PSYCHOMETRIKA—VOL. 84, NO. 4, 1018–1046
DECEMBER 2019
https://doi.org/10.1007/s11336-019-09682-5

MODELING DIFFERENCES BETWEEN RESPONSE TIMES OF CORRECT AND INCORRECT RESPONSES

MARIA BOLSINOVA

ACTNEXT

JESPER TIJMSTRA

TILBURG UNIVERSITY

While standard joint models for response time and accuracy commonly assume the relationship between response time and accuracy to be fully explained by the latent variables of the model, this assumption of conditional independence is often violated in practice. If such violations are present, taking these residual dependencies between response time and accuracy into account may both improve the fit of the model to the data and improve our understanding of the response processes that led to the observed responses. In this paper, we propose a framework for the joint modeling of response time and accuracy data that allows for differences in the processes leading to correct and incorrect responses. Extensions of the standard hierarchical model (van der Linden in Psychometrika 72:287–308, 2007. https://doi.org/10.1007/s11336-006-1478-z) are considered that allow some or all item parameters in the measurement model of speed to differ depending on whether a correct or an incorrect response was obtained. The framework also allows one to consider models that include two speed latent variables, which explain the patterns observed in the responses times of correct and of incorrect responses, respectively. Model selection procedures are proposed and evaluated based on a simulation study, and a simulation study investigating parameter recovery is presented. An application of the modeling framework to empirical data from international large-scale assessment is considered to illustrate the relevance of modeling possible differences between the processes leading to correct and incorrect responses.

Key words: conditional dependence, hierarchical model, joint modeling, response times.

1. Introduction

With the advance of computerized tests, it has become common for test administrators to not just record the accuracy of the responses provided to the items, but also their response times (RTs). The benefit of considering RT in addition to response accuracy (RA) in the context of ability measurement can generally be considered to be twofold: RTs may provide collateral information for the estimation of ability (van der Linden, Klein Entink, & Fox, 2010), and RTs may also shed further light on the cognitive processes that led to the observed response (van der Maas & Jansen, 2003; Partchev & De Boeck, 2012). As such, the question of how to model RT and RA data has received a lot of attention in the recent psychometric literature. Most authors currently agree that the most appropriate way of modeling RT and RA in educational tests is the joint modeling of the two outcome variables of the response process. This allows one to treat both RA and RT as the outcome of stochastic processes, which on the person side can be explained by latent variables, for which a joint distribution can be considered. The joint distribution of RAs and RTs is typically modeled conditional on the latent ability and speed variables:

$$f(\mathbf{X}=\mathbf{x}, \mathbf{T}=\mathbf{t} \mid \boldsymbol{\Theta}=\boldsymbol{\theta}, \mathbf{H}=\boldsymbol{\eta}), \tag{1}$$

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11336-019-09682-5) contains supplementary material, which is available to authorized users.

Correspondence should be made to Maria Bolsinova, ACTNext, 500 ACT dr., Iowa City, IA 52243, USA. Email: maria.bolsinova@act.org

where $\mathbf{X}$ is a random vector of responses with realizations $x = 1$ if it is correct and $x = 0$ if it is incorrect for each element $X_i, \forall i \in [1 : K]$, $\mathbf{T}$ is a random vector of RTs with realizations $\mathbf{t}$, and $\mathbf{\Theta}$ and $\mathbf{H}$ are random vectors of latent variables that are commonly referred to as representing ability and speed, with realizations $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$, respectively. It should be noted that $\mathbf{H}$ captures *response speed* rather than cognitive speed, since it is a vector of random effects explaining patterns in the RTs and the cognitive process itself is not observed.[1] For generality, here $\mathbf{\Theta}$ and $\mathbf{H}$ are vectors that may have more than one element, although in practice often a single latent variable is considered for $\mathbf{\Theta}$ and for $\mathbf{H}$. For notational convenience throughout the paper, we will use $f(\mathbf{x}, \mathbf{t} \mid \boldsymbol{\theta}, \boldsymbol{\eta})$ instead of (1) for the joint distribution of the vectors of RAs and RTs, and $f(x_i, t_i \mid \boldsymbol{\theta}, \boldsymbol{\eta})$ for the joint distribution of the RA and the RT on a particular item i.

In most latent variable models used in educational measurement, the observed variables are usually assumed to be independent of each other given the latent variable(s) in the model. In the context of models that only consider RA, such as standard item response theory (IRT) models, this is captured by the assumption of local independence (see, e.g., Lord & Novick, 1968), stating that the RA on different items is independent of each other given the ability latent variable(s). Similarly, in models that only consider RT, one commonly takes the RTs of different items to be independent given the speed latent variable (van der Linden, 2009). When considering both RA and RT data, one additionally needs to address the question of how for each item the two types of outcomes are related. A statistically and conceptually attractive answer is to assume conditional independence (CI) of the outcome variables of different types given the latent variables in their measurement models, similar to the assumptions of local independence made by standard RA- or RT-only models. Mathematically, the assumption of CI of RA and RT can be presented as:

$$f(x_i, t_i \mid \boldsymbol{\theta}, \boldsymbol{\eta}) = f(x_i \mid \boldsymbol{\theta}, \boldsymbol{\eta}) f(t_i \mid \boldsymbol{\theta}, \boldsymbol{\eta}). \tag{2}$$

Arguably, the most commonly considered approach for the joint modeling of RT and RA is the hierarchical model (HM) proposed by van der Linden (2007), which assumes CI of RA and RT. The HM specifies separate measurement models for ability and speed, which results in a simple structure: On the person side, RA only depends on the ability latent variable(s), while RT only depends on the speed latent variable(s). The model connects the two outcome variables on the higher level: An observed nonzero correlation between the RTs and RAs on a single item is assumed to be explained by a correlation between speed and ability, while an observed nonzero correlation between the RTs and RAs of the same person to different items is assumed to be explained by the correlation between the item parameters in the two measurement models, usually referred to as item difficulty (the higher it is the lower the proportion of correct responses to the item) and item time intensity (the higher it is the more time on average respondents spend on an item). Consequently, the model assumes that once the person and item parameters are taken into account, RT and RA are independent.

The hierarchical modeling framework provides one with a flexible approach for modeling RA and RT, in the sense that it allows one to consider a variety of measurement models for both ability and speed. In this paper, we consider a rather simple version of the HM in which the two-parameter normal-ogive (2PNO) model (Birnbaum, 1968) is used as the measurement model for ability and the one-factor model for log-RTs is used as the measurement model for speed (Klein Entink, Fox, & van der Linden, 2009). Alternatively, for example, the three-parameter normal-ogive model (Klein Entink, Fox, & van der Linden, 2009), logistic IRT models (Bolsinova, De Boeck, & Tijmstra, 2017), and cognitive diagnostic models (Zhan, Jiao, & Liao, 2017) have been

[1] Since these models only consider observed RT, the latent variable in the measurement model for RT only captures response speed: The degree to which a person displays the tendency to provide responses quickly rather than slowly. This tendency should not be equated with cognitive speed (however defined exactly).

used as measurement models for ability within the HM. As alternatives to the linear factor model for log-RTs, for example, a Box-Cox transformation can be used (Klein Entink, van der Linden, & Fox, 2009) or a Weibull distribution for RTs (Rouder, Sun, Speckman, Lu, & Zhou, 2003).

Under the considered specification of the HM, the joint distribution of the RA and RT for item i conditional on the latent variables is:

$$f(x_i, t_i \mid \theta, \eta) = \Phi(\alpha_i\theta + \beta_i)^{x_i}(1 - \Phi(\alpha_i\theta + \beta_i))^{1-x_i} \ln\mathcal{N}(t_i; \xi_i - \lambda_i\eta, \sigma_i^2), \tag{3}$$

where α_i and β_i are the slope and the intercept of the item characteristic curve, and $\Phi(\cdot)$ denotes the cumulative standard normal distribution function, and where RT is lognormally distributed with the mean parameter equal to the difference between the item's time intensity, denoted by ξ_i, and the speed latent variable weighted by the item factor loading, denoted by λ_i, and where the residual variance is denoted by σ_i^2. The ability and speed latent variables are assumed to be random effects, and their joint distribution is assumed to be a bivariate normal. For identification, the mean vector of the latent variables is constrained to zero, and their variances are constrained to one.

While statistically appealing, in practice it is often not realistic to assume that after taking ability and speed into account the accuracy of a response is independent of the time it took to provide that response (Ranger & Ortner, 2012; Bolsinova & Tijmstra, 2016). Residual associations between RA and RT are likely to occur in many applications for a wide variety of possible reasons, an overview of which has been provided by Bolsinova, Tijmstra, De Boeck, and Molenaar (2017b). For example, respondents may speed up during the test, or may show a temporary lapse in concentration, which may result in negative or positive residual associations between RA and RT, respectively. Differential item functioning may affect both the difficulty and time intensity of an item, which can also create dependencies. Additionally, respondents may differ in their problem solving strategies, also resulting in possible residual associations between RA and RT. Thus, it may often be necessary to acknowledge that there are likely to be dependencies between RA and RT that cannot be explained by the person and item parameters in the HM. However, one can consider extending the HM to take these dependencies into account with the aim of increasing both model fit and our understanding of the relationship between RT and RA and of the underlying response processes (Bolsinova, Tijmstra, Molenaar, & De Boeck, 2017b).

Different approaches can be taken when extending the hierarchical modeling framework for RT and RA to allow for conditional dependence (CD) between the outcome variables: (1) the joint distribution of RT and RA to the same item can be modeled as a bivariate distribution with a nonzero dependence parameter; (2) the joint distribution of RT and RA can be factorized as the product of a marginal distribution of RT and a conditional distribution of RA given RT, and (3) the joint distribution can be factorized as the product of the marginal distribution of RA and the conditional distribution of RT given RA. While differing in structure, each of these three approaches abandons the assumption that RA and RT on the same item are independent conditional on the latent variables in the model and allow for more complex associations between RA and RT than are possible under the standard HM.[2]

The first approach was proposed by Ranger and Ortner (2012) who modeled the joint distribution of a transformation of RA and RT as a bivariate normal:

$$f(x_i^*, t_i^* \mid \theta, \eta) = \mathcal{N}_2\left(\begin{bmatrix}\alpha_i\theta + \beta_i\\ \xi_i - \lambda_i\eta\end{bmatrix}, \begin{bmatrix}1 & \rho_i\sigma_i\\ \rho_i\sigma_i & \sigma_i^2\end{bmatrix}\right). \tag{4}$$

[2] For clarification it may be relevant to point out that each of these approaches models conditional dependence between RA and RT, which may arise due to many sources and need not be reducible or even linked to a speed–accuracy trade-off (Bolsinova, Tijmstra, Molenaar, & De Boeck, 2017b). Hence, none of these approaches specifically attempt to model the speed–accuracy trade-off.

where x_i^* is the underlying continuous response ($x_i = \mathcal{I}(x_i^* > 0)$) and t_i^* is log-RT, and where the marginal distributions of RA and RT, respectively, are the 2PNO model and the lognormal model, the same as in the HM presented earlier. Here, the CD between RT and RA is quantified by the conditional correlation parameter ρ_i, which varies across items. Meng, Tao and Cheng (2015) have further extended this model to allow the conditional correlation to vary not only across items, but also across persons.

Under the second approach, one models the joint distribution of RT and RA as follows:

$$f(x_i, t_i \mid \boldsymbol{\theta}, \boldsymbol{\eta}) = f(t_i \mid \boldsymbol{\theta}, \boldsymbol{\eta}) f(x_i \mid t_i, \boldsymbol{\theta}, \boldsymbol{\eta}). \tag{5}$$

This approach has been explored by Bolsinova, Tijmstra and Molenaar (2017a). In their modeling framework, the marginal distribution of RT is assumed to follow the lognormal model (van der Linden, 2006), and the conditional model for RA given RT is a 2PNO model with the intercept being a linear function of the standardized difference between the observed and expected log-RT:

$$f(x_i \mid t_i, \theta, \eta) = \Psi\left(\alpha_i\theta + \beta_{i0} + \beta_{i1}\frac{\ln t_i - (\xi_i - \eta)}{\sigma_i}; x_i\right), \tag{6}$$

where β_{i0} is the baseline intercept and β_{1i} is the linear effect of standardized residual log-RT on the intercept of the item characteristic curve, and $\Psi(\cdot; x_i) = \Phi(\cdot)^{x_i}(1 - \Phi(\cdot))^{1-x_i}$. Furthermore, the model has been extended to allow not only for between-item differences in CD, but also between-person differences in CD. Alternatively it has been proposed to include a linear effect on the intercept and on the log-transformed slope of the item characteristic curve to model CD (Bolsinova, De Boeck, & Tijmstra, 2017).

If one would adopt the third approach, one can specify the joint distribution of RA and RT as follows:

$$f(x_i, t_i \mid \boldsymbol{\theta}, \boldsymbol{\eta}) = f(x_i \mid \boldsymbol{\theta}, \boldsymbol{\eta}) f(t_i \mid x_i, \boldsymbol{\theta}, \boldsymbol{\eta}). \tag{7}$$

While the first two approaches have been extensively explored and various models have been proposed allowing for between-item and between-person differences in the relationship between RT and RA, this third approach has not received much attention. To our knowledge, the specification in Equation 7 has only been considered in the work of van der Linden and Glas (2010), who used a model with separate time intensity parameters for the correct and incorrect responses mainly as an alternative model to test the assumption of CI against, rather than considering the model as substantively relevant in its own right. Thus, the merit of this third approach for jointly modeling RA and RT has to our knowledge not been explored.

Adopting the third approach rather than one of the other two may be appealing if one considers that there are likely important qualitative differences between the response processes leading to a correct response versus those leading to an incorrect response. At the very least, the response processes differ in how they are terminated. That is, a correct response is likely the outcome of successfully following the intended solution strategy, while an incorrect response may have been produced for a variety of reasons, such as following the intended solution strategy unsuccessfully, following a different solution strategy than the one intended, giving up on the item after trying one's best, or failing to attempt to solve the item (e.g., skipping). To a large extent, one can (on well-designed tests without multiple choice) expect such between- and within-person differences in the employed response processes to be reflected in the correctness of the responses, with successfully following the intended solution strategy generally resulting in a correct response and following inappropriate solution strategies (including giving up or skipping) generally resulting in an incorrect response. This makes adopting the third approach to modeling CD intuitive and appealing, since it allows for different RT models for correct and incorrect responses, matching the fact that correct and incorrect responses are likely the result of different response processes. The other two approaches to modeling CD model are—by design—not equipped to capture such differences in the response processes between correct and incorrect responses, making it important to consider this third approach.

Adopting a modeling approach based on the specification in Equation 7 has the potential to bring relevant phenomena to light that cannot easily be studied when using the other two approaches. For example, it may be that RTs of correct responses show more structural patterns than those of incorrect responses, where the latter may have larger residual variances. Such differences may be expected if correct responses are thought to be the product of relatively homogeneous response processes (i.e., with everyone employing the intended solution strategy) while incorrect responses may have been the outcome of a more heterogeneous set of response processes (including skipping, giving up, and employing a wrong solution strategy). Similarly, it may be that the RTs of correct responses are more strongly related to the speed latent variable than the RTs of incorrect responses, which might also be plausible if incorrect responses are the product of a more heterogeneous set of response processes than correct responses. This would lead to a difference in factor loading of RT on speed depending on RA, which can be modeled under such an approach. This approach would also allow one to study whether the time intensity of items differs depending on RA, which allows one to take the possibility into account that for persons with the same ability and speed levels correct responses to an item are on average faster or slower than incorrect responses. It may be considered plausible that the expected RT of a person on an item differs depending on whether that person employs the intended solution strategy or whether the person employs an inadequate solution strategy, which would need to be accommodated by allowing correct and incorrect responses to an item to be modeled using separate time intensity parameters. These possible differences between the RTs correct and incorrect responses and the extent to which models of the first two approaches are unable to capture them will be explored when considering an empirical example further on in the manuscript.

While joint models for RA and RT have generally assumed a single latent variable to be sufficient for explaining the patterns in the RTs, it may be that there are structural differences between the response speed of correct versus incorrect responses: Some respondents may spend a relatively long time on items with which they have difficulties, while others may spend little time on items that they expect to fail and dedicate most of their time to solving items that they expect to be able to solve. In such cases the RTs cannot fully be explained by a single latent variable, but rather two different latent variables are needed to explain the patterns in the RTs. Adopting an approach in line with Equation 7 allows one to study whether the speed with which correct responses are given differs from the speed with which incorrect responses are given, which may suggest qualitative differences between persons in how they allocate their time. If this is the case, it may, for example, be worthwhile to study which type of respondent has a higher response speed for correct responses than for incorrect responses, and how this difference relates to their ability level. Furthermore, if two speed latent variables are needed, the question arises how they relate to each other, and whether their relationship to the ability latent variable(s) is the same. The fact that response speed for correct responses may not be linked to ability in the same way as the response speed for incorrect responses has already been recognized before in the literature. For example, in a test of numerical reasoning Semmes, Davidson and Close (2011) examined correlations between ability and median RT separately for correct and incorrect responses and found that while there was no correlation for correct RTs, there was a positive correlation for incorrect RTs. Another example comes from the Amsterdam Chess Test, where van der Maas and Wagenmakers (2005) found that ability is negatively correlated with the average RT of correct responses, but is not correlated with the average RT of incorrect responses. Hence, there is empirical support for the need of considering separate measures of response speed for correct and incorrect responses. Such issues cannot be directly addressed using models in line with the first two approaches, but require one to consider a modeling framework that makes use of the modeling structure in Equation 7.

In this paper, we propose a modeling framework in line with the third approach (Equation 7), which allows one to work with different RT models for correct and incorrect responses. The modeling framework considers extensions of the HM, where model parameters are allowed to differ depending on the RA. Both models with a single speed latent variable and models with separate speed latent variables for correct and incorrect responses are considered. Using this modeling framework should allow practitioners to test whether there are qualitative differences between the processes leading to correct and incorrect responses, to gain a more complete understanding of these response processes, and to obtain a more complete picture of both the respondents that take the test and the items that the test consists of.

The remainder of the paper is organized as follows. In Section 2, we present a general framework for modeling differences in the distribution of RT for correct and incorrect responses, for which twelve partially nested models are proposed. An estimation procedure is presented, and different model comparison tools are considered. In Section 3, we present two simulation studies, which evaluate the parameter recovery of the procedure and the effectiveness of the different model comparison tools, respectively. Section 4 presents an application of the procedure to empirical data from international large-scale assessment, which illustrates the relevance of the proposed framework in practice. Section 5 presents an additional simulation study which evaluates parameter recovery of the model parameters when their true values are the same as the estimates in the empirical example. The paper concludes with a discussion.

2. Modeling the Differences Between Correct and Incorrect Responses

Equation 7 is general in form and leaves it open which models are considered for $f(x_i \mid \boldsymbol{\theta}, \boldsymbol{\eta})$ and $f(t_i \mid x_i, \boldsymbol{\theta}, \boldsymbol{\eta})$, and how exactly the dependence of t_i on x_i is specified. For the marginal distribution of RA, we propose using the same IRT model as in the standard HM assuming CI presented in Equation 3, that is, a 2PNO model. For the conditional distribution of RT, we take the one-factor model for log-RTs in Equation 3 as a starting point, but allow the model parameters and the speed latent variable to potentially depend on the RA:

$$f(t_i \mid x_i, \boldsymbol{\theta}, \boldsymbol{\eta}) = f(t_i \mid x_i, \eta_0, \eta_1) = \ln\mathcal{N}(t_i; \xi_{ix_i} - \lambda_{ix_i}\eta_{x_i}, \sigma^2_{ix_i}), \tag{8}$$

where η_0 and η_1 denote speed latent variables for incorrect and for correct responses, respectively; $\{\xi_{i0}, \lambda_{i0}, \sigma^2_{i0}\}$ and $\{\xi_{i1}, \lambda_{i1}, \sigma^2_{i1}\}$ are the time intensity, factor loading and residual variance of item i for incorrect and correct responses, respectively.

In addition to the full model in Equation 8 in which the latent variable and the three item parameters differ across correct and incorrect responses, one can also consider constrained versions of the model, in which some of the parameters are equal for different values of RA. In the context of modeling possible differences in the RTs between correct and incorrect responses, we propose to consider twelve models in line with Equation 8: six models with a single speed latent variable, and six models with two speed latent variables. Both for the two-dimensional (i.e., one ability and one speed latent variable) and three-dimensional models (i.e., one ability and two speed latent variables), in addition to the models in which all three item parameters depend on RA ($\mathcal{M}_{4a}$) and in which all three item parameters do not differ for correct and incorrect responses ($\mathcal{M}_1$), models with the following constraints are considered: $\lambda_{i0} = \lambda_{i1}$ and $\sigma^2_{i0} = \sigma^2_{i1}$ ($\mathcal{M}_2$); 2) $\xi_{i0} = \xi_{i1}$ and $\lambda_{i0} = \lambda_{i1}$ ($\mathcal{M}_{3b}$); 3) $\sigma^2_{i0} = \sigma^2_{i1}$ ($\mathcal{M}_{3a}$); and 4) $\lambda_{i0} = \lambda_{i1}$ ($\mathcal{M}_{4b}$). Here, we do not consider constrained models in which factor loadings are different while the time intensities are

Table 1.
Overview of the models obtained when placing equality constraints on some of the item parameters in Equation 8

		$\sigma^2_{i0} = \sigma^2_{i1}$	$\sigma^2_{i0} \neq \sigma^2_{i1}$
$\xi_{i0} = \xi_{i1}$	$\lambda_{i0} = \lambda_{i1}$	$\mathcal{M}_1$	$\mathcal{M}_{3b}$
	$\lambda_{i0} \neq \lambda_{i1}$	-	-
$\xi_{i0} \neq \xi_{i1}$	$\lambda_{i0} = \lambda_{i1}$	$\mathcal{M}_2$	$\mathcal{M}_{4b}$
	$\lambda_{i0} \neq \lambda_{i1}$	$\mathcal{M}_{3a}$	$\mathcal{M}_{4a}$

Note: Model labels are assigned in line with the model selection steps presented in Figure 1.

the same, because the difference in factor loadings can be seen as an interaction effect between RA and speed on log-RT and it may not be desirable to include an interaction effect in the model without including the main effect (i.e., a difference in time intensities). It may be noted that $\mathcal{M}_1$ specifies a CI model in line with the standard HM, if one would consider a $\mathcal{M}_1$ with a single speed latent variable. The model extension of the HM used by van der Linden and Glas (2010) to test for CI matches $\mathcal{M}_2$ if a single speed latent variable is used.[3] An overview of the six types of models that are considered for a given dimensionality is presented in Table 1.

In any statistical model that considers multiple latent variables, a joint distribution for the latent variables need to be specified. When only one speed latent variable is used, $\{\theta, \eta\}$ are assumed to have a bivariate normal distribution with a zero mean vector $\boldsymbol{\mu}$ and a 2×2 covariance matrix $\boldsymbol{\Sigma}$ in which the variances are constrained to 1. When two speed latent variables are used, $\{\theta, \eta_0, \eta_1\}$ are assumed to have a multivariate normal distribution with a mean vector $\boldsymbol{\mu}$ and a 3×3 covariance matrix $\boldsymbol{\Sigma}$. The first two elements of $\boldsymbol{\mu}$ and the first two variances of $\boldsymbol{\Sigma}$ are constrained to 0 and 1, respectively, for all six models. $\mu_3 \equiv 0$ only when $\xi_{i0} \neq \xi_{i1}$, and $\Sigma_{3,3} \equiv 1$ only when $\lambda_{i0} \neq \lambda_{i1}$.

2.1. Estimation

The proposed models for the differences between RTs of correct and incorrect responses can be estimated by sampling from the joint posterior distribution of the model parameters, which is proportional to the product of the density of the data and the prior distribution. Averages of the sampled values for each parameter can be used as an approximation of its posterior mean, which can be used as a point estimate of the parameter. The 2.5% and 97.5% percentiles of the sampled values can be used as approximations of the 95% credible interval, which can be used to quantify the uncertainty about the parameters.

The density of the RT and the RA data for the two-dimensional model is the following:

$$f(\mathbf{x}, \mathbf{t} \,|\, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\sigma}^2, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{p=1}^{N} \int \int \prod_{i=1}^{K} \Psi\left(\alpha_i \theta + \beta_i ; x_{pi}\right) \ln\mathcal{N}\left(t_{pi} ; \xi_{ix_i} - \lambda_{ix_i}\eta_{x_i}, \sigma^2_{ix_i}\right) \mathcal{N}_2(\theta, \eta; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \mathrm{d}\theta \mathrm{d}\eta, \tag{9}$$

[3] Note that van der Linden and Glas (2010) considered a constrained version of the one-factor model in which the factor loadings of all items are equal to 1, and the variance of the speed latent variable is freely estimated.

and for the three-dimensional model it is:

$$f(\mathbf{x},\mathbf{t}\,|\,\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi},\boldsymbol{\lambda},\boldsymbol{\sigma}^2,\boldsymbol{\mu},\boldsymbol{\Sigma}) \\ =\prod_{p=1}^{N}\int\int\int\prod_{i=1}^{K}\Psi\left(\alpha_i\theta+\beta_i;x_{pi}\right)\ln\mathcal{N}\left(t_{pi};\xi_{ix_i}-\lambda_{ix_i}\eta_{x_i},\sigma^2_{ix_i}\right)\mathcal{N}_3(\theta,\eta_0,\eta_1;\boldsymbol{\mu},\boldsymbol{\Sigma})\mathrm{d}\theta\mathrm{d}\eta_0\mathrm{d}\eta_1, \tag{10}$$

where $\mathbf{x}$ and $\mathbf{t}$ are the $N\times K$ matrices of RAs and RTs of N persons to K items, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are K-length vectors with item slopes and intercepts for all items, $\boldsymbol{\xi}$, $\boldsymbol{\lambda}$, and $\boldsymbol{\sigma}^2$ are the $K\times 2$ matrices of time intensities, factor loadings, and residual variances for incorrect (first column) and correct responses (second column) of all items.

For the item parameters we use independent semi-conjugate low-informative priors:

$$f(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi},\boldsymbol{\lambda},\boldsymbol{\sigma}^2) \\ =\prod_{i=1}^{K}\mathcal{N}(\alpha_i;0,100^2)\mathcal{N}(\beta_i;0,100^2)\prod_{k=\{0,1\}}\mathcal{N}(\xi_{ik};0,100^2)\mathcal{N}(\lambda_{ik};0,100^2)\mathcal{IG}(\sigma^2_{ik};0.001,0.001). \tag{11}$$

We decided to use independent low-informative priors because the use of strong prior information or hierarchical priors is not needed to achieve reasonable estimates of the item parameters (see Simulation Study 1, see also Molenaar, Tuerlinckx, & van der Maas, 2015b, 2015a). Alternatively, one may opt to include possible relationships between the item parameters in the model in the form of a hierarchical prior distribution.

While the mean vector and the covariance matrix are (partially) constrained, to improve convergence of the model we sample them freely but for each sample from the posterior rescale all the parameters such that the constraints of the mean vector and covariance matrix hold (see Supplementary materials for details). For the mean vector we use an improper prior ($\propto 1$) and for the covariance matrix we use an inverse-Wishart distribution with $d+2$ degrees of freedom, where d is the number of latent variables in the model, and $\mathbf{I}_d$ is the scale parameter. With $N>>d$, the posterior distribution is dominated by the data (Hoff, 2009).

For sampling from the joint posterior distribution of the model parameters, we use a Gibbs Sampler in which the model parameters are consecutively sampled from their full conditional posterior distributions. Data augmentation (Tanner & Wong, 1987) is implemented to simplify the conditional posteriors. First, for each person p person-specific θ_p, η_{p0} and η_{p1} are introduced ($\eta_{p0}=\eta_{p1}=\eta_p$ in the case of the two-dimensional models). Second, for each combination of a person and an item an augmented continuous response $y_{pi}\sim\mathcal{N}(y_{pi};\alpha_i\theta_p+\beta_i,1)$ is introduced (Albert, 1992) which is defined in such a way that $x_{pi}=\mathcal{I}(y_{pi}\geq 0)$.

To start the Gibbs Sampler, the model parameters need to be initialized. The person parameters are sampled from $\mathcal{N}(0,1)$. The item slopes and factor loadings are sampled from $\mathcal{N}(1,0.1^2)$, the item intercepts are sampled from $\mathcal{N}(0,0.1^2)$, the time intensities are sampled from $\mathcal{N}\left(\frac{1}{NK}\sum_p\sum_i\ln t_{pi},0.1^2\right)$ (i.e., the mean is equal to the average log-RT in the data), and the residual variances are sampled from $\mathcal{U}(0.1,0.3)$. $\mathbf{0}$ and $\mathbf{I}_d$ are used for the mean vector and covariance matrix of the person parameters. The augmented responses y_{pi}s do not need to be initialized because they are sampled in the first step of the sampler. The exact steps of the Gibbs Sampler are presented in the Supplementary materials.

2.2. Model Selection

The different models presented in Section 2 differ in their complexity and will likely differ in how well they are able to describe the data. To determine which model should be preferred for modeling the data, model selection criteria can be considered. Both selection procedures based on the Akaike information criterion (AIC) and on the Bayesian information criterion (BIC) are considered. For the computation of either of these measures, the log-likelihood of the data given the estimates of the model parameters needs to be obtained. As a Bayesian estimation procedure is considered, the log-likelihood will be evaluated at the posterior mean of the parameters rather than at their maximum likelihood estimate, meaning that a modified version of the AIC and BIC is considered: the mAIC and mBIC, respectively.[4] Computing the log-likelihood requires integrating over the latent variables, which here is done using Gauss–Hermite quadrature with 10 nodes in each dimension. The values of log-likelihood for the two-dimensional models ($\ln \mathcal{L}_{2\mathrm{dim}}$) for the three-dimensional models ($\ln \mathcal{L}_{3\mathrm{dim}}$) are approximated as follows:

$$\begin{aligned}&\ln \mathcal{L}_{2\mathrm{dim}}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\sigma}}^2, \hat{\Sigma}_{12}; \mathbf{x}, \mathbf{t})\\ &\quad\approx \sum_{p=1}^{N}\sum_{g=1}^{10}\sum_{h=1}^{10}\frac{w_g}{\sqrt{\pi}}\frac{w_h}{\sqrt{\pi}}\prod_{i=1}^{K}\Psi\left(\hat{\alpha}_i v_{ghj1} + \hat{\beta}_i; x_{pi}\right)\ln\mathcal{N}\left(t_{pi}; \hat{\xi}_{ix_i} - \hat{\lambda}_{ix_i} v_{ghj2}, \hat{\sigma}^2_{ix_i}\right);\end{aligned} \tag{12}$$

and

$$\begin{aligned}&\ln \mathcal{L}_{3\mathrm{dim}}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\sigma}}^2, \hat{\Sigma}_{12}, \hat{\Sigma}_{13}, \hat{\Sigma}_{23}, \hat{\Sigma}_{33}, \hat{\mu}_3; \mathbf{x}, \mathbf{t})\\ &\quad\approx \sum_{p=1}^{N}\sum_{g=1}^{10}\sum_{h=1}^{10}\sum_{j=1}^{10}\frac{w_g}{\sqrt{\pi}}\frac{w_h}{\sqrt{\pi}}\frac{w_j}{\sqrt{\pi}}\prod_{i=1}^{K}\Psi\left(\hat{\alpha}_i v_{ghj1} + \hat{\beta}_i; x_{pi}\right)\ln\mathcal{N}\left(t_{pi}; \hat{\xi}_{ix_i} - \hat{\lambda}_{ix_i} v_{ghj(2+x_i)}, \hat{\sigma}^2_{ix_i}\right),\end{aligned} \tag{13}$$

with

$$\begin{aligned}v_{ghj1} &= \sqrt{2}y_g,\\ v_{ghj2} &= \sqrt{2(1-\hat{\Sigma}^2_{12})}y_h + \hat{\Sigma}_{12}v_{ghj1},\\ v_{ghj3} &= \sqrt{2\left(\hat{\Sigma}_{33} - \frac{\hat{\Sigma}^2_{13}+\hat{\Sigma}^2_{23}-2\hat{\Sigma}_{12}\hat{\Sigma}_{13}\hat{\Sigma}_{23}}{1-\hat{\Sigma}^2_{12}}\right)}y_j + \hat{\mu}_3 + \frac{\hat{\Sigma}_{13}-\hat{\Sigma}_{12}\hat{\Sigma}_{23}}{1-\hat{\Sigma}^2_{12}}v_{ghj1} + \frac{\hat{\Sigma}_{23}-\hat{\Sigma}_{12}\hat{\Sigma}_{13}}{1-\hat{\Sigma}^2_{12}}v_{ghj2},\end{aligned} \tag{14}$$

where $\mathbf{y} = \{y_1, \ldots, y_{10}\}$ and $\mathbf{w} = \{w_1, \ldots, w_{10}\}$ are the quadrature nodes and weights, respectively.

The number of parameters for the models is determined as follows: On the item side for each of the models there are $5K$ parameters (αs, βs, ξs, λs, and σ^2s), in addition to that there are an extra K parameters for the models with $\xi_{i0} \neq \xi_{i1}$, an extra K parameters for the models with $\lambda_{i0} \neq \lambda_{i1}$, and an extra K parameters for the models with $\sigma^2_{i0} \neq \sigma^2_{i1}$; on the population side there is one covariance in the models with $\eta_0 = \eta_1$, and three covariances for the models with $\eta_0 \neq \eta_1$, and in addition to that there is one freely estimated mean for the models with $\eta_0 \neq \eta_1$ and $\xi_{i0} = \xi_{i1}$, and one freely estimated variance for the models with $\eta_0 \neq \eta_1$ and $\lambda_{i0} = \lambda_{i1}$.

[4] Since we do not include any strong prior information in the estimation of the parameters, the posterior means would be very close to the maximum likelihood estimates, meaning that the difference between both the AIC and mAIC and the BIC and mBIC can be expected to be minimal.

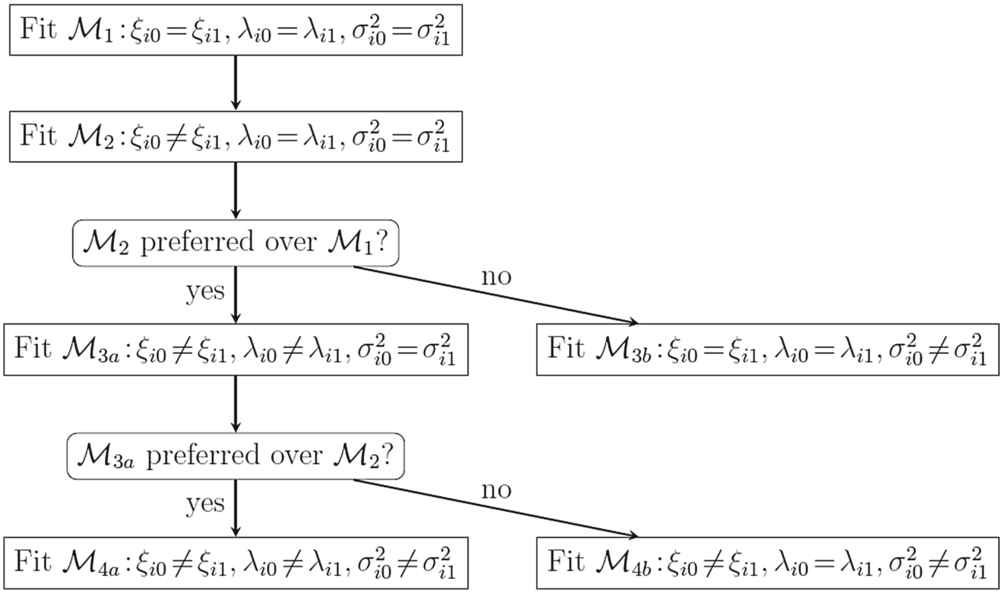
Fit $\mathcal{M}_1 : \xi_{i0} = \xi_{i1}, \lambda_{i0} = \lambda_{i1}, \sigma^2_{i0} = \sigma^2_{i1}$
Fit $\mathcal{M}_2 : \xi_{i0} \neq \xi_{i1}, \lambda_{i0} = \lambda_{i1}, \sigma^2_{i0} = \sigma^2_{i1}$
$\mathcal{M}_2$ preferred over $\mathcal{M}_1$?
yes
no
Fit $\mathcal{M}_{3a} : \xi_{i0} \neq \xi_{i1}, \lambda_{i0} \neq \lambda_{i1}, \sigma^2_{i0} = \sigma^2_{i1}$
Fit $\mathcal{M}_{3b} : \xi_{i0} = \xi_{i1}, \lambda_{i0} = \lambda_{i1}, \sigma^2_{i0} \neq \sigma^2_{i1}$
$\mathcal{M}_{3a}$ preferred over $\mathcal{M}_2$?
yes
no
Fit $\mathcal{M}_{4a} : \xi_{i0} \neq \xi_{i1}, \lambda_{i0} \neq \lambda_{i1}, \sigma^2_{i0} \neq \sigma^2_{i1}$
Fit $\mathcal{M}_{4b} : \xi_{i0} \neq \xi_{i1}, \lambda_{i0} = \lambda_{i1}, \sigma^2_{i0} \neq \sigma^2_{i1}$

FIGURE 1.
A scheme for the stepwise evaluation of the six models based on a model information criterion, considered separately for the set of two-dimensional models and the set of three-dimensional models. At any terminal node, the model that is selected from the considered set as the preferred model is the one with the lowest value for the information criterion.

While it is possible to fit all twelve models to the data, it may in practice be preferable to do stepwise model selection to reduce computation time. In line with other stepwise procedures such as those used in the measurement invariance testing literature (see, e.g., Van de Schoot, Lugtig, & Hox, 2012), it may be sensible to consider a selection scheme in which the stepwise relaxation of the item-parameter equality constraints in $\mathcal{M}_1$ is considered. A proposal for such a scheme is presented in Figure 1, which shows the steps in which item-parameter constraints can be relaxed. The scheme can be applied separately to the set of two-dimensional models and the set of three-dimensional models, resulting in a choice for the preferred two-dimensional model and preferred three-dimensional model.

While of these two models, one can opt to prefer the model with the lowest mAIC or mBIC, when comparing the two- and three-dimensional models the mAIC and mBIC might not work properly because the models differ in a parameter that in the two-dimensional model is restricted at the boundary of its parameter space (i.e., the correlation between η_0 and η_1 is restricted to be 1). Pilot studies (results not reported) showed that selecting models purely based on either the mAIC or the mBIC may lead one to favor a three-dimensional model even in cases where the true model only has one speed latent variable. In such cases solutions are generally found for the three-dimensional model where the two speed latent variables that are obtained are highly correlated, such that distinguishing between these two dimensions would not be practically relevant even if there would in fact have been two speed latent variables underlying the responses. To address this issue, we propose to add an additional procedure which evaluates whether the person-level heterogeneity in RTs of the correct versus incorrect responses can adequately be accounted for by the best two-dimensional model, or whether the best three-dimensional model is needed to properly account for such differences.

A posterior predictive check (X.-L. Meng, 1994; Gelman, Meng, & Stern, 1996) can be used to evaluate whether the best two-dimensional model adequately captures the relevant patterns in the data. As the statistic of interest, we propose using the correlation between persons' average log-RT of the correct responses and their average log-RT on the incorrect responses. This statistic is calculated for the observed data and for G replicated data sets generated using samples from the posterior distribution of the parameters of the best two-dimensional model. The proportion of data sets in which the replicated statistic is larger than the observed statistic is the approximation of the posterior predictive p-value. Our statistic of interest is designed such that posterior predictive p-values close to 1 indicate model misfit (i.e., the correlation predicted by the two-dimensional model is higher than the observed correlation). If the posterior predictive p-value is below a certain threshold (e.g., 0.95) the two-dimensional model can be seen as adequately capturing the person-specific heterogeneity of the RTs, and would therefore be selected as the best model overall. If the posterior predictive p-value is above the threshold, then the best three-dimensional model would be selected as the best model overall.

3. Simulation Studies

Two simulation studies were performed: One to evaluate parameter recovery, and one to evaluate the model selection procedures. In the first study, parameter recovery is evaluated for the two-dimensional and three-dimensional models in which all item parameters differ across correct and incorrect responses ($\mathcal{M}_{4a}$, Equation 8). In the second study, data are generated under all twelve models considered in this paper, to which the different model selection procedures described in Section 2.2 were applied.

3.1. *Simulation Study 1: Parameter Recovery*

3.1.1. Method For the two-dimensional and three-dimensional models, parameter recovery was evaluated under a baseline condition, as well as under extra conditions in which one factor was changed compared to the baseline condition. In the baseline condition, the sample size (N) was equal to 1000, the number of items (K) was equal to 16, the correlation(s) between the speed latent variable(s) and the ability latent variable (Σ_{12} and Σ_{13}), was (or were) equal to 0, and in the case of the three-dimensional model the correlation between the two speed latent variables (Σ_{23}) was equal to .7. This correlation was picked such that the two speed latent variables are quite strongly linked, but still only share roughly 50% of their variance, representing a situation where the response speed of correct and incorrect responses is strongly linked but not identical. Conditions with a sample size twice as large (2000) and twice as small (500) as in the baseline condition were considered. Additionally, a condition with 32 items instead of 16 items was added. Furthermore, a condition was used where a correlation of .5 between ability parameter and speed parameter(s) was used instead of the correlation of 0 in the baseline condition, representing a situation where response speed provides collateral information for the estimation of ability. Finally, for the three-dimensional model a condition with a larger correlation between the two speed latent variables (.9) was used to explore the performance of the procedure when faced with two different speed latent variables that are highly similar. Thus, in total five conditions were used for the two-dimensional models, and six conditions were used for the three-dimensional models (i.e., the effect of each of the changes to the baseline condition on parameter recovery was evaluated separately instead of using a full factorial design).

For each of the item parameters (except the item intercept in the RA model), we used two different values (relatively low and relatively high) which are similar to what has commonly been found in empirical data sets when using the HM in educational measurement: For α_i values of 0.5 and 1 were used, representing an item that functions somewhat poorly and an item that functions well, respectively; for $\{\lambda_{0i}, \lambda_{1i}\}$ the pairs of values $\{0.3, 0.4\}$ and $\{0.4, 0.3\}$ were used, which ensure that on all items RTs are informative of response speed, and that items only differ somewhat in whether the link with response speed is stronger for correct versus incorrect responses; for $\{\xi_{i0}, \xi_{i1}\}$ the pair of values {4, 4.1} and {4.1, 4} were used, resulting in average RTs that roughly match those found in empirical studies and resulting in small differences between the average RT for correct and incorrect responses; for $\{\sigma^2_{i0}, \sigma^2_{i1}\}$ the pairs of values {0.3, 0.2} and {0.2, 0.3} were used, resulting in a distribution of the RTs that roughly matches what one finds in empirical studies and in distributions that differ somewhat in variance for correct versus incorrect responses. Each combination of these values for α_i and these pairs of values of the time-related parameters were used, resulting in 16 unique item parameter combinations. The item intercept parameters were equally spaced between -1.5 and 1.5, such that a wide range of item difficulties is considered and the effect of item difficulty on the recovery of the other item parameters can be evaluated. In the condition with 32 items, the same item parameters were used twice.

For each of the first five conditions, 500 data sets were replicated using both the two-dimensional and three-dimensional version of $\mathcal{M}_{4a}$. For the last condition, 500 data sets were replicated using the three-dimensional $\mathcal{M}_{4a}$. For each data set to be generated, for each person the latent variables were sampled from $\mathcal{N}_2\left(\mathbf{0}, \begin{bmatrix} 1 & \Sigma_{12} \\ \Sigma_{12} & 1 \end{bmatrix}\right)$ for the conditions with $\eta_0 = \eta$, and from $\mathcal{N}_3\left(\mathbf{0}, \begin{bmatrix} 1 & \Sigma_{12} & \Sigma_{13} \\ \Sigma_{12} & 1 & \Sigma_{23} \\ \Sigma_{13} & \Sigma_{23} & 1 \end{bmatrix}\right)$ for the conditions with $\eta_0 \neq \eta_1$. The RA data were generated under the 2PNO model, and the RT data were generated using Equation 8. The joint model for differences in RTs of correct and incorrect responses (using the correct dimensionality) was fitted to each of the generated data sets using a Gibbs Sampler with 6000 iterations. The first 1000 iterations were removed as burn-in, and the estimates of the item parameters and of the correlations between the latent variables were computed based on each second iteration after the burn-in (i.e., the estimates were based on 2500 sampled values). For each type of item parameter and for each correlation between the latent variables the (average) absolute bias, variance and mean squared error were approximated based on the 500 sets of estimates.

3.1.2. Results The results of the simulation study are presented in Table 2. As the table shows, the (average) absolute bias, variance, and mean squared error (MSE) of each of the model parameter types appear to be small, suggesting that all item parameters as well as the latent correlations are recovered well. The largest MSE was generally observed for α_i, suggesting that this parameter is more difficult to recover than the other item parameters. The recovery of the parameters improves with sample size. Performance was at its worst for the smallest sample size, but also in this condition all model parameters seem to be recovered rather well, with the largest observed MSE being 0.014 (α_i, $N = 500$). Increasing sample size reduced the bias and variance. Increasing test length had only a slight impact on parameter recovery, with both the bias and variance either staying the same or improving a little. Having two highly correlated speed latent variables ($\Sigma_{23} = .9$) only seems to have resulted in a higher (but still small) bias for the estimated correlation, with the bias and variance being comparable to the baseline condition for all other parameters.

Since the results presented in Table 2 only consider the patterns observed per item type, a graphical representation of the results at the level of individual items may provide relevant additional information about the recovery of the item parameters for each specific item. Figure 2 presents these results for each of the parameters of the sixteen items included in the baseline condition. These results show that for most item parameter types, recovery of the parameters is comparable for the sixteen items. None of the items show any notable bias, suggesting that with sufficient data they would all be recovered well. There do, however, appear to be some relevant differences between the items in terms of the variance of the estimates of some parameters. For

TABLE 2.
Results of Simulation Study 1: average absolute bias, variance and mean squared error for the item parameters and the freely estimated parameters of the covariance matrix of the person parameters (based on 500 replications)

Condition		α	β	ξ	λ	σ^2	Σ_{12}	Σ_{13}	Σ_{23}
		Bias							
$\eta_0 = \eta_1$	Baseline	0.008	0.007	0.001	0.001	0.002	0.001	–	–
	$N = 500$	0.021	0.014	0.002	0.002	0.003	0.001	–	–
	$N = 2000$	0.005	0.003	0.001	0.000	0.001	0.001	–	–
	$K = 32$	0.008	0.006	0.001	0.001	0.002	0.000	–	–
	$\Sigma_{12} = .5$	0.009	0.007	0.001	0.001	0.002	0.004	–	–
$\eta_0 \neq \eta_1$	Baseline	0.008	0.006	0.001	0.001	0.001	0.001	0.001	0.003
	$N = 500$	0.018	0.014	0.002	0.002	0.004	0.000	0.000	0.007
	$N = 2000$	0.004	0.003	0.001	0.001	0.001	0.002	0.000	0.002
	$K = 32$	0.009	0.006	0.001	0.001	0.002	0.001	0.000	0.002
	$\Sigma_{12} = \Sigma_{13} = .5$	0.010	0.007	0.001	0.001	0.002	0.002	0.001	0.001
	$\Sigma_{23} = .9$	0.011	0.008	0.001	0.001	0.001	0.000	0.002	0.011
		Variance							
$\eta_0 = \eta_1$	Baseline	0.006	0.004	0.001	0.001	0.001	0.001	–	–
	$N = 500$	0.013	0.009	0.002	0.002	0.001	0.003	–	–
	$N = 2000$	0.003	0.002	0.000	0.001	0.000	0.001	–	–
	$K = 32$	0.005	0.004	0.001	0.001	0.001	0.001	–	–
	$\Sigma_{12} = .5$	0.007	0.004	0.001	0.001	0.001	0.001	–	–
$\eta_0 \neq \eta_1$	Baseline	0.007	0.004	0.001	0.001	0.001	0.002	0.002	0.001
	$N = 500$	0.014	0.009	0.002	0.002	0.001	0.003	0.004	0.001
	$N = 2000$	0.003	0.002	0.001	0.001	0.000	0.001	0.001	0.000
	$K = 32$	0.005	0.004	0.001	0.001	0.001	0.001	0.001	0.000
	$\Sigma_{12} = \Sigma_{13} = .5$	0.006	0.004	0.001	0.001	0.001	0.001	0.001	0.001
	$\Sigma_{23} = .9$	0.007	0.004	0.001	0.001	0.001	0.002	0.002	0.000
		Mean squared error							
$\eta_0 = \eta_1$	Baseline	0.006	0.004	0.001	0.001	0.001	0.001	–	–
	$N = 500$	0.014	0.009	0.002	0.002	0.001	0.003	–	–
	$N = 2000$	0.003	0.002	0.000	0.001	0.000	0.001	–	–
	$K = 32$	0.005	0.004	0.001	0.001	0.001	0.001	–	–
	$\Sigma_{12} = .5$	0.007	0.004	0.001	0.001	0.001	0.001	–	–
$\eta_0 \neq \eta_1$	Baseline	0.007	0.004	0.001	0.001	0.001	0.002	0.002	0.001
	$N = 500$	0.014	0.010	0.002	0.002	0.001	0.003	0.004	0.001
	$N = 2000$	0.003	0.002	0.001	0.001	0.000	0.001	0.001	0.000
	$K = 32$	0.005	0.004	0.001	0.001	0.001	0.001	0.001	0.000
	$\Sigma_{12} = \Sigma_{13} = .5$	0.006	0.004	0.001	0.001	0.001	0.001	0.001	0.001
	$\Sigma_{23} = .9$	0.007	0.005	0.001	0.001	0.001	0.002	0.002	0.000

ξ_{i0}, λ_{i0}, and σ^2_{i0}, items with a high intercept (β_i, i.e., easy items) show relatively large variance. In contrast, for ξ_{i1}, λ_{i1}, and σ^2_{i1}, items with a low intercept (i.e., difficult items) show relatively large variance. This can be explained by the fact that when an item is relatively easy or relatively difficult, there will be a lot of correct responses or incorrect responses, respectively. With few incorrect responses, there is relatively little information available for the estimation of ξ_{i0}, λ_{i0}, and σ^2_{i0}, resulting in a higher variance of the estimate. Conversely, for ξ_{i1}, λ_{i1}, and σ^2_{i1} this occurs for items for which there are few correct responses.

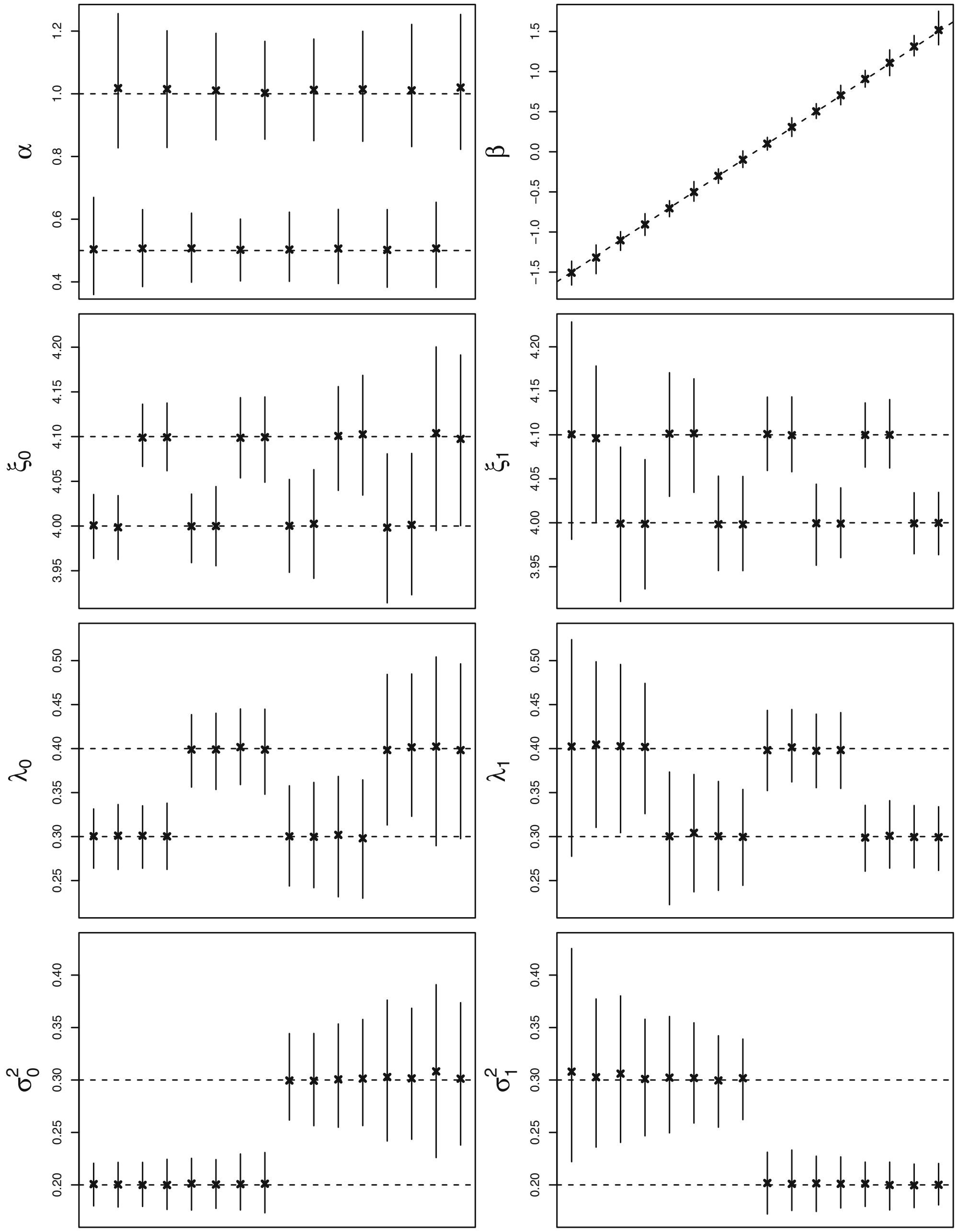

FIGURE 2.
Average estimates and the intervals between the 2.5th and 97.5th percentile of the distribution of the estimates of the items parameters in the baseline condition (on the y-axis) plotted against the true values of the parameters (on the x-axis). Dotted lines indicate where the estimate would be equal to the true value.

TABLE 3.
Simulation conditions used in Simulation Study 2: N—sample size, K—number of items Σ_{23}—correlation between the speed latent variable (used only for the three-dimensional models)

Condition	N	K	Σ_{23}
A (baseline)	1000	20	.7
B	1000	**10**	.7
C	1000	**40**	.7
D	**500**	20	.7
E	**2000**	20	.7
F	1000	20	**.9**

Note: For each non-baseline condition the factor that differentiates it from the baseline condition is in bold. Condition F was used only for the three-dimensional models.

3.2. Simulation Study 2: Model Selection

3.2.1. Method The second simulation study focused on evaluating the four model selection procedures presented in Section 2.2: mAIC, mBIC, mAIC in combination with the posterior predictive check, and mBIC in combination with the posterior predictive check. Data were generated under the 12 different models (6 two-dimensional and 6 three-dimensional models) for RT and RA. For each procedure, the models were fitted according to the model-fitting scheme presented in Figure 1, and the best model was selected. Five different conditions were considered for the two-dimensional models, and six different conditions were considered for the three-dimensional models. First, a baseline condition (A) with $N = 1000$, $K = 20$, and in the case of the three-dimensional models a correlation of .7 between the two speed latent variables ($\Sigma_{23}/\sqrt{\Sigma_{33}} = .7$) was considered. Second, two additional conditions with twice as few ($K = 10$) and twice as many ($K = 40$) items as in the baseline condition and $N = 1000$ were used (conditions B and C, respectively). Third, two conditions with the sample size twice as small ($N = 500$) and twice as large ($N = 2000$) as in the baseline condition and $K = 20$ were used (conditions D and E, respectively). Finally, in the case of the three-dimensional models a condition with a stronger correlation between the speed latent variables ($\Sigma_{23}/\sqrt{\Sigma_{33}} = .9$) was used (condition F). For clarity the conditions are summarized in Table 3. Similarly to Simulation Study 1, the effect of each of the changes to the baseline condition was evaluated separately and the full factorial design was not used.

In each condition 50 data sets were generated under each of the twelve models. In each replication item parameters were generated as follows: $\beta_i \sim \mathcal{N}(0, 0.5^2)$, $\alpha_i \sim \mathcal{N}(1, 0.2^2)$, for models with equal time intensities $\xi_i \sim \mathcal{N}(4, 0.5^2)$, otherwise

$$[\xi_{i0}\,\xi_{i1}]^T \sim \mathcal{N}_2\left(\begin{bmatrix}4\\4.1\end{bmatrix}, 0.25\begin{bmatrix}1 & 0.7\\0.7 & 1\end{bmatrix}\right), \tag{15}$$

for models with equal factor loadings $\lambda_i \sim \mathcal{N}(0.4, 0.1^2)$, otherwise

$$[\lambda_{i0}\,\lambda_{i1}]^T \sim \mathcal{N}_2\left(\begin{bmatrix}0.4\\0.4\sqrt{0.8}\end{bmatrix}, 0.01\begin{bmatrix}1 & 0.7\\0.7 & 1\end{bmatrix}\right), \tag{16}$$

for models with equal residual variances $\sigma_i^2 \sim \mathcal{U}(0.2, 0.3)$, otherwise $\sigma_{i0}^2 \sim \mathcal{U}(0.2, 0.3)$ and $\sigma_{i1}^2 \sim \mathcal{U}(0.15, 0.25)$; and person parameters were sampled from $\mathcal{N}_2\left(\mathbf{0}, \mathbf{I}_2\right)$ for the two-dimensional models, and from $\mathcal{N}_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu} = \mathbf{0}$ for models with $\xi_{i0} \neq \xi_{i1}$ and $\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \\ 0.1 \end{bmatrix}$ for models with $\xi_{i0} = \xi_{i1}$, and $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & \Sigma_{23}/\sqrt{\Sigma_{33}} \\ 0 & \Sigma_{23}/\sqrt{\Sigma_{33}} & 1 \end{bmatrix}$ for models with $\lambda_{i0} \neq \lambda_{i1}$ and $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & \Sigma_{23}/\sqrt{\Sigma_{33}}\sqrt{0.8} \\ 0 & \Sigma_{23}/\sqrt{\Sigma_{33}}\sqrt{0.8} & 0.8 \end{bmatrix}$ for models with $\lambda_{i0} = \lambda_{i1}$.

The models were estimated with Gibbs Samplers with 6000 iterations including 1000 burn-in, each second iteration after the burn-in was used for computing the estimates of the parameters. For each data set the best model was selected using mAIC and mBIC, both with and without applying the additional posterior predictive check to the best two-dimensional model (performed using 100 samples from the posterior distribution of the model parameters).

3.2.2. Results The results of the second simulation study are presented in Table 4. It can be observed that when data are generated under a two-dimensional model, both the mAIC and mBIC have difficulty selecting the right model when the additional posterior predictive check is not implemented. For both the mAIC and mBIC procedures, adding the posterior predictive check improves the performance when two-dimensional models are considered, while the performance when three-dimensional models are considered is hardly affected, with the exception of the condition in which the two speed latent variables are very strongly correlated. In this last condition the correct three-dimensional model is still selected in more than 80% of the replications when the mAIC with the posterior predictive check is used. Thus, in the conditions considered, only preferring a three-dimensional model over a two-dimensional model if the posterior predictive check indicates that the two-dimensional model cannot account for the person-level heterogeneity of RTs between the correct and incorrect responses seems to be desirable.

When one adds the posterior predictive check, the mAIC seems to generally outperform the mBIC. Both in the baseline condition (A) and in the condition with a longer test (C) or larger sample (E), the mAIC almost always (at least 47 replications out of 50) succeeded in selecting the right model. The mBIC does succeed in correctly selecting $\mathcal{M}_1$ and $\mathcal{M}_2$, but has more difficulty selecting $\mathcal{M}_{3b}$ and $\mathcal{M}_{4b}$, and rarely selects the correct model when the true model is $\mathcal{M}_{3a}$ or $\mathcal{M}_{4a}$, unless the sample size is large (condition E). This does not appear to be a problem for the mAIC, which seems to perform well under all conditions and for all generating models. Thus, the results of the simulation study suggest that using a procedure based on the mAIC and the posterior predictive check for the person-level heterogeneity of RTs performs well and may be preferred.

4. Empirical Example: PIAAC Problem Solving

The use of the models presented in this paper is illustrated using data from the problem solving in technology-based environments domain of the Programme of International Assessment of Adult Competences (PIAAC). Problem solving in PIAAC is defined as “using digital technology, communication tools and networks to acquire and evaluate information, communicate with others and perform practical tasks” (OECD, 2012, p. 46). Sample items from the problem solving domain can be found in the OECD’s report describing the measurement framework (OECD, 2012, p. 53). The items are interactive and require a constructed response (obtaining a correct response by guessing is not plausible). The PIAAC study included two computer-based problem solving modules each consisting of 7 items (14 unique items in total). Some respondents received only one of the problem solving modules and a module from a different domain, while others received

Table 4.

Results of Simulation Study 2, displaying the number of data sets (out of 50) in which the true model was selected as the best model with different selection procedures: modified Akaike information criterion (mAIC), mAIC with the posterior predictive check (PPC), modified Bayesian information criterion (mBIC), and mBIC with the PPC; across different conditions: A—baseline condition ($N = 1000, n = 20$), B—smaller number of items ($N = 1000, K = 10$), C—larger number of items ($N = 1000, K = 40$), D—smaller sample size ($N = 500, K = 20$), E—larger sample size ($N = 2000, K = 20$), F—stronger correlation between the speed latent variables (Σ_{23}=.9; only used for models with $\eta_0 \neq \eta_1$)

			mAIC						mAIC and PPC					
			Condition						Condition					
True model		P	A	B	C	D	E	F	A	B	C	D	E	F
$\eta_0 = \eta_1$	$\mathcal{M}_1$	$5K+1$	1	41	6	0	0	–	49	48	47	47	48	–
	$\mathcal{M}_2$	$6K+1$	0	39	4	0	0	–	49	45	48	50	49	–
	$\mathcal{M}_{3a}$	$7K+1$	5	42	13	0	0	–	49	47	45	49	48	–
	$\mathcal{M}_{3b}$	$6K+1$	0	28	2	0	0	–	47	46	47	47	49	–
	$\mathcal{M}_{4a}$	$8K+1$	0	20	1	0	0	–	48	48	49	49	49	–
	$\mathcal{M}_{4b}$	$7K+1$	0	32	4	0	0	–	49	45	47	50	49	–
$\eta_0 \neq \eta_1$	$\mathcal{M}_1$	$5K+5$	49	46	49	50	50	50	49	46	49	50	50	42
	$\mathcal{M}_2$	$6K+4$	50	47	48	49	48	48	50	47	48	49	48	41
	$\mathcal{M}_{3a}$	$7K+3$	50	50	50	50	49	48	50	49	50	50	49	43
	$\mathcal{M}_{3b}$	$6K+5$	50	49	49	50	50	50	50	47	49	50	50	47
	$\mathcal{M}_{4a}$	$8K+3$	49	48	50	50	50	50	50	45	45	50	50	43
	$\mathcal{M}_{4b}$	$7K+4$	50	46	45	50	50	50	49	47	50	50	50	45

			mBIC						mBIC and PPC					
			Condition						Condition					
True model		P	A	B	C	D	E	F	A	B	C	D	E	F
$\eta_0 = \eta_1$	$\mathcal{M}_1$	$5K+1$	14	50	0	24	0	–	49	50	49	48	50	–
	$\mathcal{M}_2$	$6K+1$	3	50	0	20	0	–	49	50	48	50	50	–
	$\mathcal{M}_{3a}$	$7K+1$	3	16	0	1	2	–	24	17	1	29	48	–
	$\mathcal{M}_{3b}$	$6K+1$	3	33	0	2	0	–	40	33	12	42	49	–
	$\mathcal{M}_{4a}$	$8K+1$	0	15	0	0	0	–	27	19	3	31	47	–
	$\mathcal{M}_{4b}$	$7K+1$	0	32	0	2	0	–	42	38	9	49	49	–
$\eta_0 \neq \eta_1$	$\mathcal{M}_1$	$5K+5$	50	50	50	50	50	50	43	37	48	41	48	45
	$\mathcal{M}_2$	$6K+4$	50	50	50	50	50	50	39	38	33	38	45	43
	$\mathcal{M}_{3a}$	$7K+3$	12	10	10	3	41	16	7	8	3	7	20	14
	$\mathcal{M}_{3b}$	$6K+5$	43	38	50	17	49	43	43	38	17	50	49	41
	$\mathcal{M}_{4a}$	$8K+3$	19	9	17	0	44	40	21	9	2	15	44	36
	$\mathcal{M}_{4b}$	$7K+4$	45	42	50	18	50	45	44	42	16	50	50	42

Note: P denotes the number of free parameters in the true model.

both problem solving modules. Each module had an overall time limit of 30 minutes. Public use data files were downloaded from the OECD webpage on the 12th of June 2018. Data from one of the participating countries (Canada) were used for analysis. This country was chosen because it has the largest number of respondents in the data set (10,315).

In line with PIAAC recommendations, disengaged responses were identified using the P+>0% method described and validated for the PIAAC study by Goldhammer et al. (2016).[5]

[5] For each item the responses were divided in bins based on RTs with 5 s per bin. The proportions of correct responses were computed per bin. The lower bound of the first bin in which the proportion of correct responses was above zero was used a threshold for disengaged responses. In total, 8.35% of the observed responses were flagged as disengaged.

Responses flagged as disengaged were treated as values missing at random. In total, 10,245 respondents provided a nondisengaged response to at least one of the items and were included in the subsequent analyses. The overall percentage of nonmissing responses in the data set was 58.5%, and the average number of responses was 8 per person and 5993 per item. The items were coded as correct/incorrect, and item-level RTs were available for analysis. Examination of the eigenvalues of the correlation matrix of the RA scores suggested that one dimension should be sufficient for modeling the RA data.

First, the CI-HM ($\mathcal{M}_1$, $\eta_0 = \eta_1$) was fitted to the data using a Gibbs Sampler with 20,000 iterations (including 10,000 burn-in, and a thinning of 2 was applied). We evaluated whether the model adequately captured important aspects of the dependence between RA and RT, namely possible differences between the RTs of correct and incorrect responses. Posterior predictive checks were performed with the following three statistics of interest: D_1, the difference between the average log-RT of correct and incorrect responses; D_2, the ratio between the variances of log-RTs of correct and incorrect responses; D_3, the ratio between the first eigenvalues of the correlation matrices of log-RTs computed separately for correct and incorrect responses.[6] In the empirical data the log-RT of the correct responses was on average 0.338 higher for correct responses than for incorrect responses, the variance of the log-RTs of correct responses was 0.537 of that of the incorrect responses, and the ratio between the first eigenvalues of the correlation matrices of log-RTs of the correct and incorrect responses was equal to 1.285. The statistics of interest were calculated for 100 replicated data sets generated using samples from the posterior distribution of the parameters of the CI model. In all of the 100 generated data sets the replicated D_1 and D_3 were smaller than the observed ones, and the replicated D_2 was larger than the observed one. These results strongly suggest that the CI-HM cannot adequately represent these aspects of the relationship between RA and RT, and there is likely CD between RA and RT that is not accounted for.

Second, we fitted two CD models: one representing the first approach to modeling of CD, and another representing the second approach. The first model is from Ranger and Ortner (2012) and is presented in Equation 5. The second model is an adaptation of one of the models from Bolsinova, Tijmstra, and Molenaar (2017) in which the CD is modeled as item-specific linear effects of the standardized residual log-RT on the slope and the intercept of the RA model, which here was generalized to allow for item-specific factor loadings in the lognormal model for RTs. Gibbs Samplers (20,000 iterations, including 10,000 burn-in, and a thinning of 2) were used for estimating these models, and independent vague priors were used for the item parameters similar to the priors used for the models developed in this paper.

The mAIC for these two CD models can be found in Table 5, which suggests that both models perform better than the CI model. That is, they successfully capture at least some of the CD in the data. However, this leaves unanswered the question of whether they succeed in fully accounting for observed differences between the RTs of correct and incorrect responses, which was investigated by using the same posterior predictive checks as used for the CI model. The results can be found in Table 5. For both models all three posterior predictive p-values were extreme, with D_1 and D_3 being too small in the replicated data and D_2 being too large. That is, while these two models capture some of the CD, they do not fully account for the observed differences between the log-RTs of the correct and incorrect responses. Therefore, these models do not allow one to study the substantively interesting differences in the response processes leading to correct and incorrect responses that these checks suggest are likely to be present in the data.

Third, in line with the model selection steps described in Section 2.2 and displayed in Figure 1, the set of two-dimensional models and the set of three-dimensional models described in Section 2

[6] This measure is aimed at capturing the possible differences between the RTs of correct and incorrect response in how strongly they are correlated among each other, and hence whether depending on RA the RTs show stronger or weaker structural patterns.

Table 5.

Model selection and posterior predictive checks for the models fitted to the PIAAC problem solving data: CI—conditional independence, CD—conditional dependence, P—number of free parameters, mAIC—modified Akaike information criterion, $\bar{D}$s—means of the posterior predictive distribution of there statistics of interest (D_1—the difference between the average log-RT of correct and incorrect responses; D_2—the ratio between the variances of log-RTs of correct and incorrect responses; D_3—the ratio between the first eigenvalues of the correlation matrices of log-RTs computed separately for correct and in correct responses), ps—corresponding posterior predictive p-values; posterior predictive checks are based on 100 samples from the posterior predictive distribution of response accuracy and response time under each of the models

Model	P	mAIC	$\bar{D}_1$	$\bar{D}_2$	$\bar{D}_3$	p_1	p_2	p_3
CI model ($\mathcal{M}_1$, $\eta_0 = \eta_1$)	71	234,633.3	0.215	0.779	0.811	.00	1.00	.00
CD model from approach 1	85	229,347.3	0.323	0.887	0.847	.00	1.00	.00
CD model from approach 2	99	228,351.8	0.322	0.819	0.837	.01	1.00	.00
CD models from approach 3								
$\mathcal{M}_2$, $\eta_0 = \eta_1$	85	227,699.7	0.336	0.859	0.832	.36	1.00	.00
$\mathcal{M}_{3a}$, $\eta_0 = \eta_1$	99	225,809.6	0.335	0.707	0.650	.32	1.00	.00
$\mathcal{M}_{4a}$, $\eta_0 = \eta_1$	113	218,993.7	0.336	0.549	1.240	.34	.94	.08
$\mathcal{M}_1$, $\eta_0 \neq \eta_1$	75	229,599.2	0.377	0.652	0.663	1.00	1.00	.00
$\mathcal{M}_2$, $\eta_0 \neq \eta_1$	88	225,194.5	0.336	0.733	0.667	.34	1.00	.00
$\mathcal{M}_{3a}$, $\eta_0 \neq \eta_1$	101	224,114.3	0.336	0.693	0.796	.29	1.00	0.00
$\mathcal{M}_{4a}$, $\eta_0 \neq \eta_1$	105	216,824.9	0.337	0.541	1.250	.40	.64	.15

were fitted to the data using Gibbs Samplers (20,000 iterations, including 10,000 burn-in, and thinning of 2). In line with the details described in Section 2.2 and the conclusions based on Simulation Study 2, for both the set of two- and three-dimensional models the best fitting model was selected based on the mAIC. The mAIC results for the considered models are displayed in Table 5. This led to the selection of $\mathcal{M}_{4a}$ for both the preferred two- and the three-dimensional model. Thus, regardless of whether one considers a model with a single speed dimension or a model with two speed dimensions, the most complex model type is preferred for modeling the RTs. This suggests that the factor loadings, residual variances, and the time intensities need to be estimated freely for correct and incorrect responses to best capture the patterns observed in this data set.

The mAIC values suggest that one should prefer the three-dimensional variant of $\mathcal{M}_{4a}$ over the two-dimensional variant. However, as discussed in Section 2, the mAIC may not be ideal for determining the preferred number of dimensions, and hence the discussed additional check against overfitting was performed. A posterior predictive check was performed for the best fitting two-dimensional model to evaluate whether it adequately captures the person-level heterogeneity between the RTs of the correct and incorrect responses. In the observed data the correlation between persons' average log-RT on correct responses and their average log-RT on the incorrect responses was equal to .442. Using the samples from the posterior distribution of the parameters of the two-dimensional $\mathcal{M}_{4a}$, 100 data sets were replicated and the correlation between the average log-RT of correct and incorrect responses was computed. This correlation ranged from .481 to .526 in the replicated data, resulting in a posterior predictive p-value of 1, which indicated that the two-dimensional model does not adequately capture the heterogeneity of RTs of correct and incorrect responses. Therefore, the three-dimensional model was selected as the best model for these data.

As can be seen in Table 5, all CD models proposed in this paper (with the exception of the three-dimensional $\mathcal{M}_1$) have a smaller mAIC than the models representing the first two approaches to modeling CD. Posterior predictive checks were applied to the fitted models to evaluate whether the models successfully account for the dependence between RA and RT as captured in the three

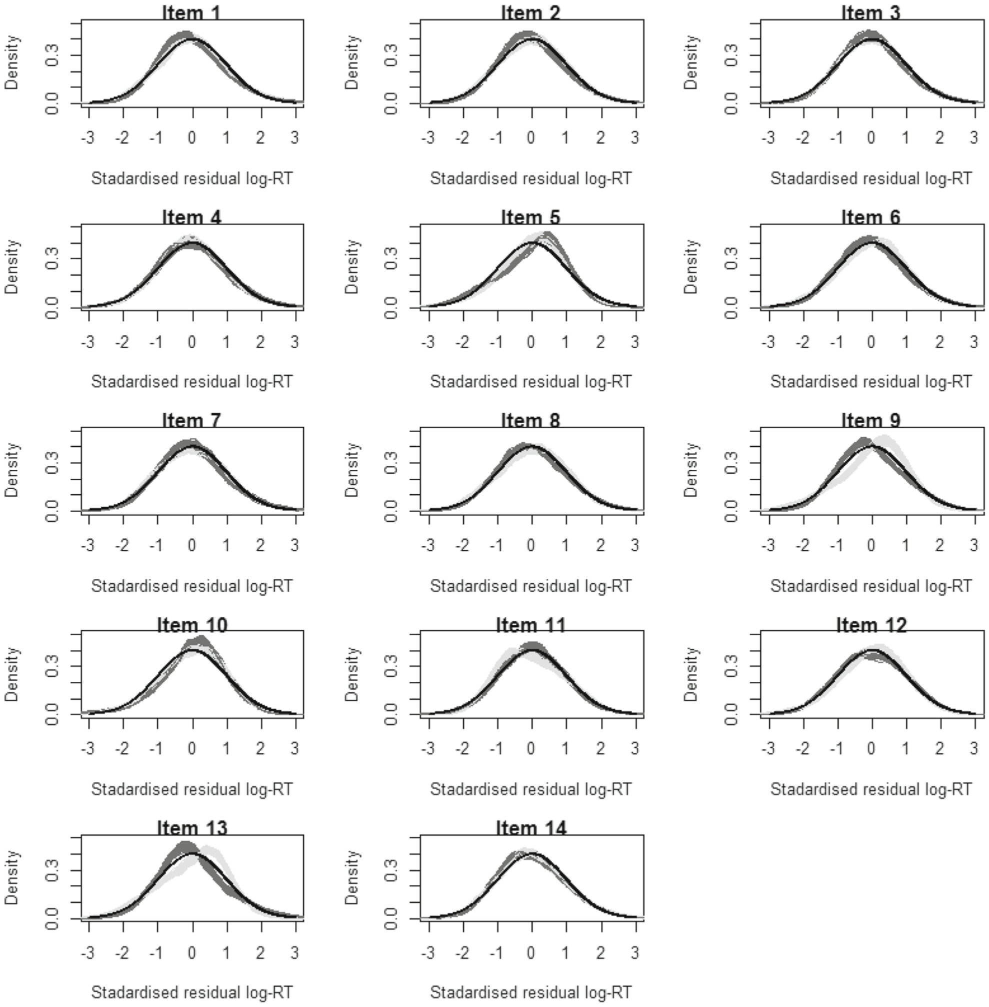

FIGURE 3.
Plots for examining the distributional assumptions of the three-dimensional $\mathcal{M}_{4a}$: A black line indicates the standard normal distribution, each light gray and each dark gray line is the empirical distribution of standardized residual of log-RTs of incorrect and correct responses, respectively, of an item computed using one sample from the posterior distribution of the model parameters.

statistic of interest (see Table 5). For all models except the three-dimensional $\mathcal{M}_1$, the observed D_1 is consistent with its posterior predictive distribution. However, the ratio of the variances of the log-RTs of the correct and incorrect responses (D_2) and the ratio between the first eigenvalues of the correlation matrices of the correct and incorrect responses (D_3) are only captured adequately by the models that allow for differences between the factor loadings and the residual variances for correct and incorrect responses (i.e., the two- and three-dimensional $\mathcal{M}_{4a}$). These results suggest that only the most general versions of the two- and three-dimensional models adequately capture these observed differences between the RTs of correct and incorrect responses.

The selected CD model has different assumptions about the distribution of RTs compared to the CI model and the CD models from the first two modeling approaches: Instead of assuming a lognormal distribution for the RT for the correct responses and the incorrect responses combined, it assumes a separate lognormal distribution for RT for the two RA outcomes. While this gives the

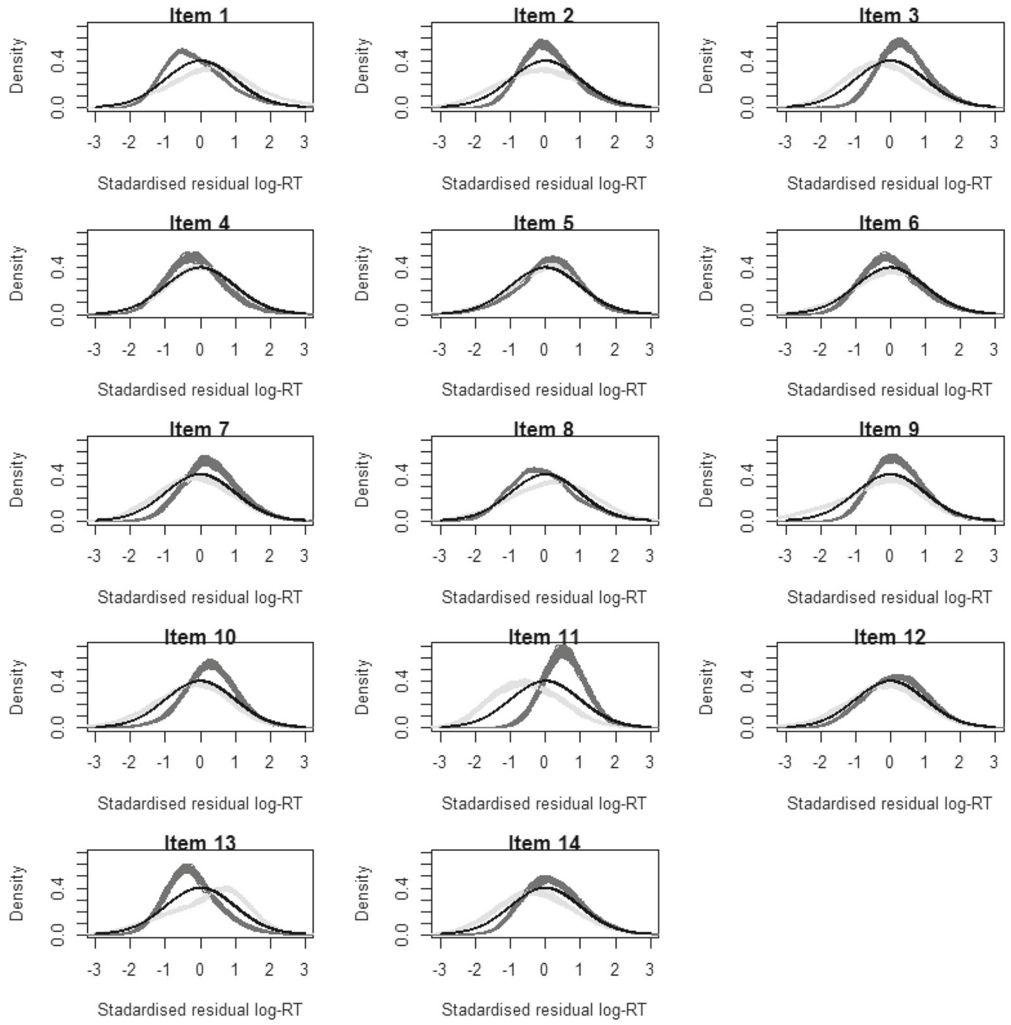

FIGURE 4.
Examining the distributional assumptions of the conditional independence model: a black line indicates the standard normal distribution, each light gray and each dark gray line is the empirical distribution of standardized residual of log-RTs of incorrect and correct responses, respectively, of an item computed using one sample from the posterior distribution of the model parameters.

model more flexibility by not requiring the overall RTs to be lognormally distributed, it still makes distributional assumptions that need to be checked. To evaluate this assumption, we examined the posterior distribution of the standardized residuals of log-RTs. If the distributional assumptions hold, then these residuals would have a standard normal distribution. Figure 3 presents a graphical check for this: For 100 samples from the posterior distribution of the model parameters under the selected model the standardized residuals of log-RTs were computed and their distribution was plotted for each item for the correct and the incorrect responses separately. For some of the items, there is a very close match between the empirical distribution of the residuals and the standard normal distribution. For other items, there are some deviations from $\mathcal{N}(0, 1)$, but these deviations are much smaller than those observed for the residuals in the CI model (see Figure 4).

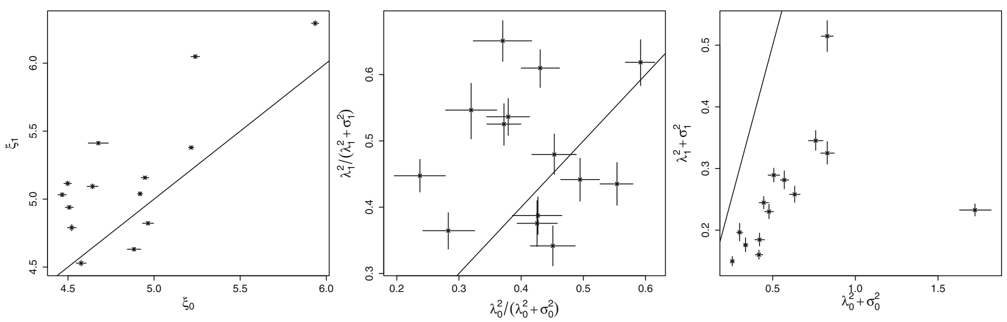

FIGURE 5.

The estimates of the RT-related item properties for the incorrect responses (on the x-axis) and for the correct responses (on the y-axis). The location of the points in the plot indicate the posterior means of the parameters and the lines attached to the point indicate the 95% credible intervals for the parameters: time intensities on the left, proportions of explained variance in the middle, and total variances on the right.

Under the selected model, three correlations between the person parameters are considered. The correlation between θ and η_0 was estimated to be -.661 (95% credible interval [-.642,-.679]), indicating that for incorrect responses there is a strong negative association between the response speed and the ability level of the respondent, with persons who give fast incorrect responses generally having a lower ability level. In contrast, for the correct responses a weak positive association was found between speed and ability, with the correlation between θ and η_1 estimated at .038 [.005,.072]. Thus, the link between response speed and ability is much weaker and also different in sign for correct responses than for incorrect responses. The correlation between η_0 and η_1 was estimated to be .689 [.662,.714], indicating that the two speed latent variables are strongly associated but still only share less than 50% of their variance. Overall, these results suggest that considering two rather than one speed latent variable is needed to capture the patterns present in the data, and that considering these will provide a more complete picture of the underlying response processes than would be possible when considering only a single speed latent variable.

Figure 5 shows the estimates and the 95% credible intervals for relevant item properties in the RT model. For all 14 items the credible interval for the difference between the time intensities of the item ($\xi_{i1} - \xi_{i0}$) excluded 0, where for 3 items time intensity is higher for incorrect responses, and for 11 items it is higher for correct responses. Thus, (conditional on ability) correct responses seem to generally take more time than incorrect responses for this test. In addition to differences in time intensity, the model also allowed each item to differ in their factor loading and residual variance for correct and incorrect responses. As both the factor loadings and the residual variances are not standardized, a direct comparison of these parameters themselves may be less informative than considering for each item differences in the proportion of variance explained by the speed latent variables $\left(\frac{\lambda_i^2}{\lambda_i^2+\sigma_i^2}\right)$ and in total variance ($\lambda_i^2 + \sigma_i^2$). For all 14 items the credible interval for the difference between the total variance of the correct log-RTs and of the incorrect log-RTs excluded zero, with all items having a higher total variance for incorrect responses. That is, correct responses were more homogeneous than incorrect responses, and the incorrect responses are likely to come from a larger variety of different response processes. For 7 items the credible interval for the difference in the proportions of explained variance for the correct and the incorrect responses excluded zero, with 5 items having a larger proportion of explained variance for incorrect responses and 2 items for correct responses. Thus, while there appear to be differences for a notable portion of the items in the covariance structure of log-RTs for correct and incorrect responses, there does not appear to be a clear pattern to these differences.

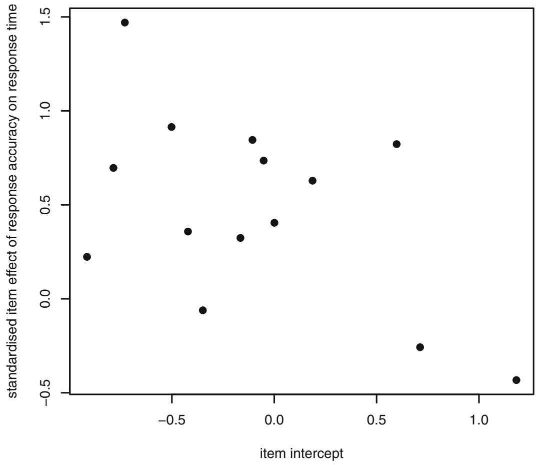

FIGURE 6.
Relationship between the estimated item easiness (on the x-axis) and the estimated effect of RA on log-RT (on the y-axis).

In their study on CD, Bolsinova, De Boeck, and Tijmstra (2017) found that difficult and easy items differ in the sign of CD between RT and RA, where difficult items generally had a positive residual association while easy items had a negative residual association. This suggests that in the present study it may also be relevant to investigate whether the observed item-level CD between RT and RA can be explained by the item easiness. To quantify the strength and direction of the CD for each of the 14 items, we computed the item-specific standardized effect of RA on log-RT: For each item an average of $\frac{\xi_{i1}-\xi_{i0}}{\sqrt{\frac{(N_{i0}-1)(\sigma^2_{i0}+\lambda^2_{i0})+(N_{i1}-1)(\sigma^2_{i1}+\lambda^2_{i1})}{N_{i0}+N_{i1}-2}}}$, was computed across iterations of the Gibbs Sampler, where the denominator is the pooled standard deviation of log-RT, N_{0i} is the number of correct responses to item i, and N_{1i} is the number of incorrect responses to item i. The sample correlation between the item easiness and the item-specific standardized effect of RA on RT was $-.512$, and the results are displayed graphically in Figure 6. These results suggest that generally for difficult items correct responses take longer than incorrect responses, while for easier items this effect may be less strong or, for very easy items, even be reversed. This general finding is in line with what has been found when applying models that incorporate the effect of residual RT on RA (e.g., Bolsinova, De Boeck, & Tijmstra, 2017; Bolsinova, Tijmstra, & Molenaar, 2017a).

Scherer, Greiff and Hautamäki (2015) explored the relationship between complex problem solving ability and complex problem solving time-on-task, for which they used the CI model. A similar research question can be explored in the context of problem solving in PIAAC. If one would use a CI-HM, one would only be able to obtain a rather limited picture of the way in which problem solving ability relates to time-on-task, since only an overall correlation between ability and response speed is considered. This correlation between the two latent variables in the CI model is estimated to be $-.668$ [$-.651$, $-.684$] for this particular data set, which would suggest that people who work faster on this test also perform worse. However, if one used the proposed CD model, one would conclude that ability is negatively correlated only with the response speed on the incorrect responses (i.e., being fast is associated with lower ability), while it is hardly (and positively) related to the response speed of the correct responses, as the results discussed above showed. Concretely, this leads one to question the general claim that one might be tempted to make based on the results for the CI model, that working fast is generally associated with having a lower ability: Here, this claim only seems to hold for incorrect responses. This is an

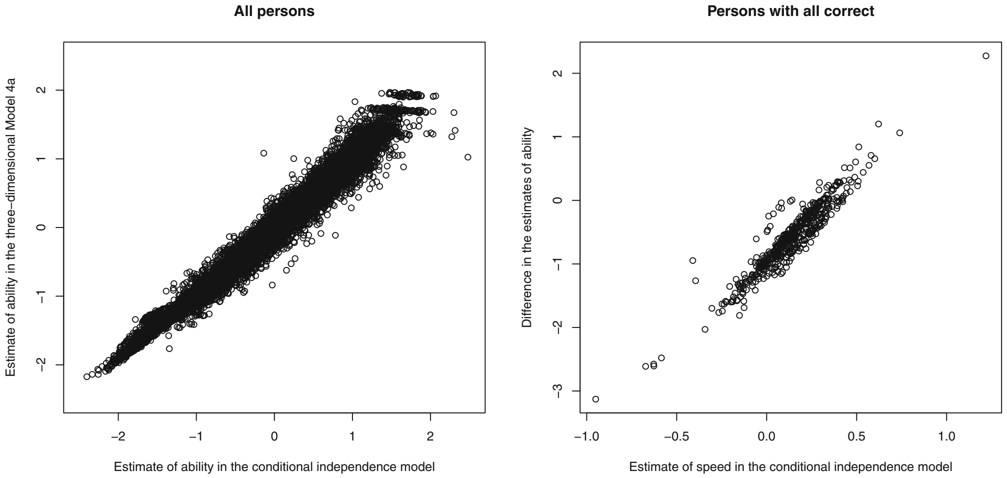

FIGURE 7.
Left panel: Comparison of the estimates of ability in the conditional independence model (on the x-axis) and in the three-dimensional $\mathcal{M}_{4a}$ (on the y-axis) on the right. Right panel: Dependence of the differences between the estimates of ability in the two models (on the y-axis) on the speed estimate from the conditional independence model (on the x-axis) for persons with all observed responses correct.

important distinction that suggests that for those persons that are following the right solution strategy (leading to a correct response), the speed at which they do so is hardly linked to their ability level, and hence concluding that people who respond quickly also generally have lower ability would misrepresent the patterns observed exactly for those persons that we can assume followed the intended solution strategy (as opposed to persons who may have followed unintended solution strategies and ended up with an incorrect response). An additional benefit of the used CD model is that it allows one to consider the item-specific residual association between RA and RT. This for example allows one to observe that for three items the CD is negative, with fast responses being more often correct than slow responses. Such fine-grained results make it possible to obtain a more complete and nuanced picture of the relationship between time-on-task and performance than is possible under the CI-HM.

To further examine the consequences of using the selected CD model instead of the CI model, we compared the estimates of ability obtained under the two models (see Figure 7). On the left panel of the figure one can see that overall the estimates of ability under the two models are rather close to each other. However, one can also see that while there is hardly any difference for persons with low ability (i.e., persons with mainly incorrect responses), there is quite a large difference on the higher end of ability. While persons with all responses correct get almost the same estimates of ability in the CD model, in the CI model there is a rather large variance of the ability estimates between these persons. The right panel of Figure 7 shows what can explain this variance and the difference between the estimates of ability in the two models: Due to the negative correlation between speed and ability in the CI model, for persons with low speed and all items correct the estimates of ability are larger under the CI model than under the CD model, while for persons with high speed the converse holds. One can consider the increase of the estimates for slower respondents with all correct responses and the decrease of the estimates for faster respondents with all correct responses resulting from the HM undesirable, since there is no negative relationship between ability and the speed of correct responses.

TABLE 6.
Results of the simulation study based on the empirical example: average absolute bias, variance and mean squared error for the item parameters and the freely estimated parameters of the covariance matrix of the person parameters (based on 500 replications)

Condition	α	β	ξ	λ	σ^2	Σ_{12}	Σ_{13}	Σ_{23}
	Bias							
$N = 500$	0.033	0.011	0.004	0.001	0.002	0.016	0.019	0.024
$N = 1000$	0.020	0.006	0.001	0.001	0.001	0.011	0.011	0.016
$N = 2000$	0.009	0.003	0.002	0.001	0.001	0.011	0.004	0.009
$N = 10{,}245$	0.004	0.001	0.001	0.001	0.000	0.005	0.005	0.006
	Variance							
$N = 500$	0.019	0.010	0.002	0.002	0.001	0.001	0.004	0.002
$N = 1000$	0.009	0.005	0.001	0.001	0.001	0.001	0.002	0.001
$N = 2000$	0.004	0.002	0.000	0.000	0.000	0.000	0.001	0.000
$N = 10{,}245$	0.002	0.001	0.000	0.000	0.000	0.000	0.000	0.000
	Mean squared error							
$N = 500$	0.021	0.010	0.002	0.002	0.001	0.001	0.004	0.002
$N = 1000$	0.010	0.005	0.001	0.001	0.001	0.001	0.002	0.001
$N = 2000$	0.004	0.002	0.000	0.000	0.000	0.000	0.001	0.001
$N = 10{,}245$	0.002	0.001	0.000	0.000	0.000	0.000	0.000	0.000

Note: The first three conditions have a complete design, and the last condition has an incomplete design with the missingness patterns matching those in the empirical example.

5. Simulation Study Based on the Empirical Example

Based on the results of the empirical example, an additional simulation study was performed that closely matched the empirical data. In the original simulation study the values for the item parameters were chosen such that item with different combinations of typical values of the item parameters were present in the test. However, such choice of the parameter values did not allow to include the amount of heterogeneity in the item parameters of each type that is common for empirical data. Furthermore, it was not possible to include a relationship between the item easiness and the CD between RA and RT in the choice of the item parameter values. For this reason, in this additional study the estimates of the item parameters and of the covariance matrix of the person parameters were used as true values of the model parameters for simulating data. Three conditions with a complete design were considered using the three sample sizes considered in the original simulation study: $N = 500$, $N = 1000$, and $N = 2000$. Additionally, a condition with the same sample size ($N = 10{,}245$) and the pattern of missingness as in the empirical example was used. In each replication person parameters were sampled from a multivariate normal distribution, and the RA and RT were generated using the 2PNO model and the three-dimensional $\mathcal{M}_{4a}$ in Equation 8, respectively. 500 simulated data sets were generated in each condition and the three-dimensional $\mathcal{M}_{4a}$ was fitted to each of these data sets using the Gibbs Sampler (6000 iterations, including 1000 burn-in, and a thinning of 2). Parameter recovery (bias, variance, and MSE) was evaluated for each item parameter and each freely estimated element of the covariance matrix of the latent variables (Table 6).

Table 6 shows the parameter recovery results. As can be observed, while the recovery of each of the parameters is worse than what was observed in the earlier parameter recovery simulation study, the bias, variance and MSE still seem to be acceptable. Since the parameter values used to generate the data are somewhat more extreme than the ones used in the earlier study, the fact that each parameter shows worse recovery may not be too surprising. As before, parameter recovery benefits greatly (both in terms of bias and variance) from an increase in sample size. For the actual sample size of the data set, there is hardly any bias or variance for any of the parameters, suggesting that parameter recovery should not have been an issue for the considered empirical example.

6. Discussion

While conceptually appealing, the standard HM makes strong assumptions about the relationship between RA and RT that often do not hold in practice. When extending the HM for RA and RT, three general approaches can be followed: including a dependence parameter in the bivariate distribution of RA and RT of each item, modeling the marginal distribution of RT and the conditional distribution of RA given RT, and modeling the marginal distribution of RA and the conditional distribution of RT given RA. This paper proposed a framework in line with the third approach, which allows users to directly investigate whether the RT model for correct responses differs from the RT model for incorrect responses, and hence whether there may be qualitative differences between the response processes leading to correct and incorrect responses. The parameter recovery simulation studies suggest that all model parameters can generally be recovered well if the model is correctly specified. The model selection simulation study suggests that the mAIC is well-suited for selecting the correct model, as long as one includes a posterior predictive check that ensures that a three-dimensional model is only selected if the best two-dimensional model cannot account for observed person-level heterogeneity of the RTs of correct versus incorrect responses. The results of the simulation study suggest that this check generally helps select the right number of dimensions, which generally performed well and only had some difficulties selecting a model with two speed latent variables when these were highly correlated (.9). One could argue that in the latter case, failing to distinguish between the two speed latent variables may not be overly problematic given the extent to which they overlap, and it should also be noted that in those conditions the procedure is still able to select the right model over 80% of the time.

The application of the modeling framework to a large-scale assessment test suggested that there may in practice be notable relevant differences between the models for the RTs of correct and incorrect responses. Based on the model selection procedure, the most complex three-dimensional model was preferred, suggesting that to best model the data two speed latent variables needed to be considered and that all RT item parameters should be estimated freely for correct and incorrect responses. These findings suggest that there may be qualitative differences between the response processes that lead to correct and incorrect responses, which would explain why for many items differences were found in the item parameters for correct and incorrect responses. This suggests that treating all responses as being the product of comparable response processes—as is implicitly assumed by the standard HM—may be too simplistic, and could possibly lead to confounding of measurement.

On the person side, two speed latent variables were needed to best model the empirical data. While these two variables were strongly correlated, a notable difference was found in the correlation of these variables with the ability latent variable, which suggested that for this test by and large only the speed at which incorrect responses are given was informative of ability. This suggests that the two speed latent variables may have substantively different interpretations. It also suggests that using a single speed latent variable when two are needed may confound measurement of ability: If it is only the speed at which incorrect responses are provided that is related to ability, it may be undesirable to ‘punish’ respondents who give many fast correct responses for having a high response speed. This suggests that it may be important to further study which psychological attributes can in practice be taken to underlie the speed latent variable in the standard HM and the two speed latent variables in the three-dimensional version of the model proposed in this paper, to gain a better picture of the optimal way of modeling the underlying phenomena and to avoid the confounding of measurement due to a possible mismatch between these phenomena and the statistical models we use to capture these phenomena.

In this paper a specific way of modeling possible differences in the RTs of correct and incorrect responses was chosen, which is based on the assumption that the lognormal model for RTs as it is commonly considered in the standard HM is appropriate for modeling the data, as long as one allows some (or all) of the item parameters (and possibly also the speed latent variable) to differ based on the RA. While this framework present a more general way of modeling the processes underlying the RT, it is not a given that the considered parametric form will always adequately capture the structure in the data. Other parametric forms for the RT model for correct and incorrect responses could be explored, and it could also be possible that the model that needs to be considered for correct responses is not of the same parametric form as the one needed to explain the RTs of incorrect responses. Furthermore, semi-parameteric models can be considered for RTs, which would allow to handle different types of RT distributions in a flexible way (Wang, Chang, & Douglas, 2013a; Wang, Fan, Chang, & Douglas, 2013b).

When using the developed framework, care should be taken to ensure that the model for the ability latent variable(s) is correctly specified, to avoid risks of possible confounding. Specifically, one should be careful with assuming a single ability latent variable to be sufficient when in fact the test is multidimensional, since situations are conceivable where this unmodeled multidimensionality on the RA side would affect the estimated model on the RT side if items matching the different ability dimensions differ in their average time intensity. Such possible misspecifications should be checked carefully before continuing to jointly model RA and RT.

While the framework presented in this paper has focused on modeling dichotomously scored responses obtained based on open-ended questions, it can readily be extended to make it possible to deal with polytomously scored responses. Similar to the dichotomous case, one could allow the model for RT to vary for some or all of the different possible item score scores. Such an approach might be appealing if one considers there to be important qualitative differences between the response processes leading to, for example, incorrect, partially correct, and fully correct responses, possibly resulting in structural differences in the RTs. Alternatively, one might consider it plausible that partially correct and fully correct responses originate from similar response processes, and only separate the RT models for incorrect responses versus responses that are correct or partially correct. As with the dichotomous case, considering an approach that lets the RT model depend on the item score will be most relevant in cases where one suspects there to be heterogeneity in the response processes leading to the different possible item scores.

The model framework considered in this paper extends the HM in a specific way: Only the measurement model of speed is generalized and a second speed latent variable is introduced. While relaxing the assumption of CI of RA and RT, the proposed models do maintain the assumption of the standard HM that conditional on the speed latent variable(s), RT is no longer informative of ability. Hence, it is still assumed that RTs only provide collateral information for the estimation of ability through the speed latent variable(s) in the model. As Bolsinova and Tijmstra (2018) have shown, this assumption may not always be plausible, and allowing for cross-loadings of RT on ability may have the potential to improve the precision with which ability is measured. In light of the findings in the empirical example, it may be relevant to merge the two approaches and extend the current model by including cross-loadings of RT on ability that may differ depending on RA. This would, for example, allow one to take the possibility into account that residual RT is only indicative of ability if a correct response was provided, and not (or less) indicative if an incorrect response was given.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

Albert, J. (1992). Bayesian estimation of Normal Ogive item response curves using Gibbs sampling. *Journal of Educational Statistics*, *17*, 251–269. https://doi.org/10.2307/1165149.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395–479). Reading: Addison-Wesley.

Bolsinova, M., De Boeck, P., & Tijmstra, J. (2017). Modelling conditional dependence between response time and accuracy. *Psychometrika*, *82*(4), 1126–1148.

Bolsinova, M., & Tijmstra, J. (2016). Posterior predictive checks for conditional independence between response time and accuracy. *Journal of Educational and Behavioral Statistics*, *41*(2), 123–145.

Bolsinova, M., & Tijmstra, J. (2018). Improving precision of ability estimation: Getting more from response times. *British Journal of Mathematical and Statistical Psychology*, *71*(1), 13–38.

Bolsinova, M., Tijmstra, J., & Molenaar, D. (2017a). Response moderation models for conditional dependence between response time and response accuracy. *British Journal of Mathematical and Statistical Psychology*, *70*(2), 257–279.

Bolsinova, M., Tijmstra, J., Molenaar, D., & De Boeck, P. (2017b). Conditional dependence between response time and accuracy: An overview of its possible sources and directions for distinguishing between them. *Frontiers in Psychology*, *8*, 202. https://doi.org/10.3389/fpsyg.2017.00202.

Gelman, A., Meng, X., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *4*, 733–807.

Goldhammer, F., Martens, T., Christoph, G., & Lüdtke, O. (2016). Test-taking engagement in PIAAC. *OECD Education Working Papers*, (133). https://doi.org/10.1787/5jlzfl6fhxs2-en

Hoff, P. D. (2009). *A first course in Bayesian statistical methods*. New York: Spinger. https://doi.org/10.1007/978-0-387-92407-6.

Klein Entink, R., Fox, J., & van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika*, *74*, 21–48.

Klein Entink, R., van der Linden, W. J., & Fox, J.-P. (2009). A Box–Cox normal model for response times. *British Journal of Mathematical and Statistical Psychology*, *62*(3), 621–640.

Lord, F., & Novick, M. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

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

Meng, X.-L. (1994). Posterior predictive p-values. *The Annals of Statistics*, *22*(3), 1142–1160.

Molenaar, D., Tuerlinckx, F., & van der Maas, H. (2015a). A bivariate generalized linear item response theory modeling framework to the analysis of responses and response times. *Multivariate Behavioural Research*, *50*, 56–74. https://doi.org/10.1080/00273171.2014.962684.

Molenaar, D., Tuerlinckx, F., & van der Maas, H. (2015b). A generalized linear factor model approach to the hierarchical framework for responses and response times. *British Journal of Mathematical and Statistical Psychology*, *68*, 197–219. https://doi.org/10.1111/bmsp.12042.

OECD. (2012). *Literacy, numeracy and problem solving in technology-rich environments*. Paris: OECD Publishing. https://doi.org/10.1787/9789264128859-en.

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

Ranger, J., & Ortner, T. (2012). The case of dependency of responses and response times: A modeling approach based on standard latent trait models. *Psychological Test and Assessment Modeling*, *54*, 128–148.

Rouder, J. N., Sun, D., Speckman, P. L., Lu, J., & Zhou, D. (2003). A hierarchical bayesian statistical framework for response time distributions. *Psychometrika*, *68*(4), 589–606.

Scherer, R., Greiff, S., & Hautamäki, J. (2015). Exploring the relation between time on task and ability in complex problem solving. *Intelligence*, *48*, 37–50.

Semmes, R., Davison, M. L., & Close, C. (2011). Modeling individual differences in numerical reasoning speed as a random effect of response time limits. *Applied Psychological Measurement*, *35*(6), 433–446.

Tanner, M., & Wong, W. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*, 528–540. https://doi.org/10.2307/2289457.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioural Statistics*, *31*, 181–204. https://doi.org/10.3102/10769986031002181.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*, 287–308. https://doi.org/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.

van der Linden, W. J., & Glas, C. A. W. (2010). Statistical tests of conditional independence between responses and/or response times on test items. *Psychometrika*, *75*(1), 120–139. https://doi.org/10.1007/s11336-009-9129-9.

van der Linden, W. J., Klein Entink, R. H., & Fox, J. P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement*, *34*, 327–347. https://doi.org/10.1177/0146621609349800.

van der Maas, H. L., & Jansen, B. R. (2003). What response times tell of children's behavior on the balance scale task. *Journal of Experimental Child Psychology*, *85*(2), 141–177.

van der Maas, H. L., & Wagenmakers, E.-J. (2005). A psychometric analysis of chess expertise. *The American journal of psychology*, *118*, 29–60.

Van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*(4), 486–492.

Wang, C., Chang, H. H., & Douglas, J. A. (2013a). The linear transformation model with frailties for the analysis of item response times. *British Journal of Mathematical and Statistical Psychology*, *66*(1), 144–168.

Wang, C., Fan, Z., Chang, H. H., & Douglas, J. A. (2013b). A semiparametric model for jointly analyzing response times and accuracy in computerized testing. *Journal of Educational and Behavioral Statistics*, *38*(4), 381–417.

Zhan, P., Jiao, H., & Liao, D. (2017). Cognitive diagnosis modelling incorporating item response times. *British Journal of Mathematical and Statistical Psychology*, *71*, 262–286.

Manuscript Received: 30 MAY 2018
Final Version Received: 23 JUL 2019
Published Online Date: 28 AUG 2019