Response Mixture Modeling: Accounting for Heterogeneity in Item Characteristics across Response Times

Molenaar, D.; de Boeck, P.

Link to publication

RESPONSE MIXTURE MODELING: ACCOUNTING FOR HETEROGENEITY IN ITEM CHARACTERISTICS ACROSS RESPONSE TIMES

DYLAN MOLENAAR

UNIVERSITY OF AMSTERDAM

PAUL DE BOECK

OHIO STATE UNIVERSITY

In item response theory modeling of responses and response times, it is commonly assumed that the item responses have the same characteristics across the response times. However, heterogeneity might arise in the data if subjects resort to different response processes when solving the test items. These differences may be within-subject effects, that is, a subject might use a certain process on some of the items and a different process with different item characteristics on the other items. If the probability of using one process over the other process depends on the subject's response time, within-subject heterogeneity of the item characteristics across the response times arises. In this paper, the method of response mixture modeling is presented to account for such heterogeneity. Contrary to traditional mixture modeling where the full response vectors are classified, response mixture modeling involves classification of the individual elements in the response vector. In a simulation study, the response mixture model is shown to be viable in terms of parameter recovery. In addition, the response mixture model is applied to a real dataset to illustrate its use in investigating within-subject heterogeneity in the item characteristics across response times.

Key words: item response theory, response time modeling, mixture modeling.

1. Introduction

Many of the current approaches to the analysis of responses and response times assume local independence of the data conditional upon the latent variables in the model (e.g., Ferrando & Lorenzo-Seva, 2007a, 2007b; Molenaar, Tuerlinckx, & Van der Maas, 2015; Ranger, & Ortner, 2011; Thissen, 1983; Van der Linden, 2007; Tuerlinckx & De Boeck, 2005; Van der Maas, Molenaar, Maris, Kievit, & Borsboom, 2011). Specifically, three types of local independence can be distinguished in these models (Bolsinova & Maris, 2016; Bolsinova & Tijmstra, 2016; Ranger & Ortner, 2012; Van der Linden, 2009; Van der Linden & Glas, 2010). That is: (1) local independence between the responses; (2) local independence between the response times; and (3) local independence between the responses and the response times.

Violations of assumptions (1) and (2) are relatively well understood as they are similar to the local independence assumption commonly posed in generalized linear latent variable models (Bartholomew, Knott, & Moustaki, 2011; Mellenbergh, 1994). That is, a residual dependence between two items may indicate that an additional skill is involved in solving these two items but which is not involved in the other items of the same latent variable. For example, it is a

well-established finding that the 'block design' and 'object assembly' subtests of the Wechsler Adult Intelligence Scale (WAIS-III) have a residual dependence while they both involve the latent variable 'perceptual organization' (see, e.g., Dolan et al., 2006). This result can be explained by the motoric component involved in both tasks but not in the others (matrices, picture arrangement, and picture completion). For the response time model, similar explanations for the violation of local independence are possible. For instance, if two items of an arithmetic test both require a large text to be read in order to answer the items, a residual dependency may arise between the response times of these items because of individual differences in reading speed.

In the present paper, we focus on assumption (3), that is, the assumption of local independence between the responses and responses times. In applications to large-scale computerized assessments, Van der Linden and Glas (2010) found this type of local independence to be violated in 56 of the 96 items and Bolsinova and Maris (2016) found this type of local independence to be violated in 25 of the 30 items. Although these dependencies may be small, question arises how such a systematic departure from local independence may occur.

A conditional association between the responses and response times of a given item arises if the item difficulties of the item responses are heterogeneous with respect to the response times (Bolsinova, Tijmstra, Molenaar, & De Boeck, 2017). Similarly, the item discriminations may not be invariant across the response time variable. Thus, the characteristics of the items may not be the same for responses that differ in their response time. Heterogeneity might arise in the item responses if subjects resort to different response processes when solving the test items. These differences may be within-subject effects, that is, a subject might use a certain process on some of the items and a different process with different item characteristics on the other items. If the probability of using one process over the other process depends on the subject's response time, within-subject heterogeneity of the item characteristics across the response times arises. Note that this is a within-subject effect: If the effect is between-subjects, that is, subjects differ in the process that is used during the test, these individual differences are captured in the latent ability and latent speed variable. Within-subject differences in the use of response processes may occur in, for instance, an arithmetic test where subjects solve some of the items by retrieving the answer from memory and other items are solved by actually performing the requested operation, see Grabner et al. 2009. In addition, Carpenter, Just, and Shell (1990) used eye tracking and verbal protocols to show that subject use different response processes in the different items of the Raven progressive matrices test. Other phenomena that indicate within-subject differences in response processes during the test include guessing (Schnipke, & Scrams, 1997), item preknowledge (McLeod, Lewis, & Thissen, 2003), and post-error slowing (Rabbitt, 1979). In all of these examples, heterogeneity in the data will arise across the response times.

A suitable procedure to account for the heterogeneity of the item responses with respect to the response times is the IRT tree model (De Boeck & Partchev, 2012; Jeon & De Boeck, 2016). In this procedure, prior to modeling, the response times are dichotomized into fast responses and slow responses by an item median split or a person median split (Partchev & De Boeck, 2012; DiTrapani, Jeon, De Boeck, & Partchev, 2016). Subsequently, to these data, a model with three latent variables is fit: a latent speed variable, a latent fast ability variable, and a latent slow ability variable. Item characteristics of the fast and slow responses can then be compared across the fast and slow ability variables.

Although valuable, the IRT tree approach requires the continuous response time data to be dichotomized. This has two important implications. First, it makes the approach deterministic. As a result, it is not straightforward to test statistical hypothesis about the exact cut-off point to use to split the response times into two categories. Main problem is that using a different cut-off point will result in a different dataset which invalidates model comparison using likelihood based fit indices like AIC or BIC. In addition, the ad hoc approach of trying out different cut-offs to infer the most optimal cut-off is from our perspective a statistically suboptimal procedure as the

same data are used repeatedly. Second, dichotomization of the response times implies that the amount of information concerning individual differences in the response times is reduced (see, e.g., Cohen, 1983; MacCallum, Zhang, Preacher, & Rucker, 2002) while one of the motivations to add the response times is to increase measurement precision (Van der Linden, Entink, & Fox, 2010).

Therefore, in this paper, we propose an explicit statistical model to account for the heterogeneity of the item characteristics with respect to the response times. The model is referred to as a response mixture model as—contrary to the traditional mixture models where each item response vector is classified into one of the classes—each element of the response vector is classified into one of the classes. The idea of response classification instead of response vector classification has previously been adopted to account for rapid guessing (Schnipke, & Scrams, 1997; Wang & Xu, 2015; Wang, Xu, & Shang, 2016) and to account for zero-inflated responses in repeated measures count data (e.g., Hall, 2000; Min & Agresti, 2005). In addition, the three-parameter model (Birnbaum, 1968) can be seen as a model that classifies responses. That is, in the three-parameter model, the probability of a correct response is modeled using a two-parameter model in one class and a uniform guessing probability in the other class.

We account for possible sources of heterogeneity in the item characteristics with respect to the response times by imposing two item-specific latent classes to underlie the responses of each item. First, the classes are allowed to differ in their item characteristics (discrimination and difficulty). Next, class membership is regressed on the response times to test whether the responses with a relatively large response time are more likely to belong to one class (a 'slower class') and the responses with a relatively small response time are more likely to belong to the other class (a 'faster class'). Doing so avoids the dichotomization of the response times in the IRT tree approach above, and the response classification is inferred from the data instead of by a deterministic ad hoc cut-off.

The aim of the present paper is to derive, test, and illustrate the response mixture modeling approach. We use the response mixture approach mainly in an 'indirect application' (e.g., Yung, 1997). That is, similarly to the indirect application of traditional mixture models, the two classes are not necessarily substantively meaningful. They are rather used as a statistical tool to detect the heterogeneity in the responses with respect to the response times. That is, there may be more classes in the data, or the difference in the item characteristics across response times might be continuous (see Bolsinova, Tijmstra, & De Boeck, in press). However, the two classes imposed in the response mixture model are used to statistically capture the most significant part of the heterogeneity in the data. Such an application might be valuable to test the assumption of homogeneity as imposed in the modeling approach discussed above or to detect aberrant responses (e.g., Van der Linden, & Guo, 2008).

The outline of this paper is as follows: First, a general item response theory approach to response time analysis is outlined which characterizes many of the current approaches. Next, the response mixture model for responses and response times is derived using this general model. Then, the viability of the response mixture model is studied in a simulation study. Next, the model is applied to a real dataset pertaining to arithmetic ability to test for heterogeneity in the item characteristic across response times. We end with a discussion of the possibilities and limitations of the present approach.

2. A Response Mixture Model for Responses and Response Times

Many of the existing item response theory models for responses and response times can be characterized by a separate measurement model for the responses and a separate measurement model for the response times which are connected in a certain way. For instance, Thissen (1983) used a two-parameter item response theory model for the responses and a linear factor model for

the log-transformed response times, and connected the models by allowing linear cross-loadings from the response times on the latent ability variable. Ferrando & Lorenzo-Seva (2007a; 2007b) proposed a comparable model but with nonlinear cross-loadings. In addition, Van der Linden (2007) used a three-parameter model and a linear factor model, but connected both measurement models by correlating the random subject and random item effects across the models. Also, mathematically more complex process models like the diffusion-based item response theory model (Tuerlinckx & De Boeck, 2005; Van der Maas et al., 2011) can be seen as consisting of a separate model for the responses (which is, under some assumptions, equivalent to a two-parameter item response theory model) and a separate model for the response times.

To formulate the response mixture model, we also use the idea of separate measurement models for the responses and response times. For the response times, we follow Van der Linden (2007) and use a normal linear homoscedastic factor model, that is,

$$\ln\left(T_{pi}\right) = \lambda_i - \tau_p + \varepsilon_{pi} \tag{1}$$

Because the raw response times T_{pi} for subject $p = 1, \ldots, N$, on item $i = 1, \ldots, n$, are bounded by zero and skewed (Luce, 1986), a log-transformation is used to make the assumption of linearity and homoscedasticity in Eq. 1 more plausible. Other possible transformations include the square root transformation and the reciprocal transformation (see Rummel, 1970 for more options). In Eq. 1 λ_i is the main effect of the item on the log-response time distribution (time intensity parameter), τ_p is the main effect of the subject on the log-response time distribution (subject speed parameter) that accounts for some subjects being on average faster than others. Finally, ε_{pi} is the error with $\text{VAR}(\varepsilon_{pi}) = \sigma_{\varepsilon i}^2$.

For the responses, we adopt the idea by Partchev and De Boeck (2012) who specified a separate measurement model for the faster responses and a separate measurement model for the slower responses. Partchev and De Boeck separated faster responses from slower responses by a median split of the item response times. Next, the faster and slower measurement models were fit as a multidimensional item response theory model to the responses and the categorized response time data. In the present paper, we avoid the necessity of dichotomization of the response time data by formulating a general measurement model for the responses which consists of a mixture of two measurement models, that is

$$\begin{aligned} P\left(X_{pi} = 1 | \theta_p, \alpha_{0i}, \alpha_{1i}, \beta_{0i}, \beta_{1i}\right) &= \pi_{pi} P\left(X_{pi} = 1 | \theta_p, \alpha_{0i}, \beta_{0i}\right) \\ &\quad + \left(1 - _{pi}\right) P\left(X_{pi} = 1 | \theta_p, \alpha_{1i}, \beta_{1i}\right). \end{aligned} \tag{2}$$

where π_{pi} is the mixing proportion which denotes the probability that the response by subject p to item i follows the measurement model in class 0, $\pi_{pi} = P(C_{pi} = 0)$ where $C_{pi} = 0, 1$, is the latent class variable underlying item i. The measurement models for classes $C_{pi} = 0$ and $C_{pi} = 1$ in Eq. 2 are given by, respectively,

$$logit\left[P\left(X_{pi} = 1 | \theta_p, \alpha_{0i}, \beta_{0i}\right)\right] = \alpha_{0i}\theta_p - \beta_{0i} \tag{3}$$

and

$$logit\left[P\left(X_{pi} = 1 | \theta_p, \alpha_{1i}, \beta_{1i}\right)\right] = \alpha_{1i}\theta_p - \beta_{1i} \tag{4}$$

where α_{0i} and β_{0i} are, respectively, the discrimination and difficulty parameters for $C_{pi} = 0$, and α_{1i} and β_{1i} are, respectively, the discrimination and difficulty parameters for $C_{pi} = 1$.

To identify the two classes in terms of the response times, class membership is regressed on the subject and item-corrected log-response times. That is,

$$logit\left[P(C_{pi} = 0|T_{pi}, \lambda_i, \tau_p, \sigma_{\varepsilon i}, \zeta_1, \zeta_0)\right] = \zeta_1 \left(\frac{\ln(T_{pi}) - (\lambda_i - {}_p)}{\sigma_{\varepsilon i}} - \zeta_0 \right). \qquad (5)$$

where λ_i, τ_p and $\sigma_{\varepsilon i}$ are the free parameters that also appear in Eq. 1. We explicitly take the main effects of the subjects (i.e., τ_p) and the main effects of the items (i.e., λ_i) on the log-response times into account in Eq. 5. Neglecting these main effects would confound the within-subject effects by the between-subject differences in basic speed (which are now separated by parameter τ_p) and the between-item differences in time intensity (which are now separated by parameter λ_i). Note that the detection of speediness (Van Der Linden, Breithaupt, Chuah, & Zhang, 2007) and aberrant responses (Van der Linden & van Krimpen-Stoop, 2003) has also been based on a similar procedure.

In Eq. 5, parameter $\zeta_1 \in [0, \infty)$ is constrained to be nonnegative to avoid label switching (i.e., switching of the statistical properties of the two classes during parameter estimation). As a result, class 1 with parameters α_{1i} and β_{1i} correspond to the item characteristics of the faster responses. Parameter ζ_0 is an intercept parameter for which smaller values denote a larger class size of the slower class 0. In the present parameterization, ζ_0 can be interpreted as the 'difficulty' to respond according to the slower class. Note that the item median split procedure used by Partchev and De Boeck to operationalize faster and slower responses does not estimate the class size from the data but assumes the slower class to contain 50% of the responses. The present approach is a statistical variant of this deterministic procedure which additionally retains the continuous information in the response times to operationalize faster and slower responses.

In our approach above, the response times enter the response mixture model in Eq. 2 as a covariate (i.e., via Eq. 5). That is, the response times are a predictor for the latent class variable, C_{pi}. Using predictor variables to identify latent classes is common practice in mixture modeling (e.g., Lubke & Muthen, 2005). Note that this does not imply a mixture distribution for the covariate. That is, the present response mixture model is a mixture model for the responses but not for the response times. The present approach therefore differs from the mixture approaches by Molenaar, Oberski, Vermunt, and De Boeck (2016), Schnipke and Scrams (1997), Wang and Xu (2015), and Wang et al., (2016) in which it is assumed that the latent class variable affects both the responses and the response time distribution.

Using the mixing proportion π_{pi}, inferences can be made about the marginal class size for class 0, $C_{pi} = 0$, for the items ($\tilde{\pi}_i = \frac{1}{N} \sum_{p=1}^{N} \pi_{pi}$), for the subjects ($\tilde{\pi}_p = \frac{1}{n} \sum_{i=1}^{n} \pi_{pi}$), or overall ($\tilde{\pi} = \frac{1}{N \times n} \sum_{p=1}^{N} \sum_{i=1}^{n} \pi_{pi}$). In the population, π_{pi} is equal across items and subjects as the mixing parameters ζ_0 and ζ_1 are not subject specific or item specific. This assumption of equal class sizes across items in the population is referred to as the assumption of time homogeneity in the longitudinal modeling literature (e.g, Bacci, Pandolfi, & Pennoni, 2014) and can be relaxed in principle. We do not consider such extensions in the present study. We return to this point in the discussion section. Note that $\tilde{\pi}_i$ and $\tilde{\pi}_p$ are implicitly assumed to be equal to 0.50 in, respectively, the item median split and the person median split procedure by Partchev and De Boeck (2012).

2.1. Estimation

Let $\boldsymbol{\eta}_i$ denote the vector of item parameters for item i,
$\boldsymbol{\eta}_i = [\ln(\alpha_{0i}), \ln(\alpha_{1i}), \beta_{0i}, \beta_{1i}, \lambda_i, \ln(\sigma_{\varepsilon i}^2)]$ for $I = 1, \ldots, n$, let $\boldsymbol{\omega}_p$ denote a vector containing the parameters for subject p, $\boldsymbol{\omega}_p = [\theta_p, \tau_p]$ for $p = 1, \ldots, N$, and let $\boldsymbol{\zeta}$ denote a vector with the mixing parameters $\boldsymbol{\zeta} = [\zeta_0, \ln(\zeta_1)]$. Note that we use logarithmic transformations for the parameters that are strictly positive. The likelihood of the data for subject p on item i given the model is then

equal to

$$L\left(X_{pi}, \ln\left(T_{pi}\right); \boldsymbol{\eta}_i, \boldsymbol{\omega}_p, \boldsymbol{\zeta}\right) = f\left(X_{pi}|\ln\left(T_{pi}\right), \alpha_{0i}, \alpha_{1i}, \beta_{0i}, \beta_{1i}, \zeta_0, \zeta_1, \theta_p\right)$$
$$g\left(\ln(T_{pi})|\lambda_i, \sigma_{\varepsilon i}^2, \tau_p\right) \quad (6)$$

where f (.) is a mixture Bernoulli probability function with the success parameter given by Eq. 2 and the mixing proportion given by Eq. 5. In addition, g (.) is a normal density function.

The model can be fit to data by assuming the item parameters $\boldsymbol{\eta}_i$ and $\boldsymbol{\zeta}$ to be fixed parameters and numerically integrating over the random parameters, θ_p and τ_p, in the likelihood function by assuming a prior distribution for these parameters. Next, this marginal likelihood is maximized for the unknown fixed parameters in $\boldsymbol{\eta}_i$ and $\boldsymbol{\zeta}$ (Bock & Aitkin, 1981). Another possibility, which is adopted here, is to assume prior distributions for all parameters in $\boldsymbol{\eta}_i$, $\boldsymbol{\omega}_p$, and $\boldsymbol{\zeta}$ and use a Markov Chain Monte Carlo approach to draw samples from the joint posterior parameter distribution (MCMC; see, e.g., Gilks, Richardson, & Spiegelhalter, 1996).

2.2. Prior Distributions

In choosing our priors, we generally follow Fox, Klein Entink, and Van der Linden (2007) and Van der Linden (2007) who discuss prior and hyperprior specification in a general hierarchical modeling framework for responses and response times. First, we parameterized a bivariate normal distribution for the vector $\boldsymbol{\omega}_p$ using

$$\theta_p \sim \text{Normal}(0, 1) \quad (7)$$
$$\tau_p \sim \text{Normal}\left(\sigma_{\theta\tau}\theta_p; \sigma_\tau^2 - \sigma_{\theta\tau}^2\right) \quad (8)$$

which can be derived by considering the distribution of θ_p conditional on τ_p in a bivariate normal distribution (see, e.g., Fox et al., 2007). Note that the above ensures that $\text{VAR}(\theta_p) = 1$ and that the covariance matrix of ω_p (denoted Σ_P) is positive definite, which identifies the model. No additional identification constraints are needed as the scale of τ_p is identified by the unit of the log-response times. Note that due to the parametrization above, besides $\sigma_{\theta\tau}$, we estimate the hyperparameter $\sigma_\tau'^2 = \sigma_\tau^2 - \sigma_{\theta\tau}^2$ instead of σ_τ^2.

For the item parameters in $\boldsymbol{\eta}_i$, we use

$$\boldsymbol{\eta}_i \sim \text{Multivariate} - \text{Normal}\left(\boldsymbol{\mu}_\eta; \boldsymbol{\Sigma}_\eta\right) \quad (9)$$

in which all hyperparameters in the mean vector $\boldsymbol{\mu}_\eta$ and covariance matrix $\boldsymbol{\Sigma}_\eta$ are estimated (see Van der Linden, 2007). Finally, for ζ_0 and $\ln(\zeta_1)$, we use

$$\zeta_0 \sim \text{Normal}(\mu_{\zeta_0}, \sigma_{\zeta_0}^2) \quad (10)$$
$$\ln(\zeta_1) \sim \text{Normal}(\mu_{\zeta_1}, \sigma_{\zeta_1}^2). \quad (11)$$

2.3. Hyperpriors

For the hyperparameters above, we specified the following hyperpriors

$$\sigma_{\theta\tau} \sim \text{Normal}\left(\mu_\rho, \sigma_\rho^2\right) \quad (12)$$
$$\sigma_\tau'^2 \sim \text{Inverse} - \text{Gamma}(v_1, v_2) \quad (13)$$

In addition, for $\boldsymbol{\mu}_\eta$ we follow Van der Linden et al (2007) and Fox et al. (2007) and specify

$$\boldsymbol{\mu}_\eta \sim \text{Multivariate} - \text{Normal}\,(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_\eta) \tag{14}$$

with

$$\boldsymbol{\Sigma}_\eta \sim \text{Inverse} - \text{Wishart}\,(v_I, \mathbf{V_I})\,. \tag{15}$$

The model can be implemented in the freely available software packages JAGS (Plummer, 2003), STAN (Stan Development Team, 2015), and BUGS (Spiegelhalter, Thomas, Best, & Gilks, 1995). Here we implemented the model in OpenBUGS (Thomas, Hara, Ligges, & Sturtz, 2006). The syntax file is given in "Appendix."

2.4. Model Fit Diagnosis

To assess the goodness-of-fit of the response mixture model, we want to compare the model to a baseline model. First, a reasonable baseline model can be derived by assuming that the responses in the two classes have the same item characteristics, that is $\alpha_i = \alpha_{0i} = \alpha_{1i}$ and $\beta_i = \beta_{0i} = \beta_{1i}$. As a result, in Eq. 2 it will hold that $P(X_{pi} = 1|\theta_p, \alpha_{0i}, \beta_{0i}) = P(X_{pi} = 1|\theta_p, \alpha_{1i}, \beta_{1i})$. Therefore, the model for the responses will reduce to a standard two-parameter item response theory model, that is,

$$logit\left[P\left(X_{pi} = 1|\theta_p, \alpha_i, \beta_i\right)\right] = \alpha_i\theta_p - \beta_i, \tag{16}$$

and consequently ζ_0 and ζ_1 from Eq. 5 cancel out of the likelihood function. The baseline model that results (i.e., Eqs. 1 and 16) is equivalent to the hierarchical model of Van der Linden (2007).

A common model fit index is the Deviance Information Criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linde, 2002). Although valuable in a variety of model selection situations in practice, it has been noted that the DIC may perform suboptimally in mixture models (e.g., Celeux, Forbes, Robert, & Titterington, 2006). The main problem is that the estimate of the number of parameters, which is needed to evaluate the DIC, may be wrong in mixture models or highly nonlinear models. In preliminary simulations (not presented), we established that the DIC and an alternative fit index, the Watanabe-Aike Information Criterion (WAIC, Watanabe, 2010), are both indeed unsuitable for the present model fit comparison.

As common model fit indices can thus not be used, we focus on the key objective for which we apply the response mixture model: accounting for heterogeneity in the item characteristic across the response times. Thus, to decide if the response mixture model is of use in a given data application, the central question is whether there is heterogeneity in the item characteristic across response times. Therefore, as a diagnostic tool to investigate parameter heterogeneity, we consider the posterior parameter distributions of $\Delta\,\alpha_i = \alpha_{0i} - \alpha_{1i}$ and $\Delta\,\beta_i = \beta_{0i} - \beta_{1i}$ and we consider the posterior probabilities that $\alpha_{0i} > \alpha_{1i}$, and $\beta_{0i} > \beta_{1i}$. If the data do not contain any heterogeneity in the item characteristics with respect to the response times, the baseline model will hold for the data and the posterior parameter distributions of $\Delta\,\alpha_i$, and $\Delta\,\beta_i$ will all have their posterior probability mass concentrated around 0. In addition, the posterior probabilities of $\alpha_{0i} > \alpha_{1i}$ and $\beta_{0i} > \beta_{1i}$ will tend to 0.5. As a consequence, the mixing parameters ζ_0 and ζ_1 will be poorly identified reflected by estimates near the parameter boundary and a large variance of the sampling distribution. Such signals of poor identification are characteristic for mixture models in general (i.e., also for more basic mixture models), that is, if the data truly consist of a single component, any two component mixture model will be poorly identified in these data. However, this problem is easily diagnosed from the sampling distributions of the mixture model parameters as we will show below.

3. Simulation Study

3.1. Data

We conducted a simulation study to investigate the parameter recovery of the response mixture model in the case that the data truly follow a response mixture model and in the case that the data actually follow a baseline model without mixtures. To this end, we simulated data according to the response mixture model (Eqs. 1, 2, and 5) and according to the baseline model (Eqs. 1 and 16). In the case of the response mixture model, we drew the vector of true item parameters for item i, $\boldsymbol{\eta}_i = [\ln(\alpha_{0i}), \ln(\alpha_{1i}), \beta_{0i}, \beta_{1i}, \lambda_i, \ln(\sigma^2_{\varepsilon i})]$, from a multivariate normal distribution with mean vector $[\ln(1), \ln(1.5), -1, 1, 2, \ln(.3)]$ and a covariance matrix with diagonal elements equal to 0.2, 0.2, 0.7, 0.7, 0.02, and 0.02, and with $COR(\alpha_{0i}, \alpha_{1i}) = 0.8$, $COR(\beta_{0i}, \beta_{1i}) = 0.8$, $COR(\beta_{0i}, \lambda_i) = 0.4$, and $COR(\beta_{0i}, \lambda_i) = 0.4$. All other parameter correlations were equal to 0. We used the same true item parameters in all replications to study the recovery of the item parameters. In each replication, we drew the vector of person parameters for person p, $\boldsymbol{\omega}_p = [\theta_p, \tau_p]$, from a bivariate normal distribution with mean vector $[0,0]$ and a covariance matrix with $\sigma^2_\theta = 1$, $\sigma^2_\tau = 0.16$, and $\sigma_{\theta\tau} = 0.16$ which implies $COR(\theta_p, \tau_p) = 0.4$. The mixing parameters were chosen to equal $\zeta_1 = 1$ and $\zeta_0 = 0.5$ which imply a marginal probability of a slow response of 0.40. The parameter values above result in untransformed response times between roughly 1 and 60 s with the mean around 9 s. In the case of the baseline model, we used $\alpha_i = \alpha_{0i}$ and $\beta_i = \beta_{0i}$, all other values (i.e., the values for λ_i, $\sigma^2_{\varepsilon i}$, σ^2_τ, and $\sigma_{\theta\tau}$) were equal to those above. We simulated data for $N = 2000$ subjects and $n = 20$ items. We conducted 80 replications.

3.2. Estimation

In our OpenBUGS implementation of the model, the hyper parameters are set as follows: We set $\mu_{\zeta 0}$, $\mu_{\zeta 1}$, and μ_ρ to be equal to 0, and $\sigma^2_{\zeta 0}$, $\sigma^2_{\zeta 1}$, and σ^2_ρ were all set equal to 10. In addition, v_I was set to 6 and $\mathbf{V}_I$ was equal to a diagonal matrix with elements equal to 0.01. Finally, v_1 and v_2 were set to 0.01, and $\boldsymbol{\mu}_\eta$ was equal to $[0,0,0,0,0]$. Note that these choices reflect vague information about $\boldsymbol{\zeta}$, $\sigma_{\theta\tau}$, σ'^2_τ, $\boldsymbol{\omega}_p$, and $\boldsymbol{\eta}_i$. We drew 8,000 samples from the posterior parameter distribution of which we discarded the first 4,000 samples as burn-in. Preliminary data simulations showed that this is sufficient to ensure convergence of the chain to its stationary distribution.

3.3. Results

3.3.1. Data Follow the Response Mixture Model   Plots of the posterior parameter means with the range between their 1th and 99th percentile across the replications are depicted for α_{0i}, α_{1i}, β_{0i}, β_{1i}, λ_i, and $\sigma^2_{\varepsilon i}$ in Fig. 1. As can be seen, the parameter recovery is generally acceptable as the posterior means fluctuate around their true value suggesting that they are not biased. For β_{1i} and α_{1i}, parameter variability is somewhat larger as compared to β_{0i} and α_{0i} due to the faster class being proportionally larger in the data (due to the true value of ζ_0 being 0.5).

The other parameters are recovered acceptably as well: For the mixing parameters, ζ_0 we found an average posterior mean across replications equal to 0.49 (1th and 99th percentile: 0.24; 0.82) where the true value equaled 0.5. For the mixing parameter ζ_1, we found an average posterior mean across replications equal to 1.03 (1th and 99th percentile: 0.86, 1.31) where the true value equaled 1.0. The average posterior mean across replications for $\sigma_{\theta\tau}$ equaled 0.16 (1th and 99th percentile: 0.14, 0.18) and 0.14 (1th and 99th percentile: 0.12, 0.14) for σ'^2_τ where the true values equaled, respectively, 0.16 and 0.14.

3.3.2. Data Follow the Baseline Model   Here we study whether we can diagnose model misfit of the full response mixture model if the data follow the baseline model without classes. In Fig. 2,
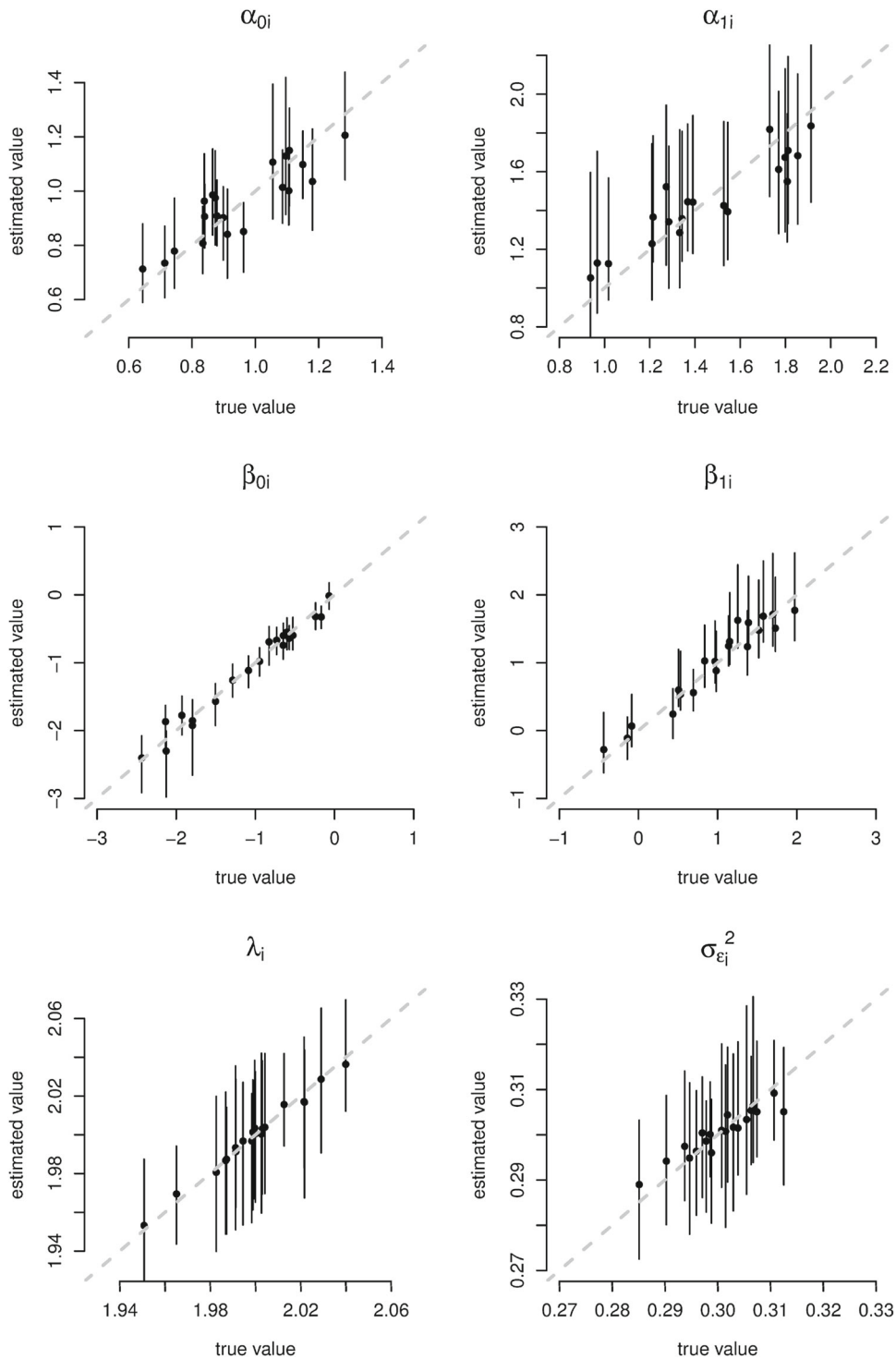
FIGURE 1.
Plots of the average posterior means across replications for α_{0i}, α_{1i}, β_{0i}, β_{1i}, λ_i, and $\sigma_{\varepsilon i}^2$ (y-axis) and the true parameters values (x-axis). The vertical lines denote the range between the 1th and 99th percentile of the posterior parameter means across replications. The striped gray line denotes a one-to-one correspondence.
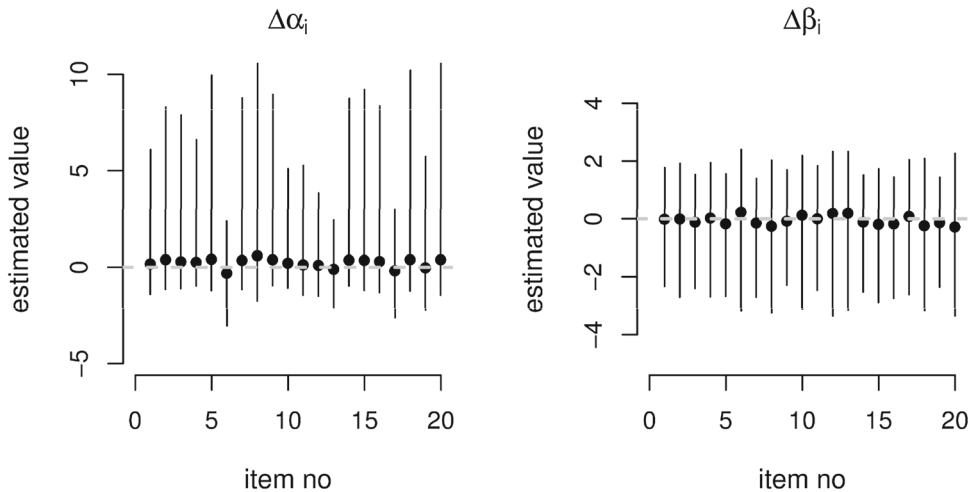
FIGURE 2.
Plots of the posterior means of $\Delta \alpha_i = \alpha_{0i} - \alpha_{1i}$ and $\Delta \beta_i = \beta_{0i} - \beta_{1i}$ across the replications in which the data follow the baseline model. The vertical lines denote the range between the 1th and 99th percentile of the posterior parameter means across replications.

plots of the posterior parameter means with the range between their 1th and 99th percentile across the replications are depicted for $\Delta \alpha_i$ and $\Delta \beta_i$. As can be seen from the figure, the posterior parameter means fluctuate around 0 indicating that $\alpha_{0i} = \alpha_{1i}$ and $\beta_{0i} = \beta_{1i}$. In addition, the posterior probabilities of $\alpha_{0i} > \alpha_{1i}$ and $\beta_{0i} > \beta_{1i}$ averaged over replications and items are, respectively, equal to 0.53 and 0.50 indicating that these inequalities do not hold and that the discrimination and difficulty parameters are equal over classes. In addition, the overall class probability $\tilde{\pi}$ is close to 0 or 1 in, respectively, 22 (27.5%) and 46 (57.5%) of the replications. For 11 replications (13.75%), $\tilde{\pi}$ was estimated to be close to 0.5 due to ζ_1 approaching 0. Only in 1 replication (1.25%), there was no indication of model misfit in $\tilde{\pi}$ as the estimate equaled 0.64 in this replication.[1] Thus, in conclusions, it seems that the model diagnostics above can indeed be used to study model misfit of the full response mixture model.

3.4. Application

3.4.1. Data   The data comprise a computerized arithmetic test which was a part of the secondary school exams in the Netherlands in 2013. These data have previously been analyzed by Bolsinova and Maris (2016) who tested the responses and response times on the assumption of conditional independence. The total dataset contains the responses and response times of 10,369 subjects to 60 items. Here, we randomly selected 2000 subjects and 20 items to match the setting in the simulation study above. See Table 1 for the item discrimination, item difficulty, time intensity, and error variance of the responses and response times in the baseline model (Eqs. 1 and 16, based on 8000 iterations with 4000 as burn-in).

3.5. Response Mixture Model

To investigate whether item characteristics are heterogeneous across response times, we resorted to response mixture modeling. All prior specifications are the same as in the simulation

---

[1] $\tilde{\pi}$ is calculated using the formula given above, $\tilde{\pi} = \frac{1}{N \times n} \sum_{p=1}^{N} \sum_{i=1}^{n} \pi_{pi}$, with π_{pi} calculated using Eq. 5 in which item parameters $\zeta_0, \zeta_1, \lambda_i$, and σ_i are substituted by their posterior means, and nuisance parameters τ_p are numerically integrated out using a normal distribution for τ_p.

TABLE 1.
Posterior Means with 99% HPD regions of the response and response time parameters in the baseline model.

| i	α_i			β_i			λ_i			σ_ε^2		
	Mean	99% HPD		Mean	99% HPD		Mean	99% HPD		Mean	99% HPD	
		Lower	Upper		Lower	Upper		Lower	Upper		Lower	Upper
1	0.85	0.68	1.04	−1.73	−1.93	−1.56	4.76	4.74	4.79	0.10	0.09	0.11
2	0.55	0.42	0.69	−0.46	−0.59	−0.34	5.03	5.00	5.05	0.12	0.11	0.13
3	1.04	0.84	1.25	−1.60	−1.80	−1.42	4.37	4.35	4.40	0.18	0.16	0.19
4	0.87	0.71	1.07	−1.42	−1.59	−1.25	3.15	3.11	3.18	0.29	0.27	0.32
5	0.51	0.36	0.67	−1.68	−1.87	−1.53	4.45	4.43	4.47	0.09	0.09	0.10
6	0.98	0.77	1.18	−2.05	−2.26	−1.84	3.72	3.69	3.75	0.30	0.28	0.33
7	0.56	0.40	0.70	1.04	0.90	1.18	5.14	5.12	5.16	0.07	0.07	0.08
8	0.79	0.63	0.98	1.29	1.13	1.46	5.18	5.16	5.20	0.12	0.11	0.13
9	1.16	0.97	1.36	0.27	0.12	0.42	5.08	5.05	5.11	0.26	0.23	0.28
10	1.13	0.94	1.33	−0.61	−0.77	−0.46	4.71	4.68	4.74	0.23	0.21	0.25
11	1.06	0.86	1.30	−1.40	−1.59	−1.24	4.72	4.69	4.74	0.14	0.13	0.15
12	1.00	0.83	1.19	−0.76	−0.92	−0.62	4.54	4.52	4.57	0.17	0.15	0.18
13	0.75	0.59	0.92	−1.49	−1.68	−1.34	4.67	4.65	4.69	0.10	0.09	0.11
14	1.13	0.92	1.34	1.05	0.89	1.21	5.17	5.14	5.20	0.23	0.22	0.25
15	1.54	1.29	1.82	0.86	0.68	1.05	4.96	4.93	5.00	0.30	0.27	0.32
16	1.05	0.89	1.24	−0.57	−0.71	−0.44	5.04	5.01	5.07	0.24	0.22	0.25
17	0.96	0.79	1.13	−0.41	−0.56	−0.27	5.01	4.98	5.03	0.17	0.15	0.18
18	0.96	0.78	1.14	−1.07	−1.23	−0.91	4.22	4.19	4.26	0.32	0.29	0.35
19	1.56	1.30	1.82	1.00	0.82	1.19	5.02	4.98	5.07	0.55	0.51	0.60
20	0.52	0.38	0.68	−1.28	−1.43	−1.14	4.32	4.29	4.34	0.13	0.12	0.15
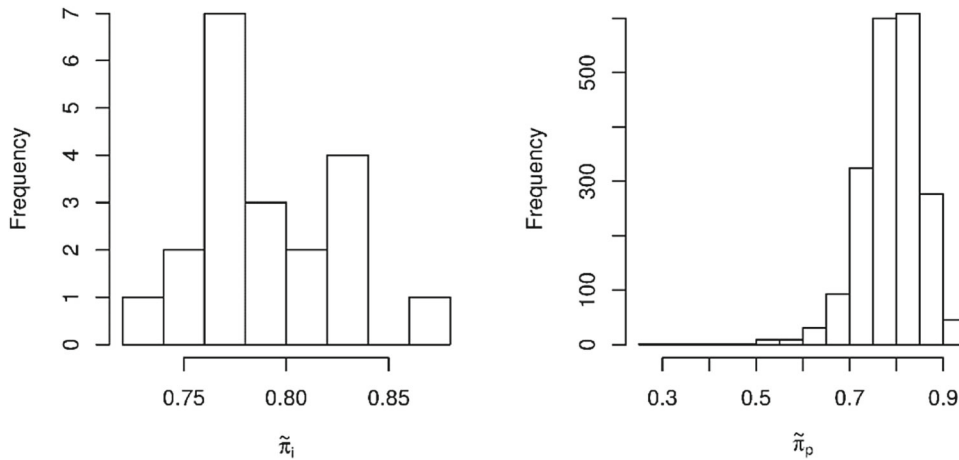
FIGURE 3.
Histograms of the item ($\tilde{\pi}_i$) and subject ($\tilde{\pi}_p$) specific marginal class probabilities.

study. We run 20,000 iterations with 10,000 as burn-in. The Gelman and Rubin statistic indicated (based on 5 chains with randomly drawn starting values) that this number of iterations was sufficient to ensure that the MCMC sampling scheme converged to its stationary distribution (Gelman & Rubin, 1992). In addition, from the trace plots all chains seemed to vary randomly around a stable average.

3.5.1. Results   The posterior mean of ζ_0 equals -0.70 with 99% HPD bounds of -0.79 and -0.62. The posterior mean of ζ_1 equals 5.785 with 99% HPD bounds of 4.25 and 7.73. Using these parameters, the posterior class probabilities, $\tilde{\pi}$, $\tilde{\pi}_p$ and $\tilde{\pi}_i$ are obtained. The overall class probability (the overall size of the slow class, $C_{pi} = 0$) equals 0.79. The item- and person-specific class probabilities are depicted in Fig. 3. Table 2 contains the posterior parameter means for the item parameters in the two classes according to the full model. The posterior means of the estimates of α_{0i}, α_{1i}, β_{0i}, and β_{1i} are plotted in Fig. 4 for the items as ordered according to their general difficulty (i.e., the difficulty as estimated using the baseline model, see Table 1). As can be seen from the figure, the slower responses are associated with uniformly smaller discrimination parameters. In addition, there appears to be an interaction effect between the overall difficulty of the items and the difference between the fast and slow difficulty. That is, for the overall more difficult items, the slower responses have a smaller difficulty (larger probability of a correct response), while for the overall less difficult items, the faster responses have a smaller difficulty (larger probability of a correct response).

To investigate how well the above observations concerning the differences between the fast and slow response characteristics are supported by the data, we depicted the posterior parameter distributions of $\Delta \alpha_i = \alpha_{0i} - \alpha_{1i}$ and $\Delta \beta_i = \beta_{0i} - \beta_{1i}$ in Fig. 5 for the items as ordered according to their general difficulty. As can be seen, for most items, the conclusion that the slower responses (class 0) are associated with smaller discrimination parameters is supported by the 99% HPD regions of $\Delta \alpha_i$. In addition, for eleven items the difficulty parameters differ between the classes according to their 99% HPD regions. As already observed above, the difference is not in the same direction for all these items. That is, for the overall easier items, the responses in the faster class are associated with smaller item difficulty parameters as compared to the difficulty parameters of the responses in the slower class ($\Delta \beta_i > 0$), while for the overall more difficult items, this difference is in the opposite direction ($\Delta \beta_i < 0$). This provides some evidence for the interaction

TABLE 2.
Posterior Means and 99% HPD regions of the response parameters in the response mixture model.

i	α_{0i} Mean	99% HPD Lower	Upper	α_{1i} Mean	99% HPD Lower	Upper	β_{0i} Mean	99% HPD Lower	Upper	β_{1i} Mean	99% HPD Lower	Upper
1	0.73	0.54	0.96	1.42	0.97	1.93	−1.74	−1.96	−1.53	−1.78	−2.32	−1.29
2	0.52	0.37	0.67	1.09	0.76	1.44	−0.71	−0.89	−0.55	0.42	0.04	0.82
3	0.88	0.68	1.10	2.10	1.48	3.05	−1.30	−1.51	−1.11	−3.39	−4.46	−2.56
4	0.79	0.61	0.98	1.79	1.19	2.57	−1.08	−1.26	−0.90	−3.57	−4.60	−2.71
5	0.42	0.28	0.57	1.06	0.65	1.49	−1.58	−1.77	−1.39	−2.19	−2.82	−1.68
6	0.87	0.65	1.10	1.41	0.98	1.90	−2.02	−2.26	−1.77	−2.14	−2.71	−1.63
7	0.49	0.35	0.64	0.91	0.48	1.45	0.72	0.57	0.88	3.64	2.53	5.00
8	0.77	0.58	0.97	1.16	0.67	1.76	0.99	0.81	1.17	4.35	2.95	6.55
9	1.10	0.88	1.35	1.35	0.95	1.85	−0.18	−0.34	0.00	3.12	2.44	3.92
10	0.92	0.74	1.11	1.83	1.40	2.38	−0.62	−0.80	−0.46	−0.70	−1.18	−0.27
11	0.90	0.69	1.11	1.96	1.43	2.74	−1.23	−1.42	−1.03	−2.33	−3.09	−1.66
12	0.91	0.72	1.13	1.49	1.15	1.88	−0.93	−1.11	−0.76	−0.08	−0.48	0.35
13	0.63	0.47	0.80	1.37	1.02	1.85	−1.46	−1.65	−1.28	−1.68	−2.26	−1.14
14	1.03	0.82	1.24	1.62	1.05	2.37	0.86	0.68	1.03	3.29	1.99	4.88
15	1.37	1.11	1.63	2.12	1.35	3.13	0.67	0.49	0.87	2.27	1.46	3.19
16	0.99	0.77	1.22	1.10	0.60	1.67	−1.12	−1.32	−0.93	3.03	2.10	4.44
17	0.85	0.68	1.04	1.75	1.33	2.31	−0.46	−0.61	−0.31	−0.13	−0.63	0.35
18	0.82	0.63	1.02	1.47	1.07	1.94	−1.16	−1.33	−0.98	−0.59	−1.05	−0.15
19	1.36	1.11	1.64	1.93	1.15	2.88	0.69	0.50	0.89	2.97	2.17	3.86
20	0.43	0.28	0.59	1.02	0.66	1.41	−1.31	−1.48	−1.13	−1.26	−1.71	−0.87
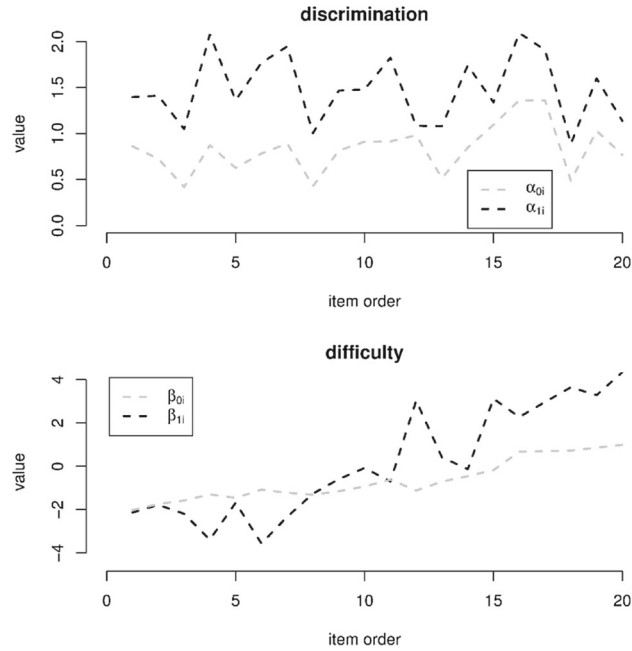
FIGURE 4.

Plots of the posterior means of the α_{0i}, α_{1i}, β_{0i}, and β_{1i} parameters (y-axis) across item as ordered on their difficulty according to the baseline model (x-axis).
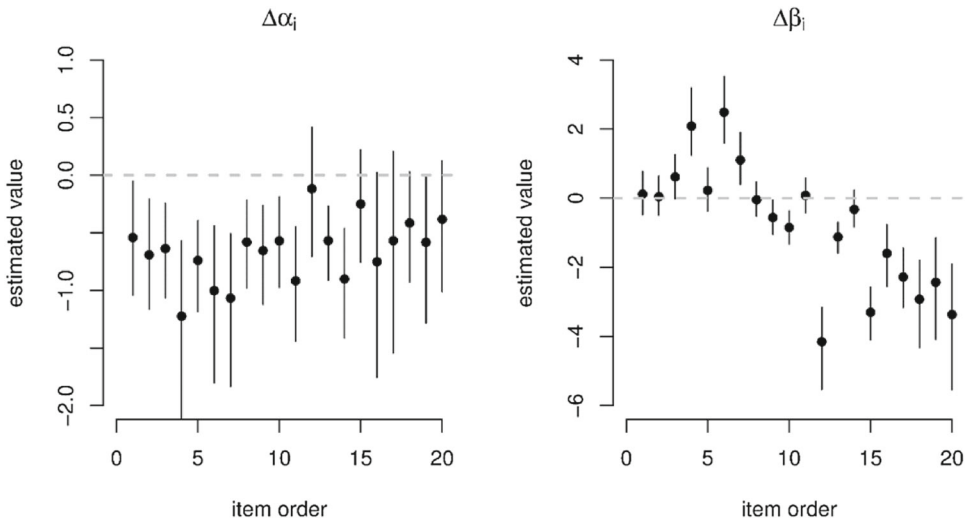

FIGURE 5.

Plots of the posterior means of $\Delta\alpha_i = \alpha_{0i} - \alpha_{1i}$ and $\Delta\beta_i = \beta_{0i} - \beta_{1i}$ for each item in the application. The vertical lines denote the range of the 99% HPD regions.

between the overall difficulty of the items and the difference between the fast and slow difficulty as already noticed in Fig. 4. Furthermore, we depicted the posterior probabilities of $\alpha_{0i} > \alpha_{1i}$ and $\beta_{0i} > \beta_{1i}$ in Table 3. As can be seen, the conclusion are the same as those from Fig. 5.

TABLE 3.
Posterior probabilities of $\alpha_{0i} > \alpha_{1i}$ and $\beta_{0i} > \beta_{1i}$.

i	$\alpha_{0i} > \alpha_{1i}$	$\beta_{0i} > \beta_{1i}$
1	0.00	0.56
2	0.00	0.00
3	0.00	1.00
4	0.00	1.00
5	0.00	1.00
6	0.00	0.68
7	0.01	0.00
8	0.03	0.00
9	0.08	0.00
10	0.00	0.65
11	0.00	1.00
12	0.00	0.00
13	0.00	0.83
14	0.00	0.00
15	0.00	0.00
16	0.30	0.00
17	0.00	0.06
18	0.00	0.00
19	0.03	0.00
20	0.00	0.39

3.6. Discussion

In the real data application, it is concluded that the faster responses are associated with larger discrimination parameters as compared to the slower responses. This finding is not line with the 'worst performance rule' from experimental psychology (Larson, & Alderton, 1990) which predicts that slower responses contain more information about individual differences in ability than faster responses. This contrasting result might suggest that the worst performance rule, which is typically established using experimental tasks with extremely fast responses (500–1000 ms), does not hold for cognitive ability tasks with responses between 5–120 s.

A second finding in the application is that there is an interaction between the overall item difficulty and the probability of a correct response in the faster and slower classes. Specifically, we found that the faster responses are more successful for the easier items while the slower responses are being more successful for the more difficult items. This finding is in line with Bolsinova et al. (in press), De Boeck, Chen, & Davison (2017), Goldhammer et al. (2014), and Partchev and De Boeck (2012) who found a similar effect. These findings suggest that for easier items, successful responses are given by more or less automated processes while if these automated processes fail, other controlled processes are being used which are more error prone. On the contrary, for the more difficult items, it holds that the responses for which subjects take more time are generally more successful than the faster responses.

4. General Discussion

In this paper, we outlined the method of response mixture modeling to account for heterogeneity in the item characteristics across response times. As compared to the hierarchical model of Van der Linden (2007), the model put forward in this paper contains two additional parameters per item (an additional discrimination and an additional difficulty) and two additional mixing

parameters. The model is thus reasonably complex reflected by relatively large posterior standard deviations of the parameter estimates for the class-specific discrimination parameters (see, e.g., Fig. 1 from the simulation study). We therefore did not consider more general cases of the present model. However, various extensions can be thought of. That is, two separate latent variables can be assumed to underlie the responses in the faster and slower classes. In doing so, the correlation between the ability in the faster and slower classes can be investigated, as well as the correlation of the faster and slower ability with the latent speed variable. Another extension might be to relax the assumption of time homogeneity (Bacci, Pandolfi, & Pennoni, 2014) by making the mixing parameters item specific. From our perspective, this is feasible, but will require large datasets.

A key assumption in our model is that there are two latent classes underlying the responses to a given test. As discussed above, we focused mainly on the indirect application (Yung, 1997) in which we do not necessarily interpret these classes substantively. That is, we treat the classes as statistical tools to capture the heterogeneity of the item characteristics across the response times. If our model is applied to data that contain more than two classes, the present approach will likely still capture the most important patterns in the data. That is, the model will capture the possible systematic differences in the discrimination and difficulty of the items for larger and smaller response times. However, if some subjects in the data do not display heterogeneity in their response characteristics, e.g., because they only use a single response process, the model does not hold for that subject (person misfit). As a result, the model will produce invalid posterior class assignments for this subject. Such person misfit can be diagnosed by consulting the posterior means of the class probabilities. If only a single class holds for a given subject, the class probabilities will tend to 0.5 indicating chance assignment.

In this paper, we did not so much focus on direct applications of the mixture model in which the two classes are indeed substantively interpreted (e.g., Titterington, Smith, & Makov, 1985; Dolan & Van der Maas, 1998). However, in the presence of a strong theory or strong empirical justification, such applications of the response mixture model are certainly possible. For instance, Jansen and Van der Maas (2001) studied two strategies to solve the balance scale task. These strategies can be derived from Piaget's theory on conservation (see Siegler, 1976). Another example of a theoretical framework where two classes are assumed to underlie response behavior is the dual processing framework (Shiffrin & Schneider, 1977; Goldhammer et al., 2014). In this framework, faster responses are assumed to reflect automated processes that are proceduralized, parallel, and do not require active control, while slower responses are assumed to reflect controlled processes that are serial and require attentional control. In addition, in cognitive psychology it has been shown that decision making may involve a slower selective search strategy or a faster pattern recognition strategy in medical decisions (Ericsson & Staszewski 1989) and in solving chess puzzles (Van Harreveld, Wagenmakers, and Van der Maas, 2007). Finally, as discussed above, in arithmetic tests it can be expected that subjects use both memory retrieval and actual calculation to solve the arithmetic problems of a test (Gabner et al., 2009).

Appendix: OpenBUS Syntax to Fit the Full Response Mixture Model

```
model{

for(p in 1:N){

for(i in 1:nit){

        logit(p.slow[p,i]) <- a.slow[i]*theta[p] - b.slow[i]          #Eq. 3

        logit(p.fast[p,i]) <- a.fast[i]*theta[p] - b.fast[i]          #Eq. 4


        mu.rt[p,i]<- icept[i] - speed[p]                             #Expected value of Eq. 1

        lrt[p,i]~dnorm(mu.rt[p,i],prec[i])                           #Follows from Eq. 1

        res[p,i]<-lrt[p,i]- mu.rt[p,i]

        logit(pi[p,i])<- time*(res[p,i] * sqrt(prec[i])-marg)        #Eq. 5 with prec[i] =σ_εi^{-2}

        prob[p,i] <- pi[p,i]*p.slow[p,i]+(1-pi[p,i])*p.fast[p,i]     #Eq. 2

        x[p,i]~dbern(prob[p,i])

}}

for(p in 1:N){

        theta[p]~dnorm(0,1)                                          # prior on θ

        mu.s[p]<-cor*theta[p]

        speed[p]~dnorm(mu.s[p],prec.s)                              # prior on τ with prec.s=σ_τ'^{-2}

}

for(i in 1: nit){                                                   # prior on η_i

        item.par[i,1:6]~dmnorm(item.mu[],item.prec[,])

        a.fast[i]<-exp(item.par[i,1])                               # α_{0i}

        a.slow[i]<-exp(item.par[i,2])                               # α_{1i}

        b.fast[i]<-item.par[i,3]                                    # β_{0i}

        b.slow[i]<-item.par[i,4]                                    # β_{1i}

        icept[i]<-item.par[i,5]                                     # λ_i

        prec[i]<-1/exp(item.par[i,6])                              # σ_εi^{-2}

}

        lntime~dnorm(0,.1)                                         #prior on ln(ζ_1)

        time<-exp(lntime)                                          #ζ_1

        marg~dnorm(0,.1)                                           #prior on ζ_0


item.mu[1:6]~dmnorm(zeros[],item.prec[,])                          # hyperprior on μ_η

item.prec[1:6,1:6]~dwish(R[,],6)                                   # hyperprior on Σ_η^{-2}


sigma2.s.tilde<-1/prec.s

sigma2.s<-sigma2.s.tilde+cor*cor                                   # σ_τ^2 = σ_τ'^2 + σ_θτ^2


cor~dnorm(0,.1)                                                    # hyperprior on σ_θτ

prec.s~dgamma(.1,.1)                                               # hyperprior on σ_τ'^{-2}

}
```

References

Bacci, S., Pandolfi, S., & Pennoni, F. (2014). A comparison of some criteria for states selection in the latent Markov model for longitudinal data. *Advances in Data Analysis and Classification*, *8*(2), 125–145.

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: An unified approach* (Vol. 904). Hoboken: Wiley.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In E. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores (chap)* (pp. 17–20). Reading, MA: Addison Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443–459.

Bolsinova, M., & Maris, G. (2016). A test for conditional independence between response time and accuracy. *British Journal of Mathematical and Statistical Psychology*, *69*(1), 62–79.

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., de Boeck, P., & Tijmstra, J. (in press). Modelling conditional dependence between response time and accuracy. *Psychometrika*.

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

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review*, *97*, 404–431.

Celeux, G., Forbes, F., Robert, C. P., & Titterington, D. M. (2006). Deviance information criteria for missing data models. *Bayesian Analysis*, *1*(4), 651–673.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, *7*, 249–253.

De Boeck, P., Chen, H., & Davison, M. (2017). Spontaneous and imposed speed of cognitive test responses. *British Journal of Mathematical and Statistical Psychology*, *70*(2), 225–237.

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

DiTrapani, J., Jeon, M., De Boeck, P., & Partchev, I. (2016). Attempting to differentiate fast and slow intelligence: Using generalized item response trees to examine the role of speed on intelligence tests. *Intelligence*, *56*, 82–92.

Dolan, C. V., Colom, R., Abad, F. J., Wicherts, J. M., Hessen, D. J., & van de Sluis, S. (2006). Multi- group covariance and mean structure modeling of the relationship between the WAIS-III common factors and sex and educational attainment in Spain. *Intelligence*, *34*, 193–210.

Dolan, C. V., & van der Maas, H. L. (1998). Fitting multivariage normal finite mixtures subject to structural equation modeling. *Psychometrika*, *63*(3), 227–253.

Ericsson, K. A., & Staszewski, J. J. (1989). Skilled memory and expertise: Mechanisms of exceptional performance. In D. Klahr & K. Kotovsky (Eds.), *Complex information processing: The impact of Herbert A. Simon*. NY: Hillsdale: Erlbaum.

Ferrando, P. J., & Lorenzo-Seva, U. (2007a). An item response theory model for incorporating response time data in binary personality items. *Applied Psychological Measurement*, *31*, 525–543.

Ferrando, P. J., & Lorenzo-Seva, U. (2007b). A measurement model for Likert responses that incorporates response time. *Multivariate Behavioral Research*, *42*, 675–706.

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

Grabner, R. H., Ansari, D., Koschutnig, K., Reishofer, G., Ebner, F., & Neuper, C. (2009). To retrieve or to calculate? Left angular gyrus mediates the retrieval of arithmetic facts during problem solving. *Neuropsychologia*, *47*(2), 604–608.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457–511.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). Introducing markov chain monte carlo. In: *Markov chain Monte Carlo in practice* (pp. 1–19). US: Springer.

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights From a computer-based large-scale assessment. *Journal of Educational Psychology*, *106*, 608–626.

Hall, D. B. (2000). Zero-inflated Poisson and binomial regression with random effects: A case study. *Biometrics*, *56*(4), 1030–1039.

Jansen, B. R., & Van der Maas, H. L. (2001). Evidence for the phase transition from Rule I to Rule II on the balance scale task. *Developmental Review*, *21*(4), 450–494.

Jeon, M., & De Boeck, P. (2016). A generalized item response tree model for psychological assessments. *Behavior Research Methods*, *48*(3), 1070–1085.

Larson, G. E., & Alderton, D. L. (1990). Reaction time variability and intelligence: A "worst performance" analysis of individual differences. *Intelligence*, *14*(3), 309–325.

Lubke, G. H., & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods*, *10*(1), 21.

Luce, R. D. (1986). *Response times*. New York: Oxford University Press.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*, 19–40.

McLeod, L., Lewis, C., & Thissen, D. (2003). A Bayesian method for the detection of item preknowledge in computerized adaptive testing. *Applied Psychological Measurement*, *27*(2), 121–137.

Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin*, *115*, 300–300.

Min, Y., & Agresti, A. (2005). Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling*, *5*(1), 1–19.

Molenaar, D., Oberski, D., Vermunt, J., & De Boeck, P. (2016). Hidden Markov item response theory models for responses and response times. *Multivariate Behavioral Research*, *51*(5), 606–626.

Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. J. (2015). A bivariate generalized linear item response theory modeling framework to the analysis of responses and response times. *Multivariate Behavioral Research*, *50*(1), 56–74.

Partchev, I., & De Boeck, P. (2012). Can fast and slow intelligence be differentiated? *Intelligence*, *40*(1), 23–32.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing (DSC 2003)* (pp. 20–22).

Rabbitt, P. (1979). How old and young subjects monitor and control responses for accuracy and speed. *British Journal of Psychology*, *70*, 305–311.

Ranger, J., & Ortner, T. M. (2011). Assessing personality traits through response latencies using item response theory. *Educational and Psychological Measurement*, *71*(2), 389–406.

Rummel, R. J. (1970). *Applied factor analysis*. Evanston: Northwestern University Press.

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

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review*, *84*, 127.

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Gilks, W. R. (1995). *BUGS: Bayesian inference using Gibbs sampling, Version 0.50*. Cambridge: MRC Biostatistics Unit.

Spiegelhalter, D. J., Best, N. G.-, Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society Series B*, *64*, 583–640.

Stan Development Team. (2015). Stan modeling language users guide and reference manual. Version *2*(9)

Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive Psychology*, *8*, 481–520.

Thissen, D. (1983). Timed testing: An approach using item response testing. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing* (pp. 179–203). New York: Academic Press.

Thomas, A., Hara, B. O., Ligges, U., & Sturtz, S. (2006). Making BUGS open. *R News*, *6*, 12–17.

Titterington, D. M., Smith, A. F. M., & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. Chicester: Wiley.

Tuerlinckx, F., & De Boeck, P. (2005). Two interpretations of the discrimination parameter. *Psychometrika*, *70*(4), 629–650.

Van Harreveld, F., Wagenmakers, E. J., & Van Der Maas, H. L. (2007). The effects of time pressure on chess skill: An investigation into fast and slow processes underlying expert performance. *Psychological Research*, *71*, 591–597.

Van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*, 287–308.

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., Breithaupt, K., Chuah, S. C., & Zhang, Y. (2007). Detecting differential speededness in multistage testing. *Journal of Educational Measurement*, *44*(2), 117–130.

Van der Linden, W. J., & Glas, C. A. (2010). Statistical tests of conditional independence between responses and/or response times on test items. *Psychometrika*, *75*, 120–139.

Van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, *73*(3), 365–384.

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*(5), 327–347.

van der Linden, W. J., & van Krimpen-Stoop, E. M. (2003). Using response times to detect aberrant responses in computerized adaptive testing. *Psychometrika*, *68*(2), 251–265.

van der Maas, H. L., Molenaar, D., Maris, G., Kievit, R. A., & Borsboom, D. (2011). Cognitive psychology meets psychometric theory: On the relation between process models for decision making and latent variable models for individual differences. *Psychological review*, *118*(2), 339–356.

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, *68*, 456–477.

Wang, C., Xu, G., & Shang, Z. (2016). A two-stage approach to differentiating normal and aberrant behavior in computer based testing. *Psychometrika*. https://doi.org/10.1007/s11336-016-9525-x.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable Information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*, 3571–3594.

Yung, Y. F. (1997). Finite mixtures in confirmatory factor-analysis models. *Psychometrika*, *62*(3), 297–330.