British Journal of Mathematical and Statistical Psychology (2016)

The British Psychological Society

www.wileyonlinelibrary.com

Special Issue Paper

Response moderation models for conditional dependence between response time and response accuracy

Maria Bolsinova[1]*, Jesper Tijmstra[2] and Dylan Molenaar[1]

[1]University of Amsterdam, The Netherlands
[2]Tilburg University, The Netherlands

It is becoming more feasible and common to register response times in the application of psychometric tests. Researchers thus have the opportunity to jointly model response accuracy and response time, which provides users with more relevant information. The most common choice is to use the hierarchical model (van der Linden, 2007, *Psychometrika*, 72, 287), which assumes conditional independence between response time and accuracy, given a person's speed and ability. However, this assumption may be violated in practice if, for example, persons vary their speed or differ in their response strategies, leading to conditional dependence between response time and accuracy and confounding measurement. We propose six nested hierarchical models for response time and accuracy that allow for conditional dependence, and discuss their relationship to existing models. Unlike existing approaches, the proposed hierarchical models allow for various forms of conditional dependence in the model and allow the effect of continuous residual response time on response accuracy to be item-specific, person-specific, or both. Estimation procedures for the models are proposed, as well as two information criteria that can be used for model selection. Parameter recovery and usefulness of the information criteria are investigated using simulation, indicating that the procedure works well and is likely to select the appropriate model. Two empirical applications are discussed to illustrate the different types of conditional dependence that may occur in practice and how these can be captured using the proposed hierarchical models.

1. Introduction

Due to the increased popularity of computerized test administration, in many applications of psychological tests the item response times (RTs) have become available in addition to the item responses. Psychometric models have been developed to jointly model RT together with response accuracy (RA) and to incorporate this additional source of information about the responses into the measurement model (e.g., van der Linden, 2007; Molenaar, Tuerlinckx & van der Maas 2015a,b). These approaches all have in common that the RT and RA are modelled hierarchically: separate measurement models for RA and RT are connected at the higher level at which the relationship between persons' overall

*Correspondence should be addressed to Maria Bolsinova, Department of Psychology, University of Amsterdam, Nieuweachtergracht 129, 1018 WS Amsterdam, The Netherlands (email: m.a.bolsinova@uva.nl).

DOI:10.1111/bmsp.12076

ability and speed on the test and the relationship between the items' difficulty and time intensity are modelled.

The hierarchical structure allows one to explain the correlation between RT and RA, but only that part of it which stems from the correlation between the ability and speed of the persons and from the correlation between the difficulty and time intensity of the items. As a consequence, these models cannot account for influences contributing to a correlation between RT and RA due, for example, to the fact that some respondents vary their speed during the test resulting in a speed–accuracy trade-off, or from the use of different response strategies (across and within persons) that affect both how much time it takes to solve an item and how likely it is to obtain a correct response (e.g., automated versus controlled processing).

To be able to draw conclusions about the association between RT and RA that go beyond assessing the correlation between the overall speed and ability across persons and beyond the correlation between the overall difficulty and time intensity, a joint model for RT and RA has to allow for conditional dependence (CD) between RT and RA. Several extensions of the hierarchical model have been proposed that capture some form of CD between RT and RA.

Ranger and Ortner (2012) proposed to include an item-specific residual correlation parameter in the hierarchical model which directly reflects the additional correlation between the RT and RA on the same item that cannot be explained by the correlation between speed and ability. Their model, however, does not allow different patterns of CD for different persons. Meng, Tao and Chang (2015) used a product of the person and the item components for the residual correlation instead of the item-specific correlation. However, in their model the item component is fixed to be non-negative, such that the items cannot differ in the sign of the residual correlation between RT and RA, which is rather restrictive given the empirical evidence suggesting that items often do differ in the direction of CD, and do so in a systematic way: the residual correlation between RT and RA is often positive for difficult items but negative for easy items (Partchev & de Boeck, 2012; Bolsinova, de Boeck & Tijmstra, 2016).

Bolsinova *et al.* (2016) proposed a model that allows the residual correlation for items to differ in sign as well as in size. They achieve this by incorporating item-specific effects of the residual RT on RA. However, their approach is limited by the assumption that there are no differences in CDs between persons, in the sense that persons with the same ability are assumed not to differ in the effect that residual RT has on RA. This means that it is not designed to deal with possible between-person differences that may influence the sign and magnitude of CD, such as differences in the extent to which a person's accuracy is influenced by an increase in speed or switches in test-taking strategies.

Some of the existing models allow for between-item difference in CD (Ranger & Ortner, 2012; Bolsinova *et al.*, 2016) but do not allow for person differences, while other models focus on the differences between persons (Meng *et al.*, 2015). However, among currently available approaches there is no general way to account for both item and person differences in residual dependence in a flexible way. In this paper we propose such a general framework that remedies the mentioned limitations of the existing models for CD and allows for a flexible combination of item- and person-specific effects, providing users with a structured way of assessing whether CD of a particular type is present. The proposed framework, which we call a response moderation framework, consists of a set of nested hierarchical models that include a moderator effect of continuous residual log-RT on RA, and have the hierarchical model assuming conditional independence (CI) as a special case. Unlike in traditional moderation modelling, in this framework the value of

the moderator is not person- or item-specific but rather response-specific, meaning that the value of the moderator (i.e., residual log-RT) differs both across items and across persons. The effect of the moderator is decomposed into an item-specific, person-specific, or item- and person-specific effect, and possibly their interaction. As will be explained, some decompositions result in models that are equivalent or similar to the previously mentioned existing models for CD of RT and RA.

The rest of the paper is organized as follows. In Section 2 we describe the specification of the hierarchical model assuming CI and propose a set of nested hierarchical response moderation models in which the effect of residual RT on RA is specified in different ways. In Section 3 a Bayesian estimation procedure is proposed and model selection is discussed. Simulation studies evaluating parameter recovery and performance of the information criteria for model selection are presented in Section 4. Two empirical examples are presented in Section 5. The paper concludes with a discussion.

2. Response moderation models

Before introducing response moderation models, we briefly describe the hierarchical model for RT and RA that assumes CI (van der Linden, 2007). A full specification of the hierarchical model for RTs and RAs includes the models for RT and for RA at the lower level, and the models for the relationship between the item parameters and for the relationship between the person parameters at the higher level. In this paper we consider a two-parameter log-normal model for RTs (van der Linden, 2006) and a two-parameter normal ogive model for RA (Lord & Novick, 1968). The log-RT is assumed to be normally distributed with the mean defined as the difference between the item time intensity (ξ_i) and person speed (τ_p) and the item-dependent residual variance (σ_i^2):

$$\ln t_{pi} \sim \mathcal{N}(\xi_i - \tau_p, \sigma_i^2). \tag{1}$$

The probability of a correct response is modelled as

$$\Pr(x_{pi} = 1) = \Phi(\alpha_i \theta_p + \beta_i) \text{ [Model 0]}, \tag{2}$$

where θ_p is the person's ability, α_i is the item discrimination which specifies the strength of the relationship between the item response and ability, β_i is the easiness parameter, and $\Phi(\cdot)$ denotes the standard normal cumulative distribution function.

The person parameters θ_p and τ_p are assumed to have a bivariate normal distribution with the mean vector constrained to zero, and $\sigma_\theta^2 = 1$ to ensure that the model is identified. The correlation $\rho_{\theta\tau}$ is taken to fully explain the correlation between the RA and RT on the same item across persons. The item parameters α_i, β_i, and ξ_i are assumed to have a truncated (i.e., $\alpha_i > 0$) multivariate normal distribution with mean vector $\boldsymbol{\mu}_\mathcal{I}$ and covariance matrix $\Sigma_\mathcal{I}$. The correlation between β_i and ξ_i is taken to fully explain the correlation between the RT and RA of the same person across items.

In practice the assumption of CI between RT and RA does not necessarily hold. There may be residual dependence left after the relationship between θ and τ and the relationship between β and ξ are taken into account. Responses that are faster than expected, given the person's speed and the item's time intensity, might be more or less often correct than the responses that are slower than expected. To quantify whether a response is relatively fast or relatively slow we use the standardized residual of log-RT from the log-normal model in Equation 1:

$$z_{pi} = \frac{\ln t_{pi} - (\xi_i - \tau_p)}{\sigma_i}. \tag{3}$$

If $z_{pi} < 0$ the response of person p to item i is relatively fast, while if $z_{pi} > 0$ the response is relatively slow. If CI holds then x_{pi} is independent of z_{pi}. Linear CD between RT and RA can in its most general form be specified in the conditional accuracy model by including the effect of z_{pi} on RA:

$$\Pr(x_{pi} = 1|z_{pi}) = \Phi\big(\alpha_i\theta_p + \beta_i + \eta_{pi}z_{pi}\big), \tag{4}$$

where η_{pi} is the effect of residual log-RT on RA of person p on item i. Let us denote by N the number of persons in the sample and by K the number of items in the test. Although theoretically η_{pi} can be considered for each combination of a person and an item, $N \times K$ response-level effects η_{pi} cannot be identified. However, we can put a structure on these effects, for example by decomposing η_{pi} into an item and/or a person component. Below we consider a set of hierarchical response moderation models that each propose a different decomposition of η_{pi} to capture the CD between RT and RA given overall speed and ability. Each of these models provides a different balance between two extremes: not modelling CD at all (i.e., $\eta_{pi} = 0$ as in the standard hierarchical model) versus modelling each person by item effect η_{pi} individually (resulting in an unidentifiable model). The different models and their nested structure are illustrated in Figure 1, and will be introduced below.

To start, one can consider the following conditional accuracy model:

$$\Pr(x_{pi} = 1|z_{pi}) = \Phi\big(\alpha_i\theta_p + \beta_i + \lambda_i z_{pi}\big) \text{ [Model 1a]}, \tag{5}$$

in which the effect of residual log-RT is item-specific ($\eta_{pi} = \lambda_i$) and there is no variation of effects across persons. The parameters α_i and β_i are the baseline discrimination and easiness, that is, the discrimination and the easiness of item i given that the log-RT is equal to its expected value.

If $\lambda_i > 0$, the probability of a correct response is higher when the response is slower than expected given the speed of the person and the time intensity of the item. If $\lambda_i < 0$,

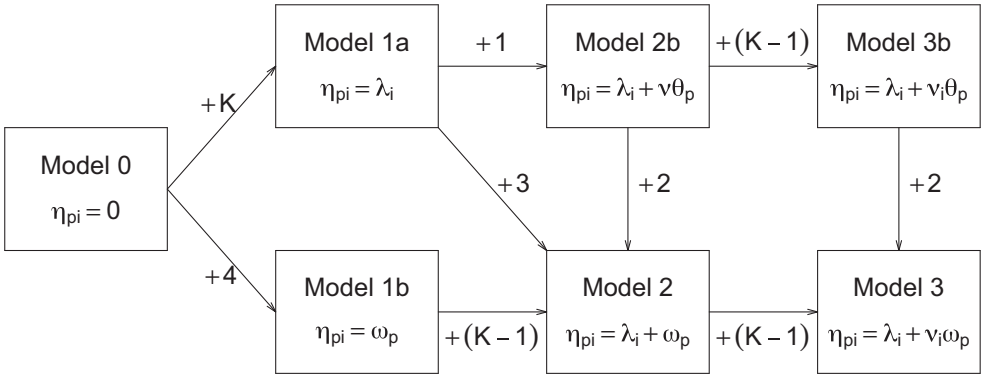

Figure 1. Relationships between the different response moderation models. An arrow indicates that a model is a special case of another model, and the number next to the arrow indicates the number of extra parameters that are added in moving to the more complex model.

the probability of a correct response is higher when the log-RT is shorter than expected. CI between RT and RA holds if $\lambda_i = 0$ for all items. The extra item parameter λ_i can be included in the hierarchical structure together with the other item parameters, that is, for Model 1a the mean vector $\boldsymbol{\mu}_{\mathcal{I}}$ is of length 4, and $\Sigma_{\mathcal{I}}$ is a 4×4 matrix.

Model 1a is equivalent to the model of Ranger and Ortner (2012) in which the residual dependence between RT and RA is modelled by introducing an item-specific residual correlation parameter ρ_i for RT and RA. The extra parameters in the two models are related to each other in the following way: $\lambda_i = \rho_i / \sqrt{1 - \rho_i^2}$.

While Model 1a introduces an item-specific component to explain residual dependence between RA and RT, one could alternatively consider extending the hierarchical model assuming CI by introducing a person-specific component,

$$\Pr(x_{pi} = 1 | z_{pi}) = \Phi\big(\alpha_i \theta_p + \beta_i + \omega_p z_{pi}\big) \text{ [Model 1b]}, \tag{6}$$

in which the effect of residual log-RT is determined by the person component ω_p. This person component (unlike the item component in Model 1a) is treated as a random effect. This way, the number of model parameters that need to be estimated (and contrasted when comparing models) remains manageable, while we still retain the information about the variance of the person-specific component ω_p and its relationship to the latent variables of interest (θ_p, τ_p). The relationship between the three person parameters can be modelled using a multivariate normal distribution. For persons with a positive ω_p responses that are slower than expected are also more accurate than expected, while they are less accurate than expected for persons with a negative ω_p. Unlike the mean of θ and τ, the mean of ω (denoted by μ_ω) can be freely estimated. CI between RT and RA holds if $\mu_\omega = \sigma_\omega^2 = 0$, where σ_ω^2 denotes the variance of ω.

As an extension of Models 1a and 1b, one can consider a model that includes both an item- and a person-specific component,

$$\Pr(x_{pi} = 1 | z_{pi}) = \Phi\big(\alpha_i \theta_p + \beta_i + (\lambda_i + \omega_p) z_{pi}\big) \text{ [Model 2]}, \tag{7}$$

that is, where the effect of residual log-RT is both item- and person-specific: $\eta_{pi} = \lambda_i + \omega_p$. For all λ_i to be freely estimated, we set $\mu_\omega = 0$.

Model 2 allows for differences between the item discriminations depending on the residual log-RT. This becomes apparent by rewriting the conditional accuracy model of Model 2 as

$$\Pr(x_{pi} = 1 | z_{pi}) = \Phi((\alpha_i + \rho_{\theta\omega}\sigma_\omega z_{pi})\theta_p + \beta_i + \lambda_i z_{pi} + \omega_p^* z_{pi}), \tag{8}$$

where ω_p is decomposed as $\rho_{\theta\omega}\sigma_\omega\theta + \omega_p^*$, and ω_p^* is independent of ability. The conditional discrimination of item i (i.e., the strength of the relationship between θ and the probability of a correct response given the residual log-RT) is equal to $\alpha_i + \rho_{\theta\omega}\sigma_\omega z_{pi}$. If the person-specific component of the effect is correlated with ability ($\rho_{\theta\omega} \neq 0$), then there will be an interaction effect between θ_p and z_{pi} on RA, that is, the relationship between the probability of a correct response and the residual log-RT will be different for persons with different levels of ability, or, to put it another way, the relationship between the probability of a correct response and the ability (item discrimination) will be different for relatively slow and relatively fast responses. If the person-specific component ω_p is fully determined by ability, then the model can be reduced to

$$\Pr(x_{pi} = 1|z_{pi}) = \Phi\big((\alpha_i + \nu z_{pi})\theta_p + \beta_i + \lambda_i z_{pi}\big) \text{ [Model 2b]}, \tag{9}$$

where ν can be viewed as the effect of residual RT on item discrimination. This model is similar to the model of Bolsinova *et al.* (2016). The differences from their model are that ν is the linear effect on the discrimination instead of a linear effect on the log-discrimination, and that the normal ogive model is used instead of the logistic model. Model 2b has two parameters fewer than Model 2, since the correlations $\rho_{\theta\omega}$ and $\rho_{\tau\omega}$ are constrained.

Model 2 can be extended by allowing the variance of the person-specific effect to differ across items, resulting in the following model:

$$\Pr(x_{pi} = 1|z_{pi}) = \Phi\big(\alpha_i\theta_p + \beta_i + (\lambda_i + \nu_i\omega_p)z_{pi}\big) \text{ [Model 3]}. \tag{10}$$

That is, the effect of residual log-RT is item- and person-specific, $\eta_{pi} = \lambda_i + \nu_i\omega_p$, but compared to Model 2 the influence of the person on the effect of z_{pi} varies across items. The person-specific effect might differ not only in magnitude but also in sign. For identification, $\sigma^2_\omega = 1$ and $\rho_{\theta\omega} > 0$, such that ν_i can be freely estimated. The new item parameter ν_i can be included in the hierarchical structure with other item parameters, that is, in Model 3 the mean vector $\boldsymbol{\mu}_\mathcal{I}$ is of length 5, and $\Sigma_\mathcal{I}$ is a 5×5 matrix.

Analogous to Model 2, Model 3 can be reduced to a model in which the person-specific effect is perfectly correlated with ability:

$$\Pr(x_{pi} = 1|z_{pi}) = \Phi\big((\alpha_i + \nu_i z_{pi})\theta_p + \beta_i + \lambda_i z_{pi}\big), \text{ [Model 3b]} \tag{11}$$

where ν_i can be viewed as the effect of residual log-RT on discrimination.

The different models presented in this section each provide a decomposition of the possible moderation effect of residual RT on RA, and are nested within each other. Figure 1 presents their relationships. The nested structure depicted in Figure 1 also provides guidance for model selection procedures that can be used to capture CD between RA and RT. We propose to consider stepwise extensions of the hierarchical model that assumes CI in order to determine which forms of CD are supported by the data, as will be discussed in the following section.

3. Estimation and model selection

3.1. Estimation

For the estimation of the models introduced in the previous section we developed Gibbs samplers (Geman & Geman, 1984; Casella & George, 1992) implemented in the R programming language (R Core Team, 2014); see Appendix for details. These algorithms are used to obtain samples from the joint posterior distribution of the model parameters. In the case of the full model (Model 3) the joint posterior distribution is

$$p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\sigma}^2, \boldsymbol{\tau}, \boldsymbol{\lambda}, \boldsymbol{\nu}, \sigma^2_\tau, \rho_{\theta\tau}, \rho_{\theta\omega}, \rho_{\tau\omega}, \boldsymbol{\mu}_\mathcal{I}, \boldsymbol{\Sigma}_\mathcal{I}|\mathbf{X}, \mathbf{T}). \tag{12}$$

Although in the covariance matrix of the person parameters the variances of θ and ω are constrained to 1, to simplify the conditional posteriors of the model parameters and to improve the convergence of the model at each iteration of the Gibbs sampler, the full covariance matrix $\Sigma_\mathcal{P}$ is sampled and at the end of each iteration the model parameters are transformed such that $\sigma^2_\theta = \sigma^2_\omega = 1$ (see Appendix for details). The following vague prior

distributions are used for the hyperparameters: a multivariate normal distribution with a zero mean vector, variances equal to 100 and correlations equal to 0 for $\boldsymbol{\mu}_{\mathcal{I}}$; inverse-Wishart distributions with five degrees of freedom and identity matrix $\mathbf{I}_3$ as the scale parameter for $\Sigma_{\mathcal{P}}$, and with seven degrees of freedom and identity matrix $\mathbf{I}_5$ as the scale parameter for $\Sigma_{\mathcal{I}}$. In the case of Model 1b the prior for μ_ω also has to be specified. We use a vague normal prior with a mean of zero and a variance equal to 100. Independent improper priors are used for σ_i^2,

$$\prod_i p(\sigma_i^2) \propto \prod_i \frac{1}{\sigma_i^2}. \tag{13}$$

Parameter recovery is evaluated in a simulation study (see Section 4.1).

3.2. Model selection

When selecting the best model for RT and RA among the models described in Section 2, both model fit and model complexity should be taken into account. We propose to use two information criteria for model selection, inspired by the Akaike information criterion (AIC, Akaike, 1974) and the Bayesian information criterion (BIC, Schwarz, 1978). We use the same terms for penalizing model complexity as these information criteria: $2P$ and $\ln(N)P$, respectively, where P is the number of extra parameters compared to Model 0. However, unlike the original AIC and BIC which use minus twice the log-likelihood computed at the maximum likelihood estimates of the parameters, we use minus twice the log-likelihood computed at the posterior means of the parameters, since we are using a Bayesian algorithm for estimation. Due to this difference we will call the information criteria used in this paper the *modified* AIC and the *modified* BIC.

The log-likelihood for Models 0, 1a, 2b, and 3b is computed as follows:

$$\begin{aligned}\ln \mathcal{L}(\mathbf{X}, \mathbf{T}) = \sum_p \ln \prod_i \frac{1}{\sqrt{2\pi}\sigma_i t_{pi}} \exp\left(-\frac{(\ln t_{pi} - \xi_i + \tau_p)^2}{2\sigma_i^2}\right) \times \\ \prod_i \left(1 - x_{pi} + (2x_{pi} - 1)\Phi\left(\alpha_i\theta_p + \beta_i + (\lambda_i + \nu_i\theta_p)\frac{(\ln t_{pi} - \xi_i + \tau_p)}{\sigma_i}\right)\right),\end{aligned} \tag{14}$$

where $\lambda_i = \nu_i = 0$ for all items in Model 0, $\nu_i = 0$ for all items in Model 1a, and $\nu_i = \nu$ for all items in Model 2b. For Models 1b, 2, and 3 the log-likelihood is computed as follows:

$$\begin{aligned}\ln \mathcal{L}(\mathbf{X}, \mathbf{T}) = \sum_p \ln \mathcal{N}(\omega|\theta_p, \tau_p, \boldsymbol{\Sigma}_{\mathcal{P}}, \mu_\omega) \prod_i \frac{1}{\sqrt{2\pi}\sigma_i t_{pi}} \exp\left(-\frac{(\ln t_{pi} - \xi_i + \tau_p)^2}{2\sigma_i^2}\right) \times \\ \prod_i \left(1 - x_{pi} + (2x_{pi} - 1)\Phi\left(\alpha_i\theta_p + \beta_i + (\lambda_i + \nu_i\omega)\frac{(\ln t_{pi} - \xi_i + \tau_p)}{\sigma_i}\right)\right) d\omega,\end{aligned} \tag{15}$$

where the extra person parameter ω is integrated out using the conditional normal distribution given the values of ability and speed; $\lambda_i = 0$ for all items in Model 1b, $\nu_i = 1$ for all items in Models 1b and 2, $\mu_\omega = 0$ in Models 2 and 3. Numerical integration can be

used to approximate the integral in Equation 15. Here we use Gauss–Hermite quadrature with 30 nodes. The number of extra parameters P compared to Model 0 is equal to K, 4, $K + 3$, $K + 1$, $2K + 2$, and $2K$ for Models 1a, 1b, 2, 2b, 3, and 3b, respectively. The performance of the modified AIC and the modified BIC is evaluated in a simulation study (see Section 4.2).

4. Simulation studies

4.1. Parameter recovery

In the first simulation study we evaluated how well the item parameters can be recovered for the most complex model introduced in this paper (Model 3; see Equation 10). Parameter recovery was evaluated in five conditions, which differed from each other in either the sample size or the number of items. As a baseline condition we used $N = 1500$ and $K = 30$. Next, two conditions were used in which the sample size was either twice as large or twice as small as in the baseline condition ($N = 3000$, $K = 30$ and $N = 750$, $K = 30$). The last two conditions were those in which the number of items was either twice as large or twice as small as in the baseline condition ($N = 1500$, $K = 60$ and $N = 1500$, $K = 15$).

The true values for the item parameters $[\alpha_i, \beta_i, \xi_i, \lambda_i, \nu_i]$ were sampled from a truncated multivariate normal distribution ($\alpha_i > 0$) with the following hyperparameters: the mean vector [1, 0, 4, 0, 0], the vector of variances [0.2, 0.5, 0.5, 0.2, 0.2], and the correlation matrix

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -.7 & -.5 & 0 \\ 0 & -.7 & 1 & .5 & 0 \\ 0 & -.5 & .5 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \tag{16}$$

Residual variances σ_i^2 were sampled from $\ln \mathcal{N}(\ln(0.3), 0.2)$. The same values of the item parameters were used in each replication. First, the values were sampled for a set of 60 items. Second, from this set a subset of 30 items was randomly sampled for the conditions with $K = 30$. Finally, from this subset of 30 items a subset of 15 items was randomly sampled for the condition with $K = 15$.

In each replication, values for the person parameters $[\theta_p, \tau_p, \omega_p]$ were sampled from a multivariate normal distribution with a zero mean vector and covariance matrix

$$\begin{bmatrix} 1 & 0 & .5 \\ 0 & .1 & 0 \\ .5 & 0 & 1 \end{bmatrix}, \tag{17}$$

and the data were generated according to Model 3 (see Equation 10). In each condition 100 data sets were simulated, Model 3 was fitted with the Gibbs sampler with 50,000 iterations (including 10,000 burn-in iterations), and the estimates of the parameters (i.e., the posterior means) were obtained using every 50th iteration (i.e., 50 thinning) after the burn-in to get rid of the effect of autocorrelation. Bias, variance, and mean squared error (MSE) of the estimates of the item parameters were computed.

Table 1. Absolute bias (Bias), variance (Var), and mean squared error (MSE) of the estimates of the item parameters averaged across items.

		α_i			β_i			ξ_i		
N	K	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
1500	30	0.014	0.005	0.005	0.007	0.004	0.004	0.002	0.000	0.000
1500	15	0.012	0.007	0.007	0.009	0.004	0.004	0.001	0.000	0.000
1500	60	0.011	0.005	0.005	0.011	0.003	0.004	0.002	0.000	0.000
750	30	0.022	0.009	0.010	0.013	0.007	0.008	0.002	0.001	0.001
3000	30	0.008	0.003	0.003	0.005	0.002	0.002	0.002	0.000	0.000
		σ_i^2			λ_i			ν_i		
		Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
1500	30	0.001	0.000	0.000	0.008	0.003	0.003	0.015	0.005	0.006
1500	15	0.001	0.000	0.000	0.006	0.003	0.003	0.019	0.007	0.008
1500	60	0.001	0.000	0.000	0.008	0.002	0.003	0.010	0.004	0.004
750	30	0.001	0.000	0.000	0.014	0.005	0.005	0.025	0.010	0.012
3000	30	0.001	0.000	0.000	0.005	0.001	0.001	0.010	0.003	0.003

In Table 1 the absolute bias, variance, and MSE are presented averaged for the α_i, β_i, ξ_i, σ_i^2, λ_i, and ν_i in the five simulation conditions. For all conditions the results were acceptable. The absolute bias decreased slightly when N increased because with the smaller N the shrinkage effect was stronger. The variance and MSE also decreased with N.

Figure 2 shows the estimates of the item parameters averaged across replications in the baseline condition plotted against their true values. The estimates lie almost perfectly on the identity line. For the α_i and ν_i the highest values were slightly underestimated, while the lowest values were slightly overestimated. This was to be expected since a hierarchical prior was used and the parameter estimates were shrunk to the mean.

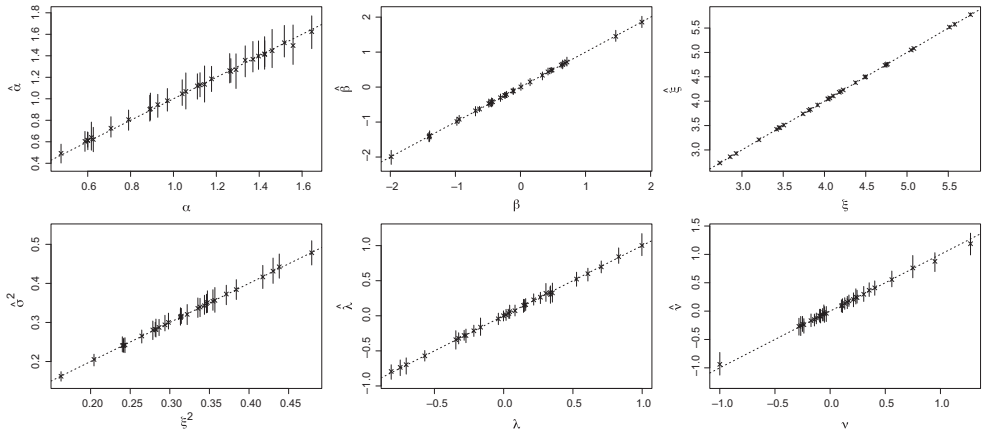

Figure 2. Estimates of the item parameters averaged across 100 replications (on the *y*-axis) against the true values of the item parameters (on the *x*-axis) in the baseline condition (N=1500, K=30). Vertical lines represent the range between the 2nd and 99th percentiles of the distribution of the estimates across replications.

4.2. Performance of the information criteria for model selection

Data were generated under each of the seven models introduced in Section 2. The item and person parameters for Model 3 were sampled from the same distributions as in the first simulation study. For the other models the parameters were sampled analogously, but with the following constraints:

$$\begin{aligned}
&\text{Model 0: } \lambda_i = \nu_i = 0, \forall i \in [1:n], \omega_p = 0, \forall p \in [1:N];\\
&\text{Model 1a: } \nu_i = 0, \forall i \in [1:n],\ \omega_p = 0, \forall p \in [1:N];\\
&\text{Model 1b: } \lambda_i = 0, \nu_i = -\sqrt{0.1}, \forall i \in [1:K];\\
&\text{Model 2: } \nu_i = -\sqrt{0.1}, \forall i \in [1:K];\\
&\text{Model 2b: } \nu_i = -\sqrt{0.1}, \forall i \in [1:K],\ \omega_p = \theta_p, \forall p \in [1:N];\\
&\text{Model 3b: } \omega_p = \theta_p, \forall p \in [1:N].
\end{aligned} \tag{18}$$

The parameters ν_i were chosen to be negative in Models 2 and 2b, because this corresponds to a negative effect of residual log-RT on item discrimination as has been found in the literature (e.g., Partchev & de Boeck, 2012; Bolsinova *et al.*, 2016); and the absolute value of ν_i was chosen to be close to the average absolute value of these parameters in Model 3.

All seven models were fitted to each data set using a Gibbs sampler (50,000 iterations, 10,000 burn-in, 50 thinning) and the modified AIC and modified BIC were computed. The same five conditions as in the first simulation study were used, and in each condition 10 replications were performed.

Table 2 shows the cross-tabs of true versus selected model based on the modified AIC and the modified BIC in each of the conditions. The modified AIC performed well (i.e., in almost all replications the true model was chosen as the best model) in all conditions except the one with the small number of items ($N = 1500$, $K = 15$), in which a more complex model was sometimes selected instead of the true model. The BIC performed well in all conditions, even when $K = 15$. Given these results, we recommend using the modified BIC for model selection.

5. Empirical examples

5.1. Math Garden

5.1.1 Data and method

The models introduced in this paper were applied to data from Math Garden, (Klinkenberg, Straatemeier & van der Maas, 2011; Straatemeier, 2014) which is an online adaptive practice system for arithmetics. Children can practise in several mathematics domains, such as addition, multiplication, or fractions. For this study the multiplication domain was used, more specifically all items from the single-digit tables of multiplication (e.g., '$2 \times 7 =$' and '$9 \times 1 =$'), meaning that 81 open-ended items were used. In Math Garden items are adaptively matched to the students for facilitation of learning and motivation. The items need to be answered within 20 seconds.

Responses collected in the period from 1 September to 1 October 2015 were used for analysis. For each person only responses given within a single day (the one with the most responses) were selected to make sure that ability and speed were relatively stable across the responses. Data from all persons with at least 15 responses were used, resulting in a

Table 2. Results of the simulation study investigating the performance of the modified AIC and the modified BIC for model selection in the response moderation framework.

		Selected model based on the modified AIC																				
		0	1a	1b	2	2b	3	3b	0	1a	1b	2	2b	3	3b	0	1a	1b	2	2b	3	3b
		N=1500, K=30							N=1500, K=15							N=1500, K=60						
True model	0	**10**	0	0	0	0	0	0	**10**	0	0	0	0	0	0	**9**	0	1	0	0	0	0
	1a	0	**9**	0	0	1	0	0	0	**6**	0	1	1	2	0	0	**10**	0	0	0	0	0
	1b	0	0	**10**	0	0	0	0	0	0	**8**	1	0	1	0	0	0	**10**	0	0	0	0
	2	0	0	0	**10**	0	0	0	0	0	0	**10**	0	0	0	0	0	0	**10**	0	0	0
	2b	0	0	0	0	**10**	0	0	0	0	0	0	**8**	0	2	0	0	0	0	**10**	0	0
	3	0	0	0	0	0	**10**	0	0	0	0	0	0	**10**	0	0	0	0	0	0	**10**	0
	3b	0	0	0	0	0	0	**10**	0	0	0	0	0	2	**8**	0	0	0	0	0	1	**9**
		N=750, K=30							N=3000, K=30													
True model	0	**10**	0	0	0	0	0	0	**10**	0	0	0	0	0	0							
	1a	0	**9**	0	0	1	0	0	0	**10**	0	0	0	0	0							
	1b	0	0	**10**	0	0	0	0	0	0	**10**	0	0	0	0							
	2	0	0	0	**10**	0	0	0	0	0	0	**10**	0	0	0							
	2b	0	0	0	0	**10**	0	0	0	0	0	0	**10**	0	0							
	3	0	0	0	0	0	**10**	0	0	0	0	0	0	**10**	0							
	3b	0	0	0	0	0	1	**9**	0	0	0	0	0	0	**10**							

Continued

Table 2. *(Continued)*

		Selected model based on the modified BIC																			
		0	1a	1b	2	2b	3	3b	0	1a	1b	2	2b	3	3b	0	1a	1b	2	2b	3	3b
		N=1500, K=30							N=1500, K=15							N=1500, K=60						
True model	0	**10**	0	0	0	0	0	0	**10**	0	0	0	0	0	0	**9**	0	1	0	0	0	0
	1a	0	**10**	0	0	0	0	0	0	**10**	0	0	0	0	9	0	**10**	0	0	0	0	0
	1b	0	0	**10**	0	0	0	0	0	0	**10**	0	0	0	0	0	0	**10**	0	0	0	0
	2	0	0	0	**10**	0	0	0	0	0	0	**10**	0	0	0	0	0	0	**10**	0	0	0
	2b	0	0	0	0	**10**	0	0	0	0	0	0	**10**	0	0	0	0	0	0	**10**	0	0
	3	0	0	0	0	0	**10**	0	0	0	0	0	0	**10**	0	0	0	0	0	0	**10**	0
	3b	0	0	0	0	0	0	**10**	0	0	0	0	0	1	**9**	0	0	0	0	0	0	**10**
		N=750, K=30							N=3000, K=30													
True model	0	**10**	0	0	0	0	0	0	**10**	0	0	0	0	0	0							
	1a	0	**10**	0	0	0	0	0	0	**10**	0	0	0	0	0							
	1b	0	0	**10**	0	0	0	0	0	0	**10**	0	0	0	0							
	2	0	0	0	**10**	0	0	0	0	0	0	**10**	0	0	0							
	2b	0	0	0	0	**10**	0	0	0	0	0	0	**10**	0	0							
	3	0	0	0	0	0	**10**	0	0	0	0	0	0	**10**	0							
	3b	0	0	0	0	0	0	**10**	0	0	0	0	0	0	**10**							

sample size of 2,989, with an average of 19 responses per person and 699 responses per item. The models were fitted using Gibbs samplers with 250,000 iterations (125,000 burn-in, 50 thinning) and compared using the modified AIC and BIC.

5.1.2 Results

Table 3 shows the information criteria for the seven models. While including between-person differences in CD improves the model (compare Model 1b with Model 0 in Table 3), including between-item differences in CD does not improve the model (compare Model 1a with Model 0, and Model 2 with Model 1b in Table 3). While the modified AIC preferred Model 2, the modified BIC preferred the more parsimonious Model 1b. Since the simulation results indicated that the BIC should be preferred, the remainder of this section will focus on the results of Model 1b.

In Table 4 the estimates of the hyperparameters are presented. The mean of ω is negative (-0.07), meaning that for an average person the probability of a correct response is larger if the response is faster than expected under the log-normal model than if it is slower than expected. In the multiplication domain the relatively faster responses might be a result of a better-automated process. That is, a child that has already learned a certain part of the multiplication table can just retrieve the answer from memory instead of having to do multiplication by repeated addition, where the latter takes more time and is more prone to error. Since ω is negatively correlated with ability ($-.24$), this effect is more pronounced among high-ability children. However, given the variance of ω (0.13), it does not hold for all children. The conditional discrimination of each item i is equal to $\alpha_i - 0.08z_{pi}$, with the 95% credible interval for the effect on discrimination being $[-0.11,-0.06]$. That is, the relationship between RA and ability is stronger for fast responses than for slow responses.

For part of the population the CD between RT and RA is positive, which might be due to the fact that these children give fast random responses when they find items too difficult. Note that children who often give a lot of random fast responses would have higher τ and likely a positive ω, which is consistent with the positive correlation $\rho_{\tau\omega} = .23$. Furthermore, for some children a positive ω might be an indication of a speed–accuracy trade-off: for some items they might have to increase their speed, resulting in lower accuracy, while for other items they answer relatively slowly, resulting in improved accuracy. Speed and ability together explain 15.1% ([10.5,20.0]%) of the variance of ω, which means that other personal characteristics (e.g., motivation, cognitive flexibility, or

Table 3. Values of the modified AIC (AICm) and modified BIC (BICm) for the Math Garden data for the seven fitted hierarchical models, with the number of extra parameters compared to Model 0 indicated by *P*.

Model	Moderation parameters	*P*	AICm	BICm
Model 0	—	0	94949.02	94949.02
Model 1a	$\boldsymbol{\lambda}$	81	94953.80	95440.02
Model 1b	$\mu_\omega, \sigma^2_\omega, \rho_{\theta\omega}, \rho_{\tau\omega}$	4	94259.81	94283.82
Model 2	$\boldsymbol{\lambda}, \sigma^2_\omega, \rho_{\theta\omega}, \rho_{\tau\omega}$	84	94131.03	94635.26
Model 2b	$\boldsymbol{\lambda}, \nu$	82	94521.57	95013.79
Model 3	$\boldsymbol{\lambda}, \boldsymbol{\nu}, \rho_{\theta\omega}, \rho_{\tau\omega}$	164	94211.50	95195.94
Model 3b	$\boldsymbol{\lambda}, \boldsymbol{\nu}$	162	94548.44	95520.88

Table 4. Posterior means and 95% credible intervals of the hyperparameters under Model 1b for the Math Garden data (for the correlations the estimates are below the diagonal and the credible intervals are above the diagonal).

		α	β	ξ		θ	τ	ω
Means		0.88	1.09	1.85		0	0	−0.07
		[0.81,0.96]	[0.91,1.25]	[1.78,1.92]		—	—	[−0.10,−0.05]
Variances		0.07	0.56	0.09		1	0.10	0.13
		[0.05,0.10]	[0.40,0.77]	[0.07,0.13]		—	[0.10,0.11]	[0.11,0.14]
Correlations	α	1	[−.73,−.35]	[.40,.73]	θ	1	[.22,.30]	[−.31,−.17]
	β	−.56	1	[−.91,−.78]	τ	.26	1	[.16,.29]
	ξ	.59	−.85	1	ω	−.24	.23	1

Table 5. Values of the modified AIC (AICm), and the modified BIC (BICm) for the arithmetic test data for the seven fitted hierarchical models, with the number of extra parameters compared to Model 0 indicated by *P*.

Model	Moderation parameters	*P*	AICm	BICm
Model 0	—	0	554736.4	554736.4
Model 1a	$\boldsymbol{\lambda}$	50	549575.0	549896.3
Model 1b	$\mu_\omega, \sigma^2_\omega, \rho_{\theta\omega}, \rho_{\tau\omega}$	4	554028.7	554054.4
Model 2	$\boldsymbol{\lambda}, \sigma^2_\omega, \rho_{\theta\omega}, \rho_{\tau\omega}$	53	549206.9	549134.6
Model 2b	$\boldsymbol{\lambda}, \nu$	51	549376.6	549704.3
Model 3	$\boldsymbol{\lambda}, \boldsymbol{\nu}, \rho_{\theta\omega}, \rho_{\tau\omega}$	102	548795.6	549451.1
Model 3b	$\boldsymbol{\lambda}, \boldsymbol{\nu}$	100	548929.7	549572.3

test taking strategies) may also be very important for explaining the person-specific CD between RT and RA.

5.2. High-stakes arithmetics test

5.2.1 Data and method

The second application is to data of a high-stakes arithmetics test that is a part of the final examination in Dutch secondary education. One of the test versions (containing 50 items) was used, and the corresponding sample size was 4,568.[1]

The CI model and the six response moderation models were fitted to the data using Gibbs samplers with 250,000 iterations (125,000 for burn-in, 50 thinning) and compared using the information criteria.

5.2.2 Results

Table 5 shows the information criteria computed for the seven fitted models. Model 3 has the lowest modified AIC and modified BIC, therefore we present the results of this model. In Table 6 the estimates of the item hyper-parameters are presented. The mean of λ_i is positive, that is, for a combination of an average person with an average item, the probability of a correct response is higher given a relatively slow response than given a relatively fast response. The item-specific effect λ_i had a strong negative correlation with the baseline intercept (−.66). Consequently, for most easy items the item-specific component of the effect is negative (i.e., fast responses are more often correct) while for all difficult items the item-specific component of the effect is positive (i.e., slow responses are more often correct), as can be observed in Figure 3 (left).

The correlations between the person parameters were as follows: $\rho_{\theta\tau} = -.10$ [−.13,−.07], $\rho_{\tau\omega} = -.14$ [−.18,−.09], and $\rho_{\theta\omega} = .62$ [.56,.67]. The mean of ν was negative (−0.11) which in combination with a positive $\rho_{\theta\omega}$ can be interpreted as follows: for an average item the conditional discrimination is higher for fast responses than for slow responses, and the effect of residual RT is negative for high-ability students and positive for low-ability students. However, as can be seen from Figure 3 (right), this does not hold for all the items.

[1] For reasons of confidentiality it is not possible to provide the details of the exact items in this test version, but example items can be found at http://www.cito.nl/onderwijs/voortgezet%20onderwijs/rekentoetsvo/voorbeeldtoetsen.

Table 6. Posterior means and 95% credible intervals of the item hyperparameters for the arithmetic test data, with the credible intervals of the correlation estimates displayed above the diagonal.

		α_i	β_i	ξ_i	λ_i	ν_i
Means		0.57	0.13	4.42	0.10	−0.11
		[0.50,0.63]	[−0.11,0.37]	[4.26,4.58]	[0.01,0.19]	[−0.16,−0.05]
Variances		0.06	0.80	0.37	0.11	0.04
		[0.04,0.09]	[0.54,1.20]	[0.25,0.56]	[0.07,0.15]	[0.03,0.06]
Correlations	α_i	1	[−.65,−.20]	[.18,.63]	[.07,.56]	[−.45,.06]
	β_i	−.45	1	[−.854,−.57]	[−.79,−.49]	[−.40,.16]
	ξ_i	.42	−.72	1	[.34,.72]	[−.19,.36]
	λ_i	.33	−.66	.55	1	[−.15,.40]
	ν_i	−.20	−.12	.09	.13	1

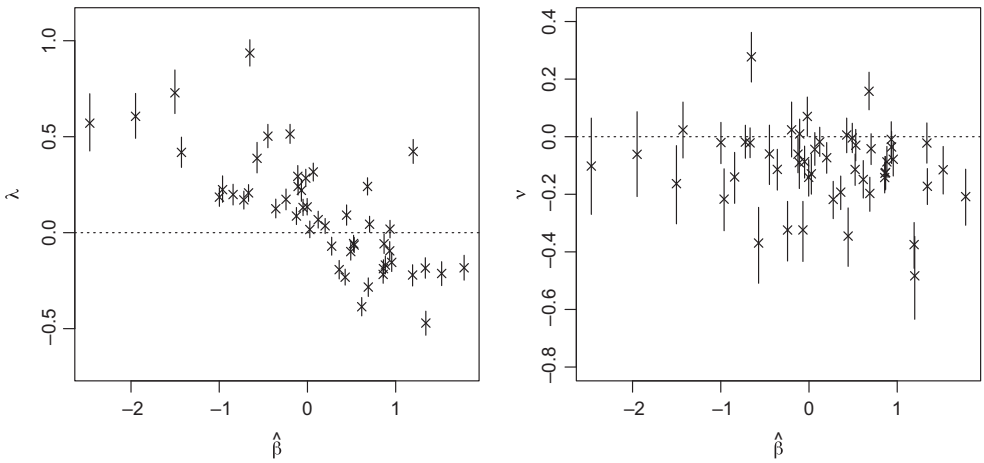

Figure 3. Posterior means of the λ_i (left) and ν_i (right) and their 95% credible intervals plotted against the Posterior means of the baseline easiness under Model 3.

6. Discussion

Standard hierarchical approaches to jointly modelling RT and RA assume CI, which means that the within-person correlation between RT and RA is taken to be be fully explained by the correlation between the item difficulty and time intensity. Likewise, it assumes that the within-item correlation between RT and RA can be fully explained through the correlation between speed and ability. These assumptions may be implausible in real-life applications, for example when people vary their effective speed or use different response strategies. This paper has proposed extensions of the hierarchical model that include both person- and item-specific effects, which can be used to explore the presence and nature of conditional dependencies. With these models, users can abandon the assumption of CI if needed, while still being able to model speed and ability by taking into account both the higher-level correlation between person latent variables and lower-level dependencies in one hierarchical model.

The different proposed models for CD are nested, providing users with a structured way of assessing whether conditional dependencies of a particular type are present. For

example, Models 1a and 1b allow users to assess whether item- or person-specific forms of CD are present. Model 2 allows for both between-item and between-person variation of CD. If support for Model 2 is found, the model can either be extended by allowing the random effect across person to have item-specific variance (Model 3), or be restricted by constraining the person-specific component to be fully determined by ability (Model 2b), which is equivalent to including an effect of the residual RT on discrimination (Bolsinova *et al.*, 2016). Model 3b extends the latter model by allowing the effect on discrimination to be item-specific.

Model 2b and 3b were considered because they match models that were proposed by Bolsinova *et al.* (2016). The two models are special cases of Model 2 and 3, respectively, with the difference being that both former models assume the person component to be perfectly correlated with ability. As the two empirical examples illustrate, the more general models that allow for a weaker correlation between the person component and ability may be more realistic in practice. This means that the between-person differences in the effect of having a relatively fast or slow response cannot be fully explained by ability and likely depend on other traits as well (e.g., some form of cognitive flexibility).

As evidenced by the first simulation study, the estimation procedure successfully recovers the model parameters under reasonable modelling conditions. Even for sample sizes that can be considered relatively small for such complex models (i.e., 750 persons), the accuracy and precision were high for each of the model parameters. The second simulation study indicates that both the information criteria reliably select the appropriate model, with the information criterion based on the BIC slightly outperforming the one based on the AIC. Since the BIC also uses a stronger penalty for adding parameters and hence more strongly favours parsimony, preferring this criterion is also in line with preventing overfitting.

As the empirical examples illustrate, which response moderation model is most appropriate for explaining violations of CI is likely to depend on the application and the sources of the deviation from CI. For example, Model 1a may be appropriate when persons vary their effective speed from item to item, but items differ in the extent to which this affects performance (i.e., expected accuracy). This amounts to stating that the speed–accuracy trade-off function may differ for different items, for example because increasing speed may be more detrimental for performance on one item than on another. If the items are very homogeneous, as in the first empirical example, there may be no relevant differences between the items in this respect, but there may still be important differences between persons with respect to their speed–accuracy trade-off function (Bolsinova & Tijmstra, 2015; Goldhammer, 2015), as captured by Model 1b. Both item and person differences may be present, as captured by Model 2 (and 2b). Furthermore, the between-person differences in CD may be more strongly present for some of the items than for others, as captured by Model 3 (and 3b). These kinds of person-by-item interactions may be expected for tests where the items are relatively heterogeneous (e.g., in type or response format), as was the case in the second empirical example.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723. doi:10.1109/TAC.1974.1100705

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

Bolsinova, M., & Tijmstra, J. (2015). Can response speed be fixed experimentally, and does this lead to unconfounded measurement of ability? *Measurement, 13* 165–168. doi:10.1080/15366367.2015.1105080

Bolsinova, M., de Boeck, P., & Tijmstra, J. (2016). Modeling conditional dependence between response time and accuracy. *Psychometrika*. Manuscript accepted for publication.

Casella, G., & George, E. (1992). Explaining the Gibbs sampler. *The American Statistician*, *46*, 167–174. doi:10.2307/2685208

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721–741. doi:10.1109/TPAMI.1984.4767596

Goldhammer, F. (2015). Measuring ability, speed, or both? challenges, psychometric solutions, and what can be gained from experimantal control. *Measurement*, *13*, 133–164. doi:10.1080/15366367.2015.1100020

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

Klinkenberg, S., Straatemeier, M., & van der Maas, H. (2011). Computer adaptive practice of Maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers and Education*, *57*, 1813–1824. doi:10.1016/j.compedu.2011.02.003

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. doi:10.1111/jedm.12060

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. doi: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. doi:10.1111/bmsp.12042

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

R Core Team (2014). *R: A language and environment for statistical computing [Computer software manual]*. Vienna, Austria. Retrieved from http://www.R-project.org/

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.

Schwarz, G. (1978). Estimating the dimension of the model. *Annals of Statistics*, *6*, 461–464. doi:10.1214/aos/1176344136

Straatemeier, M. (2014). Math *Garden: a new educational and scientific instrument*(Unpublished doctoral dissertation). University of Amsterdam.

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

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

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

Received 1 April 2016; revised version received 8 July 2016

Appendix:

Here we describe how to sample from the posterior distribution of model parameters of Model 3. Modified versions of the algorithm were also developed for Models 0, 1a, 1b, 2, 2b, and 3b (see online supporting information for the R code).

Data augmentation (Tanner & Wong, 1987) is implemented for simplifying the full conditional posteriors of the model parameters. First, for each response x_{pi} an augmented continuous variable $y_{pi} \sim \mathcal{N}(\alpha_i\theta_p + \beta_i + (\lambda_i + \nu_i\omega_p)z_{pi}, 1)$ is introduced (Albert, 1992), defined in such a way that $x_{pi} = \mathcal{I}(y_{pi} \geq 0)$. Second, parameters ω_p for each person are sampled. Hence, in the Gibbs sampler samples are obtained from the following joint posterior distribution:

$$p(\mathbf{Y}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\sigma}^2, \boldsymbol{\tau}, \boldsymbol{\lambda}, \boldsymbol{\nu}, \boldsymbol{\omega}, \sigma^2_\tau, \rho_{\theta\tau}, \rho_{\theta\omega}, \rho_{\tau\omega}, \boldsymbol{\mu}_\mathcal{I}, \boldsymbol{\Sigma}_\mathcal{I} | \mathbf{X}, \mathbf{T}). \tag{19}$$

Let us denote by $\mathbf{U}$ an $N \times K$ matrix such that $u_{pi} = 1$ if a response of person p to item i is observed and 0 otherwise.

In the Gibbs sampler the model parameters are consecutively sampled from their full conditional posterior distributions which are specified in the steps below.

Step 1. For each combination of person p with item i, if $u_{pi} = 1$ sample the augmented response y_{pi} from its full conditional posterior,

$$y_{pi} \sim \mathcal{N}(\alpha_i\theta_p + \beta_i + (\lambda_i + \nu_i\omega_p)z_{pi}, 1)\big(x_{pi}\mathcal{I}(y_{pi} \geq 0) + (1 - x_{pi})\mathcal{I}(y_{pi} < 0)\big), \tag{20}$$

which is a normal distribution truncated below zero if the response x_{pi} is correct, or above zero if the response is incorrect. Note that, given the matrix of augmented responses $\mathbf{Y}$, the model parameters are independent of the actual responses $\mathbf{X}$.

Step 2. For each person p, sample the person parameters:

Step 2a. Sample the ability parameter θ_p from its full conditional posterior,

$$\theta_p \sim \mathcal{N}\left(\frac{\sum_i u_{pi}\alpha_i \left(y_{pi} - \beta_i - (\lambda_i + \nu_i\omega_p)\dfrac{\ln t_{pi} - \xi_i + \tau_p}{\sigma_i} \right) + \dfrac{\mu^*_\theta}{\sigma^{2*}_\theta}}{\sum_i u_{pi}\alpha_i^2 + \dfrac{1}{\sigma^{2*}_\theta}}, \frac{1}{\sum_i u_{pi}\alpha_i^2 + \dfrac{1}{\sigma^{2*}_\theta}} \right), \tag{21}$$

where μ^*_θ and σ^{2*}_θ are the conditional mean and conditional variance of ability, given the other two person parameters (τ_p and ω_p).

Step 2b. Sample the speed parameter τ_p from its full conditional posterior which is a normal distribution with mean

$$\frac{\sum_i u_{pi}\left(\dfrac{\xi_i - \ln t_{pi}}{\sigma_i^2} + \dfrac{(\lambda_i + \nu_i\omega_p)}{\sigma_i}\left(y_{pi} - \alpha_i\theta_p - \beta_i - (\lambda_i + \nu_i\omega_p)\dfrac{\ln t_{pi} - \xi_i}{\sigma_i} \right) \right) + \dfrac{\mu^*_\tau}{\sigma^{2*}_\tau}}{\sum_i u_{pi}\dfrac{1 + (\lambda_i + \nu_i\omega_p)^2}{\sigma_i^2} + \dfrac{1}{\sigma^{2*}_\tau}} \tag{22}$$

and variance

$$\frac{1}{\sum_i u_{pi} \dfrac{1+(\lambda_i+\nu_i\omega_p)^2}{\sigma_i^2}+\dfrac{1}{\sigma_\tau^{2*}}}, \tag{23}$$

where μ_τ^* and σ_τ^{2*} are the conditional mean and conditional variance of speed, given the other two person parameters (θ_p and ω_p).

Step 2c. Sample the person-specific component of the effect of residual RT ω_p from its full conditional posterior which is a normal distribution with mean

$$\frac{\sum_i u_{pi} \dfrac{\nu_i(\ln t_{pi}-\xi_i+\tau_p)}{\sigma_i}\left(y_{pi}-\alpha_i\theta_p-\beta_i-\dfrac{\lambda_i(\ln t_{pi}-\xi_i+\tau_p)}{\sigma_i}\right)+\dfrac{\mu_\omega^*}{\sigma_\omega^{2*}}}{\sum_i u_{pi}\nu_i^2\dfrac{(\ln t_{pi}-\xi_i+\tau_p)^2}{\sigma_i^2}+\dfrac{1}{\sigma_\omega^{2*}}} \tag{24}$$

and variance

$$\frac{1}{\sum_i u_{pi} \dfrac{\nu_i^2(\ln t_{pi}-\xi_i+\tau_p)^2}{\sigma_i^2}+\dfrac{1}{\sigma_\omega^{2*}}}, \tag{25}$$

where μ_ω^* and σ_ω^{2*} are the conditional mean and conditional variance of ω, given the ability and the speed parameters.

Step 3. For each item i, sample the item parameters:

Step 3a. Sample α_i from its full conditional posterior,

$$\alpha_i \sim \mathcal{I}(\alpha_i>0)\mathcal{N}\left(\frac{\sum_p u_{pi}\theta_p\left(y_{pi}-\beta_i-(\lambda_i+\nu_i\omega_p)\dfrac{\ln t_{pi}-\xi_i+\tau_p}{\sigma_i}\right)+\dfrac{\mu_\alpha^*}{\sigma_\alpha^{2*}}}{\sum_p u_{pi}\theta_p^2+\dfrac{1}{\sigma_\alpha^{2*}}},\frac{1}{\sum_p u_{pi}\theta_p^2+\dfrac{1}{\sigma_\alpha^{2*}}}\right), \tag{26}$$

where μ_α^* and σ_α^{2*} are the conditional mean and conditional variance of α_i, given β_i, ξ_i, λ_i, and ν_i.

Step 3b. Sample β_i from its full conditional posterior,

$$\beta_i \sim \mathcal{N}\left(\frac{\sum_p u_{pi}\left(y_{pi}-\alpha_i\theta_p-(\lambda_i+\nu_i\omega_p)\dfrac{\ln t_{pi}-\xi_i+\tau_p}{\sigma_i}\right)+\dfrac{\mu_\beta^*}{\sigma_\beta^{2*}}}{\sum_p u_{pi}+\dfrac{1}{\sigma_\beta^{2*}}},\frac{1}{\sum_p u_{pi}+\dfrac{1}{\sigma_\beta^{2*}}}\right), \tag{27}$$

where μ_β^* and σ_β^{2*} are the conditional mean and conditional variance of β_i, given α_i, ξ_i, λ_i, and ν_i.

Step 3c. Sample ξ_i from its full conditional posterior, which is a normal distribution with mean

$$\frac{\sum_p u_{pi}\left(\frac{\tau_p + \ln t_{pi}}{\sigma_i^2} - \frac{(\lambda_i + \nu_i\omega_p)}{\sigma_i}\left(y_{pi} - \alpha_i\theta_p - \beta_i - (\lambda_i + \nu_i\omega_p)\frac{\ln t_{pi} + \tau_p}{\sigma_i}\right)\right) + \frac{\mu_\xi^*}{\sigma_\xi^{2*}}}{\sum_p u_{pi}\frac{1 + (\lambda_i + \nu_i\omega_p)^2}{\sigma_i^2} + \frac{1}{\sigma_\xi^{2*}}} \tag{28}$$

and variance

$$\frac{1}{\sum_p u_{pi}\frac{1 + (\lambda_i + \nu_i\omega_p)^2}{\sigma_i^2} + \frac{1}{\sigma_\xi^{2*}}}, \tag{29}$$

where μ_ξ^* and σ_ξ^{2*} are the conditional mean and conditional variance of ξ_i, given α_i, β_i, λ_i, and ν_i.

Step 3d. Sample λ_i from its full conditional posterior, which is a normal distribution with mean

$$\frac{\sum_p u_{pi}\frac{\ln t_{pi} - \xi_i + \tau_p}{\sigma_i}\left(y_{pi} - \alpha_i\theta_p - \beta_i - \nu_i\omega_p\frac{\ln t_{pi} - \xi_i + \tau_p}{\sigma_i}\right) + \frac{\mu_\lambda^*}{\sigma_\lambda^{2*}}}{\sum_p u_{pi}\frac{(\ln t_{pi} - \xi_i + \tau_p)^2}{\sigma_i^2} + \frac{1}{\sigma_\lambda^{2*}}} \tag{30}$$

and variance

$$\frac{1}{\sum_p u_{pi}\frac{(\ln t_{pi} - \xi_i + \tau_p)^2}{\sigma_i^2} + \frac{1}{\sigma_\lambda^{2*}}}, \tag{31}$$

where μ_λ^* and σ_λ^{2*} are the conditional mean and conditional variance of λ_i, given α_i, β_i, ξ_i, and ν_i.

Step 3e. Sample ν_i from its full conditional posterior, which is a normal distribution with mean

$$\frac{\sum_p u_{pi}\frac{\omega_p(\ln t_{pi} - \xi_i + \tau_p)}{\sigma_i}\left(y_{pi} - \alpha_i\theta_p - \beta_i - \frac{\lambda_i(\ln t_{pi} - \xi_i + \tau_p)}{\sigma_i}\right) + \frac{\mu_\nu^*}{\sigma_\nu^{2*}}}{\sum_p u_{pi}\omega_p^2\frac{(\ln t_{pi} - \xi_i + \tau_p)^2}{\sigma_i^2} + \frac{1}{\sigma_\nu^{2*}}} \tag{32}$$

and variance

$$\frac{1}{\sum_p u_{pi}\frac{\omega_p^2(\ln t_{pi}-\xi_i+\tau_p)^2}{\sigma_i^2}+\frac{1}{\sigma_\nu^{2*}}}, \tag{33}$$

where μ_ν^* and σ_ν^{2*} are the conditional mean and conditional variance of ν_i, given α_i, β_i, ξ_i, and λ_i.

Step 3f. Sample σ_i^2. Unlike other parameters the residual variance of log RT does not have a full conditional posterior of known form. To sample from its conditional posterior a Metropolis–Hastings algorithm is used. A candidate value c is sampled from a log-normal distribution $\ln\mathcal{N}(\ln(\sigma_i^2), \sqrt{0.05})$ and the acceptance probability is equal to

$$\Pr(\sigma_i^2 \to c) = \frac{\sigma_i^{\sum_p u_{pi}} \exp\left(\sum_p u_{pi}\left(\frac{(\ln t_{pi}-\xi_i+\tau_p)^2}{2\sigma_i^2}+\frac{\left(y_{pi}-\alpha_i\theta_p-\beta_i-(\lambda_i+\nu_i\omega_p)\frac{\ln t_{pi}-\xi_i+\tau_p}{\sigma_i}\right)^2}{2}\right)\right)}{c^{\frac{1}{2}\sum_p u_{pi}} \exp\left(\sum_p u_{pi}\left(\frac{(\ln t_{pi}-\xi_i+\tau_p)^2}{2c}+\frac{\left(y_{pi}-\alpha_i\theta_p-\beta_i-(\lambda_i+\nu_i\omega_p)\frac{\ln t_{pi}-\xi_i+\tau_p}{\sqrt{c}}\right)^2}{2}\right)\right)}. \tag{34}$$

Step 4. Sample the covariance matrix of the item parameters from

$$p(\boldsymbol{\Sigma}_\mathcal{I}|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi},\boldsymbol{\lambda},\boldsymbol{\nu},\boldsymbol{\mu}_\mathcal{I}) \propto p(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi},\boldsymbol{\lambda},\boldsymbol{\nu}|\boldsymbol{\Sigma}_\mathcal{I},\boldsymbol{\mu}_\mathcal{I})p(\boldsymbol{\Sigma}_\mathcal{I}). \tag{35}$$

which, given the inverse-Wishart prior, is known to be an inverse-Wishart distribution (see, for example, Hoff, 2009):

$$\boldsymbol{\Sigma}_\mathcal{I} \sim \text{Inv-Wishart}\left(7+K, \mathbf{I}_5 + \sum_i([\alpha_i,\beta_i,\xi_i,\lambda_i,\nu_i]^T-\mu_\mathcal{I})([\alpha_i,\beta_i,\xi_i,\lambda_i,\nu_i]-\mu_\mathcal{I}^T)\right). \tag{36}$$

Step 5. Sample the mean vector of the item parameters from

$$p(\boldsymbol{\mu}_\mathcal{I}|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi},\boldsymbol{\lambda},\boldsymbol{\nu},\boldsymbol{\Sigma}_\mathcal{I}) \propto p(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi},\boldsymbol{\lambda},\boldsymbol{\nu}|\boldsymbol{\mu}_\mathcal{I},\boldsymbol{\Sigma}_\mathcal{I})p(\boldsymbol{\mu}_\mathcal{I}). \tag{37}$$

With a multivariate normal prior for $\boldsymbol{\mu}_\mathcal{I}$, this conditional posterior is also a multivariate normal with mean vector

$$\left((100\mathbf{I}_5)^{-1}+K\boldsymbol{\Sigma}_\mathcal{I}^{-1}\right)^{-1}\left(\boldsymbol{\Sigma}_\mathcal{I}^{-1}\left(\sum_i\alpha_i,\sum_i\beta_i,\sum_i\xi_i,\sum_i\lambda_i,\sum_i\nu_i\right)^T\right) \tag{38}$$

and covariance matrix $\left((100\mathbf{I}_5)^{-1}+K\Sigma_\mathcal{I}^{-1}\right)^{-1}$.

Step 6. Sample the covariance matrix of person parameters from

$$p(\boldsymbol{\Sigma}_{\mathcal{P}}|\boldsymbol{\theta}, \boldsymbol{\tau}, \boldsymbol{\omega}) \propto p(\boldsymbol{\theta}, \boldsymbol{\tau}, \boldsymbol{\omega}|\boldsymbol{\Sigma}_{\mathcal{P}})p(\boldsymbol{\Sigma}_{\mathcal{P}}), \tag{39}$$

which, given the truncated inverse-Wishart prior, is also a truncated inverse-Wishart distribution (see, for example, Hoff, 2009):

$$\boldsymbol{\Sigma}_{\mathcal{P}} \sim \mathcal{I}(\rho_{\theta\omega} > 0)\text{Inv-Wishart}\left(5 + N, \mathbf{I}_3 + \sum_p (\theta_p, \tau_p, \omega_p)(\theta_p, \tau_p, \omega_p)^T\right). \tag{40}$$

To sample from this truncated distribution we use rejection sampling, that is, values from the unconstrained inverse-Wishart distribution are sampled until $\rho_{\theta\omega} > 0$.

Step 7. Rescale model parameters to equate the variances of θ and of ω to 1:

$$\begin{aligned}
\theta_p &\rightarrow \tfrac{\theta_p}{\sigma_\theta}, \quad \forall p \in [1:N],\\
\alpha_i &\rightarrow \alpha_i\sigma_\theta, \quad \forall i \in [1:K],\\
\omega_p &\rightarrow \tfrac{\omega_p}{\sigma_\omega}, \quad \forall p \in [1:N],\\
\nu_i &\rightarrow \nu_i\sigma_\omega, \quad \forall i \in [1:K],\\
\boldsymbol{\Sigma}_{\mathcal{I}} &\rightarrow \begin{bmatrix}
\sigma_\alpha^2\sigma_\theta^2 & \sigma_{\alpha\beta}\sigma_\theta & \sigma_{\alpha\xi}\sigma_\theta & \sigma_{\alpha\lambda}\sigma_\theta & \sigma_{\alpha\nu}\sigma_\theta\sigma_\omega\\
\sigma_{\alpha\beta}\sigma_\theta & \sigma_\beta^2 & \sigma_{\beta\xi} & \sigma_{\beta\lambda} & \sigma_{\beta\nu}\sigma_\omega\\
\sigma_{\alpha\xi}\sigma_\theta & \sigma_{\beta\xi} & \sigma_\xi^2 & \sigma_{\xi\lambda} & \sigma_{\xi\nu}\sigma_\omega\\
\sigma_{\alpha\lambda}\sigma_\theta & \sigma_{\beta\lambda} & \sigma_{\xi\lambda} & \sigma_\lambda^2 & \sigma_{\lambda\nu}\sigma_\omega\\
\sigma_{\alpha\nu}\sigma_\theta\sigma_\omega & \sigma_{\beta\nu}\sigma_\omega & \sigma_{\xi\nu}\sigma_\omega & \sigma_{\lambda\nu}\sigma_\omega & \sigma_\nu^2\sigma_\omega^2
\end{bmatrix},\\
\boldsymbol{\Sigma}_{\mathcal{P}} &\rightarrow \begin{bmatrix}
1 & \rho_{\theta\tau}\sigma_\tau & \rho_{\theta\omega}\\
\rho_{\theta\tau}\sigma_\tau & \sigma_\tau^2 & \rho_{\tau\omega}\sigma_\tau\\
\rho_{\theta\omega} & \rho_{\tau\omega}\sigma_\tau & 1
\end{bmatrix}.
\end{aligned} \tag{41}$$