Behavior Research Methods (2019) 51:676–696
https://doi.org/10.3758/s13428-019-01229-x

A heteroscedastic hidden Markov mixture model for responses and categorized response times

Dylan Molenaar[1] · Sandor Rózsa[2] · Maria Bolsinova[3]

Published online: 28 March 2019

Abstract

Various mixture modeling approaches have been proposed to identify within-subjects differences in the psychological processes underlying responses to psychometric tests. Although valuable, the existing mixture models are associated with at least one of the following three challenges: (1) A parametric distribution is assumed for the response times that—if violated—may bias the results; (2) the response processes are assumed to result in equal variances (homoscedasticity) in the response times, whereas some processes may produce more variability than others (heteroscedasticity); and (3) the different response processes are modeled as independent latent variables, whereas they may be related. Although each of these challenges has been addressed separately, in practice they may occur simultaneously. Therefore, we propose a heteroscedastic hidden Markov mixture model for responses and categorized response times that addresses all the challenges above in a single model. In a simulation study, we demonstrated that the model is associated with acceptable parameter recovery and acceptable resolution to distinguish between various special cases. In addition, the model was applied to the responses and response times of the WAIS-IV block design subtest, to demonstrate its use in practice.

Keywords Mixture models · Item response theory · Response times · Hidden Markov models

In psychological and educational measurement of constructs and abilities, within-subjects differences may exist in the psychological processes that resulted in the responses to the items of the test. For instance, respondents may resort to fast guessing on some of the items of an educational measurement test but use a regular response process on the other items (Schnipke & Scrams, 1997); respondents may alternate between memory retrieval and actual calculation on the items of an arithmetic test (Grabner et al., 2009); or they may use trial and error on some items of a spatial puzzle but use an analytical strategy on others (Goldstein & Scheerer, 1941).

The objective of this article is to improve on existing statistical methods to detect these within-subjects differences in response processes. In psychological and educational measurement, the dominant source of information are the item responses themselves, which indicate the accuracy of the underlying response process. In this article, we will additionally focus on the item response times as a valuable additional source of information concerning the response process as they indicate the amount of time it took for the response processes to be executed (Luce, 1986). That is, everything else being equal, a systematic difference in response time suggests a difference in the underlying response process.

Various psychometric modeling approaches based on mixture modeling have been proposed that—in addition to the item responses—use the response times to identify within-subjects differences in response processes (Molenaar, Oberski, Vermunt, & De Boeck, 2016; Schnipke & Scrams, 1997; Wang & Xu, 2015; Wang, Xu, & Shang, 2018). However, although valuable, the existing mixture models are associated with at least one of the following three challenges: (1) A parametric distribution is assumed for the response times that—if violated—may bias the results; (2) the response processes are assumed to result in equal variances (homoscedasticity) in the response times, whereas some processes may produce more variability than others (heteroscedasticity; e.g., fast guessing is commonly associated with less variance than the regular response process); and (3) the different response processes are modeled as independent

✉ Dylan Molenaar
D.Molenaar@uva.nl

[1] Department of Psychology, University of Amsterdam, Amsterdam, The Netherlands

[2] Washington University School of Medicine, St. Louis, MO, USA

[3] ACTNext, Iowa City, IA, USA

latent variables, whereas they may be related (e.g., after a guess, a subject may be more likely to guess on the next item).

Challenges 1, 2, and 3 have all been studied separately. That is, Challenge 1 has been addressed by Molenaar, Bolsinova, and Vermunt (2018), who proposed a mixture modeling approach based on the categorized response times to avoid assumptions about the specific parametric shape of the response time distribution. The approach was demonstrated to perform better than a parametric approach based on the log-normal response time distribution if the observed response time distribution departs from log-normality. In addition, Challenge 2 has been addressed by Wang and Xu (2015) and Wang et al. (2018), who proposed a model for two response processes, fast guessing and a regular solution process, in which the processes were heteroscedastic, that is, associated with differences in the underlying response time variance. Finally, Challenge 3 has been addressed by Molenaar et al. (2016), who modeled the possible relation between the response processes underlying two subsequent items using a time homogeneous hidden Markov process of order one.

Although the three challenges above have been addressed separately, in practice they may occur simultaneously. In the present article, we therefore propose a heteroscedastic hidden Markov mixture model for responses and categorized response times in which we explicitly address Challenges 1, 2, and 3 in a joint model. That is, we combine the categorized response time approach of Molenaar et al. (2018), the heteroscedastic response processes approach by Wang and Xu (2015) and Wang et al. (2018), and the Markov process approach of Molenaar et al. (2016) in a single model. The outline is as follows: First, the full model is derived and tested in a simulation study to investigate parameter recovery and the resolution to distinguish between different special cases. Next, the model is applied to a real dataset to demonstrate its use in practice.

The general mixture framework

A joint modeling approach

Within traditional item response theory models, it is assumed either that the item responses to psychometric tests are the results of a single response process (e.g., an information accumulation process; see Tuerlinckx & De Boeck, 2005; van der Maas, Molenaar, Maris, Kievit, & Borsboom, 2011) or that the response processes are homogeneous (e.g., multiple processes underlie the scores of an arithmetic test, such as subtraction and addition, but these processes are homogeneous in the sense that, statistically, they are commonly unidimensional). As a result, between-subjects differences in the accuracy of these response processes can be modeled by posing a latent ability variable, θ_p, to underlie the item responses of respondent $p = 1, \ldots, N$ to a test. Similarly, individual differences in the speed with which these processes are executed can be captured by posing a latent speed variable, τ_p, to underlie the response times to a test.

A joint psychometric model for responses and response times was proposed by van der Linden (2007). In this model, commonly referred to as "the hierarchical model," the joint density of the responses, x_{pi}, and the response times, t_{pi}, of respondent p on item $i = 1, \ldots, n$, conditional on θ_p and τ_p is denoted by $d(x_{pi}, t_{pi}| \theta_p, \tau_p) = f(x_{pi}, t_{pi}| \theta_p, \tau_p)$. By assuming that the responses and response times are independent conditional on θ_p and τ_p (see, e.g., van der Linden, 2007; van der Linden & Glas, 2010), this conditional density can be factored into a separate response part, and a separate response time part, that is,

$$f\left(x_{pi}, t_{pi} \mid \theta_p, \tau_p\right) = g\left(x_{pi} \mid \theta_p\right) \times h\left(t_{pi} \mid \tau_p\right), \tag{1}$$

where $g(.)$ denotes the conditional probability mass function of the responses, and $h(.)$ denotes the conditional density function of the response times.

Because psychometric test items commonly differ in the properties with which they measure the underlying processes, a model is specified for $g(x_{pi} \mid \theta_p)$ and $h(t_{pi} \mid \tau_p)$ in order to separate item effects and respondent effects on the responses and response times, respectively (e.g., some items are more difficulty and some respondents are faster). For instance, the three-parameter logistic item response theory model is given by

$$g\left(x_{pi}|\theta_p\right) = P\left(x_{pi} = 1|\theta_p\right)^{x_{pi}}\left[1-P\left(x_{pi} = 1|\theta_p\right)\right]^{1-x_{pi}}, \tag{2}$$

with the probability of a correct response given by

$$P\left(x_{pi} = 1|\theta_p\right) = \gamma_i + (1-\gamma_i)\omega\left(\alpha_i\theta_p + \beta_i\right), \tag{3}$$

where $\omega(.)$ is a logistic or normal ogive function, and γ_i, α_i, and β_i are the item parameters. Specifically, γ_i is a lower-asymptote parameter that accounts for correct responses due to guessing, α_i is a discrimination parameter that accounts for the degree to which the item captures differences in θ_p, and β_i is an easiness parameter that accounts for the proportion correct of the item. In Fig. 1 (left) is illustrated, for three example items, how these parameters affect the probability of a correct response, $P(x_{pi} = 1 \mid \theta_p)$ in Eq. 3. Important for the assessment of between-subjects differences in the latent ability variable is the concept of "information." That is, depending on the measurement properties of the item, an item can be more informative about θ_p for specific levels on the θ_p range. Similarly, the test as a whole does not necessarily provide an equal amount of information for each level of θ_p. See Fig. 1 (middle) for the

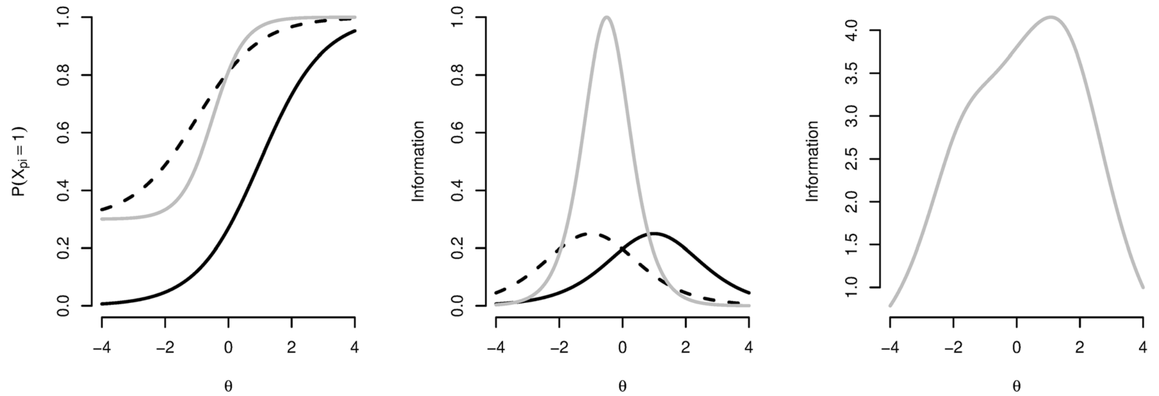

Fig. 1 (Left) Probabilities of a correct response as a function of θ_p, $P(X_{pi} = 1 \mid \theta_p)$, for different parameter configurations. (Middle) The resulting item information as a function of θ_p. Solid black line: $\alpha_i = 1$, $\beta_i = -1$, $\gamma_i = 0$; dashed black line: $\alpha_i = 1$, $\beta_i = 1$, $\gamma_i = 0$; gray line: $\alpha_i = 2$, $\beta_i = 1$, $\gamma_i = .3$. (Right) Test information as a function of θ_p for 25 example items (β_is are between − 3 and 3, α_is are between 0.5 and 1.5, and $\gamma_i = 0$)

item information as a function of θ_p for the three example items from Fig. 1 (left). See Fig. 1 (right) for the test information as a function of θ_p for an example test of 25 items.

For the response times, similar approaches exist that separate between the latent speed variable, τ_p, and the measurement properties of the response time variables. For instance, the log-normal model is given by

$$h\left(t_{pi}|\tau_p\right) = \frac{1}{\sigma_i t_{pi}} \varphi\left\{\frac{\ln\left(\mathrm{t_{pi}}\right) - \left(\nu_i - \lambda_i \times \tau_p\right)}{\sigma_i}\right\}, \tag{4}$$

where $\varphi(.)$ is the standard normal distribution function and v_i, λ_i, and σ_i are the item parameters. Specifically, ν_i is an intercept that accounts for the time intensity of the item (i.e., some items require more time irrespective of the difficulty, because of, for instance, a large text that has to read), λ_i is a factor loading that accounts for the degree with which the item captures differences in τ_p, and σ_i is the standard deviation of the residual, which contains measurement error and misfit. As for the responses, the model has implications for the information about τ_p in the response times. That is, the information is constant over the τ_p range and only depends on λ_i and σ_i (see Mellenbergh, 1994).

A mixture joint modeling approach

The general idea of the mixture approach by Schnipke and Scrams (1997), Wang and Xu (2015), Wang et al. (2018), and Molenaar et al. (2016) is to model within-subjects differences in response processes by extending the joint model above to include item-specific latent class variables, ζ_{pi}, with two states $c = 0, 1$ to underlie the responses and response times of item i. The two states either correspond to a discrete difference in two qualitative response processes that produce heterogeneity in the data (e.g., memory retrieval and logical reasoning) or the two states correspond to two statistical states that capture heterogeneity in the data that is due to discrete differences in multiple response processes (e.g., multiple solution strategies) or due to continuous differences in one or more response processes (e.g., motivation or fatigue).

If the response processes are indeed heterogeneous, the measurement properties of θ_p and τ_p will be different across states. Therefore, in the general mixture framework, the joint conditional density of the responses, x_{pi}, and the response times, t_{pi}, is a mixture of the joint conditional densities of x_{pi} and t_{pi} within the two states, that is

$$d\left(x_{pi}, t_{pi}|\theta_p, \tau_p\right) = \sum_{\zeta_{pi}=0}^{1} P\left(\zeta_{pi}\right) f_c\left(x_{pi}, t_{pi}|\theta_p, \tau_p, \zeta_{pi}\right) \tag{5}$$

where $f_c(.)$ is the joint density function within state $\zeta_{pi} = c$, and $P(\zeta_{pi})$ is the state probability. Within each state, the responses and response times are still assumed to be independent conditional on θ_p and τ_p, that is

$$f_c\left(x_{pi}, t_{pi}|\theta_p, \tau_p, \zeta_{pi}\right) = g_c\ \left(x_{pi}|\theta_p, \zeta_{pi}\right) \times h_c\ \left(t_{pi} \mid \tau_p, \zeta_{pi}\right), \tag{6}$$

where $g_c(.)$ denotes the conditional probability mass function of the responses in state c and $h_c(.)$ denotes the conditional density function of the response times in state c. In the general mixture framework, for the within-state response time density, the log-normal linear model from Eq. 4 is used as follows

$$h_c\left(t_{pi}|\tau_p, \zeta_{pi} = c\right) = \frac{1}{\sigma_{ci} t_{pi}} \varphi\left\{\frac{\ln\left(\mathrm{t_{pi}}\right) - \left(\nu_{ci} - \lambda_{ci} \times \tau_p\right)}{\sigma_{ci}}\right\}, \tag{7}$$

where the item parameters are allowed to differ across states as indicated by index c. For the responses, the

Table 1 Parameter restrictions in the general mixture framework necessary to obtain special cases from the literature

			Response Times			Responses		
Model	References	c	ν_{ci}	λ_{ci}	σ_{ci}	γ_{ci}	α_{ci}	β_{ci}
Hierarchical model (baseline)	van der Linden (2007)	0	ν_{0i}	1	σ_{0i}	γ_{0i}	α_{0i}	β_{0i}
		1	–	–	–	–	–	–
Standard mixture model	Schnipke and Scrams (1997)	0	ν_{0i}	0	σ_{0i}	–	–	–
		1	ν_{1i}	0	σ_{1i}	–	–	–
Common-guessing mixture model	Schnipke and Scrams (1997)	0	ν_0	0	σ_0	–	–	–
		1	ν_{1i}	0	σ_{1i}	–	–	–
Mixture hierarchical model	Wang and Xu (2015); Wang et al. (2018)	0	ν_0	0	σ_0	0	0	β_{0i}
		1	ν_{1i}	1	σ_{1i}	γ_{ci}	α_{1i}	β_{1i}
Independent-states mixture model	Molenaar et al. (2016)	0	ν_{0i}	1	σ_i	0	α_{0i}	β_{0i}
		1	$\nu_{0i}+\delta_1$	1	σ_i	0	α_{1i}	β_{1i}

"–" denotes that this part of the general model is omitted (i.e., for the hierarchical model by van der Linden, 2007, there is no Class 1 in the model, and for the models by Schnipke & Scrams, 1997, there is no measurement model for the responses)

three-parameter item response theory model from Eq. 3 is used:

$$g_c\left(x_{pi}|\theta_p,\zeta_{pi}=c\right)$$
$$=P\left(x_{pi}=1|\theta_p,\zeta_{pi}=c\right)^{x_{pi}}\left[1-P\left(x_{pi}=1|\theta_p,\zeta_{pi}=c\right)\right]^{1-x_{pi}}, \tag{8}$$

with

$$P\left(x_{pi}=1|\theta_p,\zeta_{pi}=c\right)=\gamma_{ci}+(1-\gamma_{ci})\omega\left(\alpha_{ci}\theta_{pi}+\beta_{ci}\right), \tag{9}$$

where the item parameters are again allowed to differ across states. The framework given by Eqs. 5, 6, 7, 8, and 9 is very general, in the sense that it includes many parameters that are not identified simultaneously and that are yet difficult to interpret. However, various special cases within this general framework have been considered in the literature. See Table 1 for the exact restrictions needed to arrive at these special cases.[1]

From the table it can be seen that the first model, the hierarchical model by van der Linden (2007) discussed above, arises by specifying a log-normal model with λ_{0i} = 1 for the response times, and a three-parameter model for the responses in state 0 and leaving state 1 empty. Because this model assumes a single state only, it corresponds to a single-process model or homogeneous process model that can be used as a baseline in drawing inferences about within-subjects differences in response processes in the data. Note that the factor loadings are constrained to be equal to 1 in the single-state model and in all other models that include τ_p, which is an essentially tau-equivalent factor model (Lord & Novick, 1968). This assumption has been relaxed in the hierarchical model by, for instance Fox, Klein Entink, and van der Linden (2007) and Molenaar, Tuerlinckx, and van der Maas (2015).

The next two models in Table 1 are by Schnipke and Scrams (1997). These models consider response times only. As can be seen, both models do not include a latent speed variable as λ_{ci} = 0 in both states. In the standard mixture model, the intercept and variance are estimated for each item in both states. In the common-guessing mixture model, the intercepts and variances in Class 0 (the guessing class) are restricted to be equal across items. Although these models by Schnipke and Scrams are not latent variable models, to our knowledge, these models have been the first to include a within-subjects mixture component for response times. In addition, the idea of common-guessing has been adopted by Wang and Xu (2015) and Wang et al. (2018), who proposed a common-guessing latent-variable model for both responses and response times. As can be seen in Table 1, the response time model includes a latent speed variable in state 1 (i.e., λ_{1i} = 1) with item-specific intercepts and residual variances, and a common intercept and residual variance in state 0, but without a latent speed variable. In addition, the response model includes a three-parameter latent-variable model for the responses in state 1 and a fast-guessing parameter β_{0i} in state 0 without a latent variable. Finally, Molenaar et al. (2016) proposed a model with a latent speed variable in both states (i.e., λ_{0i} = 1 and λ_{1i} = 1), in which the item-specific intercepts in state 1 are equal to the intercepts of state 0 shifted by a

[1] Note that the restrictions provided in Table 1 result in models equivalent to the models discussed in the text [i.e., equivalent in terms of the likelihood of the model. The exact parameterization in the corresponding articles is for some cases slightly different. For instance, Schnipke and Scrams (1997) estimated $\ln(\nu_{ci})$ instead of ν_{ci}, and Wang and Xu (2015) used $\alpha_{1i}(\theta_p-\beta_{1i})$ in the three-parameter model, instead of $\alpha_{1i}\theta_p+\beta_{1i}$.

common scalar, δ_1. In addition, the residual standard deviation is assumed to be equal across states ($\sigma_{ci} = \sigma_i$). For the responses, a two-parameter model is used in both states ($\gamma_{ci} = 0$).

Challenges and a possible solution

The response time distribution

The mixture approaches discussed above are all associated with one of the following challenges. First, the approaches all assume a log-normal distribution for the response times within the states. As has been argued by Vermunt (2011) for standard mixture models, and demonstrated by Bauer and Curran (2003) for growth mixture models and by Molenaar et al. (2018) for the independent states mixture model in Table 1, violations of the assumed within-states distribution may result in (1) spurious states—that is, states that are not actually in the data but appear as a significant source of variation in the modeling to capture the misfit in the data distribution—and (2) biased true states—that is, differences between true states (that are actually in the data) may seem smaller or larger depending on the source of the misfit in the data distribution (e.g., positive skew or negative skew, truncation, etc.).

In principle, this challenge can be solved by specifying a more appropriate response time distribution within each state. However, commonly there is no theory about the response time distribution within each state. In addition, inferring the within-state response time distribution from the data is difficult, because only the observed distribution of the response times is available, which cannot straightforwardly be used to make inferences about the parametric form of the within-state distribution as the observed response time distribution will depart from the within-state distribution by definition. Kuipers, Visser, and Molenaar (2018) proposed a test on log-normality of the within-state response time distribution. However, if the log-normality assumption fails, the above mixture models are not suitable for the data.

As a solution, Molenaar et al. (2018) proposed to categorize the continuous response times so that the resulting response time distribution could be better captured using category-specific threshold parameters. Specifically, Molenaar et al. (2018) proposed to replace the log-normal linear model above by a partial-credit model (Masters, 1982), which is an adjacent-category model for ordered categories, or any other model for ordered categories (e.g., the graded response model [Samejima, 1969], which is a cumulative probability model). With respect to the categorization of the response times, Molenaar et al. (2018) proposed to use an item-wise categorization procedure using the observed percentiles. For five or seven categories, this approach worked well in terms of both parameter recovery and power.

Dependency between the states

In the general model in Eq. 5, it is assumed that the latent class variables underlying the items, ζ_{pi}, are independent. However, various examples show why the ζ_{pi} variables can be dependent. First, if a respondent guesses on one item, it may be more likely that this respondent will also guess on the next item. A similar example includes response strategies in general. That is, if multiple solution strategies are possible that differ in their efficiency, using an efficient solution strategy on one item will probably increase the probability that this strategy will also be used on the next item. Another example includes posterror slowing (Rabbitt, 1979), which refers to the phenomenon that respondents, who know (or think) that they made an error on a given item, slow down on the next item resulting in a dependency between subsequent ζ_{pi}s.

Within the general mixture framework in Eq. 5, Molenaar et al. (2016) accounted for a possible dependency of the item specific latent class variables of item i, ζ_{pi}, on the item specific latent class variables of item $i - 1$, $\zeta_{p(i-1)}$. That is, in a model for continuous log-normal response times, the assumption of independent ζ_{pi} was relaxed by introducing a first-order Markov structure (e.g., MacDonald & Zucchini, 1997) on ζ_{pi}. Molenaar et al. (2016) showed that the presence of a Markov structure in the data can successfully be detected using fit indices BIC, CAIC, AIC with a triple penalty (AIC3), and the sample-size-adjusted BIC (saBIC). The conventional AIC (which uses a double penalty term) was associated with an increased false positive rate.

Heteroscedasticity between the states

The categorized response time model and the Markov structure thus provide a solution to the spurious-state and independency challenges of the general framework in Eq. 5. However, contrary to Wang and Xu (2015), Wang et al. (2018), and Schnipke and Scrams (1997), both models assume that the within-state response time variance is homoscedastic (equal across states). In the Markov mixture model, this assumption is explicit, as $\sigma_{0i} = \sigma_{1i}$ in the model by Molenaar et al. (2016). In the categorized response time model it is less explicit, since traditional item response theory models do not have a variance parameter. However, the same thresholds are applied in both states to categorize the response times (since the marginal response time distribution is categorized and not the within-state response time distribution, because this distribution is unknown). Therefore, heteroscedasticity across states will not be detected and will bias the results, as we will demonstrate in the simulation study below.

Proposed model

In this article, we thus propose a model that combines the categorized response time model by Molenaar et al. (2018), the Markov model by Molenaar et al. (2016), and the heteroscedastic state model by Wang and Xu (2015), Wang et al. (2018), and Schnipke and Scrams (1997) into a single model. First, to be able to accommodate the general model in Eq. 5 to include a Markov dependence among ζ_{pi}, we need to consider the conditional density of the full vector of responses, $\boldsymbol{x_p} = [x_{p1}, \ldots, x_{pn}]$, and the full vector of categorized response times, $\boldsymbol{t_p}' = [t_{p1}', \ldots, t_{pn}']$, where t_{pi}' denotes the categorized response times, $t_{pi}' = 0, 1, \ldots, T–1$. Next, Eqs. 5 and 6 change into

$$d\left(x_p, t_p^{'}|\theta_p, \tau_p\right) = \sum_{\zeta_{p1}=0}^{1}\cdots\sum_{\zeta_{pn}=0}^{1} P\left(\zeta_{p1}\right)\prod_{i=2}^{n} P\left(\zeta_{pi}|\zeta_{p(i-1)}\right)\prod_{i=1}^{n} g_c\left(x_{pi}|\theta_p, \zeta_{pi}\right) h_c\left(t_{pi}^{'}|\tau_p, \zeta_{pi}\right) \qquad (10)$$

where $P(\zeta_{p1} = 1) = \pi_1$ is the initial state parameter, and $P(\zeta_{pi} = 1| \zeta_{p(i-1)} = 0) = \pi_{1|0}$ and $P(\zeta_{pi} = 1| \zeta_{p(i-1)} = 1) = \pi_{1|1}$ are the transition parameters. Note that $P(\zeta_{p1} = 0)$, $P(\zeta_{pi} = 0| \zeta_{p(i-1)} = 0)$, and $P(\zeta_{pi} = 0| \zeta_{p(i-1)} = 1)$ can be calculated from these parameters. In addition, we assume homogeneity of the Markov structure over items. That is, the transition probabilities are invariant over all items, $P(\zeta_{pi}|\zeta_{p(i-1)}) = P(\zeta_{pj}|\zeta_{p(j-1)})$ for all i and all $j = 1, \ldots, n$. This assumption is common in Markov modeling (e.g., Bacci, Pandolfi, & Pennoni, 2014; Gudicha, Schmittmann, & Vermunt, 2016; Zucchini, MacDonald, & Langrock, 2016, p. 15). Besides being common practice, here, we also assume time homogeneity of the Markov structure to prevent the model from becoming too complex. Including a time non-homogeneous Markov structure would result in two additional parameters for each item [probability of remaining in a class, $P(\zeta_{pi} = 1 \mid \zeta_{p(i-1)} = 1)$, and the probability of switching classes, $P(\zeta_{pi} = 1 \mid \zeta_{p(i-1)} = 0)$] which makes the model very complex. Given that the model from Molenaar et al. (2018) already includes four parameters for each item response variable and $T - 1$ response time category parameters, we did not consider such an extension of the homogeneous Markov structure into a nonhomogeneous Markov structure. However, this extension is straightforward (i.e., in the syntax to fit the model in the Appendix, which will be explained later, we indicate how to drop the time homogeneity assumption). In addition, the assumption of time homogeneity can be statistically tested (see, e.g., Tan & Yılmaz, 2002).

Next, for the conditional probability function of the categorized response times, $h_c(.)$, we use the partial credit model subject to heteroscedasticity (Hedeker, Berbaum, & Mermelstein, 2006), as follows:

$$h_c\left(t_{pi}^{'} = t|\tau_p, \zeta_{pi} = c\right) = P\left(t_{pi}^{'} = t|\tau_p, \zeta_{pi} = c\right) = \frac{\exp\left(\sum_{z=0}^{t}\frac{\nu_{iz} - \delta_c - \lambda_i\tau_p}{\sigma_c}\right)}{\sum_{j=0}^{T-1}\exp\left(\sum_{z=0}^{j}\frac{\nu_{iz} - \delta_c - \lambda_i\tau_p}{\sigma_c}\right)}, \qquad (11)$$

$$\text{with}\, \delta_1, \sigma_c > 0,$$

where ν_{it} denotes the threshold of response time category t on item i with ν_{i0} arbitrarily set to 0. In Eq. 11, we assume the intercepts and loadings to be invariant across states but we model a scale and location difference between the states using, respectively, δ_c and σ_c. That is, if $\delta_0 = 0$ for identification purposes, δ_1 accounts for a location shift of the thresholds in state 1 as compared to the thresholds in state 0. This reflects that the average raw response times are different between the states. As $\delta_1 > 0$, the responses in Class 1 are on average faster than the responses in Class 0. Parameter σ_c accounts for a scale difference in state 1 as compared to state 0, which is due to the raw response times being more variable in one state than in the other (heteroscedasticity). Note that in the traditional partial-credit model with only one state $\sigma_c = \sigma$ is only identified if two thresholds are fixed (Mehta, Neale, & Flay, 2004). However, here, if $\sigma_0 = 1$ for identification purposes, parameter σ_1 is identified and represents the ratio between the residual standard deviations in the two states. Thus, in the case of homoscedasticity $\sigma_0 = \sigma_1 = 1$. In the case of heteroscedasticity, $\sigma_1 > 1$, denotes more variability in state 1 and $\sigma_1 < 1$ denotes more variability in state 0. In the model for categorized response times in Eq. 11, differences in variability between items (i.e., differences in σ_{ci} across i in the continuous response time model in Eq. 7) are captured in the thresholds, ν_i and the factor loadings, λ_i. Differences in variability between classes are captured by σ_c.

Finally, for the conditional probability mass function of the responses within each state, $g_c(x_{pi}| \theta_p, \zeta_{pi} = c)$, we use Eq. 8 with a two-parameter model for $P(x_{pi} = 1 \mid \theta_p, \zeta_{pi} = c)$, that is,

$$P\left(x_{pi} = 1 | \theta_p, \zeta_{pi} = c\right) = \omega\left(\alpha_{ci}\theta_{pi} + \beta_{ci}\right). \tag{12}$$

Note that, contrary to Wang and Xu (2015) and Wang et al. (2018), we follow Molenaar et al. (2018; Molenaar et al., 2016) and use a two-parameter model for the responses (see also Table 1). Our main reason is that we want to operate in a generalized linear modeling framework that does not include the three-parameter model as a special case.[2] Using a three-parameter model would increase our model complexity, resulting in a potentially poorly identified model. Within the generalized linear modeling framework, we are sure that the model is identified and can be estimated properly. In addition, our modeling interest is mainly in detecting possible differences in item discrimination and item easiness across the different states (suggesting different response processes). However, extending the present model to a three-parameter model would be possible in principle

The model given by Eq. 10, with $h_c(.)$ given by Eq. 11, $g(.)$ given by Eq. 8, and $P(x_{pi} = 1|\theta_p, \zeta_{pi} = c)$ given by Eq. 12, constitutes the *heteroscedastic hidden Markov mixture model*. If we assume a bivariate standard normal distribution for τ_p and θ_p with correlation ρ, and if $\boldsymbol{\eta}$ denotes the vector of free parameters in the model (i.e., α_{0i}, α_{1i}, β_{0i}, β_{1i}, and λ_i for all i, ν_{it} for all i and for $t = 1, \ldots, T-1$, and δ_1, σ_1, π_1, $\pi_{0|1}$, and ρ), then the resulting full marginal log-likelihood function of the model is given by

$$\ell(\boldsymbol{\eta}) = ln\int_{-\infty}^{\infty} \sum_{\zeta_{p1}=0}^{1} \ldots \sum_{\zeta_{pn}=0}^{1} P\left(\zeta_{p1}\right) \prod_{i=2}^{n} P\left(\zeta_{pi}|\zeta_{p(i-1)}\right) \prod_{i=1}^{n} g_c\left(x_{pi}|\theta_p, \zeta_{pi}\right) h_c\left(\overset{'}{\underset{pi}{t}}|\tau_p, \zeta_{pi}\right) k\left(\tau_p, \theta_p\right) d\theta d\tau, \tag{13}$$

where $k(.)$ is a bivariate standard normal distribution with correlation ρ.

We focus on five instances of the general model above:

1. *Baseline:* A baseline model with one state (see Table 1).
2. *Heteroscedastic Markov states:* The full model with a Markov structure on the latent class variables and heteroscedastic states.
3. *Homoscedastic Markov states:* A model with a Markov structure on the latent class variables and homoscedastic states.
4. *Heteroscedastic independent states:* A model with independent latent class variables and heteroscedastic states
5. *Homoscedastic independent states:* A model with independent latent class variables and homoscedastic states

In all models, we use categorized response times. In the simulation study below, we investigate the viability of the general model in terms of parameter recovery and the resolution to distinguish between the different models above in responses and categorized response time data.

Categorization of response times

The models proposed require categorization of the continuous response times. Because the results potentially depend on the exact categorization scheme, categorization should be done with care. In the partial credit model above, the adjacent categories logit in the baseline model (i.e., $\delta_c = 0$ and $\sigma_c = 1$ for all c) is given by

$$\log\left\{\frac{h_c\left(t'_{pi} = t|\tau_p\right)}{h_c\left(t'_{pi} = t{-}1|\tau_p\right)}\right\} = \nu_{it} - \lambda_i\tau_p.$$

In this equation, the threshold parameter ν_{it} is directly influenced by the cut-off values at which the continuous response times are categorized. In principle, this is not a problem, as the other parameters are relatively unaffected by the exact choice of the cutoff values. However, this choice does affect the power to detect differences between states. Therefore, the cutoff values should be chosen in an optimal way. Here we propose to categorize the continuous response times in such a way that the adjacent categories logits show large, but, constant differences across categories. This will result in thresholds parameters ν_{it} that are equidistant and well spread over the τ_p range so that the information about τ_p in the categorized response times is approximately constant over τ_p (at least in the interval – 3, 3). A possible way to accomplish this is to choose the cutoff values on basis of equally spaced values in a symmetrical distribution (e.g., logistic or normal distribution). Here we use – 2, – 2/3, 2/3, and 2 in a normal distribution. This corresponds to cumulative probabilities of .0228, .2525, .7475, and .9773, which are used to categorize the continuous response times (i.e., at percentiles 2.28, 25.25, 74.75, and 97.73). In Fig. 2, this procedure is illustrated for a simulated-data example. Specifically, for a single item response time variable, the figure contains a histogram of the

[2] However, note that it is possible to specify the three-parameter model as a mixture of a two-parameter model and a guessing model, which is a generalized linear model.

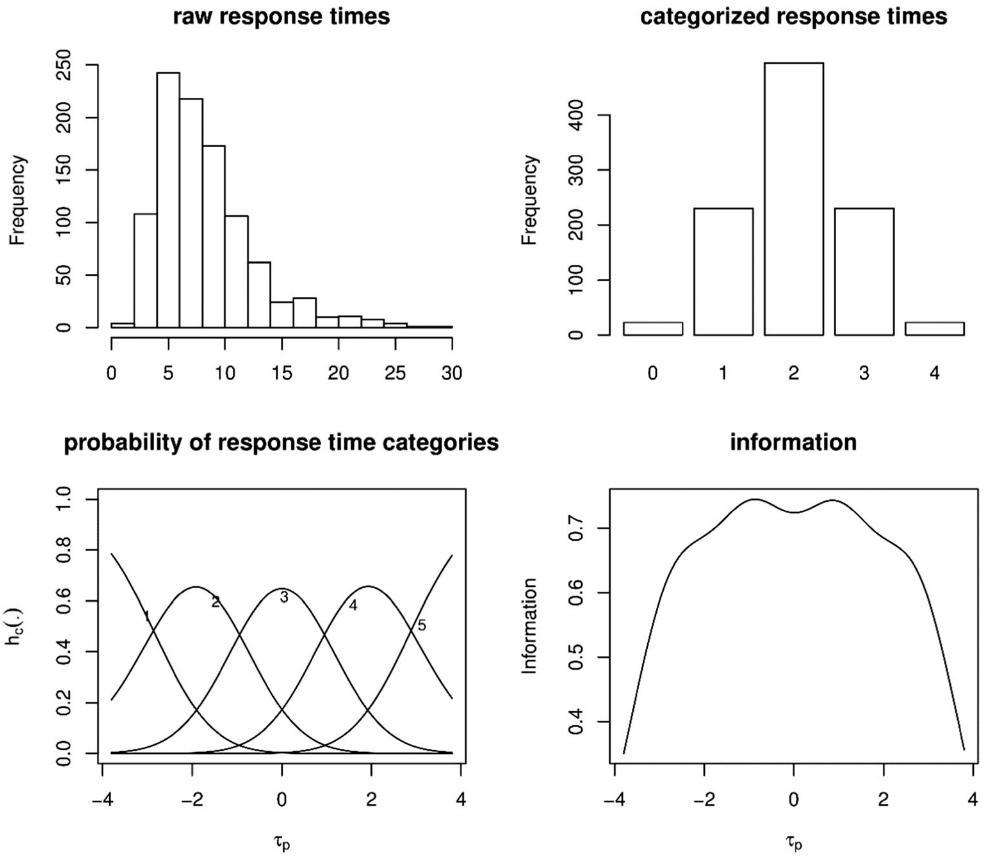

Fig. 2 Illustration of the categorization procedure, based on percentiles derived from a normal distribution (2nd, 25th, 75th, and 98th percentiles). (Top left) Histogram of the raw response times. (Top right) Bar plot of the categorized response times. (Bottom left) Conditional probabilities of each response time category. (Bottom right) Information in the categorized response times as a function of the latent speed variable τ_p

raw response times, a bar plot of the categorized response times, a plot of the conditional probability of each response time category, and the information of the categorized response times across τ_p. Applying the partial-credit model to data such as those in Fig. 2 will result in ν_{it} estimates that are well spread out over the τ_p range (at least in the – 3, 3 range), such that the information about τ_p is relatively constant in the range (– 3, 3). An alternative approach to categorizing the continuous response times may be to use equidistant percentiles like 20, 40, 60, and 80; however, as is illustrated in Fig. 3, such an approach will result in conditional response time category probabilities (bottom left plot) that are mainly centered around $\tau_p = 0$. Applying the partial-credit model to data such as those in Fig. 3 will result in ν_{it} estimates that are close together for a given item *i*. As a result, the information about the latent speed variable, τ_p, peaks at 0 and decreases relatively fast for values further away from 0. In the present study, we therefore consider the former approach (based on percentiles derived from a normal distribution at – 2, – 2/3, 2/3, and 2).

Estimation

The models above were implemented in LatentGold (Vermunt & Magidson, 2013) and estimated using marginal maximum likelihood. We optimized the marginal log-likelihood function in Eq. 13 above by numerically integrating the double integral using ten quadrature points for each dimension. Next, we used the Baum–Welch adapted EM algorithm (Baum, Petrie, Soules, & Weiss, 1970; Welch, 2003) to obtain reasonable starting values, after which we used the Newton–Raphson algorithm to find the maximum of the likelihood function. Because this procedure is full-information, missing data in the responses or the response times do not pose a problem as long as these are missing at random (Little & Rubin, 1987). The syntax to fit the full model (heteroscedastic Markov states) is available in the Appendix.

Simulation study

Design

To study the viability of the proposed models, we investigated the parameter recovery of the latent state parameters α_{ic}, β_{ic}, π_1, $\pi_{1|0}$, and $\pi_{1|1}$. We considered the situation in which the response time distribution departs from a log-normal distribution such that the continuous response time mixture model for

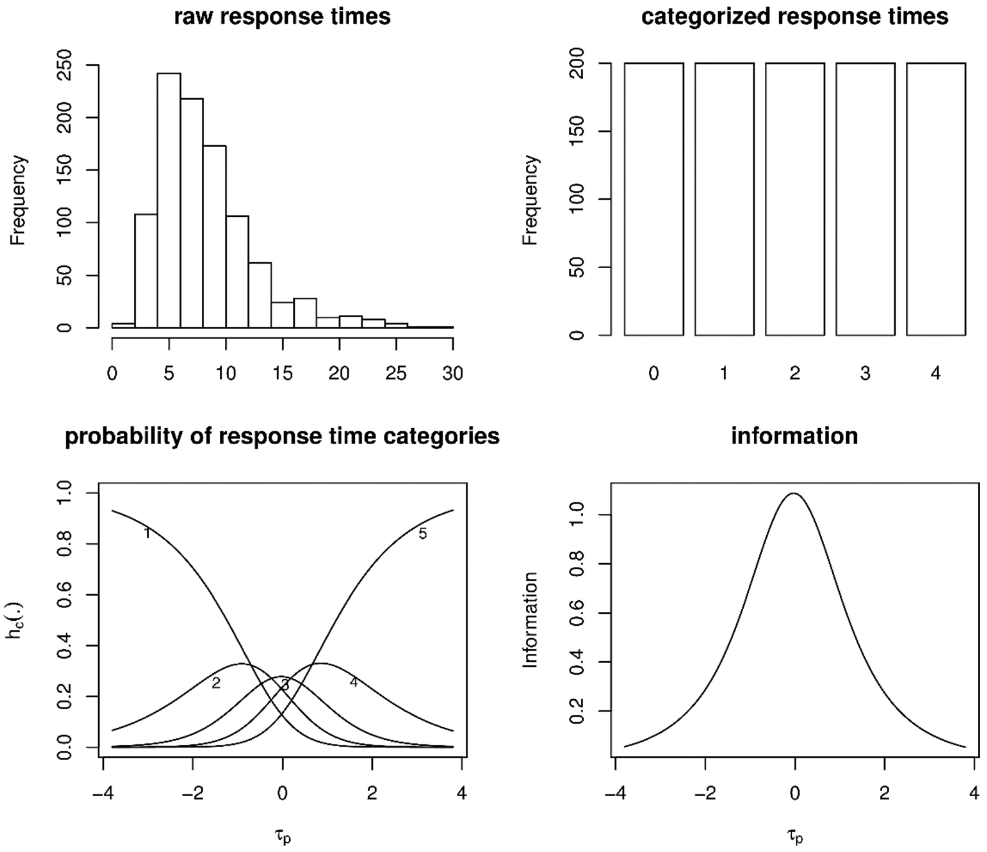

Fig. 3 Illustration of the categorization procedure based on equidistant percentiles (the 20th, 40th, 60th, and 80th percentiles). (Top left) Histogram of the raw response times. (Top right) Bar plot of the categorized response times. (Bottom left) Conditional probabilities of each response time category. (Bottom right) Information in the categorized response times as a function of the latent speed variable τ_p

the response times in Eq. 7 is unsuitable (i.e., as it will produce bias and false positives as discussed above).

The general procedure was as follows: We simulated responses and response times for 1,000 respondents on 20 items according to five scenarios that correspond to the five models above. We first simulated responses and continuous response times, after which the response times were categorized. Continuous response time data for the five scenarios were simulated according to a Box–Cox-transformed log-normal response time model that corresponds to the given scenario (e.g., for the heteroscedastic Markov states scenario, this will be a heteroscedastic Markov states model in which the partial credit model in Eq. 11 is replaced by a Box–Cox-transformed log-normal model). The Box–Cox transformation was used in order to make the response time data overly skewed, such that the response times do not follow a log-normal distribution, which invalidates models like the one in Eq. 7 discussed above. Below we discuss how we exactly simulated the responses and continuous response time data in each scenario:

Heteroscedastic Markov states To generate data for the first scenario, we used the heteroscedastic Markov states model with a continuous log-normal response time distribution with mean $\nu_i - \delta_c - \tau_p$ and standard deviation σ_c, which is the continuous version of Eq. 11 from the heteroscedastic Markov states model for categorized response times. For the mixture parameters, we used π_1 = .666 for the initial state parameter and $\pi_{0|1}$ = .231 and $\pi_{1|1}$ = .769 for the transition parameters (note that these choices imply that π_0 = .333, $\pi_{1|0}$ = .231, and $\pi_{0|0}$ = .769). These effect sizes correspond to moderately imbalanced initial state probabilities (Dias, 2006) and moderately unstable transition parameters (Bacci et al., 2014). The responses were simulated using $\alpha_{0i} = 1.5$ and $\alpha_{1i} = 1$ for all i for the discrimination parameters. For the easiness parameter, we used increasing, equally spaced values between – 2 and 0 for β_{0i} and between 0 and 2 for β_{1i}. For the response times, we simulated τ_p with $\sigma_\tau = \sqrt{0.13}$ and a correlation between τ_p and θ_p of .4. For the intercepts, we used $\nu_i = 2$ for all i, $\delta_0 = 0$, and $\delta_1 = 0.5$. For the residual standard deviations, we used $\sigma_0 = \sqrt{0.39}$ and $\sigma_1 = \sqrt{0.13}$. These choices result in communalities of .25 in Class 0 and .5 in Class 1 on the log-scale (as we simulated log-normal data; see above). In addition, the intercept differences of 0.5 between the states were considered of medium effect size by Molenaar et al. (2018). After the log-

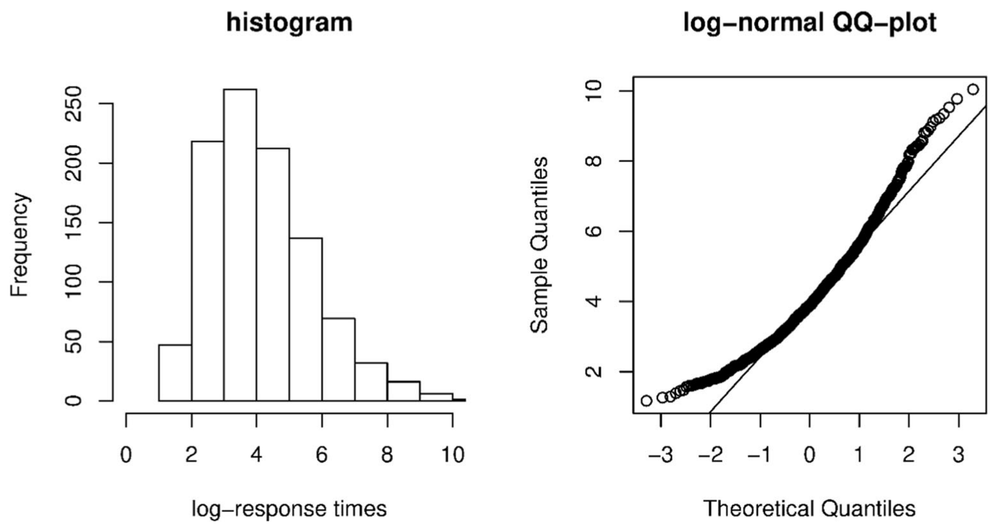

Fig. 4 Example distribution of continuous response times in the simulation study, which depart from a log-normal distribution. These continuous response times are subsequently categorized. (Left) Histogram of the log-transformed continuous response times (which should be normal if the response times followed a log-normal distribution). (Right) Log-normal QQ-plot of the response times (which should be on the straight line if the response times followed a log-normal distribution)

normal response time data were simulated, we log-transformed the simulated response times resulting in normally distributed log-response times. These log-response times were subsequently transformed using the Box–Cox transformation, $\xi(x+1)^{\zeta}$, with transformation parameter $\xi = 0.3$, such that the raw response times (i.e., the exponentially transformed Box–Cox log-response times) are overly skewed as compared to a log-normal distribution. As we mentioned, this makes these data unsuitable for mixture models like the one in Eq. 7, calling for our categorized response time mixture model. See Fig. 4 for an example response time distribution from the present simulation study.

Homoscedastic Markov states In this scenario, we used the same setup and procedure as for the Heteroscedastic-Markov-States scenario but with $\sigma_0 = \sigma_1 = \sqrt{0.13}$.

Heteroscedastic independent states In this scenario, we used the same setup and procedure as in the Heteroscedastic-Markov-States scenario but without the Markov structure on the states (i.e., $P(\zeta_{pi} = 1) = \pi_1$ for all i)

Homoscedastic independent states In this scenario, we used the same setup and procedure as in the heteroscedastic independent states scenario, but with $\sigma_0 = \sigma_1 = \sqrt{0.13}$.

Baseline In this scenario, we used a baseline model without mixture (i.e., only one state: $\delta_0 = \delta_1 = 0$, $\sigma_{0\,=} = \sigma_1 = \sigma$, $\alpha_{0i} = \alpha_{1i} = \alpha_i$, and $\beta_{0i} = \beta_{1i} = \beta_i$). For the response time parameters ν_i and σ_i, we used the parameters from state 0 in the homoscedastic independent states model above. For the responses we used $\alpha_i = 1.5$ and equally spaced values between – 2 and 2 for β_i. All other parameters were the same as in the homoscedastic independent states model above. In addition, like in the other scenarios, the response times data were transformed according to the Box–Cox transformation as explained above.

After the responses and continuous response times had been simulated, the raw response times were categorized at percentiles 2.28, 25.25, 74.75, and 97.73, resulting in five response time categories. Note that it does not make a difference whether the raw or transformed response times are categorized as the percentile scores will be the same. The percentiles that we used are obtained from a standard normal distribution at – 2, – 2/3, 2/3, and 2.

We used 50 replications for each data scenario. To the replications within each data scenario we fit the five models discussed above. Note that we thus did not fit the true model to the simulated data as the data were generated according to the Box–Cox-transformed log-normal model, and we fit a model for categorized response times. However, if the categorized model is viable, the latent state parameters α_{ic}, β_{ic}, π_1, $\pi_{1|0}$, and $\pi_{1|1}$ should be correctly recoverable despite the response times being categorized. The recovery of the response time measurement model parameters ν_{it}, λ_i, and σ_c cannot be studied as they do not have a corresponding true parameter value.

For each model we considered which of the five models is the best-fitting model according to the following fit indices: the Bayesian information criterion (BIC; Schwarz, 1978), Akaike's information criterion (AIC; Akaike, 1974), AIC3 (Bozdogan, 1993), the consistent AIC (CAIC; Bozdogan, 1987), and the sample-size-adjusted BIC (saBIC; Sclove, 1987). All these fit indices are based on the maximum marginal log-likelihood, $\ell(\hat{\boldsymbol{\eta}})$, where $\hat{\boldsymbol{\eta}}$ contain the parameter values that maximize $\ell(\boldsymbol{\eta})$ from Eq. 13. That is, the general

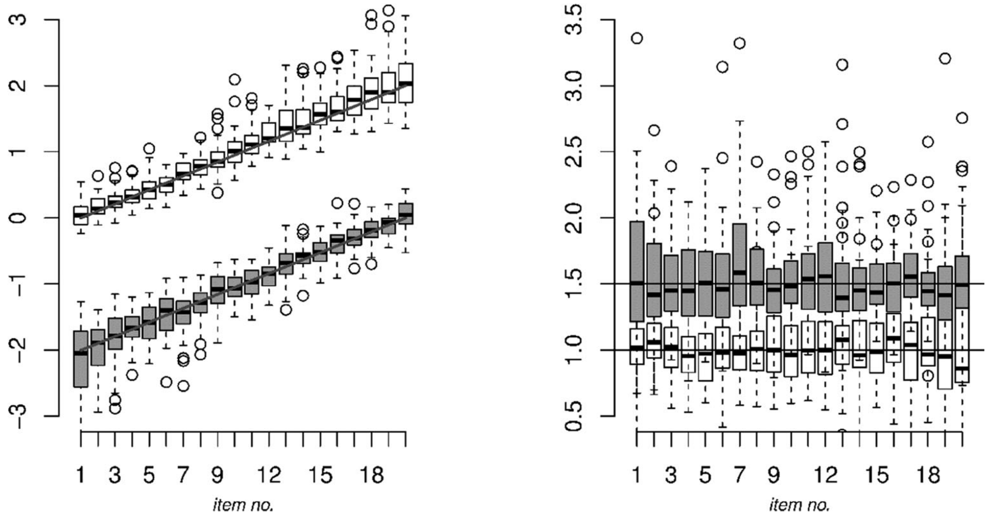

Fig. 5 Parameter recovery for the easiness parameters (left) and discrimination parameters (right) for the two states (gray: state 0, the slower state; white: state 1, the faster state), in the presence of heteroscedasticity in the response times between states that is explicitly accounted for using the scale factor

form of these fit indices is: $-2\ell(\hat{\boldsymbol{\eta}})+P$. The main difference between the fit indices above is the penalty term, *P*, that is used. That is,

$$\begin{aligned}
&\text{AIC}: P = 2\times npar,\\
&\text{BIC}: P = \log(N)\times npar,\\
&\text{AIC3}: P = 3\times npar,\\
&\text{CAIC}: P = 2\times npar(npar-1)/(N-npar-1),\\
&\text{saBIC}: P = \log((\text{N}+2)/24)\times npar,
\end{aligned}$$

where *npar* denotes the number of estimated parameters in a given model. For all the fit indices it holds that a smaller value indicates a better model fit.

Results

Parameter recovery We limit our presentation of the parameter recovery results to the most complex model (heteroscedastic Markov states model) as this is the model of key interest and the most challenging model to fit in terms of the number of parameters, but the results for the other, more parsimonious, models are comparable.

To study the parameter recovery of the model, Fig. 5 depicts box plots of the item parameter estimates β_{0i}, β_{1i}, α_{0i}, and α_{1i} across replications for the heteroscedastic Markov states model in the heteroscedastic Markov states scenario. As can be seen, all parameters seem to be recovered acceptably, with more variability in the discrimination parameters than in the easiness parameters. In addition, overall, the parameter estimates in state 0 (gray in the figure) are associated with somewhat more variability than the parameter estimates in state 1, as state 0 is smaller than state 1.

Statistics concerning the parameter recovery of the Markov parameters (π_1, $\pi_{1|0}$, and $\pi_{1|1}$) and the correlation between θ_p and τ_p (ρ) of the heteroscedastic Markov states model in the heteroscedastic Markov states scenario is depicted in Table 2. As can be seen, all parameters seem unbiased, with acceptable sampling properties (in terms of the 95% coverage rates, and the standard deviations and RMSEs of the estimates as compared to the mean standard error), although the coverage rate of $\pi_{1|0}$ is somewhat too small (.900 instead of .950). However, overall, we think the results do not indicate any problems with the model.

To study the effects of unmodeled heteroscedasticity between the states, Fig. 6 depicts box plots of the parameter

Table 2 Recovery results for the Markov parameters and for ρ

Parameter	True	MEAN(Est)	*SD*(Est)	RMSE	MEAN *SE*	Coverage
ρ	−. 400	−. 420	. 033	. 038	. 032	. 940
π_1	. 667	. 661	. 085	. 085	. 073	. 940
$\pi_{1\|0}$	. 231	. 222	. 014	. 016	. 015	. 900
$\pi_{1\|1}$	. 769	. 768	. 018	. 017	. 014	. 960

"Est" denotes the estimates of the corresponding parameter across the different replications, RMSE is the root-mean squared error, "*SE*" refers to the analytical standard errors of the parameter estimates (Est), and "Coverage" refers to the 95% coverage rates

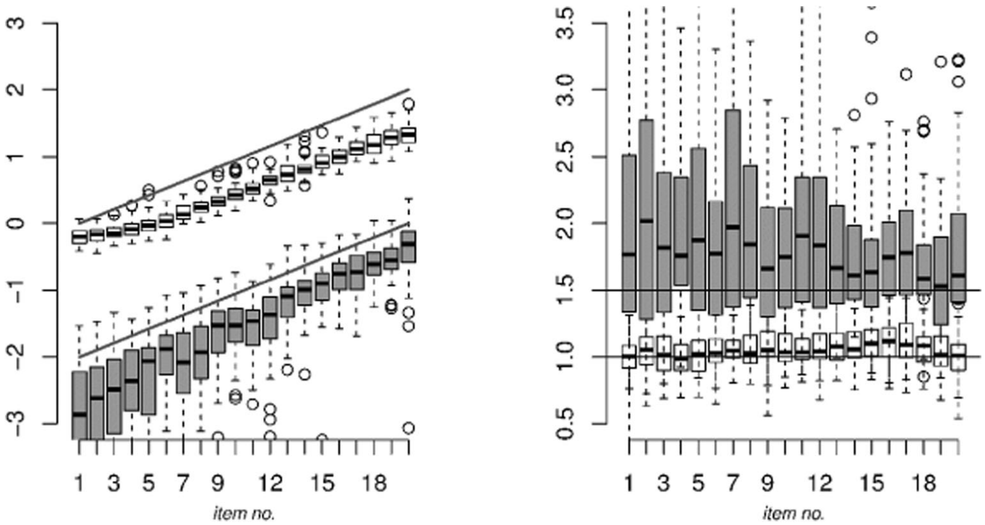

Fig. 6 Parameter recovery for the easiness parameters (left) and discrimination parameters (right) for the two states (gray: state 0, the slower state; white: state 1, the faster state), in the presence of unmodeled heteroscedasticity in the response times

estimates for the discrimination and easiness parameters in the heteroscedastic Markov states scenario but for the homoscedastic Markov states model. Comparing Figs. 5 and 6, it can be seen that neglecting the heteroscedasticity between states (Fig. 6) biased the parameter estimates (most notably in α_{1i}, β_{0i}, and β_{1i}) and increased the variance of the estimates of α_{1i} and β_{1i} (as compared to Fig. 5). In addition, it can be seen that neglecting heteroscedasticity in the data decreased the variance of α_{1i} and β_{1i} as compared to the case in which heteroscedasticity was accounted for. This is due to the size of state 1 (the faster state) being overestimated: π_1 has an average estimate of .816 (*SD:* .040), where the true value equaled .666. In addition, state 0 was relatively unstable: The average estimate of transition parameter $\pi_{0|1}$ was equal to .463 (*SD:* .0452), where the true value equaled .231. State 1 was estimated to be relatively stable: The average estimate of transition parameter $\pi_{1|1}$ was equal to .844 (*SD:* .010), where the true value equaled .769. Thus, Class 1 was still relatively stable, while Class 0 appeared relatively unstable.

True positive rates See Table 3 for the detection rates of the fit indices in each data scenario. The detection rate of a given model is the proportion of replications in which that model was indicated to be the best-fitting model among the five models considered. In the table, the true positive rates of a model are marked in gray. The true positive rate of a model is the detection rate of that model in the case that the model is fit to its corresponding scenario (e.g., the baseline model to the baseline scenario).[3] All other detection rates in Table 3 are false positives, which ideally should be close to 0. We consider true positive rates between .80 and 1.00 to indicate a good true positive rate, rates between .70 and .80 as acceptable, rates between .50 and .70 as moderate, and rates below .500 as poor.

As can be seen from Table 3, for the baseline model and the heteroscedastic Markov states model, true positives are perfect (i.e., 1.00) for all fit indices, but the true positive rate for the AIC is only .24 for the baseline model. As can be seen from the false positive rate in the baseline scenario, using the AIC fit index, the baseline model is hard to distinguish from the homoscedastic Markov states model, which is associated with a false positive rate of .40. For the homoscedastic Markov states model, true positives are all acceptable to good, with values between .86 and .98. For the heteroscedastic independent states model, the true positives are also considered acceptable to good, with values between .72 and 1.00, and for the homoscedastic independent states model, the true positive rate is moderate for the AIC, with a rate of .62, but acceptable to good for the other fit indices, with values between .80 and .98.

Conclusion

In conclusion, it appeared that parameter recovery is acceptable and that all fit indices but the AIC behaved acceptably in selecting among the different models under the circumstances simulated. The poor behavior of the AIC in model selection is in line with the findings of Molenaar et al. (2016), who also found poor performance of the AIC in selecting among models that did and did not include (Markov) mixtures. In addition, we found that neglecting heteroscedasticity between classes may bias the item parameter estimates and increase their variance.

The main purpose of these simulations was a proof of principle in the sense that we wanted to show that we can adequately recover the true parameter values of the model and that we

[3] Note that we cannot speak of the “true model,” because the response time data were simulated under a different model (a categorized Box–Cox-transformed log-normal model) from the model applied to the data (a partial-credit model; see Eq. 11).

Table 3 Detection rates of the BIC, AIC, AIC3, CAIC, and saBIC for the five models in each data scenario

	Model				
Data Scenario	*Base-line*	*Hetero. Markov States*	*Homo. Markov States*	*Hetero. Indep. States*	*Homo. Indep. States*
	BIC				
Baseline	1.00	.00	.00	.00	.00
Hetero. Markov states	.00	1.00	.00	.00	.00
Homo. Markov states	.00	.02	.98	.00	.00
Hetero. independent states	.00	.00	.00	1.00	.00
Homo. independent states	.00	.00	.00	.06	.94
	AIC				
Baseline	.24	.18	.40	.04	.14
Hetero. Markov states	.00	1.00	.00	.00	.00
Homo. Markov states	.00	.14	.86	.00	.00
Hetero. independent states	.00	.28	.00	.72	.00
Homo. independent states	.00	.06	.20	.12	.62
	AIC3				
Baseline	1.00	.00	.00	.00	.00
Hetero. Markov states	.00	1.00	.00	.00	.00
Homo. Markov states	.00	.06	.94	.00	.00
Hetero. independent states	.00	.08	.00	.92	.00
Homo. independent states	.00	.02	.06	.12	.80
	CAIC				
Baseline	1.00	.00	.00	.00	.00
Hetero. Markov states	.00	1.00	.00	.00	.00
Homo. Markov states	.00	.02	.98	.00	.00
Hetero. independent states	.00	.00	.00	1.00	.00
Homo. independent states	.00	.00	.00	.02	.98
	saBIC				
Baseline	1.00	.00	.00	.00	.00
Hetero. Markov states	.00	1.00	.00	.00	.00
Homo. Markov states	.00	.04	.96	.00	.00
Hetero. independent states	.00	.02	.00	.98	.00
Homo. independent states	.00	.02	.02	.10	.86

Gray shading indicates the true positive rates (the detection rate for a model in its corresponding scenario—e.g., the baseline model in the baseline scenario); the other rates are false positive rates. In addition: "Hetero." denotes "Heteroscedastic" and "Homo." denotes "Homoscedastic"

can distinguish well between the different models given a reasonable sample size and reasonable effect sizes. However, the results above depend on the choices we made concerning parameter values. That is, true positives will decrease for decreasing differences between the states in terms of δ_c and β_{ci} and α_{ci}. In addition, if the stability of the states decreases (reflected by larger values for $\pi_{1|0}$ and smaller values for $\pi_{1|1}$) true positives will also decrease (see, e.g., Molenaar et al., 2016).

Illustration

Data

In this section, we demonstrate the viability of the present modeling approach in a real dataset. We used the responses and response times to the block design subtest of the Hungarian WAIS-IV (Nagyné Réz et al., 2008). These data

Table 4 Model fit indices for the five models considered in the application, for $T = 5$

Model	BIC	AIC	AIC3	CAIC	sBIC
Baseline	27,612	27,163	27,255	27,704	27,320
Heteroscedastic Markov states	**27,043**	**26,442**	**26,565**	**27,166**	**26,652**
Homoscedastic Markov states	27,068	26,472	26,594	27,190	26,681
Heteroscedastic independent states	27,428	26,837	26,958	27,549	27,044
Homoscedastic independent states	27,437	26,851	26,971	27,557	27,056

The best values of the fit indices are in boldface

have been analyzed by Molenaar, Bolsinova, Rósza, and De Boeck (2016), who analyzed these data using a mixture model for the responses but not for the response times. The data consist of the responses and response times of 978 respondents to 14 items. The items were designed to be decreasing in easiness. The raw response times are between 1 and 360 s. We omitted Item 1 from the analysis as this item caused numerical problems due to the high success rate (.999). We used the same procedure as in the simulation study. That is, we used the same categorization procedure for the raw response times, we considered the same models, and we used the same estimation procedure.

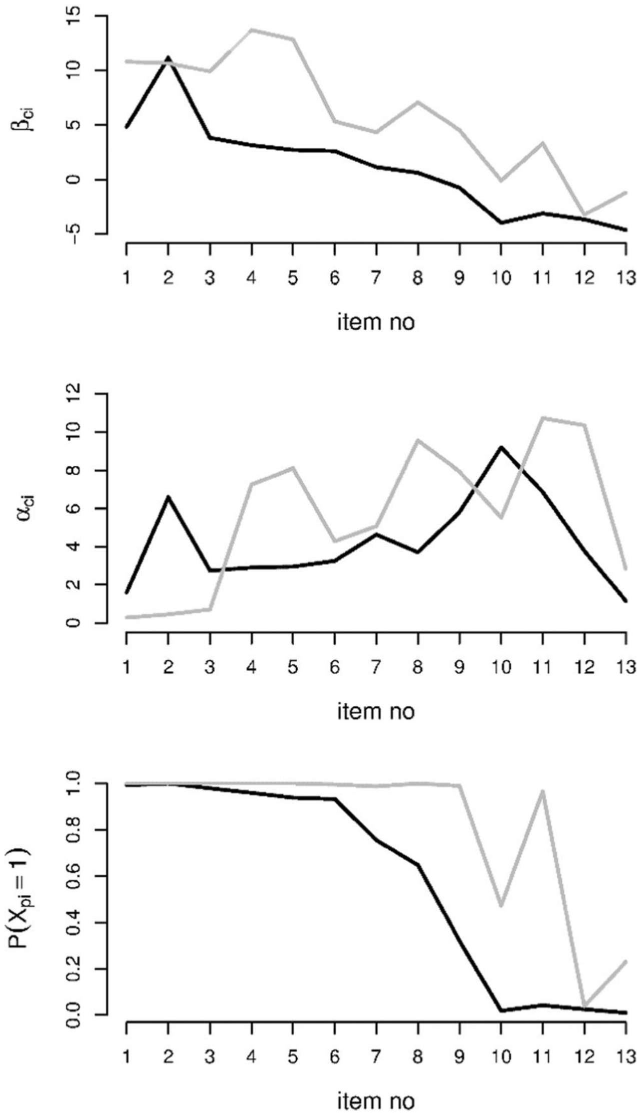

Fig. 7 Parameter estimates for the easiness parameters (top) and discrimination parameters (middle), together with the implied marginal probabilities of a correct response in Class 0 (black lines) and Class 1 (gray lines)

Results

See Table 4 for the model fit indices of the models considered. As can be seen, all fit indices indicate the heteroscedastic Markov states model to be the best-fitting model. Below we discuss the results from this model. First, it appeared that Class 1 (the faster class) is somewhat larger with an initial state parameter π_1 estimate of .617 (*SE*: 0.052). In addition, the classes seem relatively stable with transition parameters $\pi_{1|0}$ and $\pi_{1|1}$ estimated to be .124 (*SE*: 0.016) and .840 (*SE*: 0.015), respectively. In addition, δ_1 was estimated to be 3.484 (*SE*: 0.210), and the residual standard deviation in Class 1, σ_1, was 1.695 (*SE*: 0.131), indicating that Class 1 is associated with more variability in the response times.[4]

In Fig. 7, the item easiness parameters, discrimination parameters, and marginal probabilities of a correct response in the two classes are plotted. As can be seen, the easiness parameters in Class 1, β_{1i}, are generally larger than the easiness parameters in Class 0, β_{0i}. For the discrimination parameters, there is a less clear difference: It seems that the discrimination parameters in Class 1, α_{1i}, are somewhat larger than the discrimination parameters in Class 0, α_{0i}, for the items later in the test (from Item 4 onward, with Item 10 as an exception), but this effect is small.

Figure 8 depicts the raw response times, the item-wise standardized response times, and the posterior probabilities

[4] For numerical reasons, we estimated the logits of the initial-state and transition parameters. In addition, we estimated $exp(-\sigma_1)$. The reported standard errors were obtained using the delta method.

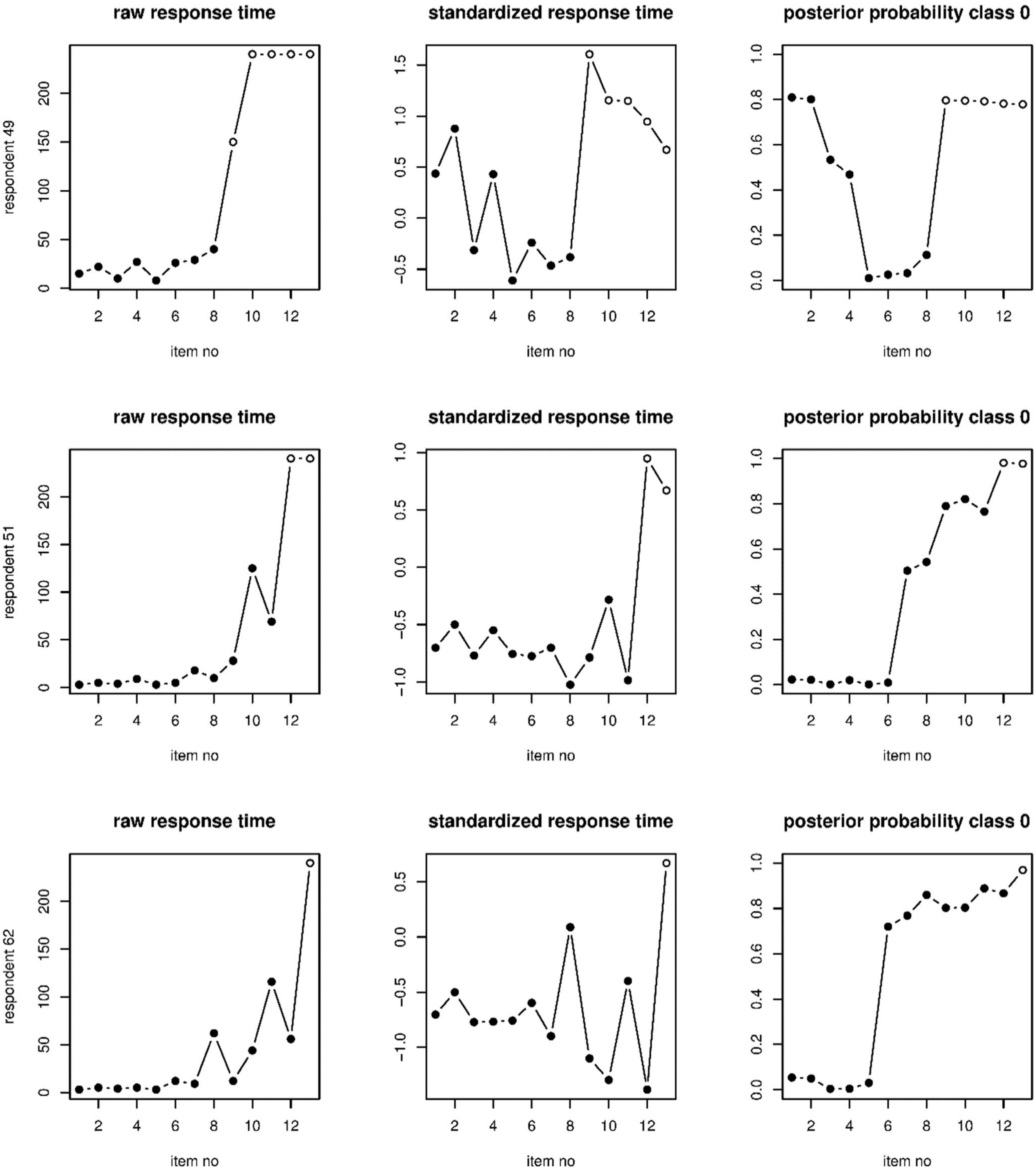

Fig. 8 Raw response times, the item-wise standardized response times, and posterior probabilities of Class 0 for three example respondents. Solid dots denote that the response to that item was correct

of Class 0 according to the heteroscedastic Markov states model for three example respondents. The raw response times are hard to interpret, as the items differ in their time intensity. The item-wise standardized response times provide an ad-hoc method to account for this confounding effect. However, besides the ad-hoc nature of this method, a drawback is that it does not account for the dependency between adjacent items and for the response outcome (correct or incorrect). As can be seen, the posterior probabilities generally give an improved picture of the response dynamics, as compared to the

Table 5 Model fit indices for the five models considered in the application for T = 3

Model	BIC	AIC	AIC3	CAIC	sBIC
Baseline	20,553	20,231	20,297	20,619	20,343
Heteroscedastic Markov states	**20,158**	**19,689**	**19,785**	**20,254**	**19,853**
Homoscedastic Markov states	20,178	19,709	19,805	20,274	19,873
Heteroscedastic independent states	20,417	19,958	20,052	20,511	20,118
Homoscedastic independent states	20,426	19,967	20,061	20,520	20,128

The best values of the fit indices are in boldface

standardized response times, with a clearer pattern. In addition, the classification is sometimes different for the posterior probabilities than for the standardized response times. For instance, for Respondent 62, the responses to Items 9, 10, and 12 are the fastest among all items according to the standardized response times, but according to the posterior probabilities, these responses are likely in Class 0 (the slower class).

Robustness analysis

To see whether the results above are robust to the exact number of response time categories used, we also conducted the above analyses using $T = 3$ and $T = 2$ response time categories. In the case of $T = 3$, we categorized the continuous response times of each item at percentiles 15.87 and 84.13 (obtained from a standard normal distribution at – 1 and 1). In the case of $T = 2$, we used a median split of the continuous response times of each item (i.e., we used a cutoff corresponding to percentile 50).

First, the estimates of parameters π_1, $\pi_{1|0}$, and $\pi_{1|1}$ are .599 (*SE:* .059), .152 (*SE:* .027), and .848 (*SE:* .016) for $T = 3$, and .582 (*SE:* .067), .241 (*SE:* .018), and .759 (*SE:* .021) for $T = 2$. As we discussed above, for $T = 5$ these estimates were, respectively, .617 (*SE:* .052), .124 (*SE:* .016), and .840 (*SE:* .015), respectively. As judged by the standard errors, these estimates do not differ importantly.

Tables 5 and 6 contain the fit measures for the different models for, respectively, $T = 3$ and $T = 2$. As can be seen, all fit measures favor the full model in both the $T = 3$ and $T = 2$ data. This is in line with the conclusions draw above for the $T = 5$ case (see Table 4). To compare the parameter estimates from the $T = 5$, $T = 3$, and $T = 2$ data, we plotted the person parameter estimates of θ_p and τ_p (Fig. 9) and the item parameter estimates of β_{0i}, β_{1i}, α_{0i}, and α_{1i} (Fig. 10) for the $T = 5$, $T = 3$, and $T = 2$ data. As can be seen from Fig. 9, there is a strong one-to-one correspondence between the person parameter estimates obtained from the different datasets. In Fig. 10, it can be seen that for the item parameters, the correspondence between the $T = 5$, $T = 3$, and $T = 2$ parameter estimates is best for β_{0i} and β_{1i}. For α_{0i}, the correspondence is associated with somewhat more noise than for β_{0i} and β_{1i}. For α_{1i} the correspondence is noisiest. This has to do with the relatively large standard error of the α_{1i} parameters as compared to the other item parameters. However, for the item parameters overall, there does not seem to be a systematic difference between the parameter estimates from the different datasets. We therefore conclude that the robustness of the results across the different numbers of response time categories is acceptable.

Discussion

In this article, we presented a mixture model to detect heterogeneity in the response processes underlying psychometric test items. The new model combines the strengths of previous mixture models by Schnipke and Scrams (1997), Wang and Xu (2015), Wang et al. (2018), Molenaar et al. (2016), and Molenaar et al. (2018). In our modeling approach we used mixture modeling in an indirect application (Yung, 1997). That is, the mixture components in our model are not

Table 6 Model fit indices for the five models considered in the application for T = 2

Model	BIC	AIC	AIC3	CAIC	sBIC
Baseline	18,603	18,344	18,397	18,656	18,435
Heteroscedastic Markov states	**18,096**	**17,686**	**17,770**	**18,180**	**17,829**
Homoscedastic Markov states	18,102	17,696	17,779	18,185	17,838
Heteroscedastic independent states	18,408	18,008	18,090	18,490	18,148
Homoscedastic independent states	18,404	18,008	18,089	18,485	18,147

The best values of the fit indices are in boldface

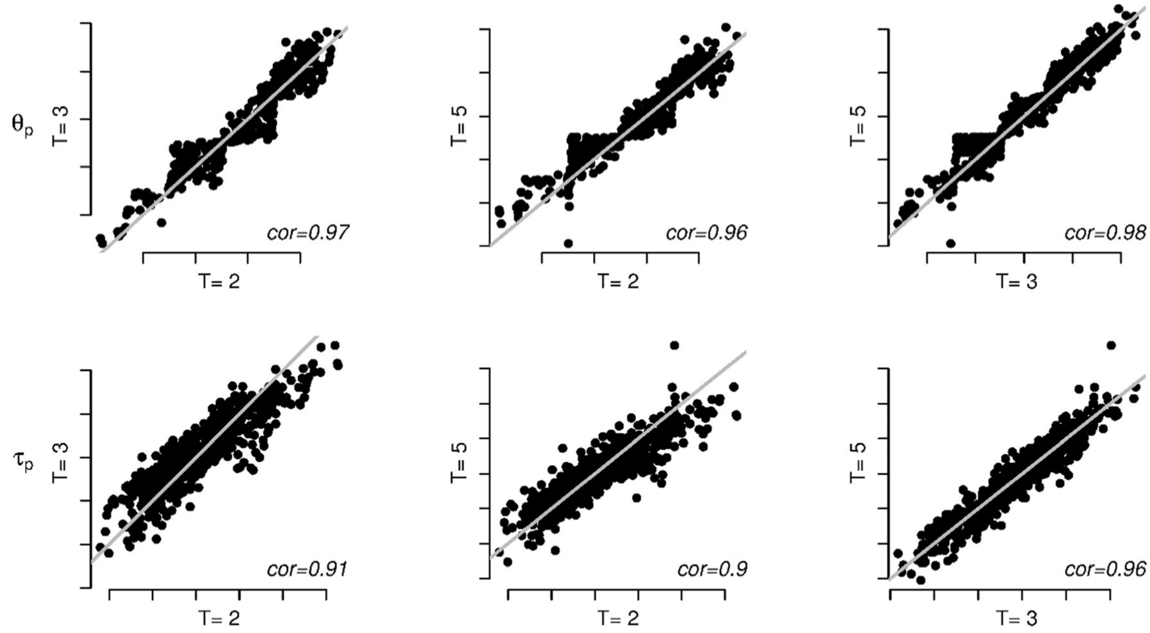

Fig. 9 Plot of the estimates for θ_p (first row of plots) and τ_p (second row of plots) for different numbers of response time categories, T (left plots: $T = 2$ vs. $T = 3$; middle plots: $T = 2$ vs. $T = 5$; right plots: $T = 3$ vs. $T = 5$). The solid gray lines denote one-to-one correspondences

necessarily substantively interpretable but are rather statistical tools to detect heterogeneity in the data that is due to differences in response processes. This is different from the modeling perspective by for instance Wang and Xu who used mixture modeling in a direct application (Dolan & van der Maas, 1998) in which the mixture components are substantively interpreted. Specifically, Wang and Xu distinguished between a fast guessing process and a solution process. Regardless of the nature of the mixture application (direct or indirect), the methodology presented in this article is equally amenable to the modeling of fast guessing and solution behavior. That is, if the measurement model for the responses in the faster state is restricted to represent fast guessing (i.e., discrimination equal to 0, see Table 1), the model is in essence the model by Wang and Xu, but with Markov-dependent states. Other restrictions are possible, which we will illustrate below. However, such restrictions need a strong theory about the response processes, which is not always available.

Throughout this article, we have assumed two latent states to underlie the item responses and response times, this has mainly a pragmatic reason in the sense that we think that two states can capture the most important patterns in the data. In addition, some theories describe binary processing, for instance the automated versus controlled processing theory (Shiffrin & Schneider, 1977), and the fast versus slow intelligence theory (DiTrapani, Jeon, De Boeck, & Partchev, 2016; Partchev & De Boeck, 2012). However, it can certainly be that some situations require more than two states (e.g., if three clearly distinct solution strategies underlie the response behavior of the respondents). In principle, it is straightforward to extend the present model to include three or more item specific states. However, the number of parameters rapidly grows. That is, for three item specific states, six parameters need to be estimated for each response variable (three discriminations and three easiness parameters). In such a situation, either the sample sizes should be very large, or one should incorporate reasonable model restrictions. That is, model restrictions can be thought of that are either pragmatically defendable or that are derived from theory. For instance, Molenaar et al. (2018) considered a model in which the item parameters have an overall difference across states and not an item specific difference (as in the models considered in the present article). In addition, Molenaar et al. (2016) used the restrictions that van der Maas and Jansen (2003) derived from the developmental theory by Siegler (1981) to distinguish different solution strategies underlying the Piagetian balance scale task. Using these restrictions, Molenaar et al. (2016) identified up to five states in a hidden Markov model for responses and continuous response times.

To solve the problem of spurious mixtures, we followed Molenaar et al. (2018) and categorized the continuous response times. This approach is pragmatic but shown effective in countering false positives in the case of distributional misfit. However, the approach has the drawback that information about individual differences is decreased such that the power to detect an effect may depend on arbitrary choices concerning the number and location of the cut-off values. It is therefore advisable to always investigate the robustness of the results with respect to the cut-off values as was illustrated in our real data example.

Another aspect of the general mixture modeling framework considered in this article (Table 1) is the operationalization of

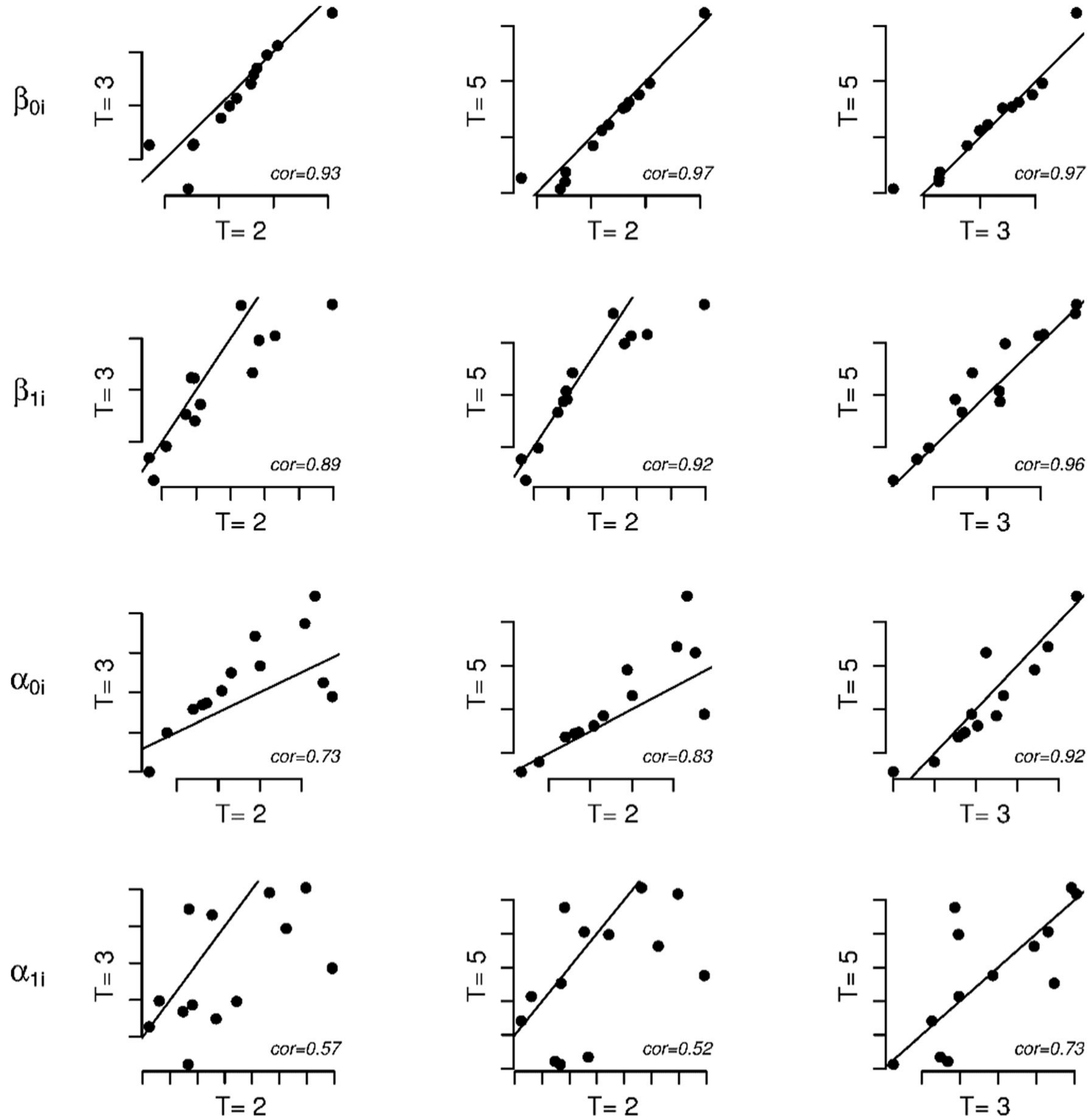

Fig. 10 Plot of the estimates for β_{0i} (first row of plots), β_{1i} (second row of plots), α_{0i} (third row of plots), and α_{1i} (fourth row of plots) for different numbers of response time categories, T (left plots: $T = 2$ vs. $T = 3$; middle plots: $T = 2$ vs. $T = 5$; right plots: $T = 3$ vs. $T = 5$). The solid black lines denote one-to-one correspondences

response processes in terms of the item properties (discrimination and easiness) and the expected response times. That is, a response process difference is assumed to be characterized by (1) a difference in the discrimination and/or easiness parameter and (2) a difference in the expected response time. This operationalization in Difference 1 can be justified by the statistical theory about measurement invariance (Mellenbergh, 1989; Meredith, 1993), which dictates that a difference in measurement model parameters indicates a difference in the interpretation of the underlying latent variable. That is, if faster responses are associated with different measurement parameters (discrimination and/or easiness) as compared to the slower responses, the latent variable has a different interpretation for these responses indicating a different response process. As we discussed before, the operationalization in Difference 2 can be justified by the theory about response times in experimental psychology (e.g., Luce, 1986), which dictates that the response times indicate the time that is needed for a certain psychological process to be executed. A difference in expected response time thus indicates a different process (all other things being equal).

An alternative to the statistical operationalizations of response processes adopted here are process-modeling operationalizations from mathematical psychology. In this framework, stronger assumptions are made about the response process (e.g., a response process consists of noisy information accumulation that stops if enough information for one of the response alternatives is gathered). From these assumptions, a mathematical model can be derived that is fit to the data. Examples of such models include the diffusion model (Ratcliff, 1978), the linear accumulator model (Brown & Heathcote, 2008), and the race model (Audley & Pike, 1965). However, these models are mathematically more complex, which made them less suitable to the aims of the present article. Yet it will certainly be interesting to consider models from mathematical psychology in light of the present mixture modeling framework.

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

Appendix

The syntax below can be used to fit the heteroscedastic hidden Markov mixture model to responses and categorized response times in LatentGOLD.

```
 options
   maxthreads=all;
  algorithm
    tolerance=1e-008 emtolerance=0.01 emiterations=250 nriterations=50 ;
  startvalues
    seed=0 sets=16 tolerance=1e-005 iterations=50;
  bayes
    categorical=1 variances=1 latent=1 poisson=1;
  montecarlo
    seed=0 sets=0 replicates=500 tolerance=1e-008;
  quadrature nodes=10;
  missing  includeall;
  output
    parameters=first standarderrors estimatedvalues=model classification=posterior;
variables
  caseid ID;
  dependent X, catT ordinal;
  independent item nominal;
  latent
    Ability  continuous,
    Speed  continuous,
    Cluster nominal 2 dynamic;
equations
 (1) Ability;
 (1) Speed;
  Ability <-> Speed;
  Cluster[=0] <- 1;
  Cluster <- 1 a Cluster[-1];      // replace by "Cluster <- 1 | Cluster[-1] Item" to drop the assumption of
                                   // time homogeneity
  X <- 1 | Item Cluster + (+) Ability | Item Cluster;
  catT <- 1 | item + (-) Cluster +  Speed|item;
  catT<<- cluster;
end model
```

References

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

Audley, R. J., & Pike, A. R. (1965). Some alternative stochastic models of choice. *British Journal of Mathematical and Statistical Psychology, 18* , 207–225.

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*, 125–145.

Bauer, D. J., & Curran, P. J. (2003). Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes. *Psychological Methods*, *8*, 338–363. https://doi.org/10.1037/1082-989X.8.3.338

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

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*(3), 345–370.

Bozdogan, H. (1993). Choosing the Number of Component Clusters in the Mixture Model Using a New Informational Complexity Criterion of the Inverse-Fisher Information Matrix. In (Eds., O. Opitz, B. Lausen, & R. Klar), *Information and Classification* (pp 40–54). Heidelberg: Springer-Verlag.

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*, 153–178. https://doi.org/10.1016/j.cogpsych.2007.12.002

Dias, J. (2006). Latent class analysis and model selection. In M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nurnberger, & W. Gaul (Eds.), *From data and information analysis to knowledge engineering* (pp. 95–102). Berlin, Germany: Springer.

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

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

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

Goldstein, K., & Scheerer, M. (1941). Abstract and concrete behavior an experimental study with special tests. *Psychological Monographs*, *53*(2), 1–151.

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

Gudicha, D. W., Schmittmann, V. D., & Vermunt, J. K. (2016). Power computation for likelihood ratio tests for the transition parameters in latent Markov models. *Structural Equation Modeling*, *23*, 234–245.

Hedeker, D., Berbaum, M., & Mermelstein, R. (2006). Location-scale models for multilevel ordinal data: Between- and within-subjects variance modeling. *Journal of Probability and Statistical Science*, *4*, 1–20.

Kuipers, R., Visser, I., & Molenaar, D. (2018). *Testing the within-class distribution in mixture models for responses and response times*. Manuscript in preparation.

Little, R. J., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York, NY: Wiley.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison Wesley.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization* (No. 8). Oxford, UK: Oxford University Press.

MacDonald, I. L., & Zucchini, W. (1997). *Hidden Markov and other models for discrete-valued time series* (Vol. 110). New York, NY: CRC Press.

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

Mehta, P. D., Neale, M. C., & Flay, B. R. (2004). Squeezing interval change from ordinal panel data: latent growth curves with ordinal outcomes. *Psychological Methods*, *9*, 301–333.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*, 127–143.

Mellenbergh, G. J. (1994). A unidimensional latent trait model for continuous item responses. *Multivariate Behavioral Research*, *29*, 223–236.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika*, *58*, 525–543.

Molenaar, D., Bolsinova, M., & Vermunt, J. K. (2018). A semi-parametric within-subject mixture approach to the analyses of responses and response times. *British Journal of Mathematical and Statistical Psychology*, *71*, 205–228. https://doi.org/10.1111/bmsp.12117

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

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

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

Nagyné Réz, I., Lányiné Engelmayer, Á., Kuncz, E., Mészáros, A., Mlinkó, R., Bass, L., . . . Kõ, N. (2008). *WISC-IV: A Wechsler Gyermek Intelligenciateszt Legújabb Változata* (Hungarian Version of the Wechsler Intelligence Scale for Children—Fourth Edition, WISC-IV). Budapest: OS Hungary Tesztfejlesztõ.

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

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

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59–108. https://doi.org/10.1037/0033-295X.85.2.59

Samejima, F. (1969). *Psychometric monographs: Vol. 17. Estimation of ability using a response pattern of graded scores*. Austin, TX: Psychometric Society.

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

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

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

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review*, *84*, 127–190. https://doi.org/10.1037/0033-295X.84.2.127

Siegler, R. S. (1981). Developmental sequences within and between concepts. *Monographs of the Society for Research in Child Development*, *46*(2). https://doi.org/10.2307/1165995

Tan, B., & Yılmaz, K. (2002). Markov chain test for time dependence and homogeneity: An analytical and empirical evaluation. *European Journal of Operational Research*, *137*, 524–543. https://doi.org/10.1016/S0377-2217(01)00081-9

Tuerlinckx, F., & De Boeck, P. (2005). Two interpretations of the discrimination parameter. *Psychometrika*, *70*, 629–650. https://doi.org/10.1007/s11336-000-0810-3

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72*(3), 287.

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

van der Maas, H. L. J., & Jansen, B. R. (2003). What response times tell of children's behavior on the balance scale task. *Journal of Experimental Child Psychology*, *85*, 141–177.

van der Maas, H. L. J., 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*, 339–356. https://doi.org/10.1037/a0022749

Vermunt, J. K. (2011). K-means may perform as well as mixture model clustering but may also be much worse: Comment on Steinley and Brusco (2011). *Psychological Methods, 16*(1), 82–88.

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

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

Wang, C., Xu, G., & Shang, Z. (2018). A two-stage approach to differentiating normal and aberrant behavior in computer based testing. *Psychometrika*, *83*, 223–254. https://doi.org/10.1007/s11336-016-9525-

Welch, L. R. (2003). Hidden Markov models and the Baum–Welch algorithm. *IEEE Information Theory Society Newsletter*, *53*, 10–13.

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

Zucchini, W., MacDonald, I. L., & Langrock, R. (2016). *Hidden Markov models for time series: An introduction using R*. New York, NY: Chapman & Hall/CRC.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.