A semi-parametric within-subject mixture approach to the analyses of responses and response times

Molenaar, D.; Bolsinova, M.; Vermunt, J.K.

DOI
10.1111/bmsp.12117

Publication date
2018

Document Version
Final published version

Published in
British Journal of Mathematical & Statistical Psychology

British Journal of Mathematical and Statistical Psychology (2018), 71, 205–228

The British Psychological Society

www.wileyonlinelibrary.com

A semi-parametric within-subject mixture approach to the analyses of responses and response times

Dylan Molenaar[1]*, Maria Bolsinova[1] and Jeroen K. Vermunt[2]

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

In item response theory, modelling 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 resorts to categorized response times. This approach is shown to hardly produce false positives and parameter bias. In addition, the semi-parametric approach results in approximately the same power as the parametric approach.

1. Introduction

The interest in response times in psychometrics dates back many decades (Thorndike, Bregman, Cobb, & Woodyard, 1926). Since then, effort has been devoted to the development of item response theory (IRT) models for responses and response times (e.g., Roskam, 1987; Thissen, 1983; see Schnipke & Scrams, 2002; Kyllonen & Zu, 2016; for a more comprehensive overview). Recently, work in this area has been boosted by the development of a general modelling framework for responses and response times (Van Der Linden, 2007, 2009a). 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. A 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; Ranger & Kuhn, 2012; Ranger & Ortner, 2012a, 2013; Wang, Chang, & Douglas, 2013; Wang, Fan, Chang, & Douglas, 2013), and models for personality data

*Correspondence should be addressed to Dylan Molenaar, Psychological Methods, Department of Psychology, University of Amsterdam, Postbus 15906, 1001 NK Amsterdam, The Netherlands (email: D.Molenaar@uva.nl).

DOI:10.1111/bmsp.12117

(Ferrando & Lorenzo-Seva, 2007a,b). Also, some of the earlier approaches (e.g., Roskam, 1987; Thissen, 1983) are special cases.

The main purpose of incorporating 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 (Marianti, Fox, Avetisyan, Veldkamp, & Tijmstra, 2014; Van Der Linden & Guo, 2008), guessing (Schnipke & Scrams, 1997), differences in the adopted solution strategy (Van Der Maas & Jansen, 2003), item pre-knowledge (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 instance, with respect to improving measurement precision, it has been shown within the general framework that the benefits of adding the response times are limited (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, Bolsinova, Rozsa, & De Boeck, 2016).

With respect to the latter, effort has been devoted to developing 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 & De Boeck, 2016; Molenaar *et al.*, 2016; Partchev & De Boeck, 2012; Wang & Xu, 2015; Wang, Xu, & Shang, 2016). In one class the item properties of the faster responses are modelled, and in the other class the item properties of the slower responses are modelled. 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 pre-knowledge on some of the items). However, the classes do not necessarily need to be substantively interpretable. They can also be seen as 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 Fox & Marianti, 2016), but 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 cut-off 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, Obserski, Vermunt, and De Boeck (2016) proposed an approach based on mixture modelling (see also Wang & Xu, 2015; Wang *et al.*, 2016). In this approach, the classes are operationalized by a two-component multivariate mixture distribution on the responses and response times. 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 modelling approach, the distribution of the response times within each class needs to be specified. Molenaar *et al.*, Wang and Xu, and Wang *et al.* 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 modelling framework is sensitive to violations of the assumed response time distribution. That is, if the response time distribution departs from the assumed distribution, then spurious classes may be detected if there are no classes underlying the data, and parameter estimates are biased if there are truly different classes in the data. The key to 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 to show that the problem outlined above can be remedied by adopting a semi-parametric within-subject mixture modelling approach. This is a practical and effective approach in which the distributional assumption on the response times is relaxed by categorizing the response times into an arbitrary number of classes. Next, a suitable within-subject mixture model is applied to the responses and categorized response times. We refer to this approach as 'semi-parametric' as the assumption on the response time distribution is less stringent than in the parametric (log-normal modelling) approach. In a simulation study we show that the semi-parametric approach rarely 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 scarcely affected in the semi-parametric approach as compared to the parametric approach.

The paper is organized as follows. In Section 2 we present the parametric within-subjects mixture model with log-normal response times within the classes. In Section 3 we show in a simulation study that this model is associated with false positives and parameter bias if the assumption of log-normal response times is violated. In Sections 4 and 5 we present the semi-parametric alternative and show on the same simulated data sets as above that this approach rarely suffers from false positives and parameter bias. In Section 6 we apply the parametric and semi-parametric approaches to a real data set pertaining to logical reasoning. Section 7 concludes with a general discussion.

2. 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.*, 2016; Wang & Xu, 2015; Wang *et al.*, 2016). 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 $\mathbf{c}_p = [C_{p1}, C_{p2}, \ldots, C_{pn}]$ where n denotes the number of items. The probability of observing response vector $\mathbf{x}_p = [X_{p1}, X_{p2}, \ldots, X_{pn}]$ is then given by

$$P(\mathbf{x}_p|\theta_p, \mathbf{c}_p) = \prod_{i=1}^{n} \omega(\Omega_{pi})^{X_{pi}} \omega(-\Omega_{pi})^{1-X_{pi}}, \tag{1}$$

with

$$\Omega_{pi} = [\alpha_{0i}(1 - C_{pi}) + \alpha_{1i}C_{pi}]\theta_p + \beta_{0i}(1 - C_{pi}) + \beta_{1i}C_{pi},$$

where θ_p is the latent ability, $\omega(.)$ 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, $\ln(\mathbf{t}_p) = [\ln(T_{p1}), \ln(T_{p2}), \ldots, \ln(T_{pn})]$, can be modelled using a conditional multivariate normal distribution with uncorrelated dimensions, that is,

$$f(\ln(\mathbf{t}_p)|\tau_p, \mathbf{c}_p) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2_{\varepsilon i}}} \exp\left\{ -\frac{1}{2} \frac{[\ln(T_{pi}) - E(ln(T_{pi})|\tau_p, C_{pi})]^2}{\sigma^2_{\varepsilon i}} \right\}, \tag{2}$$

with

$$E(\ln(T_{pi})|\tau_p, C_{pi}) = \nu_i - \delta C_{pi} - \tau_p, \qquad \delta > 0, \tag{3}$$

where τ_p is the latent speed, $\sigma^2_{\varepsilon i}$ 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 $C_{pi} = 1$ corresponds to the faster state (i.e., response times in this state are smaller).

In the model given by equations (1)–(3), it is assumed that the item effects are fixed and the subject effects are random (see Molenaar, Tuerlinckx, & van Der Maas, 2015; Ranger & Ortner, 2012b; Van Der Linden & Guo, 2008; Wang, Chang, *et al.*, 2013; Wang, Fan, *et al.*, 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^2_θ and σ^2_τ, and covariance $\sigma_{\theta\tau}$. For identification reasons, $\mu_\theta = \mu_\tau = 0$ and $\sigma^2_\theta = 1$. 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

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

Thus, it is assumed that the item states are independent and time homogeneous (i.e., the item states have equal state probabilities across items) with $P(C_{pi} = 1) = \pi$ for all i. It is possible to relax the independence assumption by introducing a time-homogeneous first-

order Markov structure on the item states (e.g., MacDonald & Zucchini, 1997; Vermunt, Langeheine, & Bockenholt, 1999). 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 free parameters in the parametric ISM include α_{0i}, α_{1i}, β_{0i}, β_{1i}, δ, ν_i, $\sigma^2_{\varepsilon i}$, σ^2_{τ}, $\sigma_{\theta\tau}$, and π. If the parameters are collected in model parameter vector $\boldsymbol{\eta}$, then the log marginal likelihood of response vector $\mathbf{x}_p$ and the log-response time vector $\ln(\mathbf{t}_p)$ for the parametric ISM is given by

$$\ell(\mathbf{x}_p, \ln(\mathbf{t}_p); \boldsymbol{\eta}) = \ln \int_{-\infty}^{\infty}\int \sum_{C_{p1}=0}^{1} \sum_{C_{p2}=0}^{1} \cdots \sum_{C_{pm}=0}^{1} P(\mathbf{x}_p|\theta_p, \mathbf{c}_p) f(\ln(\mathbf{t}_p)|\tau_p, \mathbf{c}_p) P(\mathbf{c}_p) g(\theta_p, \tau_p) \mathrm{d}\theta_p \mathrm{d}\tau_p, \tag{5}$$

where $P(\mathbf{x}_p|\theta_p, \mathbf{c}_p)$ is given by equation (1), $f(\ln(\mathbf{t}_p)|\tau_p, \mathbf{c}_p)$ is given by equation (2), $P(\mathbf{c}_p)$ is given by equation (4), and $g(.)$ is the bivariate normal density function.

2.1. Related models

The ISM as presented above is related to existing models. First, the approach by Partchev and De Boeck (2012) to separate within-subjects effects from between-subject effects in responses and response times can be seen as a special case of the ISM. Specifically, in Partchev and De Boeck, the class variables, C_{pi}, are treated as observed variables which are obtained from dichotomizing the observed response times. 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, the state size π depends on the cut-off point used to dichotomize the response times.

Second, the models by Wang and Xu (2015) and Wang *et al.* (2016) to separate solution behaviour, fast guessing, and cheating are related to the ISM. Specifically, fast guessing can be incorporated into the ISM by specifying $\alpha_{1i} = 0$ for the faster state ($C_{pi} = 1$). As a result, the distribution of $\mathbf{x}_p$ does not depend on θ_p, and β_{1i} reflects the logit-guessing probability. In Wang *et al.*, an additional procedure is proposed to detect cheating behaviour. Specifically, after separating fast guessing from regular solution behaviour using the model above (the first stage), cheating can be detected from the model residuals in the regular solution state 0 (the second stage). Such an approach is in principle equally amenable to the ISM.

2.2. Baseline model

To enable inferences about the relative goodness of fit of the ISM, a baseline model is needed. 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 *et al.*, 2015; Ranger & Ortner, 2012b). We will simply refer to this model as the *baseline model* (BM).

3. Simulation study A

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

3.1. Method

3.1.1. Scenarios

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

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 as $\nu_i = 2$ for all i and the residual response time variances are chosen as $\sigma^2_{\varepsilon i} = 0.13$ for all i. In addition, $\sigma^2_\tau = 0.13$ and $\sigma_{\theta\tau} = 0.144$. These parameter values result in a correlation between θ_p and τ_p equal to $\rho_{\theta\tau} = .4$, an R^2 of 0.50 in the log-response times, and untransformed response times between 1 and 50 s. See the top row in Figure 1 for a normal quantile–quantile plot and a histogram of the log-response times to an example item within this scenario.

S2b: a truncated BM. In this scenario, the data are generated using the same set-up as in S1b. However, instead of the normal distribution for the log-response times, a right-truncated normal distribution is used with truncation at ln(12) such that the untransformed response time distribution is right-truncated at 12 s. See the middle row in Figure 1 for a normal quantile–quantile 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 set-up 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 $\ln(T_{pi})' = (\lambda \ln(T_{pi}) + 1)^\lambda$, such that the transformed log-response times, $\ln(T_{pi})'$, are skewed. For the transformation parameter λ we use 0.3. See the bottom row in Figure 1 for a normal quantile–quantile plot and a histogram of the response times to an example item within this scenario.

In the remaining three scenarios (S1s, S2s, and S3s) 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):

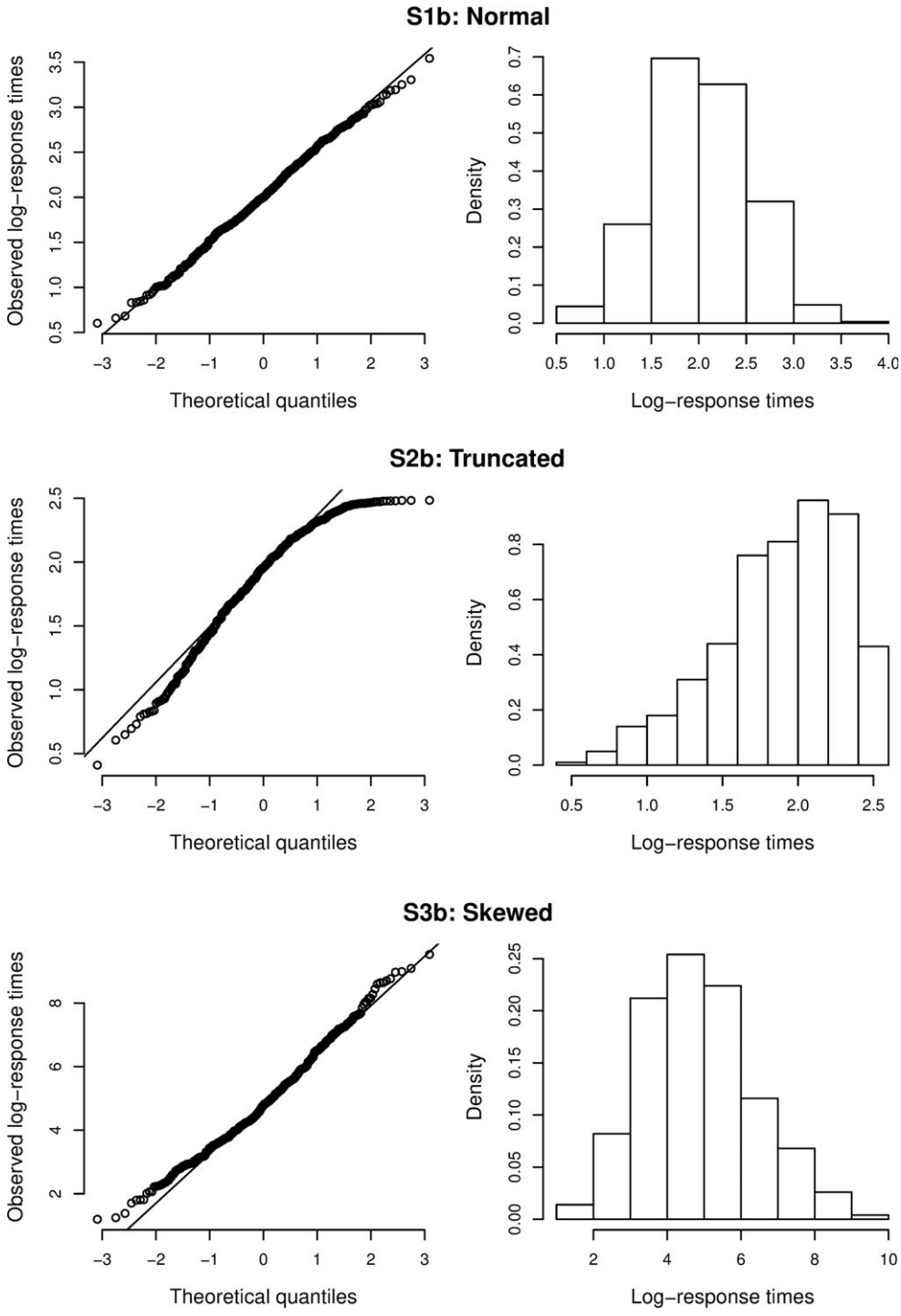

Figure 1. Normal quantile–quantile plots and histograms of the log-response time distribution for an example item within the baseline scenarios (S1b, S2b, and S3b).

S1s: a normal ISM. In this scenario, the data are generated using the ISM model given by equations (1)–(4). The true parameter values are chosen as follows. First, we chose $\delta = 0.5$ and $\pi = .5$. For the discrimination parameters, we used $\alpha_{0i} = 1.5$ and $\alpha_{1i} = 1.0$. 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 .11, which are reasonable. For instance, Molenaar *et al.* (2016) found residual correlations between .07 and .16 in the standardization data of the Hungarian WISC-IV block design test. The response time parameters ν_i, $\sigma^2_{\varepsilon i}$, σ^2_{τ}, and $\sigma_{\theta\tau}$ 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 set-up as in S1s. However, similarly to baseline scenario S2b, we use a truncated normal distribution for the log-response times with right-truncation at ln(12).
S3s: a skewed ISM. In this scenario, the data are generated using the same set-up as in S1s. However, similarly to baseline scenario S3b, the normal log-response times are transformed using a Box–Cox transformation, with the transformation parameter $\lambda = 0.3$.

3.1.2. Procedure

We conducted 100 replications of each scenario with 20 items and 500 subjects. For the data within each replication, the parametric ISM is fitted (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, 1974), 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). 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.

3.2. Results

3.2.1. False positive and true positive rates

Table 1 contains the false positive and true positive rates of the P-ISM in the different scenarios. 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 1, for the P-ISM, there are hardly any 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 accepted in the majority of the replications (false positives rates between 0.90 and 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.

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 scenarios without item states (S1b, S2b, and S3b)

	Data	BIC	AIC	AIC3	CAIC	saBIC
False positive rate	S1b: Normal baseline	.00	.03	.00	.00	.00
	S2b: Truncated baseline	.99	1.00	1.00	.90	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

Note. AIC = Akaike information criterion; BIC = Bayesian information criterion; CAIC = consistent AIC; saBIC = size adjusted BIC.

3.2.2. Parameter recovery

Table 2 gives the means and standard deviations of the estimates for the state size parameter, π, the response time difference between the states, δ, the variance of τ_p, σ^2_τ, and the correlation between speed and ability, $\rho_{\theta\tau}$, in the scenarios where the data truly contain different item states (S1s, S2s, S3s).[1] As can be seen from the table, if the within-state distribution of the log-response times is normal (S1s), parameters are adequately recovered, although the correlation between θ_p and τ_p is slightly overestimated. In the case of truncation (S2s) or skewness (S3s) in the distribution of the log-response times, all parameters are biased except for $\rho_{\theta\tau}$, the correlation between θ_p and τ_p.

Box plots of the parameter estimates in the P-ISM for the scenarios that include item states (S1s, S2s, and S3s) are shown in Figure 2 for the item easiness parameters, β_{0i} and β_{1i}, and in Figure 3 for the discrimination parameters, α_{0i} and α_{1i}. As expected, the parameters are acceptably recovered in the P-ISM if the data are generated according to the normal item states scenario (S1s; left plot in Figures 2 and 3). However, if the data are generated according to the truncated item states scenario (S2s; middle plot in Figures 2 and 3) or skewed item states scenario (S3s; right plot in Figures 2 and 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 seems small), but β_{0i} and α_{0i} are underestimated. In the case of skewness, β_{1i} is underestimated and β_{0i} is recovered acceptably. The parameters 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.

4. 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 *ISM*. The semi-parametric model differs from the model above in that the response times are categorized, that is, the

[1] We estimate the Cholesky decomposed covariance matrix of θ_p and τ_p. However, for ease of presentation we transformed these parameters into σ^2_τ and $\rho_{\theta\tau}$. In addition, we estimated logit(π) but we present the results for π.

Table 2. Means 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 parentheses

	π (0.50)		δ (0.50)		σ^2_τ (0.13)		$\rho_{\theta\tau}$ (0.40)	
Scenario	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
S1s: Normal	0.50	0.02	0.50	0.03	0.12	0.01	0.45	0.05
S2s: Truncated	0.32	0.01	0.56	0.01	0.07	0.01	0.41	0.05
S3s: Skewed	0.80	0.02	2.00	0.07	0.82	0.08	0.42	0.04

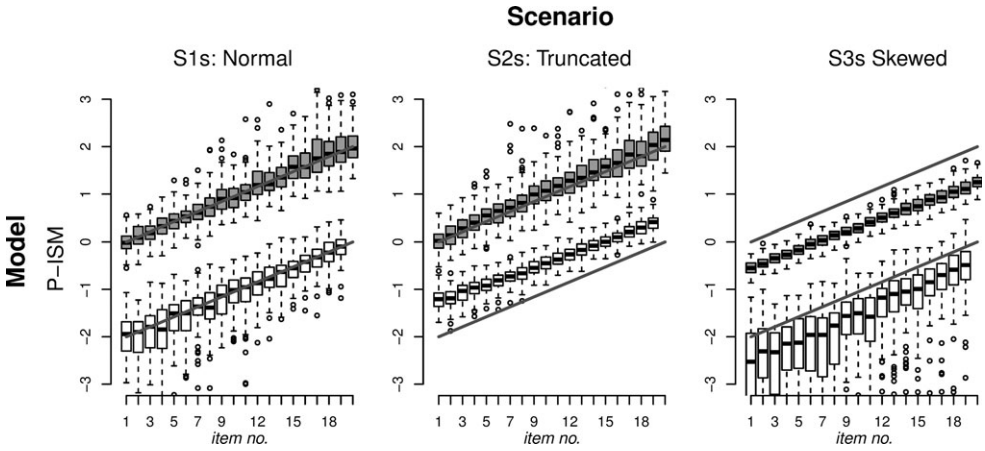

Figure 2. Box plots of the β_{0i} (white) and β_{1i} (grey) parameter estimates for the items in the parametric normal model (P-ISM) in the different scenarios that include item states (S1s, S2s, and S3s). The solid grey lines denote the true values of β_{0i} (lower grey line) and β_{1i} (upper grey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

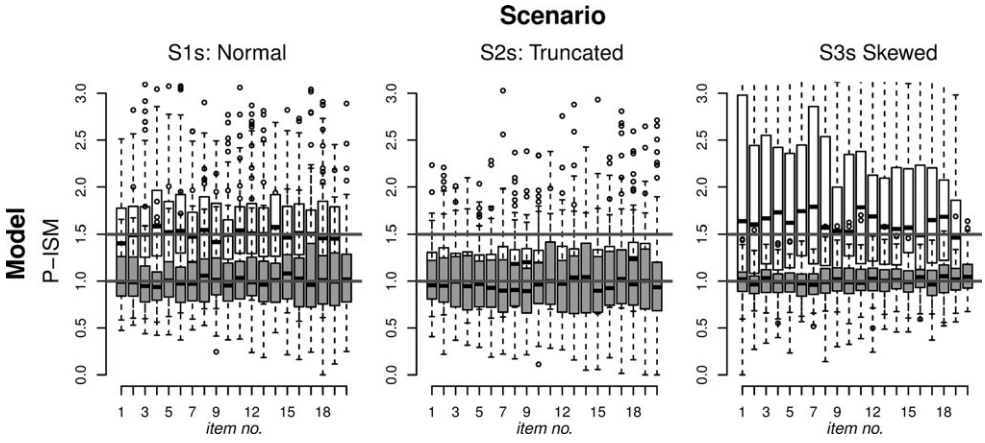

Figure 3. Box plots of the α_{0i} (white) and α_{1i} (grey) parameter estimates for the items in the parametric normal model (P-ISM) in the different scenarios that include item states (S1s, S2s, and S3s). The solid grey lines denote the true values of α_{0i} (upper grey line) and α_{1i} (lower grey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

categorized response times, T'_{pi}, are obtained from the possibly transformed response times as follows:

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

where b_{zi} are the thresholds at which the (transformed) response times are categorized, Z denotes the number of response time categories used, and $k(.)$ is the transformation function. Both b_{zi} and Z are chosen by the researcher. But as we illustrate in the real data application, multiple options can be considered to study the robustness of the results. We leave open the option to transform the response times by function $k(.)$ prior to categorization to include, among others, the possibility of categorizing the log-response times (i.e., $k(T_{pi}) = \ln(T_{pi})$), the reciprocal response times (i.e., $k(T_{pi}) = 1/T_{pi}$), or the raw response times ($k(T_{pi}) = T_{pi}$). We need this possibility later, to facilitate the demonstration that the model above is a generalization of the hierarchical model of Van Der Linden (2007). However, in practice, it does not matter whether the raw or the transformed response times are categorized (as long as $k(.)$ is a monotome function).

Next, within the semi-parametric ISM, the probability of the vector of categorized response times, $\mathbf{t}'_p = \left[T'_{p1}, T'_{p2}, \ldots, T'_{pn}\right]$, is subjected to a generalized linear IRT model with a suitable link function (see, e.g., Mellenbergh, 1994). Specifically, if dummy variable d_{piz} codes whether T'_{pi} is in category z (i.e., $d_{piz} = 1$) or not ($d_{piz} = 0$), the generalized linear IRT model for the categorized response times is given by

$$h[E(d_{piz}|\tau_p, \mathbf{c}_p)] = \gamma_{zi} - \delta C_{pi} - \varphi_i \tau_p, \delta > 0, \tag{7}$$

where the γ_{zi} are response time category parameters for category z of the response times of item i. In this generalized linear model for the categorized response times, a slope parameter, φ_i, is added. This is necessary as differences in the residual variances, $\sigma^2_{\varepsilon i}$, across items will be absorbed in this parameter and in the response time category parameters, γ_{zi}. Omitting the item-specific slope parameter results in misfit if $\sigma^2_{\varepsilon i}$ differs across items. If $\sigma^2_{\varepsilon i}$ is equal across items, the effect of $\sigma^2_{\varepsilon i}$ will be absorbed in σ^2_τ. However, this is unlikely in practice. Due to the extra slope parameters φ_i, the scale of τ_p needs to be identified. This can be done either by fixing σ^2_τ or by fixing φ_i for some i. All other identification constraints are similar to the parametric case.

In the model for categorized response times in equation (7), $h(.)$ is the link function. Although initial simulations (not presented) showed that the choice for $h(.)$ hardly affects results, there are conceptual differences between the models that arise for different forms of $h(.)$.

Cumulative categories model. If $h(.)$ is chosen to be the cumulative probit of category z, that is, $h\left[E(d_{piz}|\tau_p, \boldsymbol{c}_p)\right] = \Phi^{-1}\left[\sum_{a=z}^{z-1} E(d_{pia}|\tau_p, \mathbf{c}_p)\right]$, a cumulative categories model arises for the categorized response times, from which it follows that

$$P(\mathbf{t}'_p|\tau_p, \mathbf{c}_p) = \prod_{i=1}^{n}\left[\Phi(\gamma_{(T'_{pi}+1)i} - \delta C_{pi} - \varphi_i\tau_p) - \Phi(\gamma_{(T'_{pi})i} - \delta C_{pi} - \varphi_i\tau_p)\right], \tag{8}$$

where $\gamma_{(T'_{pi}+1)i}$ is the response time category parameter, γ_{zi}, for category $z = T'_{pi} + 1$, and similarly, $\gamma_{(T'_{pi})i}$ is the response time category parameter for $z = T'_{pi}$. For numerical reasons,

an approximation using the cumulative logit function can also be considered. The model in equation (8) is equivalent to a graded response model (Samejima, 1969). If this model is adopted for $\mathbf{t}'_p$, the full model given by equations (1) and (8) is a generalization of the hierarchical model of Van Der Linden (2007) for categorized response times with $k(T_{pi}) = \ln(T_{pi})$. That is, if the continuous response times T_{pi} are log-normally distributed, the probability that a log-response time, $\ln(T_{pi})$, falls into the interval $(b_{zi}, b_{(z+1)i})$ is given by the graded response model in equation (8). That is, for normal $\ln(T_{pi})$, the response time category parameters γ_{zi} are a function of the categorization thresholds b_{zi}, the residual variances $\sigma^2_{\varepsilon i}$, and the variance of the speed factor σ^2_τ. Thus, the approach above assumes that a normal distribution underlies the categorized response times. Departures from normality in $\ln(T_{pi})$ will be captured by the response time category parameters γ_{zi} and not result in spurious item states, as we will show in the simulation study below.

Adjacent categories model. If $h(.)$ is chosen to be the adjacent categories logit, that is, $h[E(d_{piz}|\ \tau_p,\ \mathbf{c}_p)] = \ln[E(d_{piz}|\tau_p,\ \mathbf{c}_p)/E(d_{pi(z-1)}|\tau_p,\ \mathbf{c}_p)]$, an adjacent categories model arises for the categorized response times, from which it follows that

$$P(\mathbf{t}'_p|\tau_p, \mathbf{c}_p) = \prod_{i=1}^{n} \frac{\exp\left(\sum_{z=0}^{T'_{pi}} \gamma_{zi} - \delta C_{pi} - \varphi_i \tau_p\right)}{\sum_{j=0}^{Z-1} \exp\left(\sum_{z=0}^{j} \gamma_{zi} - \delta C_{pi} - \varphi_i \tau_p\right)}, \tag{9}$$

where the category parameter γ_{0i} may be chosen in such a way that

$$\sum_{z=0}^{Z-1} -\delta - \varphi_i \tau_p + \gamma_{zi} = 0. \tag{10}$$

This model is equivalent to the partial credit model (Masters, 1982). Contrary to the cumulative probit model above, there is not an obvious response time model that will generate equation (9). In that sense, choosing the partial credit model for the categorized response times is a pragmatic choice.

Equation (7) with an appropriate choice for $h(.)$, together with the model for the responses in equation (1) and the bivariate normal distribution for θ_p and τ_p, constitutes the full model. The free parameters in the semi-parametric ISM include α_{0i}, α_{1i}, β_{0i}, β_{1i}, γ_{zi}, φ_i, δ, σ^2_τ, $\sigma_{\theta\tau}$, 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 $\mathbf{x}_p$ and the categorized response time vector $\mathbf{t}'_p$ for the semi-parametric ISM is given by

$$\ell(\mathbf{x}_p, \mathbf{t}'_p; \boldsymbol{\zeta}) = \ln \iint_{-\infty}^{\infty} \sum_{C_{p1}=0}^{1} \sum_{C_{p2}=0}^{1} \cdots \sum_{C_{pn}=0}^{1} P(\mathbf{x}_p|\theta_p, \mathbf{c}_p) P(\mathbf{t}'_p|\tau_p, \mathbf{c}_p) P(\mathbf{c}_p) g(\theta_p, \tau_p) \mathrm{d}\theta \mathrm{d}\tau, \tag{11}$$

where $P(\mathbf{x}_p|\theta_p, \mathbf{c}_p)$ is given by equation (1), $P(\mathbf{c}_p)$ by equation (4), and $P(\mathbf{t}'_p|\tau_p, \mathbf{c}_p)$ depends on the choice for $h(.)$ in equation (7) (e.g., equation (8) in the case of a cumulative probit function and equation (9) in the case of an adjacent categories logit).

4.1. Baseline model

For the semi-parametric ISM, the baseline model can be derived in a similar way to the parametric normal model above. The resulting model is a latent variable model with a two-parameter model for the responses and a model for the categorized response times and correlated random subject effects.

5. Simulation study B

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

5.1. Method

We used the same 100 replications of the six scenarios as in simulation study A. To these data, we fitted 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, respectively). In addition, we fitted the corresponding baseline models (S-BM7, S-BM5, and S-BM3). In these models for responses and categorized response times, we identified the scale of τ_p by fixing φ_i to 1 for item 1.

As regards categorization, we chose to categorize the raw response times (i.e., $k(T_{pi}) = T_{pi}$), therefore in equation (6), b_{0i} and b_{Zi} are 0 and ∞ by definition. The remaining thresholds, $b_{1i}, b_{2i}, \ldots, b_{(Z-1)i}$ 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 quantile approach to categorizing the response times as desirable because it results in thresholds that 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 because the resulting categorization will be equivalent (but the thresholds will be different, i.e., the thresholds obtained with the percentile method for the log-response times are the log-transformed thresholds that will be obtained on the raw response times).

For each data set, the fit of the three item state models (S-ISM7, S-ISM5, S-ISM3) is compared to its corresponding baseline model (S-BM7, S-BM5, S-BM3). We used the cumulative categories model in equation (8) for the categorized response times. 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 studies. Syntax to fit the semi-parametric model is available in the Appendix.

5.2. Results

5.2.1. False positives

In Table 3, the false positive rates are depicted for the ISMs (S-ISM7, S-ISM5, S-ISM3) in the scenarios in which the data do not contain item states (S1b, S2b, S3b). As can be seen from the table, the semi-parametric models hardly suffer from false positives, with false positive rates close to 0 for all fit indices except the AIC. The AIC fit index is associated with unacceptable false positive rates for the semi-parametric model with $Z = 7$ and $Z = 5$ (rates between .22 and .70). For $Z = 3$, the false positive rates for the AIC seem acceptable, with rates between .02 and .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 scenarios 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	.00	**.70**	**.02**	.00	**.01**
	S2b: Truncated baseline	.00	**.38**	.00	.00	.00
	S3b: Skewed baseline	.00	**.63**	.00	.00	.00
S-ISM5: Semi-par. item states with $Z = 5$	S1b: Normal baseline	.00	**.33**	.00	.00	.00
	S2b: Truncated baseline	.00	**.29**	**.01**	.00	**.01**
	S3b: Skewed baseline	.00	**.22**	.00	.00	.00
S-ISM3: Semi-par. item states with $Z = 3$	S1b: Normal baseline	.00	**.05**	.00	.00	.00
	S2b: Truncated baseline	.00	**.03**	.00	.00	.00
	S3b: Skewed baseline	.00	**.02**	.00	.00	.00

Notes. Non-zero rates are in bold.
AIC = Akaike information criterion; BIC = Bayesian information criterion; CAIC = consistent AIC; saBIC = size adjusted BIC.

5.2.2. True positives

All true positives rates are equal to 1.00 for all scenarios. This indicates that in all replications, the item states in the data have been successfully detected by the semi-parametric ISM irrespective of the distribution of the raw response times.

5.2.3. Parameter recovery

Table 4 gives the means and standard deviations of the estimates for the state size parameter, π, the response time difference between the states, δ, the variance of τ_p, σ^2_τ, and the correlation between speed and ability, $\rho_{\theta\tau}$, in the scenarios where the data truly contain different item states (S1s, S2s, S3s). As can be seen from the table, π is recovered adequately in all scenarios. The correlation parameter seems slightly overestimated. However, the overestimation is also evident in the normal scenario (S1s) and is thus not related to the truncation or skewness in the log-response times. The mean estimates of σ^2_τ are not close to the true parameter value. However, this is not surprising as σ^2_τ is

Table 4. Means 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 parentheses

		π (0.50)		δ (0.50)		σ^2_τ (0.13)		$\rho_{\theta\tau}$ (0.40)	
Model	Scenario	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
S-ISM7	S1s: Normal	0.50	0.03	1.09	0.06	2.95	0.51	0.44	0.05
	S2s: Truncated	0.53	0.04	1.01	0.05	2.20	0.41	0.43	0.04
	S3s: Skewed	0.50	0.02	1.08	0.05	3.13	0.54	0.46	0.05
S-ISM5	S1s: Normal	0.50	0.03	3.62	0.26	2.92	0.56	0.43	0.05
	S2s: Truncated	0.53	0.04	3.44	0.31	2.27	0.46	0.42	0.04
	S3s: Skewed	0.50	0.03	3.59	0.26	3.06	0.59	0.45	0.05
S-ISM3	S1s: Normal	0.49	0.04	2.39	0.31	2.89	0.68	0.42	0.05
	S2s: Truncated	0.52	0.04	2.38	0.38	2.40	0.55	0.42	0.05
	S3s: Skewed	0.50	0.03	2.40	0.32	3.11	0.72	0.43	0.05

dependent upon our identification constraint $\varphi_i = 1$ for item 1 (see above). The correlation $\rho_{\theta\tau}$, which is calculated from σ^2_τ, is not affected by the scaling. In addition, from the table it appears that δ depends on the number of response time categories that are used. This is due to the scale of δ being not the same as the scale of τ_p (see equation (7)), that is, the scale of δ depends on the scale of γ_{zi}, which is in turn dependent on the number of response time categories.

Box plots of the parameter estimates of the 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 shown in Figure 4 for the item easiness parameters, β_{0i} and β_{1i}, and in Figure 5 for the discrimination parameters, α_{0i} and α_{1i}. Note again that these models have been fitted to the same simulated data sets as used for the parametric model in Figures 2 and 3. To provide a reference for the results in Figures 4 and 5, see Figure 6 for box plots of the easiness and discrimination parameter estimates based on the response data only for scenario S1s. Note that for the other scenarios these plots will look the same because the scenarios differ only in the response time data but not in the response data. As can be seen from Figure 4 for the easiness parameters and Figure 5 for the discrimination parameters, the ISM estimates tend to be unbiased for all semi-parametric models and all scenarios. The standard errors are slightly smaller for the $Z = 5$ and $Z = 7$ models than for the $Z = 3$ model, which is due to the larger variance in the categorized response times for more response time categories.

5.2.4. Overall conclusion

As appears from the results of simulation study A and B, if the log-response time distribution departs from normality but a normal ISM is applied nevertheless, spurious item 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 ISM is still able to detect these, but parameter estimates are biased. The proposed class of semi-parametric models with $Z = 7$, $Z = 5$, and $Z = 3$ was shown to not suffer from the problem of spurious states (except for the AIC) or bias in the parameter estimates, while the power to detect different item states in the data is hardly affected. As the standard errors were found to be smaller for $Z = 5$ and $Z = 7$, it is generally advisable to consider at least five response time categories if the shape of the response time distribution and the sample size allow this.

5.3. Discussion

In the simulation study, we did not manipulate the effect size of the item states in the data. We chose a relatively optimal setting (equal state sizes and differences between the states in terms of β_{0i}, β_{1i}, α_{0i}, α_{1i}, and δ that were not too small) to be able to demonstrate what the potential problem is (spurious latent states in the case of departures from normality in the transformed response times) and to facilitate demonstration of the feasibility of our solution (categorizing the response times). It should, however, be noted that in practice, similarly as in more traditional mixture models, the power to detect different item states in the data will depend on the state size π (with smaller power for unequal state sizes due to larger standard errors in the smaller state), the size of the differences between the states (i.e., β_{0i}, β_{1i}, α_{0i}, α_{1i}, and δ), and the number of subjects and items.

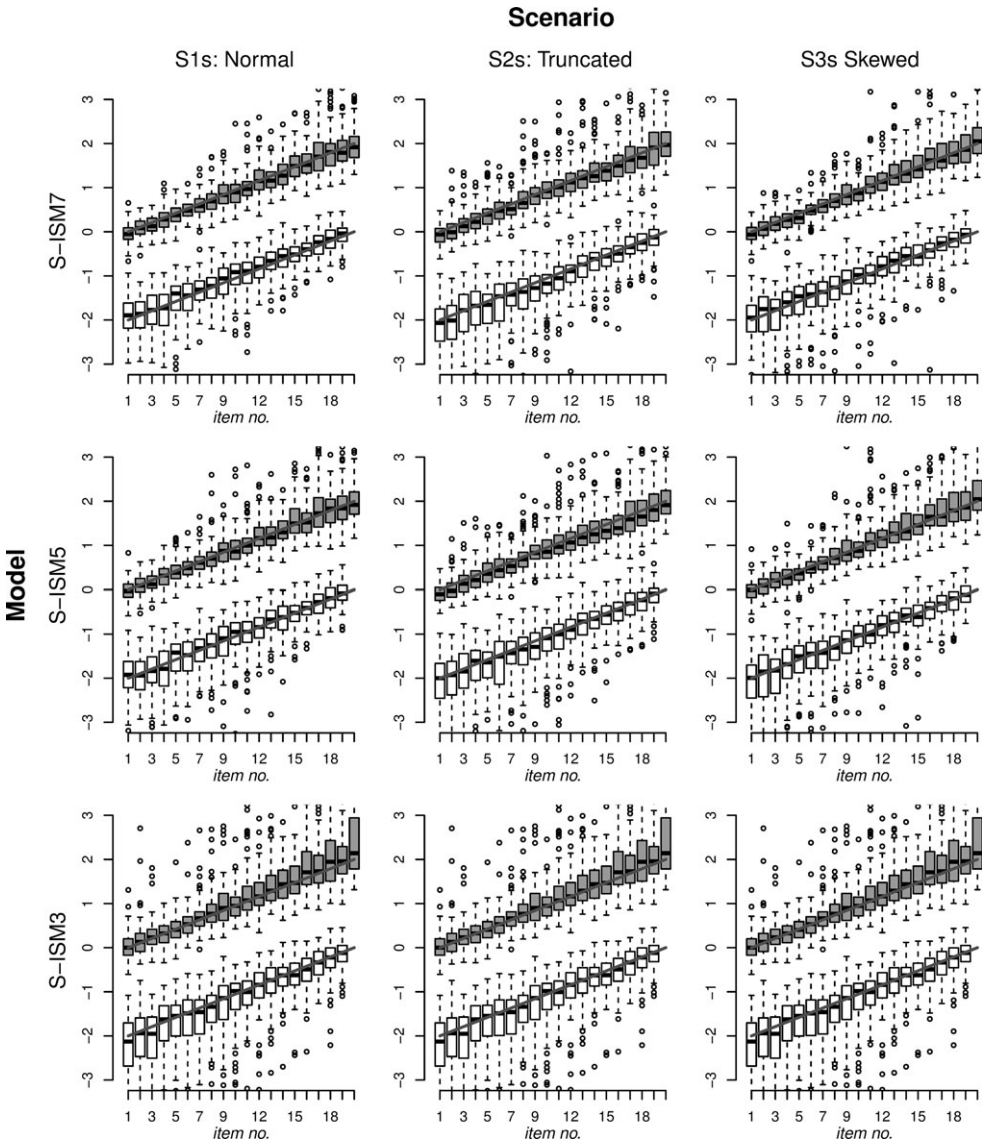

Figure 4. Box plots of the β_{0i} (white) and β_{1i} (grey) parameter estimates of the items in the different semi-parametric models (S-ISM7, S-ISM5, and S-ISM3) in the different scenarios that include item states (S1s, S2s, and S3s). The solid grey lines denote the true values of β_{0i} (lower grey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

6. Illustration

6.1. 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 five optional elements would complete the pattern. The items are administered using a 40 s deadline.

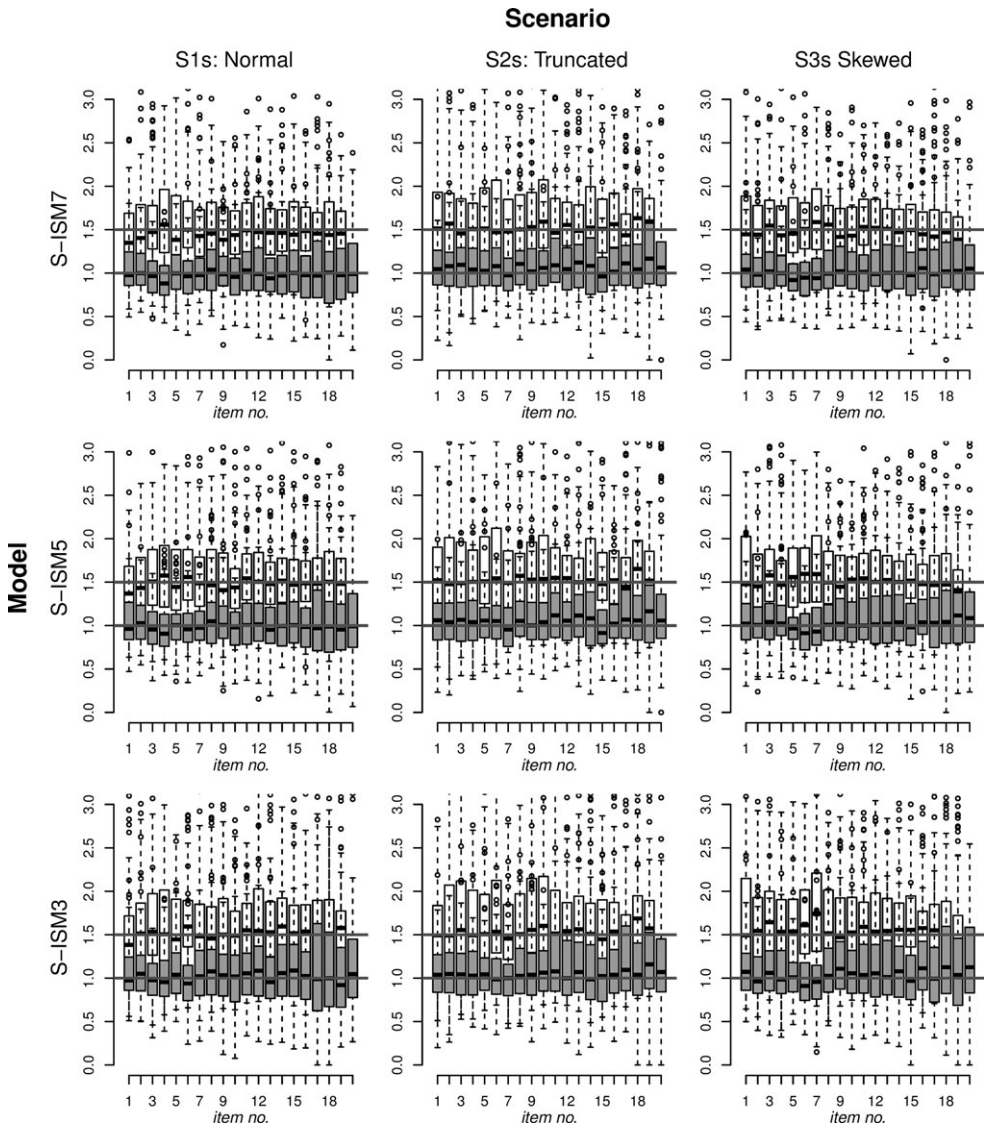

Figure 5. Box plots of the α_{0i} (white) and α_{1i} (grey) parameter estimates for the items in the different semi-parametric models (S-ISM7, S-ISM5, and S-ISM3) in the different scenarios that include item states (S1s, S2s, and S3s). The solid grey lines denote the true values of α_{0i} (upper grey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

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. Thirty-six respondents are omitted from the analysis because they showed suspiciously short response times (1 s or faster), resulting in a sample size of 628 respondents.

To the data we fitted the same parametric and semi-parametric baseline and ISMs as considered in the simulation studies. 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 are conducted using the same procedure as

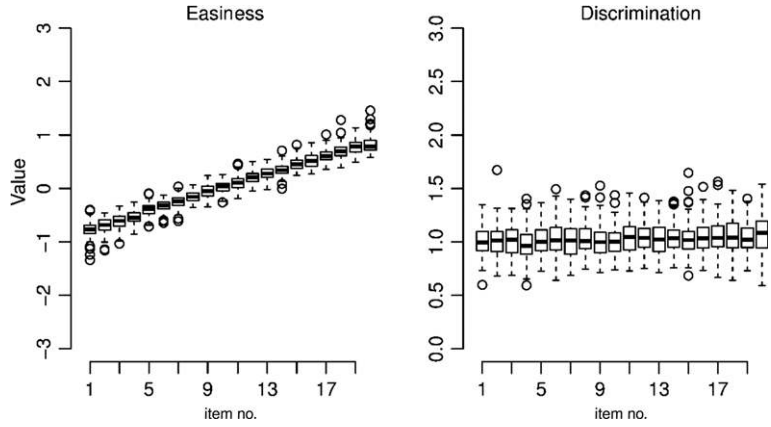

Figure 6. Box plots of the easiness and discrimination parameter estimates of the response data only for the S1s scenario.

outlined in the simulation studies. For the categorized response times, we use the adjacent categories model from equation (9).

6.2. Results

Table 5 gives the model fit indices of the different models. As can be seen, for all semi-parametric and parametric approaches, the ISM is the better-fitting model according to the indices considered. We therefore accept the ISM model and look into the parameter estimates within this model for the semi-parametric and parametric approach.

Table 6 contains the parameters estimates of the state size parameter, π, the response time difference between the states, δ, the variance of τ_p, σ^2_τ, and the correlation between speed and ability, $\rho_{\theta\tau}$, in the ISM models. As can be seen, in the parametric model (P-ISM), the estimate of the faster state size, π, is substantially smaller than in the semi-parametric models (S-ISM), 0.16 versus 0.38–0.43. In addition, the estimate of π is relatively stable across the semi-parametric models. Similarly to what was shown in the simulation studies,

Table 5. Model fit indices for the different parametric and semi-parametric models in the illustration

	Z	Model	BIC	AIC	AIC3	CAIC	saBIC
Parametric	–	P-ISM	**34,752**	**34,122**	**34,264**	**34,894**	**34,302**
		P-BM	35,493	35,075	35,169	35,587	35,194
Semi-parametric	7	S-ISM7	**67,983**	**66,845**	**67,101**	**68,239**	**67,170**
		S-BM7	68,493	67,667	67,853	68,679	67,903
	5	S-ISM5	**58,518**	**57,585**	**57,795**	**58,728**	**57,852**
		S-BM5	58,932	58,310	58,450	59,072	58,487
	3	S-ISM3	**44,744**	**44,016**	**44,180**	**44,908**	**44,224**
		S-BM3	44,959	44,541	44,635	45,053	44,660

Note. For each pair of ISM and BM models, the smaller fit index is in bold.
AIC = Akaike information criterion; BIC = Bayesian information criterion; CAIC = consistent AIC; P-BM = parametric baseline model; P-ISM = parametric ISM; saBIC = size adjusted BIC.

Table 6. 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^2_τ, and the correlation between speed and ability, $\rho_{\theta\tau}$

	π		δ		σ^2_τ		$\rho_{\theta\tau}$	
Model	Est.	*SE*	est.	*SE*	est.	*SE*	est.	*SE*
P-ISM	0.16	0.01	_0.74	0.01	0.13	0.01	−.52	0.02
S-ISM7	0.43	0.02	_1.20	0.07	0.60	0.09	−.53	0.05
S-ISM5	0.43	0.03	_1.48	0.11	0.91	0.15	−.60	0.07
S-ISM3	0.38	0.03	_2.30	0.18	2.15	0.39	−.54	0.07

the estimate of the response time difference, δ, fluctuates between the semi-parametric models due to scale differences in γ_{zi}. In addition, the correlation between θ_p and τ_p (i.e., $\rho_{\theta\tau}$, which we calculated from the estimates of $\sigma_{\theta\tau}$ and σ^2_τ) is stable across the semi-parametric models and does not differ significantly between the parametric and semi-parametric approaches.

In Figure 7 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

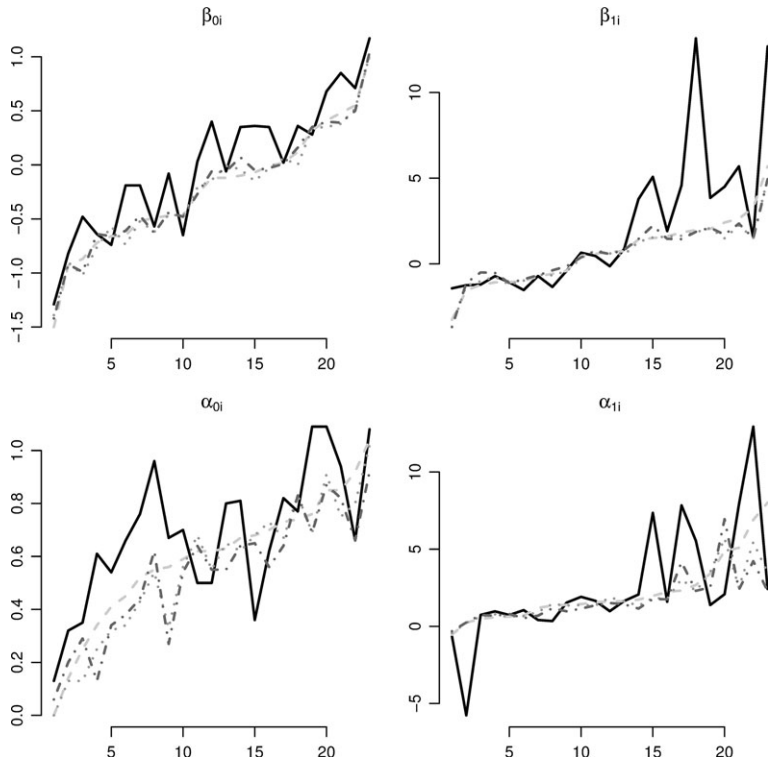

Figure 7. Plots of the β_{0i}, β_{1i}, α_{0i}, and α_{1i} parameter estimates for the normal item states model (P-ISM, solid black line) and the semi-parametric item states model (S-ISM7, S-ISM5, and S-ISM3; dotted dashed, dotted, and dashed lines respectively). In each plot, the items are ordered on basis of the estimates in S-ISM3 for clarity.

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 studies.

To conclude, results seem to be stable between the semi-parametric approaches. That is, the exact number of response time categories does not significantly affect the results. There are, however, notable differences between the semi-parametric approach and the parametric approach in the state size parameter, π, and the item parameters. Nevertheless, as we know from the simulation studies 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.

7. Discussion

In the simulation studies we established that the parametric ISM 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.

Categorization of continuous variables generally 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, however, it can be desirable to categorize the response times such that violations of the assumed distribution do not affect the results. In addition, we showed that the power is hardly affected. However, a disadvantage of the categorization adopted in the present approach is that the number and location of the categorization thresholds are arbitrary. In the simulation studies, it was shown that for the configurations of the parameters we used, the number of response time categories hardly influenced the results in terms of power or parameter recovery. However, in practice, it is still advisable to fit the semi-parametric approach using different numbers of response time categories to investigate the stability of the results. If the results are stable, one can choose a definite number of categories by considering some criterion (e.g., the standard errors).

The model as presented in this paper can be seen as a semi-parametric alternative to the ISM presented by Molenaar *et al.* (2016). Because in Molenaar *et al*. the variance of the response times is assumed to be equal across states, we retained this assumption in the present semi-parametric model. It could, however, be argued that for some response processes there are important differences in the variance of the response times. For instance, fast guessing is commonly associated with less variance than the regular response process. In the present model, it is straightforward to allow for such differences by estimating the response time category parameters separately in each group. Other extensions of the present approach include the use of the mid-points within each response time category. By doing so, the categorized distribution resembles the observed response time distribution better than in the case of percentiles (for which the distribution is uniform).

We adopted a categorized response time model as it is a relatively easy and effective method. However, we note that other semi-parametric possibilities exist, including the proportional hazards model (Kang, 2017; Loeys *et al.*, 2014; Ranger &

Ortner, 2012b, 2013; Wang, Fan, *et al.*, 2013) and the linear transformation model (Wang, Chang, *et al.*, 2013). An advantage of adopting these models over our model is that they do not rely on arbitrary decision about the number of thresholds Z and the position of the thresholds b_{zi}. However, although feasible, these models are relatively challenging to estimate even for a baseline model (without mixtures). For a discussion on these challenges, see, for example, Kang (2017) for the proportional hazards model and Wang, Chang, *et al.* (2013) for the linear transformation model. Thus, we do not rely here on the proportional hazards model or the linear transformation model as it is less straightforward to extend these approaches to include item-specific latent class variables. The main advantage is that the present approach of categorized response times remains in the framework of generalized linear modelling which is relatively well understood and the models are relatively well estimable. However, we acknowledge that the semi-parametric approaches discussed above are also amenable to the present undertaking in principle.

Acknowledgements

The research by Dylan Molenaar was made possible by a grant from the Netherlands Organization for Scientific Research (NWO VENI- 451-15-008).

References

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

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, *26*, 211–252.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*, 345–370. https://doi.org/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 O. Opitz, B. Lausen & R. Klar (Eds.), *Studies in classification, data analysis, and knowledge organization* (pp. 40–54). Berlin, Germany: 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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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*, 170–173. https://doi.org/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. https://doi.org/10.3758/s13428-015-0631-y

Kang, H. A. (2017). Penalized partial likelihood inference of proportional hazards latent trait models. *British Journal of Mathematical and Statistical Psychology*, *70*, 187–208. https://doi.org/10.1111/bmsp.12080

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. https://doi.org/10.1007/s11336-008-9075-y

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

Kyllonen, P. C., & Zu, J. (2016). Use of response time for measuring cognitive ability. *Journal of Intelligence*, *4*(4), 14.

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. https://doi.org/10.1111/bmsp.12020

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. https://doi.org/10.1037//1082-989X.7.1.19

MacDonald, I. L., & Zucchini, W. (1997). *Hidden Markov and other models for discrete-valued time series*. London, UK: Chapman & Hall.

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*, 426–451. https://doi.org/10.3102/1076998614559412

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

Maxwell, S. E., & Delaney, H. D. (1993). Bivariate median splits and spurious statistical significance. *Psychological Bulletin*, *113*, 181–190. https://doi.org/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*, 121–137. https://doi.org/10.1177/0146621602250534

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

Molenaar, D. (2015). The value of response times in item response modeling. *Measurement: Interdisciplinary Research and Perspectives*, *13*, 177–181.

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. https://doi.org/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*, 197–219. https://doi.org/10.1111/bmsp.12042

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

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

Ranger, J. (2013). A note on the hierarchical model for responses and response times in tests of van Der Linden (2007). *Psychometrika*, *78*, 538–544. https://doi.org/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. https://doi.org/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. https://doi.org/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*, 128–148.

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*, 334–349. https://doi.org/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*, 503–533. https://doi.org/10.1080/00273171.2013.796280

Raven, J. C. (1962). *Advanced progressive matrices. Set II*. London, UK: H. K. Lewis & Co.

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, the Netherlands: North-Holland.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometric Monograph Supplement No. 17). Richmond, VA: Psychometric Society.

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. https://doi.org/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. https://doi.org/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, NY: 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. https://doi.org/10.1007/s11336-006-1478-z

Van Der Linden, W. J. (2009a). Conceptual issues in response-time modeling. *Journal of Educational Measurement*, *46*, 247–272. https://doi.org/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. https://doi.org/10.1177/0146621607314042

Van Der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, *73*, 365–384. https://doi.org/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*, 327–347. https://doi.org/10.1177/0146621609349800

Van Der Maas, H. L., & Jansen, B. R. (2003). What response times tell of children's behavior on the balance scale task. *Journal of Experimental Child Psychology*, *85*, 141–177. https://doi.org/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*, 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. https://doi.org/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*, 381–417. https://doi.org/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. https://doi.org/10.1111/bmsp.12054

Wang, C., Xu, G., & Shang, Z. (2016). A two-stage approach to differentiating normal and aberrant behavior in computer based testing. *Psychometrika*, *83*, 223–254. https://doi.org/10.1007/s11336-016-9525-x

Received 11 October 2016; revised version received 28 June 2017

Appendix: Syntax to fit the semi-parametric ISM using LatentGOLD

```
model

options
 maxthreads=all;
 algorithm
  tolerance=1e-008 emtolerance=0.01 emiterations=250 nriterations=50 ;
 startvalues
  seed=0 sets=16 tolerance=1e-005 iterations=50;
 bayes
  categorical=1 variances=1 latent=1 poisson=1;
 montecarlo
  seed=0 sets=0 replicates=500 tolerance=1e-008;
 quadrature nodes=10;
 missing  excludeall;
 output
  parameters=first standarderrors estimatedvalues=model;

 variables
  caseid ID;
  dependent X, kT cumlogit;
  independent item nominal;
  latent
   Ability  continuous,
   Speed  continuous,
   Cluster nominal 2 dynamic;
 equations
   (1) Ability;
   Speed;
   Ability <-> Speed;
   Cluster <- 1;
   X <- 1 | Item Cluster + (+) Ability | Item Cluster;
   kT <- 1 | item + (-) Cluster + (aa) Speed;
   aa[1,1]=-1;
 end model
```