is 17.15% of the test takers.

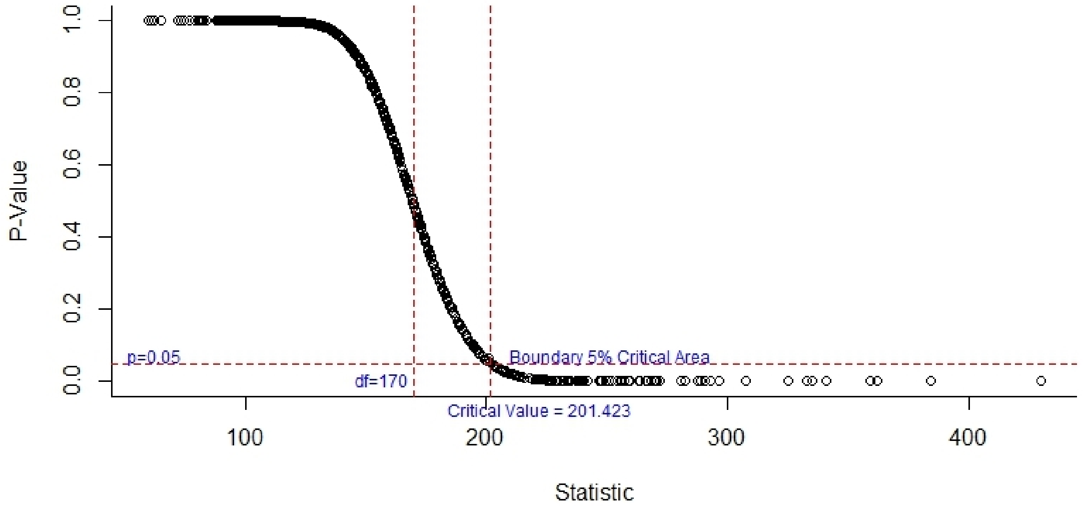

Figure 3. Estimated person fit statistics l_t with respect to the RT patterns plotted against the corresponding posterior significance probability.

In Figure 4, the estimated person-fit statistic values l_s^y are plotted against the posterior significance probability. The critical area is above the statistic value of 1.645, when considering a significance level of .05. Test takers with a statistic value higher than 1.645, are located in the critical region. In this study, 16 persons (around 2.2%) are identified in the critical region and hence, are detected as persons with aberrant RA patterns.

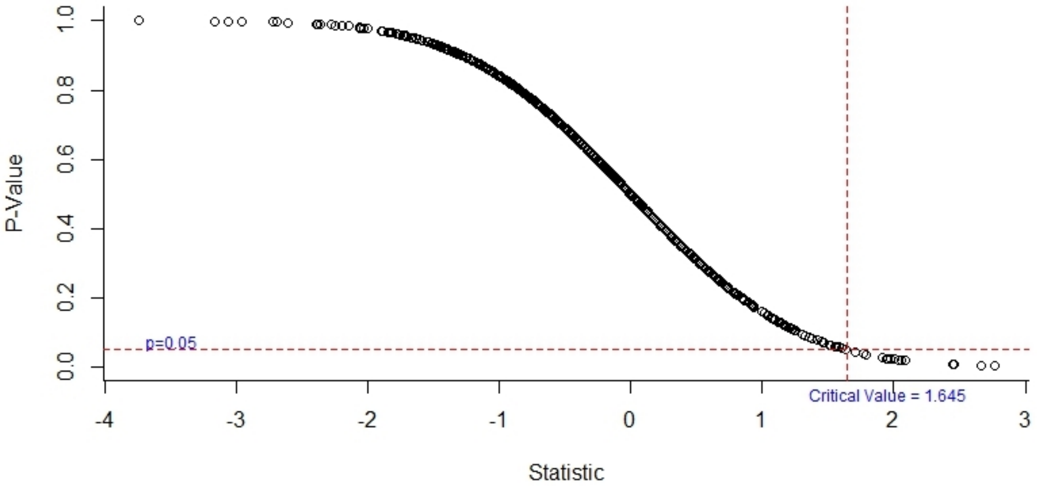

Figure 4. Estimated person-fit statistics l_s^y with respect to RA patterns plotted against the corresponding posterior significance probability.

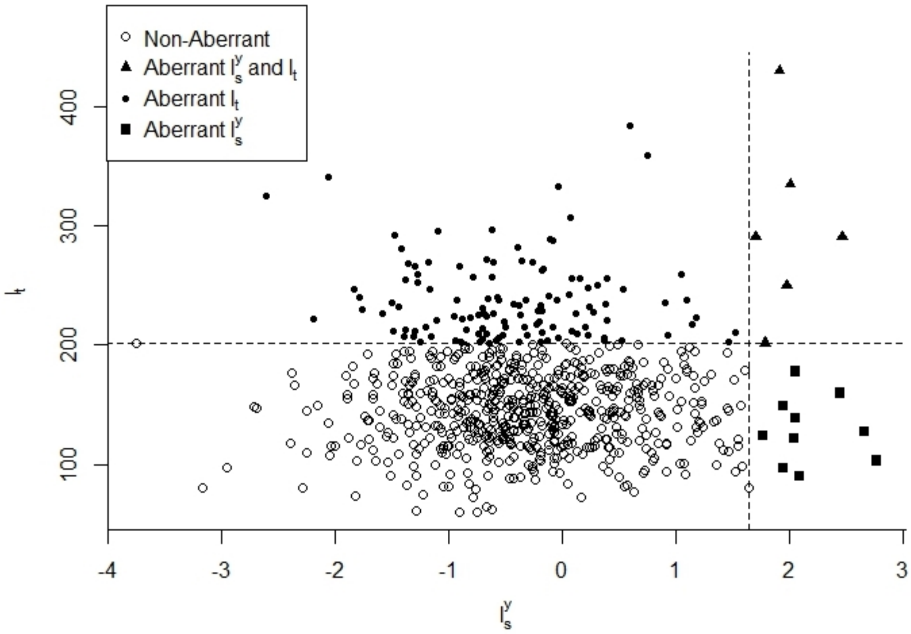

Figure 5. Person-fit statistic l_t (related to RT) plotted against l_s^y (related to RA).

In Figure 5, the person-fit statistic l_s^y *(x-axis)* is plotted against the person-fit statistic l_t (y-axis). For both statistics, the threshold value of the significant area is marked with a dotted line. It can be seen that with respect to aberrant RT patterns, l_t, a serious number of test takers are marked as aberrant, since their value is above the threshold of 201.4. A few test takers are marked as aberrant with respect to their RA pattern, since their statistic value is above 1.645. Those marked as aberrant with respect to their RA and RT pattern are represented by a triangle. Only 6 test takers are marked as aberrant for both patterns.

The plotted statistic scores concerning RT and RA patterns do not seem to be related. In theory, this relationship is possible, since in the computation of the person-fit statistics structural relationships between parameters are taken into account. Therefore, it would be possible that differences between aberrant and non-aberrant patterns are explained by a relationship between speed and ability or by a relationship between item characteristics.

The speed-accuracy trade-off in the population was investigated by plotting the estimated ability against the speed values. In Figure 6, the relationship between speed and ability for the identified non-aberrant and aberrant test takers is plotted. An aberrant group of 124 test takers was identified according to the l_t (significance level of .05). It can be seen that both groups show a comparable positive correlation between speed and ability. Three regression lines of ability on speed are represented in Figure 6, a dotted line for the aberrant test takers, a dashed line for the non-aberrant test takers, and a straight line for all test takers. It can be seen that the correlation between speed and ability is just slightly smaller for the aberrant test takers. The aberrant RT patterns do not strongly influence the estimated relationship between speed and ability.

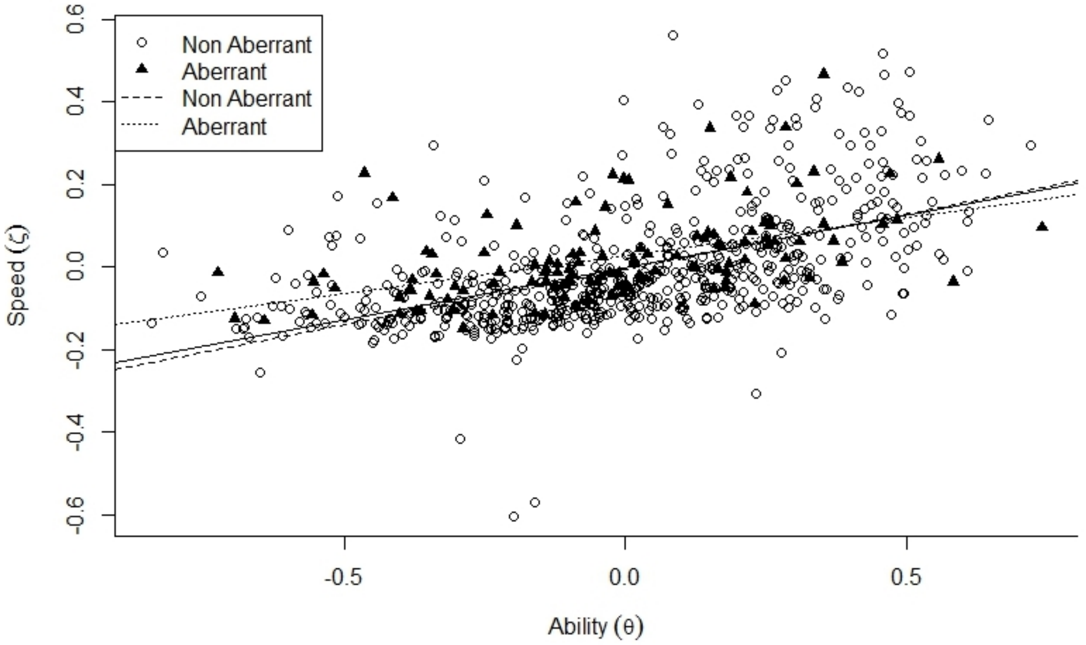

Figure 6. Estimated level of ability plotted against speed for the identified non-aberrant and aberrant test takers according to the l_t .

In Table 2, the covariance and correlation estimates are given of the population parameters of the joint model under the label “LNIRT”. For all test takers, it can be seen that the estimated correlation between ability and speed, when speed is constant, is around .486. The positive correlation indicates that the high-ability test takers worked faster than the low-ability test takers. The variation in speed values across test takers is around .022, which is rather small, since the variation in time intensities is .10 and almost 5 times larger. The test takers’ speed values range from -0.60 to 0.56. Most of the variation between RTs is explained by the differences in time intensities.

There exists a high correlation between item discrimination and time discrimination, and item difficulty and time intensity, around .501 and .464, respectively. This means that the discriminating items with respect to ability also discriminate well with respect to speed. The positive relation between the item difficulty and time intensity means that the time-intensive items are the more difficult items.

TABLE 2

Covariance and correlation estimates of person and item population parameters of the joint model (LNIRT).

Variance Components		LNIRT		
		Mean	SD	Cor.
Person covariance matrix				
Ability	σ_θ^2	0.093	0.006	
	$\rho_{\theta\zeta_0}$	0.022	0.002	0.486
Speed	$\sigma_{\zeta_0}^2$	0.022	0.001	
Item Covariance Matrix				
Discrimination	Σ_{11}	0.287	0.041	
	Σ_{12}	-0.062	0.022	-0.236
	Σ_{13}	0.091	0.018	0.501
	Σ_{14}	0.013	0.014	0.076
Difficulty	Σ_{22}	0.24	0.027	
	Σ_{23}	-0.07	0.015	-0.421
	Σ_{24}	0.073	0.013	0.464
Time Discrimination	Σ_{33}	0.115	0.016	
	Σ_{34}	-0.035	0.01	-0.322
Time Intensity	Σ_{44}	0.103	0.011	

Discussion

The joint modeling of RA and RT data can be used to make inferences about ability and speed given educational test data. The joint model receives more and more attention due to the increase in computer-based testing. To make correct inferences from the joint model, statistical tests have been developed to evaluate the model fit.

Person-fit tests have been developed to identify aberrant RA and/or RT patterns under the joint model. A Bayesian significance testing procedure has been proposed to evaluate the extremeness of computed person-fit statistics. The posterior probability of the extremeness of an RA pattern is used to decide whether a pattern should be flagged as aberrant. The developed person-fit statistics can be used to identify aberrant test takers with respect to their RT pattern, their RA pattern, or both of their patterns.

It was shown in a simulation study that under different conditions aberrant test takers were correctly identified. Two different types of aberrant behavior were considered in this study, but a more comprehensive simulation study is needed to fully investigate the performance of the proposed person-fit tests for joint models.

For RA data missing at random, missing data can be generated under the model, and the fit statistics can be applied to the complete response patterns. When the missing data are not missing at random, the person-fit statistics can only be applied to the observed patterns. However, missing responses to items which are not reached, might be related to aberrant behavior. This should be taken into account, when making inferences about response behavior. Missing data by design (i.e., non-administered items) can be ignored in the person-fit analysis.

From a theoretical point of view, a re-analysis of the test results using the tools for the joint model can provide more insight in the test characteristics. In some test settings, the person-fit statistics can be the only tool to identify (statistical) irregularities, for instance, when physical activities of the test takers cannot be observed. The person-fit statistics do not directly provide information about the type of aberrant behaviour. However, they can provide insight in the prevalence of irregular behaviour and identify items with a high level of irregular RA and RT data. Furthermore, the effects of time limits can be investigated. For instance, when running out of time, test takers might change their current strategy to work faster or adapt their speed of working due to fatigue (Fox & Marianti, 2016).

In practice, aberrant response behavior can seriously diminish the validity of the test results and affect test results of other test takers. As argued by Sinharay and Johnson (2016), test companies and programs require advanced technology such as video surveillance, but also seating charts and follow-up interviews, to prevent and detect inappropriate behavior of test takers. They should support test integrity and actively prevent and detect fraudulent or deceptive response behavior, since the test results can have important consequences for test takers. The person-fit tests can be used to detect inappropriate behavior by identifying patterns, which show irregularities and/or extreme RA and/or RT data.

Although statistical evidence is useful, it is not sufficient to conclude that fraudulent (response) behavior occurred. Statistical evidence should be complemented with other sources of information to obtain conclusive evidence of fraudulent behavior. However, it can also be concluded that a final score could not be computed due to irregularities in the RA and/or RT pattern. The test taker can be asked to retake the test or to provide additional information (e.g., follow-up interview) such that his/her test results are reconsidered for scoring (Sinharay & Johnson, 2016; van der Linden & Jeon, 2012).

References

Albert, J. H., & 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.

Bolsinova, M. & Maris, G. (2016). A test for conditional independence between response time and accuracy. *British Journal for Mathematical and Statistical Psychology, 69, 62-79*. DOI: 10.1111/bmsp.12059.

Béguin, A. A., & Glas, C. A. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, *66*(4), 541-561. DOI: 10.1111/bmsp.12059.

Chang, Y. W., Tsai, R. C., & Hsu, N. J. (2014). A speeded item response model:

Leave the harder till later. *Psychometrika, 79(2)*, 255-274. DOI: 10.1007/S11336-007-9031-2.

Cizek, G. J., Wollack, J. A. (2016). Exploring cheating on tests: The context, the concern, and the challenge. In G. Cizek & J. A. Wollack (Eds.), *Handbook of Detecting Cheating on Tests.* New York, NY: Routeledge.

De Boeck, P., & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software*, *48*(1), 1-28.

Dimitrov, D. M., & Smith, R. M. (2006). Adjusted Rasch person-fit statistics. *Journal of Applied measurement*, *7(2)*, 170-183.

Drasgow, F., Levine, M. V., & Williams, E. A. (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, J.-P. (2010). *Bayesian Item Response Modeling: Theory and Applications*. New York: Springer. DOI: 10.1007/978-1-4419-0742-4

Fox, J.-P., Klein Entink, R. H., & Linden, W. J. (2007). Modeling of responses and response times with the package CIRT. *Journal of Statistical Software*, 20(7), 1-14.

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

Glas, C. A. W., & Meijer, R. R. (2003). A Bayesian approach to person fit analysis in item response theory models. *Applied Psychological Measurement*, *27(3)*, 217-233. DOI: 10.1177/0146621603027003003.

Goldhammer, F. (2015). Measuring ability, speed, or both? Challenges, psychometric solutions, and what can be gained from experimental control. *Measurement, 13(3-4)*, 133–164. DOI: 10.1080/15366367.2015.1100020.

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, R. H., Fox, J.-P., & van der Linden, W. J. (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.

Klein Entink, R. H., Kuhn, J. T., Hornke, L. F., & Fox, J.-P. (2009). Evaluating cognitive theory: A joint modeling approach using responses and response times. *Psychological Methods*, *14(1)*, 54. DOI: 10.1037/a0014877.

Lee, Y.-H., & Chen, H. (2011). A review of recent response-time analysis in educational testing. *Psychological Test and Assessment Modeling, 53(3)*, 359-379.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational and Behavioral Statistics*, *4(4)*, 269-290. DOI: 10.3102/10769986004004269.

Marianti, S., Fox, J.-P., Avetisyan, M., Veldkamp, B. P., & 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.

Meng, X.-B., Tao, J., & Chang, H.-H. (2015). A conditional modeling approach for locally dependent item responses and response times. *Journal of Educational Measurement, 52*, 1-27. DOI: 10.1111/jedm.12060.

Meijer, R. R. (1996). Person-fit research: An introduction. *Applied Measurement in Education*, *9*(1), 3-8. DOI:10.1207/s15324818ame0901_2.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25*(2), 107-135. DOI: 10.1177/01466210122031957.

Molenaar, D., Tuerlinckx, F., & Maas, H. L. (2015). A generalized linear factor model approach to the hierarchical framework for responses and response times. *British Journal of Mathematical and Statistical Psychology*, *68*(2), 197-219. DOI:10.1111/bmsp.12042.

Partchev, I., & De Boeck, P. (2012). Can fast and slow intelligence be differentiated? *Intelligence*, *40*(1), 23-32. DOI:10.1016/j.intell.2011.11.002.

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence Diagnosis and Output Analysis for MCMC, *R News*, *vol. 6*, 7-11.

Ranger, J., & Kuhn, J. T. (2012). A flexible latent trait model for response times in tests. *Psychometrika*, *77*(1), 31-47. DOI: 10.1007/S11336-011-9231-7.

Rupp, A. A. (2013). A systematic review of the methodology for person fit research in item response theory: Lessons about generalizability of inferences from the design of simulation studies. *Psychological Test and Assessment Modeling, 55(1)*, 3–38.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement, 34(3)*, 213-232.

Sinharay, S., & Johnson, M. S. (2016). Three new methods for analysis of answer changes. *Educational and Psychological Measurement*. DOI: 10.1177/0013164416632287.

St-Onge, C., Valois, P., Abdous, B., & Germain, S. (2011). Accuracy of person-fit statistics: A Monte Carlo study of the influence of aberrance rates. *Applied Psychological Measurement*, *35(6)*,419-432. DOI: 10.1177/0146621610391777.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics, 31,* 181-204.

van der Linden, W. J. (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, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement*, *46(3)*, 247-272. DOI: 10.1111/j.1745-3984.2009.00080.x.

van der Linden, W. J. (2011). Modeling response times with latent variables: Principles and applications. *Psychological Test and Assessment Modeling, 53,* 334-358.

van der Linden, W. J. & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika, 73*, 365-384. DOI:10.1007/s11336-007-9046-8.

van der Linden, W. J., & Jeon, M. (2012). Modeling answer changes on test items. *Journal of Educational and Behavioral Statistics, 37*, 180-199. DOI:10.3102/1076998610396899.

van der Linden, W. J., & Lewis, C. (2015). Bayesian checks on cheating on tests. *Psychometrika, 80*, 689-706. DOI:10.1007/s11336-014-9409-x.