Hidden Markov Item Response Theory Models for Responses and Response Times

Molenaar, D.; Oberski, D.; Vermunt, J.; de Boeck, P.

DOI
10.1080/00273171.2016.1192983

Publication date
2016

Document Version
Final published version

Published in
Multivariate Behavioral Research

MULTIVARIATE BEHAVIORAL RESEARCH
2016, VOL. 51, NO. 5, 606–626
http://dx.doi.org/10.1080/00273171.2016.1192983

Routledge
Taylor & Francis Group

Hidden Markov Item Response Theory Models for Responses and Response Times

Dylan Molenaar[a], Daniel Oberski[b], Jeroen Vermunt[b], and Paul De Boeck[c]

[a]University of Amsterdam; [b]Tilburg University; [c]Ohio State University

ABSTRACT

Current approaches to model responses and response times to psychometric tests solely focus on between-subject differences in speed and ability. Within subjects, speed and ability are assumed to be constants. Violations of this assumption are generally absorbed in the residual of the model. As a result, within-subject departures from the between-subject speed and ability level remain undetected. These departures may be of interest to the researcher as they reflect differences in the response processes adopted on the items of a test. In this article, we propose a dynamic approach for responses and response times based on hidden Markov modeling to account for within-subject differences in responses and response times. A simulation study is conducted to demonstrate acceptable parameter recovery and acceptable performance of various fit indices in distinguishing between different models. In addition, both a confirmatory and an exploratory application are presented to demonstrate the practical value of the modeling approach.

KEYWORDS

Conditional independence; dynamic modeling; hidden Markov modeling; item response theory; latent class models; response time modeling

Inferences about individual differences in psychological abilities have traditionally been based on latent variables that are operationalized using measurement models for the responses to test items. Popular measurement models include for example the Rasch model (Rasch, 1960), the two-parameter model (Birnbaum, 1968), and the graded response model (Samejima, 1969). Due to the increased popularity of computerized testing, response times have become available in addition to the responses. Such response times may aid in estimating the latent ability because of the "speed-accuracy tradeoff"—that is, faster responses may tend to be less thought out.

Research has focused on how to incorporate this additional source of information concerning individual differences in the existing measurement models. Main motivations to include the response times in the measurement model have been to increase measurement precision about the latent ability (e.g., Ranger & Ortner, 2011; Van der Linden, Entink, & Fox, 2010), to test substantive theories about cognitive processes (e.g., Klein Entink, Kuhn, Hornke, & Fox, 2009; Van der Maas, Molenaar, Maris, Kievit, & Borsboom, 2011) and personality constructs (Ferrando & Lorenzo-Seva, 2007a; 2007b), and to improve test construction (item calibration, item selection in adaptive testing, etc.; Van der Linden, 2007).

Currently, the dominant approach to the analysis of responses and response times is the hierarchical generalized linear modeling approach. In this approach, a latent speed variable is operationalized using a measurement model for the response times. This measurement model is subsequently connected to the measurement model for the responses. For example, the person and item parameters from both measurement models can be considered as random variables that have a common multivariate normal distribution across the models (Glas and van der Linden, 2010; Klein Entink, Fox, & van der Linden, 2009; Loeys, Legrand, Schettino, & Pourtois, 2014; Van der Linden, 2007, 2009a). Other researchers have simplified this model by only assuming a common distribution for the speed and ability variables (Molenaar, Tuerlinckx, & Van der Maas, 2015a; Ranger & Ortner, 2012; Wang, Chang, & Douglas, 2013; Wang, Fan, Chang, and Douglas, 2013). Alternatively, the speed and ability variables are assumed to be uncorrelated, but with linear cross-loadings of the response times on the ability variable (Furneaux, 1961; Molenaar, Tuerlinckx, & Van der Maas, 2015b, Thissen, 1983). This approach has been extended to include nonlinear cross-loadings to accommodate personality data (Ferrando and Lorenzo-Seva, 2007a; 2007b; Molenaar et al., 2015b; Ranger, 2013).

CONTACT Dylan Molenaar D.Molenaar@uva.nl Department of Psychology, University of Amsterdam, Weesperplein 4, 1018 XA, Amsterdam, The Netherlands.

Between-subject differences

All these approaches have in common that they solely model differences between subjects in ability and speed. That is, the main effects of the respondents' speed and the respondents' ability are captured by the latent speed and latent ability variables. As these latent variables are static variables, speed and ability are assumed to be constant within subjects (Goldhammer & Kroehne, 2014; Meng, Tao, & Chang, 2015; Van der Linden, 2009a). Thus, it is assumed that respondents work with a constant speed and a constant ability through the test. Statistically, this assumption is relatively unproblematic as violations of this assumption can—at least partly—be accommodated by modeling the conditional dependence of the responses and the response times of a given item (see Meng, Tao, & Chang, 2015; Molenaar et al., 2015b; Ranger & Ortner, 2012). In addition, the effect of differential speededness can be controlled for by design (Goldhammer, 2015; Goldhammer & Kroehne, 2014; Van der Linden, 2009b; Van der Linden, Scrams, & Schnipke, 1999). However, these approaches, while accounting for violations of the assumptions, do not allow the researcher to study how speed and ability develop within subjects.

Within-subject differences

There are various reasons why a researcher might be interested in within-subject differences in speed and ability (Molenaar, 2015). First, the researcher may want to assign different scores to different speed–ability compromises (see Maris & Van der Maas, 2012). That is, a fast correct response might be given more credit as compared to a slow correct response. A second reason, which is the focus of this article, is that a researcher might be interested in the underlying process that resulted in the response. That is, there may be differences in the response process of an individual throughout the test administration. These differences may be due to the use of different psychologically relevant solution strategies, for example, if different cognitive strategies are being used to solve the test items (Van der Maas & Jansen, 2003). Or, there might be undesirable strategies such as faking on some of the items of a test (Holden & Kroner, 1992) or the use of item preknowledge (McLeod, Lewis, & Thissen, 2003). Other examples include differences due to factors related to "testing," for instance, learning and practice effects (Carpenter, Just, & Shell, 1990), posterror slowing (Rabbit, 1979), and fatigue and motivation issues (Mollenkopf, 1950).

If the differences in response processes are large enough and if the response processes differ in their execution time, the measurement properties of the faster responses will differ from those of the slower responses reflecting that a different process underlies the measurement. If respondents stick to the same response process on all items of a test, this effect will be captured by the between-subjects speed and ability variables. However, if respondents switch between response processes during test administration, this is a within-subjects effect.

Existing approaches

As discussed in the preceding, in the hierarchical generalized linear models, the within-subject effects are absorbed in the residual of the model. Therefore, researchers have focused on detecting different response processes by considering the residual response times. These residuals can be tested for aberrances. For example, extreme residuals may suggest the use of different response strategies (Van der Linden & Guo, 2008), or trends in the residuals may suggest an effect related to testing, such as learning during the test or a decreased motivation (Van der Linden, Breithaupt, Chuah, & Zhang, 2007) or warming up and slowing down effects (Van der Linden, 2009b).

Besides the residual response time approach, the item response theory (IRT) tree approach is suitable to detect within-subject differences in responses and response times (Partchev & De Boeck, 2012). In this approach, the continuous response times are dichotomized into a fast and slow category. As a result, the fast responses can be investigated separately from the slow responses to reveal possible differences among them. Finally, a suitable approach to detect within-subject differences in speed and ability due to rapid guessing behavior is the hierarchical mixture modeling approach by Wang and Xu (2015). In this model, faster responses are assumed to be the result of a guessing process, which is modeled separately from the slower responses.

Aim of the present study

In this article, we adopt a dynamic modeling approach to separate the between-subjects variability from the within-subjects variability (Molenaar, 2004). Specifically, using a hidden Markov modeling framework (e.g., MacDonald & Zucchini, 1997; Vermunt, Langeheine, & Bockenholt, 1999), we distinguish the between-subjects ability and speed variables from the within-subjects states variables (Hamaker, Nesselroade, & Molenaar, 2007). That is, respondents are assumed to work at an overall speed and overall ability level through the test, but for each item, the response may be the result of a different state. The states are Markov dependent and may differ in their measurement properties. As a result, inferences can be made about the nature of the response processes underlying a given test. The present approach is embedded in the hierarchical generalized linear modeling framework and

therefore applicable to the modeling instances discussed in the preceding. This approach has advantages over the existing approaches: (a) It combines the response and response time information into a single measure for inferences about dynamic response behavior instead of only considering the residual response times; (b) it takes the dependency of the responses and response times to subsequent items into account (e.g., if a respondent guesses on a given item, he or she may be more likely to guess on the next item); (c) it enables the formulation of an explicit model-based multivariate approach to test for dynamics in the response behavior of a given test administration; (d) it enables researchers to specify theoretical constraints to identify specific answer strategies; (e) it takes the possible differences in the measurement properties of the different solution strategies or response processes into account; (f) it avoids the dichotomization of the continuous response times as in the IRT tree approach, whereby we retain all information about individual differences in the response times; and (g) it provides a statistically justified distinction between faster and slower responses instead of an ad hoc chosen cutoff point. All these possibilities will be demonstrated in this article.

The outline is as follows: First we derive the hidden Markov modeling approach to the analysis of responses and response times. Next, we present a simulation study to establish the parameter recovery of the model and to study the performance of various fit indices in distinguishing between models with and without different item states. Then, we present an exploratory application to the knowledge subtest of the Intelligence Structure Test (Amthauer, Brocke, Liepmann, & Beauducel, 2001) and a confirmatory application of the model to data on the balance scale task in children (Van der Maas & Janssen, 2003). We end with a general discussion of the results.

Hidden Markov modeling of responses and response times

To account for within-subject differences in the measurement properties of faster and slower responses, we assume an item-specific latent class variable, C_{pi}, to underlie the response, X_{pi}, and the response time, T_{pi}, of respondent p on item i. In the following, we assume that the response times follow a log-normal distribution such that the log-response times are normally distributed (see, e.g., Thissen, 1983; Van der Linden, 2007). The latent states of the latent class variable, $C_{pi} = 0, \ldots, K\text{-}1$, may represent different response processes or different solution strategies, where K represents the number of states, which is chosen by the researcher as will be explained in the following. As the latent state on item i may depend on the latent state on item i-1, we assume a Markov dependency of order 1 for C_{pi}. Let $\boldsymbol{x}_p$ denote the vector of item responses, $\boldsymbol{x}_p = [X_{p1}, X_{p2}, \ldots X_{pn}]$; let $\boldsymbol{t}_p$ denote the vector of log-response times, $\boldsymbol{t}_p = [lnT_{p1}, lnT_{p2}, \ldots, lnT_{pn}]$; and let $\mathbf{c}_p$ denote the vector of item states,

$\mathbf{c}_p = [C_{p1}, C_{p2}, \ldots, C_{pn}]$. Then the joint data density is given by

$$h\left(\boldsymbol{x}_p, \boldsymbol{t}_p\right) = \sum_{C_{p1}=0}^{K-1} \sum_{C_{p2}=0}^{K-1} \cdots \sum_{C_{pn}=0}^{K-1} f\left(\boldsymbol{x}_p, \boldsymbol{t}_p | \boldsymbol{c}_p\right) \times P(C_{p1}) \times \prod_{i=2}^{n} P\left(C_{pi} | C_{p(i-1)}\right) \quad (1)$$

where $P(C_{p1})$ is the initial state probability, which models the probability that a response belongs to a given state at item 1. In addition, $P(C_{pi}|C_{p(i-1)})$ is the transition probability, which models the dependency between the states on subsequent items. Note that our approach of introducing Markov-dependent item states is one possibility to account for dynamic behavior; for other possibilities, see Hamaker et al. (2007), Kempf (1977), Verhelst and Glas (1993), and Wang, Berger, and Burdick (2013).

In this model, differences in item and person properties are not taken into account. The latent class variables in $\mathbf{c}_p$ will be conflated by differences between respondents in overall ability (θ_p) and overall speed (τ_p) and by differences between items in overall easiness (β_i) and overall time intensity (ν_i). That is, we need to specify a measurement model for the responses and the response times to separate item effects, person effects, and the effect of the latent state C_{pi}. We follow Molenaar et al. (2015a), Ranger (2013), Ranger and Ortner (2012), Wang, Chang et al. (2013), and Wang, Fan et al. (2013) and treat the item parameters as fixed and the respondent parameters as random. Note that Glas and Van der Linden (2010), Klein Entink et al. (2009), and Van der Linden (2007) proposed measurement models for responses and response times incorporating both random person and random item effects.

By assuming independence between $\mathbf{x}_p$ and $\mathbf{t}_p$ conditional on θ_p and τ_p within the states, $\boldsymbol{c}_p$, the bivariate data density function factors as follows:

$$f\left(\boldsymbol{x}_p, \boldsymbol{t}_p | \theta_p, \tau_p, \boldsymbol{c}_p\right) = P\left(\boldsymbol{x}_p | \theta_p, \boldsymbol{c}_p\right) \times f\left(\boldsymbol{t}_p | \tau_p, \boldsymbol{c}_p\right).$$

As a result, we can specify separate measurement models for the responses and the response times within the latent states. In the hierarchical generalized linear modeling approach, models that have been considered for the responses are the Rasch model (Loeys et al., 2014), the two-parameter model (Thissen, 1983; Molenaar et al., 2015a, 2015b), the graded response model (Molenaar et al., 2015b; Ranger, 2013), the linear factor model (Ferrando & Lorenzo-Seva, 2007b), and the three-parameter

model (Klein Entink et al., 2009; Van der Linden, 2007). Here we specify a two-parameter model for binary item scores within each state, that is

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

where $\omega(.)$ is the logistic function. Parameters α_{si} and β_{si} denote the discrimination and easiness parameters in state $C_{pi} = s$ on item i.

For the response times, measurement models that have been considered are a log-normal model (Thissen, 1983; Van der Linden, 2007), a proportional hazards model (Loeys et al., 2014; Wang, Fan et al., 2013), a linear transformation model (Wang, Chang et al., 2013), and a categorical model for discretized time (Ranger & Kuhn, 2012; 2013). Here we specify a log-normal model for continuously distributed response times conditional on τ_p within each state; that is, the vector of log-response times, $\mathbf{t}_p$, is assumed to have a conditional multivariate normal distribution with uncorrelated components (lnT_{pi}), that is,

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

with $\mu_{pi}|\tau_p, C_{pi} = E(lnT_{pi}|\tau_p, C_{pi}) = \nu_i - \delta_s - \tau_p$ with $\delta_0 = 0$, and $\delta_0 \leq \delta_1 \leq \cdots \leq \delta_{K-1}$, where ν_i is the time intensity parameter and ${\sigma_{\varepsilon i}}^2$ is the residual log-response time variance. In addition, δ_s denotes the expected speed for state $C_{pi} = s$. For identification reasons, we fix $\delta_0 = 0$. In addition, the constraint $\delta_0 \leq \delta_1 \cdots \delta_{K-1}$ ensures that the states are decreasing in their expected response time. Thus, if we assume two states, δ_1 denotes the mean difference in expected speed between states 0 and 1, where state 1 has a larger speed (and smaller expected response time). Therefore, a response X_{pi} has a higher probability to be from state 0 if the corresponding log-response time is closer to $\nu_i - \tau_p$, and the response has a higher probability to be from state 1 if the corresponding log-response time is closer to $\nu_i - \delta_s - \tau_p$.

Likelihood function

We focus on the log marginal likelihood of the data to facilitate marginal maximum likelihood estimation discussed later. To this end, we assume a bivariate normal distribution for the continuous latent variables θ_p and τ_p, with $VAR(\tau_p) = {\sigma_\tau}^2$, covariance $\sigma_{\theta\tau}$, and $VAR(\theta_p) = {\sigma_\theta}^2 = 1$. The log marginal likelihood of response vector $\boldsymbol{x_p}$ and the log-response time vector $\boldsymbol{t_p}$ given the model parameter vector, $\boldsymbol{\eta}$, is then given by

$$\ell\left(\boldsymbol{x}_p, \boldsymbol{t}_p;\right) = ln \int\int_{-\infty}^{\infty} \sum_{C_{p1}}^{K-1} \sum_{C_{p2}}^{K-1} \cdots \sum_{C_{pn}}^{K-1} P\left(\boldsymbol{x}_p|\theta_p, \boldsymbol{c}_p\right) \times f\left(\boldsymbol{t}_p|\tau_p, \theta_p, \boldsymbol{c}_p\right) P\left(C_{p1}\right) \times \prod_{j=2}^{n} P\left(C_{pj}|C_{p(j-1)}\right) g\left(\theta_p, \tau_p\right) d\theta d\tau,$$

where the initial state probability is parameterized as $\beta_{i1} = \beta_{i0} + \Delta\beta$. In addition, the transition probabilities are parameterized as $P(C_{pj} = s|C_{p(j-1)} = r) = \pi_{s|r}$ with $\pi_{0|r} = 1 - \sum_{s=0}^{K-1} \pi_{s|r}$. That is, there are K initial state probabilities and $K \times (K-1)$ transition probabilities. Note that we assume time homogeneity of the Markov chain (Bacci, Pandolfi, & Pennoni, 2014); that is, the transition probabilities are equal for all subsequent items.

The Markov-dependent item states model

The model described in the preceding is referred to as the *Markov-dependent item states model*. The Markov structure of this model is thus characterized by time homogeneity of the n latent class variables each with K states. The free parameters in $\boldsymbol{\eta}$ for this model are the $n \times (2 \times K)$ item parameters for each state (α_{si}, β_{si}), the $2 \times n$ response time parameters (ν_i, and ${\sigma_{\varepsilon i}}^2$), the K-1 state response time parameters (δ_s, for $s \neq 0$), the K-1 initial state parameters (π_s, for $s \neq 0$), the $K \times (K - 1)$ transition parameters ($\pi_{s|r}$, for $s \neq 0$), and the population parameters (${\sigma_\tau}^2$ and $\sigma_{\theta\tau}$). The total number of parameters is thus equal to $n \times (2 \times K) + 2 \times n + 2 \times (K - 1) + K \times (K - 1) + 2$. See Figure 1 for a schematic representation of the model.

In the Markov dependent item states model, two features are worth mentioning. The first feature is that the response time parameters are assumed to be equal across states while the item response parameters are allowed to differ across states. Our main interest is to study the differences across states in the item discrimination and item easiness parameters to make inferences about the response processes underlying the item responses. The response times are used as a tool to accomplish this. Differences between states in the item time intensity and residual variance are therefore not our main interests and make the model needlessly complex (it would increase the number of item parameters from $2 \times K + 2$ to $4 \times K + 2$). Instead, we used the δ_s parameterization (which can be seen as a uniform difference across states in the time intensities) to identify—in a parsimonious way—the faster states in terms of the faster responses and the slower states in terms of the slower responses. We used

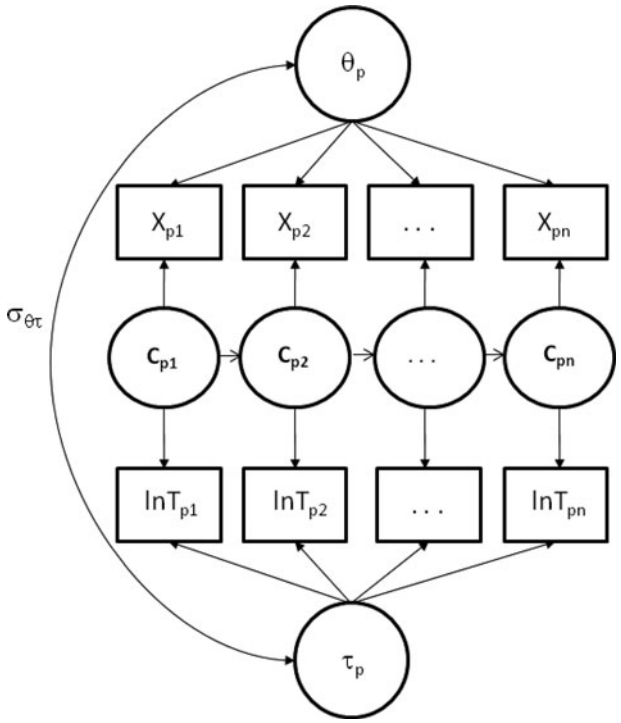

Figure 1. Graphical representation of the Markov-dependent item states model. Dependency of the ability and speed loadings on the latent states C_{pi} has been omitted from the graph for clarity.

a similar line of reasoning for the assumed homogeneity of the Markov states. That is, we could have introduced item-specific transition parameters making the Markov states time heterogeneous. However, this would increase model complexity severely while these additional parameters will not directly contribute to our substantive understanding of the differences between faster and slower responses. We note, however, that if it is of theoretical or practical interest, it is certainly possible to relax the time homogeneity assumption and the assumption of state-independent time intensity and residual variance parameters.

The second feature involves the fixed item effects in the preceding model. In the log-normal model by Van der Linden (2007), the item effects are considered as random effects (see also Glas & Van der Linden, 2010). Here, we follow Molenaar, Tuerlinckx, and Van der Maas (2015b), Ranger and Ortner (2012), Van der Linden and Guo (2008), Wang, Chang et al. (2013), and Wang, Fan et al. (2013) and treat the item effects as fixed. Note that Molenaar et al. (2015a) have shown that neglecting the randomness in the item parameters does not notably bias the results of the log-normal model for 20 or 40 items. We do however note that random item effects may be valuable in some research situations (see De Boeck, 2008).

Special cases

We consider two special cases of the Markov-dependent item states model in the preceding. The first special case arises when the Markov dependencies between the latent class variables in $\mathbf{c}_p$ are dropped from the model, that is,

$$P\left(\boldsymbol{c}_p\right) = P\left(C_{p1}\right) P\left(C_{p2}\right) \ldots P\left(C_{pn}\right),$$

with time homogenous state probabilities

$$P\left(C_{pi} = s\right) = \pi_s \quad \text{for } i = 1, \ldots, n.$$

Thus, for each item, the respondents have a probability of π_s to respond according to the state s measurement model. We will refer to this model as the *independent item states model*. This model follows from the Markov-dependent item states model by omitting the $K \times (K-1)$ transition parameters (i.e., if $\pi_{s|r} = \pi_s$ and $\pi_{s|s} = \pi_s$, the transition parameters cancel out of the likelihood equation). That is, the probability to be in class s at a given item does not depend on the state at the previous item and equals π_s for all items: for example, this will be the case when the response strategy is chosen for each item independently, irrespective of the strategy employed to answer the previous items. The model contains $n \times (2 \times \mathrm{K}) + 2 \times \mathrm{n} + 2 \times (\mathrm{K}-1) + 2$ parameters and may be a useful baseline model to make inferences about the presence of dependencies between the item states by comparing its fit to the fit of the Markov-dependent item states model. If, in the independent item states model, one specifies $\mathrm{K} = 2$, $\alpha_{0\mathrm{i}} = \alpha_0 = 0$, and $\beta_{0\mathrm{i}} = \beta_0 = \omega^{-1}(g)$ where g is a guessing probability, then the resulting model is equivalent to the mixture model by Wang and Xu (2015) for rapid guessing behavior.

The second special case arises when we omit all latent class variables, $\mathbf{c}_p$, from the model. That is, in the independent item states model, we specify $\pi_1 = 1$ and $\alpha_{si} = \alpha_{\mathrm{ri}}$, $\beta_{si} = \beta_{\mathrm{ri}}$, and $\delta_{\mathrm{s}} = 0$ for all s and r. The model thus includes $4 \times n + 2$ parameters and is equivalent to the hierarchical generalized linear model of Van der Linden (2007) with random person effects and fixed item effects (see Molenaar et al., 2015a; Ranger, 2013; Ranger & Ortner, 2012; Wang, Chang et al., 2013; and Wang, Fan et al., 2013). This will be appropriate when any given respondent chooses one single response strategy for all items and sticks to it during the entire test, for example. This model without item states constitutes a useful baseline model to make inferences about the presence of dynamic item states by comparing its fit to the Markov-dependent and independent item states models.

Exploratory and confirmatory use

The preceding dynamic item states models can be used in an exploratory and confirmatory application. In an exploratory application, there are no expectations about the item states underlying the data. In such a case, all parameters are estimated freely to infer differences in measurement properties across faster and slower item responses. As the Markov-dependent item states model contains $n \times (2 \times \mathrm{K}) + 2 \times \mathrm{n} + 2 \times (\mathrm{K} - 1) + \mathrm{K} \times (\mathrm{K} - 1) + 2$ parameters and the independent item states model contains $n \times (2 \times \mathrm{K}) + 2 \times \mathrm{n} + 2 \times (\mathrm{K} - 1) + 2$ parameters, the full unconstrained model becomes very demanding for $\mathrm{K} > 2$. In exploratory settings we therefore advise that for moderate sample sizes (500–1,000), $\mathrm{K} = 2$ is used. In the following simulation study, we demonstrate that this model is feasible for $N = 500$ and $N = 1{,}000$. We think that, in an exploratory setting, additional states ($\mathrm{K} > 2$) will not capture substantial patterns in the data that are missed by the $\mathrm{K} = 2$ model for sample sizes around 500–1,000. That is, if the data truly contain five states differing in their expected response times, the $\mathrm{K} = 2$ exploratory model will be a reasonable approximation that captures the most important patterns in the data. The slow state will contain the measurement properties of the slower states in the data, and the fast state will contain the measurement properties of the faster states in the data. If a researcher wants to know how many states are truly in the data, either a very large sample size should be used, or ideally, a theory about the number of states is considered to enable a confirmatory application, as explained in the following.

In exploratory settings, the item parameters α_{si} and β_{si} and δ_{s} are used to quantify the differences in measurement properties between the faster and slower item responses. That is, if α_{si} and/or β_{si} are unequal across states, measurement invariance is violated, indicating that the faster responses measure a psychometrically different state than the slower responses. Therefore, in some cases the differences among the states in α_{si} and β_{si} may be used to interpret the item states. For example, fast guessing is characterized by discrimination parameters that approach 0 and easiness parameters that approach the guessing level in the fast state (e.g., $\omega^{-1}(0.25) = -1.10$ in the case of a multiple-choice test with four answer options). In addition, item preknowledge will be reflected by small discrimination and high easiness parameters in the fast class. However, as with measurement invariance and differential item functioning research, sometimes it is unclear why a violation occurs. This can then be addressed in follow-up research (e.g., explaining the invariance using covariates, see Steinmayr, Bergold, Margraf-Stiksrud, & Freund, 2015).

In a confirmatory setting, the number of parameters can be decreased substantially by introducing constraints (e.g., by fixing the item easiness parameters to reflect a fast-guessing state). Hereby, models with $\mathrm{K} > 2$ are feasible depending on the number of states expected theoretically. Identification of these models depends on the exact constraints that are introduced by the researcher. It is therefore important that in confirmatory applications of the model, the modeling results are carefully checked on signals of nonidentification (e.g., large standard errors and ill conditioning of the Hessian matrix). We also encourage researchers to use multiple sets of starting values in both the exploratory and confirmatory applications of the present model.

Another issue in confirmatory applications is related to testing the number of states in the data. That is, multiple theories might exist that predict a different number of states. In the simulation study, we identify fit indices that are suitable to select among models with either $\mathrm{K} = 1$ (i.e., a static model) or $\mathrm{K} = 2$. Similarly as in our simulation study to follow, for $\mathrm{K} > 2$, it has been found that the Bayesian information criterion (BIC; Schwarz, 1978) performs satisfactorily while the Akaike information criterion (AIC; Akaike, 1974) tends to underpenalize model complexity (Celeux & Durand, 2008; Visser, Raijmakers, & Molenaar, 2002). The bootstrapped likelihood ratio statistic is also known to be suitable to determine the number of states in a Markov model (Gudicha, Schmittman, Tekle, & Vermunt, 2015).

Estimation

The parameters from the models can be estimated using marginal maximum likelihood estimation (MML; Bock & Aitkin, 1981). To this end, the preceding models are implemented in the LatentGOLD 5.0 software package (Vermunt & Magidson, 2013). The integrals in the likelihood function are approximated using Gauss-Hermite quadratures with 10 nodes per dimension (100 in total). This function is optimized using the EM algorithm and the Newton-Raphson algorithm. In the E-step of the EM algorithm, the Baum-Welch forward-backward algorithm is used to avoid the computation of the joint density of the latent class variables in $\mathbf{c}_p$, which is numerically demanding (Baum, Petrie, Soules, & Weiss, 1970; Vermunt, Tran, & Magidson, 2008). To facilitate parameter estimation, the logit of the initial state probability parameters, π_{s}', and the transition probability parameters, $\pi_{\mathrm{s|r}}$', are estimated. In addition, ${\sigma_\tau}^2$ and $\sigma_{\theta\tau}$ are estimated by estimating the corresponding elements from the Cholesky decomposition of the covariance matrix of θ and τ, denoted ${\sigma_\tau}^{2}$' and $\sigma_{\tau\theta}$'. Note that ${\sigma_\theta}^2$ is not a free parameter in the

model. The syntax to fit the different models is available from the site of the first author.

Simulation study

The simulation study presented here served multiple purposes. First, we wanted to establish whether true parameter values can satisfactorily be recovered for the Markov-dependent item state model and the independent item state model. In addition, we wanted to establish the performance of various fit indices in distinguishing between the models with and without item states. As the dynamic models become increasingly complex for increasing K, we focus on $K = 2$ here. That is, a fast and a slow state. Adding more states to the model is possible in principle, but additional constraints to identify each state are needed. We will illustrate this in the application section.

Design

We simulated data according to the Markov-dependent item state model, the independent item state model, and the model without item states. We used 100 replications. In the case of the models with item states, we manipulated the expected speed difference between the two states, δ_1, into three levels, which we refer to as a "small" ($\delta_1 = 0.4$), medium" ($\delta_1 = 0.5$), and "large" ($\delta_1 = 0.6$) effect. In addition, we manipulated the state stability into two levels, "stable" and "unstable" states. The stability of the states is defined by the transition probabilities, $\pi_{1|0}$ and $\pi_{1|1}$. Larger values for $\pi_{1|1}$ and smaller values for $\pi_{1|0}$ indicate more stable states. We chose $\pi_{1|0} = 0.15$ and $\pi_{1|1} = 0.85$ for the stable condition and $\pi_{1|0} = 0.3$ and $\pi_{1|1} = 0.7$ for the unstable condition. Finally, we manipulated the sample size to be $N = 500$ and $N = 1{,}000$.

The remaining parameters are not manipulated. We used 20 items. For the easiness parameters in state 1, β_{1i}, we used increasing, equally spaced values between -2 and 2 for the 20 items. For the easiness parameters in state 0, β_{0i}, we used increasing, equally spaced values between -1.5 and 2.5. Possibly, the difference between β_{0i} and β_{1i} might also affect the success with which differences between the two states are detected by the fit indices. However, as the two states are formally defined in terms of a difference in speed and not in terms of a difference in easiness (e.g., a "hard" and "easy" state), we chose to manipulate δ_s instead of manipulating the difference between β_{0i} and β_{1i}.

The discrimination parameters in state 1, α_{1i}, were all chosen to equal 1 for the odd items and 2 for the even items. For the discrimination parameters in state 0, α_{0i}, we chose 1.5 for the odd items and 2.5 for the even items. For the time intensity parameters in the response time model, ν_i, we used 2 for the odd items and 3 for the even items. The residual response time variance, ${\sigma_{\varepsilon i}}^2$, was chosen to equal 0.2. The latent speed factor variance, ${\sigma_\tau}^2$, was equal to .01, and $\sigma_{\theta\tau}$ was chosen to be 0.07 such that the correlation between θ_p and τ_p equaled .7. We chose this positive correlation because of our own experiences with response and response time modeling. However, we note that the correlation between speed and ability can also be negative (see Van der Linden, 2009a). In addition, the values for ${\sigma_{\varepsilon i}}^2$ and ${\sigma_\tau}^2$ may seem small but they are reasonable in a response time modeling setting. That is, they result in untransformed response times between approximately 2 and 14 seconds. Finally, for the dynamic models, $\pi_1 = .7$ was used. For the static model without item states, we used the values of ν_i, ${\sigma_{\varepsilon i}}^2$, and ${\sigma_\tau}^2$ discussed in the preceding together with α_{1i} and β_{1i} from class 1 and with $\delta_1 = 0$.

Model selection

To see whether we can successfully distinguish the different models, we study the performance of various fit indices in indicating the best fitting model under the different conditions of the simulation study. The three models considered here are nested according to the restrictions discussed in the preceding. However, as these restrictions include various boundary constraints, we do not consider the power of the likelihood ratio test to distinguish between the different models. Although the performance of fit indices such as AIC and BIC may also suffer from such boundary constraints (see, e.g., Greven & Kneib, 2010), we identify specific fit indices that can be successfully used to separate between competing models despite these boundary constraints. Specifically, we focus on the AIC and BIC previously discussed and the consistent AIC (CAIC; Bozdogan, 1987), the AIC3 (Bozdogan, 1993), and the sample-size-adjusted BIC (saBIC; Sclove, 1987). For these fit indices it holds that a smaller value indicates better model fit.

For each model-fitting attempt, we ran 16 different sets of random starting values. If the estimation algorithm did not converge, we reran the model at most two more times, again using 16 different sets of random starting values. Only for a few cases, convergence issues remained after these 3 estimation attempts. The Markov-dependent item states model failed to converge in 13 of the 2,000 simulated data sets; the independent item states model failed to converge in 29 of the 2,000 simulated data sets. These nonconverged cases concerned mainly cases in which the true model was the static model without item states. We retained the nonconverged cases as these will hardly affect the results presented in the following.

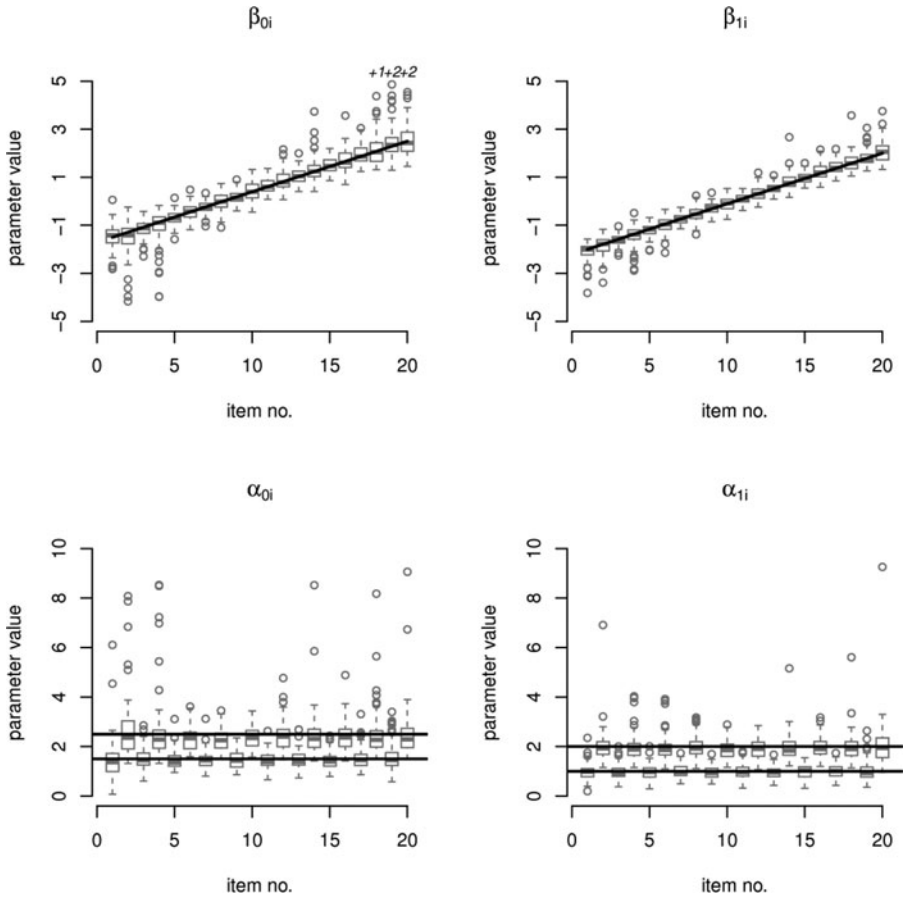

Figure 2. Boxplot of the parameter estimates of the Markov-dependent item states model in the medium effect size and stable classes condition for $N = 500$. The solid line denotes the true parameter values. For the discrimination parameters α_{0i} and α_{1i}, the odd items correspond to the upper line and the even items correspond to the lower line.

Results

Parameter recovery

Item parameters

We limit the presentation of the results for the recovery of the item parameters to the medium effect size condition. The recovery for the small and the large effect size condition follows a similar pattern of results. We study parameter recovery of the item parameters by means of a box plot of the parameter estimates in the case that the true model is fit to the data. For the Markov-dependent item states model, these box plots are displayed in Figure 2 ($N = 500$) and Figure 3 ($N = 1{,}000$) for the easiness parameters (β_{0i} and β_{1i}) and the discrimination parameters (α_{0i} and α_{1i}). For the independent item states model, the box plots of the parameter estimates are in Figure 4 ($N = 500$) and Figure 5 ($N = 1{,}000$). As can be seen, the item parameters seem to be generally unbiased; that is, the parameter estimates scatter around the true parameter values for all items and for both dynamic models. The discrimination parameters have generally more variability as compared to the easiness parameters. In addition, the parameter estimates in state 1 are associated with somewhat less variability as compared to the estimates in state 0 for both the easiness parameters and the discrimination parameters. This is due to state 0 being proportionally smaller ($\pi_1 = .7$). Finally, the parameter recovery in the Markov-dependent item states model is generally associated with less variability in the estimates as compared to the independent item states model.

Class parameters

For the initial state probability (Markov-dependent item states model) or state probability (independent item states model) parameter π_1, the transition parameters $\pi_{1|0}$ and $\pi_{1|1}$ (Markov-dependent item states model), and the state response time parameter, δ_1, the results concerning parameter recovery are in Table 1 for the Markov-dependent item states model and in Table 2 for the independent item states model. That is, the mean, standard deviation, and mean standard errors are depicted for the parameter estimates in the true model (Markov-dependent or independent item state model) for the different configurations of the parameters for $N = 500$. Note

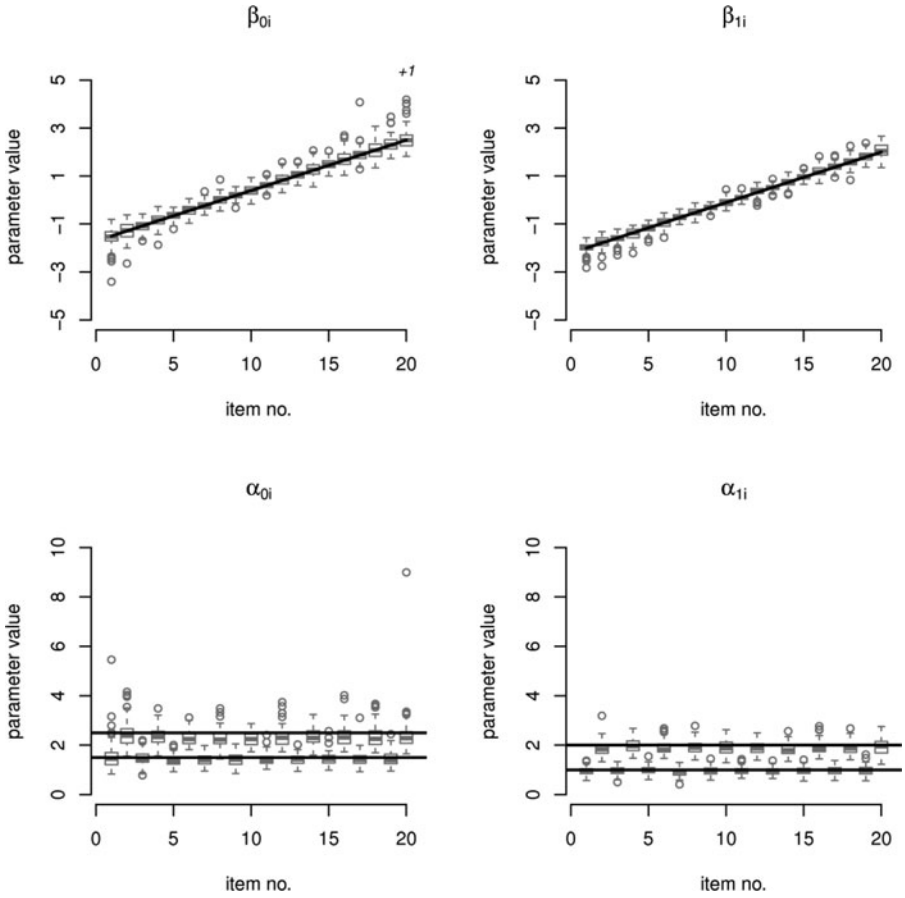

Figure 3. Boxplot of the parameter estimates of the Markov-dependent item states model in the medium effect size and stable classes condition for $N = 1{,}000$. The solid line denotes the true parameter values. For the discrimination parameters α_{0i} and α_{1i}, the odd items correspond to the upper line and the even items correspond to the lower line.

that we estimate the logit transformed parameters, $\pi_1^{'}$, $\pi^{'}_{1|0}$, and $\pi^{'}_{1|1}$ as discussed in the preceding; however, we provide the parameter recovery results in terms of the original parameterization (i.e., π_1, $\pi_{1|0}$, and $\pi_{1|1}$). The reported standard errors are obtained by the univariate delta method. As can be seen from the table, true parameter values are generally recovered well, with slightly better recovery for larger effect sizes and hardly any difference between the recovery in the independent item states model and in the Markov-dependent item states model. In general, standard errors decrease as the effect size increases. For $N = 1{,}000$, results are similar. In addition, the recovery of the true parameter values for $\sigma_\tau{}^2$ and $\sigma_{\tau\theta}$ is good (not depicted).

Model selection

Tables 3, 4, and 5 contain the "selection rates" of the different models. We defined a selection rate as the proportions of replications in which the different models are identified as the best fitting model by the different fit indices when the true model is a Markov-dependent item states model, an independent item states model, or a static model without item states. The selection rate in the case that a given model is the true model is referred to as "hit rate"; the selection rate in the case that a given model is not the true model is referred to as "false positive rate."

First, we focus on the hit rates and the false positive rates when the true model is the Markov-dependent item states model, see Table 3. As can be seen, in general, the hit rates of the Markov-dependent item states model increase for increasing N and δ_1, and the hit rates decrease for increasing $\pi_{1|0}$. The hit rate of the BIC fit index is conservative for $\pi_{1|0} = 0.3$ and poor for $\pi_{1|0} = 0.15$ and $\delta_1 = 0.4$, but the hit rates are acceptable for the other cases (between 0.77 and 1.0). The AIC fit index has hit rates close to 1 in all cases. The AIC3 index has hit rates close to 1.0 in all cases except the case of $N = 500$, $\pi_{1|0} = 0.30$, and $\delta = 0.4$, where the hit rate equals 0.12. The CAIC fit index has acceptable hit rates in the cases that $\pi_{1|0} = 0.15$ (close to 1.0 for medium and large effect sizes in δ_1 but 0 for a small effect size). However, in the case that $\pi_{1|0} = 0.30$, the hit rate is close to 0.0 for $N = 500$ and conservative for $N = 1{,}000$ (only large for large effect in δ_1). Finally, the saBIC fit index is associated with acceptable to good hit rates (only small for small effects in δ_1).

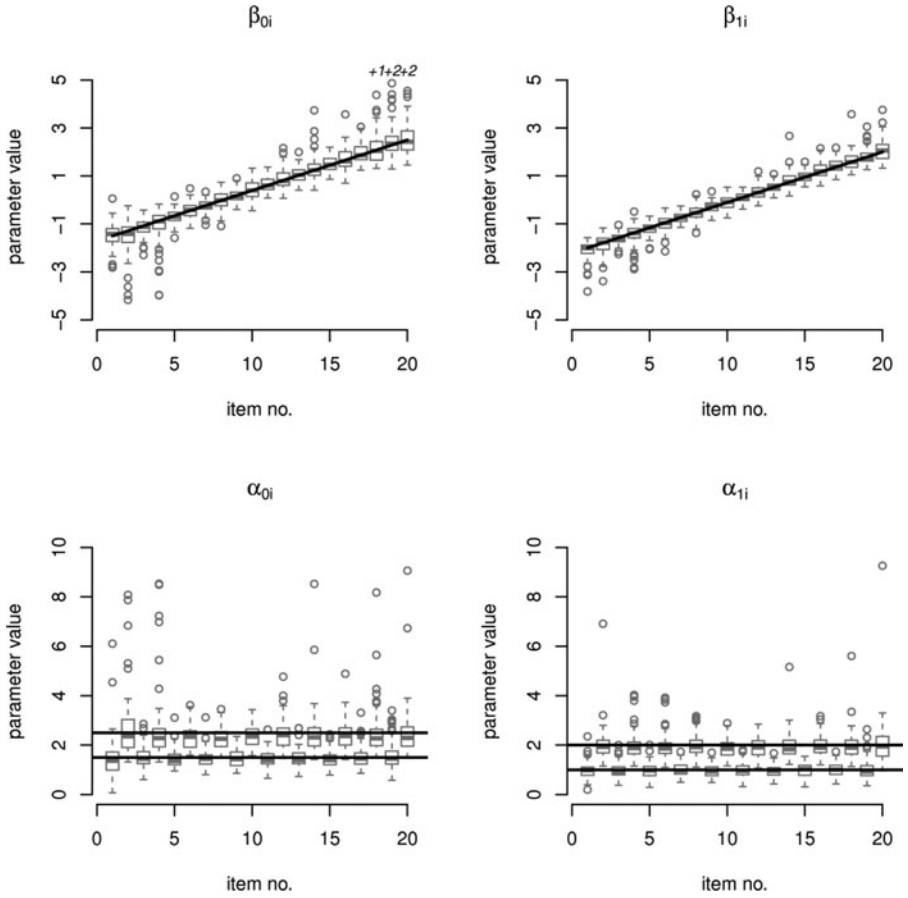

Figure 4. Boxplot of the parameter estimates of the independent item states model in the medium effect size and stable classes condition for $N = 500$. The solid line denotes the true parameter values. For the discrimination parameters α_{0i} and α_{1i}, the odd items correspond to the upper line and the even items correspond to the lower line.

Next, we focus on the hit rates and false positive rates of the independent item states model, see Table 4. As can be seen from the table, in general, the hit rates are poor. The hit rate is only acceptable for the AIC in the case that $\delta_1 = 0.6$, and in the case $\delta_1 = 0.5$ for $N = 1{,}000$. For the AIC3, the hit rate is only acceptable for $N = 1{,}000$ and $\delta_1 = 0.6$. For all other fit indices and all other conditions, the hit rates are unacceptable. As can be seen, the static model without item states is in general the preferred model over the independent item states model.

Finally, Table 5 contains the hit rates and false positive rates when the true model is the static model without item states. Ideally, the Markov-dependent item states and the independent item states models are not detected by any of the fit indices as the best fitting model as this would indicate that the fit indices might be biased in favor of the dynamic models. As can be seen from the table, this is not the case for the BIC, AIC3, CAIC, and saBIC. The AIC indicates in 13% ($N = 500$) and 17% ($N = 1{,}000$) of the replications wrongfully that the Markov-dependent item states model underlies the data while the static model is the true model. This fit index is thus associated with a slightly increased false positive rate.

From the preceding, it appears that the independent item states model is difficult to distinguish from a static model while the Markov-dependent item states model can be acceptably distinguished using the fit indices considered. That is, when the true model is an independent item states model and a static model is fit to that data, the misfit is only minor, causing the selection rates to be small. On the contrary, when the true model is a Markov-dependent item states model, fitting a static model to the data causes more severe misfit, which results in larger selection rates. To study the source of misfit, we investigated which parameters in the static model are biased systematically when the true model is a Markov-dependent item states model, but which are not biased (or only mildly) when the true model is an independent item states model. It appeared that misfit is most evident in the covariance between θ and τ. To see this, we scaled the estimates of the covariance $\sigma_{\tau\theta}$ into correlations $\rho_{\theta\tau}$ (see Table 6). As can be seen, when the true model is the independent item state model, but a static model is fit to the data, $\rho_{\theta\tau}$ is hardly affected. That is, the true value for the correlation equals 0.7, and this value is acceptably recovered with mean estimates of

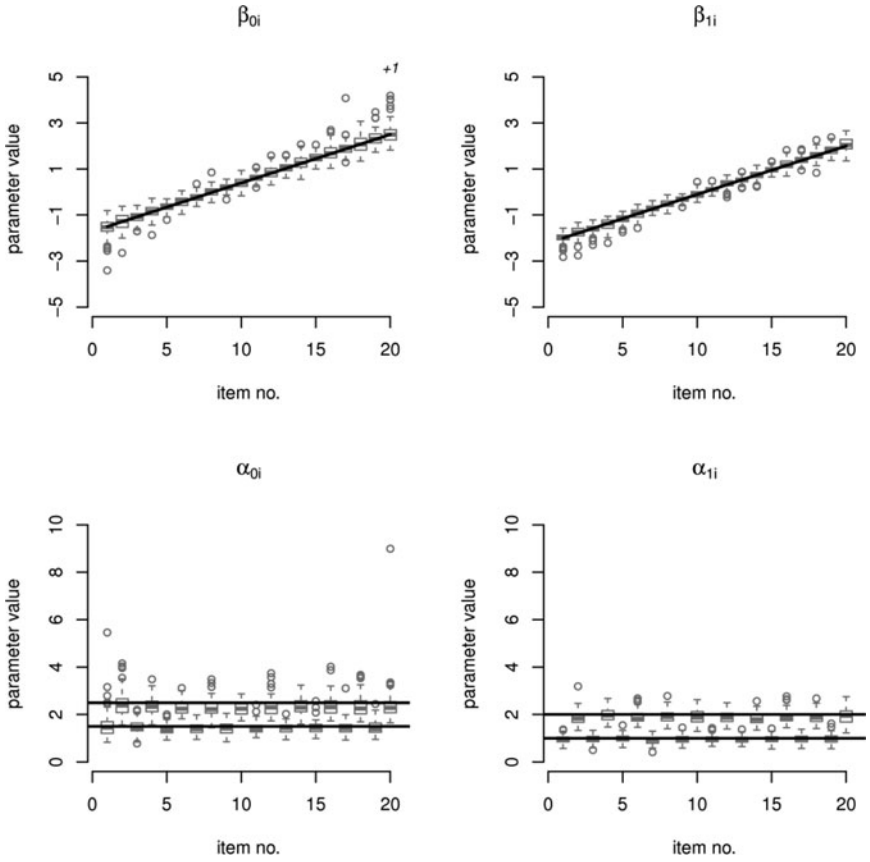

Figure 5. Boxplot of the parameter estimates of the independent item states model in the medium effect size and stable classes condition for $N = 1{,}000$. The solid line denotes the true parameter values. For the discrimination parameters α_{0i} and α_{1i}, the odd items correspond to the upper line and the even items correspond to the lower line.

Table 1. Mean, standard deviation (*SD*), and mean standard error (*MSE*) for the estimates of the class parameter, δ, π_0, $\pi_{1|0}$, and $\pi_{1|1}$ in the simulation study for the different configurations of δ, π_0, $\pi_{1|0}$, and $\pi_{1|1}$ in the Markov-dependent item states model for $N = 500$.

δ				π_0				$\pi_{1\|0}$				$\pi_{1\|1}$			
True	Mean	*SD*	*MSE*	True	Mean	*SD*	*MSE*	True	Mean	*SD*	*MSE*	True	Mean	*SD*	*MSE*
0.4	0.40	0.03	0.02	0.7	0.65	0.16	0.11	0.15	0.15	0.04	0.03	0.85	0.84	0.04	0.03
0.5	0.50	0.02	0.02	0.7	0.68	0.10	0.08	0.15	0.15	0.02	0.02	0.85	0.85	0.02	0.02
0.6	0.60	0.02	0.02	0.7	0.70	0.06	0.06	0.15	0.15	0.02	0.01	0.85	0.85	0.02	0.01
0.4	0.37	0.09	0.04	0.7	0.61	0.21	0.15	0.3	0.28	0.12	0.05	0.7	0.73	0.11	0.05
0.5	0.50	0.04	0.04	0.7	0.65	0.17	0.12	0.3	0.30	0.04	0.04	0.7	0.71	0.05	0.04
0.6	0.60	0.03	0.03	0.7	0.69	0.13	0.09	0.3	0.30	0.03	0.03	0.7	0.70	0.03	0.03

about 0.67. However, when the Markov independent item states model is the true model, the estimates of $\rho_{\theta\tau}$ are greatly affected, with values between 0.39 and 0.58. Thus, correlations between responses and response times are underestimated in the static model, causing model misfit. This misfit is not apparent if the true model is the independent item states model, causing this model to be hard to distinguish from the static model. However, it should be noted that the effect size for the independent item states model is small. We return to this point in the discussion section.

Table 2. Mean, standard deviation (SD), and mean standard error (MSE) for the estimates of the class parameter, δ, π_0, $\pi_{1|0}$, and $\pi_{1|1,}$ in the simulation study for the different configurations of δ, π_0, $\pi_{1|0}$, and $\pi_{1|1}$ in the independent item states model for $N = 500$.

δ				π_0			
True	Mean	*SD*	*MSE*	True	Mean	*SD*	*MSE*
0.4	0.36	0.13	0.05	0.7	0.62	0.14	0.07
0.5	0.49	0.08	0.04	0.7	0.68	0.06	0.05
0.6	0.60	0.03	0.03	0.7	0.70	0.03	0.03

In conclusion, from the parameter recovery results, we can conclude that the true values are retrieved generally satisfactorily. In addition, from results of the fit indices, it appears that the Markov-dependent model is generally well separable from the independent item states model

Table 3. True model: Markov-dependent item states model. Selection rates for the Markov-dependent item states model, the independent item states model, and the static model without item states for the various conditions in the simulation study.

N	$\pi_{1\|0}$	δ	Model fitted	BIC	AIC	AIC3	CAIC	saBIC
500	.15	.4	Static	1.00	0.00	0.02	1.00	0.02
			Independent	0.00	0.00	0.00	0.00	0.00
			Markov	0.00	1.00	0.98	0.00	0.98
		.5	Static	0.06	0.00	0.00	0.32	0.00
			Independent	0.00	0.00	0.00	0.00	0.00
			Markov	0.94	1.00	1.00	0.68	1.00
		.6	Static	0.00	0.00	0.00	0.00	0.00
			Independent	0.00	0.00	0.00	0.00	0.00
			Markov	1.00	1.00	1.00	1.00	1.00
	.30	.4	Static	1.00	0.08	0.88	1.00	0.89
			Independent	0.00	0.00	0.00	0.00	0.00
			Markov	0.00	0.92	0.12	0.00	0.11
		.5	Static	1.00	0.01	0.11	1.00	0.13
			Independent	0.00	0.00	0.00	0.00	0.00
			Markov	0.00	0.99	0.89	0.00	0.87
		.6	Static	0.83	0.00	0.00	0.99	0.00
			Independent	0.00	0.00	0.00	0.00	0.00
			Markov	0.17	1.00	1.00	0.01	1.00
1,000	.15	.4	Static	0.23	0.00	0.00	0.70	0.00
			Independent	0.00	0.00	0.00	0.00	0.00
			Markov	0.77	1.00	1.00	0.30	1.00
		.5	Static	0.00	0.00	0.00	0.00	0.00
			Independent	0.00	0.00	0.00	0.00	0.00
			Markov	1.00	1.00	1.00	1.00	1.00
		.6	Static	0.00	0.00	0.00	0.00	0.00
			Independent	0.00	0.00	0.00	0.00	0.00
			Markov	1.00	1.00	1.00	1.00	1.00
	.30	.4	Static	1.00	0.00	0.09	1.00	0.52
			Independent	0.00	0.00	0.00	0.00	0.00
			Markov	0.00	1.00	0.91	0.00	0.48
		.5	Static	0.77	0.00	0.00	0.98	0.00
			Independent	0.00	0.00	0.00	0.00	0.00
			Markov	0.23	1.00	1.00	0.02	1.00
		.6	Static	0.00	0.00	0.00	0.00	0.00
			Independent	0.00	0.00	0.00	0.00	0.00
			Markov	1.00	1.00	1.00	1.00	1.00

Note. BIC = Bayesian information criterion; AIC = Akaike information criterion; AIC3 = AIC with a penalty weight of 3 instead of 2; CAIC = consistent AIC; saBIC = sample-size-adjusted BIC.

Table 4. True model: Independent item states model. Selection rates for the Markov-dependent item states model, the independent item states model, and the static model without item states for the various conditions in the simulation study.

N	δ	Model fitted	BIC	AIC	AIC3	CAIC	saBIC
500	.4	Static	1.00	0.81	1.00	1.00	1.00
		Independent	0.00	0.05	0.00	0.00	0.00
		Markov	0.00	0.14	0.00	0.00	0.00
	.5	Static	1.00	0.61	0.99	1.00	0.99
		Independent	0.00	0.26	0.01	0.00	0.01
		Markov	0.00	0.13	0.00	0.00	0.00
	.6	Static	1.00	0.08	0.88	1.00	0.91
		Independent	0.00	0.72	0.12	0.00	0.09
		Markov	0.00	0.20	0.00	0.00	0.00
1,000	.4	Static	1.00	0.55	0.99	1.00	1.00
		Independent	0.00	0.29	0.01	0.00	0.00
		Markov	0.00	0.16	0.00	0.00	0.00
	.5	Static	1.00	0.09	0.80	1.00	0.99
		Independent	0.00	0.75	0.17	0.00	0.01
		Markov	0.00	0.16	0.03	0.00	0.00
	.6	Static	1.00	0.00	0.01	1.00	0.26
		Independent	0.00	0.83	0.89	0.00	0.69
		Markov	0.00	0.17	0.10	0.00	0.05

Note. BIC = Bayesian information criterion; AIC = Akaike information criterion; AIC3 = AIC with a penalty weight of 3 instead of 2; CAIC = consistent AIC; saBIC = sample-size-adjusted BIC.

Table 5. True model: Static model without item states. Selection rates for the Markov-dependent item states model, the independent item states model, and the static model without item states for the various conditions in the simulation study.

		Fit index				
N	Model fitted	BIC	AIC	AIC3	CAIC	saBIC
500	Static	1.00	0.86	1.00	1.00	1.00
	Independent	0.00	0.01	0.00	0.00	0.00
	Markov	0.00	0.13	0.00	0.00	0.00
1,000	Static	1.00	0.83	1.00	1.00	1.00
	Independent	0.00	0.00	0.00	0.00	0.00
	Markov	0.00	0.17	0.00	0.00	0.00

Note. BIC = Bayesian information criterion; AIC = Akaike information criterion; AIC3 = AIC with a penalty weight of 3 instead of 2; CAIC = consistent AIC; saBIC = sample-size-adjusted BIC.

and static model without item states using the BIC, AIC, CAIC, AIC3, and the saBIC. However, the AIC is associated with a slightly increased false positive rate and should thus be interpreted with care. The independent item states model, on the other hand, is hard to separate from the static model. All fit indices generally favored the static model when the true model was in fact an independent item states model.

Application 1: Identifying within-subject differences in the response process

Data

We now demonstrate how our model can be applied to explore possible differences in solution strategies used by the respondents. The data comprise the responses and response times of 389 psychology freshmen on the 28 items of the knowledge subtest of the Dutch Intelligence Structure Test (Amthauer et al., 2001).

Modeling

To the responses and response times, we applied the Markov-dependent item states model with two states (K = 2), the independent item states model with two states (K = 2), and the static model without item states. Within the independent and Markov-dependent item states model we also studied the degree to which the item parameters differ across states. To this end we considered a model in which we parameterized the discrimination parameters in state 1 as follows:

$$\alpha_{i1} = \alpha_{i0} + \Delta\alpha.$$

That is, we allowed for a uniform difference between the discrimination parameters in state 1 as compared to the discrimination parameters in state 0. We also considered a model in which we specified a similar effect on the item easiness parameters in addition to the uniform effect on the discriminations; that is,

$$\beta_{i1} = \beta_{i0} + \Delta\beta.$$

In addition, we studied a model with uniform differences on both the discrimination and easiness parameters.

Using the best fitting model, we will illustrate how the modeling results can be used to make inferences about the within-subject differences in the response process. To this end, we compare (a) the raw log-response times; (b) the standardized residual log-response time; and (c) the estimated state probabilities. First, the raw log-response times are simply the observed log-transformed response times on the items. Making inferences based on the raw log-response times is difficult because these response times conflate item and respondent main effects. Therefore, we also consider the standardized residual log-response times. A comparable Bayesian version of this statistic has been proposed by Van der Linden and Guo (2008) to investigate aberrant response times. If the standardized residual of a given response time is large, the response time deviates from the model expectations given ν_i and τ_p. This might suggest that a different response process underlies this response (e.g., guessing, item preknowledge, different response strategy, etc). Here we calculate the standardized residual log-response times as follows:

$$z_{pi} = \frac{lnT_{pi} - (\hat{\nu}_i - \hat{\tau}_p)}{\hat{\sigma}_{\varepsilon i}},$$

where $\hat{\nu}_i$, and $\hat{\sigma}_{\varepsilon i}$ are the MML estimates from the static model and $\hat{\tau}_p$ is the EAP estimate from the static model. Finally, we consider the expected a posteriori (EAP) state probabilities in the best fitting model. The information included in these probabilities differs from the information in the residuals z_{pi} in the sense that the EAP state probabilities also directly include information about the responses (correct or false) and that they incorporate the restrictions introduced in the Markov-dependent item

Table 6. Mean, standard deviation (*SD*), and root mean squared error (RMSE) of the correlation between θ and τ (i.e., $\rho_{\theta\tau}$) in the static model without item states for different true models for N = 500. The value of $\rho_{\theta\tau}$ equals 0.7 in all true models.

True model	$\pi_{1\|0}$	Mean	*SD*	RMSE
Static	—	0.69	0.05	0.05
Independent	—	0.67	0.06	0.06
	—	0.68	0.06	0.06
	—	0.67	0.06	0.07
Markov-dependent	0.15	0.45	0.06	0.25
		0.39	0.05	0.31
		0.35	0.05	0.36
	0.3	0.58	0.05	0.13
		0.54	0.05	0.17
		0.50	0.05	0.21

Table 7. Model fit results for Application 1.

Model	npar	ℓ	BIC	AIC	AIC3	CAIC	saBIC
1: Static	114	−12,787	26,254	25,802	25,916	26,368	25,892
2a: Full Independent	172	−12,474	25,973	25,291	25,463	26,145	25,427
2b: Uniform difference in αsi	145	−12,497	25,859	25,285	25,430	**26,004**	25,399
2c: Uniform difference in βsi	145	−12,547	25,958	25,384	25,529	26,103	25,498
2d: Uniform difference in αsi and βsi	118	−12,629	25,962	25,494	25,612	26,080	25,588
3a: Full Markov dependent	174	−12,468	25,973	25,284	25,458	26,147	25,421
3b: Uniform difference in αsi	147	−12,490	**25,857**	**25,274**	**25,421**	**26,004**	**25,390**
3c: Uniform difference in βsi	147	−12,539	25,955	25,373	25,520	26,102	25,489
3d: Uniform difference in αsi and βsi	120	−12,620	25,956	25,480	25,600	26,076	25,575

Note. BIC = Bayesian information criterion; AIC = Akaike information criterion; AIC3 = AIC with a penalty weight of 3 instead of 2; CAIC = consistent AIC; saBIC = sample-size-adjusted BIC; npar = number of parameters in the model; ℓ = value of the log marginal likelihood function at the solution. The smallest values are in boldface.

states model (in this case the fast and slow restriction imposed in δ).

Results

The model fit results are in Table 7. As can be seen, all fit indices indicate Model 3b as the best fitting model (a Markov-dependent item states model with uniform difference in α_{si} between the states and separate β_{si} parameters in each state). Note that although the CAIC for this model is equal to the value of Model 2b, all other fit indices favor Model 3b. We therefore select Model 3b as the best fitting model. For this model, the estimate of $\Delta\alpha$ (which represents the uniform difference between α_{0i} and α_{1i}) is equal to 0.02 (SE = 0.21) indicating that the faster and slower responses do not differ in their discrimination. In addition, the difference in mean speed between the two states, δ_1, was estimated to be 0.53 (SE = 0.02). For the initial state and transition parameters in Model 3b, see Table 8. As can be seen from the initial probabilities, the slow state is larger at the first item. However, from the transition probabilities, it can be seen that the slow state is relatively unstable, and a large portion of the respondents switch to the fast state during the test. In Figure 6 (top), the marginal probability of a correct response, $P(X_{pi} = 1|C_{pi})$, is depicted for the fast (s = 1) and slow (s = 0) states. The responses in the fast state are associated with larger probabilities of a correct response as compared to the responses in the slow state. In Figure 6 (bottom), it is illustrated how the difference in marginal probability in the different states is related to the violation of local independence between the responses and the response times. To this end, we estimated the residual correlations between the responses and response times of the same item using weighted least squares estimation in Mplus (Muthén & Muthén, 2007) and plotted these against the difference in $P(X_{pi}|C_{pi})$ in the slow state and the fast state. As can be seen, most items are associated with negative residual correlations, indicating that the faster responses (smaller $\ln T_{pi}$) are associated with higher probability of a correct response.

In Figures 7 and 8, the raw and standardized residual log-response times are displayed for four respondents together with the estimated slow-state probabilities on all items (based on Model 3b). First, in Figure 7, two examples are given for respondents that speed up during the testing. As appears from the figure, the speeding-up effect can hardly be seen from the raw response times as the item differences in ν_i mask the effect. From the standardized

Table 8. Parameter estimates (standard errors) of the initial state probabilities, π_s, and the transition probabilities, $\pi_{s|r}$.

		s = 0	s = 1
π_s		0.77 (0.12)	0.23 (0.12)
$\pi_{s|r}$	r = 0	0.26 (0.02)	0.74 (0.02)
	r = 1	0.22 (0.01)	0.78 (0.01)

Note. Standard errors are obtained from the standard errors of π_s' and $\pi_{s|r}$' using the delta method.

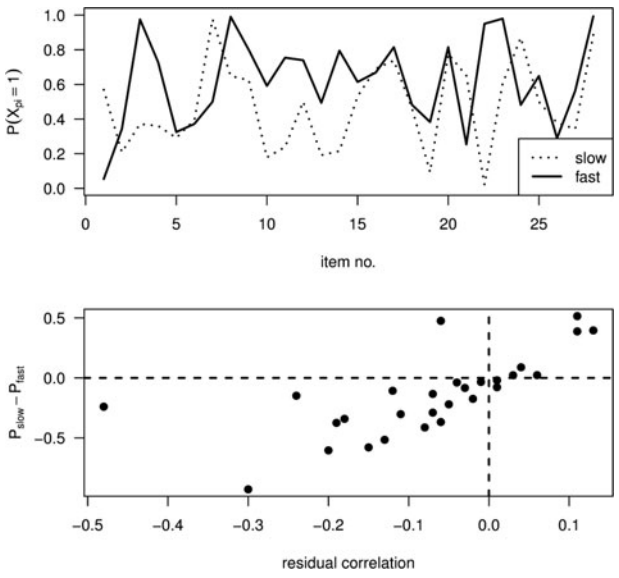

Figure 6. Top: Plot of the marginal probability of a correct response for the fast and slow state for each of the items. Bottom: The difference in the marginal probability of a correct response between the states, denoted by $P_{slow} - P_{fast}$, as a function of the residual correlation as estimated in the static model without item states.

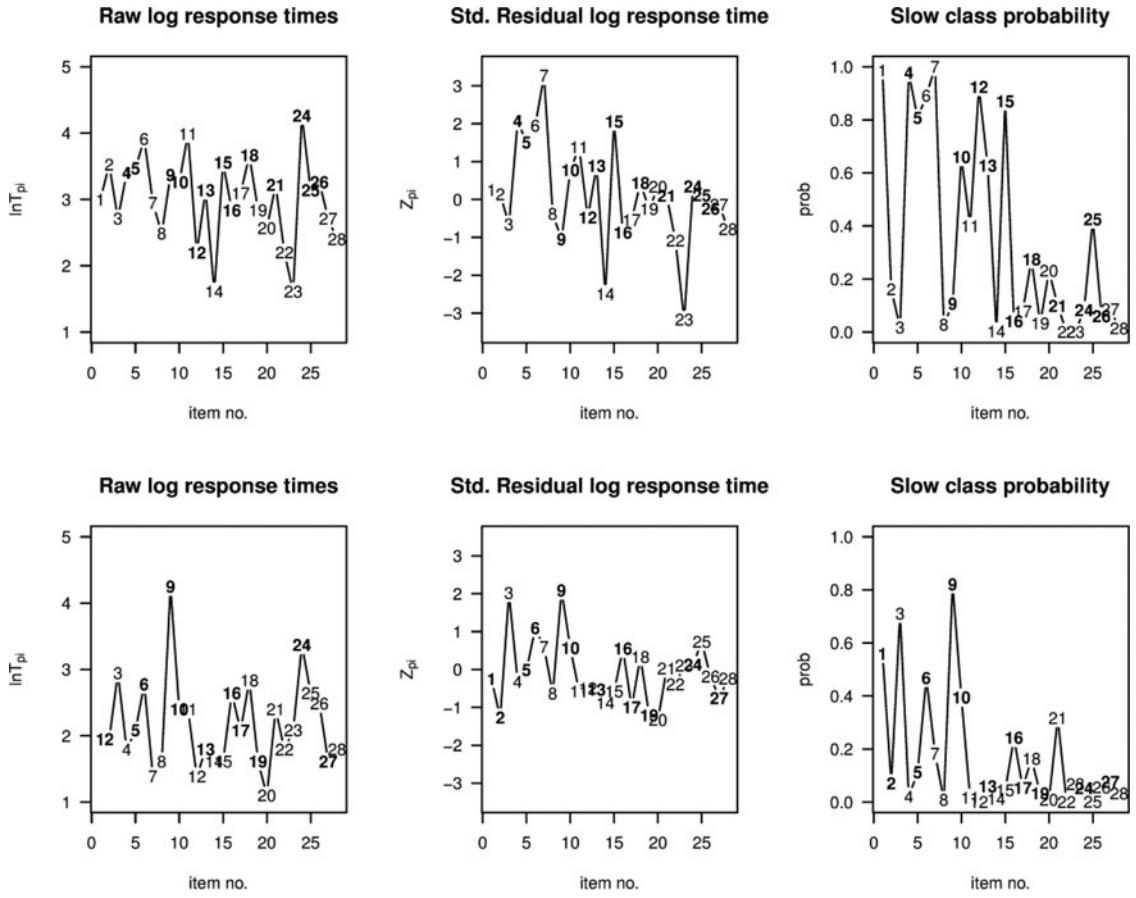

Figure 7. The estimated probability of a slow state response for two example respondents who show speeding up. Item numbers in bold indicate that the response was incorrect.

residuals, the effect is noticeable for the first respondent (top of the figure) but not so much for the second respondent (bottom figure). A statistical test might be needed to test the presence of the effect. From the state probabilities, however, the effect is clear, with the responses to the first half of the test items being generally more probable in the slow state than in the fast state.

In Figure 8, two other example respondents are given that have aberrances in their responding. For the first respondent (top), item 23 stands out for the residual response time that is not evident from the raw response times. This item stands out also for the class probabilities. However, the class probabilities also indicate that the response to item 1 has a large probability of being in the slow state. This is not apparent for z_{pi}, as for the residuals, item 1 is about average. The difference between the class probability and the residual for item 1 is due to the class probability taking into account that the response to item 1 has been correct. As can be seen from Figure 6, for item 1, the slow state is associated with a higher probability of correct. Therefore, the fact that the respondent did item 1 correctly increases the probability of the response being from the slow state. A similar example can be found for the second respondent (bottom). Judged by the standardized residuals, item 17 also stands out, which is also evident from the class probabilities. However, judged by the class probabilities, item 10 also stands out, which is not evident from the standardized residuals. Thus, the results from the Markov-dependent states model can be a valuable addition to the standardized residual method in detecting differences in the response process within subjects.

Application 2: Detecting within-subject differences in solution strategies

Our model can be used not only to "discover" differing response strategies, as demonstrated in the previous section, but also to investigate and test psychological theories about previously hypothesized strategies. The present application demonstrates this confirmatory feature according to a theory by Siegler (1981; see also Van der Maas & Jansen, 2003).

Data

We analyzed the balance scale task data of Van der Maas and Jansen (2003). These data comprise the responses and response times of 191 respondents (mean age 11.84, minimum 6, maximum 25) to 76 balance scale items.

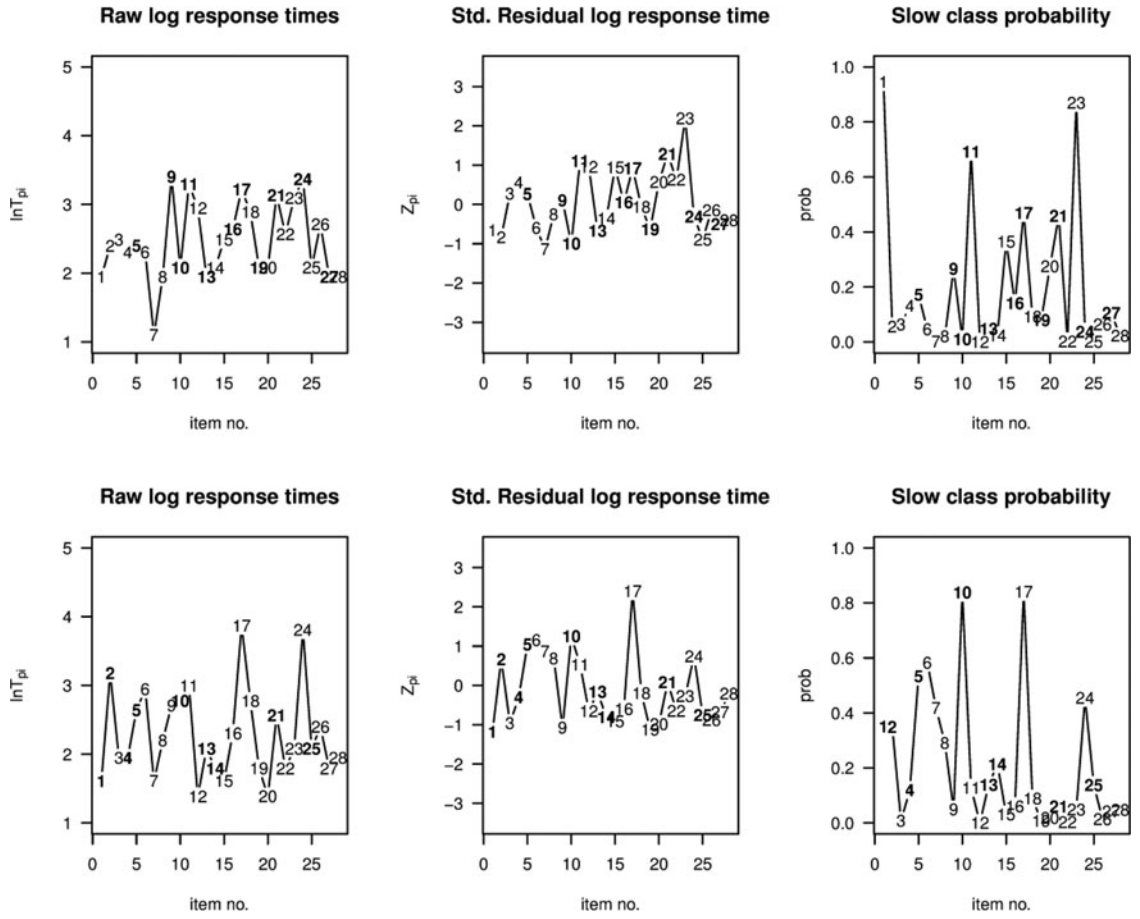

Figure 8. The estimated probability of a slow state response for two example respondents who show aberrances in their responding. Item numbers in bold indicate that the response was incorrect.

Each item displayed a picture of a balance scale with equally heavy weights placed at pegs situated at equal distance from the fulcrum. The items differed in how many weights are placed at each arm and at which pegs the weights were placed. The numbers and distances of the weights are altered according to eight different schemes resulting in eight different item types ("simple balance," "simple weight," "simple distance," "conflict balance A," "conflict balance B," "conflict weight," "conflict distance," and "weight-distance"). For each item type, 10 items exist except for the "weight-distance" scheme, where only 6 items exist.

The data were analyzed previously by Van der Maas and Jansen (2003) using cluster analysis and regression analysis and by Molenaar et al. (2015a) using generalized linear latent variable modeling. Both studies focused on the differences between respondents in their use of solution strategies. In this article, we investigate whether there are differences within respondents in their solution strategies. That is, do respondents consistently apply the same solution strategy to all items, or do they switch between different solution strategies?

Modeling

Van der Maas and Jansen (2003) discussed five solution strategies of the balance scale items that can be derived from the theory of Siegler (1981). Each strategy has different predictions about the proportion correct across the eight different item types (see Table 9). The strategies are ordered according to their complexity. That is, strategy I is considered the least complex strategy as it involves the least number of steps, while strategy V is the most complex strategy as it involves all steps necessary to solve all items correctly. Besides predictions about the correct

Table 9. Predicted item score when using one of the strategies for each item type.

Item type	Strategy I	Strategy II	Strategy III	Strategy IV	Strategy V
Simple balance	1	1	1	1	1
Simple weight	1	1	1	1	1
Simple distance	0	1	1	1	1
Conflict balance A	0	0	1/3	1	1
Conflict balance B	0	0	1/3	0	1
Conflict distance	0	0	1/3	1	1
Conflict weight	1	1	1/3	0	1
Weight-distance	1	1	1	1	1

Note. 0 = incorrect; 1 = correct; 1/3 = guess.

Table 10. Parameter configuration for the latent class response model.

		Strategy				
Item type	*n*	I	II	III	IV	V
Simple balance	10	β_{11}	β_{11}	β_{11}	β_{11}	β_{11}
Simple weight	10	β_{12}	β_{12}	β_{12}	β_{12}	β_{12}
Simple distance	10	$-\beta_{13}$	β_{13}	β_{13}	β_{13}	β_{13}
Conflict balance A	10	$-\beta_{14}$	$-\beta_{14}$	−0.66*	β_{14}	β_{14}
Conflict balance B	10	$-\beta_{15}$	$-\beta_{15}$	−0.66*	$-\beta_{15}$	β_{15}
Conflict distance	10	$-\beta_{16}$	$-\beta_{16}$	−0.66*	β_{16}	β_{16}
Conflict weight	10	β_{17}	β_{17}	−0.66*	$-\beta_{17}$	β_{17}
Weight-distance	6	β_{18}	β_{18}	β_{18}	β_{18}	β_{18}

*This value is fixed to reflect guessing; that is, $P(X_{pi} = 1|C_{pi}) = \omega(\beta_{si}) = \omega(-0.66) \approx 0.34$.

proportion, the different strategies also differ in their predicted response times. That is, the more complex a strategy, the more time children need to apply it as it involves more steps. Molenaar et al. (2015a) translated these predictions into constraints in a latent variable model with a categorical ability factor and a continuous speed factor. The model is however solely between subjects. Here, we use the predictions discussed in the preceding to identify a Markov-dependent item states model with five item states where each item state represents a solution strategy (see Table 10). As the number of items is relatively large compared to the number of respondents, we assume that items configured according to the same scheme (e.g., the simple balance scheme) have equal easiness. In addition, as the theory predicts that more complex strategies require more time, we released the constraint on δ_s (i.e., the constraint that $\delta_0 \leq \delta_1 \cdots \delta_{K-1}$). We did consider the constraint $\delta_0 \geq \delta_1 \geq \cdots \geq \delta_4$, which is in line with the theoretical predictions; however, this model did not converge due to a clear violation of this constraint (which will be shown in the following). We therefore estimated $\delta_1, \ldots, \delta_4$ freely (as $\delta_0 = 0$ for identification purposes).

According to the theory by Siegler (1981), in this application, the balance scale items measure a categorical ability (i.e., the solution strategies). If the responses to these items would have been analyzed without the response times, one would fit a five-component latent class measurement model to these data (subject to the constraints as discussed in Table 10). Adding the response times and the Markov structure will not change this: The main ability measured by the items is categorical. Therefore, as both the ability and the states are categorical, they coincide. Therefore, we do not include a main effect for θ as this effect coincides with C_{pi} in this application. Adding a continuous ability variable will not make sense from the theory by Siegler (1981) as this theory does not predict a continuum to underlie the item responses. Thus, the measurement model for the responses in this application equals

$$P\left(\boldsymbol{x}_p|C_{p1}, \ldots, C_{pn}\right) = \prod_{i=1}^{n} \omega(\beta_{si})^{x_{pi}} \omega(-[\beta_{si}])^{1-x_{pi}},$$

where β_{si} are subject to the constraints in Table 10.

The full model including the response, the response times, and the Markov structure may seem numerically demanding because of the many states (K = 5) and many items ($n = 76$); however, the model is highly restricted, containing only 53 parameters (as compared to the full Markov-dependent item states model, which would have contained 461 parameters for K equal to only 2 and $n = 76$). In addition, we carefully checked the results on convergence issues (ill conditioning of the Hessian matrix and extreme standard errors), but we have no reason to doubt the final solution of the models.

Table 11. Model fit results for Application 2.

Model	npar	ℓ	BIC	AIC	AIC3	CAIC	saBIC
Static	33	−13,670	27,514	27,406	27,439	27,547	27,409
Independent	33	−14,807	29,788	29,681	29,714	29,821	29,684
Markov dependent	53	−13,226	**26,729**	**26,557**	**26,610**	**26,782**	**26,562**

Note. BIC = Bayesian information criterion; AIC = Akaike information criterion; AIC3 = AIC with a penalty weight of 3 instead of 2; CAIC = consistent AIC; saBIC = sample-size-adjusted BIC; npar = number of parameters in the model; ℓ = value of the log marginal likelihood function at the solution. The smallest values are in boldface.

Results

We fit the static baseline model (including θ_p and τ_p), the independent item states model, and the Markov-dependent item states model to the data. The results concerning model fit are in Table 11. As can be seen, the Markov-dependent item states model is identified as the best fitting model according to all fit indices. The parameter estimates for the initial state probabilities and the state speed parameter are in Table 12; the parameter estimates for the transition probabilities are in Table 13. As can be seen, Strategy I is highly stable. Children adopting this strategy do not change to a different strategy. On the contrary, Strategy III is relatively unstable; children adopting this strategy are likely to switch to a different strategy. From the estimates for δ_s, it can be seen that the predictions by Siegler (1981) about the response times (i.e., that

Table 12. Parameter estimates (standard errors) of the initial state probabilities π_s and state speed parameter δ_s.

Strategy	I	II	III	IV	V
π_s	0.20 (0.04)	0.17 (0.06)	0.06 (0.03)	0.00 (0.00)	0.57 (0.07)
δ_s	0*	−0.22 (0.03)	−1.14 (0.03)	−0.19 (0.02)	−0.22 (0.03)

Note. Standard errors are obtained from the standard errors of π_s' using the delta method.
*This parameter is fixed for identification purposes.

Table 13. Parameter estimates (standard errors) of transition probabilities $\pi_{s|r}$.

	C_{pi}	I	II	III	IV	V
$C_{p(i-1)}$	I	0.99 (0.00)	0.00 (0.00)	0.01 (0.00)	0.00 (0.00)	0.00 (0.00)
	II	0.01 (0.00)	0.95 (0.01)	0.04 (0.01)	0.00 (0.00)	0.00 (0.00)
	III	0.07 (0.01)	0.19 (0.03)	0.28 (0.02)	0.32 (0.03)	0.15 (0.03)
	IV	0.00 (0.00)	0.00 (0.00)	0.08 (0.01)	0.78 (0.01)	0.13 (0.01)
	V	0.01 (0.01)	0.00 (0.00)	0.15 (0.02)	0.26 (0.03)	0.58 (0.03)

Note. Standard errors are obtained from the standard errors of $\pi_{s|r}'$ using the delta method.

the more complex strategies have larger response times and thus smaller δ_s) only hold partly. That is, Strategies I, II, and III are indeed decreasing in δ_s; however, strategies IV and V require approximately as much time as Strategy II.

Discussion

We presented a hidden Markov IRT modeling approach for responses and response times. In this model, respondents are assumed to switch between different item states from item to item. The simulation study showed that the proposed model is feasible and yields good parameter recovery. Moreover, the example application to the intelligence data demonstrated how our approach is useful to explore differing response strategies, while the application to the balance scale task demonstrated its use for testing psychological theories regarding response strategies.

A dynamic model without Markov dependencies between the item states was shown to be less successful in detecting dynamic aspects in the response process. However, it should be noted that in the simulation study, the difference in item easiness between the fast and slow states was minor. Therefore, residual correlations between the responses and response times were only around .02–.03, which is very small. For larger differences in item easiness between the item states, the hit rates will be larger. Judged by the results of the simulation study, the hidden Markov model is viable. However, in order to ensure identification of the model, we assumed that the transition probabilities are homogenous over time. This assumption should ideally be tested, which is possible in principle in the current modeling framework. However, as the resulting model with time heterogeneity will include n x n transition parameters, maximizing the resulting likelihood function will be a challenging endeavor. A feasible ad hoc approach might be to test the assumption on the residual response times (z_{pi}, see application 2) and the residual responses (obtained in a similar way as z_{pi}) simultaneously. The advantage is that no measurement model parameters need to be estimated, only the $(n-1) \times K \times (K-1)$ transition parameters.

An implication of the assumed Markov structure for the item states is that the items should be administered in the same order for all respondents. Therefore, the present model cannot be applied to adaptive test data. However, for such applications, the independent item states model will constitute a suitable alternative as it does not assume a structure among subsequent items.

In the simulation study, we compared the dynamic models to a static model with local independence between the responses and response times. It would be interesting to see how the fit of the dynamic models would compare to the fit of a static model with residual correlations (as applied with weighted least squares to the data of the application). However, we did not do this because the static model with residual correlations is not yet developed within a marginal maximum likelihood framework. This hampers the direct comparison of the static model with residual correlations to the dynamic models presented here.

In the model selection analysis, we manipulated the mean speed between the states, the transition probabilities, and the sample size. However, it should be noted that the number of items will also affect the ability to distinguish between the different models, with more items resulting in better separable models. In addition, we investigated model selection only by considering information-criteria-based fit indices such as the AIC and BIC. This has the disadvantage that the difference in likelihood of the different models is not taken into account. That is, the size of the AIC differences between the models is not used to calculate the hit rate. The only information taken into account is which of the models has the smallest fit index. As the models considered in this article are all nested, it would be interesting to see how likelihood-based (bootstrap) methods will perform (Feng, & McCulloch, 1996; Gudicha et al., 2015; McLachlan, 1987). These methods do take the difference in likelihood between two models into account by calculating the power to reject the more constrained model in favor of the less constrained model. Finally, researchers may be interested in relating the estimated latent states of respondents to external covariates such as age, teacher, or educational program. The extension of our model to investigate such research questions is straightforward although its application remains a topic for future study.

Finally, a remark can be made about the applicability of the methodology presented in this article to test model misspecification in IRT. That is, the item states model can be applied as a means to study person misfit (e.g., Reise, 2000) or item misfit. For example, in a given sample, if for some but not all respondents the two-parameter model is violated, the deviating respondents may form a separate state on the items due to different properties of their responses and response times. These respondents can then be detected using posterior class probabilities. In addition, if for some but not all

items the two-parameter model is violated (e.g., a three-parameter model holds for these items), the malfunctioning items can be detected by considering the differences in item characteristics between the states. If these differences are large, the item can be considered to misfit the two-parameter model.

Article information

Conflict of Interest Disclosures: Each author signed a form for disclosure of potential conflicts of interest. JV reports personal fees (royalties for contributions to the LatentGOLD software package) from Statistical Innovations, outside the submitted work.

Ethical Principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was supported by the following grants from the Dutch Organization for Scientific Research (NWO): VENI-451-15-008 awarded to DM, VENI-451-14-017 awarded to DO, and VICI 453-10-002 awarded to JV.

Role of the Funders/Sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Acknowledgments: The authors thank Han van der Maas and Harrie Vorst for providing the data used in the applications. The ideas and opinions expressed herein are those of the authors alone, and endorsement by the authors' institutions is not intended and should not be inferred.

References

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

Amthauer, R., Brocke, B., Liepmann, D., & Beauducel, A. (2001). *Intelligenz-struktur-test 2000 R*. Göttingen, Germany: Hogrefe.

Bacci, S., Pandolfi, S., & Pennoni, F. (2014). A comparison of some criteria for states selection in the latent Markov model for longitudinal data. *Advances in Data Analysis and Classification*, *8*(2), 125–145. doi:10.1007/s11634-013-0154-2

Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, *41*(1), 164–171.

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

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459. doi:10.1007/BF02293801

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*(3), 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* (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. doi:10.1037/0033-295X.97.3.404

Celeux, G., & Durand, J. B. (2008). Selecting hidden Markov model state number with cross-validated likelihood. *Computational Statistics*, *23*(4), 541–564. doi:10.1007/s00180-007-0097-1

De Boeck, P. (2008). Random item IRT models. *Psychometrika*, *73*(4), 533–559. doi:10.1007/s11336-008-9092-x

Feng, Z. D., & McCulloch, C. E. (1996). Using bootstrap likelihood ratios in finite mixture models. *Journal of the Royal Statistical Society. Series B*, *58*, 609–617.

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

Furneaux, W. D. (1961). Intellectual abilities and problem solving behavior. In H. J. Eysenck (Ed.), *The handbook of abnormal psychology*. London, UK: Pitman.

Glas, C. A., & Linden, W. J. (2010). Marginal likelihood inference for a model for item responses and response times. *British Journal of Mathematical and Statistical Psychology*, *63*, 603–626. doi:10.1348/000711009X481360

Goldhammer, F., & Kroehne, U. (2014). Controlling Individuals' Time Spent on Task in Speeded Performance Measures: Experimental Time Limits, Posterior Time Limits, and Response Time Modeling. *Applied Psychological Measurement*, *38*, 255–267. doi:10.1177/0146621613517164

Goldhammer, F. (2015). Measuring Ability, Speed, or Both? Challenges, Psychometric Solutions, and What Can Be Gained from Experimental Control. *Measurement: Interdisciplinary Research and Perspectives*, *13*(3–4), 133–164. doi:10.1080/15366367.2015.1100020

Greven, S., & Kneib, T. (2010). On the behaviour of marginal and conditional AIC in linear mixed models. *Biometrika*, *97*, 773–789. doi:10.1093/biomet/asq042

Gudicha, D. W., Schmittman, V. D., Tekle, F. B., & Vermunt, J. K. (2015). Power Analysis for the Likelihood-Ratio Test in Latent Markov Models: Short-cutting the bootstrap p-value based method. *Manuscript submitted for publication*.

Hamaker, E. L., Nesselroade, J. R., & Molenaar, P. C. (2007). The integrated trait—state model. *Journal of Research in Personality*, *41*(2), 295–315. doi:10.1016/j.jrp.2006.04.003

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

Kempf, W. (1977). Dynamic models for the measurement of 'traits' in social behavior. In W. Kempf, B. H. Repp (Eds.), *Mathematical models for social psychology* (pp. 14–58). New York, NY: Wiley.

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*, 21–48. doi:10.1007/s11336-008-9075-y

Klein Entink, R. H., Kuhn, J. T., Hornke, L. F., & Fox, J. P. (2009). Evaluating cognitive theory: a joint modeling approach using responses and response times. *Psychological Methods*, *14*(1), 54–75. doi:10.1037/a0014877

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*(2), 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). London, UK: Chapman & Hall.

Maris, G., & van der Maas, H. (2012). Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika*, *77*(4), 615–633. doi:10.2307/2347790

McLachlan, G. (1987). On bootstrapping the likelihood ratio test stastistic for the number of components in a normal mixture. *Applied Statistics*, *36*, 318–324. doi:10.2307/2347790

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.

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

Molenaar, D. (2015). The value of response times in item response modeling. *Measurement: Interdisciplinary Research and Perspectives*, *13*(3–4), 177–181. doi:10.1080/15366367.2015.1105073

Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. J. (2015a). 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. doi:10.1111/bmsp.12042

Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. J. (2015b). A bivariate generalized linear item response theory modeling framework to the analysis of responses and response times. *Multivariate Behavioral Research*, *50*(1), 56–74. doi:10.1080/00273171.2014.962684

Molenaar, P. C. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement*, *2*(4), 201–218. doi:10.1207/s15366359mea0204_1

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

Muthén, L. K., & Muthén, B. O. (2007). *Mplus user's guide* (3rd ed.). Los Angeles, CA: Muthén & Muthén.

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

Rabbitt, P. (1979). How old and young subjects monitor and control responses for accuracy and speed. *British Journal of Psychology*, *70*, 305–311. doi:10.1111/j.2044-8295.1979.tb01687.x

Ranger, J. (2013). Modeling responses and response times in personality tests with rating scales. *Psychological Test and Assessment Modeling*, *55*, 361–382.

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., & Kuhn, J. T. (2013). Analyzing response times in tests with rank correlation approaches. *Journal of Educational and Behavioral Statistics*, *38*(1), 61–80. doi:10.3102/1076998611431086

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

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

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danmarks Paedogogiske Institut.

Reise, S. P. (2000). Using multilevel logistic regression to evaluate person-fit in IRT models. *Multivariate Behavioral Research*, *35*, 543–568. doi:10.1207/S15327906MBR3504_06

Samejima, F. (1969). Estimation of ability using a response pattern of graded scores. *Psychometrika Monograph* (vol. 17). Richmond, VA: The Psychometric Society.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*, 461–464. doi:10.1214/aos/1176344136

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

Siegler, R. S. (1981). Developmental sequences within and between concepts. *Monographs of the Society for Research in Child Development*, *46*(2).

Steinmayr, R., Bergold, S., Margraf-Stiksrud, J., & Freund, P. A. (2015). Gender differences on general knowledge tests: Are they due to Differential Item Functioning?. *Intelligence*, *50*, 164–174. doi:10.1016/j.intell.2015.04.001

Thissen, D. (1983). Timed testing: An approach using item response theory. In D. J. Weiss (Ed.), New horizons in

testing: Latent trait test theory and computerized adaptive testing (pp. 179–203). New York, NY: Academic Press.

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. (2009a). Conceptual issues in response-time modeling. *Journal of Educational Measurement*, *46*, 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., Breithaupt, K., Chuah, S. C., & Zhang, Y. (2007). Detecting differential speededness in multistage testing. *Journal of Educational Measurement*, *44*(2), 117–130. doi:10.1111/j.1745-3984.2007.00030.x

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 Linden, W. J., Scrams, D. J., & Schnipke, D. L. (1999). Using response-time constraints to control for differential speededness in computerized adaptive testing. *Applied Psychological Measurement*, *23*(3), 195–210. doi:10.1177/01466219922031329

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

Van der Maas, H. L., Molenaar, D., Maris, G., Kievit, R. A., & Borsboom, D. (2011). Cognitive psychology meets psychometric theory: on the relation between process models for decision making and latent variable models for individual differences. *Psychological review*, *118*(2), 339–356. doi:dx.doi.org/10.1037/a0022749

Verhelst, N. D., & Glas, C. A. (1993). A dynamic generalization of the Rasch model. *Psychometrika*, *58*(3), 395–415. doi:10.1007/BF02294648

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. doi:10.3102/10769986024002179

Vermunt J. K., & Magidson, J. (2013). *Technical guide for latent GOLD 5.0: Basic, advanced, and syntax*. Belmont, MA: Statistical Innovations Inc.

Vermunt, J. K., Tran, B., & Magidson, J. (2008). Latent class models in longitudinal research. In S. Menard (eds.), *Handbook of longitudinal research: design, measurement, and analysis* (pp. 373–385). Burlington, MA: Elsevier.

Visser, I., Raijmakers, M. E., & Molenaar, P. (2002). Fitting hidden Markov models to psychological data. *Scientific Programming*, *10*(3), 185–199.

Wang, X., Berger, J. O., & Burdick, D. S. (2013). Bayesian analysis of dynamic item response models in educational testing. *The Annals of Applied Statistics*, *7*(1), 126–153. doi:10.1214/12-AOAS608

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