RUNNING HEAD: Semi-Parametric RT models

A Semi-Parametric Within-Subject Mixture Approach to the Analyses of Responses and Response Times

Dylan Molenaar[1], Maria Bolsinova[1], & Jeroen Vermunt[2]

[1] University of Amsterdam, The Netherlands

[2] Tilburg University, The Netherlands

Correspondence concerning this manuscript should be addressed to: Dylan Molenaar, Psychological Methods, Department of Psychology, University of Amsterdam, Postbus 15906,
1001 NK, Amsterdam, The Netherlands, telephone: +31 205256584, email: D.Molenaar@uva.nl.

* The research by Dylan Molenaar was made possible by a grant from the Netherlands Organization for Scientific Research (NWO VENI- 451-15-008). The latentGOLD syntax to fit the models is available from www.dylanmolenaar.nl

Abstract

In item response theory, modeling the item response times in addition to the item responses may improve the detection of possible between- and within-subject differences in the process that resulted in the responses. For instance, if respondents rely on rapid guessing on some items but not on all, the joint distribution of the responses and response times will be a multivariate within-subject mixture distribution. Suitable parametric methods to detect these within-subject differences have been proposed. In these approaches, a distribution needs to be assumed for the within-class response times. In this paper, it is demonstrated that these parametric within-subject approaches may produce false positives and biased parameter estimates if the assumption concerning the response time distribution is violated. A semi-parametric approach is proposed which hardly produces false positives and parameter bias. In addition, the semi-parametric approach has approximately the same power to detect within-subject differences in responses and response times as compared to the parametric approach.

Keywords: Item Response Theory, Response times, Mixture Modeling;

The interest in response times in item response theory modeling (IRT) dates back to many decennia ago (Thorndike, Bregman, Cobb, & Woodyard, 1926). Since then, effort has been devoted to the development of IRT models for responses and response times (e.g., Roskam, 1987; Thissen, 1983; see Schnipke & Scrams, 2002, for a more comprehensive overview). Recently, the work in this area was boosted by the development of a general modeling framework for responses and response times (Van der Linden 2007; 2009). In this framework, measurement models are specified for the responses and response times separately, after which these models are connected by correlating the random effects across the models. Key characteristic of this framework is that the responses and response times are independent conditional on the underlying latent speed and latent ability variables. Various instances and extensions of the general approach have been developed since then, including, for instance: multilevel models (Klein Entink, Fox, & Van Der Linden, 2009), models for different distributions of the response times (Klein Entink, Van der Linden, & Fox, 2009; Loeys, Legrand, Schettino, & Pourtois, 2014; Wang, Chang, Douglas, 2013; Wang, Fan, Chang, & Douglas, 2013; Ranger & Kuhn, 2012; Ranger & Ortner, 2012a, 2013), and models for personality data (Ferrando & Lorenzo-Seva, 2007a; 2007b). Also, some of the earlier approaches (e.g., Roskam, 1987; and Thissen 1983) are special cases.

The main purpose to incorporate the response times as an additional source of information about individual differences in the existing IRT models has been twofold (see Molenaar, 2015). First, it has been shown that the response times may improve measurement precision of the latent ability in traditional IRT models (Ranger & Ortner, 2011; Van der Linden, Klein Entink, & Fox, 2010). Second, the response times may shed light on differences in the psychological process that resulted in the responses. That is, the response times have been used to detected aberrant responses (Van der Linden & Guo, 2008; Marianti, Fox, Avetisyan, Veldkamp, & Tijmstra, 2014), guessing (Schnipke & Scrams, 1997), differences in the adopted solution strategy (Van der Maas & Jansen, 2003), item preknowledge (McLeod, Lewis, & Thissen, 2003), warming-up and slowing down effects (Van der Linden, 2009b), effects related to testing (Carpenter, Just, & Shell, 1990), and faking on personality items (Holden & Kroner, 1992).

Although response times have been successfully used for the two purposes above, some challenges still remain. For instances, with respect to improving the measurement precision, it has been shown within the general framework that the benefits of adding the response times are limited and largely apply to the easier items only (Ranger, 2013). Furthermore, with respect to detecting differences in the response process, inferences have been hampered by the focus on models for between-subject inferences only (Molenaar, Oberski, Vermunt, & De Boeck, 2016).

With respect to the latter, effort has been devoted to develop IRT models that explicitly take into account the within-subject differences in responses and response times. The conventional between-subject approaches assume that the item and person properties are constant within a given respondent. In the within-subject approaches, this is not necessarily the case. Specifically, item and/or person properties are allowed to be different for responses that differ in their response time. As a result, conditional independence between the responses and response times is violated.

To model within-subject differences, research has focused on models with two item specific classes underlying the responses and response times (DiTrapani, Jeon, De Boeck, & Partchev, 2016; Jeon and De Boeck, 2016; Molenaar, Oberski, Vermunt, & De Boeck, in press; Molenaar, Bolsinova, Rozsa, and De Boeck, 2016; Partchev & De Boeck, 2012; Wang & Xu, 2015;). In one class the item properties of the faster responses are modeled, and in the other class, the item properties of the slower responses are modeled. Next, class membership may vary from item to item for each respondent. In this way, within-subject differences are captured by the class variables enabling inferences about differences in the underling response processes. Thus, in these approaches, within-subject differences arise because of discrete differences in the response process. These differences may reflect true discrete differences in the response process (e.g., guessing and non-guessing, two different solution strategies, or item preknowledge on some of the items). However, the classes do not necessarily need to be substantively interpretable. They can also be seen a statistical tool to capture the heterogeneity of the responses with respect to the response times. That is, there may be more classes in the data, or the measurement properties may differ continuously across the response times (see Bolsinova, Tijmstra, & De Boeck, 2016; Bolsinova, Tijmstra, & Molenaar, 2016; Fox, & Marianti, 2016), however, the two classes in the model are used to statistically capture the most important patterns in the data.

In the models for discrete within-subject differences, Partchev and De Boeck (2012), DiTrapani et al. (2016), and Jeon and De Boeck (2016) operationalized the faster and slower classes by dichotomizing the response times to obtain the item class variables for each respondent. This approach results in deterministic classes with the class size chosen by the researcher (i.e., depending on the cutoff point that is used to dichotomize the response times). In addition, the amount of information in the continuous response times is reduced. To this end, Molenaar et al. (in press) proposed an approach based on mixtures modeling (see also Wang & Xu, 2015). In this approach, the classes are operationalized by a two component multivariate mixture distribution on the responses and response times simultaneously. As a result, the classes are stochastic with the class sizes estimated from the data. In addition, the continuous nature of the response times is retained. However, to enable such a mixture modeling approach, the distribution of the response times within each class needs to be specified. Molenaar et al. and Wang and Xu presented approaches for log-normal response time distributions within each class.

The aim of the present study is twofold. First, it will be demonstrated that the within-subject mixture modeling framework is sensitive to violations of the assumed response time distribution. That is, if the response time distribution departs from the assumed distribution: 1) spurious classes may be detected if there are no classes underlying the data; and 2) parameter estimates are biased if there are truly different classes in the data. Key of the problem is the misspecification of the response time distribution which can obviously be solved by specifying a more appropriate response time distribution for the data. However, doing so is challenging as it is hard to infer the true distribution within each class from the data. That is, the observed response time distribution will depart from the within-class distribution by definition because of the mixture of the two within-class distributions. For instance, if the within-class distribution is log-normal, the observed marginal response time distribution will depart from a log-normal distribution. Thus, it is unclear whether departures from log-normality reflect a mixture of two classes or whether the departures reflect a misspecified response time distribution . Therefore, it is hard to infer a plausible distribution for the within-class response time distributions from the marginal response time data.

A second aim of the present study is that it will be shown that the problem outlined above can be remedied by adopting a semi-parametric within-subject mixture modeling approach. This is a practical but effective approach in which the distributional assumption on the response times is relaxed by categorizing the response times into an arbitrary number of categories. Next, to the responses and categorized response times, a suitable within-subject mixture model is applied that takes the categorical nature of the response times into account. We refer to this approach as ‘semi-parametric’ as the assumption on the response time distribution is less stringent as compared to the parametric (log-normal modeling) approach. In a simulation study we show that the semi-parametric approach hardly results in false positives or parameter bias even if the response time distribution is truncated or highly skewed. In addition, it is shown that the power to detect the different classes in the data is not affected in the semi-parametric approach as compared to the parametric approach.

The outline is as follows: First, we present the parametric within-subjects mixture model with log-normal response times within the classes. Next, in a simulation study we show that this model is associated with false positives and parameter bias if the assumption of log-normal response times is violated. Then, we present the semi-parametric alternative and we show on the same simulated datasets as above that this approach does hardly suffer from false positives and parameter bias. Then, we apply the parametric and semi-parametric approaches to a real dataset pertaining to logical reasoning. We end with a general discussion.

The Parametric Within-Subject Mixture Model

In the parametric within-subject mixture approach, a latent class variable C_{pi} is assumed to underlie the response of respondent *p* on item *i* (Molenaar et al., in press; Wang & Xu, 2015). In principle, C_{pi} can have multiple levels, referred to as states. Here, we focus on two states, a slower state $C_{pi} = 0$, and a faster state, $C_{pi} = 1$, which are all collected in the state vector $\mathbf{c}_p = [C_{p1}, C_{p2}, \ldots, C_{pn}]$. The probability of response vector $\mathbf{x}_p = [X_{p1}, X_{p2}, \ldots, X_{pn}]$ is then given by

$$P(\boldsymbol{x}_p|\theta_p, \boldsymbol{c}_p) = \prod_{\boldsymbol{i=1}}^{\boldsymbol{n}} \omega(\alpha_{si} \times \theta_p + \beta_{si})^{x_{pi}} \omega(-[\alpha_{si} \times \theta_p + \beta_{si}])^{1-x_{pi}} \qquad (1)$$

where θ_p is the latent ability, ω(.) is the logistic function, α_{si} is the discrimination of item *i* in state *s* = 0, 1, and β_{si} is the easiness of item *i* in state *s*. Next, within each state, the response times are assumed to have a log-normal distribution such that the vector of log-transformed response times, $\mathbf{t}_p = [\ln T_{p1}, \ln T_{p2}, \ldots, \ln T_{pn}]$ can be modeled using a conditional multivariate normal distribution with uncorrelated dimensions, that is,

$$f(\boldsymbol{t}_p|\tau_p, \boldsymbol{c}_p) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_{\varepsilon i}^2}} \exp\left[-\frac{1}{2}\frac{(lnT_{pi} - \mu_{pi}|\tau_p, C_{pi})^2}{\sigma_{\varepsilon i}^2}\right] \qquad (2)$$

with

$$\mu_{pi}|\tau_p, C_{pi} = E\left(lnT_{pi}\middle|\tau_p, C_{pi}\right) = \nu_i - \delta \times C_{pi} - \tau_p \qquad \text{with } \delta > 0 \tag{3}$$

where τ_p is the latent speed, $\sigma_{\varepsilon i}^2$ is the residual variance, ν_i is the time intensity, and δ is the difference in log-response time between the states $C_{pi} = 0$ and $C_{pi} = 1$. The constraint $\delta > 0$ is imposed to ensure that state $C_{pi} = 1$ correspond to the faster state (i.e., response times in this state are smaller).

In the model given by Equations 1, 2, and 3, it is assumed that the item effects are fixed and the subject effects are random (see Molenaar, Tuerlinckx, and Van der Maas, 2015; Ranger and Ortner, 2012b; Van der Linden & Guo, 2008; Wang, Chang, & Douglas, 2013; Wang, Fan, Chang, & Douglas 2013). For the random subject effects, θ_p and τ_p, a bivariate normal distribution is assumed with means μ_θ and μ_τ, with variances σ_θ^2 and σ_τ^2, and covariance $\sigma_{\theta\tau}$. For identification reasons $\mu_\theta = \mu_\tau = 0$ and $\sigma_\theta^2 = 1$. No further constraints are needed to identify the model. The latent class variable, C_{pi}, is assumed to be distributed according to a Bernoulli distribution with success probability π, such that

$$P(\boldsymbol{c}_p) = \prod_{i=1}^{n} \boldsymbol{\pi}^{\boldsymbol{C_{pi}}}(1-\pi)^{\boldsymbol{1-C_{pi}}}. \tag{4}$$

Thus, it is assumed that the item states are independent and time homogenous (i.e., the item states have equal state probabilities across items) with $P(C_p=1) = \pi$. It is possible to relax the independence assumption by introducing a time homogenous first-order Markov structure on the item states (e.g., MacDonald, & Zucchini, 1997; Vermunt, Langeheine, & Bockenholt, 1999), see Molenaar et al. (in press). We will refer to the model above as the *Parametric Item States Model* (*ISM*). Note that in data for which the model above holds, the assumption of conditional independence that is commonly imposed in the framework of Van der Linden (2007) is violated.

The approach by Partchev and De Boeck (2012) to separate within-subjects from between-subject effects in responses and response times can be seen as a special case of the ISM where the class variables, C_{pi}, are observed variables. That is, the observed response times are dichotomized to obtain C_{pi}. In this way, β_{0i}, β_{1i}, α_{0i} and α_{1i} from Equation 1 can be estimated using standard IRT packages (see De Boeck & Partchev, 2012; Jeon & De Boeck, 2016). As discussed above, this approach does not take into account the measurement error in the assessment of C_{pi}. In addition, π depends on the cutoff point used to dichotomize the response times.

The free parameters in the parametric ISM include: α_{0i}, α_{1i}, β_{0i}, β_{1i}, δ, ν_i, σ_ε^2 σ_τ^2, $\sigma_{\theta\tau}$ and π for all *i*. If the parameters are collected in model parameter vector, $\boldsymbol{\eta}$, then the log marginal likelihood of response vector $\boldsymbol{x_p}$ and the log- response time vector $\boldsymbol{t_p}$ for the parametric ISM is given by

$$\ell(\boldsymbol{x}_p, \boldsymbol{t_p}; \boldsymbol{\eta}) = ln \iint_{-\infty}^{\infty} \sum_{C_{p1}}^{2} \sum_{C_{p2}}^{2} \cdots \sum_{C_{pn}}^{2} P(\boldsymbol{x}_p | \theta_p, \boldsymbol{c}_p) f(\boldsymbol{t}_p | \tau_p, \boldsymbol{c}_p) P(C_p)\, g(\theta_p, \tau_p)\, d\theta d\tau \quad (5)$$

where $P(\boldsymbol{x}_p | \theta_p, \boldsymbol{c}_p)$ is given by Equation 1, $f(\boldsymbol{t}_p | \tau_p, \boldsymbol{c}_p)$ is given by Equation 2, and g(.) is the bivariate normal density function.

Baseline model

To enable inferences about the relative goodness-of-fit of the item states model, a baseline model is needed (see Molenaar et al., in press). To derive a baseline model, the slower state is assumed to be empty (i.e., $\pi = 1$) with equal discrimination and easiness parameters in both states (i.e., $\alpha_i = \alpha_{0i} = \alpha_{1i}$ and $\beta_i = \beta_{0i} = \beta_{1i}$). In addition, $\delta = 0$. The resulting model is a latent variable model with a two parameter model for the responses and a linear model for the response times and correlated random subject effects. This model is identical to the hierarchical model for responses and response times of Van der

Linden (2007) with fixed item effects (see Molenaar, Tuerlinckx, & van der Maas, 2015; Ranger & Ortner, 2012b). We will simply refer to this model as the *Baseline Model* or *BM*.

Simulation Study 1A

In simulation study 1A we show that 1) the parametric ISM model is viable if the response times are truly log-normal; 2) if the response time distribution departs from a log-normal distribution, the parametric ISM produces false positives and biased parameter estimates.

Method

Scenarios

We simulated data according to 6 scenarios. The first 3 scenarios (S1b, S2b, and S3b) concern baseline scenarios in which the data do not include item states. The scenarios differ in the exact distribution that is used for the log-transformed responses times. These are either normal, truncated, or skewed. Specifically, we consider the following scenario's:

S1b: A normal BM. In this scenario, the data are generated using a baseline model with normally distributed log-response times. In this normal baseline model, we used $\alpha_i = 1$ for all *i*. For the easiness parameters, β_i, we used increasing, equally spaced values between -2 and 2. The time intensity parameters are chosen to $\nu_i = 2$ for all *i* and the residual response time variances are chosen to $\sigma_{\varepsilon i}^2 = 0.2$ for all *i*. In addition, $\sigma_\tau^2 = 0.0625$ and $\sigma_{\theta\tau} = 0.1$ such that the correlation between θ_p and τ_p equals $\rho_{\theta\tau} = .4$. See the top row in Figure 1 for a normal QQ-plot and a histogram of the response times to an example item within this scenario.

S2b: A truncated BM. In this scenario, the data are generated using the same setup as in S1b. However, instead of the normal distribution for the log-response times, a truncated normal distribution is used with truncation at the upper limit, $\ln T_{pi} = \log(12)$ such that the untransformed response time distribution is truncated at 12 seconds. See the middle row in Figure 1 for a normal QQ-plot and a histogram of the response times to an example item within this scenario.

S3b: A skewed BM. In this scenario, the data are generated using the same setup as in S1b. However, the normal log-response times are transformed using a Box-Cox transformation (Box & Cox, 1964). Commonly the Box-Cox transformation, $X' = (X^\lambda - 1) / \lambda$, is used to transform skewed variables (X in this case), such that the transformed variable, X', is closer to a normal distribution. Here, we use the transformation the other way around. That is, we transform the normally distributed log-response times using $\ln T_{pi}' = (\lambda \times \ln T_{pi} + 1)^\lambda$, such that the transformed log-response times, $\ln T_{pi}'$, are skewed. For transformation parameter λ we use 0.3. See the bottom row in Figure 1 for a normal QQ-plot and a histogram of the response times to an example item within this scenario.

The remaining 3 scenarios (S1s, S2s, and S3s) are scenario's in which the data do include different item states. The scenarios differ in the exact distribution that is used for the log-transformed response times. That is, each scenario corresponds to a baseline scenario above (S1b, S2b, or S3b). That is:

S1s: A normal ISM. In this scenario, the data are generated using the ISM model given by Equations 1, 2, 3, and 4. The true parameter values are chosen as follows. First, we chose $\delta = 0.5$ and $\pi = .5$. For the discrimination parameters, we used $\alpha_{0i} = 1$ and $\alpha_{1i} = 1.5$. For the easiness parameters, we used increasing, equally spaced values between -2 and 0 for β_{0i} and between 0 and 2 for β_{1i}. These differences may seem large, but together with the other parameter choices above, these values resulted in residual correlations between the responses and the log-response times of around 0.11 which are reasonable. For instance, Molenaar et al. (2016) found residual correlations between 0.07 and 0.16 in the standardization data of the Hungarian WISC-IV block design test. The response time parameters ν_i, $\sigma_{\varepsilon i}^2$, σ_τ^2, $\sigma_{\theta\tau}$ are given the same values as in the normal baseline scenario S1b.

S2s: A truncated ISM. In this scenario, the data are generated using the same setup as in S1s. However, similar as in baseline scenario S2b, we use a truncated normal distribution for the log-response times with truncation at the upper limit, $\ln T_{pi} = \log(12)$.

S3s: A skewed ISM. In this scenario, the data are generated using the same setup as in S1s. However, similar as in baseline scenario S3b, the normal log-response times are transformed using a Box-Cox transformation, with the transformation parameter, λ, equal to 0.3.

Procedure

We conducted 100 replications of each scenario. For the data within each replication, the Parametric ISM is fit (P-ISM) together with its corresponding parametric baseline model (P-BM). Next, the model fit of the P-ISM and the P-BM are compared using the Akaike Information Criterion (AIC; Akaike, 1987), the Bayesian Information Criterion (BIC; Schwarz, 1978), the AIC3 (Bozdogan, 1993), the Consistent AIC (CAIC; Bozdogan, 1987), and the sample size adjusted BIC (saBIC; Sclove, 1987). We used 20 items and 500 subjects. Models are estimated using marginal maximum likelihood estimation in the LatentGOLD software package (Vermunt & Magidson, 2013). We used 100 nodes to approximate the two integrals in the likelihood function (10 nodes for each dimension). Syntax to fit the different models is available from the website of the first author.

Results

False positive and true positive rates.

Table 1 contains the false positive and true positive rates of the P-ISM in the different scenario's. First, the false positive rate is obtained by considering the acceptance rates of the P-ISM over the P-BM in the scenarios in which the data do not contain item states (S1b, S2b, and S3b). As can be seen from Table 3, for the P-ISM, there are no false positives in the case of a baseline model with normally distributed log-response times. However, if the log-response time distribution is either truncated (S2b) or skewed (S3b) the P-ISM is never rejected (false positives rate of 1.00) despite the fact that the data do not include item states. Similarly, the true positive rate is obtained by considering the acceptance rates of the P-ISM over the P-BM in the scenarios in which the data do indeed contain different item states (S1s, S2s, and S3s). As can be seen from Table 3 the true positive rate is 1.00 in all cases.

Parameter recovery

See Table 2 for the means and standard deviations of the estimates for the class size parameter, π, the response time difference between the states, δ, the variance of τ_p, σ_τ^2, and the correlation between speed and ability, ρ, in the scenario's where the data truly contain different item states (S1s, S2s, S3s).[1] As can be seen from the table, if the within-class distribution of the log-response times is normal (S1s), parameters are adequately recovered. However, in the case of truncation (S2s) or skewness (S3s) in the distribution of the log-response times, all parameters are biased except for ρ, the correlation between θ_p and τ_p.

Box plots of the parameter estimates of the odd items in the P-ISM for the scenarios that include item states (S1s, S2s, and S3s) are depicted in Figure 2 for the item easiness parameters, β_{0i} and β_{1i}, and Figure 3 for the discrimination parameters, α_{0i} and α_{1i}. As expected, the parameters are

[1] We estimate the Cholesky decomposed covariance matrix of θ_p and τ_p. However, for the ease of presentation we transformed these parameters into σ_τ^2 and ρ. In addition, we estimated logit(π) but we present the results for π.

acceptably recovered in the P-ISM if the data is generated according to the normal item states scenario (S1s; left plot in Figure 2 and Figure 3). However, if the data is generated according to the truncated item states scenario (S2s; middle plot in Figure 2 and Figure 3) or skewed item states scenario (S3s; right plot in Figure 2 and Figure 3), the parameters are systematically biased in the P-ISM. Specifically, the difference between the faster and slower states is underestimated: In the case of truncation, β_{1i} and α_{1i} are recovered acceptably (i.e., bias seem small), but β_{0i} and α_{0i} are underestimated. In the case of skewness, β_{1i} is underestimated and β_{0i} is recovered acceptably. Parameter α_{0i} and α_{1i} seem to be hardly biased in the case of skewness but the estimates of α_{0i} have very large standard errors.

A Semi-Parametric Item States Model

As we showed in the simulation study above, the parametric model is sensitive to violations of the normality assumption in Equation 2. That is, if the distribution of the response times departs from the log-normal (e.g., the response time distribution is truncated due to an item time limit), spurious item states may be detected and parameters are biased.

As a solution, we propose a semi-parametric item states model. The semi-parametric model differs from the model above in that the response times are categorized, that is, the categorized response times, T_{pi}', are obtained from the raw response times, T_{pi}, as follows:

$$T'_{pi} = z \quad \text{if} \quad T_{pi} \in (b_{zi}, b_{(z+1)i}) \quad \text{with } z = 0, 1, \ldots, Z\text{-}1 \qquad (6)$$

where b_{zi} are the thresholds at which the response times are categorized with $b_{0i}=0$ and $b_{(Z-1)i}=\infty$, and Z denotes the number of categories that is used. Both the thresholds b_{zi} and the number of response time categories, Z, are chosen by the researcher. But as we illustrate in the real data application, multiple option can be considered to study the robustness of the results.

Next, within the semi-parametric item states model, the probability of the vector of categorized response times, $\mathbf{t}_p' = [T_{p1}', T_{p2}', \ldots, T_{pn}']$, is subjected to an adjacent categories model

$$P\left(\boldsymbol{t}_p' \middle| \tau_p, \boldsymbol{c}_p\right) = \prod_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}} \frac{\boldsymbol{exp}\left(\sum_{\boldsymbol{z}=\mathbf{0}}^{\boldsymbol{T'_{pi}}} \gamma_{\mathrm{zi}} - \delta \times C_{pi} - \tau_p\right)}{\sum_{\boldsymbol{j}=\mathbf{0}}^{\boldsymbol{Z-1}} \boldsymbol{exp}\left(\sum_{\boldsymbol{z}=\mathbf{0}}^{\boldsymbol{j}} \gamma_{\mathrm{zi}} - \delta \times C_{pi} - \tau_p\right)} \qquad \text{with } \delta > 0 \tag{7}$$

where γ_{zi} are response time category parameters for category *z* of the response times of item *i*. Category parameter γ_{0i} is chosen in such a way that

$$\sum_{\boldsymbol{z}=\mathbf{0}}^{\mathbf{0}} -\delta_s - \tau_p + \gamma_{0\mathrm{i}} = 0. \tag{8}$$

Equation 7 together with the model for the responses in Equation 1 and the bivariate normal distribution for θ_p and τ_p constitute the full model. The free parameters in the semi-parametric ISM include: α_{0i}, α_{1i}, β_{0i}, β_{1i}, γ_{zi}, δ, σ_τ^2, $\sigma_{\theta\tau}$, and π for all *i* and all *z > 0.* If these parameters are collected in model parameter vector, $\boldsymbol{\zeta}$, then the log marginal likelihood of response vector $\boldsymbol{x_p}$ and the categorized response time vector $\boldsymbol{t_p'}$ for the semi-parametric ISM is given by

$$\ell\left(\boldsymbol{x}_p, \boldsymbol{t}_p'; \boldsymbol{\zeta}\right) = ln \iint_{-\infty}^{\infty} \sum_{C_{p1}}^{2} \sum_{C_{p2}}^{2} \cdots \sum_{C_{pn}}^{2} P\left(\boldsymbol{x}_p \middle| \theta_p, \boldsymbol{c}_p\right) P\left(\boldsymbol{t}_p' \middle| \tau_p, \boldsymbol{c}_p\right) P\left(C_p\right) g(\theta_p, \tau_p) d\theta d\tau \tag{9}$$

where $P\left(\boldsymbol{x}_p \middle| \theta_p, \boldsymbol{c}_p\right)$ is given by Equation 1 and $P\left(\boldsymbol{t}_p' \middle| \tau_p, \boldsymbol{c}_p\right)$ is given by Equation 7.

Baseline model

For the semi-parametric item states model, the baseline model can be derived in a similar way as was done for the parametric normal model above. The resulting model is a latent variable model with a two parameter model for the responses and a partial credit model for the categorized response times and correlated random subject effects. This model can be seen as a generalization of the hierarchical model for responses and response times of Van der Linden (2007) for categorical response times and fixed item effects.

Simulation Study 1B

In this simulation study we analyze the same datasets as in simulation study 1A. We show in these data that 1) the semi-parametric approach as discussed above hardly suffers from the increased false positive rate or the parameter bias as was found for the parametric approach; while 2) the semi-parametric approach is still capable of detecting truly different item states in the data with acceptable true positive rates.

Method

Procedure

We used the same 100 replications of the 6 scenarios as in simulation study 1a. To these data, we fit the three Semi-parametric ISMs with respectively Z=7, Z=5, and Z=3 response time categories (referred to as S-ISM7, S-ISM5, and S-ISM3). In addition, we fit the corresponding baseline models (S-BM7, S-BM5, and S-BM3).

For the response time categorization in Equation 6, b_{0i} and b_{Zi} are 0 and ∞ by definition. The remaining thresholds, b_{1i}, b_{2i}, …, $b_{(Z-1)i}$ are chosen at the Z-quantiles of the observed response time distribution of item *i*, where Z is the number of thresholds used to categorize the response times as defined above. We consider this specific procedure to categorize the response times as desirable because the thresholds depend on the shape of the response time distribution. In addition, by using this approach, it does not matter whether the raw response times or the log-response times are categorized as the resulting categorization will be equivalent.

For each dataset, the fit of the three item state models (S-ISM7, S-ISM5, and S-ISM3) is compared to its corresponding baseline model (S-BM7, S-BM5, S-BM3). All other details concerning model estimation and model fit (i.e., the fit indices used, the software, the estimation algorithm, and the number of nodes) are the same as in the simulation study 1a. Syntax to fit the different models is available from the website of the first author.

Results

False positives.

In Table 3, the false positive rates are depicted for the item states models (S-ISM7, S-ISM5, and S-ISM3) in the scenarios in which the data do not contain item states (S1b, S2b, and S3b). As can be seen from the table, the semi-parametric models do not suffer from false positives with false positive rates of 0.00 for all fit indices except the AIC. The AIC fit index is associated with an increased false positive rate for the semi-parametric model with rates between 0.02 and 0.08.

True Positives.

In Table 4, the true positives rates are depicted for the item state models in the case of the scenarios in which the data truly contain item states (S1s, S2s, and S3s). True positive rates of 0.80 or larger are considered as acceptable. As can be seen from the table, generally, the true positive rate is acceptable for all models. An exception is the true positive rate of 0.54 for the CAIC of the semi-parametric item states model with Z=3 in the case of a truncated response time distribution (scenario S2s).

Parameter recovery

See Table 5 for the means and standard deviations of the estimates for the class size parameter, π, the response time difference between the states, δ, the variance of τ_p, σ_τ^2, and the correlation between speed and ability, ρ, in the scenario's where the data truly contain different item states (S1s, S2s, S3s). As can be seen from the table, π and ρ are recovered adequately in all scenario's. However, the mean estimates of δ and σ_τ^2 are not close to the true parameter value. However, this is not surprising as both

δ and σ_τ^2 are dependent upon the scale of the categorized response times which differs for different number of response time categories and different thresholds, β_{zi}. But note that ρ, the correlation between θ_p and τ_p, which is calculated from σ_τ^2 is unaffected by this scale difference. This parameter is adequately recovered.

Box plots of the parameter estimates of the odd items in the semi-parametric item state models (S-ISM7: top row; S-ISM5: middle row; and S-ISM3: bottom row) for the scenarios that include item states (S1s, S2s, and S3s) are depicted in Figure 4 for the item easiness parameters, β_{0i} and β_{1i}, and Figure 5 for the discrimination parameters, α_{0i} and α_{1i}. Note again that these models have been fit to the same simulated data sets as used for the parametric model in Figure 2 and Figure 3. As can be seen, for all scenarios and all semi-parametric models, the estimates tend to be unbiased with reasonable standard errors. That is, the parameters are acceptably recovered irrespective of the distribution of the response times.

Overall conclusion

As appears from the results of simulation study 1A and 1B above, if the log-response time distribution departs from normality but a normal item states model is applied nevertheless, spurious i tem states may be detected by the AIC, BIC, AIC3, CAIC, and saBIC if the data do not contain different item states. If the data do contain different item states, the normal item states model is still able to detect these, however, parameter estimates are biased. The proposed class of semi-parametric model with Z=7, Z=5, and Z=3 were shown to not suffer from these problems while the power to detect different item states in the data was hardly affected.

Illustration

Data

The data comprise the responses and response times of 664 Dutch high school students to the 23 items of the so-called “puzzles” test. This test is based on the Raven progressive matrices test (Raven, 1962).

Each item consists of a matrix that constitutes a pattern but with one element missing. The respondents have to indicate which of 5 optional elements would complete the pattern. The items are administered using a 40 seconds deadline. As a result, the observed response times show truncation effects with the severity of the effect increasing for the later items because the items are of increasing difficulty. 36 respondents are omitted from the analyses because they showed suspiciously small response times (1 second or faster) resulting in a sample size of 628 respondents.

To the data we fitted the same parametric and semi-parametric baseline and item states models as considered in the simulation study. We were interested to see whether the results (parameter estimates and model fit) are similar across the different approaches. Parameter estimation and assessment of model fit is conducted using the same procedure as outlined in the simulation study section.

Results

See Table 6 for the model fit indices of the different models. As can be seen, for all semi-parametric and parametric approaches, the ISM is the best fitting model according to the indices considered. One exception is the S-BM3 which is favored over S-ISM3 by the CAIC. However, in the simulation study, the CAIC was already shown to have poor power in the case of Z=3 and truncation, see Table 4. We therefore accept the ISM model and look into the parameter estimates within this model for the semi-parametric and parametric approach.

In Table 7 for the parameters estimates of the class size parameter, π, the response time difference between the states, δ, the variance of τ_p, σ_τ^2, and the correlation between speed and ability, ρ, in the ISM models. As can be seen, in the parametric model (P-ISM), the estimate of the faster class size, π, is substantially smaller than in the semi-parametric models (S-ISM), .16 versus .38-.44. In addition, the estimate of π is relatively stable across the semi-parametric models. The estimate of the response time difference, δ, fluctuates between the semi-parametric models. However, this is expected as the scale of τ_p on which δ is a parameter, depends on the number of response time categories. This is also reflected in the estimates of the variance of τ_p which differs across the semi-parametric models. The correlation between θ_p and τ_p (i.e., ρ, which we calculated from the estimates of $\sigma_{\theta\tau}$ and σ_τ^2) is however stable across the semi-parametric models. In addition, the estimated correlation does not differ importantly between the parametric and semi-parametric approaches.

In Figure 6 parameter estimates of β_{0i}, β_{1i}, α_{0i}, and α_{1i} are depicted for the different models. In the figure, the items are ordered according to the estimates in S-ISM3 for clarity. As can be seen, the estimates of the semi-parametric models are close to each other. The estimates of the parametric approach deviate most notably from the semi-parametric approach for β_{0i} and α_{0i}. This is congruent with what we found in the truncation scenario of the simulation study.

To conclude, results seem to be stable between the semi-parametric approaches. That is, the exact number of response time categories does affect the results importantly. There are, however, notable differences between the semi-parametric approach and the parametric approach in the class size parameter, π, and the item parameters. Nevertheless, as we know from the simulation study that the semi-parametric models are less sensitive to violations of normality in the log-response times, and because the results of the semi-parametric models are largely insensitive to the number of response time categories, we trust the results from the semi-parametric better than those of the parametric model.

Discussion

In the simulation study we established that the parametric item states model is associated with a substantial false positive rate and parameter bias if the log-response times are not normally distributed. The proposed solution to this problem, a semi-parametric model for the responses and categorized response times was shown to not suffer from this problem, while the true positive rates are still comparable to those of the parametric model.

Generally, categorization of continuous variables is discouraged due to the loss of information about individual differences, smaller power, and the arbitrary nature of the thresholds (Cohen, 1983; MacCallum, Zhang, Preacher, & Rucker, 2002; Maxwell & Delaney, 1993). In the present mixture framework it can however be desirable to categorize the response times such that violations of the assumed distribution do not affect the results. In addition, we showed that although the power is indeed affected, for our parameter choices in the simulation study, this effect was not large. However, in other situations not covered by the simulation study, the loss of power may be larger. The present approach can therefore be seen as a conservative approach to the within-subject analysis of responses and response times. Furthermore, although the number and the location of the thresholds are indeed arbitrary, in the simulation study and the real data application, we showed that results are largely consistent across models with different numbers of response time categories. In practice we thus advice to always fit the semi-parametric approach using different numbers of response time categories to investigate the stability of the results.

With respect to the exact categorization of the response times, we chose a quantile-based approach resulting in equal-distant scores that are uniformly distributed. This approach was shown to perform well in the simulation study in terms of parameter recovery and power. However, an alternative approach might be to use the mid-points within each category such that the categorized distribution resembles the observed response time distribution better.

In the present paper, we demonstrated that if the data do not contain classes (item states) with different response and response time properties and a normal distribution is wrongfully assumed for the log-transformed response times, spurious classes may arise. The same will hold for the case where there are two classes underlying the data, if a normal log-response time model is applied to these data, additional classes may be detected.

In the present undertaking, we assumed the classes to be independent. However, it would be interesting to consider relaxing this assumption in future work by extending the present approach to include a Markov structure on the item states.

References

Akaike, H. (1974). A new look at the statistical model identification. Automatic Control, IEEE Transactions on, 19, 716-723.

Bolsinova, M., Tijmstra, J., & De Boeck, P. (in press). Modeling conditional dependence between response time and accuracy. *Psychometrika*.

Bolsinova, M., Tijmstra, J., & Molenaar, D. (in press). Response moderation models for conditional dependence between response time and accuracy. *British Journal of Mathematical and Statistical Psychology.*

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B*, 211-252.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*, 345-370. doi:10.1007/BF02294361

Bozdogan, H. (1993). Choosing the number of component clusters in the mixture model using a new informational complexity criterion of the inverse Fisher information matrix. In: Opitz, O., Lausen, B., Klar, R., eds. *Studies in Classification, Data Analysis, and Knowledge Organization* (pp. 40–54). Heidelberg: Springer

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. DOI: 10.1037/0033-295X.97.3.404

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

De Boeck, P., & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software*, *48*(1), 1-28. DOI: 10.1.1.302.6429

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. Doi: 10.1016/j.intell.2016.02.012

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. DOI: 10.1177/0146621606295197

Ferrando, P. J., & Lorenzo-Seva, U. (2007b). A measurement model for Likert responses that incorporates response time. *Multivariate Behavioral Research, 42*, 675-706. DOI:10.1080/00273170701710247

Fox, J. P., & Marianti, S. (2016). Joint Modeling of Ability and Differential Speed Using Responses and Response Times. *Multivariate behavioral research*, *51*, 530-553. DOI: 10.1080/00273171.2016.1171128

Holden, R. R., & Kroner, D. G. (1992). Relative efficacy of differential response latencies for detecting faking on a self-report measure of psychopathology.*Psychological Assessment*, *4*(2), 170. DOI: 10.1037/1040-3590.4.2.170

Jeon, M., & De Boeck, P. (2016). A generalized item response tree model for psychological assessments. *Behavior research methods*, *48*, 1070-1085. DOI: 10.3758/s13428-015-0631-y

Klein Entink, R. H., Fox, J. P., & Van Der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers.*Psychometrika*, *74*(1), 21-48. DOI:10.1007/s11336-008-9075-y

Klein Entink, R. H., Linden, W. J., & Fox, J. P. (2009). A Box–Cox normal model for response times. *British Journal of Mathematical and Statistical Psychology*,*62*(3), 621-640. DOI: 10.1348/000711008X374126

Loeys, T., Legrand, C., Schettino, A., & Pourtois, G. (2014). Semi-parametric proportional hazards models with crossed random effects for psychometric response times. *British Journal of Mathematical and Statistical Psychology*, 67, 304–327. DOI: 10.1111/bmsp.12020

MacDonald, I. L., & Zucchini, W. (1997). *Hidden Markov and other models for discrete-valued time series* (Vol. 110). CRC 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. DOI: 10.1037//1082-989X.7.1.19

Marianti, S., Fox, J. P., Avetisyan, M., Veldkamp, B. P., & Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *Journal of educational and behavioral statistics*, *39*(6), 426-451. DOI: 10.3102/1076998614559412

Maxwell, S. E., & Delaney, H. D. (1993). Bivariate median splits and spurious statistical significance. *Psychological Bulletin, 113*, 181–190. DOI: 10.1037/0033-2909.113.1.181

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. DOI: 10.1177/0146621602250534

Molenaar, D., Oberski, D., Vermunt, J., De Boeck, P. (in press). Hidden Markov IRT Models for Responses and Response Times. *Multivariate Behavioral Research*.

Molenaar, D., Bolsinova, M., Rozsa, S., & De Boeck, P. (2016). Response Mixture Modeling of Intraindividual Differences in Responses and Response Times to the Hungarian WISC-IV Block Design Test. *Journal of Intelligence*,*4*(3), 10. DOI:10.3390/jintelligence4030010

Molenaar, D., Tuerlinckx, F., & van der Maas, H.L.J. (2015). A Generalized Linear

Factor Model Approach to the Hierarchical Framework for Responses and Response Times. *British Journal of Mathematical and Statistical Psychology*, *68*(2), 197-219.

Mollenkopf, W. G. (1950). An experimental study of the effects on item-analysis data of changing item placement and test time limit. *Psychometrika, 15*, 291-315. DOI:10.1007/BF02289044

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

Raven. J. C. (1962). Advanced Progressive Matrices. Set II. London: H. K. Lewis & Co. Distributed in the USA by The Psychological Corporation. San Antonio. Texas.

Ranger, J. (2013). A Note on the Hierarchical Model for Responses and Response Times in Tests of van der Linden (2007). *Psychometrika*, *78*(3), 538-544. DOI:10.1007/s11336-013-9324-6

Ranger, J., & Kuhn, J. T. (2012). A flexible latent trait model for response times in tests. *Psychometrika*, *77*(1), 31-47. DOI: 10.1007/s11336-011-9231-7

Ranger, J., & Ortner, T. (2011). Assessing personality traits through response latencies using item response theory. *Educational and Psychological Measurement, 71*, 389–406. DOI: 10.1177/0013164410382895

Ranger, J., & Ortner, T. (2012a). The case of dependency of responses and response times: A modeling approach based on standard latent trait models. *Psychological Test and Assessment Modeling*, *54*(2), 128.

Ranger, J., & Ortner, T. (2012b). A latent trait model for response times on tests employing the proportional hazards model. *British Journal of Mathematical and Statistical Psychology*, *65*(2), 334-349. DOI: 10.1111/j.2044-8317.2011.02032.x

Ranger, J., & Ortner, T. M. (2013). Response time modeling based on the proportional hazards model. *Multivariate behavioral research*, *48*(4), 503-533. DOI:

10.1080/00273171.2013.796280

Roskam, E. E. (1987). Toward a psychometric theory of intelligence. In E. E. Roskam & R. Suck (Eds.), *Progress in mathematical psychology* (pp. 151–171). Amsterdam: North-Holland.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, *34*(3), 213-232. DOI: 10.1111/j.1745-3984.1997.tb00516.x

Schnipke, D.L., & Scrams, D.J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In C.N. Mills, M. Potenza, J.J. Fremer & W. Ward (Eds.), Computer Based Testing: Building the Foundation for Future Assessments (pp. 237–266). Hillsdale, NJ: Lawrence Erlbaum Associates.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics, 6*, 461-464.

Sclove, L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika, 52*, 333-343. doi:10.1007/BF02294360

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.

Thorndike, E. L., Bregman, E. O., Cobb, M. V., & Woodyard, E. (1926). The measurement of intelligence. New York, NY: Teachers College Bureau of Publications.

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

Van der Linden, W. J. (2009). Conceptual Issues in Response-Time Modeling.*Journal of Educational Measurement*, *46*(3), 247-272. DOI: 10.1111/j.1745-3984.2009.00080.x

Van der Linden, W. J. (2009b). Predictive control of speededness in adaptive testing. *Applied Psychological Measurement, 33(1)*, 25-41. DOI:10.1177/0146621607314042

Van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, *73*(3), 365-384. DOI: 10.1007/s11336-007-9046-8.

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. DOI: 10.1177/0146621609349800

Van der Maas, H. L., & Jansen, B. R. (2003). What response times tell of children's behavior on the balance scale task. *Journal of Experimental Child Psychology*, *85*(2), 141-177. DOI: 10.1016/S0022-0965(03)00058-4

Vermunt, J. K., Langeheine, R., & Bockenholt, U. (1999). Discrete-time discrete-state latent Markov models with time-constant and time-varying covariates. *Journal of Educational and Behavioral Statistics*, *24*(2), 179-207.

Vermunt J.K., & Magidson, J. (2013). Technical Guide for Latent GOLD 5.0: Basic, Advanced, and Syntax. Belmont, MA: Statistical Innovations Inc.

Wang, C., Chang, H. H., & Douglas, J. A. (2013). The linear transformation model with frailties for the analysis of item response times. *British Journal of Mathematical and Statistical Psychology*, *66*(1), 144-168. DOI: 10.1111/j.2044-8317.2012.02045.x

Wang, C., Fan, Z., Chang, H. H., & Douglas, J. A. (2013). A semiparametric model for jointly analyzing response times and accuracy in computerized testing. *Journal of Educational and Behavioral Statistics*, *38*(4), 381-417. DOI: 10.3102/1076998612461831

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. DOI: 10.1111/bmsp.12054

Table 1.

False positive rates and true positive rates of the P-ISM as compared to its baseline model, P-BM for the different data scenario's without item states (S1b, S2b, and S3b).

	Data	**BIC**	**AIC**	**AIC3**	**CAIC**	**saBIC**
False positive rate	S1b: Normal baseline	0.00	0.00	0.00	0.00	0.00
	S2b: Truncated baseline	1.00	1.00	1.00	1.00	1.00
	S3b: Skewed baseline	1.00	1.00	1.00	1.00	1.00
True positive rate	S1s: Normal item states	1.00	1.00	1.00	1.00	1.00
	S2s: Truncated item states	1.00	1.00	1.00	1.00	1.00
	S3s: Skewed item states	1.00	1.00	1.00	1.00	1.00

Table 2.

Means (me) and standard deviations (sd) of the parameter estimates in the P-ISM in the cases where the data truly contain item states (S1s, S2s, S3s). The true parameter values are in brackets.

Scenario	π (0.50)		δ (0.50)		σ_τ^2 (0.06)		ρ (0.40)	
	me	*sd*	*me*	*sd*	*me*	*sd*	*me*	*sd*
S1s: Normal	0.50	0.04	0.50	0.05	0.06	0.01	0.40	0.05
S2s: Trunc	0.29	0.02	0.67	0.01	0.03	0.00	0.38	0.05
S3s: Skewed	0.84	0.01	2.53	0.08	0.38	0.03	0.39	0.05

Table 3.

False positive rates of the different item states models (S-ISM7, S-ISM5, and S-ISM3) as compared to their baseline models without item states (S-BM7, S-BM-5, and S-BM3) for the different data scenario's without item states (S1b, S2b, and S3b).

Model	Data	BIC	AIC	AIC3	CAIC	saBIC
S-ISM7: Semi-par. item states with Z=7	S1b: Normal baseline	0.00	**0.03**	0.00	0.00	0.00
	S2b: Truncated baseline	0.00	**0.08**	0.00	0.00	0.00
	S3b: Skewed baseline	0.00	**0.04**	0.00	0.00	0.00
S-ISM5: Semi-par. item states with Z=5	S1b: Normal baseline	0.00	**0.01**	0.00	0.00	0.00
	S2b: Truncated baseline	0.00	**0.06**	0.00	0.00	0.00
	S3b: Skewed baseline	0.00	**0.02**	0.00	0.00	0.00
S-ISM3: Semi-par. item states with Z=3	S1b: Normal baseline	0.00	**0.01**	0.00	0.00	0.00
	S2b: Truncated baseline	0.00	**0.01**	0.00	0.00	0.00
	S3b: Skewed baseline	0.00	**0.01**	0.00	0.00	0.00

Note. Non-zero rates are in boldface

Table 4.

True positive rates of the different item states models (S-ISM7, S-ISM-5 and S-ISM3) as compared to their baseline models without item states (S-BM7, S-BM5, and S-BM3) for the different data scenario's with item states (S1s, S2s, and S3s).

Model	Data	BIC	AIC	AIC3	CAIC	saBIC
S-ISM7: Semi-par. item states with Z=7	S1s: Normal item states	1.00	1.00	1.00	0.99	1.00
	S2s: Truncated item states	1.00	1.00	1.00	0.88	1.00
	S3s: Skewed item states	1.00	1.00	1.00	1.00	1.00
S-ISM5: Semi-par. item states with Z=5	S1s: Normal item states	1.00	1.00	1.00	0.99	1.00
	S2s: Truncated item states	0.99	1.00	1.00	0.82	1.00
	S3s: Skewed item states	1.00	1.00	1.00	1.00	1.00
S-ISM3: Semi-par. item states with Z=3	S1s: Normal item states	0.99	1.00	1.00	0.91	1.00
	S2s: Truncated item states	0.94	1.00	1.00	**0.54**	1.00
	S3s: Skewed item states	1.00	1.00	1.00	0.95	1.00

Note. Rates smaller than 0.80 are in bold face

Table 5.

Means (me) and standard deviations (sd) of the parameter estimates in the P-ISM in the cases where the data truly contain item states (S1s, S2s, S3s). The true parameter values are in brackets.

Model	Scenario	π (0.50)		δ (0.50)		σ_τ^2 (0.06)		ρ (0.40)	
		me	*sd*	*me*	*sd*	*me*	*sd*	*me*	*sd*
S-ISM7	S1s: Normal	0.50	0.04	2.32	0.76	0.16	0.05	0.40	0.05
	S2s: Trunc	0.50	0.07	1.57	0.80	0.09	0.04	0.40	0.05
	S3s: Skewed	0.49	0.04	2.30	0.78	0.16	0.05	0.40	0.05
S-ISM5	S1s: Normal	0.50	0.05	1.09	0.36	0.26	0.08	0.40	0.05
	S2s: Trunc	0.50	0.07	0.76	0.35	0.16	0.05	0.39	0.05
	S3s: Skewed	0.49	0.04	1.09	0.37	0.26	0.08	0.40	0.05
S-ISM3	S1s: Normal	0.49	0.06	1.03	0.32	0.50	0.12	0.40	0.05
	S2s: Trunc	0.49	0.07	0.87	0.33	0.37	0.09	0.40	0.05
	S3s: Skewed	0.48	0.05	1.04	0.33	0.51	0.12	0.40	0.05

Table 6.

Model fit indices for the different parametric and semi-parametric models in the illustration.

	Z	Model	BIC	AIC	AIC3	CAIC	saBIC
Parametric	-	P-ISM	**34752**	**34122**	**34264**	**34894**	**34302**
		P-BM	35493	35075	35169	35587	35194
Semi-parametric	7	S-ISM7	**68359**	**67320**	**67554**	**68593**	**67616**
		S-BM7	68493	67667	67853	68679	67903
	5	S-ISM5	**58826**	**57991**	**58179**	**59014**	**58229**
		S-BM5	58932	58310	58450	59072	58487
	3	S-ISM3	**44921**	**44290**	**44432**	45063	**44470**
		S-BM3	44959	44541	44635	**45053**	44660

Note. For each pair of ISM and BM models, the smallest fit indices are in bold face.

Table 7.

Parameter estimates (est.) and standard errors (se) of the class size parameter, π, the response time difference between the states, δ, the variance of the latent speed variable, σ_τ^2, and the correlation between speed and ability, ρ.

Model	π		δ		σ_τ^2		ρ	
	est	*se*	*est*	*se*	*est*	*se*	*est*	*se*
P-ISM	0.16	0.01	-0.74	0.01	0.13	0.01	-0.52	0.02
S-ISM7	0.44	0.04	-1.05	0.08	0.72	0.07	-0.48	0.04
S-ISM5	0.44	0.04	-1.32	0.11	1.16	0.10	-0.46	0.04
S-ISM3	0.38	0.04	-2.06	0.18	2.53	0.24	-0.49	0.05

Figure Captions

Figure 1. Normal QQ-plots and histograms of the log-response time distribution for an example item within the baseline scenarios (S1b, S2b, and S3b).

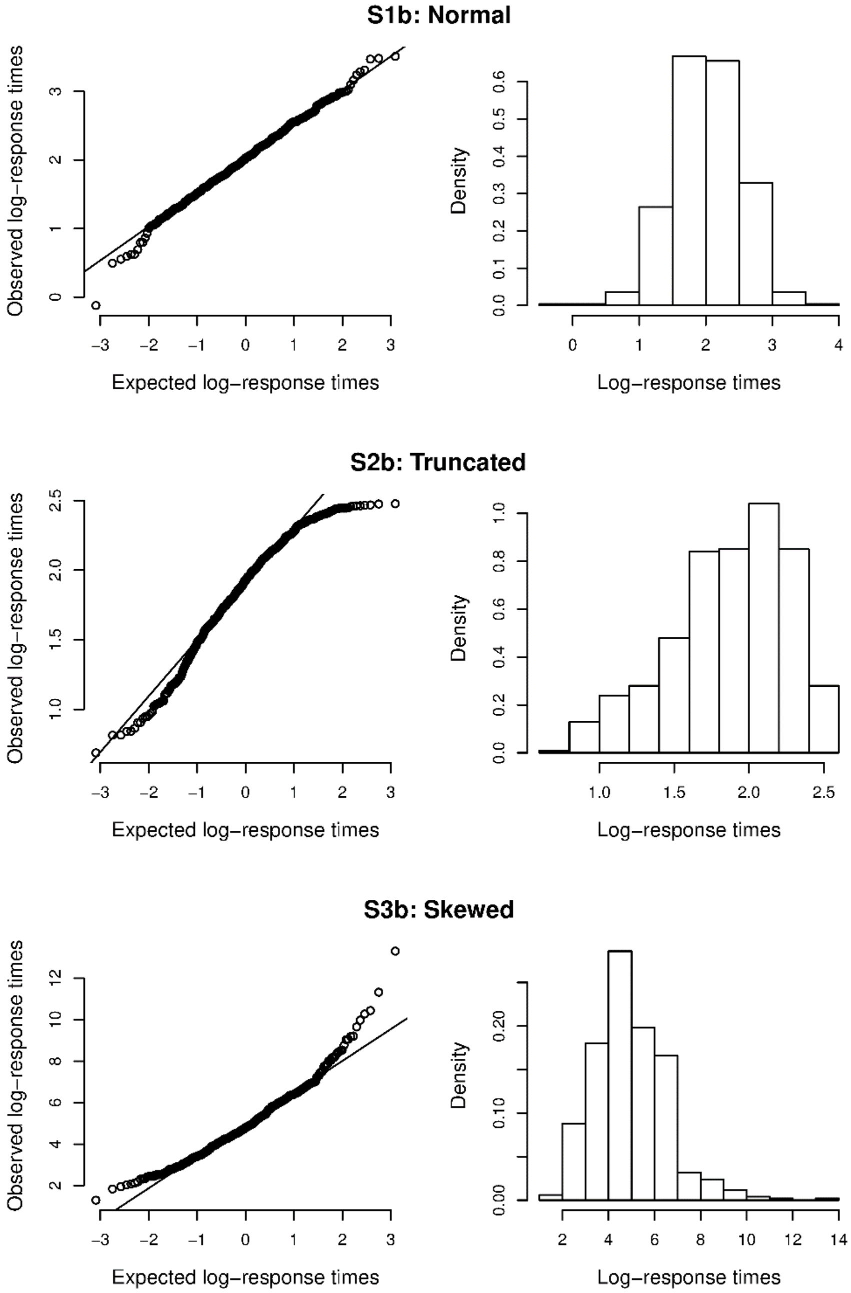
Figure 2. Box plots of the β_{0i} (white) and β_{1i} (grey) parameter estimates for the odd items in the parametric normal model (P-ISM) in the different scenarios that include item states (S1s, S2s, and S3s). The solid grey line denotes the true values of β_{0i} (lower grey line) and β_{1i} (upper grey line).

Figure 3. Box plots of the α_{0i} (white) and α_{1i} (grey) parameter estimates for the odd items in the parametric normal model (P-ISM) in the different scenarios that include item states (S1s, S2s, and S3s). The solid grey line denotes the true values of α_{0i} (upper grey line) and α_{1i} (lower grey line).

Figure 4. Box plots of the β_{0i} (white) and β_{1i} (grey) parameter estimates of the odd items in the different semi-parametric models (S-ISM7, S-ISM5, and S-ISM3) in the different scenarios that include item states (S1s, S2s, and S3s). The solid grey line denotes the true values of β_{0i} (lower grey line) and β_{1i} (upper grey line).

Figure 5. Box plots of the α_{0i} (white) and α_{1i} (grey) parameter estimates for the odd items in the different semi-parametric models (S-ISM7, S-ISM5, and S-ISM3) in the different scenarios that include item states (S1s, S2s, and S3s). The solid grey line denotes the true values of α_{0i} (upper grey line) and α_{1i} (lower grey line).

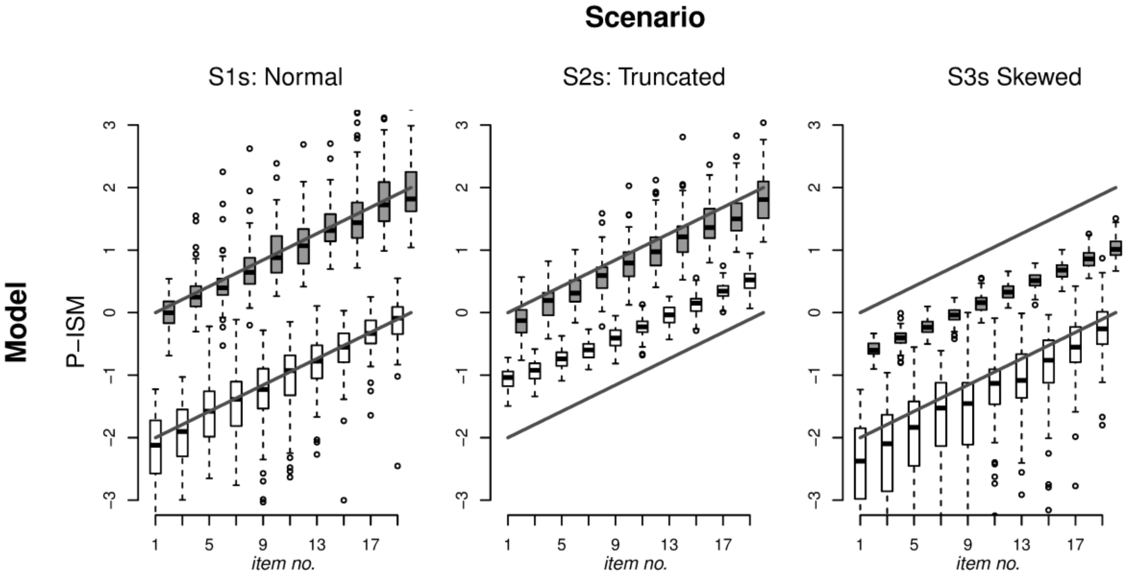
Figure 6. Plots of the β_{0i} β_{1i}, α_{0i}, and α_{1i} parameter estimates for the normal item states model (P-ISM; solid black line) and the semi-parametric item states model (S-ISM7, S-ISM5, and S-ISM3; striped grey lines). In each plot, the items are ordered on basis of the estimates in S-ISM3 for clarity.

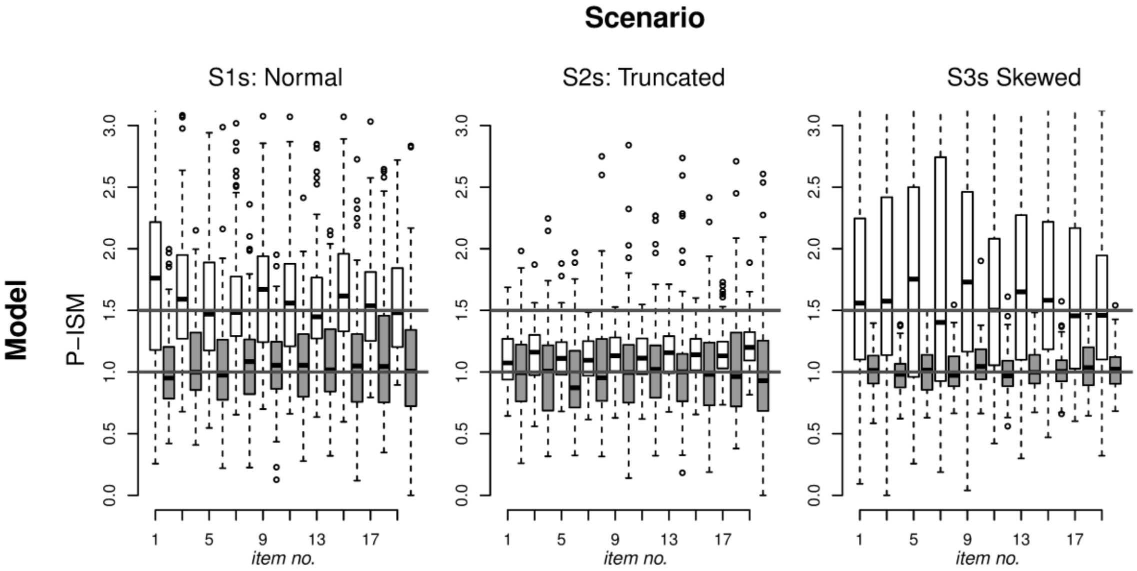


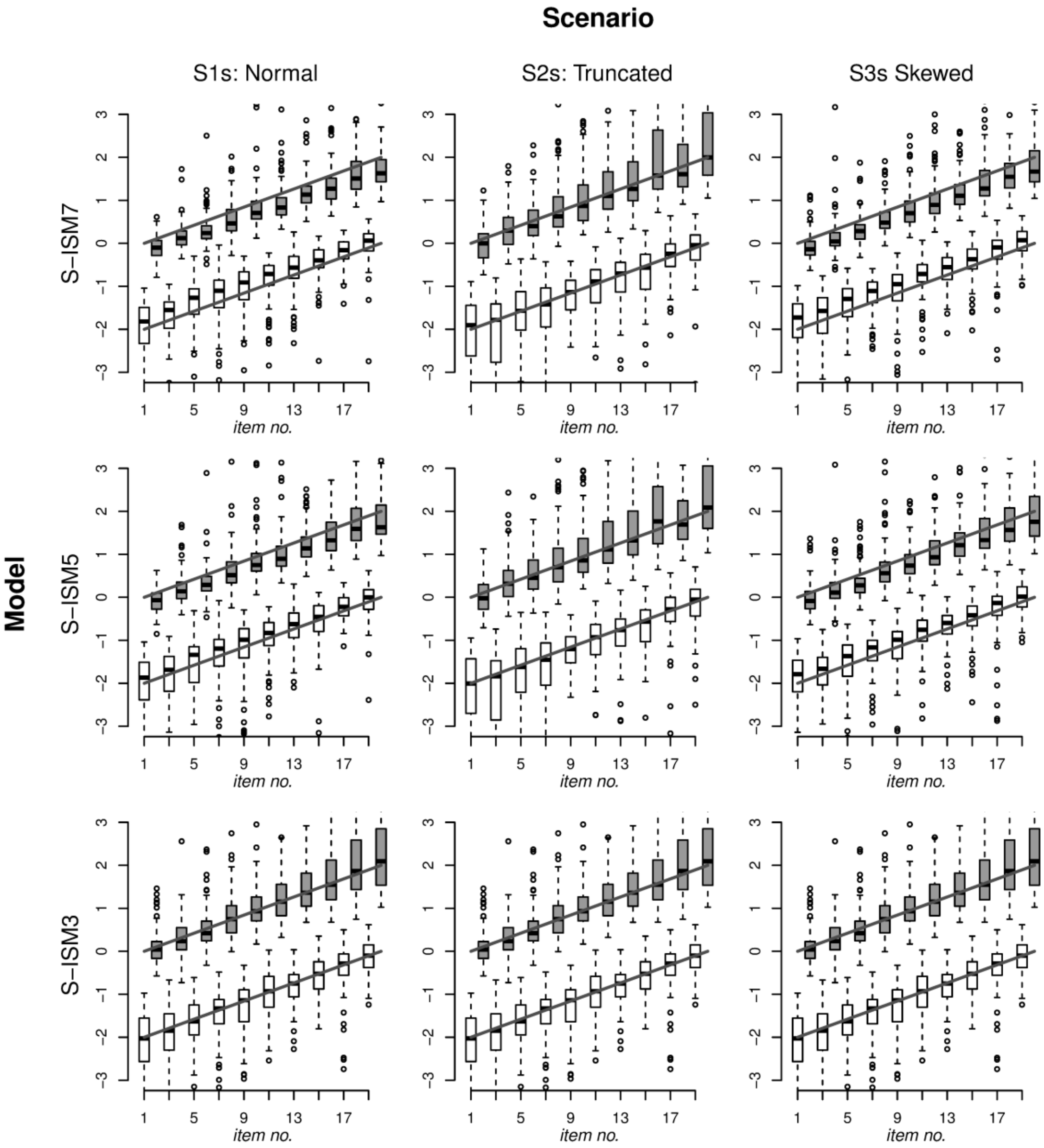
S1b: Normal

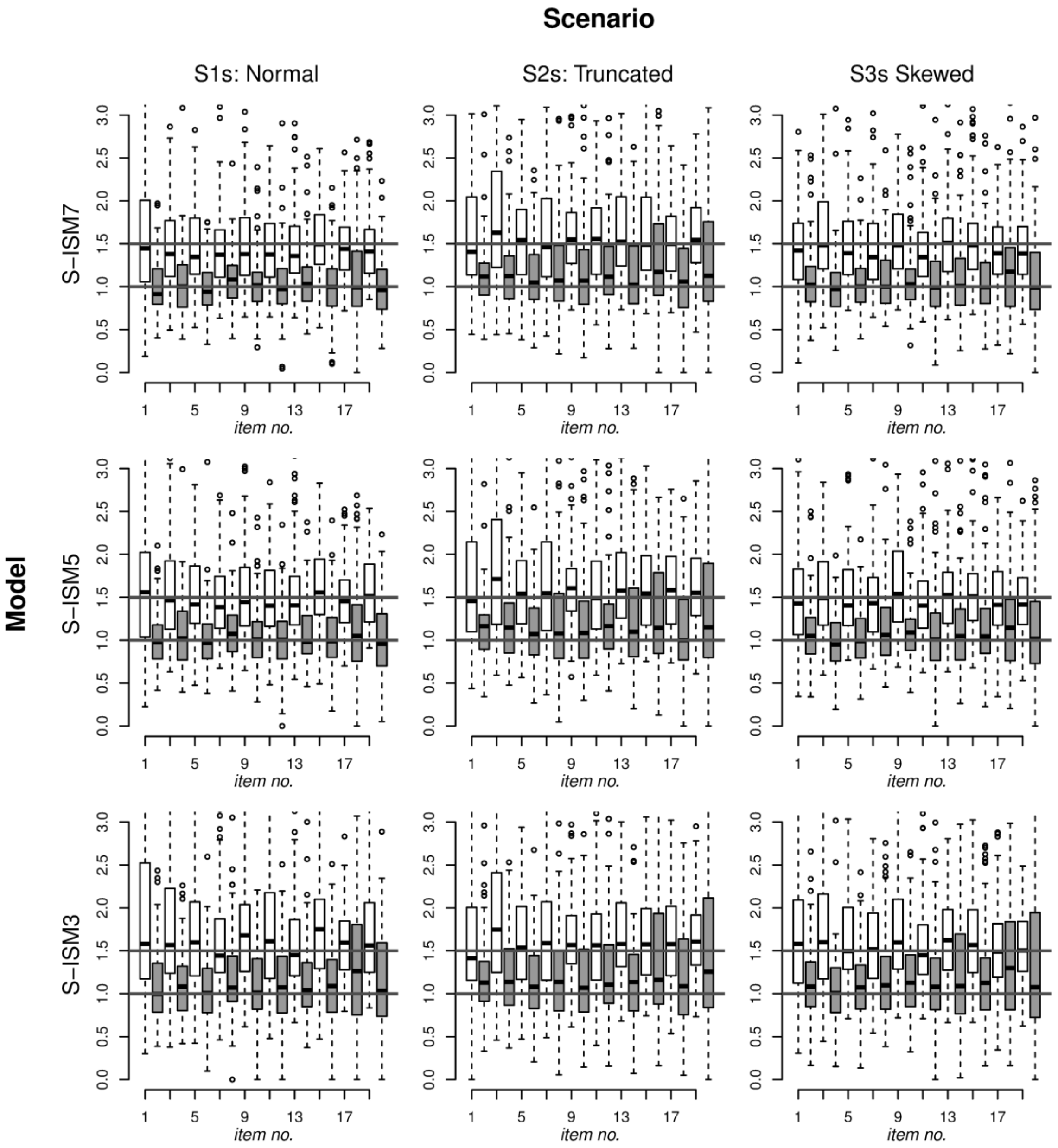
S2b: Truncated

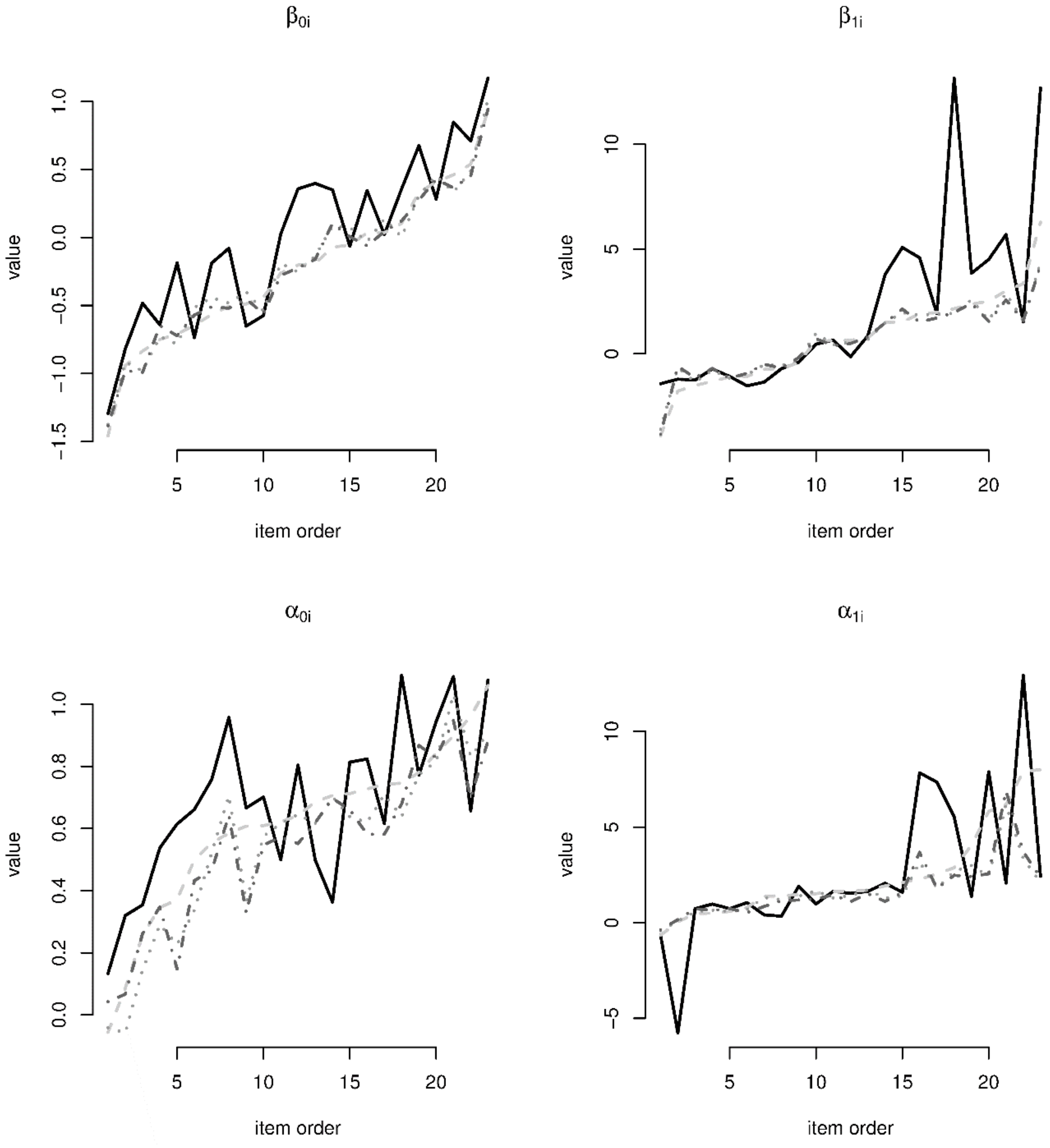
S3b: Skewed


Scenario
S1s: Normal
S2s: Truncated
S3s Skewed
Model
P–ISM
3
2
1
0
−1
−2
−3
1
5
9
13
17
item no.


Scenario
S1s: Normal
S2s: Truncated
S3s Skewed
Model
P−ISM
0.0
0.5
1.0
1.5
2.0
2.5
3.0
1
5
9
13
17
item no.


Scenario
S1s: Normal
S2s: Truncated
S3s Skewed
Model
S−ISM7
S−ISM5
S−ISM3
3
2
1
0
−1
−2
−3
1
5
9
13
17
item no.


Scenario
S1s: Normal
S2s: Truncated
S3s Skewed
Model
S−ISM7
S−ISM5
S−ISM3
0.0
0.5
1.0
1.5
2.0
2.5
3.0
1
5
9
13
17
item no.


β_{0i}
value
−1.5
−1.0
−0.5
0.0
0.5
1.0
5
10
15
20
item order
β_{1i}
value
0
5
10
5
10
15
20
item order
α_{0i}
value
0.0
0.2
0.4
0.6
0.8
1.0
5
10
15
20
item order
α_{1i}
value
−5
0
5
10
5
10
15
20
item order