Quantitative and Computational Methods in Behavioral Sciences

PsychOpen GOLD

Fundamental Research Articles

Disengage When Returns Diminish: Modeling Solution Behavior and Disengagement in Achievement Items With a Single-Process Model for Responses and Response Times

Sören Much [1,2,3], Augustin Mutak [2], Steffi Pohl [2], Jochen Ranger [1]

[1] *Institut für Psychologie, Martin-Luther-Universität Halle-Wittenberg, Halle (Saale), Germany.* **[2]** *Department of Education and Psychology, Freie Universität Berlin, Berlin, Germany.* **[3]** *Wilhelm-Wundt-Institute for Psychology, Leipzig University, Leipzig, Germany.*

Quantitative and Computational Methods in Behavioral Sciences, 2026, Article e14893, https://doi.org/10.5964/qcmb.14893

Received: 2024-06-25 • **Accepted:** 2025-12-11 • **Published (VoR):** 2026-06-16

Handling Editor: Tina Braun, Charlotte Fresenius Hochschule, Wiesbaden, Germany

Corresponding Author: Sören Much, Wilhelm-Wundt Institute for Psychology, Leipzig University, Neumarkt 9-19, 04109 Leipzig, Germany. Tel: +49 (0)341 97 35924. E-mail: soeren.much@uni-leipzig.de

Supplementary Materials: Code, Data, Materials [see Index of Supplementary Materials]

Abstract

Valid assessments with achievement tests hinge on test-takers being motivated to take the test. Existing latent trait models attempt to disentangle competence and motivational influences, but have theoretical limitations. We propose a single-process accumulator model based on the idea that test-takers accumulate information to solve an item at a continuously decreasing rate. A correct response is generated once the information exceeds a solution threshold. The model incorporates disengagement which is governed by the solution process. Once the accumulation rate falls below a critical level, test-takers stop working on the item. Due to the computational intensity of an analytic solution, we compare maximum likelihood, neural network and Bayesian estimators that use a simulation-based likelihood in a simulation study. Using two empirical examples, the model demonstrates good fit to accuracy and response times of individual items and is able to capture various forms of dependencies between accuracy and response times, including non-linear dependencies.

Keywords

disengagement, accumulator model, response times, process modeling

The performance in achievement tests depends on the competence of the test takers, but also on their motivation to take the test (Eklof, 2010; Silm et al., 2020; Thurstone, 1937). Test takers with little motivation, for example, are prone to give hasty responses based on heuristic solutions or rapid guesses (e.g., Wise & Kong, 2005) or to omit the response entirely (Ulitzsch et al., 2020b). As a consequence, any measure of the test takers' latent capacity to solve the items does not reflect their competence level in pure form, but is contaminated by motivational influences. When the competence of the test takers is the aim of the assessment, this is a serious validity problem. It challenges the results from all large-scale assessment programs that are low-stakes tests for test-takers (Butler & Adams, 2007; Eklof, 2010; Wise & DeMars, 2005). In fact, this issue is part of the achievement testing research from its beginning (Spearman, 1927; Webb, 1915).

Several efforts have been made in order to disentangle the different factors underlying the test performance. In the following review, the focus is on latent trait models that relate test performance to traits representing different aspects of test taking. In specific, we review latent trait models for responses and response times (RTs) that aim at separating competence from motivational confounders. As a consequence, we do not review literature on the detection of rapid guesses (e.g., Hong et al., 2020; Wise & Gao, 2017), robust estimation (e.g., Hong & Cheng, 2019; Ranger et al., 2019) or the speed accuracy trade-off (e.g., Domingue et al., 2022; Guo et al., 2019).

The latent trait models can broadly be classified into three classes, mixture models, race models and single-process models. In the mixture models (e.g., Lu et al., 2020; Meyer, 2010; Molenaar et al., 2018; Ulitzsch et al., 2020a; Wang & Xu, 2015), two modes of responding are assumed, a regular response mode and a disengaged response mode. Responses generated within the regular response mode are reflective of the test takers' competence level. Regular responses are modeled by a standard latent trait model. Responses generated within the disengaged response mode are supposed to be unrelated to the test takers' competence level. Race models (Lee & Ying, 2015; Lu & Wang, 2020; Ranger & Kuhn, 2014) are similar to cure mixture models in survival analysis. The models assume a race between the response process and a process of disengagement. In case a test taker requires more time for the solution of an item than he/she is willing to spend, the solution process is interrupted and either an incorrect response or no response at all is given. Wrong responses can thus be generated by the regular response process, provided it is terminated, or by its preliminary termination. Single-process models claim to model different components of the response process directly. The most popular single-process models are psychometric variants of the diffusion model (Kang et al., 2022; Molenaar et al., 2015; Tuerlinckx & De Boeck, 2005; van der Maas et al., 2011; Vandekerckhove et al., 2011). In the diffusion model, the response process is construed as an evidence accumulation process. The accumulation of evidence over time is represented by a Wiener process with a linear drift. Test takers respond in the moment the accumulated evidence exceeds a decision boundary. The linear drift accounts for the

capability to process information and reflects both the test takers' competence and processing speed. The decision boundary is supposed to reflect the test takers' cautiousness, and implies a speed-accuracy trade-off (Tuerlinckx et al., 2016; van der Maas et al., 2011).

All these models are notable efforts to improve the estimation of the competence level of a test taker. Despite their strengths, they also have limitations that should be addressed. Mixture models distinguish only two levels of engagement, thereby ignoring more fine-grained nuances of motivational levels between test takers, between items, or both. A dichotomic view also ignores the possibility to respond on the basis of partial knowledge (Bechger et al., 2005). Race models assume that test takers will interrupt the response process in case it takes too long to find the solution. This accounts for individual differences in effort or persistence. In all race models proposed so far, the time when test takers disengage is set independently from the response process. Disengagement thus acts as a *deus ex machina* that has no relation to the progress a test taker has made so far. This is not realistic as test takers will not stop when they are making progress. Some race models that are based on evidence accumulation also assume that all items can be solved in infinite time, which is another unrealistic assumption (Ranger & Kuhn, 2014).

The diffusion model has its origins in experimental psychology where it has been used successfully in simple perceptual tasks. Whether these models can be transferred to cognitive items in achievement tests, however, is questionable. Diffusion models assume a linear drift which implies that on average the evidence increases or decreases linearly without limit. Such an evolvement over time might be realistic in perceptual decision tasks with a constant stimulus. In items of cognitive tests or problem-solving tasks, however, the knowledge of a test taker will finally be exhausted and no new insights can be gained anymore. This implies an upper boundary for the evidence level that can be achieved. The standard diffusion model also requires binary decisions and can only be used for items with single choice. The application for open questions requires additional non-trivial modifications (van der Maas et al., 2011).

In this paper, we propose a new process model for responses and RTs on test items — the differential ballistic accumulator (DBA) model. The model is an information accumulation model that is based on a ballistic variant of the Ornstein-Uhlenbeck (OU) process (Uhlenbeck & Ornstein, 1930). Unlike a Wiener process with a constant drift, the OU process includes a mean-reverting drift term that translates to a decay of the accumulation rate. According to the DBA model, the test takers accumulate information when processing an item. The momentary level of information of a test taker is represented by an increasing non-linear function with an upper asymptote. As soon as the momentary information exceeds a critical threshold, the correct response is given. However, as a consequence of the upper asymptote, not all test takers are capable to reach the critical threshold. These test takers will terminate the solution process at some time-point. We assume that test takers give up when the growth rate of the information accumulation

process becomes too low. This reflects the idea that test takers will give up in case they feel stuck and spending more time on the problem will be useless. By proposing this model, we hope to overcome some of the limitations of earlier models, like the coarse distinction between two motivational levels in the mixture models, the weak relation between disengagement and the response process in the race models and the linear and unbounded increase of evidence in the diffusion model. We examine the model in a basic form that is applied to single test items to deliver a proof of concept that is not contaminated by additional complexities that arise from the estimation of a full model based on marginalization across items, conditional on stationary trait levels. We outline the associated challenges and possible solutions in the Discussion.

The DBA Model for Responses and Response Times on Test Items

In the paper, we interpret the response process as an information accumulation process. We assume that the amount of information a test-taker has about a test item's solution accumulates over time. We model this evolvement with a ballistic variant of the OU process. This means that the momentary level of information is modeled as the expectation of an OU process and momentary random fluctuations around the expectation are ignored. Also, contrary to a general OU process, the level of information is bound to be positive and starts at zero for $t = 0$. Based on these assumptions, the accumulated information $I(t)$ of a test taker in an item is a positive, strictly increasing and bounded function of the processing time t:

$$I(t) = \mu \cdot (1 - exp(-\alpha \cdot t)). \quad (1)$$

Parameter μ ($\mu \in R^+$) is the upper asymptote of the information accumulation process and is reached when $t \to \infty$. Parameter μ thus determines the maximum level of information that can be reached and reflects aspects of power to solve problems in the sense of Thurstone (1937). Parameter α ($\alpha \in R^+$) determines how fast the upper asymptote is reached. Parameter α reflects the speed of information processing in the sense of Thurstone (1937).

Whether an item is solved depends on whether the correct solution is found before a test taker gives up. In this sense, the response process can be interpreted as a race between the information accumulation process and a disengagement process, even though disengagement is dependent on the accumulation rate. Denote by the solution time t_s the time needed to find the correct solution. It is the time-point when the accumulated information reaches a critical threshold c_1. The critical threshold c_1 represents the information that is necessary in order to solve an item and is related to the item difficulty. Mathematically, the solution time is determined by the relation $I(t_s) = c_1$ when a solution exists. Otherwise the solution time is set to $t_s = \infty$. As the accumulated information is a positive, continuous and strictly increasing function of time, the solution time is unique.

Note that higher values of c_1 imply longer solution times. Additionally, denote by the disengagement time t_d the time when a test taker stops working. It is the time-point when the rate of information accumulation drops below the critical threshold c_2. This is motivated by the idea that test takers stop working when they are hardly making anymore progress. Parameter c_2 is the minimal increase of information per time a test taker is willing to tolerate. Mathematically, the disengagement time is determined by the relation $I(t_s)' = c_2$ where

$$I'(t) = \alpha \cdot \mu \cdot exp(-\alpha \cdot t) \tag{2}$$

is the rate of information accumulation, that is, the first derivative of $I(t)$ with respect to t. Should $I'(t = 0) < c_2$, the disengagement time t_d is set to zero. This reflects the case that the motivation of a test taker is too low to even start working. As the rate of information accumulation is a positive, continuous and strictly decreasing function of time, the disengagement time is unique.

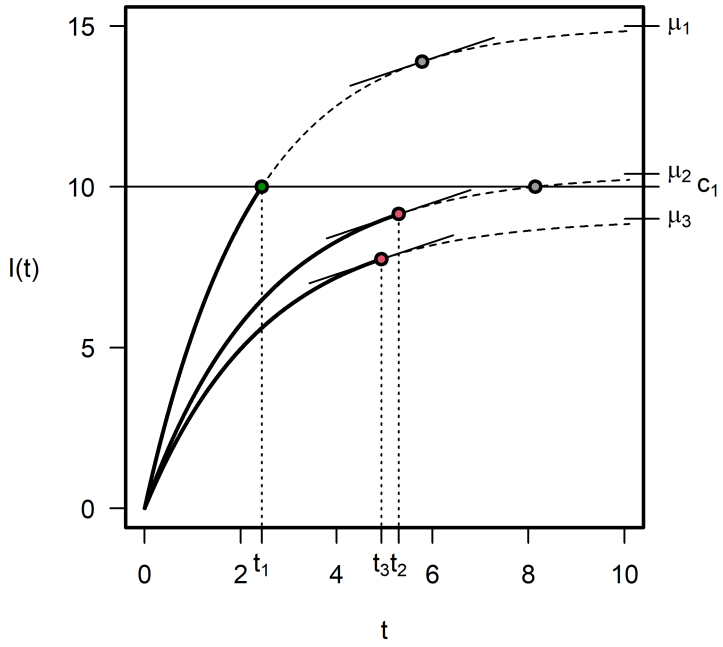
The observed response x and RT t are determined by the solution time t_s and the disengagement time t_d as follows. In case the solution time is shorter than the disengagement time ($t_s < t_d$), the correct response is given ($x = 1$). Otherwise, ($t_d > t_s$) an incorrect response is given in a single choice item or no response is given in an open question ($x = 0$). The observed RT is the shorter of the two times $t = \min(t_S, t_D)$.

The response process is illustrated in Figure 1 for three combinations of μ and α.

Figure 1 visualizes the trajectory of the information accumulation process over time. In addition to the momentary level of information, the accumulation rate is added as a straight line at the points where it drops below the critical threshold. In the first case (μ_1), the process of information accumulation reaches the threshold $c_1 = 10$ at $t_s = 2.2$. The rate of information acquisition drops below the critical level $c_2 = 0.5$ at $t_d = 5.4$. As the solution time is shorter than the disengagement time, the test taker responds correctly at the solution time. In the second case (μ_2), the rate of information acquisition would reach the critical level $c_1 = 10$ at $t_s = 8.1$. The rate of information acquisition, however, drops below the critical level $c_2 = 0.5$ at $t_d = 5.5$. As the disengagement time is shorter than the solution time, the test taker responds incorrectly at the disengagement time. Note that the test taker would finally have been capable to solve the item in case he/she had persisted more. In the third case (μ_3), the accumulated information will never cross the critical level c_1. The solution time t_s is infinite. The rate of accumulation, however, drops below the critical level of $c_2 = 0.5$ at $t_d = 4.9$. The test taker responds incorrectly at the disengagement time. Notably, in this basic model, disengagement always results in an incorrect response. Correct responses only result from successful and complete solutions. There is no mechanism to account for informed or random guessing after disengagement.

Figure 1

Illustration of the Assumed DBA Response Process



Note. The assumed response process for three different combinations of the model parameters (μ = (15, 10.4, 9), α = (.45, .40, .40), c_2 = 0.5). Accumulation process 1 has a high level of maximum information due to a large μ value and is very fast due to both a larger α value and the high asymptote. In contrast, accumulation processes 2 and 3 are slower in accumulating information ($\alpha_2 = \alpha_3$) and do not reach a high maximum level. Process 2 has a higher maximum level of information than process 3 and is therefore a little faster. Even though the maximum level is above the solution threshold c_1 = 10 for process 2, the response process discontinues before reaching this threshold as the tangent slope falls below the disengagement threshold c_2 at time-point t_2.

Up to this point, the model is fully deterministic as the effective model parameters uniquely determine the response and the observed RT. The model cannot account for random variation in the responses and RTs over test takers and items (or trials). This is similar to other ballistic accumulator models (Brown & Heathcote, 2005, 2008). Variability of the responses and RTs can be generated by introducing three sources of random variability connected with the effective parameters α, μ and c_2 of the model. In this basic variant of the DBA, we choose an item-wise perspective and model items separately. This means that there can be no separation between person and item parameters. We assume that the values of the effective parameters in a specific combination of item and test taker are realizations of random variables that account for both individual

differences between the test takers and trial-wise fluctuations. In particular, we assume that parameters in an item are distributed over the test takers in the sample as

$$\begin{aligned} \mu &= exp(\theta) \; with \; \theta \sim N(m_A, s_A^2) \\ \alpha &= exp(\omega) \; with \; \omega \sim N(m_B, s_B^2) \\ c_2 &= exp(\gamma) \; with \; \gamma \sim N(m_C, s_C^2) \end{aligned} \tag{3}$$

In Equation 3, θ, ω and γ are assumed to be independent random variables, following normal distributions. This is a simplifying assumption for the basic model that could be relaxed in further model extensions. As items are modelled separately, random variables are unrelated across a set of items. The parameters m_A, m_B and m_C are related to the expected parameters of the model in an item and a specific population of test takers. The parameters account for the average information level that can be reached, the average solution time and the average disengagement time, respectively. The values of the parameters are determined by characteristics of the item like its difficulty, time demand, and importance and by characteristics of the population of test takers like their average competence, speed and persistence. The standard deviations s_A, s_B and s_C account for the amount of individual differences in the population of the test takers and for unsystematic fluctuations within an item.

In contrast to α, μ and c_2, we assume parameter c_1 to be an item specific constant. This is motivated by the assumption that this parameter represents the level of information that is needed in order to solve an item, a quantity that should be identical for all test takers. Furthermore, the value of c_1 has to be fixed as the model parameters are not identified otherwise. Note that the effects of multiplicative transformations of c_1 on the observed RT can be compensated by corresponding multiplicative transformations of parameter μ and c_2. This parallels similar complications in the linear ballistic accumulator model and the diffusion model, where response thresholds, drift rates, and the variances of process parameters are not identified simultaneously as well. As a consequence, the value of parameter c_1 can be fixed to any arbitrary value greater than zero, for example to $c_1 = 10$. This identification restriction identifies all remaining parameters of the model.

Distribution of the Responses and RTs

The responses and RTs in a test item are a function of the solution time and the disengagement time which in turn are functions of the effective model parameters μ, α and c_2. The distribution of the responses and RTs are thus determined by the distributions of μ, α and c_2 which are given in Equation 3.

Setting Equation 1 and Equation 2 to the thresholds c_1 and c_2 and solving for the solution time t_s and the disengagement time t_d results in

$$t_S = -(1/\alpha) \cdot log(1 - c_1/\mu)$$
$$t_D = -(1/\alpha) \cdot log(c_2/(\alpha \cdot \mu)). \quad (4)$$

Applying standard variable transformation techniques, the distribution of the response and disengagement time can be derived. The distribution of the observed response and RTs then follows via standard findings for the distribution of the minimum. This yields the subdensities for the RTs.

With $\mu^*(t_1, \alpha) = c_1 \bullet 1/(1 - \exp(-\alpha \bullet t_1))$ and $c^*_2(t_1, t_2, \alpha) = c_1 \bullet \alpha \bullet \exp(-\alpha \bullet t_2)/(1 - \exp(-\alpha \bullet t_1))$, and the Jacobians of the transformation $\mathcal{J}_1(t_1, t_2, \alpha) = c_1 \bullet \alpha^2 \bullet \exp(-\alpha \bullet t_2)/(1 - \exp(-\alpha \bullet t_1))$ and $\mathcal{J}_2(t_1, \alpha) = c_1 \bullet \alpha \bullet \exp(-\alpha \bullet t_1)/(1 - \exp(-\alpha \bullet t_1))^2$, the subdensity of the RT for an incorrect response is

$$f(t, x = 0) = \iint Ind((\mu^*(t_1, \alpha) > 10) \& (c_2^*(t_1, t_2, \alpha)/(\alpha \cdot \mu^*(t_1, \alpha)) < 1)) \cdot J_1(t_1, t, \alpha) \cdot J_2(t_1, \alpha) \cdot \phi(\alpha; m_B, s_B) \cdot \phi(\mu^*(t_1, \alpha); m_A, s_A) \cdot \phi(c_2^*(t_1, t, \alpha); m_C, s_C) \cdot dt_1 \cdot d\alpha, \quad (5)$$

and the subdensity of a correct response is

$$f(t, x = 1) = \iint Ind((\mu^*(t, \alpha) > 10) \& (c_2^*(t, t_2, \alpha)/(\alpha \cdot \mu^*(t, \alpha)) < 1)) \cdot J_1(t, t_2, \alpha) \cdot J_2(t, \alpha) \cdot \phi(\alpha; m_B, s_B) \cdot \phi(\mu^*(t, \alpha); m_A, s_A) \cdot \phi(c_2^*(t, t_2, \alpha); m_C, s_C) \cdot dt_2 \cdot d\alpha, \quad (6)$$

where $\phi(x; m, s)$ denotes the density of a lognormal distribution with corresponding scale and shape parameters and Ind(x) is the indicator function.

The distribution has an atom at $t = 0$, $x = 0$ which occurs whenever the initial rate of information accumulation at $t = 0$ is below the threshold c_2. This event occurs when $c_2 > (\alpha \bullet \mu)$ and has probability

$$P(t = 0, x = 0) = \Phi(0, (m_A + m_B - m_C), \sqrt{s_A^2 + s_B^2 + s_C^2}) \quad (7)$$

where $\Phi(x; m, s)$ denotes the distribution function of a normal distribution.

Simulation Study

Estimators

While there is a tractable likelihood function for the model, evaluating the likelihood for the estimation is not feasible as the integral in Equation 5 and Equation 6 has to be approximated numerically for each test taker in the sample, which becomes numerically intensive as this has to be done for each iteration of any estimation procedure. In the simulation study, we explore six simulation-based estimators with different approaches

to this problem and compare their performance in terms of bias and efficiency. We examine two maximum likelihood estimators, one network-based approach and three Bayesian estimators with different methods of likelihood approximation.

Maximum-Likelihood Estimators

We examine two ML estimators to find the item parameters that maximize the likelihood of the joint empirical distributions of responses and RTs. Estimator *ML1* is based on RT categories and fits the deciles of simulated RT distributions from the parameter proposals to the deciles of the empirical RT distribution. *ML2* is based on continuous RTs and fits the log kernel density (Jones et al., 2018) of simulated conditional RT distributions for correct responses, incorrect responses and zero-time responses to the empirical conditional RT distributions. While *ML1* loses information due to the categorization, the approximation of decile distributions is more stable than the density approximation in *ML2*. Inference is challenging for both estimators as the likelihood is not calculated analytically. Also, for both estimators, the approximation is contaminated by simulation error.

Network-Based Estimator

Even simulation-based approaches can be computationally intensive with repeated costs for simulating data and approximating the likelihood. Neural network approaches are an alternative that have one-off costs to train a parameter estimation network on extensive simulated data and are then fast to output parameter estimates from data input.

The basic idea of a parameter estimation network is to train the connection of given data, i.e. item responses and RTs to point estimates of the six model parameters (Lenzi et al., 2023). During training, the network is exposed to simulated data along with the randomly drawn parameters that generated it. This process aims to enable the network to accurately predict parameter values for any new data that follows the same structure. The estimation network we examine (*NN*) is based on a network architecture that contains convolutional layers with local connectivity which allows to uphold the data structure in the hidden layers. That is, the response and RT of each trial from the input data is connected to a common node in the first hidden layer. *NN* contains three convoluted layers, followed by a global average pooling layer and two fully connected layers. Input is generated in batches that contain 1000 pairs of 6 parameters and 1000 simulated trials with responses and RTs. The network is trained on parameter ranges that are derived from the parameter values used in the simulation study (see the subsection on the simulation study's design), taking the range of each parameter expanded by a length of 1 to each side, with 0 being the minimal value for the parameters s_A , s_B , and s_C. For the training, the parameters are randomly drawn from uniform distributions with these ranges:

- m_A ~ u(1.26, 3.84) s_A ~ u(0,1.48)

- m_A ~ u(3.87, −0.29) s_B ~ u(0,1.56)
- m_C ~ u(−4.02, −0.68) s_C ~ u(0,2.99)

The advantage of a network-based approach lies in the rapid estimation once the network is trained. Since simulation-based training is not dependent on empirical data, its one-off costs amortize across the application to any dataset. Inference is possible using a bootstrap approach.

Ideally, the hidden convolutional layers of an estimation network learn sufficient statistics for the estimation. However, since it operates as a black box, it remains uncertain whether this occurs and there is no control over the network's performance for all possible parameter combinations.

Bayesian Estimators

The third group of estimators under examination are Bayesian estimators. The advantage of a Bayesian approach is the availability of credible intervals for parameter inference. Bayesian estimators also accommodate for the variability that is introduced by the simulation.

For all three Bayesian estimators, we employ a differential evolution Markov chain Monte Carlo algorithm (DE-MCMC) (ter Braak, 2006; Turner et al., 2013) that was shown to be an efficient sampler for fitting choice-RT models whose parameters are highly correlated (Turner et al., 2013). The datasets are fitted in 9 chains, each with 20,000 iterations, including a burn-in period of 5,000 iterations. Each of the three estimators uses a unique simulation-based method to approximate the likelihood function, that is implemented in the sampler. As uninformative priors, we chose uniform distributions in parameter ranges that are identical to the ones used to train the network-based estimator *NN*, shown above.

The first Bayesian estimator we examine is based on one of the neural network approaches to approximate the likelihood function as described by Fengler et al. (2021) (*BN*). For this, a fully connected network is trained to predict the likelihood value for given parameters, RT and response type (correct or incorrect).

For the second estimator, the likelihood is approximated with a Beta-Weibull distribution whose parameters are predicted by a neural network that has been trained on simulated data (*BW*).

The third estimator is based on a histogram approach to simulated data to approximate the likelihood function (*BH*). In this approach, the simulated RTs are split for each response type into 20 item-specific ventiles.

Computational Costs

While the examined estimation approaches avoid the computationally expensive analytical likelihood, they are based on simulations that can become computationally costly. The neural network approach has the advantage of becoming cost-efficient after large

one-off costs to train the network. The maximum likelihood approach is in general faster than the Bayesian approach but with variable costs depending on convergence. Bayesian approaches are more costly due to the MCMC sampler, but have the advantage of offering better parameter inference with credible intervals.

Design

In order to investigate the feasibility and performance of the estimators, we conducted a simulation study. We simulated 100 datasets for $p = 5$ items and $N = 1000$ subjects. The values of the model parameters for the five items can be found in Table 1. These values are derived from maximum likelihood estimates of the model fit to the Amsterdam Chess data set (van der Maas & Wagenmakers, 2005, see section Empirical Example for more details). The parameters imply expected RTs between 4s and 13s, and expected solution probabilities ranging from .29 to .59.

Table 1

True Parameter Values Used in the Simulation Study

Parameter	Item 1	Item 2	Item 3	Item 4	Item 5
m_A	2.84	2.72	2.35	2.27	2.26
m_B	-1.36	-2.35	-1.29	-2.87	-1.81
m_C	-1.68	-2.04	-3.02	-2.02	-2.57
s_A	0.31	0.30	0.04	0.48	0.09
s_B	0.08	0.46	0.56	0.08	0.25
s_C	1.16	0.70	1.99	0.46	1.46

Data generation and all estimators were implemented in R (Version 4.2.0, R Core Team, 2022). Network models were implemented with the *keras* package (Version 2.9, Allaire & Chollet, 2022), depending on *tensorflow* (Version 2.9, Allaire & Tang, 2022) and *python* (Version 3.8, van Rossum & Drake, 2009). Simulation Study code and neural network models are available at Much and Ranger (2026).

The datasets were generated as follows. First, for each item 1000 samples of the random variables θ, ω, and ɣ were drawn from independent normal distributions with means of m_A, m_B, and m_C as well as standard deviations of s_A, s_B, and s_C. From these, the corresponding effective parameters μ, α, and c_2 were calculated. For each subject-item-combination, potential solution and disengagement times were derived using Equation 4, with the minimum of these times yielding the observed response and RT.

The model parameters for each of the 100 datasets were then estimated with the estimators described above: the maximum likelihood estimators *ML1* and *ML2*, the network-based estimator *NN*, and the Bayesian estimators *BN*, *BW*, and *BH*.

Results

Convergence of Bayesian Estimators

The convergence of the Bayesian estimators is assessed by examining the $\widehat{R}$ values, considering values smaller than 1.1 to be satisfactory (Gelman & Shirley, 2011). To compare the performance, we assess convergence rates: the proportion of parameter-level $\widehat{R}$ values smaller than 1.1. Convergence on the parameter level was worst for estimator *BN*, with an average convergence rate of .64 across all items and parameters, with wide divergence from .42 for Item 2 to .94 for Item 4. Considerably better convergence is found for estimator *BW*, with an overall average of .95, ranging from .92 (Item 1) to 1.00 (Item 4). The best convergence is found for estimator *BH*, with an overall average of .99, with a minimum rate of .96 for parameters of Item 5.

A similar result can be found regarding model-level convergence, where *BN* converges for all parameters in only 52 percent of all 500 item fits (100 datasets with 5 items each), with better rates of .86 and .95 for *BW* and *BH*, respectively.

Parameter Recovery

The results of the parameter recovery can be found in Tables 2 and 3. In Table 2, the means and standard deviations of point estimates across all 100 samples are reported with the true values. The means of posterior distributions are used as point estimates for the Bayesian estimators. Table 3 shows the coverage frequency of 0.95 credible intervals for the three Bayesian estimators.

Bias — Bias depends to a large extent on the kind of parameter and the "item", i.e. the set of parameter values. Bias — as the absolute difference of mean point estimates across all datasets and true values — is very small for all estimators for parameter m_A , but larger for m_C , s_A , and s_B. Similarly, the parameter estimation for Item 4 was on average considerably less biased than the estimation for Item 1.

Comparing estimators across items and parameters, *ML2* has the lowest average absolute bias of 0.08, with a single considerable divergence of 0.46 for $log(s_B)$ in Item 4. The impact of the divergence, however, is exaggerated by the log-scale. Retransformed, s_B is estimated as 0.120 while the true value is 0.076. The worst performance is shown by estimator *BN* with an average absolute bias of 0.37, with severely large divergences of up to 2.39 for $log(s_A)$ in Item 2. Both *NN* and *ML1* have non-satisfactory bias, with the exception of estimates for m_A. Both *BW* and *BH* perform reasonably well, with a slightly better performance of *BH*. Both perform less well on Item 1. Like *ML2*, they have virtually no bias in Items 2 and 4.

Table 2

Model Parameter Recovery: True Values (TV), Means (M), and Standard Deviations (SD) for Maximum Likelihood Estimators (ML1, ML2); the Neural Network Estimator (NN), and Bayesian Estimators (BN, BW, and BH)

Item	Statistic	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
				m_A					m_B					m_C		
TV		2.84	2.72	2.35	2.27	2.26	-1.36	-2.35	-1.29	-2.87	-1.81	-1.68	-2.04	-3.02	-2.02	-2.57
ML1	M	2.82	2.70	2.45	2.21	2.21	-1.27	-2.25	-1.76	-2.77	-1.98	-2.06	-2.22	-2.05	-2.22	-2.41
	SD	(0.28)	(0.18)	(0.08)	(0.36)	(0.12)	(0.50)	(0.42)	(0.30)	(0.38)	(0.30)	(1.56)	(0.64)	(0.50)	(0.95)	(0.55)
ML2	M	2.81	2.73	2.37	2.27	2.24	-1.31	-2.36	-1.40	-2.86	-1.92	-1.77	-2.03	-2.76	-2.02	-2.45
	SD	(0.11)	(0.09)	(0.04)	(0.04)	(0.03)	(0.20)	(0.17)	(0.18)	(0.08)	(0.20)	(0.41)	(0.21)	(0.41)	(0.05)	(0.30)
NN	M	2.83	2.68	2.19	2.32	2.18	-1.29	-2.31	-1.21	-2.72	-2.14	-1.93	-2.41	-2.41	-2.31	-2.32
	SD	(0.08)	(0.08)	(0.06)	(0.05)	(0.06)	(0.16)	(0.15)	(0.18)	(0.10)	(0.13)	(0.23)	(0.19)	(0.16)	(0.09)	(0.12)
BN	M	2.69	2.49	2.36	2.29	2.28	-1.05	-1.70	-1.27	-2.78	-1.63	-2.41	-2.88	-3.48	-2.10	-3.23
	SD	(0.08)	(0.15)	(0.07)	(0.04)	(0.02)	(0.20)	(0.52)	(0.32)	(0.23)	(0.26)	(0.37)	(1.12)	(0.81)	(0.22)	(0.53)
BW	M	2.74	2.71	2.40	2.27	2.25	-1.09	-2.30	-1.55	-2.85	-1.95	-2.19	-2.12	-2.51	-2.04	-2.43
	SD	(0.06)	(0.09)	(0.03)	(0.04)	(0.02)	(0.18)	(0.17)	(0.12)	(0.07)	(0.18)	(0.38)	(0.19)	(0.35)	(0.06)	(0.31)
BH	M	2.75	2.71	2.38	2.27	2.24	-1.18	-2.30	-1.46	-2.86	-1.91	-2.11	-2.10	-2.63	-2.03	-2.50
	SD	(0.06)	(0.08)	(0.02)	(0.04)	(0.03)	(0.11)	(0.17)	(0.13)	(0.07)	(0.17)	(0.34)	(0.21)	(0.32)	(0.05)	(0.28)
				$\log(s_A)$					$\log(s_B)$					$\log(s_C)$		
TV		-1.18	-1.20	-3.19	-0.74	-2.37	-2.53	-0.77	-0.58	-2.58	-1.38	0.15	-0.35	0.69	-0.77	0.38
ML1	M	-1.45	-1.46	-2.51	-0.93	-1.96	-2.05	-0.83	-0.61	-1.82	-1.53	-0.19	-0.51	0.19	-0.93	0.19
	SD	(0.65)	(0.72)	(0.63)	(1.04)	(0.67)	(0.85)	(0.52)	(0.47)	(0.61)	(0.53)	(1.43)	(0.78)	(0.72)	(1.41)	(0.57)
ML2	M	-1.26	-1.22	-2.99	-0.76	-2.14	-2.30	-0.77	-0.57	-2.12	-1.41	0.17	-0.38	0.59	-0.79	0.28
	SD	(0.19)	(0.26)	(0.36)	(0.08)	(0.44)	(0.62)	(0.14)	(0.06)	(0.40)	(0.16)	(0.15)	(0.15)	(0.17)	(0.17)	(0.20)
NN	M	-1.37	-1.20	-3.14	-0.99	-1.76	-2.22	-0.83	-0.36	-2.08	-1.70	0.20	-0.25	0.53	-0.92	0.21
	SD	(0.14)	(0.18)	(0.27)	(0.13)	(0.22)	(0.34)	(0.21)	(0.05)	(0.26)	(0.20)	(0.11)	(0.14)	(0.11)	(0.22)	(0.17)
BN	M	-1.61	-3.59	-3.79	-0.89	-3.20	-1.62	-0.49	-0.47	-2.97	-1.18	0.42	-0.32	0.63	-0.79	0.45
	SD	(0.40)	(0.39)	(0.30)	(0.29)	(0.46)	(0.30)	(0.13)	(0.06)	(1.28)	(0.41)	(0.06)	(0.50)	(0.28)	(0.48)	(0.27)
BW	M	-1.41	-1.28	-2.83	-0.75	-2.08	-1.95	-0.83	-0.57	-2.49	-1.40	0.33	-0.34	0.44	-0.82	0.21
	SD	(0.12)	(0.25)	(0.38)	(0.06)	(0.26)	(0.32)	(0.20)	(0.05)	(0.55)	(0.17)	(0.15)	(0.14)	(0.12)	(0.16)	(0.18)
BH	M	-1.39	-1.34	-2.96	-0.76	-2.19	-2.02	-0.78	-0.56	-2.40	-1.44	0.28	-0.36	0.52	-0.78	0.28
	SD	(0.11)	(0.28)	(0.32)	(0.08)	(0.39)	(0.27)	(0.17)	(0.05)	(0.38)	(0.17)	(0.13)	(0.15)	(0.13)	(0.15)	(0.18)

Table 3

Model Parameter Recovery: Coverage Frequencies (p) for the Bayesian Estimators Based on Neural Network Approximation BN, Weibull Approximation BW, and a Histogram Approach BH

Item	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
			m_A					m_B					m_C		
BN	0.01	0.28	0.39	0.88	0.23	0.01	0.49	0.40	0.32	0.23	0.02	0.29	0.41	0.78	0.24
BW	0.82	0.92	0.78	0.98	0.93	0.83	0.89	0.36	0.95	0.88	0.87	0.88	0.78	0.96	0.90
BH	0.92	0.98	0.96	0.93	0.93	0.91	0.96	0.97	0.90	0.95	0.92	0.97	0.97	0.92	0.93
			$\log(s_A)$					$\log(s_B)$					$\log(s_C)$		
BN	0.01	0.00	0.17	0.51	0.23	0.10	0.10	0.21	0.29	0.22	0.02	0.26	0.53	0.28	0.56
BW	0.71	0.91	0.88	0.97	0.81	0.78	0.92	0.95	0.93	0.97	0.87	0.90	0.55	0.92	0.85
BH	0.94	0.95	0.96	0.96	0.94	0.99	0.91	0.91	0.97	0.96	0.95	0.95	0.95	0.91	0.94

Efficiency — Regarding the variation of estimates, there is a similar parameter dependence. All estimators show the best performance for parameter m_A with mostly small standard deviations across the simulation samples. *ML1* performs poorly in terms of efficiency with standard deviations larger than 0.50 for all parameters except m_A and m_B. The lowest estimation variability has estimator *NN* which is hardly surprising since the simulation error is the only source of variation as the net is deterministic. Estimator *BN*, on the other hand, is also based on a deterministic neural net, but performs rather poorly. The comparison with the other Bayesian estimators shows that this must be attributed to the likelihood approximation. The lowest variability across the simulation samples is shown by estimator *BH*. Its outperformance is consistent across all items and parameters.

Coverage Frequencies — Recovery inference is examined for the Bayesian estimators using 95% credibility intervals. *BN* performs very poor with coverage frequencies less than 0.60 with the exception of some parameters in Item 4. *BW* shows a better performance but mostly falls short of reaching acceptable coverage frequencies. *BH* has acceptable coverage frequencies ranging from 0.90 to 0.99.

Based on the comprehensive results of the parameter recovery simulation study, we recommend to use estimator *BH* for fitting the DBA model. We consider it to be a suitable and efficient estimator, demonstrating satisfactory coverage frequency and minimal bias. Outside of a Bayesian framework, estimator ML_2 is a feasible option for model estimation.

Empirical Example

Amsterdam Chess Test

Data Description and Data Preparation

As a first data set to illustrate the use of the DBA model, we chose the 10 items from the Amsterdam Chess Test (ACT; van der Maas & Wagenmakers, 2005) that measure endgame skill in test form "Choose a move A". The participants were chess players that took part in a tournament where the data were collected. Subjects were required to find the best move in a given chess position within a time limit of 30s. The response format was essentially open. Responses were scored as correct or incorrect and RTs were recorded in seconds with a precision of 1/60 of a second. The dataset is available in the *LNIRT R* package (Fox et al., 2021). Our analysis code, estimation results and diagnostic plots are available at Much and Ranger (2026).

From the original data set with 259 participants, 22 participants with a missing ELO rating were excluded. Responses and RTs to the endgame items were missing from 2 additional subjects. The resulting data set included responses and RTs from $N = 235$ participants.

We used the Bayesian estimator based on the histogram approach (*BH*) to fit the DBA model to the data. Starting values, as well as the true parameter values for the simulation study[1], were determined with an earlier version of Estimator *ML1*. Differing from the simulation study, Estimator *BH* was adapted to the empirical datasets by extending the number of histogram bins to 40.

Results

Convergence was satisfactory for most items, with exception of Item 7 and Item 6 with $\widehat{R} > 1.10$ for parameter s_C in both items. Applying a stricter convergence criterion by Vehtari et al. (2021), for only 2 out of 10 items all parameter $\widehat{R}$ values were below 1.01, with a total of 35 out of 60 parameter estimations yielding $\widehat{R}$ values below 1.01. Traceplot inspections show a substantial autocorrelation for all items, except for Item 2. Across all model parameters, Item 2 has a minimum effective sample size (ESS) of 2561 (Parameter m_A). Other items have a considerably smaller minimal ESS in the range of 408 (m_C in Item 4) to 1470 (s_A in Item 1). All values of $\widehat{R}$ (Table A1) and ESS (Table A2) can be found in the Appendix.

We examined the fit of the model by comparing model implied and empirical response accuracy, RT distributions, and conditional accuracy functions (CAFs), i.e. mean accuracies conditional on RT quantiles with posterior predictive checks.

1) Estimates for ACT Items 2, 4, 5, 10, and 6 were used as Items 1 to 5 in the simulation study.

Response accuracy prediction was very high with a correlation of $r = 0.999$ between empirical response accuracy and mean response accuracy in posterior predictive samples. Bias was < 0.01 for 9 out of 10 items, and 0.037 for Item 10. All empirical response accuracy rates are covered by the corresponding .80 credible interval of the predicted response accuracy rate.

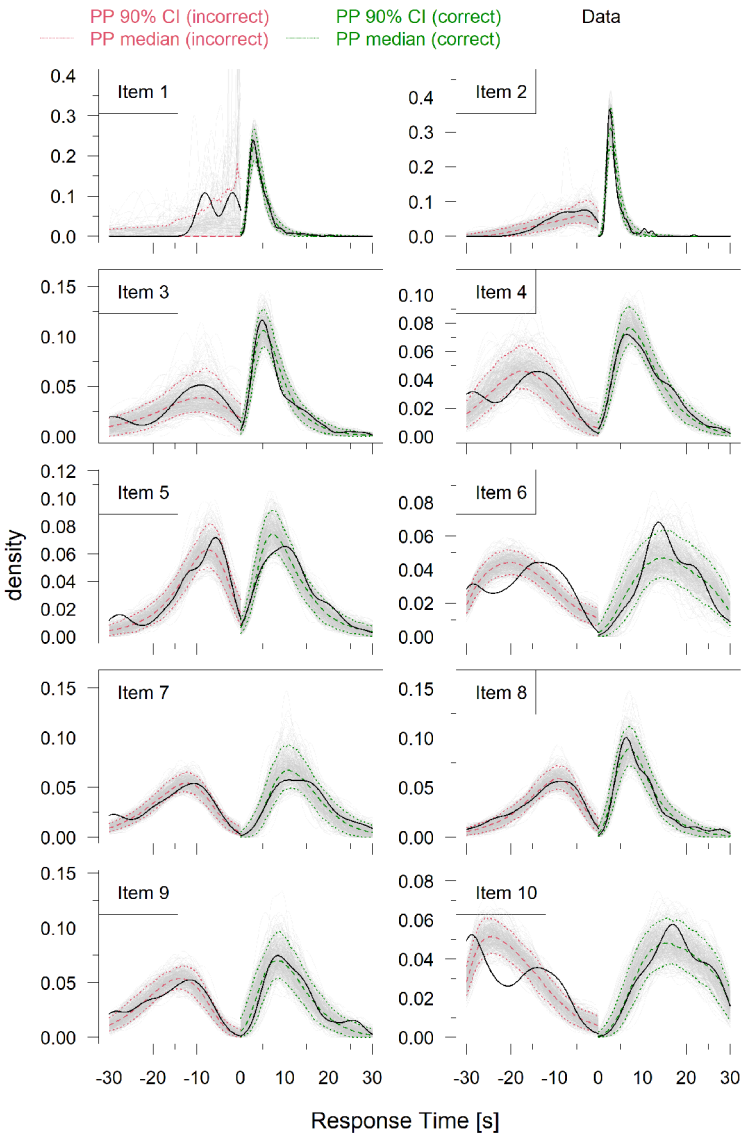
The item distributions of RTs, conditional on response type, are shown in Figure 2 where the empirical densities are compared to the densities from posterior predictive samples. The RT distributions were captured very well for relatively fast responses from a larger number of participants, e.g., correct responses for the easier Items 1 and 2. Mismatches and larger variations in the predictions can be seen when responses were scarce (e.g., only two incorrect responses in Item 1) and when empirical data is bimodal. The bimodality originates in the time-limit of 30 seconds, that led participants to give a response shortly before the time ends, yielding mostly incorrect responses. Hence, bimodal distributions are more dominant for incorrect responses, most notable in Items 10, 6, and 4 where the model predictions deviate significantly from the empirical data. Predicted distributions for correct responses fit the data rather well.

Figure 3 shows the item-wise conditional accuracy functions (CAFs) calculated from the data and the posterior predictive samples. In the empirical data, three items show a negative relationship between RT and accuracy (Items 3, 4 and 9), that means slower responses tend to be less correct, five items show an inverted-U-shaped relationship (Items 2, 5, 6, 8 and 10), and the other two a mostly constant accuracy across RTs (Items 1 and 7). While the variation of model predictions can become quite large, the overall shape and magnitude of the conditional accuracy is well captured for all items. The 0.90 prediction interval does not include the data for only two points: one in Item 6 and one in Item 10.

The posterior medians as point estimates for each parameter and item are shown in Table 4. Regarding m_A, put on the scale of accumulated information by taking the exponential, the largest value corresponds to the easiest item that was solved by 99% of the participants and the smallest value corresponds to the hardest item (Item 6). Deviations of an ordering by solution frequency result from an interplay of effective speed α (represented by Parameters m_B and s_B) and stopping thresholds c_2 (represented by Parameters m_C and s_C). E.g., according to the model Item 4 is predicted to be easily solvable with expected maximum ability values around 18.91, well above the solution threshold $c_1 = 10$. However, a relatively large expected c_2 results in an earlier stop of the solution process and hence more incorrect responses. A correspondence of response times and parameters governing effective speed α (m_B and s_B) is also palpable as the item with the fastest responses (Item 2) yields the highest m_B and the item with the slowest responses (Item 10) yields the lowest values for m_B.

Figure 2

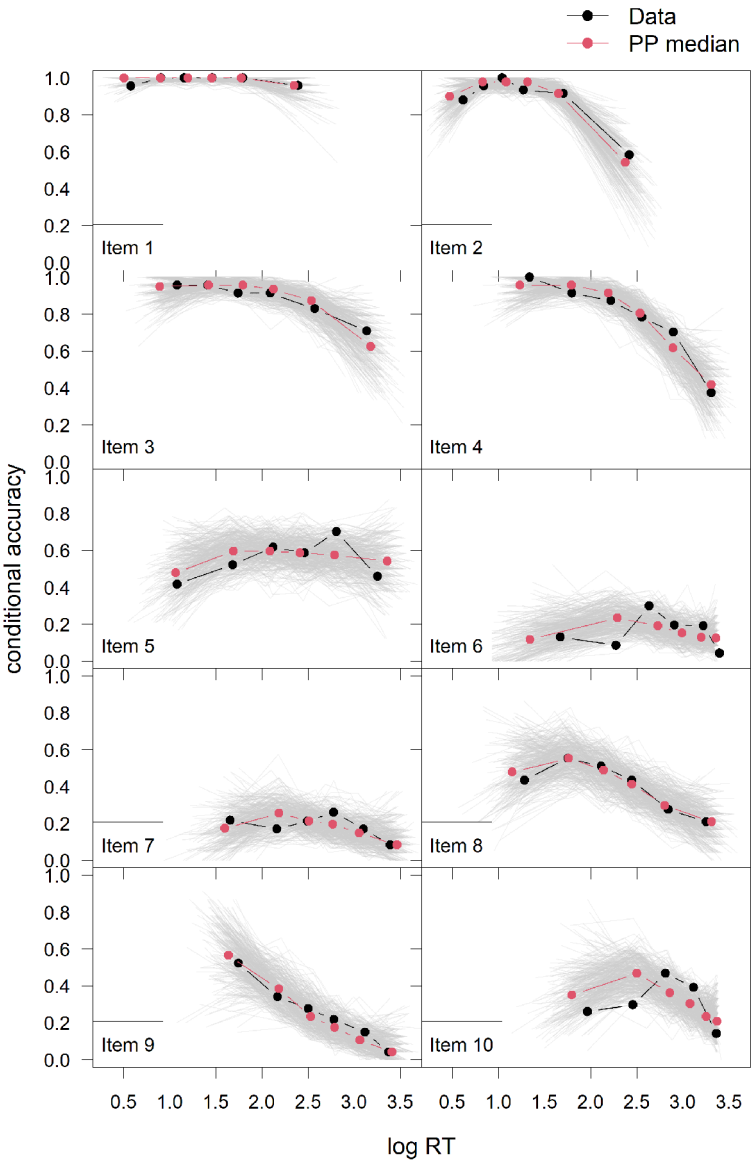
Response Time Distributions Amsterdam Chess Test



Note. Predicted conditional RT densities for correct (positive RTs) and incorrect (flipped RT distributions) responses to the 10 end-move items of the Amsterdam Chess test. Black lines depict the empirical densities. Single posterior predictions are depicted in gray. Dashed lines depict the median of predicted densities, dotted lines depict 90% credible intervals of the densities, green lines refer to correct responses, red lines refer to incorrect responses.

Figure 3

Conditional Accuracy Functions Amsterdam Chess Test



Note. Predicted conditional accuracy functions (CAFs) for the 10 end-move items of the Amsterdam Chess test. Each line plots mean accuracy against mean RT within six RT bins, defined by the 0.1, 0.3, 0.5, 0.7, and 0.9 RT quantiles. Gray lines depict the CAFs from posterior predictive samples, the black line depicts the empirical CAF, and the red line depicts the median of posterior predictive samples.

Table 4

Point Estimates (Posterior Medians), Mean Response Time, and Accuracy for the 10 End-Move Items of the Amsterdam Chess Test

Item	$exp(m_A)$	$exp(m_B)$	$exp(m_C)$	s_A	s_B	s_C	$\bar{t}$	$\bar{x}$
1	29.29	0.11	0.03	0.36	0.22	1.56	4.46	0.99
2	15.26	0.32	0.06	0.25	0.09	1.81	4.16	0.90
3	13.42	0.19	0.04	0.17	0.51	1.44	8.81	0.89
4	18.91	0.07	0.17	0.48	0.27	0.48	12.36	0.79
5	10.35	0.30	0.06	0.02	0.63	1.57	11.28	0.57
6	6.08	0.06	0.11	0.67	0.01	0.01	16.91	0.17
7	9.80	0.17	0.11	0.10	0.43	1.04	15.30	0.19
8	10.26	0.26	0.08	0.10	0.49	1.44	11.49	0.42
9	9.62	0.17	0.10	0.15	0.62	0.53	15.41	0.25
10	9.98	0.05	0.14	0.56	0.01	0.00	19.51	0.30

Discussion

The time-limit of 30 seconds is challenging for the model. It occurred that participants responded just a few seconds before the time limit which resulted in bimodal empirical RT distributions. This suggests that participants checked for other, possibly better solutions before making their final response. The model does not account for this: it is predicted that the correct response is given only once the threshold is hit and that a incorrect response is given only once the progress is too slow. A potential integration of a time-limit in the model, e.g., by censoring the predicted RTs at the time-limit, could explain bimodality in the latency of incorrect responses but not for the correct responses that occur in the data, presumably from informed or random guessing. Despite these limitations, the model fits accuracy and RT data reasonably well and accurately reflects multiple forms of the relationship between RT and accuracy.

Matrix Reasoning With IMak-17

Data Description and Data Preparation

The model was also fit to matrix reasoning data from Myszkowski et al. (2022), who collected responses from $N = 555$ adult participants to Big Five personality items and a progressive matrices test in a low-stakes context. The 17 matrix items were automatically generated with the *IMak R* package (Blum & Holling, 2018). Data were collected in an online study and total screen times were recorded for the IMak items in full seconds. The IMak test was reported to be reliable, its validity was examined using ability estimates from an intelligence test that has been administered to all participants before the study. The dataset, as well as test information and the original analysis are available in the

repository of the original study at Myszkowski and Storme (2022). Our analysis code, estimation results and diagnostic plots are available at Much and Ranger (2026).

Before fitting the data to the DBA model, we screened the data for outliers in the timing data. This was necessary because of the occurrence of unusually long screen times, presumably including longer non-response time in the online study when participants were distracted from working on the test. For each item, participant data were removed when the RT exceeded the range of 2.5 interquartile ranges from the upper or the lower quartile. This reduced the sample size to a number between 519 and 552 responses for each item. The RTs were rescaled with a factor of 1/6 to calibrate them with the RTs achievable by the model within the parameter range examined in the simulation study.

Again, the model parameters were estimated with estimator *BH*, with 40 histogram bins. Starting values were randomly drawn from uniform distributions around estimates from estimator *ML1*. Due to convergence issues and poor estimator efficiency with the setup used in the simulation study and the first empirical illustration, we extended the number of iterations to 40,000. Five thousand (5,000) iterations were discarded as burn-in.

Results

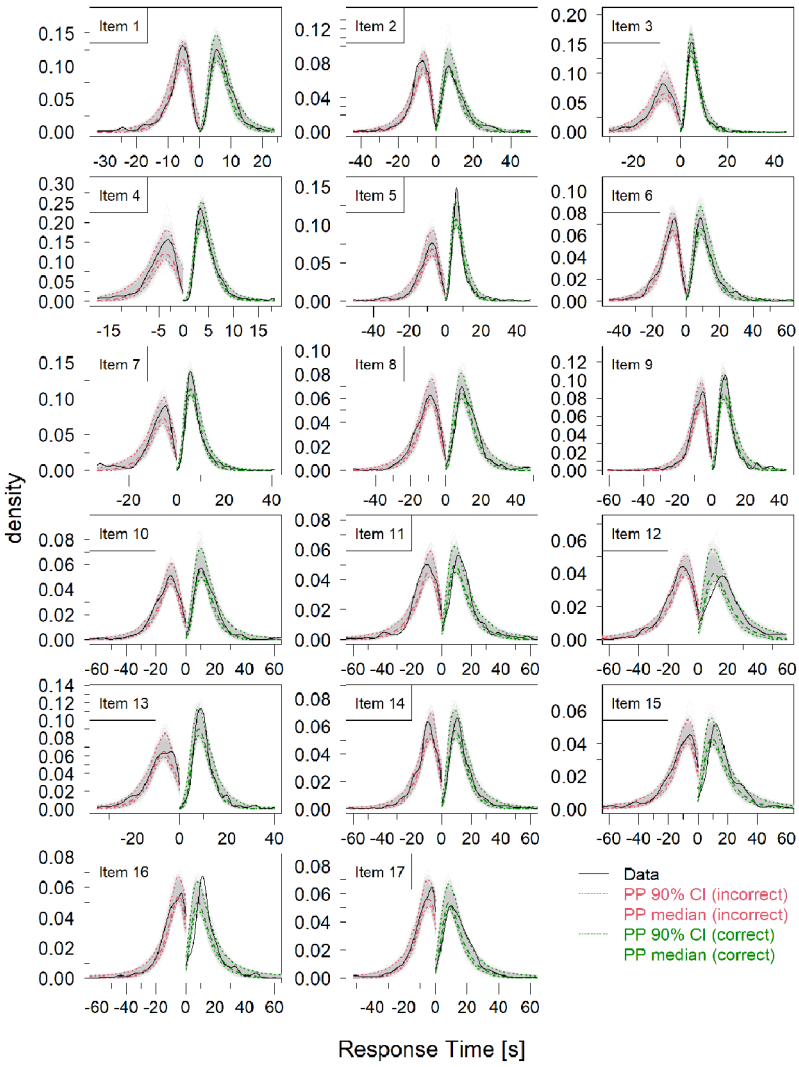
Convergence was satisfactory for all parameters in all items. All $\widehat{R}$ values were smaller than 1.10. For 8 out of 17 items, all parameter $\widehat{R}$ values were below the stricter criterion of 1.01 (Vehtari et al., 2021), with a total of 78 out of 102 parameter estimations yielding $\widehat{R}$ values below 1.01. Traceplot inspections reveal, however, substantial autocorrelations, most notably in Items 1, 13, and 17, for which minimum effective sample sizes were below 1000. Sufficiently low autocorrelations can be found for Items 5, 6, and 15, for which even the minimal effective sample sizes were larger than 1500. All values of $\widehat{R}$ (Table A3) and ESS (Table A4) can be found in the Appendix.

Model fit was again examined with posterior predictive checks. Response accuracy prediction was very high with a virtually perfect correlation of $r = 0.999$ between empirical response accuracy and mean response accuracy in posterior predictive samples. Bias was < 0.006 for all items. All empirical response accuracy rates are covered by the corresponding .80 credible interval of the predicted response accuracy rate.

Item-wise conditional RTs distributions are shown in Figure 4. They are captured quite well by the model predictions. However, there is a larger offset for the correct responses on Item 12, which could only be solved by 121 participants (22.7%), and a larger offset for the correct responses on Item 16, that shows the largest difference in RT means between correct and incorrect responses (101.76*s* vs. 60.51*s*). Regarding the incorrect responses, some empirical distributions (Items 13 and 16) show a concentration of responses at very short RTs. This suggests a significant number of fast errors.

Figure 4

Response Time Distributions IMak-17



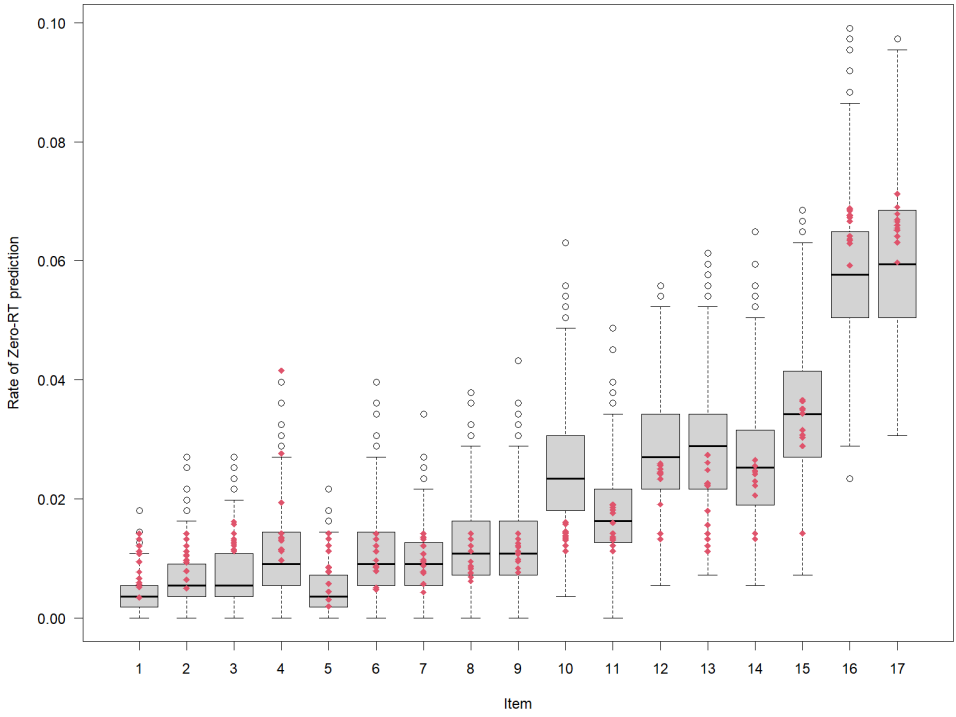
Note. Predicted conditional RT densities for correct (positive RTs) and incorrect (flipped RT distributions) responses to the 17 items of the IMak-17. Black lines depict the empirical densities. Single posterior predictions are depicted in gray. Dashed lines depict the median of predicted densities, dotted lines depict 90% credible intervals of the densities, green lines refer to correct responses, red lines refer to incorrect responses.

One type of fast errors are rapid guesses or non-responses that reflect strong disengagement. In the DBA, if the initial accumulation rate is already below the disengagement

threshold c_2 , the DBA predicts zero-RTs. This is a reflection of maximum disengagement. In Figure 5, the posterior predicted rates of these events are plotted for each item. There is an overall increase across the test with nearly no occurrences predicted for the first items, up to rates of around 0.06 for Items 16 and 17. When comparing this to the threshold-based disengagement indicator Response Time Effort (RTE, Wise & Kong, 2005), we find correlations of 0.834 and 0.977 between the mean predicted zero-RT rates per item and RTE values based on plausible thresholds between 4 and 15 seconds.

Figure 5

Boxplots of Posterior Predicted Zero-Response-Times by Item, IMak-17



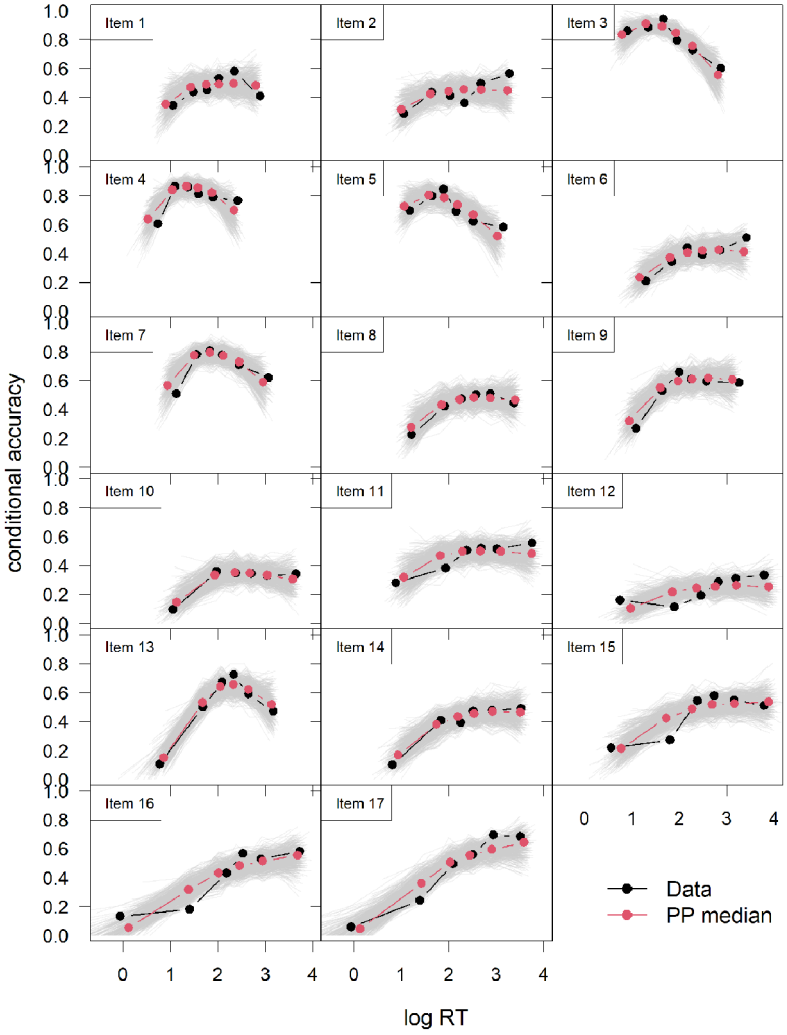
Note. Boxplots of posterior predicted zero-RTs for each IMak item. Red dots indicate RTE estimations based on thresholds between 4 and 15 seconds, scaled to the pooled sample of all predicted zero-RT rates across all items. Higher RTE values correspond to larger thresholds.

Item-wise CAFs are shown in Figure 6. Again, Items 12 and 16 as well as Item 15 stand out as exceptions to a generally good fit of conditional accuracy. For most items, there is a positive correlation of RT and accuracy, with some inverted-U-shaped forms for Items 4, 7, and 13, that are well captured by the DBA model predictions. Misfitting items

feature a step-like form indicating a sudden rise in the solution probability that cannot be accounted for by the model.

Figure 6

Conditional Accuracy Functions IMak-17



Note. Predicted conditional accuracy functions (CAFs) for the 17 items of the IMak-17. Each line plots mean accuracy against mean RT within six RT bins, defined by the 0.1, 0.3, 0.5, 0.7, and 0.9 RT quantiles. Gray lines depict the CAFs from posterior predictive samples, the black line depicts the empirical CAF, and the red line depicts the median of posterior predictive samples.

Posterior medians as point estimates for each parameter and item are shown in Table 5. The parameters governing the effective ability μ, m_A and s_A, show a small variance between items, with m_A estimates ranging from 2.31 to 2.55, translating to expectation values for μ of 10.04 and 12.76, and small estimates for the standard deviation within items, s_A, between 0.002 to 0.15. However, again the largest estimates for m_A correspond to the easiest Items (3, 4, and 5) while for the hardest Item (12), m_A is estimated with the lowest value. The correspondence between parameters governing effective speed α, m_B and s_B, is not as straightforward due to larger variations s_B and the effect of stopping thresholds c_2, reflected by model parameters m_C and s_C.

Table 5

Point Estimates (Posterior Medians), Mean Response Time in Seconds, and Accuracy for the 17 IMak Items

Item	$exp(m_A)$	$exp(m_B)$	$exp(m_C)$	s_A	s_B	s_C	$\bar{t}$	$\bar{x}$
1	10.16	0.53	0.08	0.01	0.51	1.45	45.87	0.47
2	10.09	0.43	0.05	0.00	0.61	1.65	62.62	0.42
3	12.76	0.25	0.12	0.15	0.44	1.25	42.71	0.81
4	11.54	0.45	0.12	0.07	0.40	1.56	29.76	0.80
5	10.71	0.34	0.05	0.05	0.48	1.53	55.33	0.72
6	10.16	0.30	0.07	0.01	0.58	1.50	73.39	0.39
7	10.39	0.43	0.02	0.03	0.47	2.13	51.21	0.72
8	10.31	0.25	0.09	0.01	0.57	1.37	75.95	0.44
9	10.28	0.36	0.07	0.01	0.55	1.68	59.08	0.56
10	10.26	0.20	0.10	0.03	0.56	1.42	88.97	0.32
11	10.12	0.29	0.04	0.00	0.71	1.94	90.75	0.47
12	10.04	0.26	0.05	0.00	0.73	1.95	98.74	0.23
13	10.45	0.30	0.07	0.03	0.43	1.97	60.45	0.55
14	10.24	0.26	0.08	0.01	0.62	1.66	77.74	0.41
15	10.08	0.29	0.03	0.00	0.75	2.47	93.93	0.46
16	10.07	0.35	0.04	0.00	0.75	2.68	77.30	0.41
17	10.11	0.34	0.04	0.00	0.69	2.73	72.29	0.47

Discussion

Myszkowski et al. (2022) analyzed the data with the dynamic speed joint hierarchical model of van der Linden and Fox (2016) and could achieve an acceptable model fit. An analysis of the empirical RT distributions, however, indicate a violation of the model's log-normality assumption. In this situation, it seems reasonable to fit the data to a more flexible model in terms of RT distributions.

While the DBA is flexible to accommodate for shifts in RT distributions between correct and incorrect responses, it has its limitations with larger time differences. Predictions are drawn towards the RT distribution of the more frequent response. Regarding

smaller offsets for incorrect responses, the variation of model predictions covers concentrations of fast responses very well, even though individual predictions fail to fully capture the resulting shape of the distribution. The DBA predicts zero-RTs when a test-taker disengages before the information accumulation starts. An extension of our baseline model that includes a transformation of these zero-RTs to realistic non-response times could even improve the prediction of outstanding fast errors. However, even without this extension, the rate of predicted zero-RTs can be interpreted as a model-based indicator of disengagement. This indicator correlates highly to the threshold-based RTE and was capable to detect the common finding of decreasing test-taking engagement over the course of a test.

Regarding the estimated parameter values, the results indicate a surprisingly small variance for the parameters that relate to a person's maximum ability. This concerns both the between-trial variance s_A and the variance of m_A between the items. It is noticeable that the estimates for m_A relate to effective parameter values of μ that are close to the fixed solution threshold $c_1 = 10$. This suggests that the model more efficiently represents the variance in the data with variances of processing speed and persistence. Assuming the model is valid, this indicates that given enough time and persistence, virtually everyone would be able to solve the items.

General Discussion

In this paper, we proposed a new process model for responses and RTs on test items, based on an information accumulation mechanism that reflects a decline of solution progress with a differentially decreasing accumulation rate that governs a test-taker's persistence. The proposed DBA model addresses some limitations of previous models that tackle the problem of disengaged test-taking behavior and motivational confounders of ability measures.

First of all, the DBA model reflects disengagement with the continuous effective parameter c_2 which allows for a more nuanced quantification of disengagement than the either/or perspective of mixture models. As a process model, the DBA offers an interpretation of the test-taking behavior that is based on underlying psychological processes. In contrast, models in an IRT framework are agnostic towards the process of generating item responses. They are typically derived based on beneficial statistical and measurement properties and bear no assumptions about the underlying psychological processes. While this is advantageous for flexible application, it is desirable for a model's content validity to explain how the described outcomes emerge (van der Maas et al., 2011).

In contrast to race models, the DBA is a single-process model as it assumes only one underlying process that generates the response. In our model, we offer an explanation of terminating the solution process, as the point of disengaging with an item is dependent

on the solution progress. The decreasing accumulation rate towards zero also lets go of the problematic assumption that a test-taker can solve every item given infinite time and persistence.

The assumption of a constant accumulation rate is also present in single-process models, e.g., the psychometric extensions of the diffusion model (Kang et al., 2022; van der Maas et al., 2011). While this is a convincing assumption in the context of constant stimuli that require quick decisions, we would argue that the non-linear accumulation is more plausible in the context of complex cognitive items, as test-takers will reach a point during the item solution where they are not able to gather more information. Another feature that eases wider application of the DBA model is that it offers a flexible interpretation of the item solution process. It can be applied to a large variety of item formats and is not bound to binary decisions. The model also provides a framework for modeling a test-taker's confidence and the application to rating scale items by taking the accumulator values at the time of response into account.

While the model can overcome limitations of other models, some theoretical issues remain. Most notably, it relies on a specific functional form for the accumulation process: a strictly monotonic increase with exponentially decreasing increments. This implies that test-takers consistently move towards a correct solution, with the most significant progress occurring early in the process. While useful, this provides a simplified view of test-taking behavior. Starting from other process models of item responses and RTs, it adds the feature of information depletion, represented by a relatively simple exponential function. Additionally, the DBA's single-process assumption does not capture scenarios where multiple solution strategies or sequential processes are involved. This limitation that is shared with previous models could not be overcome. The interpretation of modeling achievement item response processes is therefore limited to a simplifying, rather metaphorical view and calls for a pragmatic and cautious application.

Another limitation in the current state of the model is that it is only developed to analyze individual items and hence does not allow — as in IRT models — for a separation of person and item parameters. Future developments that ought to go hand-in-hand with improved and fine-tuned estimation algorithms might overcome this. A straightforward approach to achieve this, such as expressing effective parameters as functions of separate person and item parameters, as in the psychometric diffusion models (Tuerlinckx & De Boeck, 2005; van der Maas et al., 2011), would be computationally infeasible using standard methods such as marginal maximum likelihood estimation. Promising candidates for likelihood-free estimation algorithms involve recent developments in simulation-based inference enhanced with machine learning techniques (e.g., Cranmer et al., 2020; Lueckmann et al., 2021) and were, among others, successfully applied to other evidence accumulation models such as the leaky competing accumulator model (Radev et al., 2020) or the diffusion model (von Krause et al., 2022), though not yet to a psychometric extension like the DBA. Therefore, the suitability for the DBA needs

to be investigated and the chosen method needs calibration for our proposed model. Another promising technique, suggested by an anonymous reviewer, is variational inference (e.g., Blei et al., 2017) with automatic differentiation (Kucukelbir et al., 2017) that might perform well with certain distributional assumptions about the person parameters (Kingma & Welling, 2022). However, we believe that already the analysis of single-items can generate insights to test-taking behavior and that the model offers ideas to improve theory-based modeling of the response process.

Future developments of the DBA model should also explore the feasibility of extending the model to secondary guessing processes, similar to the sequential process interpretation of Birnbaum's (1968) 3PL model (Hutchinson, 1991; San Martín et al., 2006). In the current form, disengagement always results in incorrect responses, irrespective of partial knowledge or options for random guessing. This is rather implausible as test takers most of the times will try to maximize their score giving correct responses on chance level, or after solution progress was already made, with a probability above chance. This would also capture the frequent finding that test takers perform above chance level when they first omit an item but are given a second chance to respond (Hutchinson, 1991).

It is worth noting that in contrast to similar models of evidence accumulation, there is no parameter for non-decision time in the model. In other models, this reflects the parts of the response process where no accumulation is assumed to take place, i.e. stimulus encoding and response output processes (Ratcliff et al., 2016). Hence, the DBA in its current form predicts RTs too close to zero to be realistic. However, a feature of the DBA predictions is the occurence of zero-RTs for incorrect responses. This occurs when the disengagement threshold c_2 exceeds the inital accumulation rate. For the sake of computational simplicity and constrained model complexity, we did not include a non-decision time or a mechanism for zero-RTs and aimed at examining the model in a more pure form without additional parameters. We believe that future research on non-decision time inclusion and on different options for exploiting the occurrence of zero-RTs can be fruitful. In its current form however, the prediction of zero-RTs offers a way of quantifying predicted disengagement. The application to the IMak dataset could replicate the common finding of increasing disengagement over the course of the test (e.g., Nagy et al., 2022; Penk & Richter, 2017; Sachse et al., 2023; Weirich et al., 2017; Wu et al., 2019) and correlated highly with the established threshold-based disengagement indicator RTE.

Despite the limitations and the raw state of the model, the simulation study demonstrated the identifiability and estimability of the model, both with maximum likelihood and Bayesian techniques. We identified computationally feasible estimators with good parameter recovery. Further fine-tuning as well as exploration of likelihood approximation techniques could even improve estimation.

In two empirical applications, the model was successfully fit to capture item accuracy, conditional RT distributions and CAFs. In particular, the model was able to capture different forms of conditional accuracy: positive, negative as well as inverted-U-shaped dependencies of accuracy and RT. To our knowledge, the flexibility of the model's CAFs is unique for single-process models. The diffusion model extension with randomly varying drift rates and starting points by Kang et al. (2022) generates non-linear CAFs that can be non-monotonic, but with an asymptote at the probability of a random guess as it is characteristic of diffusion model approaches (van der Maas et al., 2011). This produces boundedness in the CAF, as the function can only be U-shaped for accuracy rates below chance level and inverted-U-shaped for accuracy rates above chance level, which is often too restrictive. Latent variable models of non-linear conditional dependence rely on modeling residuals or classes (Bolsinova & Molenaar, 2018; Naumann & Goldhammer, 2017) and are regularly embedded in a dual-process interpretation (Chen et al., 2018; De Boeck & Jeon, 2019). The DBA offers a single-process interpretation that can explain non-linear dependencies between responses and RT that are not bounded by chance levels.

Funding: The study was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) within the project "Test-taking engagement and test-taking behavior: Modeling the processes underlying item nonresponse and guessing" (DFG Project Number 28872689, Grants RA 3453/1-2 and PO 1655/3-2).

Acknowledgments: The authors have no additional (i.e., non-financial) support to report.

Competing Interests: The authors have declared that no competing interests exist.

Related Versions: An earlier version of this article appears in the first author's dissertation, *Developing psychological process models of test-taking engagement*, accessible at http://dx.doi.org/10.17169/refubium-51420.

Data Availability: Dataset, test information, and original analysis are available at Myszkowski and Storme (2022). The dataset for the Amsterdam Chess Test is available at Fox et al. (2021). The analysis code, estimation results, and diagnostic plots are available at Much and Ranger (2026).

Supplementary Materials

Type of supplementary material	Availability/Access
Data	
Study data	Myszkowski and Storme (2022)
Amsterdam Chess Test dataset	Fox et al. (2021)
Code	
Code for Differential Ballistic Accumulator Model	Much and Ranger (2026)
Material	
IMak-17 Test	Myszkowski and Storme (2022)
Study/Analysis preregistration	
Study was not preregistered	—
Other	
Supplementary tables	Myszkowski and Storme (2022)
Original study analysis	Myszkowski and Storme (2022)
Estimation results	Much and Ranger (2026)
Diagnostic plots	Much and Ranger (2026)

References

Allaire, J. J., & Chollet, F. (2022). *Keras: R interface to 'Keras'*. R Project for Statistical Computing. https://CRAN.R-project.org/package=keras

Allaire, J. J., & Tang, Y. (2022). *Tensorflow: R interface to 'TensorFlow'*. R Project for Statistical Computing. https://CRAN.R-project.org/package=tensorflow

Bechger, T. M., Maris, G., Verstralen, H., & Verhelst, N. D. (2005). The Nedelsky model for multiple-choice items. In L. A. van der Ark, M. A. Croon & K. Sijtsma (Eds.), *New developments in categorical data analysis for the social and behavioral sciences* (pp. 187-206). Psychology Press.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Addison-Wesley.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association, 112*(518), 859–877. https://doi.org/10.1080/01621459.2017.1285773

Blum, D., & Holling, H. (2018). Automatic generation of figural analogies with the *IMak* package. *Frontiers in Psychology, 9*, Article 1286. https://doi.org/10.3389/fpsyg.2018.01286

Bolsinova, M., & Molenaar, D. (2018). Modeling nonlinear conditional dependence between response time and accuracy. *Frontiers in Psychology, 9*, Article 1525. https://doi.org/10.3389/fpsyg.2018.01525

Brown, S. D., & Heathcote, A. (2005). A ballistic model of choice response time. *Psychological Review, 112*(1), 117–128. https://doi.org/10.1037/0033-295X.112.1.117

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology, 57*(3), 153–178. https://doi.org/10.1016/j.cogpsych.2007.12.002

Butler, J., & Adams, R. J. (2007). The impact of differential investment of student effort on the outcomes of international studies. *Journal of Applied Measurement, 8*(3), 279–304.

Chen, H., De Boeck, P., Grady, M., Yang, C.-L., & Waldschmidt, D. (2018). Curvilinear dependency of response accuracy on response time in cognitive tests. *Intelligence, 69*, 16–23. https://doi.org/10.1016/j.intell.2018.04.001

Cranmer, K., Brehmer, J., & Louppe, G. (2020). The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences of the United States of America, 117*(48), 30055–30062. https://doi.org/10.1073/pnas.1912789117

De Boeck, P., & Jeon, M. (2019). An overview of models for response times and processes in cognitive tests. *Frontiers in Psychology, 10*, Article 102. https://doi.org/10.3389/fpsyg.2019.00102

Domingue, B. W., Kanopka, K., Stenhaug, B., Sulik, M. J., Beverly, T., Brinkhuis, M., Circi, R., Faul, J., Liao, D., McCandliss, B., Obradovic, J., Piech, C., Porter, T., Project iLEAD Consortium, Soland, J., Weeks, J., Wise, S. L., & Yeatman, J. (2022). Speed-accuracy trade-off? Not so fast: Marginal changes in speed have inconsistent relationships with accuracy in real-world settings. *Journal of Educational and Behavioral Statistics, 47*(5), 576–602. https://doi.org/10.3102/10769986221099906

Eklof, H. (2010). Skill and will: Test-taking motivation and assessment quality. *Assessment in Education: Principles, Policy & Practice, 17*(4), 345–356. https://doi.org/10.1080/0969594X.2010.516569

Fengler, A., Govindarajan, L. N., Chen, T., & Frank, M. J. (2021). Likelihood approximation networks (LANs) for fast inference of simulation models in cognitive neuroscience. *eLife, 10*, Article e65074. https://doi.org/10.7554/eLife.65074

Fox, J.-P., Klotzke, K., & Klein Entink, R. (2021). *LNIRT: Lognormal response time item response theory models*. R Project for Statistical Computing. https://CRAN.R-project.org/package=LNIRT

Gelman, A., & Shirley, K. (2011). Inference from simulations and monitoring convergence. In S. Brooks, A. Gelman, G. Jones & X.-L. Meng (Eds.), *Handbook of Markov chain Monte Carlo* (pp. 163-174). Chapman Hall/CRC.

Guo, X., Luo, Z., & Yu, X. (2019). A speed-accuracy tradeoff hierarchical model based on cognitive experiment. *Frontiers in Psychology, 10*, Article 2910. https://doi.org/10.3389/fpsyg.2019.02910

Hong, M. R., & Cheng, Y. (2019). Robust maximum marginal likelihood (RMML) estimation for item response theory models. *Behavior Research Methods, 51*(2), 573–588. https://doi.org/10.3758/s13428-018-1150-4

Hong, M. R., Steedle, J. T., & Cheng, Y. (2020). Methods of detecting insufficient effort responding: Comparisons and practical recommendations. *Educational and Psychological Measurement, 80*(2), 312–345. https://doi.org/10.1177/0013164419865316

Hutchinson, T. P. (1991). *Ability, partial information, guessing: Statistical modelling applied to multiple-choice tests*. Rumsby Scientific.

Jones, A. T., Nguyen, H. D., & McLachlan, G. J. (2018). logKDE: Log-transformed kernel density estimation. *Journal of Open Source Software, 3*(28), Article 870. https://doi.org/10.21105/joss.00870

Kang, I., De Boeck, P., & Ratcliff, R. (2022). Modeling conditional dependence of response accuracy and response time with the diffusion item response theory model. *Psychometrika, 87*(2), 725–748. https://doi.org/10.1007/s11336-021-09819-5

Kingma, D. P., & Welling, M. (2022). *Auto-encoding variational bayes*. arXiv. https://doi.org/10.48550/arXiv.1312.6114

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., & Blei, D. M. (2017). Automatic differentiation variational inference. *Journal of Machine Learning Research, 18*(14), 1–45. http://jmlr.org/papers/v18/16-107.html

Lee, Y.-H., & Ying, Z. (2015). A mixture cure-rate model for responses and response times in time-limit tests. *Psychometrika, 80*(3), 748–775. https://doi.org/10.1007/s11336-014-9419-8

Lenzi, A., Bessac, J., Rudi, J., & Stein, M. L. (2023). Neural networks for parameter estimation in intractable models. *Computational Statistics & Data Analysis, 185*, Article 107762. https://doi.org/10.1016/j.csda.2023.107762

Lu, J., & Wang, C. (2020). A response time process model for not-reached and omitted items. *Journal of Educational Measurement, 57*(4), 584–620. https://doi.org/10.1111/jedm.12270

Lu, J., Wang, C., Zhang, J., & Tao, J. (2020). A mixture model for responses and response times with a higher-order ability structure to detect rapid guessing behaviour. *British Journal of Mathematical and Statistical Psychology, 73*(2), 261–288. https://doi.org/10.1111/bmsp.12175

Lueckmann, J.-M., Boelts, J., Greenberg, D. S., Goncalves, P. J., & Macke, J. H. (2021). *Benchmarking simulation-based inference*. arXiv. https://doi.org/10.48550/arXiv.2101.04653

Meyer, J. P. (2010). A mixture rasch model with item response time components. *Applied Psychological Measurement, 34*(7), 521–538. https://doi.org/10.1177/0146621609355451

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*(2), 205–228. https://doi.org/10.1111/bmsp.12117

Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. J. (2015). Fitting diffusion item response theory models for responses and response times using the R package *diffIRT*. *Journal of Statistical Software, 66*(4), 1–34. https://doi.org/10.18637/jss.v066.i04

Much, S., & Ranger, J. (2026). *Code for Differential Ballistic Accumulator Model* [Analysis code to replicate the study, estimation results, and diagnostic plots]. PsychOpen GOLD. https://doi.org/10.23668/psycharchives.21579

Myszkowski, N., & Storme, M. (2022). *Exploring the associations between personality and response speed trajectories in low-stakes intelligence tests* [OSF project page containing dataset, test information, and the original analysis]. Open Science Framework. https://doi.org/10.17605/OSF.IO/UGE2W

Myszkowski, N., Storme, M., Kubiak, E., & Baron, S. (2022). Exploring the associations between personality and response speed trajectories in low-stakes intelligence tests. *Personality and Individual Differences, 191*, Article 111580. https://doi.org/10.1016/j.paid.2022.111580

Nagy, G., Ulitzsch, E., & Lindner, M. A. (2022). The role of rapid guessing and test-taking persistence in modelling test-taking engagement. *Journal of Computer Assisted Learning, 39*(3), 751–766. https://doi.org/10.1111/jcal.12719

Naumann, J., & Goldhammer, F. (2017). Time-on-task effects in digital reading are non-linear and moderated by persons' skills and tasks' demands. *Learning and Individual Differences, 53*, 1–16. https://doi.org/10.1016/j.lindif.2016.10.002

Penk, C., & Richter, D. (2017). Change in test-taking motivation and its relationship to test performance in low-stakes assessments. *Educational Assessment, Evaluation and Accountability, 29*(1), 55–79. https://doi.org/10.1007/s11092-016-9248-7

R Core Team. (2022). *R: A language and environment for statistical computing*. R Project for Statistical Computing. https://www.R-project.org/

Radev, S. T., Mertens, U. K., Voss, A., & Kothe, U. (2020). Towards end-to-end likelihood-free inference with convolutional neural networks. *British Journal of Mathematical and Statistical Psychology, 73*(1), 23–43. https://doi.org/10.1111/bmsp.12159

Ranger, J., & Kuhn, J.-T. (2014). An accumulator model for responses and response times in tests based on the proportional hazards model. *British Journal of Mathematical and Statistical Psychology, 67*(3), 388–407. https://doi.org/10.1111/bmsp.12025

Ranger, J., Wolgast, A., & Kuhn, J.-T. (2019). Robust estimation of the hierarchical model for responses and response times. *British Journal of Mathematical and Statistical Psychology, 72*(1), 83–107. https://doi.org/10.1111/bmsp.12143

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences, 20*(4), 260–281. https://doi.org/10.1016/j.tics.2016.01.007

Sachse, K. A., Weirich, S., Mahler, N., & Rjosk, C. (2023). Explaining performance decline over the course of taking comprehensive proficiency tests: The roles of effort and omission propensity. *International Journal of Testing, 24*(1), 1–23. https://doi.org/10.1080/15305058.2023.2250889

San Martín, E., del Pino, G., & De Boeck, P. (2006). IRT models for ability-based guessing. *Applied Psychological Measurement, 30*(3), 183–203. https://doi.org/10.1177/0146621605282773

Silm, G., Pedaste, M., & Taht, K. (2020). The relationship between performance and test-taking effort when measured with self-report or time-based instruments: A meta-analytic review. *Educational Research Review, 31*, Article 100335. https://doi.org/10.1016/j.edurev.2020.100335

Spearman, C. (1927). *The abilities of man*. Macmillan.

ter Braak, C. J. F. (2006). A Markov chain Monte Carlo version of the genetic algorithm differential evolution: Easy Bayesian computing for real parameter spaces. *Statistics and Computing, 16*(3), 239–249. https://doi.org/10.1007/s11222-006-8769-1

Thurstone, L. L. (1937). Ability, motivation, and speed. *Psychometrika, 2*(4), 249–254. https://doi.org/10.1007/BF02287896

Tuerlinckx, F., & De Boeck, P. (2005). Two interpretations of the discrimination parameter. *Psychometrika, 70*(4), 629–650. https://doi.org/10.1007/s11336-000-0810-3

Tuerlinckx, F., Molenaar, D., & van der Maas, H. L. J. (2016). Diffusion-based response-time models. In W. J. van der Linden (Ed.), *Handbook of item response theory, Volume 1: Models* (pp. 283-300). CRC Press/Taylor & Francis.

Turner, B. M., Sederberg, P. B., Brown, S. D., & Steyvers, M. (2013). A method for efficiently sampling from distributions with correlated dimensions. *Psychological Methods, 18*(3), 368–384. https://doi.org/10.1037/a0032222

Uhlenbeck, G. E., & Ornstein, L. S. (1930). On the theory of the Brownian motion. *Physical Review, 36*(5), 823–841. https://doi.org/10.1103/PhysRev.36.823

Ulitzsch, E., von Davier, M., & Pohl, S. (2020a). A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level non-response. *British Journal of Mathematical and Statistical Psychology, 73*(Suppl. 1), 83–112. https://doi.org/10.1111/bmsp.12188

Ulitzsch, E., von Davier, M., & Pohl, S. (2020b). Using response times for joint modeling of response and omission behavior. *Multivariate Behavioral Research, 55*(3), 425–453. https://doi.org/10.1080/00273171.2019.1643699

van der Linden, W. J., & Fox, G. J. (2016). Joint hierarchical modeling of responses and response times. In W. J. van der Linden (Ed.), *Handbook of item response theory, Volume 1: Models* (pp. 481-500). CRC Press/Taylor & Francis. https://doi.org/10.1201/9781315374512-30

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*(2), 339–356. https://doi.org/10.1037/a0022749

van der Maas, H. L. J., & Wagenmakers, E.-J. (2005). A psychometric analysis of chess expertise. *American Journal of Psychology, 118*(1), 29–60.

van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual*. CreateSpace.

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two-choice response times. *Psychological Methods, 16*(1), 44–62. https://doi.org/10.1037/a0021765

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Burkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved R for assessing convergence of MCMC (with discussion). *Bayesian Analysis, 16*(2), 667–718. https://doi.org/10.1214/20-BA1221

von Krause, M., Radev, S. T., & Voss, A. (2022). Mental speed is high until age 60 as revealed by analysis of over a million participants. *Nature Human Behaviour, 6*(5), 700–708. https://doi.org/10.1038/s41562-021-01282-7

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology, 68*(3), 456–477. https://doi.org/10.1111/bmsp.12054

Webb, E. (1915). *Character and intelligence: Volume 1*. University Press.

Weirich, S., Hecht, M., Penk, C., Roppelt, A., & Bohme, K. (2017). Item position effects are moderated by changes in test-taking effort. *Applied Psychological Measurement, 41*(2), 115–129. https://doi.org/10.1177/0146621616676791

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*(1), 1–17. https://doi.org/10.1207/s15326977ea1001_1

Wise, S. L., & Gao, L. (2017). A general approach to measuring test-taking effort on computer-based tests. *Applied Measurement in Education, 30*(4), 343–354. https://doi.org/10.1080/08957347.2017.1353992

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(2), 163–183. https://doi.org/10.1207/s15324818ame1802_2

Wu, Q., Debeer, D., Buchholz, J., Hartig, J., & Janssen, R. (2019). Predictors of individual performance changes related to item positions in PISA assessments. *Large-Scale Assessments in Education, 7*, Article 5. https://doi.org/10.1186/s40536-019-0073-6

Appendix

Table A1

$\widehat{R}$ Values for the 6 Model Parameters for the 10 End-Move Items of the Amsterdam Chess Test

Item	m_A	m_B	m_C	s_A	s_B	s_C
1	1.01	1.01	1.00	1.01	1.00	1.00
2	1.00	1.00	1.00	1.00	1.00	1.00
3	1.01	1.00	1.00	1.01	1.01	1.00
4	1.03	1.04	1.06	1.03	1.02	1.01
5	1.01	1.00	1.01	1.01	1.00	1.01
6	1.01	1.01	1.01	1.00	1.01	1.13
7	1.01	1.05	1.10	1.02	1.00	1.11
8	1.00	1.00	1.00	1.00	1.00	1.00
9	1.02	1.02	1.03	1.02	1.01	1.01
10	1.02	1.02	1.02	1.02	1.01	1.02

Note. Estimation was done with a DE-MCMC algorithm with 9 chains and 20,000 iterations. 5,000 burn-in iterations were discarded.

Table A2

Effective Sample Sizes for the 6 Model Parameters for the 10 End-Move Items of the Amsterdam Chess Test

Item	m_A	m_B	m_C	s_A	s_B	s_C
1	2,134	2,089	2,403	1,470	1,789	2,513

Item	m_A	m_B	m_C	s_A	s_B	s_C
2	2,561	2,637	2,628	2,870	2,662	2,675
3	897	869	1,192	917	988	1,400
4	495	431	408	518	699	740
5	741	685	647	1,149	1,042	646
6	1,175	1,171	1,138	1,137	2,559	2,921
7	921	751	709	771	1,189	767
8	1,545	970	840	1,120	1,611	984
9	816	515	441	517	518	456
10	518	528	547	496	1,308	1,673

Note. Estimation was done with a DE-MCMC algorithm with 9 chains and 20,000 iterations. 5,000 burn-in iterations were discarded.

Table A3

$\widehat{R}$ Values for the 6 Model Parameters for the 17 Items of the IMak-17

Item	m_A	m_B	m_C	s_A	s_B	s_C
1	1.02	1.01	1.02	1.01	1.01	1.01
2	1.02	1.01	1.01	1.01	1.00	1.01
3	1.01	1.01	1.02	1.01	1.00	1.01
4	1.01	1.01	1.01	1.01	1.01	1.01
5	1.00	1.00	1.00	1.00	1.00	1.00
6	1.00	1.01	1.01	1.01	1.00	1.01
7	1.01	1.01	1.01	1.01	1.00	1.00
8	1.02	1.01	1.01	1.01	1.00	1.01
9	1.01	1.01	1.01	1.00	1.00	1.01
10	1.01	1.01	1.01	1.01	1.00	1.01
11	1.00	1.00	1.00	1.00	1.00	1.00
12	1.01	1.01	1.01	1.00	1.00	1.01
13	1.01	1.00	1.00	1.00	1.00	1.00
14	1.00	1.01	1.01	1.00	1.00	1.01
15	1.00	1.00	1.00	1.00	1.00	1.00
16	1.01	1.00	1.00	1.01	1.00	1.00
17	1.03	1.03	1.02	1.03	1.02	1.02

Note. Estimation was done with a DE-MCMC algorithm with 9 chains and 40,000 iterations. 5,000 burn-in iterations were discarded.

Table A4

Effective Sample Sizes for the 6 Model Parameters for the 17 Items of the IMak-17

Item	m_A	m_B	m_C	s_A	s_B	s_C
1	770	850	811	1,311	2,961	908
2	1,167	1,216	1,211	1,599	3,919	1,345
3	1,633	1,604	1,501	1,733	2,246	1,668
4	1,231	1,269	1,241	1,309	2,484	1,375
5	1,537	1,555	1,582	1,850	3,920	1,714
6	1,474	1,481	1,343	2,117	4,384	1,496
7	1,450	1,738	1,874	1,658	4,864	2,233
8	1,580	1,626	1,518	2,478	4,944	1,629
9	1,286	1,217	1,190	1,951	4,196	1,343
10	1,801	1,712	1,365	2,134	3,840	1,477
11	1,470	1,501	1,431	1,902	4,159	1,606
12	2,106	1,705	1,546	1,516	3,314	1,608
13	918	1,076	1,155	1,141	3,115	1,186
14	1,507	1,638	1,547	2,432	4,518	1,556

Item	m_A	m_B	m_C	s_A	s_B	s_C
15	2,164	2,523	2,424	2,885	4,395	2,625
16	1,529	1,888	1,883	2,555	3,126	1,811
17	406	452	402	486	513	388

Note. Estimation was done with a DE-MCMC algorithm with 9 chains and 40,000 iterations. 5,000 burn-in iterations were discarded.