Psychological Test and Assessment Modeling, Volume 63, 2021 (3), 361-395

A Continuous HYBRID IRT Model for Modeling Changes in Guessing Behavior in Proficiency Tests

Gabriel Nagy[1], *Alexander Robitzsch*[2]

Abstract

The results of low-stakes assessments are sensitive to individuals' persistence in maintaining a constant level of effort and precision over the course of a test. In this paper we present an item response theory (IRT) model that includes test-taking persistence as an additional latent variable. The proposed model is a continuous variant of the HYBRID IRT model. In contrast to Yamomoto's (1989) HYBRID model, our model allows for nondeterministic changes from solution to guessing behavior. Our model assumes that, over the course of a test, individuals might change their response behavior from solution behavior to random guessing behavior. Individual differences in the turning points are used to assess persistence. Individual differences in persistence can be correlated with proficiency, as well as with additional individual-level covariates. The new model is specified as a multilevel mixture IRT model and can be estimated by means of marginal maximum likelihood via the expectation maximization algorithm. The model was scrutinized in a simulation study that showed that the continuous HYBRID model provides good results in a variety of conditions. An empirical application provided further support for the model's utility because the essence of test-taking persistence was replicated in two test forms assessing science achievement.

Keywords: Item Response Theory, Multilevel Mixture Modeling, Guessing, Persistence, Test Engagement

[1] *Correspondence concerning this article should be addressed to:* Gabriel Nagy, Leibniz Institute for Science and Mathematics Education, Olshausenstraße 62, 24118 Kiel, Germany.
E-mail: nagy@ipn.uni-kiel.de

[2] Leibniz Institute for Science and Mathematics Education, Kiel, Centre for International Student Assessment, Germany

For proficiency tests administered under low-stakes conditions there is ample evidence that individuals reduce their test effort over the course of a test (Meyers, Miller, & Way, 2009). Research on rapid guessing behaviour (RGB; Wise & Kong, 2005) indicates that the reduction of effort is reflected in an increasing prevalence of RGB, which means that items presented later in a test are more likely to be guessed (Lindner, Lüdtke, & Nagy, 2019; Wise, Pastor, & Kong, 2009). As such, individual differences in the onset point of guessing behaviour might be regarded as an indicator of individuals' *test-taking persistence* (cf. Debeer, Buchholz, Hartig, & Janssen, 2014).

Many popular item response theory (IRT) models that deal with the declining probabilities of correct responses over the course of a test do not recur on changes in response strategies (e.g., Debeer & Janssen, 2013), although this idea is not new in IRT. Yamamoto (1989) proposed the HYBRID model, which combines the two-parameter logistic (2PL) IRT model for solution-based responses with a latent class analysis model for guessed responses. The HYBRID model assumes deterministic (i.e., sudden) changes from solution to guessing behaviour. However, at least in the area of low-stakes tests, a continuous decrease in the probability of solution behaviour appears to be more realistic (e.g., Goegebeur, De Boeck, Wollack, & Chen, 2008).

In the present article we present a continuous version of Yamamoto's (1989) HYBRID model, which we denote by the abbreviation C-HYBRID. The new model relaxes the assumption of a deterministic switching point from solution to guessing behavior. The model is specified as a multilevel mixture model (MMM) that can be estimated by software packages for multilevel latent variable modeling such as M*plus* (Muthén & Muthén, 1998-2017), Latent Gold (Vermunt & Magidson, 2013), and JAGS (Plummer, 2017). The MMM framework allows for extensions that, for example, enable covariate relationships with the switching point to be studied.

In the next section we introduce the conceptual aspects of guessing behaviour on which the HYBRID and the C-HYBRID models build. We then present both models and discuss the specification and estimation of the C-HYBRID model with the M*plus* software. These sections are followed by a simulation study and an application of the model to real data. We end the article by summarizing the results and discussing extensions to the suggested MMM framework.

Solution and Guessing Behavior in Proficiency Tests

In our framework the probability of a correct response of individual i ($i = 1, 2, \ldots, N$) to a dichotomously scored item j ($j = 1, 2, \ldots, J$), $P(Y_{ij} = 1)$ is represented as

$$P(Y_{ij} = 1) = s_{ij} r_{ij} + (1 - s_{ij}) g_j \,, \qquad (1)$$

where $r_{ij} \in [0,1]$ denotes the probability that individual i knows the response to item j, $s_{ij} \in [0,1]$ is the probability that i applies a solution strategy to item j (i.e., she or he does not guess), and $g_j \in [0,1]$ is the success probability of solving the item j by

guessing. The g-term is assumed to be person-independent because we assume a pure random guessing process.

Equation 1 is quite general. For example, it reduces to the one- or two-parameter logistic model (1PL or 2PL) when, for each *i* and *j*, s_{ij} is fixed to one, and r_{ij} is specified according to a 1PL or 2PL model. In addition, Equation 1 results in a three-parameter logistic model (3PL) when the g-terms are fixed to one (g_j = 1 for all j = 1, 2, …, J), and the s-terms are not allowed to differ between individuals ($s_{ij} = s_j$ for all i = 1, 2, …, N) (e.g., von Davier, 2009).

Other IRT models build upon all components included in Equation 1. For example, Mislevy and Verhelst (1990) suggested a model in which r_{ij} is represented by the 1PL model, and the probabilities of solution behavior, s_{ij}, are restricted to be constant within individuals, so that individuals either show solution behavior on all items or guess all items in a test (i.e., $s_{ij} = s_i$ for all j = 1, 2, .., J with s_i = 0, 1). The authors suggested fixing the g-terms to the success probabilities expected under pure random guessing. Yamamoto (1989) suggested the HYBRID model, where r_{ij} is represented by a 2PL structure, and the g-terms of Equation 1 can be either fixed or freely estimated. In the HYBRID model, s_{ij} can vary across items within individuals. However, the probabilities are restricted to be either zero or one (s_{ij} = 0, 1), and only irreversible switches from solution to guessing behaviour over the course of a test are allowed. Therefore, individuals' switching points to guessing behaviour can be regarded as indicators of their test-taking persistence. However, the assumption of a sudden switch appears to be too rigid, especially in low-stakes testing situations. In such situations, it appears more reasonable to assume smooth declines of s_{ij} over the course of a test.

The HYBRID Model and the Continuous HYBRID Model

In the following sections we introduce the HYBRID model in a parameterization that differs from Yamomoto's (1989) original formulation. We do so to better outline the connection between the HYBRID model and the subsequently introduced C-HYBRID model.

The HYBRID Model

One possibility to distinguish solution from guessing behaviour is to introduce a latent class variable C_{ij} that is allowed to vary between individuals and items. The value of C_{ij} indicates whether individual *i* responds to item *j* by applying solution ($C_{ij} = 1$) or random guessing behaviour ($C_{ij} = 0$). This means that the component s_{ij} of Equation 1 corresponds to the probability of person i belonging to the class $C_{ij} = 1$ when

working on item j, $P(C_{ij} = 1) = s_{ij}$. Based on the latent class variable, we can define the success probabilities as

$$\begin{aligned} P(Y_{ij} = 1|C_{ij} = 1) &= r_{ij}\,, \\ P(Y_{ij} = 1|C_{ij} = 0) &= g_j\,. \end{aligned} \tag{2}$$

The success probabilities under solution behaviour are represented by a 2PL model:

$$P(Y_{ij} = 1|C_{ij} = 1) = \frac{\exp(\alpha_j\theta_i - \nu_j)}{1 + \exp(\alpha_j\theta_i - \nu_j)}, \tag{3}$$

where α_j is an item discrimination, ν_j is an item threshold parameter, and θ is the continuous proficiency variable. Thresholds can be converted to item difficulties as $\beta_j = \nu_j/\alpha_j$.

Under guessing behaviour, the probability of a correct response corresponds to

$$P(Y_{ij} = 1|C_{ij} = 0) = \frac{\exp(-\tilde{\nu}_j)}{1 + \exp(-\tilde{\nu}_j)}, \tag{4}$$

which means that the success probabilities under random guessing could be item-specific. Such differences might, for example, arise when multiple-choice items differ in their number of response options. Indeed, in the case of multiple-choice items, we propose fixing the $\tilde{\nu}$-parameters because random guessing implies a "blind" selection of response options (Rogers, 1999). For an item j with K_j response options, the success probability under random guessing corresponds to $g_j = 1/K_j$, which implies a $\tilde{\nu}$-parameter (Equation 4) of $\tilde{\nu}_j = \log(K_j - 1)$.

In Yamomoto's (1989) HYBRID model, class membership is handled in a deterministic way by introducing an individual's specific switching point D_i that indicates the item up to which the individual i has applied a solution strategy. The variable D is integer valued in the range of $1 \leq D \leq J$, and is related to latent class membership as

$$\begin{aligned} C_{ij} &= 1, \quad \text{if} \quad D_i \geq j\,, \\ C_{ij} &= 0, \quad \text{if} \quad D_i < j\,. \end{aligned} \tag{5}$$

Equation 5 means that the value of D_i refers to the last item in the sequence on which individual i has shown solution behaviour. At the minimum value of $D_i = 1$ the individual has started to guess after the first item, whereas at the maximum value of $D_i = J$ the individual has shown solution behavior on all items. Because Equation 5 represents a deterministic assignment of individuals to the response modes, the switching point variable can by itself be considered as a between-individual latent class variable that replaces the within-individual latent class variable C_{ij}.

A problem faced in applications of the HYBRID model stems from the large number of latent classes that make it necessary to estimate $J - 1$ latent class proportions. As a solution, Cao and Stokes (2008) applied a probability function to the distribution of D that allows the latent class proportions to be expressed as a function of two parameters. A second challenge in applications of the HYBRID model is the estimation of

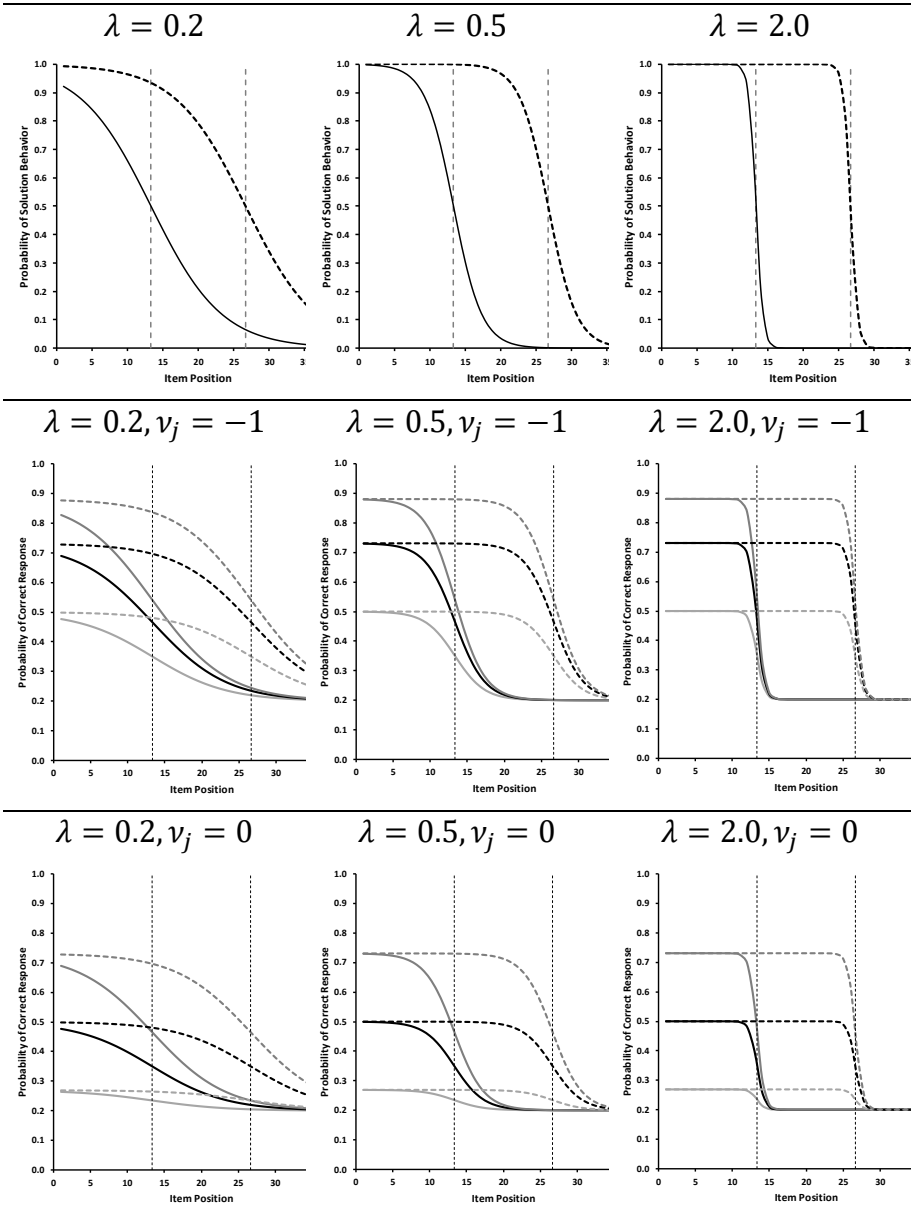
the joint distribution of latent variables. Following the advice of Yamamoto and Everson (1997), List, Robitzsch, Lüdtke, and Nagy (2017) specified θ to be linearly related to D (i.e., the longer an individual exhibits solution behaviour, the higher or lower their proficiency is expected to be).

The Continuous HYBRID Model

The C-HYBRID model shares two assumptions of the HYBRID model, namely (1) the distinction between two types of response behaviour, and (2) the serial order of item responses. Our model builds upon the latent class variable C_{ij} (Equation 2) and describes the responses given under the two response modes as in the HYBRID model (Equations 3 and 4). The key difference to the HYBRID model is that we assume a stochastic relationship between class membership and switching points. To this end we introduce a continuous latent variable δ that represents individual differences in test-taking persistence. Formally, δ_i indicates the (switching) point in the test where individual *i* has an equal probability of showing either solution or guessing behaviour. The persistence variable δ is related to the probability of solution behavior as

$$P(C_{ij} = 1) = \frac{\exp[\lambda(\delta_i - j)]}{1 + \exp[\lambda(\delta_i - j)]}. \quad (6)$$

Equation 6 includes a discrimination parameter λ that relates the difference $\delta_i - j$ to the probability of showing solution behavior. When $\delta_i = j$, an equal chance of applying either a solution strategy or a guessing strategy to item j exists. For positive differences $\delta_i - j > 0$, the probability of providing a solution-based response dominates, whereas in the opposite case of $\delta_i - j < 0$, the probability of an individual providing a random guess is higher. The parameter λ controls how quickly individuals are expected to switch from the solution to the guessing mode. Because the persistence variable and the items' positions are given on the same scale, λ has two interpretations, which do not contradict each other. First, for a fixed value of δ, λ stands for the change in the log-odds of applying solution behaviour when moving from item j to the next item $j + 1$. Second, for item j, λ represents the difference in the log-odds of working in the solution state for two individuals who differ by one unit in δ. Therefore, we refer to λ as a parameter that refers to the *process discrimination*.



Continued on the next page

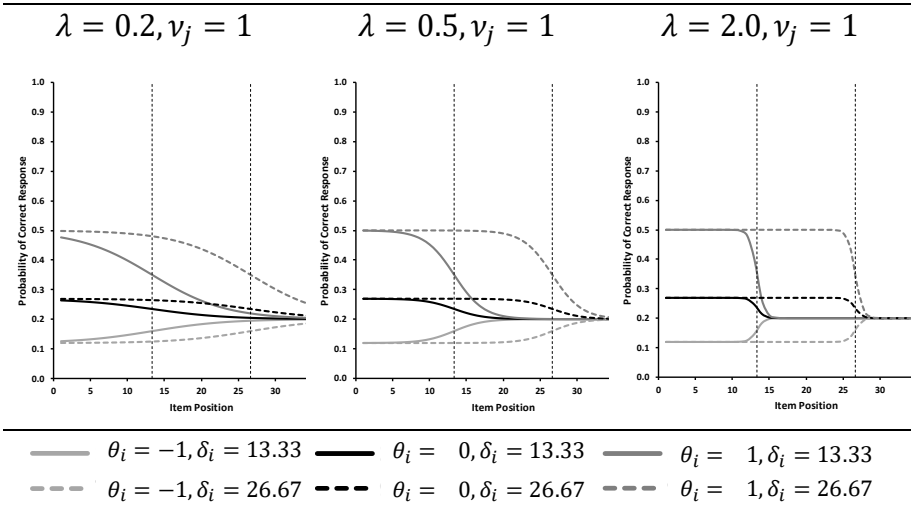


Figure 1

Upper panels: Probabilities of solution behavior by item position. Lower panels: Probabilities of correct response by item position. Examples for different levels of process discrimination (λ; columns) for easy, intermediate, and hard items (ν; rows). Probability curves for three different levels of proficiency (θ) and two levels of persistence (δ)

When λ is large, almost all individuals with a value of δ slightly higher than j (for instance, $\delta_i = j + 0.33$) apply a solution strategy to item j and randomly guess the responses to all items that follow j. In this case, individuals are expected to suddenly switch from the solution-based response state to the guessing state (as assumed in the HYBRID model). In contrast, when λ is small, only a slightly higher proportion of the individuals with the aforementioned value of δ shows solution behaviour rather than guessing behaviour on item j, while for the following items, the proportion of guessed responses exceeds the proportion of solution-based responses. This means that, for a fixed level of δ, the C-HYBRID model predicts the prevalence of guessed responses to increase over the course of the test without precluding solution-based responses in item positions higher than the individuals' values of δ. The upper panels in Figure 1 exemplify the role of the process discrimination parameter λ by means of plots of the probabilities of solution behaviour for two levels of test-taking persistence δ. Note that although item positions are discrete, the probabilities of providing a solution-based response in each position are connected to highlight the logistic shape of their evolvement across positions. The value of δ describes the point over the course of a test at which solution behaviour is expected to occur with a probability of 0.5, whereas λ represents the steepness of the curves at this point.

The C-HYBRID model states that, in the solution-based state, the probabilities of correct responses depend on the individuals' proficiencies, $P(Y_{ij} = 1|C_{ij} = 1)$ (Equation 3) and that, in the guessing state, they correspond to the chance level of success, $P(Y_{ij} = 1|C_{ij} = 0)$ (Equation 4). Because the latent class variable is stochastically related to persistence (upper panels in Figure 1), the model implies that, for a given combination of θ_i and δ_i, the probability of a correct response to item *j* is in between the values of $P(Y_{ij} = 1|C_{ij} = 1)$ and $P(Y_{ij} = 1|C_{ij} = 0)$ (Equation 1). This issue is visualized in the lower panels in Figure 1, which include examples of the probabilities of correct responses for three levels of proficiency (θ), which are combined with two levels of persistence (δ). To highlight the key aspects, the panels refer to a situation where all item discriminations and item difficulties take the same values. As shown in Figure 1, $P(Y_{ij} = 1|C_{ij} = 1)$ and $P(Y_{ij} = 1|C_{ij} = 0)$ are the boundary values of the probabilities of correct responses, the value of δ indicates the position in the test in which the probability lies exactly in the middle of the boundary values (Equation 1), and the process discrimination parameter λ controls how quickly the boundary values are approached. As shown in the figure, this interpretation does not depend on the items' difficulties. However, the items' difficulties, together with the individuals' proficiencies, determine the value of $P(Y_{ij} = 1|C_{ij} = 1)$. For easy items, $P(Y_{ij} = 1|C_{ij} = 1)$ is larger than $P(Y_{ij} = 1|C_{ij} = 0)$ for almost all individuals. In this case, low levels of persistence imply declining solution probabilities over the course of the test for almost all individuals. In hard items, low persistence can be associated with increases in performance in individuals with low proficiencies because $P(Y_{ij} = 1|C_{ij} = 1)$ can be smaller than $P(Y_{ij} = 1|C_{ij} = 0)$. Therefore, the C-HYBRID model appears to be more difficult to estimate in tests composed of many hard items. We investigated this issue in a simulation presented in the later parts of the article.

Multilevel Mixture Model Specification in Mplus. IRT models can be formulated and estimated as multilevel models with item responses represented in a long format (e.g., Van den Noortgate, De Boeck, & Meulders, 2003). The item responses of each individual *i* are stacked in a $J \times 1$ vector $\boldsymbol{y}_i$ and are regressed on *J* variables that indicate the items to which the responses in $\boldsymbol{y}_i$ belong. The main effects of the item indicators represent item thresholds, and their interaction effects with the proficiency variable represent item discriminations. The multilevel setup can be extended to a MMM by the inclusion of an item-level latent class variable. Asparouhov and Muthén (2008) have shown how the 3PL model can be specified, and Pokropek (2016) has demonstrated how item-level covariates can be used to predict latent class membership.

The C-HYBRID model can be estimated on the basis of the typical MMM setup, for example, by using the M*plus* software. However, in our experience, at least when the M*plus* software is used, the inclusion of item indicators and their interactions drastically increases the computational burden and therefore leads to long estimation times. As a solution to this problem, we propose a rearrangement of the input data that avoids the use of item indicator variables. In our setup the item responses are organized in a diagonal format. Each individual *i*'s responses are represented in a matrix $\boldsymbol{Y}_i$ of order

$J \times J$. The elements of this matrix are denoted as y_{ijk}, with missing responses in the off-diagonal entries ($j \neq k$). Individuals' arrays of item responses are augmented by an additional column in which the item positions are coded (vector $\boldsymbol{l}_i$). In order to avoid empirical identification problems arising from the fact that guessing behavior typically does not occur in the first positions of a test, we suggest treating the very last position J as the reference position (i.e., $l_{ij} = 0$), and expressing the remaining positions as $l_{ij} = J - j$. Table A1, presented in Appendix A, provides an example of a data array for one individual.

The next step is to accommodate the measurement Equations 3 and 4 to fit into the MMM framework. To this end, we specify the item parameters to be located on the item level (within-individual level), and the proficiency variable to be located on the (between) individual level. To this end, Equation 3 is modified to

$$P\left(Y_{ijk} = 1 | C_{ij} = 1\right) = \frac{\exp\left(\alpha_k w_{ij} - \nu_k\right)}{1 + \exp\left(\alpha_k w_{ij} - \nu_k\right)}, \tag{7}$$

where $\alpha_k = \alpha_j$ and $\nu_k = \nu_j$ for $j = k$, whereas Equation 4 is changed to

$$P\left(Y_{ij} = 1 | C_{ij} = 0\right) = \frac{\exp\left(\tilde{\alpha}_k w_{ij} - \tilde{\nu}_k\right)}{1 + \exp\left(\tilde{\alpha}_k w_{ij} - \tilde{\nu}_k\right)}, \tag{8}$$

with $\tilde{\alpha}_k = 0$ for all $k = 1, 2, \ldots, J$ and $\tilde{\nu}_k = \tilde{\nu}_j$ for $j = k$.

In Equations 7 and 8, the terms w_{ij} stand for the individual i's scores on a variable that is fully determined by a $J \times 1$ unit vector $\boldsymbol{u}_i$ with elements $u_{ij} = 1$:

$$w_{ij} = \theta_i u_{ij}, \tag{9}$$

which means that the proficiency variable θ is represented by a random effect located at the between-individual level. In order to ensure model identification, we specify θ to follow a standard normal distribution ($\mu_\theta = 0$ and $\sigma_\theta^2 = 1$).

In the MMM specification, latent class membership is modelled as

$$P\left(C_{ij} = 1\right) = \frac{\exp\left(\tau + \lambda l_{ij} + \zeta_i\right)}{1 + \exp\left(\tau + \lambda l_{ij} + \zeta_i\right)}, \tag{10}$$

where τ is a logistic regression intercept, and λ is a logistic regression weight that corresponds to the process-discrimination parameter of Equation 6. Finally, ζ_i is a normally distributed individual-specific disturbance with mean of zero and variance σ_ζ^2. In our setup, θ and ζ are specified to follow a bivariate normal distribution with zero mean vector and a partially structured covariance matrix: $\begin{bmatrix}\theta \\ \zeta\end{bmatrix} \sim BVN\left(\begin{bmatrix}0 \\ 0\end{bmatrix}, \begin{bmatrix}1 & \\ \rho_{\theta,\zeta}\sigma_\zeta & \sigma_\zeta^2\end{bmatrix}\right)$.

Note that Equation 10 does not refer to the δ-variable of Equation 6. It rather includes a rescaled version of δ, which might be defined as $\delta_i^* = \tau + \zeta_i$, such that

$$\delta_i = J + \frac{\tau + \zeta_i}{\lambda}. \tag{11}$$

Therefore, the estimates derived in the MMM can be converted to receive the mean and variance of the δ-variable (Equation 6), with mean given by

$$\mu_\delta = J + \frac{\tau}{\lambda}\,, \tag{12}$$

and variance represented as

$$\sigma_\delta^2 = \frac{\sigma_\zeta^2}{\lambda^2}. \tag{13}$$

The C-HYBRID model can be estimated by marginal maximum likelihood (MML) via the expectation maximization (EM) algorithm. MML estimation follows from the procedures outlined by Vermunt (2003). The log-likelihood to be maximized is given by

$$\log L = \sum_{i=1}^{N} \log P(\boldsymbol{Y}_i|\boldsymbol{l}_i)\,. \tag{14}$$

As only the diagonal entries of $\boldsymbol{Y}_i$ include information, Equation 14 is equivalent to the log-likelihood expression in which item responses are represented in a $J \times 1$ vector $\boldsymbol{y}_i$. Therefore, the likelihood of the data remains the same regardless how item responses are organized.

The case-wise likelihood function provided by the C-HYBRID model can be written as

$$P(\boldsymbol{Y}_i|\boldsymbol{l}_i) = \int_\theta \int_\zeta \left[\prod_{j=1}^{J} P\left(Y_{ij} = y_{ij}|l_{ij}, \theta, \zeta; \boldsymbol{\Psi}\right)\right] f(\theta, \zeta; \boldsymbol{\Psi}) d\theta d\zeta\,, \tag{15}$$

where $\boldsymbol{\Psi}$ refers to the collection of model parameters.

Special Cases of the C-HYBRID Model. The C-HYBRID model assumes a positive process discrimination parameter ($\lambda > 0$). When the λ-parameter is zero, the persistence variable is not defined (Equation 11), and in the case of a negative λ-parameter, the δ-variable would not match our interpretation of test-taking persistence (i.e., the probability of solution behaviour would increase over the course of the test). However, in the MMM setup presented, different special cases of the C-HYBRID model can be formulated by imposing specific parameter constraints.

First, as already explicated, the C-HYBRID model includes the 2PL model as a special case. The 2PL model can be deduced by fixing the $\tilde{\nu}$-parameters (Equation 8) to an arbitrary value, the process discrimination to $\lambda = 0$, the variance and the covariance of the disturbance to $\sigma_\zeta^2 = \rho_{\theta,\zeta}\sigma_\zeta = 0$, and the logistic intercept to a large positive value, such as $\tau = 30$ (Equation 10). These restrictions imply that all responses reflect solution behaviour.

Second, when the λ-parameter is fixed to a very high positive value, the C-HYBRID model becomes similar to Yamamoto's (1989) HYBRID model in the sense that switches from solution to guessing behaviour can be considered to be deterministic (Figure 1). However, this version of the HYBRID model assumes a bivariate normal joint distribution of proficiency and switching points to guessing behaviour, whereas in typical applications of the HYBRID model, different distributional assumptions are imposed (e.g., List et al., 2017).

Third, when λ is fixed to zero, when the variance of ζ is fixed to a very large value, and when the covariance $\rho_{\theta,\zeta}\sigma_\zeta$ is fixed to zero, the item-level latent class variable C is essentially turned into a between-individual latent class variable (Asparouhov & Muthén, 2008). In this model a subset of individuals is expected to show solution behaviour on all items, whereas a second group is expected to guess all responses. This model reflects a 2PL version of the guessing mixture model of Mislevy and Verhelst (1990). An interesting variant of this model arises when the (co-)variance structure of ζ is freely estimated. In this case, ζ governs individual differences in the propensity of guessing any item in the test.

Fourth, when the variance of the individual-level component ζ (including $\rho_{\theta,\zeta}\sigma_\zeta$) is fixed to zero, the model turns into a constrained variant of the four-parameter logistic model (4PL; Barton & Lord, 1981). Referring to Equation 1, the lower probability asymptote of a correct response is given by $\left(1 - s_j\right)g_j$, whereas the upper asymptote corresponds to $s_j + \left(1 - s_j\right)g_j$. When λ is different from zero, the s-terms have a logistic relationship with item positions.

Taken together, the MMM framework in which the C-HYBRID model is specified is flexible, as it allows a variety of models that appear as special cases to be specified. Not all special models are substantively appealing in every application but the constrained models might serve as a statistical benchmark that could be used to scrutinize the C-HYBRID model.

A Simulation Study

The main goal of the simulation study was to examine the C-HYBRID model's capability of recovering the data-generating parameters in a variety of conditions that are likely to be encountered in practice (different sample sizes, levels of test difficulty, and proportions of guessed responses). In order to keep the simulation manageable, we decided to focus on the bias and the accuracy of parameter estimates, including the item parameters, the parameters pertaining to the persistence variable (i.e., process discrimination parameter λ, and the mean and the variance of the δ-variable), and the relationships of test-taking persistence with proficiency and a covariate. We did not evaluate the accuracy of the standard errors and the inferences based on them (e.g., coverage rates). This endeavour would have required a much larger number of replications, each involving additional and time-consuming computational effort to

calculate the standard errors. Given the large number of conditions, the additional burden could not be handled with the available resources.

The second goal of the simulation was to study the impact of neglecting guessing behaviour by using a misspecified 2PL model. We chose the 2PL model for two reasons. First, this model is probably the IRT model most commonly used in large-scale assessments. Therefore, evaluating the robustness of the 2PL model's results in face of guessing behaviour is also of substantive interest. Second, the 2PL model is a natural competitor to the C-HYBRID model, as it is a special case of it. For this reason, we also examined whether the models can be distinguished on the basis of the Bayesian information criterion (BIC), which has been recommended to evaluate (multilevel) mixture IRT models (Li, Cohen, Kim, & Cho, 2009; Sen, Cohen, & Kim, 2019).

We assumed a multiple-choice test with five response options per item, such that the probability of randomly guessing the correct response was g_j = .20 for all items. In line with our former arguments, we fixed the parameter $\tilde{v}_j$ (Equation 4) to result in the theoretical success probability under random guessing [i.e., $\tilde{v}_j = \log(5-1)$ for all j = 1, 2, …, J]. An example of an M*plus* input file that includes one covariate is provided in the Appendix B.

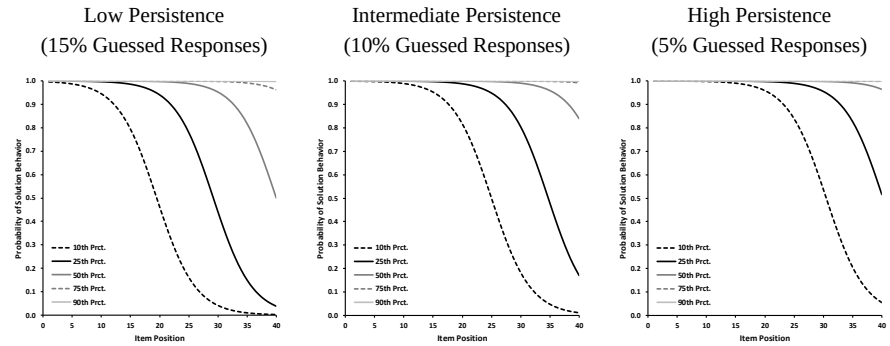


Figure 2

Evolvement of solution behaviour by percentiles of persistence as specified in the simulation study in conditions reflecting low, intermediate, and high persistence.

We simulated data reflecting 18 different conditions with R = 100 replications per condition in a test consisting of J = 40 items. The first factor manipulated was the level of test-taking persistence. We specified three different scenarios in which approximately 15%, 10%, and 5% of item responses were guessed in the total sample.

To this end we set the mean of the δ-variable to μ_δ = 40.0, 45.5, and 51.0, respectively. The standard deviation of the δ-variable was set to σ_δ = 16, and the process discrimination was set to λ = 0.3 in all conditions. Figure 2 presents the progression of solution behaviour used in this simulation.

The second factor manipulated was the difficulty of the test. Here, we specified tests with average item difficulties of $\bar{\beta} = -0.5$, 0.0, and 0.5 (easy, intermediate, and hard). Item difficulties were drawn from a standard normal distribution in the intermediate-difficulty condition and were then shifted upward and downward in the remaining two conditions. Item discriminations were drawn from a uniform distribution and were set to be invariant across conditions. We manipulated the difficulty factor because it is plausible that test-taking persistence is better identified in relatively easy tests. In such cases, the change from solution to guessing behaviour is accompanied by stronger performance declines (Figure 1). The data-generating item parameter values are given in Table 1.

The third factor manipulated was sample size. We considered two scenarios with N = 1,000 and N = 2,000. These sample sizes are rather small for typical large-scale assessments.

The correlations between the proficiency variable θ, the persistence variable δ, and the covariate x were set to $\rho_{\theta,\delta}$ = 0.4, $\rho_{\theta,x}$ = 0.3, and $\rho_{\delta,x}$ = 0.4, respectively. These variables were repeatedly (i.e., R = 100 times) sampled from a multivariate normal distribution and were used to generate item responses with success probabilities as represented in the C-HYBRID model. Within each condition, the C-HYBRID model was estimated two times: one time without the inclusion of the covariate, and one time in which the covariate was included in the model. The data-generating population values were used as starting values in the simulation without considering multiple random starting values, as it is common practice when fitting mixture models (Lubke & Muthén, 2005). Our decision to do so was motivated by our desire to save computational time. In a number of replications, we checked whether our procedure coincided with the results obtained by using multiple starting values, and found this to be the case.

All parameter estimates $\hat{\psi}$ were examined with respect to average bias, defined as

$$Bias(\psi) = \frac{\sum_{r=1}^{R}(\hat{\psi}_r - \psi)}{R}, \tag{16}$$

and with respect to the root mean square error (RMSE):

$$RMSE(\psi) = \sqrt{\frac{\sum_{r=1}^{R}(\hat{\psi}_r - \psi)^2}{R}}. \tag{17}$$

Table 1

Population Values of Item Parameters used in the Simulation Study by Level of Test Difficulty (Easy, Intermediate, and Hard).

	All Tests	Easy Test	Interm. Test	Hard Test		All Tests	Easy Test	Interm. Test	Hard Test
Item	α_j	β_j	β_j	β_j	Item	α_j	β_j	β_j	β_j
1	0.61	-0.85	-0.35	0.15	21	0.54	-2.01	-1.51	-1.01
2	1.18	1.01	1.51	2.01	22	1.05	-0.92	-0.42	0.08
3	0.75	0.38	0.88	1.38	23	0.62	-2.25	-1.75	-1.25
4	0.88	-0.28	0.22	0.72	24	1.12	1.66	2.16	2.66
5	1.41	-1.70	-1.20	-0.70	25	1.22	-0.59	-0.09	0.41
6	0.75	-1.21	-0.71	-0.21	26	0.90	-1.84	-1.34	-0.84
7	1.41	0.48	0.98	1.48	27	1.37	0.58	1.08	1.58
8	1.32	-2.66	-2.16	-1.66	28	0.87	-0.66	-0.16	0.34
9	1.01	-1.29	-0.79	-0.29	29	0.51	-0.22	0.28	0.78
10	0.82	-0.78	-0.28	0.22	30	1.36	-1.13	-0.63	-0.13
11	0.68	0.06	0.56	1.06	31	0.83	-1.06	-0.56	-0.06
12	1.18	-1.58	-1.08	-0.58	32	0.95	-0.99	-0.49	0.02
13	0.76	1.25	1.75	2.25	33	1.10	-0.53	-0.03	0.47
14	1.05	0.29	0.79	1.29	34	1.36	-1.48	-0.98	-0.48
15	0.51	-0.47	0.03	0.53	35	0.97	0.84	1.34	1.84
16	1.20	-0.34	0.16	0.66	36	1.44	0.70	1.20	1.70
17	0.66	-0.15	0.35	0.85	37	0.98	-0.08	0.42	0.92
18	1.27	-0.72	-0.22	0.28	38	1.46	-1.38	-0.88	-0.38
19	0.78	-0.41	0.09	0.59	39	0.55	0.21	0.71	1.21
20	1.25	0.13	0.63	1.13	40	1.07	-0.02	0.49	0.99

Note. α_j = Item discriminations, β_j = Item difficulties ($\beta_j = \nu_j/\alpha_j$).

To study the impact of neglecting guessing behaviour on the individuals' proficiency estimates, we compared the proficiencies implied by the 2PL model with the population distribution that was simulated to be in accordance with the C-HYBRID model. To this end, we sampled N = 100,000 proficiency and persistence scores (θ and δ) from a bivariate normal distribution as specified in each simulation condition (i.e., three persistence levels times three test-difficulty levels). For each individual combination of θ_i and δ_i, we then estimated the proficiency score expected for the 2PL model, $\hat{\theta}_{i2PL}$, by minimizing the loss function

$$\sum_{j=1}^{J}\{\text{logit}[P(Y_{ij}=1|\theta_i,\delta_i)]-\text{logit}[P(Y_{ij}=1|\hat{\theta}_{i2PL})]\}^2, \quad (18)$$

where $P(Y_{ij}=1|\theta_i,\delta_i)$ was defined on the basis of the population values of the C-HYBRID model. In contrast, $P(Y_{ij}=1|\hat{\theta}_{i2PL})$ was based on the average parameter estimates given by the 2PL model at a sample size of N = 2,000. As the values of $\hat{\theta}_{i2PL}$ were not on the metric of the original θ-variable, the estimates $\hat{\theta}_{i2PL}$ were equated to the metric of θ by utilizing a quadratic loss function of the form

$$\sum_{j=1}^{J}\{\text{logit}[P(Y_{ij}=1|\hat{\theta}_{i2PL})]-\text{logit}[P(Y_{ij}=1|\hat{\theta}^*_{i2PL})]\}^2, \quad (19)$$

where $P(Y_{ij}=1|\hat{\theta}_{i2PL})$ was derived on the basis of Equation 18, and $P(Y_{ij}=1|\hat{\theta}^*_{i2PL})$ was based on the population values of the item difficulties and discriminations. This procedure allowed us to derive values of $\hat{\theta}^*_{2PL}$ that were on the same metric as the original θ-variable. Differences between the means and dispersions of $\hat{\theta}^*_{2PL}$ and the population values (0 and 1) estimate the bias in estimated population means and dispersions that is caused by the misspecification of the 2PL model.

Results

Convergence and Separability from the 2PL Model. All models converged. However, for the C-HYBRID model in the easy-test conditions at a sample size of N = 1,000, six replications provided anomalous item parameter estimates for one item with a very low difficulty, which indicated that virtually all individuals who employed a solution strategy provided correct responses. These results reflected empirical identification problems rather than shortcomings of the C-HYBRID model. Therefore, the six replications were replaced with new draws. In all replications, within each of the 18 conditions, the BIC was in favor of the C-HYBRID model (results available upon request).

Item Parameter Estimates. Figures 3 and 4 provide displays of the bias and RMSEs of the item discriminations (Figure 3) and the item difficulties (Figure 4) derived from the C-HYBRID and the 2PL models. In the C-HYBRID model, item bias did not exceed |10%| in any condition, whereas the item parameters taken from the 2PL model were strongly biased. In the case of the discrimination parameters (Figure 2), some estimates for items with large population parameters located near to the end of the test provided by the C-HYBRID model were associated with somewhat higher levels of bias in the easy-test-low-persistence condition at a sample size of N = 1,000. In the case of the 2PL model, bias was not affected by sample size, but was found to be related to the level of persistence and test difficulty. Interestingly, the more difficult the test was, the more discrimination parameters on average were underestimated by the 2PL model.

In the case of the C-HYBRID model, the precision of the discrimination parameter estimates appeared to be a function of the number of solution-based responses. As such, RMSEs decreased at higher levels of persistence and in the larger samples. In the 2PL model, precision appeared to mainly reflect the degree of model misspecification. In the 2PL model, RMSEs decreased at higher levels of persistence but were not affected by sample size.

In the case of the item difficulty estimates, the results were as follows (Figure 4). The C-HYBRID model provided essentially unbiased results in all conditions, whereas the 2PL model resulted in positively biased estimates. Here, item difficulties were on average overestimated, whereas the degree of bias depended on the degree of model misspecification. The results for the precision of estimates, as measured by the RMSE statistic, mirrored the results for the discrimination parameters. In the C-HYBRID model, RMSEs decreased as a function of the number of solution-based responses, which means that they decreased at higher levels of persistence and in a larger sample size. In the 2PL model, precision depended almost solely on the degree of model misspecification.

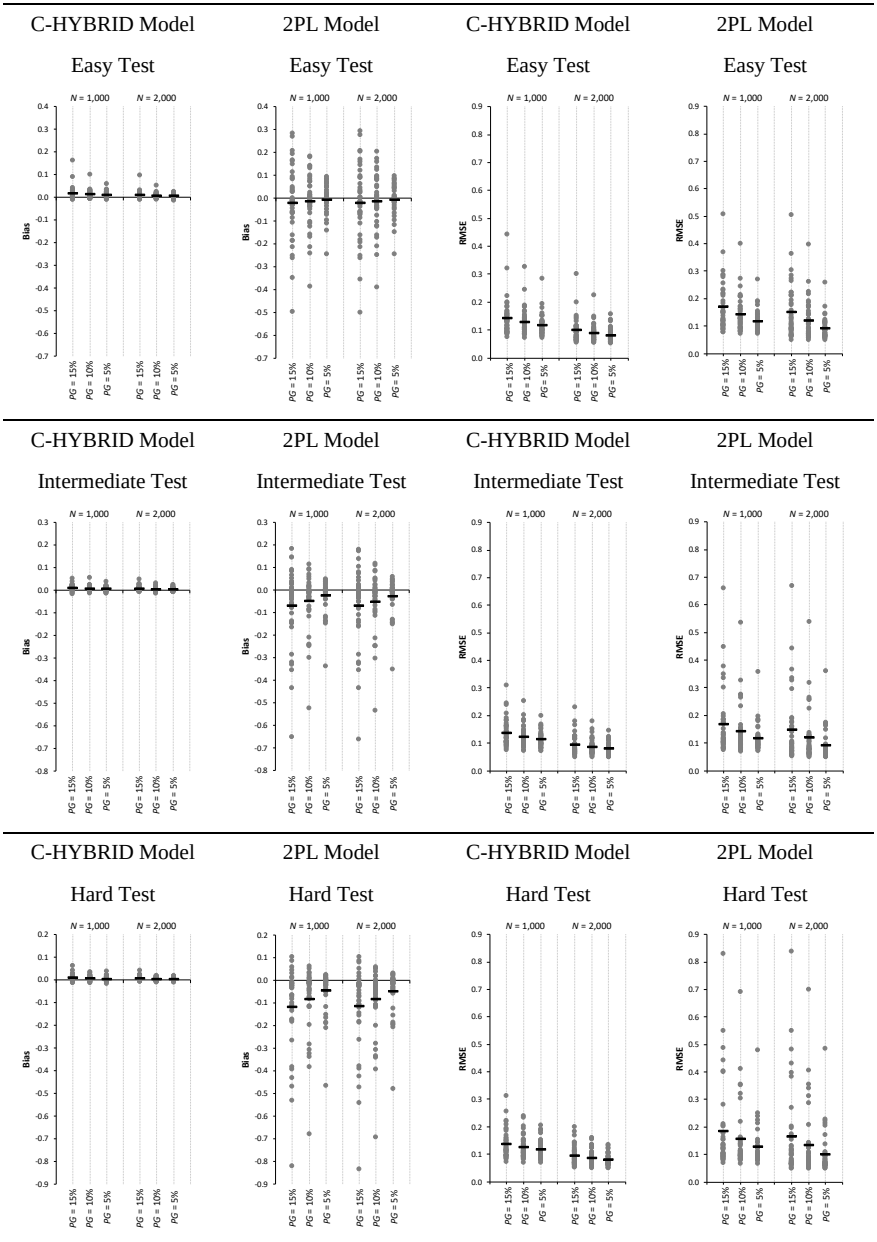


Figure 3

Bias and RMSEs of item discriminations estimated by the C-HYBRID and 2PL models in tests of different difficulty, with different sample sizes (N), and different rates of guessing behavior (PG).

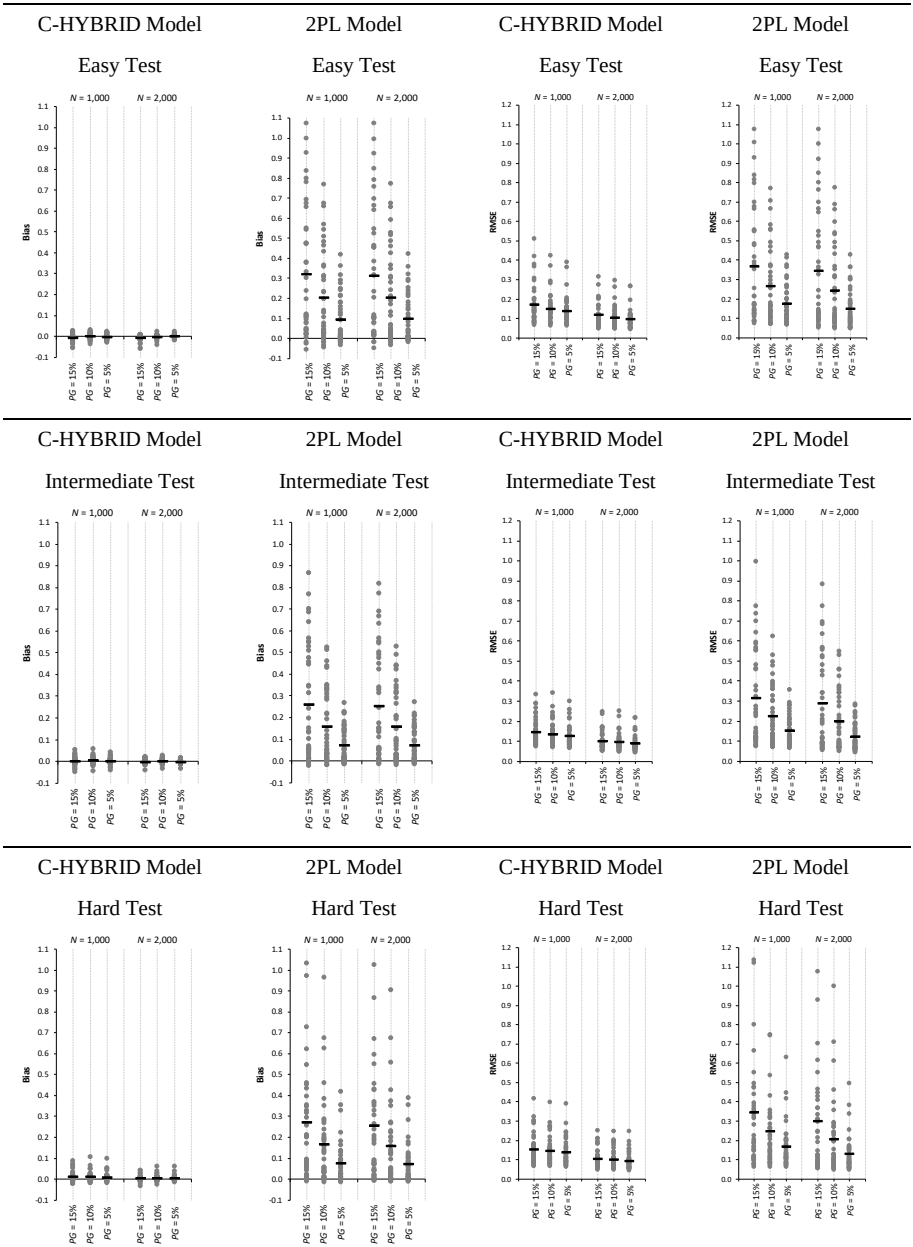


Figure 4

Bias and RMSEs of item difficulties estimated by the C-HYBRID and 2PL models in tests of different difficulty, with different sample sizes (*N*), and different rates of guessing behavior (*PG*).

Assessment of Test-Taking Persistence. We now turn to the C-HYBRID model's capability of recovering the process discrimination parameter λ, the means and standard deviations of the δ-variable, μ_δ and σ_δ, and its correlation with proficiency, $\rho_{\theta,\delta}$. As can be seen in Table 2, the parameter estimates were essentially unbiased in all conditions. The precision of the estimates, as measured by the RMSE, appeared to depend on the number of items solved under guessing behavior. The RMSEs of all parameters were smaller when test persistence was on average low and sample size was large.

Covariate Relationships. Here, we report the results derived on the basis of models that were augmented by a covariate. The inclusion of the covariate had a negligible effect on the item parameter estimates of the C-HYBRID model, and the precision of the estimates improved only slightly by adding the extra information (results available upon request).

Table 3 reports the bias and RMSE statistics of the correlations with the covariate. The results obtained by the C-HYBRID model were unbiased. The precision of the correlation of persistence with the covariate depended to some degree on the amount of guessing behavior and the sample size, such that the precision was higher the more responses were guessed. The precision of this correlation also depended on test difficulty, with higher precision being found in easier tests. In the case of the correlation of proficiency with the covariate, the precision as measured by the RMSE was only affected by sample size.

The results provided in Table 3 demonstrate that when the 2PL model was used, the relationships of proficiency with the covariate were positively biased in certain conditions. The bias and RMSE statistics were affected by the degree of persistence and the difficulty of the test. In easy tests, the relative bias reached or even exceeded 20% at intermediate or low levels of persistence (10% and 15% guessed responses), and was still over 10% in situations with higher test persistence (5% guessed responses). In tests of intermediate difficulty, the relative bias was about 20% in conditions where 15% of the items were guessed, and exceeded 10% at intermediate levels of guessing (10% guessed responses). Finally, in hard tests, only conditions with a low level of test persistence revealed nonnegligible biases (relative bias greater than 10%).

Table 2

Bias and RMSE of Estimates of Process Discrimination (λ), Mean Persistence (μ_δ), Standard Deviation of Persistence (σ_δ), and Correlation of Persistence and Proficiency ($\rho_{\theta,\delta}$).

	Easy Test		Intermediate Test		Hard Test	
	Bias	RMSE	Bias	RMSE	Bias	RMSE
15% *Guessed Responses*; N = 1,000						
λ	0.00	0.03	0.01	0.04	0.01	0.05
μ_δ	0.06	1.18	0.24	1.64	0.18	2.54
σ_δ	−0.08	1.14	−0.09	1.49	−0.09	1.78
$\rho_{\theta,\delta}$	0.00	0.06	−0.01	0.07	−0.01	0.08
15% *Guessed Responses*; N = 2,000						
λ	0.00	0.02	0.00	0.03	0.01	0.04
μ_δ	−0.21	0.86	0.11	1.14	−0.02	1.35
σ_δ	−0.20	0.79	0.00	0.91	−0.11	0.94
$\rho_{\theta,\delta}$	0.01	0.05	0.00	0.05	0.00	0.06
10% *Guessed Responses*; N = 1,000						
λ	0.01	0.04	0.02	0.07	0.02	0.07
μ_δ	0.32	1.38	0.42	2.01	0.65	3.52
σ_δ	0.07	1.33	0.09	1.84	0.40	2.55
$\rho_{\theta,\delta}$	−0.01	0.07	−0.01	0.07	−0.02	0.10
10% *Guessed Responses*; N = 2,000						
λ	0.00	0.03	0.01	0.05	0.01	0.05
μ_δ	0.08	0.92	0.44	1.45	0.67	2.19
σ_δ	−0.01	0.79	0.25	1.11	0.59	1.74
$\rho_{\theta,\delta}$	0.00	0.04	−0.01	0.05	−0.01	0.06
5% *Guessed Responses*; N = 1,000						
λ	0.01	0.05	0.02	0.09	0.03	0.08
μ_δ	0.23	1.63	0.26	2.24	−0.01	3.73
σ_δ	−0.09	1.59	−0.12	1.87	−0.39	2.48
$\rho_{\theta,\delta}$	0.00	0.07	0.00	0.08	0.00	0.09
5% *Guessed Responses*; N = 2,000						
λ	0.00	0.04	0.01	0.06	0.02	0.07
μ_δ	0.14	0.96	0.26	1.35	0.11	2.26
σ_δ	0.03	0.88	0.12	1.12	−0.13	1.61
$\rho_{\theta,\delta}$	0.00	0.04	0.00	0.05	0.01	0.06

Table 3

Bias and RMSE of Estimated Covariate Correlations with Proficiency ($\rho_{\theta,x}$) and Persistence ($\rho_{\delta,x}$) in the C-HYBRID (CH) Model and the 2PL Model (2PL).

	Easy Test		Intermediate Test		Hard Test	
	Bias	RMSE	Bias	RMSE	Bias	RMSE
15% *Guessed Responses*; N = 1,000						
$\rho_{\delta,x}$ (CH)	0.00	0.04	0.00	0.05	0.00	0.05
$\rho_{\theta,x}$ (CH)	0.00	0.04	0.00	0.04	0.00	0.04
$\rho_{\theta,x}$ (2PL)	0.07	0.08	0.06	0.06	0.04	0.05
15% *Guessed Responses*; N = 2,000						
$\rho_{\delta,x}$ (CH)	0.00	0.03	0.00	0.03	0.00	0.04
$\rho_{\theta,x}$ (CH)	0.00	0.02	0.00	0.02	0.00	0.02
$\rho_{\theta,x}$ (2PL)	0.07	0.08	0.06	0.06	0.04	0.04
10% *Guessed Responses*; N = 1,000						
$\rho_{\delta,x}$ (CH)	0.00	0.05	−0.01	0.06	0.00	0.07
$\rho_{\theta,x}$ (CH)	0.00	0.04	0.00	0.03	0.00	0.04
$\rho_{\theta,x}$ (2PL)	0.06	0.06	0.04	0.05	0.02	0.04
10% *Guessed Responses*; N = 2,000						
$\rho_{\delta,x}$ (CH)	0.00	0.03	−0.01	0.04	0.00	0.04
$\rho_{\theta,x}$ (CH)	0.00	0.02	0.00	0.02	0.00	0.02
$\rho_{\theta,x}$ (2PL)	0.06	0.06	0.04	0.04	0.02	0.03
5% *Guessed Responses*; N = 1,000						
$\rho_{\delta,x}$ (CH)	0.00	0.07	−0.01	0.08	0.00	0.09
$\rho_{\theta,x}$ (CH)	0.00	0.04	0.00	0.03	0.00	0.04
$\rho_{\theta,x}$ (2PL)	0.03	0.05	0.02	0.04	0.01	0.03
5% *Guessed Responses*; N = 2,000						
$\rho_{\delta,x}$ (CH)	0.00	0.04	0.00	0.05	0.00	0.06
$\rho_{\theta,x}$ (CH)	0.00	0.02	0.00	0.02	0.00	0.02
$\rho_{\theta,x}$ (2PL)	0.03	0.04	0.02	0.03	0.01	0.02

Proficiency Estimates in the 2PL Model. The means and standard deviations of the equated proficiencies θ^*_{2PL} (Equations 18 and 19) are reported in Table 4. The degree of test-taking persistence affected the means most strongly in the easy-test conditions. Here, proficiency means were most strongly underestimated in the presence of a high proportion of guessed responses. In hard tests, means were almost unaffected by persistence. Standard deviations showed a different relationship. Here, the impact of test-taking persistence was strongest in the hard-test conditions and rather negligible in the easy-test conditions.

Table 4

Means and Standard Deviations of Proficiencies Expected in the 2PL Model ($\hat{\theta}^{*}_{2PL}$) by Test Difficulty and Level of Test-Taking Persistence (Percentage of Guessed Responses).

	Easy Test		Intermediate Test		Hard Test	
	M	(*SD*)	*M*	(*SD*)	*M*	(*SD*)
15% Guessed Responses	−0.21	(0.96)	−0.12	(0.91)	−0.02	(0.87)
10% Guessed Responses	−0.14	(0.97)	−0.07	(0.94)	0.00	(0.90)
5% Guessed Responses	−0.06	(0.98)	−0.03	(0.96)	0.01	(0.94)

These results were not unexpected. Guessing is more likely to reduce test performance in easy tests than in hard tests because, in easy tests, even individuals with low proficiencies reduce their success probabilities when they guess. This pattern does not hold in hard tests because individuals with low proficiencies might even increase their success probabilities when they guess. Therefore, means can be expected to be most strongly underestimated in easy tests where a large proportion of individuals guess. However, in the case of dispersions, the effect is reversed. In hard tests, individuals with low proficiencies are likely to improve their test score by guessing, so that individual differences in proficiency become blurred.

Summary

The results of the simulation study provided support for the C-HYBRID model. In the conditions studied, the model provided reasonable levels of parameter recovery even in samples of N = 1,000 cases and in relatively difficult tests. The C-HYBRID model was reliably separated from the 2PL model and prevented biases in item parameter estimates that occurred in the 2PL model. In addition, the C-HYBRID model appeared as a useful tool for studying the relationships of test-taking persistence with proficiency and other individual covariates. Therefore, the simulation study indicates the C-HYBRID model's potential (1) to examine the prevalence of suboptimal levels of test-taking persistence, (2) to study the impact of test-taking persistence on the estimates of item parameters, covariate relationships, and individual proficiencies, and (3) to assess the relationships of individual differences in persistence with proficiency and covariates.

An Empirical Application of the C-HYBRID Model

In this section we demonstrate the C-HYBRID model's feasibility on the basis of real data. To this end, we compared the results obtained by the C-HYBRID model when it was applied to two test forms assessing proficiency in science. Both test forms were administered to students from the same population, whereby students were randomly assigned to the test forms. If the C-HYBRID model assessed the essence of test-taking persistence adequately, it should provide similar estimates of parameters pertaining to the persistence variable in each test form.

The data was taken from the KESS 7 study (Kompetenzen und Einstellungen von Schülerinnen und Schülern an Hamburger Schulen zu Beginn der Jahrgangsstufe 7 [Competencies and attitudes of pupils at Hamburg schools at the start of Grade 7]; Bos, Bonsen, & Gröhlich, 2009), which is a low-stakes assessment of seventh-grade students conducted in Hamburg (Germany). The school structure in Hamburg is composed of different secondary school tracks, with academic track students typically achieving the highest scores in standardized tests (e.g., Maaz, Trautwein, Lüdtke, & Baumert, 2008). As the curriculum in the academic track differs in many respects from the curriculum in nonacademic tracks (e.g., Baumert, Stanat, & Watermann, 2006), academic track students were excluded. In addition, we included only students who had completed a cognitive ability measure that was based on figural analogies tasks (Heller & Perleth, 2000); this measure was used as a covariate. The sample sizes for test forms A and B were N_A = 2,008 and N_B = 1,858, respectively. The groups that responded to the different test forms did not differ in the distribution of nonacademic school tracks and cognitive ability.

Both test forms comprised 38 multiple-choice items, each with four response options. Of these items, 19 items were included in both test forms in the same positions (6 to 24). Prior to the analyses, we subjected all items to a conventional 2PL model. Based on the results, we eliminated three items with discrimination parameters near to zero from the test. One item was common to both test forms, and two items were specific to test form A. In the main analyses, we fitted a 2PL model, as well as a C-HYBRID model (with success probability under guessing behavior fixed to g_j = 0.25 for all j = 1, 2, …, J items) separately to each test form. Both models included the cognitive ability measure as a covariate that was allowed to correlate with all latent variables. We employed multiple starting values to check whether the best log-likelihood value could be replicated. This was the case for both test forms.

The main question was whether the C-HYBRID model that was fitted separately to each test form provided similar estimates of parameters related to the persistence variable. This research question was evaluated on the basis of Wald-χ^2 tests that were used to compare the estimates of the process discrimination parameters (λ), of the means and standard deviations of the persistence variable (μ_δ and σ_δ), as well as of the correlations of persistence with proficiency and cognitive ability ($\rho_{\theta,\delta}$ and $\rho_{\delta,x}$). These analyses were supplemented by a C-HYBRID model that was fitted to the combined sample (N_{A+B} = 3,866). This model not only assumed that the (joint) distribution

of the persistence variable was unrelated to the test form, but also that the item parameters of the common items were invariant across test forms.

Results

Judged on the information criteria, the C-HYBRID model provided a better fit than the 2PL model to both test forms (Table 5). In addition, the C-HYBRID model provided similar results for both test forms.

Table 5

Model-Data Fit of 2PL and C-HYBRID Models Fitted Separately and Jointly to the Test Forms of the Science Test.

	# Par.	LL	AIC	BIC	SBIC
Test Form A					
2PL	73	-44486.04	89118.1	89527.2	89295.3
C-HYBRID	78	-44423.61	89003.2	89440.4	89192.6
Test Form B					
2PL	77	-43284.32	86722.6	87148.2	86903.6
C-HYBRID	82	-43195.72	86555.4	87008.7	86748.2
Combined Sample					
C-HYBRID[A]	160	-87619.33	175558.7	176560.3	176051.8
C-HYBRID (Invariant)[B]	116	-87637.14	175506.3	176232.4	175863.8

Note. LL = Model log likelihood, AIC = Akaike information criterion, BIC = Bayesian information criterion, SBIC = Sample size adjusted BIC, A = The log-likelihood value corresponds to the sum of C-HYBRID models fitted separately to each test form, B = The model assumes full invariance of common item parameters and the joint distribution of the proficiency, persistence, and cognitive ability variables.

The results indicated that, in both test forms, a substantive number of item responses reflected guessing behavior (Form A: 16%, Form B: 14%). As shown in Table 6, the estimates of the λ-parameters, of the means and dispersions of the δ-variable, as well as of the relationship of δ with proficiency and cognitive ability were similar across samples, and no estimate differed significantly between test forms. The results indicated (1) that the probabilities of providing solution-based responses decreased smoothly over the course of the test (λ-parameters relatively close to zero), and (2) that many students did not show low test-taking persistence (means of δ larger than the number of items), although (3) there were clear individual differences in the degree of test-taking persistence (reliable variability in δ). In addition, the results showed (4) that students with higher proficiency tended to be more persistent (positive

correlations of δ with proficiency), and (5) that the same relationship held for cognitive ability (positive correlations of δ with cognitive ability).

Table 6

Parameter Estimates Taken from C-HYBRID Models Fitted Separately to Test Forms A and B, Wald-χ^2 Test Statistic (df = 1) of Comparison of Separately Estimated Parameters, and Parameter Estimates Based on a Fully Invariant C-HYBRID Model Fitted Simultaneously to Both Test Forms.

	Test Form A	Test Form B	Comparison	Joint Estimates
	Est. (*SE*)	*Est.* (*SE*)	Wald-χ^2	*Est.* (*SE*)
λ	0.12 (0.02)	0.15 (0.03)	0.83	0.14 (0.01)
μ_δ	43.99 (3.42)	44.54 (3.69)	0.01	44.40 (1.83)
${\sigma_\delta}^{\mathrm{A}}$	16.51 (2.27)	16.29 (2.19)	0.01	16.20 (1.36)
$\rho_{\theta,\delta}$	0.47 (0.11)	0.37 (0.16)	0.30	0.40 (0.07)
$\rho_{\delta,x}$	0.24 (0.05)	0.19 (0.05)	0.37	0.20 (0.03)
$\rho_{\theta,x}$	0.48 (0.03)	0.44 (0.03)	1.12	0.46 (0.02)

Note. A = Comparison of dispersions was carried out on the basis of the standard errors of the logarithms of the variance estimates [$\log(\hat{\sigma}_\delta^2)$].

The invariance of results was further supported by a C-HYBRID model that was fitted to the combined data. As shown in Table 5, the model fitted to the combined data set did not result in a decrement of fit as compared to the overall fit derived on the basis of the test-form-specific C-HYBRID models (information criteria derived on the basis of the sum of the log-likelihood values of the models fitted separately to each test form). The joint model provided more favorable values of information criteria, and did not result in a statistically significant decrement of fit as judged on the basis of a likelihood-ratio test [$\chi^2(df = 44) = 35.61$, $p = .812$]. As expected, the estimates pertaining to the persistence variable were very close to the estimates derived on the basis of the test-form-specific C-HYBRID models (Table 5).

The upper left panel of Figure 5 provides a graphical description of the evolvement of solution behavior as predicted by the C-HYBRID model fitted to the full sample (displays for various percentiles of the distribution of the δ-variable). A sizable proportion of students were expected to show an early reduction in solution behavior. The lower left panel in Figure 5 compares the item threshold parameters provided by the 2PL model with the estimates derived by the C-HYBRID model. As expected, the more closely the items were located towards the end of the test, the more different the

parameters provided by the two models were. For these items, most estimates were lower in the C-HYBRID model, thereby indicating that the 2PL model might overestimate the difficulty of items located in later parts of the test (see also Figure 4). The lower right panel of Figure 5 compares the item discriminations provided by the 2PL and the C-HYBRID models. The estimates for the items located in the first half of the test (Positions 1 to 17) were in good agreement, whereas the estimates for the items presented later in the test diverged. This finding is again in accordance with the findings of the simulation study (Figure 3).

Finally, the upper right panel of Figure 5 compares the proficiency estimates (EAP scores) derived on the basis of the 2PL and the C-HYBRID models. Here, proficiency estimates from the 2PL model were equated to the metric of the C-HYBRID model (Equation 19). Proficiency estimates for students whose persistence was estimated to be high showed a high correspondence, but proficiency estimates for students with lower levels of persistence were estimated to be lower in the 2PL model. As such, the 2PL model provided, on average, lower proficiency estimates. Because of the positive relationship of proficiency and persistence (Table 6), the systematic discrepancy was especially pronounced in students with lower achievement. Again, this finding was in line with the results of the simulation study that showed that proficiencies were underestimated by the 2PL model in easy tests (Table 4).

Summary

The application documented the usefulness of the C-HYBRID model in low-stakes assessments. The model provided a strong indication that the performance declines were in line with the proposed conceptualization of test-taking persistence. In addition, the application provided some evidence for the robustness of the C-HYBRID model because the estimates pertaining to the persistence variable showed a good agreement between two test forms, and the pattern of results referring to the differences between the 2PL and the C-HYBRID model were well in line with the results of the simulation study. The findings once again document that the estimates for items located nearer to the end of a test might be biased when a significant proportion of individuals start to guess item responses (e.g., Oshima, 1994).

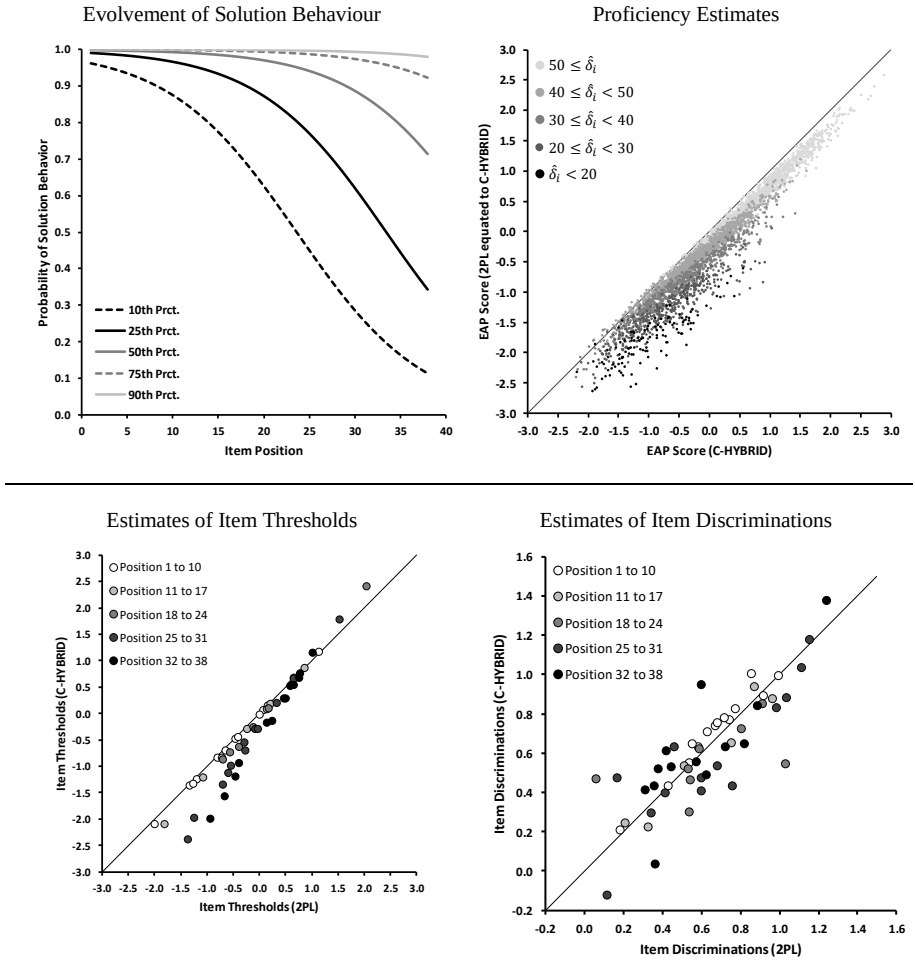


Figure 5

Evolvement of solution behavior by percentiles of persistence (upper left panel), comparison of proficiency estimates derived by the C-HYBRID and the 2PL model by levels of estimates of persistence (upper right panel), comparison of item parameter estimates given by the C-HYBRID and the 2PL model by item positions (lower left panel: item thresholds, lower right panel item discriminations).

Furthermore, the application demonstrated the C-HYBRID model's usefulness for studying individual-level correlates of test persistence. Such applications are

interesting from an applied perspective because they have the potential to shed light on individuals' reasons for reducing their test effort (e.g., Lindner, Lüdtke, & Nagy, 2019). In the present case, the results suggest science proficiency to be the more important factor than general cognitive ability because proficiency correlated more strongly with persistence, and the partial correlation between persistence and cognitive ability accounted for science proficiency was close to zero ($\hat{\rho}_{\delta,x\cdot\theta}$ = .02). However, this result certainly does not mean that there are no other individual difference variables that have a meaningful relationship with persistence over and above proficiency (e.g., Nagy, Nagengast, Frey, Becker, & Rose, 2019).

Discussion

The aim of the present article was to introduce a further development of the HYBRID model that was initially proposed by Yamamoto (1989). The C-HYBRID model relaxes the HYBRID model's assumption of sudden and irreversible switches from solution to guessing behavior that can occur over the course of a test. Our model assumes a stochastic relationship between the probability of applying solution (vs. guessing) behavior and item position, which means that the model acknowledges that individuals might switch back and forth between solution and guessing behavior in a limited part of the item sequence of a test. As such, the C-HYBRID model is likely to be appropriate for low-stakes tests where individuals might switch to guessing behavior for reasons other than running out of time (i.e., test speededness).

On the item level the C-HYBRID model assumes a gradual reduction in solution behavior and allows differences between individuals in the onset of this process. This specification is in line with current substantive research. For example, research based on self-reports of motivation and effort has found these characteristics to decline over the course of a test (Lindner, Nagy, & Retelsdorf, 2018). Similar findings have been reported with respect to RGB (e.g., Wise, Pastor, & Kong, 2009), with some findings indicating that the onset point of RGB is the main characteristic defining individual differences in response time effort (Lindner, Lüdtke, & Nagy, 2019).

Aspects of Empirical Applications

Besides its connection to well-validated patterns of test-taking behavior, the C-HYBRID model has several features that make it interesting for applied researchers. First, the model appears to provide results of reasonable accuracy even in samples of modest size ($N \approx 1{,}000$). We are not aware of how other models perform at this sample size, although there is some evidence that mixture models for test speededness perform reasonably in samples of $N = 2{,}000$ examinees (Suh, Cho, & Wollack, 2012), a sample size in which the C-HYBRID model provided good results. However, more simulation studies in which the sample size, as well as other factors, is varied are needed to gain

a more complete understanding of the model's behavior. In this vein, further simulations should also examine other aspects, such as standard error bias and coverage rates. In addition, future studies might also examine likelihood-based approaches of model selection, such as the Lo-Mendell-Rubin Test, which provide *p*-values that could aid decisions in situations where the BIC does not provide a conclusive answer.

Second, the C-HYBRID model can be estimated on the basis of a single test form. In contrast, IRT models for item position effects that have been discussed as a method for modeling test-taking persistence (Debeer et al., 2014) require rotated booklet designs in which items are presented in different positions. In addition, such IRT models do not built upon the concept of random guessing, which means that item responses impacted by low persistence are assumed to depend on proficiency (List et al., 2017). In contrast, the C-HYBRID model is not limited to fixed test forms and can be straightforwardly altered to accommodate rotated assessment designs. We expect that rotated designs can increase the efficiency and precision of the estimates provided by the C-HYBRID model (Weirich, Hecht, & Böhme, 2014), although more research is needed to explore the potential of more complex designs.

Other mixture IRT models that have been suggested in the context of test speededness can also be estimated on the basis of fixed test forms (e.g., List et al., 2017; Suh et al., 2012). However, most of these models make assumptions that are unlikely to hold in reality, such as assumptions about (1) deterministic change points (e.g., Yamomoto, 1989), (2) the existence of a single class of guessers with a known onset point of performance decline (e.g., Bolt, Cohen, & Wollack, 2002), (3) effects of proficiency on the probabilities of correct solutions at low levels of motivation (e.g., Jin & Wang, 2014), and (4) the same success probability of guessing under solution and guessing behavior (e.g., Goegebeur et al., 2008). We believe that the C-HYBRID model has desirable features that are not included in other models, although future research should more thoroughly compare the existing models, possibly on the basis of existing data sets.

Third, the C-HYBRID model offers the possibility to include individual-level covariates in the model. In the present article we focused on the correlations of covariates with persistence. However, the model can be easily altered by specifying covariates (including proficiency) to predict persistence. Alternatively, covariates can be expressed as dependent variables that are predicted by persistence. This feature allows the core elements of theories of test-taking motivation to be tested (e.g., Wise, 2017), an endeavor that cannot be easily carried out on the basis of existing mixture IRT models for test speededness.

Model Extensions and Future Research

Although we believe that the C-HYBRID model is well suited to a broad array of applications, the model is not without restrictions. Many restrictions are a consequence of our desire to keep the model parsimonious, as well as to ensure the

interpretability of its key components. However, in certain situations, the model's restrictions might be in conflict with the goals of an investigation.

We focused on situations in which individuals are regarded as members of one population, which means that the model's parameters are assumed to apply to all subpopulations included in a sample. For example, we treated process discrimination λ as a fixed parameter, so that it reflects a property of the test in a population (similar to an item discrimination that represents an item property). However, this restriction can be overcome by specifying the C-HYBRID model as a multigroup model that allows for group differences in the λ-parameters as well as in the means and (co-)variances of the δ-variable. This can be achieved by introducing a between-individual latent class variable with known class memberships (e.g., List et al., 2017). In addition, individual-level latent class membership can be treated as unknown, thereby providing a method for dividing the whole population into several subpopulations in which low test-taking persistence is reflected in different patterns of performance decline (e.g., quicker and slower changes from solution to guessing behavior).

An alternative way for modeling heterogeneity in process discriminations is to express the λ-parameter as a normally distributed random parameter that varies across individuals. Such models can be estimated, but, in our experience, MML estimation via the EM algorithm is computationally very demanding. In addition, the specification of λ as a random parameter comes at the price of blurring the interpretation of the δ-variable as an indicator of test-taking persistence (i.e., individual turning points from solution-based behavior to guessing behavior). Therefore, in order to explore heterogeneity in λ, more work is needed to derive a feasible modeling strategy.

This work should also consider Bayesian estimation techniques as an alternative to MML. Bayesian estimation allows for a large number of random effects and may therefore provide a feasible route for examining heterogeneity in λ. However, although Bayesian estimation has many advantages, it comes at the price of long estimation times. This could render Bayesian estimation impractical in the case of large sample sizes, which are typical for large-scale assessments. Therefore, more research is needed that compares the MML estimation with the Bayesian estimation procedures.

We considered only covariates that are located on the individual level. As many researchers are interested in item-level indicators of response behavior (e.g., response times), it could be interesting to extend the C-HYBRID model to include such information. Pokropek (2016) described a grade-of-membership IRT model that allows guessing behavior to be related to response times and is, in some respects, similar to the C-HYBRID model. Indeed, recent work has demonstrated that such information can be included in the MMM framework in which the C-HYBRID model is specified, which means that the C-HYBRID model can be easily extended to include item-level covariates (Nagy & Ulitzsch, 2021).

References

Asparouhov, T. & Muthén, B. (2008). Multilevel mixture models. In G. R. Hancock and K. M. Samuelson (Eds.), *Advances in Latent Variable Mixture Models* (pp. 27-51). Charlotte, NC: In-formation Age Publishing.

Barton, M. A., & Lord, F. M. (1981). *An upper asymptote for the three-parameter logistic item-response model*. Princeton, NJ: Educational Testing Service.

Baumert, J., Stanat, P., & Watermann, R. (2006). Schulstruktur und die Entstehung differenzieller Lern- und Entwicklungsmilieus [School structure and the development of differential environments for learning and development]. In J. Baumert, P. Stanat, & R. Watermann (Eds.), *Herkunftsbedingte Disparitäten im Bildungswesen: Differenzielle Bildungsprozesse und Probleme der Verteilungsgerechtigkeit* (pp. 95-188). Wiesbaden: VS für Sozialwissenschaften.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement, 39*, 331-348. doi:10.1111/j.1745-3984.2002.tb01146.x.

Bos, W., Bonsen, M., & Gröhlich, C. (2009). *KESS 7–Kompetenzen und Einstellungen von Schülerinnen und Schülern an Hamburger Schulen zu Beginn der Jahrgangsstufe 7 [KESS 7-Competences and attitudes of pupils at Hamburg schools at the start of Grade 7]*. Münster: Waxmann.

Cao, J., & Stokes, S. L. (2008). Bayesian IRT guessing models for partial guessing behaviors. *Psychometrika, 73*, 209-230. doi: 10.1007/s11336-007-9045-9

Debeer, D., Buchholz, J., Hartig, J., & Janssen, R. (2014). Student, school, and country differences in sustained test-taking effort in the 2009 PISA reading assessment. Journal of Educational and Behavioral Statistics, 39, 502-523. doi: 10.3102/1076998614558485

Debeer, D., & Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement, 50*, 164-185. doi: 10.1111/jedm.12009

Goegebeur, Y., De Boeck, P., Wollack, J. A., & Cohen, A. S. (2008). A speeded item response model with gradual process change. *Psychometrika*, 73, 65-87. doi: 10.1007/s11336-007-9031-2

Heller, K. A., & Perleth, C. (2000). *Kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision*. Göttingen: Beltz.

Jin, K.-Y., & Wang, W.-C. (2014). Item response theory models for performance decline during testing. *Journal of Educational Measurement, 51*, 178-200. doi:10.1111/jedm.12041.

Li, F., Cohen, A. S., Kim, S.-H., & Cho, S.-J. (2009). Model selection methods for mixture dichotomous IRT models. *Applied Psychological Measurement, 33*, 353–373. doi: 10.1177/0146621608326422.

Lindner, C., Nagy, G., & Retelsdorf, J. (2018). The need for self-control in achievement tests: Changes in students' state self-control capacity and effort investment. *Social Psychology of Education, 21*, 1113-1131. https://doi.org/10.1007/s11218-018-9455-9

Lindner, M. A., Lüdtke, O., & Nagy, G. (2019). The onset of rapid-guessing behavior over the course of testing time: A matter of motivational and cognitive resources. *Frontiers in Psychology*, Provisionally accepted. doi: 10.3389/fpsyg.2019.01533

List, M. K., Robitzsch, A., Lüdtke, O., Köller, O., & Nagy, G. (2017). Performance decline in low-stakes educational assessments: different mixture modeling approaches. *Large-scale Assessments in Education, 5*. doi: 10.1186/s40536-017-0049-3

Lubke, G. H., & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods, 10*, 21-39. doi:10.1037/1082-989X.10.1.21.

Maaz, K., Trautwein, U., Lüdtke, O., & Baumert, J. (2008). Educational transitions and differential learning environments: How explicit between-school tracking contributes to social inequality in educational outcomes. *Child Development Perspectives, 2*, 99-106. doi:10.1111/j.1750-8606.2008.00048.x

Meyers, J. L., Miller, G. E., & Way, W. D. (2009). Item position and item difficulty change in an IRT-based common item equating design. *Applied Measurement in Education, 22*, 38-60. doi: 10.1080/08957340802558342

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55*, 195-215. doi: 10.1007/bf02295283

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus User's Guide. Eighth Edition*. Los Angeles, CA: Muthén & Muthén.

Nagy, G., Nagengast, B., Frey, A., Becker, M., & Rose, N. (2019). A multilevel study of position effects in PISA achievement tests: student-and school-level predictors in the German tracked school system. Assessment in Education: *Principles, Policy & Practice*, 26, 422-443. doi: 10.1080/0969594x.2018.1449100

Nagy, G., & Ulitzsch, E. (2021). A multilevel mixture IRT framework for modeling response times as predictors or indicators of response engagement in IRT models. *Educational and Psychological Measurement.* Advance online publication. doi: 10.1177/00131644211045351

Oshima, T. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement, 31*, 200-219. doi:10.1111/j.1745-3984.1994.tb00443.x.

Plummer, M. (2017). *JAGS Version 4.3.0 Manual*. International Agency for Research on Cancer, Lyon, France.

Pokropek, A. (2016). Grade of Membership Response Time Model for Detecting Guessing Behaviors. *Journal of Educational and Behavioral Statistics, 41*, 300-325. doi: 10.3102/1076998616636618

Rogers, H. J. (1999). Guessing in multiple-choice tests. In G. N. Master & J. P. Keeves (Eds.), *Advances in Measurement in Educational Research and Assessment* (pp. 235-243). Amsterdam: Pergamom.

Sen, S., Cohen, A. S., & Kim, S. H. (2019). Model selection for multilevel mixture Rasch models. *Applied psychological measurement, 43*, 272-289. doi: 10.1177/0146621618779990

Suh, Y., Cho, S.-J., & Wollack, J. A. (2012). A comparison of item calibration procedure in the presence of test speededness. *Journal of Educational Measurement, 49*, 285-311. doi:10.1111/j.1745-3984.2012.00176.x.

Van den Noortgate, W., De Boeck, P., & Meulders, M. (2003). Cross-classification multilevel logistic models in psychometrics. *Journal of Educational and Behavioral Statistics, 28*, 369-386. doi: 10.3102/10769986028004369

Vermunt, J. K. (2003). Multilevel latent class models. *Sociological Methodology, 33*, 213-239. doi: 10.1111/j.0081-1750.2003.t01-1-00131.x

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

von Davier, M. (2009). Is there need for the 3PL model? Guess what? *Measurement, 7*, 110–114. doi:10.1080/15366360903117079

Weirich, S., Hecht, M., & Böhme, K. (2014). Modeling item position effects using generalized linear mixed models. *Applied Psychological Measurement, 38*, 535-548. doi:10.1177/0146621614534955

Wise, S. L. (2017). Rapid-Guessing Behavior: Its Identification, Interpretation, and Implications. *Educational Measurement: Issues and Practice, 36*, 52–61. doi: 10.1111/emip.12165

Wise, S. L., & Kong, X. J. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*, 163-183. doi: 10.1207/s15324818ame1802_2

Wise, S. L., Pastor, D. A., & Kong, X. J. (2009). Correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice. *Applied Measurement in Education, 22*, 185–205. doi:10.1080/08957340902754650

Yamamoto, K. (1989). *Hybrid Model of IRT and Latent Class Models* (ETS Research Rep. No. RR-89-41). Princeton, NJ: Educational Testing Service.

Yamamoto, K., & Everson, H. (1996). Modeling the effects of test length and test time on parameter estimation using the HYBRID model. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 89-98). New York: Waxmann

Appendix A: Example of the Diagonal Arrangement of Item Responses

Table A1 provides an example of the diagonal data array for one individual that is used to estimate the C-HYBRID model in the MMM setup. The matrix $\boldsymbol{Y}_i$ of order $J \times J$ contains the individual i's item responses with elements denoted as y_{ijk}. All off-diagonal entries ($j \neq k$) in $\boldsymbol{Y}_i$ are treated as missing (denoted as "-99"). The matrix $\boldsymbol{Y}_i$ is augmented by a column vector $\boldsymbol{l}_i$ in which the item positions are coded with entries $l_{ij} = J - j$.

Table A1

Example of Data Array for one Individual in a Hypothetical Test with J = 30 Items

	y_{i1}	y_{i2}	y_{i3}	⋯	y_{i28}	y_{i29}	y_{i30}	l_i
y_{i1}	1	−99	−99	⋯	−99	−99	−99	29
y_{i2}	−99	1	−99	⋯	−99	−99	−99	28
y_{i3}	−99	−99	0	⋯	−99	−99	−99	27
⋮	⋮	⋮	⋮	⋱	⋮	⋮	⋮	⋮
y_{i28}	−99	−99	−99	⋯	0	−99	−99	2
y_{i29}	−99	−99	−99	⋯	−99	1	−99	1
y_{i30}	−99	−99	−99	⋯	−99	−99	0	0

Note. "−99" denotes missing values.

Appendix B: M*plus* Syntax for Estimating the C-HYBRID Model

The syntax refers to an application with $J = 40$ items and one individual-level covariate x. We consider a situation where all items have five response options, and success probabilities when guessing are fixed to the value of $g_j = 0.20$ for all $j = 1, 2, \ldots, J$ items.

```
Title:      Example of C-HYBRID Model with one Covariate
Data:       file is Exampledata.dat;
Variable:   names are case y01-y40 pos x;
            usevariables are y01-y40 pos x;
            missing are all (-99);
            categorical are y01-y40;
            within are y01-y40 pos;
            between are x;
            cluster is case;
            classes are c(2);
Analysis:   type is twolevel random mixture;
            estimator is ml;
            starts = 80 16;
            processors = 8;
            integration = standard(21);
Model:      %within%
            %overall%
            unit by;
            node by y01-y40;
            [node@0 unit@1];
            node@0 unit@0;
            [y01$1-y38$1];
            theta | node on unit;
            c#1 ON pos (b);
            [c#1*0.72770 ] (a);
            %c#2%
            node by y01-y40@0;
            [y01$1-y38$1*1.386] (nu);
            %between%
            %overall%
            theta@1 x (vx);
            [theta@0 x];
            c#1 (v);
            c#1 with theta (cdt);
            c#1 with x (cdx);
            theta with x (ctx);
Model Constraint:
            new(md vd sd codt codx cotx);
            md = a/b + 40;
            vd = v*(b**(-2));
            sd = sqrt(vd);
            codt = cdt/sqrt(v);
            codx = cdx/sqrt(v*vx);
            cotx = ctx/sqrt(vx);
            nu = ln(4);
Output:     tech1 tech8;
```