Psychometrika (2026), **91**, 1232–1254
doi:10.1017/psy.2026.10112

PSYCHOMETRIC SOCIETY

THEORY AND METHODS

Extending the Diffusion Model by a Process Accounting for the Persistence of Test Takers

Jochen Ranger[1], Sören Much[1,2], Niklas Neek[3], Anett Wolgast[4], Augustin Mutak[5,6], José Luis Gaviria Soto[7] and Steffi Pohl[5]

[1]Research Methods, Martin-Luther-Universität Halle-Wittenberg, Germany; [2]Wilhelm Wundt Institute of Psychology, Leipzig University; [3]Applied Sport Psychology, Martin-Luther-Universität Halle-Wittenberg, Germany; [4]University of Applied Sciences FHM Hanover, Germany; [5]Methods and Evaluation, Freie Universität Berlin, Germany; [6]Faculty of Humanities and Social Sciences, University of Zagreb, Croatia; [7]Departamento de Investigación y Psicología en Educación, Universidad Complutense de Madrid, Spain

Corresponding author: Jochen Ranger; Email: jochen.ranger@psych.uni-halle.de

(Received 9 December 2024; revised 27 March 2026; accepted 7 April 2026)

Abstract

We propose a latent trait model for the responses and response times on tests that separates capability from persistence. Core of the model is a race between a diffusion process and a censoring process. The diffusion process represents item-level cognitive processing and determines the processing time of a test taker. The censoring process sets the maximal time a test taker is willing to invest into an item. If the processing time is shorter than the maximal time, the response is generated by the diffusion process; otherwise, the response is generated differently. In the first version of the model, the response is generated by a random guess. In the second version of the model, the response is determined by the actual level of the diffusion process. We relate the diffusion process to the capability and response caution of a test taker and the censoring process to his willingness to invest time. Similar to models for rapid guessing, the proposed models take account of disengaged responding, but allow for individual differences in persistence and informed guessing. The model also provides a mathematical specification of persistence. In a simulation study, we investigate model fitting by marginal maximum likelihood estimation. We also apply the model to two empirical data sets.

Keywords: ability-based guessing; diffusion model; item response theory; persistence; response time

The performance in a test does not only reflect a test taker's level of capability, but also his motivation to do well (Knekta, 2017). When test takers have little motivation to take a test, they tend to respond prematurely in order to avoid mental work as soon as test items become harder (Debeer et al., 2014; Wise & Smith, 2011). This results in an overestimation of the item difficulties, an underestimation of the average capability level of the test takers, and threatens the validity of the test when the test scores are interpreted in terms of capability solely (Oshima, 1994; Rios, 2022; van Laar & Braeken, 2022).

Premature responses are often conceived as rapid guesses (Wise, 2017). Rapid guesses are fast responses based on an incomplete response process that does not reflect the capability of a test taker (Wise & DeMars, 2006). Rapid guesses are typically accounted for by mixture models that assume two

© The Author(s), 2026. Published by Cambridge University Press on behalf of Psychometric Society.
This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial licence (https://creativecommons.org/licenses/by-nc/4.0), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original article is properly cited. The written permission of Cambridge University Press or the rights holder(s) must be obtained prior to any commercial use.

https://doi.org/10.1017/psy.2026.10112 Published online by Cambridge University Press

qualitatively different modes of responding (dual-process assumption).[1] In the first response mode, a test taker employs his capabilities and responds in a regular way, engaging with the content of the item. Such regular responses are modeled by an item response model. In the second response mode, the test taker does not employ his capabilities and responds without any serious effort by a guess on chance level. Several mixture models for rapid guessing have been proposed in the past. In some models, all responses of a rapid guesser are rapid guesses (Liu et al., 2020; Meyer, 2010), while in others, it is just a subset (Wang & Xu, 2015). Some models consider solely responses (Bolt et al., 2002; Boughton & Yamamoto, 2007), some solely the response times (Schnipke & Scrams, 1997), and some responses and response times (Alós-Ferrer, 2018; Molenaar et al., 2018). In some models, the likelihood of rapid guesses depends on the test taker (Lu et al., 2020), in some on the item (Wang & Xu, 2015), and in some on both (Ulitzsch et al., 2019).

Alternatively, the phenomenon of premature responding can be interpreted as an incomplete response process, whereby the regular response process is interrupted when it exceeds the time a test taker is willing to invest into an item. This can be modeled with race models that assume a race between the response process and a process of disengagement. In the race models, a regular response is given when the response process is completed before disengagement occurs, whereas a premature response is given when the process of disengagement wins. Measurement models based on a race have been proposed by Guo et al. (2025), Lee and Ying (2015), Lu and Wang (2020), Ranger and Kuhn (2014), and Ranger et al. (2025); see also Glickman et al. (2005) and Hawkins and Heathcote (2021) for the application of race models in experimental psychology. The race models differ in their assumptions about the race and its consequences. In some models, the regular response is always correct (Ranger & Kuhn, 2014), while in others, the regular response is modeled by an item response model (Guo et al., 2025; Lee & Ying, 2015; Lu & Wang, 2020). In some models, the premature response is a wrong response (Lee & Ying, 2015; Ranger & Kuhn, 2014), a guess (Ranger et al., 2025), or an omission (Guo et al., 2025; Lu & Wang, 2020). The models also differ in their distributional assumptions about the response process and the process of disengagement.

Mixture models and race models have been used successfully in psychological and educational assessment. However, they rely on a simplified view of the process that produces premature responses. The models allow for individual differences in the frequency of premature responses either by allowing for individual differences in the mixing proportion or the willingness to invest time. The response process underlying the premature responses, however, does not depend on characteristics of the test takers or the time on the task as premature responses are either wrong responses, omissions, or random guesses; but see the model of Ranger et al. (2025) for an exception. Guesses are rarely entirely random (Burton, 2002; Noventa et al., 2024). Even though the response process was incomplete, the test takers might still have partial knowledge that could be used for an informed guess, either by a comparative evaluation of the response options or an elimination of distractors (Gonthier, 2023; Lau et al., 2011; Weber et al., 2023). Mixture and race models thus capture only an extreme form of premature responding—one in which test takers either lack any relevant information or disregard the information when responding.

In this article, we propose a latent trait model for the responses and the response times on a test with a single-choice format. The model accounts for premature responses based on a race between an information accumulation process and a process of disengagement. The information accumulation process is modeled by a diffusion-based item response theory (diffIRT) model (Molenaar et al., 2015; Tuerlinckx et al., 2016; van der Maas et al., 2011). We propose two versions of the model. In the first version, premature responses are random guesses. In a second version, the premature responses are informed by the accumulated information. The proposed model makes several contributions. In contrast to the mixture and race models reviewed above, the model allows for estimating item- and person-specific parameters related to the disengagement process, which may serve as indicators of

[1] Rapid guesses are not necessarily due to low motivation, but also occur in tests with a strict time limit when test takers run out of time (Wise, 2017). This form of prematurely responding is not considered in the manuscript.

https://doi.org/10.1017/psy.2026.10112 Published online by Cambridge University Press

test-taking engagement. The model also accomodates how the guessing process depends on test taker characteristics and the time spent on task. In contrast to the race model of Ranger et al. (2025), the model incorporates within-trial noise in evidence accumulation and does not predict responses and response times deterministically from trial-wise parameters (Kang et al., 2023). Moreover, the diffusion model allows for a more detailed description of test-taking engagement than race model of Ranger et al. (2025) by considering the response caution of a test taker. Finally, the proposed model is an alternative to the mixture-based diffusion model introduced by Alós-Ferrer, (2018).

The outline of this manuscript is as follows. First, we describe the model in detail. Then, we investigate whether the values of the item parameters and the latent traits can be recovered by means of a simulation study. Finally, we illustrate the application of the model with two empirical examples.

1. The extended diffIRT model

The proposed model builds on the diffusion model that is first extended by a censoring process to account for prematurely responding and then embedded within a hierarchical framework to account for individual differences among test takers. We first review the diffusion model, then describe its extension by the censoring process, and finally address its integration into a hierarchical model.

1.1. The diffusion model

The diffusion model was originally developed in experimental psychology for binary decision tasks (Ratcliff, 1978), but has been used lately for responses and response times from psychological tests (Kang et al., 2022; Tuerlinckx & De Boeck, 2005; Tuerlinckx et al., 2016; van der Maas et al., 2011). In these applications, the response is interpreted as a binary decision between the correct and an incorrect response. The diffusion model is based on an information accumulation process. The momentary level of information at time t is represented by the random variable $X(t)$ ($X(t) \in \mathbb{R}$). It may be interpreted as the relative evidence in favor of the correct ($X(t) > 0$) or an incorrect response ($X(t) < 0$). In the unbiased version of the diffusion model, the test takers start without information ($X(t = 0) = 0$). The change in $X(t)$ over time is characterized by the stochastic differential equation of a Wiener process with drift

$$dX(t) = v \cdot dt + dB(t), \tag{1}$$

where v is the drift rate and $dB(t)$ is a Brownian increment. The drift rate determines the expected level of information at point of time t via the relation $\mathrm{E}\big(X(t)\big) = v \cdot t$. The Brownian increment represents noise and generates deviations of the actual level from the expected level of information. Its variability is determined by a parameter that is usually fixed to a specific value for the sake of model identification.

Test takers accumulate information until the momentary level of information $X(t)$ hits either an upper or a lower boundary. In case the upper boundary is hit, the test taker selects the correct response. In the other case, an incorrect response is selected. Upper and lower boundaries are denoted as a_C ($a_C \in \mathbb{R}^+$) and a_I ($a_I \in \mathbb{R}^-$) in the following. The time required to select a response is the first hitting time. In practice, the processing response time is longer than the first hitting time. Some extra time is required for reading the item or giving the response. This extra time is denominated as the non-decision time d ($d \in \mathbb{R}^+$). The processing time is the sum of the non-decision time and the first hitting time.

An illustration of the response process is given in Figure 1, left plot (diffusion model). The straight black line represents the expected level of information of a test taker with drift rate $v = 1$ as a function of time. The saw-toothed line represents the actual level of information $X(t)$. The actual level deviates from the expected level due to the Brownian increment. The diffusion process starts with a delay of $d = 0.5$ time units. It requires $t^* = 1.5$ time units in order to hit a boundary for the first time. As the upper boundary $a_C = 2$ is hit, the evidence for the correct response outweighs the evidence for the incorrect response and the test taker selects the correct response. The processing time is the non-decision time plus the first hitting time $t = t^* + d = 1.5 + 0.5 = 2.0$. In Figure 1, the non-decision time delays the start

https://doi.org/10.1017/psy.2026.10112 Published online by Cambridge University Press

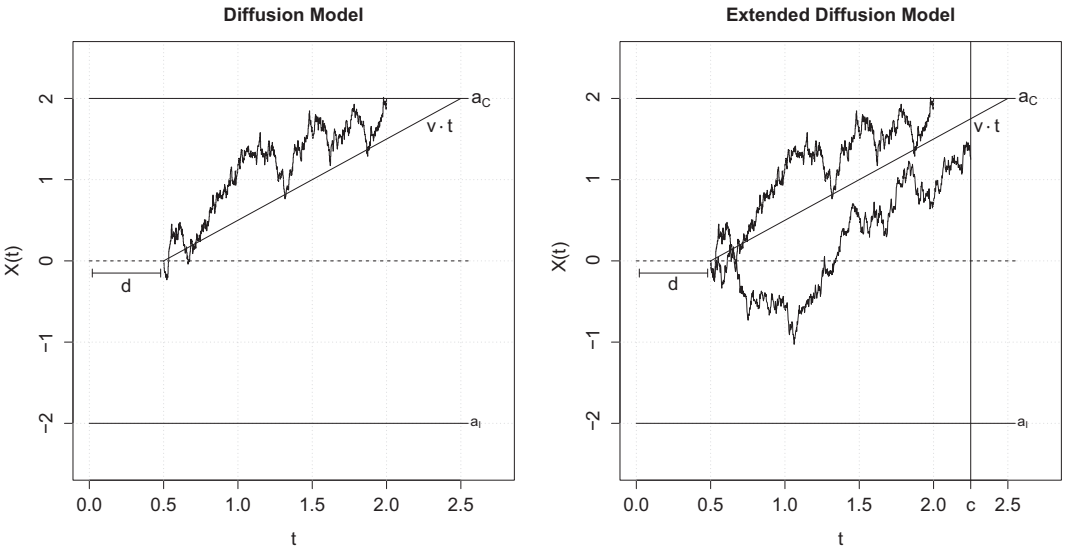


Figure 1. Illustration of the response process implied by the diffusion model (left plot) and the extended diffusion model with censoring (right plot). The straight black line represents the expected level of information for a test taker with drift rate $v = 1$ and non-decision time $d = 0.5$. The saw-tooth line represents the actual level of information of a test taker. Boundaries are denoted by a_C and a_I. The censoring time is represented by c.

of the information accumulation process, rather than constituting the post-decision phase of giving the response. We consider this more plausible in psychological assessment as reading an item requires considerably more time than the physical act of responding.[2]

In the following, we assume that the thresholds are located symmetrically around zero ($a_I = -a_C$). This avoids bias in the response process and is plausible in case test takers do not have prior knowledge of which of the response options is correct (Cizek & Wollack, 2017). When boundaries are symmetric, the diffusion process can be parameterized in terms of two parameters, the drift rate v and the boundary separation $a = a_C - a_I$ ($a \in \mathbb{R}^+$).[3] Both parameters characterize different aspects of the information accumulation process. The drift rate determines the expectation of the accumulated information. It represents the information processing capability of a test taker. Test takers with high levels of v tend to give correct responses fast. The boundary separation a is related to the level of information that is required for the test taker to respond. As test takers with high levels of a tend to respond slowly, but correctly, a is interpreted as the response caution of a test taker (Boehm et al., 2021). The non-decision time is a shift parameter that is usually not considered as a parameter pertaining to the information accumulation process.

The diffusion model was originally proposed for binary responses. The model can be applied to single-choice tasks in case all incorrect response options are grouped. This requires the assumption that the diffusion process reflects the comparison of the correct response with the most attractive incorrect response option. The assumption is, for example, reasonable in case all incorrect response options are roughly equally attractive (Boehm et al., 2021). Alternatively, the diffusion model can be adapted to items with single-choice format (van der Maas et al., 2011) or can be used in order to

[2]The question whether the non-decision time delays the start of the information accumulation process or prolongates the post-decision phase is usually irrelevant for the diffusion model. It, however, becomes relevant when the process of information accumulation is censored.

[3]In the manuscript, we assume that test takers start at $X(t = 0) = 0$ and accumulate information until either reaching $a_I = -a/2$ or $a_C = a/2$. Often, the diffusion process is parameterized in terms of $X(t = 0) = a/2$, $a_I = 0$, and $a_C = a$. Both parameterizations are equivalent. We chose the former as it is the more natural one when informed guessing is considered.

https://doi.org/10.1017/psy.2026.10112 Published online by Cambridge University Press

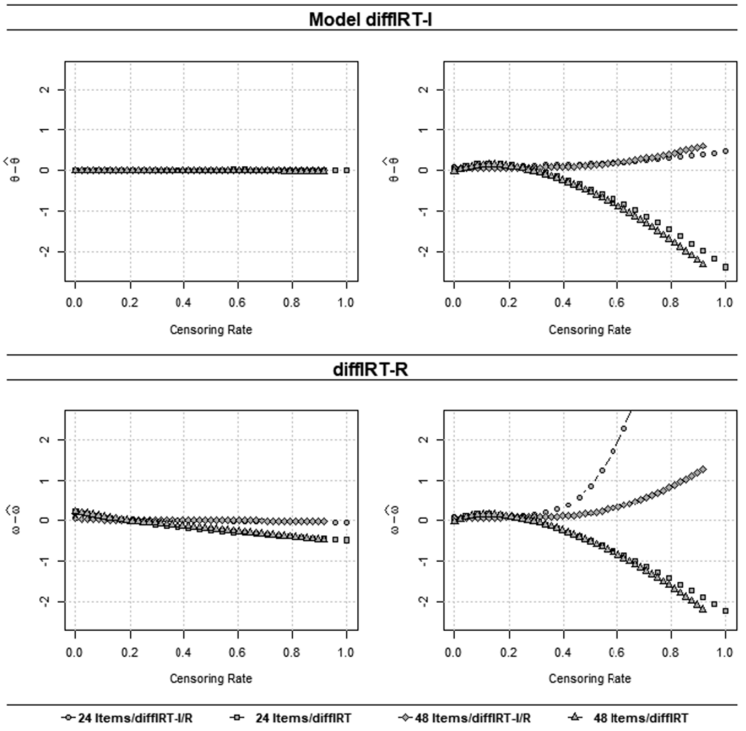


Figure 2. Bias of the estimates of capability θ and response caution ω when the data-generating model and when the standard diffusion model is used for estimation as a function of the rate of premature responses (censoring rate).

model response-specific information accumulation (Tillman et al., 2020), but this requires further extensions.

1.2. The extension of the diffusion model by a censoring process

The time a test taker is willing to spend on the items is an important aspect of test-taking engagement that is closely related to the tendency to respond prematurely (Baumert & Demmerich, 2001; Eklöf, 2010; Knekta, 2017). In the diffusion model, there is no direct counterpart to the willingness to invest time. Motivational aspects of test-taking are represented by the boundary separation. The boundary separation, however, is conceptually different to the time a test taker is willing to invest. It primarily reflects the aspired level of information that might not be reached in case information accumulation is slow and test takers give up. The boundary separation is also no substitute for the willingness to invest time pragmatically as it is indeed related to the expected response time, but does not limit the actual time on the task.

We propose an extension of the diffusion model in order to account for the willingness to invest time. In the extension, we assume that each test taker sets a censoring time c. The censoring time c is the maximal time, a test taker would spend on an item. The censoring time censors the response process as follows. In case the time required to select the response (first hitting time of the diffusion process plus the non-decision time) does not exceed the censoring time, the response process is ended regularly and the response is generated by the diffusion model. Otherwise, the response process is interrupted prematurely at the censoring time c. Premature responses are generated according to two alternative procedures. In the first version of the model, we assume that the response is randomly selected among the available response options. The correct response is selected with the fixed guessing probability π. The momentary level of information $X(c)$ has no impact on the selection as it is completely ignored. In the second version of the model, the test taker responds based on the information that was accumulated

https://doi.org/10.1017/psy.2026.10112 Published online by Cambridge University Press

until the interruption. In case the level of the accumulated information is positive ($X(c) > 0$), the test taker selects the correct response. Otherwise, the test taker selects the incorrect response. Note that only the sign of $X(c)$ matters, not the absolute value $|X(c)|$. Due to the randomness in $X(c)$, the model implies that responses are fast and almost on chance level when the censoring time is short. To account for the randomness in the process of disengagement, we assume that the censoring time c is the realization of a log-normally distributed random variable with scale parameter μ and shape parameter σ. The parameters of the log-normal distribution determine the expected censoring time and will later be related to characteristics of the test taker and the item.

An illustration of the response process is given in Figure 1, right plot (extended diffusion model). The plot depicts the accumulated information of two test takers. Both test takers are willing to spend at most $c = 2.25$ time units on the item. This is the maximal time the delayed process of information accumulation is allowed to last. One of the test takers manages to finish the response process regularly within the time limit by hitting the upper boundary at $t = 2.00$. This test taker responds after $t = 2.00$ time units correctly. The other test taker gives up and stops information accumulation at $c = 2.25$. In the first version of the model, the test taker would guess randomly. In the second version of the model, the test taker would choose the more plausible response. As the accumulated information at $c = 2.25$ time units is positive, the test taker would respond correctly after $t = 2.25$ time units.

In the second version of the extended diffusion model, the premature responses depend on partial knowledge as premature responses are determined by the sign of the accumulator when the response process is interrupted (sign rule). This response process has some resemblance to the interrogation paradigm (Bogacz et al., 2006). In the interrogation paradigm, the test takers are obliged to respond at a fixed point of time. This fixed point of time, however, is set by the experimenter while in our model, the censoring time is a random variable depending on characteristics of the test taker. Furthermore, in the interrogation paradigm, the test takers are supposed to accumulate information until they are allowed to respond by the experimenter, while in our model, the test takers may respond earlier. Selecting a response based on the sign guarantees that the more likely response is selected (Bogacz et al., 2006). The sign rule can also be justified on psychological grounds. In the diffusion model, the accumulator represents the impression that one response option is more plausible than the other (Fehr & Rangel, 2011), the net evidence in favor of one alternative against the other (Baldassi et al., 2020) or the balance of the support for the two alternatives (Kvam, 2019). It can be interpreted as the difference between two accumulators representing evidence for each response option (Webb, 2019). The decision rule is relative as a response is given as soon as the evidence for one alternative over the other exceeds a threshold. Even when this threshold is not reached, as it happens when the process of information accumulation is interrupted, the relative decision rule still requires that the alternative with more net evidence (as indicated by the sign of the accumulator) is chosen. This response mechanism is different to the one in the model of Ranger et al. (2025). There, the premature responses were informed guesses with the guessing probability depending on the actual level of the accumulator. We consider this implementation to be inadequate for the diffusion model for two reasons. First, in the model of Ranger et al. (2025), the accumulator represents the actual progress toward the solution, not the relative evidence. Some progress might enable test takers to eliminate distractors, thereby increasing the guessing probability over chance level. Second, the accumulator in the diffusion model is supposed to fully characterize the response process. In case the interruption would trigger a random guess based on the actual level of the accumulator, the same level could result into different responses. Hence, there have to be differences in the knowledge state that are not fully captured by the accumulator and would require a random mechanism beyond the Brownian motion.

1.3. A hierarchical extension of the diffusion model with censoring

The parameters of the extended diffusion model are related to latent traits and item effects via a hierarchical model in order to account for differences between test takers and items. Denote by v_{ij} the drift rate and by a_{ij} the boundary separation of test taker i ($i = 1, \ldots, N$) on item j ($j = 1, \ldots, J$) of a test.

https://doi.org/10.1017/psy.2026.10112 Published online by Cambridge University Press

We relate the drift rate via a linear model to a first trait θ_i and the boundary separation via a transformed linear model to a second trait ω_i:

$$\begin{aligned} v_{ij} &= v_j(\theta_i) = v_{0j} + v_{1j}\theta_i \\ a_{ij} &= a_j(\omega_i) = \log\left(1 + \exp(a_{0j} + a_{1j}\omega_i)\right). \end{aligned} \tag{2}$$

In Equation (2), the latent traits θ_i ($\theta \in \mathbb{R}$) and ω_i ($\omega \in \mathbb{R}$) are the capability and the response caution of test taker i, respectively. The capability comes close to what Thurstone (1937) called the power of a test taker and is usually the target trait of the assessment. The response caution has a close resemblance to the effort that is needed for a successful termination of the problem-solving process in the theory of test-taking motivation of Knekta (2017) and represents the aspiration level of a test taker. The item parameters v_{0j} ($v_{0j} \in \mathbb{R}$) and a_{0j} ($a_{0j} \in \mathbb{R}$) determine the drift rate and boundary separation of a reference test taker with trait levels $\theta_i = 0$ and $\omega_i = 0$. Item parameter v_{0j} reflects the difficulty of an item, as it is related to the ease of information accumulation. Item parameter a_{0j} reflects the perceived importance of an item as it determines the typical level of response caution. The item parameters v_{1j} ($v_{1j} \in \mathbb{R}$) and a_{1j} ($a_{1j} \in \mathbb{R}$) determine the interindividual variance over the test takers and can be interpreted similar to the loadings in factor analysis. The transformation of the linear predictor $a_{0j} + a_{1j}\omega_i$ guarantees that the resulting boundary separation a_{ij} is positive. Although in diffIRT models, a log-link is usually used for this purpose, we prefer the transformation given in Equation (2) as it is almost linear for higher levels of the linear predictor. This makes the model numerically more stable. It, however, changes the interpretation as the relation of $a_j(\omega)$ to ω in Equation (2) is not multiplicative as in the log-link. The non-decision time d_j is treated as an item parameter, as in the diffIRT model of Molenaar et al. (2015). The parameter that determines the variability of the Brownian motion is set to one for the sake of model identification (Ratcliff, 1978).

The scale parameter μ_{ij} of the log-normal distribution generating the censoring time of test taker i in item j is related to a third latent trait τ_i by a linear model:

$$\mu_{ij} = \mu_j(\tau_i) = c_{0j} + c_{1j}\tau_i. \tag{3}$$

In Equation (3), the latent trait τ_i ($\tau \in \mathbb{R}$) is the willingness of a test taker to invest time into an item. It maps to the persistence of a test taker in the theory of test-taking motivation by Knekta (2017) and accounts for the stable individual differences in test-taking engagement that have recently been found in large-scale assessment (Buchholz et al., 2006). The item parameters c_{0j} ($c_{0j} \in \mathbb{R}$) and c_{1j} ($c_{1j} \in \mathbb{R}$) are interpreted as in a linear model, with c_{0j} being the intercept and c_{1j} the regression coefficient. The shape parameter σ_j is considered as an item parameter, as in the log-normal factor model of van der Linden (2006). The hierarchical extension of the diffusion model with censoring will be denoted as diffIRT-R model in case the premature responses are generated by random guessing and as diffIRT-I model in case the premature responses are generated by the sign rule.

1.4. The distribution of the responses and the response times in a test

The proposed model is a race model where the information accumulation process and the censoring process compete. The observed response time $t_{ij} = \min(t'_{ij}, c_{ij})$ is the minimum of the processing time (first hitting time plus non-decision time) t'_{ij} and the censoring time c_{ij}. The distribution of the observed responses and response times can be derived from standard findings for race models.

In the diffIRT-R model, a specific response is given in case the diffusion process wins by hitting the corresponding boundary before the censoring time or in the case the censoring process wins and the given response is selected by a random guess. Random guesses are generated by a Bernoulli experiment with fixed guessing probability π. The defective density function of the response x_{ij} (1: Correct, 0: Incorrect) and the observed response time t_{ij} of a test taker i with fixed latent traits $(\theta_i, \omega_i, \tau_i)$ in

https://doi.org/10.1017/psy.2026.10112 Published online by Cambridge University Press

item j is

$$\begin{aligned}&\mathrm{f}\left(x_{ij},t_{ij};v_j(\theta_i),a_j(\omega_i),d_j,\mu_j(\tau_i),\sigma_j,\pi\right)\\&\quad=\mathrm{f_D}\left(x_{ij},t_{ij};v_j(\theta_i),a_j(\omega_i),d_j\right)\cdot\mathrm{S_C}\left(t_{ij};\mu_j(\tau_i),\sigma_j\right)\\&\quad\quad+\mathrm{f_C}\left(t_{ij};\mu_j(\tau_i),\sigma_j\right)\cdot\mathrm{S_D}\left(t_{ij};v_j(\theta_i),a_j(\omega_i),d_j\right)\cdot(\pi)^{x_{ij}}\cdot(1-\pi)^{1-x_{ij}}.\end{aligned}\tag{4}$$

Functions $\mathrm{f_C}\left(t_{ij};\mu_j(\tau_i),\sigma_j\right)$ and $\mathrm{S_C}\left(t_{ij};\mu_j(\tau_i),\sigma_j\right)$ are the density function and the survival function of the log-normal distribution (Crow & Shimizu, 1988) with scale parameter $\mu_j(\tau_i)$ and shape parameter σ_j, respectively. Function $\mathrm{f_D}\left(x_{ij},t_{ij};v_j(\theta_i),a_j(\omega_i),d_j\right)$ is the defective density function of the response and the response time implied by the diffusion process with drift rate $v_j(\theta_i)$, boundary separation $a_j(\omega_i)$, and shift parameter d_j (Tuerlinckx & De Boeck, 2005, Equation 3). Function $\mathrm{S_D}\left(t_{ij};v_j(\theta_i),a_j(\omega_i),d_j\right)$ is the survival function of the diffusion process. In Equation (4), the first term accounts for the case that the censoring process loses and the response is determined by the diffusion process, while the second term accounts for the case that the diffusion process loses and the response is determined by a random guess (see Supplement 1 for more details).

In the diffIRT-I model, a specific response is given in case the diffusion process wins by hitting the corresponding boundary before the censoring time or the censoring process wins and the given response is determined by the sign rule. The defective density function of the response x_{ij} (1: Correct, 0: Incorrect) and the observed response time t_{ij} of a test taker i with fixed latent traits $(\theta_i,\omega_i,\tau_i)$ in item j is

$$\begin{aligned}&\mathrm{f}\left(x_{ij},t_{ij};v_j(\theta_i),a_j(\omega_i),d_j,\mu_j(\tau_i),\sigma_j\right)\\&\quad=\mathrm{f_D}\left(x_{ij},t_{ij};v_j(\theta_i),a_j(\omega_i),d_j\right)\cdot\mathrm{S_C}\left(t_{ij};\mu_j(\tau_i),\sigma_j\right)\\&\quad\quad+\mathrm{f_C}\left(t_{ij};\mu_j(\tau_i),\sigma_j\right)\cdot\mathrm{S_D^*}\left(x_{ij},t_{ij};v_j(\theta_i),a_j(\omega_i),d_j\right).\end{aligned}\tag{5}$$

As before, function $\mathrm{f_C}\left(t_{ij};\mu_j(\tau_i),\sigma_j\right)$ is the density function of the log-normal distribution, function $\mathrm{S_C}\left(t_{ij};\mu_j(\tau_i),\sigma_j\right)$ is the survival function, and function $\mathrm{f_D}\left(x_{ij},t_{ij};v_j(\theta_i),a_j(\omega_i),d_j\right)$ is the defective density function of the response and the response time implied by the diffusion process. Function $\mathrm{S_D^*}\left(x_{ij},t_{ij};v_j(\theta_i),a_j(\omega_i),d_j\right)$ denotes the probability that the hitting time of the diffusion process is longer than t_{ij} and the value of the accumulator is positive ($X(t_{ij})>0$) when $x_{ij}=1$ or negative ($X(t_{ij})<0$) when $x_{ij}=0$ at t_{ij}. This probability follows from the distribution of the accumulator in non-terminated processes at the censoring time (Ratcliff, 1980, Equation 12/13). The distribution has to be integrated from 0 to $a/2$ in case the response is correct and from 0 to $-a/2$ in case the response is incorrect. In Equation (5), the first term accounts for the case that the censoring process loses and the response is determined by the diffusion process, while the second term accounts for the case that the diffusion process loses and the response is determined by the sign of the accumulator (see Supplement 1 for more details).

The functions in Equation (4) or Equation (5) are the defective density functions of the response and the response time of test takers with fixed latent traits $(\theta_i,\omega_i,\tau_i)$ on a single item. They specify the distribution of the response and response time conditional on the latent traits. The joint distribution of the responses and response times of a test taker with fixed trait levels on all items of a test follows from the factorization theorem if one makes the conditional independence assumption. The marginal distribution of the responses and response times is the unconditional distribution when latent traits are also sampled randomly. It is obtained by integrating the conditional distribution over the distribution of the latent traits in the population of the potential test takers. This requires an assumption about the distribution of the latent traits. Here, we assume that the latent traits are multivariate normally distributed, with expected values of zero, variances of one, and correlation matrix $\mathbf{\Sigma}$. Assumptions about the expectation and variance are necessary in order to identify the scale of the latent traits.

https://doi.org/10.1017/psy.2026.10112 Published online by Cambridge University Press

2. Simulation study

We conducted three simulation studies. In the first study, we investigated parameter recovery with marginal maximum likelihood (MML) estimation. In the second study, we investigated whether information criteria accurately identify the data-generating model among several alternatives. In the third study, we investigated the performance of the proposed models as measurement models of the latent traits and their capability to detect premature responses.

2.1. Simulation study I: Parameter recovery

In the first study, we generated data according to the extended diffIRT models, estimated the values of the models' parameters with MML estimation, and compared the estimates to the true values of the parameters. We considered 2×2 simulation conditions defined by the model (diffIRT-R/diffIRT-I) and the sample size ($N = 500$/ $N = 1{,}000$). The number of items was set to $J = 24$. For each simulation condition, we analyzed 100 simulation samples.

Data were generated as follows. We first simulated values for the latent traits of fictitious test takers by random draws from a standardized multivariate normal distribution. For all fictitious test takers and items, we determined the parameters of the diffusion model ($v_j(\theta_i)$, $a_j(\omega_i)$) and the log-normal distribution ($\mu_j(\tau_i)$) according to Equations (2) and (3). We then drew censoring times from a log-normal distribution with parameters $\mu_j(\tau_i)$ and σ_j and simulated the diffusion process with parameters $v_j(\theta_i)$, $a_j(\omega_i)$, and d_j using the `Sim.DiffProc` package (Guidoum & Boukhetala, 2020) from the statistical computing environment `R` (R Core Team, 2021). The responses and the response times were determined as illustrated in Figure 1. In the diffIRT-R model, the premature responses were random guesses generated according to a Bernoulli experiment with parameter $\pi = 0.2$. In the diffIRT-I model, the premature responses were generated according to the sign rule.

For data generation, we used the following model parameters. The correlations of the traits were set to $\rho_{\theta\omega} = 0.40$, $\rho_{\theta\tau} = 0.20$, and $\rho_{\omega\tau} = 0.30$. We used positive values as we assumed that test takers with higher capability are also more engaged. The item parameters of the diffusion model were set to the values $a_{0j} \in \{2.5, 3.0, 3.5\}$, $a_{1j} \in \{0.8, 1.2\}$, $v_{0j} \in \{-1.0, -0.5, 0.5, 1.0\}$, $v_{1j} \in \{0.8, 1.2\}$, and $d_j \in \{0.25\}$. The item parameters of the log-normal model were set to the values $c_{0j} \in \{0.9, 1.1\}$, $c_{1j} \in \{0.3\}$, and $\sigma_j \in \{0.4\}$. The guessing probability of the diffIRT-R model was set to $\pi = 0.2$. When generating the data, we intended to mimic the data from a test on mental rotation (Gaviria, 2005). We, however, extended the range of v_{0j} in order have more variation in the item difficulties than the mental rotation items had. We also took care to induce a rather low amount of disengaged responding as even in low stakes tests, most test takers are motivated to do well (e.g., Michaelides et al., 2020; Wise, 2017). A detailed list of the parameters values in all items is given in Supplementary Table S2.1.

For each data set, the parameters of the data-generating model were estimated using MML estimation. In MML estimation, one determines those values of the model parameters that maximize the marginal likelihood function (Berger et al., 1999, see Integrated Likelihood). The marginal likelihood function was implemented in the `Julia` programming language (Bezanson et al., 2017). The density function of the diffusion model was evaluated with the `DiffModels.jl` package (Drugowitsch, 2017). The function $S_D^*\left(x_{ij}, t_{ij}; v_j(\theta_i), a_j(\omega_i), d_j\right)$ was approximated by limiting the infinite sum to 30 terms. This approximation was sufficient for the present application. Integrals over the latent traits were approximated with the Mislevy quadrature using 20 nodes in case of the diffIRT-R model and 15 nodes in case of the diffIRT-I model (Kim et al., 2015). We reduced the number of nodes in the diffIRT-I model in order to save computation time. The marginal likelihood function was maximized within the statistical computing environment `R`. We used the package `optim` and the `BFGS` algorithm (Nocedal & Wright, 2006). In addition to point estimates, we also determined Wald-type confidence intervals for a confidence level of $c = 0.95$ ($\alpha = 0.05$) using the observed information matrix. The observed information matrix was approximated by the variance–covariance matrix of the score values.

https://doi.org/10.1017/psy.2026.10112 Published online by Cambridge University Press

Table 1. True value (TV), average estimate (M), standard error of estimation (SE), and coverage frequency (CI) of confidence intervals of the parameter estimates for different samples sizes N in the diffIRT-R model with random guessing

	$N = 500$				$N = 1{,}000$			
Par	TV	M	SE	CI	TV	M	SE	CI
a_0	2.500	2.536	0.111	0.990	2.500	2.522	0.069	0.980
	3.000	3.027	0.138	0.996	3.000	3.016	0.099	0.965
	3.500	3.527	0.186	0.986	3.500	3.515	0.126	0.975
a_1	0.800	0.804	0.088	0.993	0.800	0.802	0.059	0.979
	1.200	1.211	0.097	0.992	1.200	1.208	0.067	0.966
v_0	−1.000	−1.011	0.086	0.980	−1.000	−1.003	0.056	0.970
	−0.500	−0.507	0.069	0.987	−0.500	−0.502	0.049	0.967
	0.500	0.497	0.082	0.978	0.500	0.496	0.052	0.968
	1.000	1.001	0.083	0.978	1.000	1.001	0.056	0.972
v_1	0.800	0.809	0.070	0.988	0.800	0.801	0.050	0.964
	1.200	1.213	0.086	0.991	1.200	1.204	0.059	0.974
d	0.250	0.254	0.017	0.977	0.250	0.252	0.011	0.967
c_0	0.900	0.911	0.064	0.983	0.900	0.906	0.045	0.964
	1.100	1.107	0.073	0.979	1.100	1.105	0.049	0.972
c_1	0.300	0.300	0.074	0.983	0.300	0.304	0.049	0.963
σ	0.400	0.386	0.058	0.982	0.400	0.393	0.039	0.965
π	0.200	0.201	0.020	0.960	0.200	0.201	0.013	0.970
$\rho_{\theta\omega}$	0.400	0.402	0.047	0.990	0.400	0.401	0.034	0.960
$\rho_{\theta\tau}$	0.200	0.185	0.080	0.980	0.200	0.200	0.064	0.940
$\rho_{\omega\tau}$	0.300	0.283	0.092	0.980	0.300	0.290	0.065	0.960

Note: Results are based on 100 simulation samples. Results for parameters have been averaged over the items of with the same parameter values. Confidence intervals were determined for a confidence level of $c = 0.95 (\alpha = 0.05)$. Standard error of estimation refers to the parameter estimates.

The results of the simulation study are summarized in Table 1 for the diffIRT-R model and in Table 2 for the diffIRT-I model. In both tables, we report the true value (TV), the average estimate (M), the standard of error of estimation of the parameters (SE), and the coverage frequency (CI) of confidence intervals. For ease of presentation, the results concerning the parameters have been averaged over items with the same value.

In the diffIRT-R model, the MML estimator is virtually unbiased in most parameters. There is a slight bias in parameter a_0 for both sample sizes and in parameter σ for samples of $N = 500$ subjects. The bias, however, is not large and should not have any significance in practice. The coverage frequencies of the confidence intervals exceed the intended level of $c = 0.95$. In samples of $N = 1{,}000$, the coverage frequencies are closer to the intended level of $c = 0.95$, but still slightly too high.

In the diffIRT-I model, the MML estimator is virtually unbiased in all parameters except parameter σ. Results concerning the coverage frequencies of the confidence intervals depend on the sample size. In samples of $N = 500$, the coverage frequencies are slightly too high in the parameters of the diffusion model, but slightly too low in the parameters of the log-normal model and the coefficients of correlation. In samples of $N = 1{,}000$, the coverage frequencies are close to the intended level of $c = 0.95$.

https://doi.org/10.1017/psy.2026.10112 Published online by Cambridge University Press

Table 2. True value (TV), average estimate (M), standard error of estimation (SE), and coverage frequency (CI) of confidence intervals of the parameter estimates for different samples sizes N in the diffIRT-I model where premature responses are generated by the sign rule

	$N = 500$				$N = 1{,}000$			
Par	TV	M	SE	CI	TV	M	SE	CI
a_0	2.500	2.499	0.124	0.969	2.500	2.516	0.080	0.960
	3.000	3.001	0.144	0.975	3.000	3.009	0.102	0.966
	3.500	3.498	0.202	0.970	3.500	3.504	0.140	0.959
a_1	0.800	0.798	0.086	0.989	0.800	0.793	0.060	0.971
	1.200	1.199	0.101	0.983	1.200	1.185	0.072	0.923
v_0	−1.000	−1.007	0.087	0.967	−1.000	−1.004	0.056	0.935
	−0.500	−0.508	0.077	0.938	−0.500	−0.504	0.048	0.942
	0.500	0.499	0.078	0.948	0.500	0.497	0.050	0.937
	1.000	0.995	0.080	0.960	1.000	0.995	0.054	0.940
v_1	0.800	0.796	0.065	0.980	0.800	0.796	0.047	0.948
	1.200	1.199	0.085	0.974	1.200	1.195	0.058	0.957
d	0.250	0.254	0.018	0.980	0.250	0.253	0.012	0.974
c_0	0.900	0.913	0.086	0.966	0.900	0.909	0.055	0.954
	1.100	1.115	0.111	0.928	1.100	1.106	0.068	0.963
c_1	0.300	0.309	0.113	0.917	0.300	0.302	0.071	0.940
σ	0.400	0.370	0.100	0.927	0.400	0.386	0.059	0.951
$\rho_{\theta\omega}$	0.400	0.405	0.050	0.880	0.400	0.407	0.036	0.910
$\rho_{\theta\tau}$	0.200	0.181	0.103	0.720	0.200	0.203	0.064	0.950
$\rho_{\omega\tau}$	0.300	0.262	0.135	0.650	0.300	0.295	0.073	0.940

Note: Results are based on 100 simulation samples. Results for parameters have been averaged over the items of with the same parameter values. Confidence intervals were determined for a confidence level of $c = 0.95 (\alpha = 0.05)$. Standard error of estimation refers to the parameter estimates.

2.2. Simulation study II: Model selection

In the second simulation study, we addressed the question whether the diffIRT models can be distinguished empirically. For the study, we used the data sets with $N = 1{,}000$ subjects that were generated in the first simulation study. We did not consider the data sets with $N = 500$ subjects due to the high time demand of the data analysis. Each data set was analyzed with the diffIRT, diffIRT-I, and diffIRT-R models. We estimated the item parameters of each model with MML estimation as in the first simulation study. We then determined the Akaike information criterion (AIC) of Akaike, (1992) and the Bayes information criterion (BIC) of Schwarz, (1978) for the three models. The model with the lowest AIC or BIC value was selected as the one that was presumed to have generated the data (Burnham & Anderson, 2013). The relative frequencies by which each model was selected are reported in Table 3 separately for the two information criteria conditional on the true data-generating model.

In all simulation samples, the data-generating model is identified correctly. This suggests that the models can be distinguished empirically. The diffIRT model, for example, predicts generally longer response times than the diffIRT-I and diffIRT-R models. The diffIRT-I models and the diffIRT-R in turn differ in the accuracy of premature responses.

https://doi.org/10.1017/psy.2026.10112 Published online by Cambridge University Press

Table 3. Relative frequencies by which a model was selected as the alleged data-generating model for the information criteria AIC and BIC conditional on the true data-generating model

Criterion		AIC			BIC		
Selected model		diffIRT	diffIRT-R	diffIRT-I	diffIRT	diffIRT-R	diffIRT-I
True model	diffIRT-R	0.00	1.00	0.00	0.00	1.00	0.00
	diffIRT-I	0.00	0.00	1.00	0.00	0.00	1.00

Note: Results are based on 100 simulation samples. diffIRT: Diffusion model without censoring; diffIRT-R: Diffusion model with censoring and random guessing; diffIRT-I: Diffusion model with censoring and premature responding based on the sign rule.

2.3. Simulation study III: Trait recovery

In the third simulation study, we investigated whether the latent traits can be estimated reliably with maximum likelihood estimation, how trait recovery depends on the censoring rate and whether premature responses can be detected.

The simulation study was implemented as follows. By fully crossing nine trait levels $\{-2.0, -1.5, \ldots, 1.5, 2.0\}$, we defined 729 fictitious test takers who were characterized by a unique combination of capability, response caution, and willingness to invest time. We used fixed trait levels in order to study whether the maximum likelihood estimator is conditionally unbiased. For each test taker, we generated 50 responses and response time patterns with the diffIRT-I and diffIRT-R models. We considered a test with 24 items and a test with 48 items. For the test with 24 items, we used the same item parameters as in the first simulation study on parameter recovery. For the test with 48 items, we used each item twice. Data generation resulted in 729×50 response and response time pattern for each condition defined by the data-generating model and the length of the test.

We estimate the latent traits of all fictitious test takers with maximum likelihood estimation. The values of the item parameters were set to the true values. The true values of the item parameters were used to prevent confounding of item parameter and trait estimation. In addition to point estimates of the trait values, we determined Wald-type confidence intervals for a confidence level of $c = 0.95$ ($\alpha = 0.05$). The standard errors of estimation were determined by inverting the negative Hessian matrix. The Hessian matrix was approximated using numerical methods. We evaluated the maximum likelihood estimator with respect to the average estimate, the standard error of estimation, and the coverage frequency of the confidence intervals. The results are given in Table 4. Results for one trait have been aggregated over the other traits.

In the condition with 24 items, the estimator of capability has a small bias and large standard errors of estimation. The standard errors would correspond to reliability coefficients between 0.22 and 0.95 in case the squared standard errors were interpreted as the error variance. The large standard errors are partly due to a small proportion of extreme estimates. Coverage frequencies of the confidence intervals are close to the intended level of 0.95. In the condition with 48 items, the estimator is virtually unbiased and the standard errors of estimation are considerably smaller. The implied reliability coefficients are in the range from 0.77 to 0.98. The willingness to invest time does not impact the estimation of capability much (see Supplementary Figures S3.1 and S3.4). Though, the standard errors tend to be slightly larger when the willingness to invest time is low, especially for the diffIRT-R model. Presumably, this is because censored responses provide little information about a test taker's capability when they are generated through random guessing.

The estimator of response caution is biased in all conditions. The bias is smaller in the condition with 48 items than in the condition with 24 items. Standard errors are high. The corresponding reliability coefficients range from 0.02 to 0.62 in the test with 24 items and from 0.15 to 0.98 in the test with 48 items. Standard errors are higher and reliabilities are lower for high levels of response caution. Low levels of response caution, however, can be estimated well. Supplementary Figures S3.2 and S3.5 suggest that the reliability depends on the willingness to invest time. The standard error of estimation is highest when

https://doi.org/10.1017/psy.2026.10112 Published online by Cambridge University Press

Table 4. True value (TV), average estimate (M), standard error of estimation (SE), and coverage frequency (CI) of confidence intervals of the trait estimates for different item number J and the two extended diffIRT models

		diffIRT-I						diffIRT-R					
		$J = 24$			$J = 48$			$J = 24$			$J = 48$		
Par	TV	M	SE	CI	M	SE	CI	M	SE	CI	M	SE	CI
θ	−2.0	−2.20	1.36	0.93	−2.06	0.29	0.94	−2.14	0.78	0.93	−2.06	0.30	0.95
	−1.5	−1.60	0.53	0.94	−1.53	0.22	0.95	−1.62	1.06	0.93	−1.55	0.35	0.95
	−1.0	−1.05	0.28	0.94	−1.02	0.18	0.95	−1.10	1.39	0.93	−1.03	0.48	0.94
	−0.5	−0.52	0.24	0.95	−0.51	0.16	0.95	−0.54	1.88	0.93	−0.51	0.54	0.95
	0.0	0.01	0.23	0.95	0.00	0.16	0.94	0.06	1.20	0.93	0.04	0.41	0.95
	0.5	0.52	0.25	0.94	0.51	0.16	0.95	0.61	0.94	0.93	0.54	0.25	0.94
	1.0	1.05	0.27	0.95	1.02	0.18	0.95	1.09	0.46	0.93	1.05	0.30	0.94
	1.5	1.59	0.40	0.95	1.54	0.22	0.95	1.61	0.40	0.93	1.55	0.25	0.94
	2.0	2.16	0.71	0.95	2.05	0.26	0.95	2.14	0.42	0.94	2.07	0.28	0.94
ω	−2.0	−1.88	0.77	0.94	−1.94	0.11	0.94	−1.89	1.27	0.94	−1.94	0.11	0.93
	−1.5	−1.28	2.17	0.94	−1.43	0.25	0.94	−1.32	1.86	0.95	−1.42	0.85	0.95
	−1.0	−0.72	1.65	0.95	−0.90	0.46	0.95	−0.79	1.91	0.96	−0.93	0.19	0.95
	−0.5	−0.08	3.14	0.95	−0.37	0.76	0.95	−0.22	2.33	0.94	−0.42	0.35	0.95
	0.0	0.37	2.35	0.95	0.15	0.72	0.94	0.54	4.11	0.94	0.11	0.40	0.95
	0.5	0.87	1.38	0.95	0.71	0.97	0.95	1.13	4.57	0.94	0.68	1.31	0.95
	1.0	1.43	1.49	0.94	1.23	1.04	0.95	2.13	6.90	0.93	1.25	1.97	0.95
	1.5	2.03	2.20	0.94	1.77	1.13	0.94	2.62	6.14	0.93	1.75	1.43	0.94
	2.0	2.53	2.02	0.92	2.31	1.32	0.94	3.61	8.08	0.92	2.38	2.47	0.94
τ	−2.0	−1.89	1.08	0.91	−1.92	0.83	0.92	−1.84	1.08	0.92	−1.89	0.83	0.93
	−1.5	−1.40	1.14	0.91	−1.41	0.92	0.92	−1.33	1.23	0.92	−1.38	0.92	0.95
	−1.0	−0.98	1.25	0.90	−0.92	1.03	0.91	−0.80	1.37	0.91	−0.86	1.06	0.93
	−0.5	−0.50	1.36	0.90	−0.42	1.15	0.91	−0.28	1.49	0.91	−0.36	1.21	0.93
	0.0	−0.06	1.51	0.88	0.05	1.25	0.91	0.19	1.60	0.91	0.15	1.29	0.93
	0.5	0.34	1.62	0.86	0.48	1.37	0.90	0.64	1.74	0.90	0.62	1.57	0.92
	1.0	0.67	1.76	0.85	0.94	1.52	0.89	1.10	1.88	0.88	1.10	1.61	0.90
	1.5	1.11	1.91	0.84	1.32	1.62	0.87	1.48	1.99	0.87	1.54	1.73	0.89
	2.0	1.35	2.05	0.80	1.68	1.83	0.85	1.85	2.16	0.85	1.99	1.90	0.87

Note: Results for a trait have been averaged over the other traits. Results are based on 450 data patterns. Confidence intervals were determined for a level of confidence of $c = 0.95 (\alpha = 0.05)$. diffIRT-R: Diffusion model with censoring and random guessing; diffIRT-I: Diffusion model with censoring and premature responses based on the sign rule.

a high level of response caution is combined with a low willingness to invest time. This combination is especially problematic for the diffIRT-R model. In that case, almost all of the responses are random guesses that have no direct connection to the diffusion process.

Results are similar for the willingness to invest time. The estimator has a large bias, especially in the diffIRT-R model and the condition with 24 items. Standard errors are high. Corresponding reliability coefficients range from 0.19 to 0.46 in the test with 24 items and from 0.21 to 0.59 in the test with

https://doi.org/10.1017/psy.2026.10112 Published online by Cambridge University Press

Table 5. Rate of false alarms (FA), detection rate (DR), and correlation (R) between the true number of premature responses and the predicted number when responses are flagged as premature with the extended diffIRT models and the estimated trait levels depending on the item number J

	diffIRT-I		diffIRT-R	
	$J = 24$	$J = 48$	$J = 24$	$J = 48$
FA	0.11	0.08	0.07	0.06
DR	0.58	0.58	0.66	0.67
R	0.73	0.86	0.83	0.91

Note: Results are based on 36,450 data patterns and are averaged over the items. diffIRT-R: Diffusion model with censoring and random guessing; diffIRT-I: Diffusion model with censoring and premature responding based on the sign rule.

48 items. The coverage frequencies of the confidence intervals are below the intended level of 0.95. The reliability of the estimates depends on the level of response caution. Supplementary Figures S3.3 and S3.6 suggest that high levels of the willingness to invest time cannot be estimated well when the response caution is low. In this case, the diffusion process wins in most cases.

In an additional analysis, we compared the trait estimates from the extended diffIRT models to those from a misspecified diffIRT model. For this purpose, we determined the diffIRT model that best approximated the extended versions in the Kullback–Leibler sense (White, 1982). We then used this misspecified diffIRT model in order to estimate the capability and the response caution with the generated data. The bias of the trait estimates is visualized in Figure 2 as a function of the rate of premature responding. We also report the bias of the trait estimates when the data-generating model (diffIRT-R/diffIRT-I) is used for trait estimation.

When data are generated according to the diffIRT-I model, the estimates of capability are virtually unbiased irrespective of whether the diffIRT-I or the misspecified diffIRT model is used as the measurement model. This is even the case when almost all responses are generated prematurely. The estimates of the response caution, however, are affected by the misspecification of the measurement model. When data are generated according to the diffIRT-R model, the estimates of capability are biased when the misspecified diffIRT model is used as the measurement model. The estimates provided by the diffIRT-R model remain unbiased, even when almost all responses are premature. The estimates of response caution, however, have a bias in both models when the rate of premature responses exceeds 50%. The bias, however, decreases in the diffIRT-R model when more items are used for trait estimation.

Finally, we investigated whether premature responses can be detected. We used the data-generating model (diffIRT-I/diffIRT-R) and the trait estimates in order to determine the posterior probability that a response was premature. Responses were flagged as premature when the posterior probability exceeded 0.5. Table 5 contains the detection rate, the rate of false alarms, and the correlation between the true number of premature responses and the predicted number of premature responses. A plot of the true number of premature responses against the expected number is given in the Supplementary Material (see Supplement 2 and Supplementary Figure S3.9).

The models are relatively successful in detecting premature responses. The number of premature responses can be predicted well, especially with the diffIRT-R model. The models are also capable to detect more than half of the premature responses. The rate of false alarms is relatively low.

https://doi.org/10.1017/psy.2026.10112 Published online by Cambridge University Press

3. Empirical application

We analyzed two data sets with the extended diffIRT models. The data sets consisted of the responses and response times of test takers on different cognitive tests. The tests varied in their cognitive demands and response formats. We selected the data sets to explore the conditions under which the model can used. This question is particularly relevant for process models that rely on strong assumptions about the response process and may not be suitable in all contexts (Krämer et al., 2025). We used the following data sets.

3.1. SPT

The SPT data set comprised responses and response times of 773 university students on 16 items of a test on visuo-spatial perspective taking (Kessler & Thomson, 2010). Each item depicted a target person seated at a round table from a bird's-eye perspective. The person's position at the table was rotated between 120 °, 160 °, 200 °, or 240 ° relative to the test taker's point of view. Test takers were asked to judge whether a specific object on the table was closer to the person's left or right arm. Data were collected by Wolgast et al., (2020) in a low-stakes setting and are available on OSF (https://osf.io/nmxq7/). Before analyzing the data, we removed data from test takers with irregular long response times. Response times were defined as irregular long when they exceeding the median time by 2.5 times the inter quartile range. Such long response times likely reflect interruptions of the solution process that fall outside the scope of the model. Data cleaning reduced the data set to 744 subjects. In the cleaned data set, the solution frequencies ranged from 0.92 to 0.94 and the average response times varied between 1.64 to 2.02 seconds. We selected the data set as we considered it well-suited for an analysis with the diffusion model: The response process involves basic mental operations and a binary decision. The applicability of the diffusion model to data from mental rotation tests has been demonstrated in prior research (Molenaar et al., 2015; van der Maas et al., 2011).

3.2. PERC

The PERC data set comprised the responses and response times of 2,978 high school students on 15 Raven matrix tasks measuring abstract reasoning (Raven et al., 1998). Each item consisted of a matrix of figures which was generated by certain rules. One element of the matrix was missing. Test takers had to choose from a set of figures the one that complemented the matrix consistently. The items were grouped into five blocks. The first block consisted of easy tasks. Its purpose was to create an experience of success. The second block consisted of medium difficult tasks. After the tasks, the test takers received feedback and could access tips on how to improve their performance. Usage of the tips is interpreted as a measure of effort. The third block consisted of very difficult tasks. The time on task in the third block is interpreted as a measure of persistence. The final block consisted of easy tasks that were matched in difficulty to the tasks in the first set. Differences in performance between the first block and the last block are interpreted as a measure of resilience. Data were collected by Porter et al. (2020) in a low-stakes setting and are publicly available on OSF (https://osf.io/8xcjp/). For the analysis, we only considered the 1,680 cases with complete data that had previously been analyzed by Domingue et al. (2022). We removed test takers with unusually long response times as in the SPT data set. This reduced the data set to 1,560 subjects. In the reduced data set, the solution frequencies ranged from 0.10 to 0.97 and the average response times varied between 8.48 and 57.08 seconds. We included the data set as we wanted to assess whether the model can be used when tests involve more complex mental operations. Furthermore, we expected censoring in the difficult items.

We fitted several models to both data sets. We fitted a standard diffIRT model with ($d_j \neq 0$) and without ($d_j = 0$) a non-decision time. We fitted both versions of the extended diffIRT model, also either with ($d_j \neq 0$) and without ($d_j = 0$) a non-decision time. In the diffIRT-R model with random guessing, the guessing probabilities across the items were constrained to be equal. This restriction simplifies parameter estimation and is justified in case guessing is indeed random. For the sake of

https://doi.org/10.1017/psy.2026.10112 Published online by Cambridge University Press

Table 6. Value of the marginal log-likelihood function (LL), number of parameters (NP), AIC-Index (AIC), difference of AIC-Index (ΔAIC), BIC-Index, and difference of BIC-Index (ΔBIC) for six versions of the DiffIRT model (diffIRT) and four versions of the hierarchical model (HLM) in the two data sets

Data	Model	Guessing	Shift	LL	NP	AIC	ΔAIC	BIC	ΔBIC
SPT	diffIRT	None	No	−15,745.52	65	31,621.03	1,623.90	31,918.87	1,321.49
			Yes	−15,369.40	81	30,900.79	903.66	31,271.94	674.56
		Sign rule	No	−15,507.77	115	31,245.54	1,248.41	31,772.47	1,175.10
			Yes	**−14,867.57**	131	**29,997.13**	0.00	**30,597.38**	0.00
		Random	No	−15,545.24	116	31,322.49	1,325.36	31,854.00	1,256.62
			Yes	−14,881.11	132	30,026.22	29.09	30,631.05	33.67
	HLM	None	No	−15,574.06	81	31,310.12	1,312.98	31,681.26	1,083.88
			Yes	−15,375.84	97	30,945.68	948.55	31,390.14	792.76
		Random	No	−15,574.07	82	31,312.14	1,315.00	31,687.86	1,090.48
			Yes	−15,375.85	98	30,947.71	950.57	31,396.75	799.37
PERC	diffIRT	None	No	−94,952.46	61	190,026.9	7,775.56	190,353.4	7,438.36
			Yes	−93,102.95	76	186,357.9	4,106.54	186,764.7	3,849.63
		Sign rule	No	−92,041.47	108	184,298.9	2,047.57	184,877.0	1,961.93
			Yes	−91,561.38	123	183,368.8	1,117.39	184,027.1	1,112.04
		Random	No	−91,312.50	109	182,843.0	591.64	183,426.4	511.35
			Yes	**−91,001.68**	124	**182,251.4**	0.00	**182,915.1**	0.00
	HLM	None	No	−92,817.46	76	185,786.9	3,535.55	186,193.7	3,278.63
			Yes	−92,583.75	91	185,349.5	3,098.13	185,836.6	2,921.50
		Random	No	−92,814.88	77	185,783.8	3,532.39	186,195.9	3,280.82
			Yes	−92,580.90	92	185,345.8	3,094.44	185,838.2	2,923.16

Note: Differences are with respect to the best fitting model with lowest AIC. SPT: Test on perspective taking; PERC: Test with Raven matrices. The model with the best model fit concerning AIC, BIC or LL is highlighted.

comparison, we also fitted the hierarchical model proposed by van der Linden (2007) to the data. The hierarchical model is one of the most popular models for responses and response times on tests and was included as a benchmark. It combines an item response theory (IRT) model for the responses and a log-normal factor model for the response times. We considered several versions of the hierarchical model, using either a two-parameter logistic (2-PL) or a three-parameter logistic (3-PL) model for the responses, and a log-normal factor model with or without a shift parameter for the response times. All models were fitted with MML estimation. We compared the relative fit of the models with the AIC index and the BIC index. Results concerning relative model fit are given in Table 6 where we report the marginal log-likelihood (LL), the number of model parameters (NP), the AIC index, and the BIC index, as well as the difference between the information criterion for each model and that for the best fitting model (ΔAIC and ΔBIC). Note that models with lower AIC and BIC have a better fit.

In the following, we will briefly discuss the major findings separately for the two data sets. The discussion is based on the implied censoring rates (P_c), the solution probabilities of premature ($P_{x|c}$) and regular ($P_{x|u}$) responses, and the corresponding expected response times ($M_{t|c}$, $M_{t|u}$) on the items of the test. These quantities are reported in Table 7 for the model with the lowest AIC index.

https://doi.org/10.1017/psy.2026.10112 Published online by Cambridge University Press

Table 7. Solution frequency P_x, average response time M_t, implied censoring rate P_c, implied solution frequency in regular and premature responses ($P_{x|c}$, $P_{x|u}$), as well as expected response time in regular and premature responses ($M_{t|c}$, $M_{t|u}$) in the items (I) of the two data sets

SPT								PERC							
I	P_x	M_t	P_c	$P_{x\|c}$	$P_{x\|u}$	$M_{t\|c}$	$M_{t\|u}$	I	P_x	M_t	P_c	$P_{x\|c}$	$P_{x\|u}$	$M_{t\|c}$	$M_{t\|u}$
1	0.92	2.01	0.02	0.53	0.91	1.32	2.08	1	0.96	14.66	0.04	0.07	0.99	22.21	14.64
2	0.92	1.83	0.03	0.58	0.92	1.44	1.91	2	0.81	19.85	0.15	0.10	0.93	29.69	17.44
3	0.94	1.67	0.02	0.55	0.93	1.12	1.71	3	0.96	8.48	0.04	0.10	1.00	12.24	8.46
4	0.93	2.02	0.03	0.61	0.91	1.71	2.08	4	0.86	12.25	0.13	0.10	0.96	15.03	11.73
5	0.93	1.99	0.02	0.55	0.91	1.53	2.03	5	0.72	28.07	0.29	0.09	0.97	25.79	28.88
6	0.93	1.87	0.01	0.55	0.92	0.99	1.93	6	0.17	57.08	0.76	0.11	0.38	50.71	80.46
7	0.93	1.75	0.03	0.66	0.93	1.55	1.79	7	0.46	40.77	0.52	0.11	0.86	35.41	47.02
8	0.93	1.90	0.05	0.60	0.93	2.07	1.99	8	0.95	20.79	0.06	0.10	1.00	19.36	21.03
9	0.94	1.78	0.05	0.60	0.94	1.73	1.83	9	0.24	46.70	0.79	0.10	0.73	42.22	59.47
10	0.94	1.84	0.01	0.52	0.93	0.76	1.93	10	0.48	33.33	0.47	0.09	0.83	28.52	36.66
11	0.93	1.84	0.05	0.58	0.92	2.02	1.91	11	0.75	22.15	0.22	0.10	0.94	17.87	23.17
12	0.93	1.83	0.05	0.58	0.93	1.76	1.86	12	0.10	42.88	0.81	0.10	0.22	35.53	71.66
13	0.93	1.64	0.05	0.53	0.93	1.59	1.71	13	0.97	10.30	0.03	0.09	0.98	6.27	10.25
14	0.92	1.99	0.04	0.56	0.91	1.99	2.00	14	0.97	10.28	0.04	0.14	1.00	8.00	10.39
15	0.92	1.70	0.04	0.57	0.93	1.65	1.75	15	0.90	9.59	0.10	0.11	0.99	8.35	9.84
16	0.94	1.98	0.04	0.57	0.93	1.75	2.04								

Note: Implied statistics are determined with the best fitting model and the MML estimates. SPT: Test on perspective taking; PERC: Test with Raven matrices.

3.3. SPT

The model with the best relative model fit was the diffIRT-I model. This model also had an acceptable absolute model fit in terms of the conditional accuracy functions that could be recovered well apart from some deviations in long response times (see Supplementary Figure S3.7). These deviations, however, should not be overinterpreted due to the small number of incorrect responses. The estimates of the model parameters are reported in Supplementary Table S2.2 along with the standard errors of estimation. According to the results in Table 7, between 2% and 5% of the responses were given prematurely on basis of partial information. Responding prematurely, however, was not the only cause for errors despite the easiness of the items. Even in case the response process was regular, the solution probability $P_{x|u}$ was below 100%. This conforms to the observation that some test takers had a test score of zero. A part of the wrong responses thus seems to be due to a misconception. The solution probability under censoring $P_{x|c}$ was slightly above chance level. This suggests that some test takers responded prematurely without accumulating much information, although some intuitive knowledge concerning the correct response option seemed to be used.

3.4. PERC

The model with the best relative model fit was the diffIRT-R model. This model also had a good absolute model fit as the empirical and implied conditional accuracy function agreed closely (see Supplementary Figure S3.8). The estimates of the model parameters are reported in Supplementary Table S2.3. The censoring rates P_c varied over the items and were highest in the difficult items 6, 9, and 12. In these

https://doi.org/10.1017/psy.2026.10112 Published online by Cambridge University Press

items, the boundary separations were large and the drift rates close to zero (Supplementary Table S2.3). As for the item parameters of the censoring model, the values in the difficult items were comparable to those in the remaining items. This suggests that the low probability of solving the items was mostly due to the high censoring rates caused by difficulties in accumulating information. This conjecture is supported by the observation that premature responses were correct with a probability near the chance level of 1/8. In fact, this is to be expected for items that are supposed to be nearly unsolvable. Note that in a standard diffusion model, low solution probabilities only occur when drift rates are negative. Negative drift rates, however, mean that test takers accumulate information for a wrong response, which is implausible in items that do not have a solution at all. A second observation is that there was no general trend concerning the intercept parameters of the log-normal distribution generating the censoring time. This implies that test takers were resilient on general and did not reduce their willingness to invest time systematically.

4. Discussion

The performance in a test depends on the capability of a test taker as well as his test-taking engagement. These determinants of performance are partially confounded in measurement. In standard latent trait models, the latent trait represents the propensity to solve the items of a test and thus mingles capability and the test takers' willingness to apply it.

In this article, we propose a model that relates the propensity to solve items to more basic determinants of the response process. The proposed model is an extension of the diffIRT model that has recently been proposed for responses and response times on tests (Tuerlinckx et al., 2016). The extension consists in the addition of a censoring process that stops information processing at a censoring time. The model combines an information accumulation model (as represented by the diffusion model) with an urgency mechanism (as represented by the log-normal factor model) to a race model. In the model, each test taker is characterized in terms of three traits, his information processing capability, response caution, and willingness to invest time. Capability determines the maximal performance of a test taker under infinite processing time, response caution determines the desired information level when responding, and willingness to invest time determines the maximal amount of time a test taker would spend on an item. Due to this distinction of cognitive and motivational determinants of test-taking, the extended diffIRT model provides a better characterization of a test taker's capability. It also informs about aspects of test-taking engagement that might be of diagnostic relevance for educational and professional outcomes (Duckworth & Yeager, 2015; Soland et al., 2019).

The extended diffIRT model improves over previous models for rapid guessing in several ways. It improves over the standard diffusion IRT model that mingles persistence and response caution as the simulation study suggests. It improves over mixture models that are very limited in their capability to represent individual differences in persistence. In the mixture models, the test takers only differ in the response mode. Consequently, the only source of individual differences is the frequency by which the response modes occur. Test-taking engagement, however, is not an all-or-nothing decision, but rather manifests itself in varying degrees (e.g., Knekta, 2017; Wigfield & Eccles, 2000). The extended diffIRT model explicitly accounts for this by including the willingness to invest time. The extended diffIRT model also improves over previous models with censoring (Hawkins & Heathcote, 2021; Lee & Ying, 2015; Lu & Wang, 2020; Ranger & Kuhn, 2014). In the models for censoring, the censored responses are either wrong, random guesses, or omissions and thus do not contain information about the test takers per se. In the proposed model, the premature responses are informed by the information accumulation process. The premature responses therefore reflect the information processing capability of a test taker just as the regular responses. In this aspect, the proposed model is similar to models that relate the time on the task to the solution probability (e.g., Goldhammer et al., 2015; Wang & Hanson, 2005), although in our model, it is not the time that is crucial, but the information that is accumulated during the time. The model is also different to the model of Ranger et al. (2025), although both models

https://doi.org/10.1017/psy.2026.10112 Published online by Cambridge University Press

assume information accumulation and censoring. First, the models differ in the interpretation of the accumulator that either represents progress toward the solution or the relative preference for one of the two response options (Webb, 2019). This difference has implications for premature responding and its dependency on the partial knowledge. Second, the extended diffIRT models assume noisy instead of linear ballistic information accumulation which introduces a further random component to the regular and premature response process. This could also be used in order to model non-response (when the relative evidence is too weak to respond) or response confidence (which might depend on the amount of relative evidence). Finally, the extended diffIRT models are more closely related to theories of test-taking motivation as test takers are characterized in terms of persistence and their aspiration level, not just in terms of persistence as in the model of Ranger et al. (2025). For this reason, only the extended diffIRT model permits a detailed analysis of test-taking engagement.

We conducted simulation studies on parameter recovery and the recovery of the latent traits. The simulation study on parameter recovery demonstrated that the item parameters of the model can be estimated well when samples consist of 1,000 subjects and more. In this case, the estimates are virtually unbiased and the coverage frequencies of confidence intervals are close to the intended level. The simulation study on trait recovery revealed limitations of the models as measurement models for the traits. Capability, which is usually the target trait of the assessment, can be estimated well even when the rate of premature responding is high. Response caution and the willingness to invest time, however, can be estimated less well. Trait recovery is poor for low levels of the willingness to invest time in case test takers have a high level of response caution and vice versa. This limits the usefulness of the model for individual assessment in its present form. Trait estimation might be improved by an adaptive test where the time demand of the items is adapted to the willingness to invest time. Note that adaptive tests are also necessary in IRT when test are too easy or too difficult. A disadvantage of adaptive tests is that motivation cannot be estimated as a simple byproduct on the fly. The precision of the estimates might alternatively be increased by including the confidence in the response (Reynolds et al., 2020). However, not all applications of latent trait models require precise estimates of the traits on an individual level. Despite the poor recovery of the traits, the model can still be used in order to flag responses as premature. Furthermore, in standardized assessments of educational attainment (e.g., PISA), the focus is on the comparison of the educational attainment of groups and its relationship to further variables. The measurement imprecision of traits can be accounted for with structural equation models. The extended diffIRT models are also useful research tools whenever motivational aspects of test-taking have to be controlled for or are among the target constructs of the investigation. The model can also be used in order to study effects of time limits on test-taking.

We applied the model to two empirical data sets. The data sets contained data from tests varying in task complexity and response format. The aim of the study was to explore the scope of the models. The study suggests that the diffIRT-I model is suitable for tests with a binary response format and rather simple tasks. This is probably due to the fact that in such tasks, the response process comes close to an information accumulation process that weights relative evidence for the two response options. Our assumed guessing mechanism depends crucially on this assumption. When tests involve more complex mental operations or require a decision between more response options, the diffIRT-R model fits better. Remarkably, the hierarchical model of van der Linden (2007) that makes little assumptions about the response process and is sometimes considered as a standard model for responses and response times on tests had the worst relative model fit. In the empirical application, we also demonstrated that the extended diffIRT models provide additional insight into the response process of the test takers.

One of the motivations behind this article was to specify the different aspects of test-taking engagement in mathematical terms. In our model, we consider two aspects of test-taking engagement, persistence and response caution. There are, however, more aspects of test-taking engagement. Instead of assuming a censoring mechanism, one might model test-taking engagement by collapsing boundaries that become narrower over time (Voskuilen et al., 2016). This would correspond to a lowering of the

https://doi.org/10.1017/psy.2026.10112 Published online by Cambridge University Press

standards and might be due to a growing desire to finish. Diffusion models with collapsing boundaries, however, are difficult to implement. We also did not address the concentration of a test taker. By concentration, we mean the focus of a test taker on the task. A high level of concentration prevents that information processing is interrupted by task-irrelevant cognition such as mind wandering (Pieters & van der Ven, 1982). Concentration is hardly addressed by psychometric models. In models with a non-decision time component, the non-decision time might absorb time components that are not related to information processing. The non-decision time, however, cannot be interpreted as concentration unambiguously. The only model that explicitly takes task-irrelevant cognitions into account is the model proposed by Boehm et al. (2021). How concentration can be included into the extended diffIRT models is a topic of future research.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/psy.2026.10112.

Acknowledgements. We thank Moritz Beleites for his technical support in implementing the simulation study.

Data availability statement. The SPT data are available on OSF (https://osf.io/nmxq7/). The PERC data are available on OSF (https://osf.io/8xcjp/).

Funding statement. The study was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under project number 28872689.

Competing interests. The authors declare none.

References

Akaike, H. (1992). Information theory and an extension of the maximum likelihood principle. In S. Kotz, & N. Johnson (Eds.), *Breakthroughs in statistics, vol I, foundations and basic theory* (pp. 610–624). Springer. https://doi.org/10.1007/978-1-4612-0919-5_38

Alós-Ferrer, C. (2018). A dual-process diffusion model. *Journal of Behavioral Decision Making*, *31*, 203–218. https://doi.org/10.1002/bdm.1960

Baldassi, C., Cerreia-Vioglio, S., Maccheroni, F., Marinacci, M., & Pirazzini, M. (2020). A behavioral characterization of the drift diffusion model and its multialternative extension for choice under time pressure. *Management Science*, *66*, 5075–5093. https://10.1287/mnsc.2019.3475

Baumert, J., & Demmerich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education*, *16*, 441–462. https://doi.org/10.1007/BF03173192

Berger, J., Liseo, B., & Wolpert, R. (1999). Integrated likelihood methods for eliminating nuisance parameters. *Statistical Science*, *14*, 1–28. https://doi.org/10.1214/ss/1009211804

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review*, *59*, 65–98. https://doi.org/10.1137/141000671

Boehm, U., Marsman, M., van der Maas, H., & Maris, G. (2021). An attention-based diffusion model for psychometric analyses. *Psychometrika*, *86*, 938–972. https://doi.org/10.1007/s11336-021-09783-0

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, *113*, 700–765. https://doi.org/10.1037/0033-295X.113.4.700

Bolt, D., Cohen, A., & Wollack, J. (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. https://doi.org/10.1111/j.1745-3984.2002.tb01146.x

Boughton, K., & Yamamoto, K. (2007). A HYBRID model for test speededness. In M. von Davier, & C. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 147–156). Springer. https://doi.org/10.1007/978-0-387-49839-3_9

Buchholz, J., Cignetti, M., & Piacentini, M. (2006). Developing measures of engagement in PISA. (OECD Education Working Papers No. 279). OECD Publishing. https://doi.org/10.1787/2d9a73ca-en.

Burnham, K., & Anderson, D. (2013). *Model selection and inference—A practical information-theoretic approach*. Springer. https://doi.org/10.1007/978-1-4757-2917-7.

Burton, R. (2002). Misinformation, partial knowledge and guessing in true/false tests. *Medical Education*, *36*, 805–811. https://doi.org/10.1046/j.1365-2923.2002.01299.x

Cizek, G., & Wollack, J. (2017). Exploring cheating on tests: The context, the concern, and the challenges. In G. Cizek, & J. Wollack (Eds.), *Handbook of quantitative methods for detecting cheating on tests* (pp. 3–19). Routledge. https://doi.org/10.4324/9781315743097

Crow, E., & Shimizu, K. (1988). *Lognormal distributions. Theory and applications*. Marcel Dekker. https://doi.org/10.1201/9780203748664

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. https://doi.org/10.3102/1076998614558485

Domingue, B., Kanopka, K., Stenhaug, B., Sulik, M., 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*, 576–602. https://doi.org/10.3102/10769986221099906

Drugowitsch, J. (2017). *DiffModels.jl: Diffusion model simulation and first-passage time densities in Julia*. https://github.com/DrugowitschLab/DiffModels.jl

Duckworth, A., & Yeager, D. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher*, *44*, 237–251. https://doi.org/10.3102/0013189X15584327

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

Fehr, E., & Rangel, A. (2011). Neuroeconomic foundations of economic choice—Recent advances. *Journal of Economic Perspectives*, *25*, 3–30. https://doi.org/10.1257/jep.25.4.3

Gaviria, J. (2005). Increase in precision when estimating parameters in computer assisted testing using response time. *Quality & Quantity*, *39*, 45–69. https://doi.org/10.1007/s11135-004-0437-y

Glickman, M., Gray, J., & Morales, C. (2005). Combining speed and accuracy to assess error-free cognitive processes. *Psychometrika*, *70*, 405–425. https://doi.org/10.1007/s11336-002-0999-3

Goldhammer, F., Naumann, J., & Greiff, S. (2015). More is not always better: The relation between item response and item response time in Raven's matrices. *Journal of Intelligence*, *3*, 21–40. https://doi.org/10.3390/jintelligence3010021

Gonthier, C. (2023). Should intelligence tests be speeded or unspeeded? A brief review of the effects of time pressure on response processes and an experimental study with Raven's matrices. *Journal of Intelligence*, *11*, 120. https://doi.org/10.3390/jintelligence11060120

Guidoum, A., & Boukhetala, K. (2020). Performing parallel Monte Carlo and moment equations methods for Ito and Stratonovich stochastic differential systems: R package Sim.DiffProc. *Journal of Statistical Software*, *96*, 1–82. https://doi.org/10.18637/jss.v096.i02

Guo, J., Xu, X., Fang, G., Ying, Z., & Zhang, S. (2025). Jointly modeling responses and omitted items by a competing risk model: A survival analysis approach. *British Journal of Mathematical and Statistical Psychology.*, Online First. https://doi.org/10.1111/bmsp.12382.

Hawkins, G., & Heathcote, A. (2021). Racing against the clock: Evidence-based versus time-based decisions. *Psychological Review*, *128*, 222–263. https://doi.org/10.1037/rev0000259

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*, 725–748. https://doi.org/10.1007/s11336-021-09819-5

Kang, I., Molenaar, D., & Ratcliff, A. (2023). A modeling framework to examine psychological processes underlying ordinal responses and response times of psychometric data. *Psychometrika*, *88*, 940–974. https://doi.org/10.1007/s11336-023-09902-z

Kessler, K., & Thomson, L. (2010). The embodied nature of spatial perspective taking: Embodied transformation versus sensorimotor interference. *Cognition*, *114*, 72–88. https://doi.org/10.1016/j.cognition.2009.08.015

Kim, S., Bao, Y., Horan, E., Kim, M., & Cohen, A. (2015). Gauss-Hermite quadrature in marginal maximum likelihood estimation of item parameters. In L. van der Ark, D. Bolt, W. Wang, J. Douglas, & S. Chow (Eds.), *Quantitative psychology research—Springer proceedings in mathematics & statistics* (Vol. *140*, pp. 43–58). Springer. https://doi.org/10.1007/978-3-319-19977-1_4

E. Knekta (2017). *Motivational aspects of test-taking*. Umea University. http://umu.diva-portal.org/

Krämer, G., Ranger, J., Koch, M., Spinath, F., & Schmitz, F. (2025). Applicability of process models for the joint analysis of responses and response times in complex cognitive tasks. *Psychological Test and Assessment Modeling*, *67*, 90–126. https://doi.org/10.2440/001-0020

Kvam, P. (2019). A geometric framework for modeling dynamic decisions among arbitrarily many alternatives. *Journal of Mathematical Psychology*, *91*, 14–37. https://doi.org/10.1016/j.jmp.2019.03.001

Lau, P., Lau, S., Hong, K., & Usop, H. (2011). Guessing, partial knowledge, and misconceptions in multiple-choice tests. *Journal of Educational Technology & Society*, *14*, 99–110. http://www.jstor.org/stable/jeductechsoci.14.4.99

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

Liu, Y., Cheng, Y., & Liu, H. (2020). Identifying effortful individuals with mixture modeling response accuracy and response time simultaneously to improve item parameter estimation. *Educational and Psychological Measurement*, *80*, 775–807. https://doi.org/10.1177/0013164419895068

https://doi.org/10.1017/psy.2026.10112 Published online by Cambridge University Press

Lu, J., & Wang, C. (2020). A response time process model for not-reached and omitted items. *Journal of Educational Measurement*, *57*, 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*, 261–288. https://doi.org/10.1177/00131644211045

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

Michaelides, M., Ivanova, M., & Nicolaou, C. (2020). The relationship between response-time effort and accuracy in PISA science multiple choice items. *International Journal of Testing*, *20*, 187–205. https://doi.org/10.1080/15305058.2019.1706529

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

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

Nocedal, J., & Wright, S. J. (2006). *Numerical optimization*. Springer. https://doi.org/10.1007/978-0-387-40065-5

Noventa, S., Heller, J., & Kelava, A. (2024). Toward a unified perspective on assessment models, part I: Foundations of a framework. *Journal of Mathematical Psychology*, *122*, 102872. https://doi.org/10.1016/j.jmp.2024.102872

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

Pieters, J., & van der Ven, A. (1982). Precision, speed, and distraction in time-limit tests. *Applied Psychological Measurement*, *6*, 93–109. https://doi.org/10.1177/014662168200600110

Porter, T., Catalán Molina, D., Blackwell, L., Roberts, S., Quirk, A., Lee Duckworth, A., & Trzesniewski, K. (2020). Measuring mastery behaviors at scale: The persistence, effort, resilience and challenge-seeking task (PERC). *Journal of Learning analytics*, *7*, 5–18. https://doi.org/10.18608/jla.2020.71.2

R Core Team. (2021). R: A language and environment for statistical computing [computer software manual]. Vienna, Austria. http://www.R-project.org

Ranger, J., & Kuhn, J. (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*, 388–407. https://doi.org/10.1111/bmsp.12025

Ranger, J., Much, S., Neek, N., Mutak, A., & Pohl, S. (2025). Accounting for persistence in tests with linear ballistic accumulator models. *Psychometrika*, *90*, 1111–1135. https://doi.org/10.1017/psy.2025.10026

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

Ratcliff, R. (1980). A note on modeling accumulation of information when the rate of accumulation changes over time. *Journal of Mathematical Psychology*, *21*, 178–184. https://doi.org/10.1016/0022-2496(80)90006-1

Raven, J., Raven, J., & Court, J. (1998). *Manual for Raven's progressive matrices and vocabulary scales*. Oxford Psychologists Press.

Reynolds, A., Kvam, P., Osth, F., & Heathcote, A. (2020). Correlated racing evidence accumulator models. *Journal of Mathematical Psychology*, *96*, 102331. https://doi.org/10.1016/j.jmp.2020.102331

Rios, J. (2022). Assessing the accuracy of parameter estimates in the presence of rapid guessing misclassifications. *Educational and Psychological Measurement*, *82*, 122–150. https://doi.org/10.1177/00131644211003640

Schnipke, D., & Scrams, D. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, *34*, 213–232. https://doi.org/10.1111/j.1745-3984.1997.tb00516.x

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464. https://www.jstor.org/stable/2958889

Soland, J., Zamarro, G., Cheng, A., & Hitt, C. (2019). Identifying naturally occurring direct assessments of social-emotional competencies: The promise and limitations of survey and assessment disengagement metadata. *Educational Researcher*, *48*, 466–478. https://doi.org/10.3102/0013189X19861356

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

Tillman, G., Van Zandt, T., & Logan, G. (2020). Sequential sampling models without random between-trial variability: The racing diffusion model of speeded decision making. *Psychonomic Bulletin & Review*, *27*, 911–936. https://doi.org/10.3758/s13423-020-01719-6

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

Tuerlinckx, F., Molenaar, D., & van der Maas, H. (2016). Diffusion-based item response theory modeling. In W. van der Linden, & R. Hambleton (Eds.), *Handbook of modern item response theory*. Chapman and Hall/CRC Press. https://doi.org/10.1201/b19166-1

Ulitzsch, E., von Davier, M., & Pohl, S. (2019). 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*, 83–112. https://doi.org/10.1111/bmsp.12188

van der Linden, W. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*, 181–204. https://doi.org/10.3102/10769986031002181

https://doi.org/10.1017/psy.2026.10112 Published online by Cambridge University Press

van der Linden, W. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*, 287–308. https://doi.org/10.1007/s11336-006-1478-z

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

van Laar, S., & Braeken, J. (2022). Random responders in the TIMSS 2015 student questionnaire: A threat to validity? *Journal of Educational Measurement*, *59*, 470–501. https://doi.org/10.1111/jedm.12317

Voskuilen, C., Ratcliff, R., & Smith, P. (2016). Comparing fixed and collapsing boundary versions of the diffusion model. *Journal of Mathematical Psychology*, *73*, 59–79. https://doi.org/10.1016/j.jmp.2016.04.008

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

Wang, T., & Hanson, B. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement*, *29*, 323–339. https://doi.org/10.1177/0146621605275984

Webb, R. (2019). The (neural) dynamics of stochastic choice. *Management Science*, *65*, 230–255. https://doi.org/10.1287/mnsc.2017.2931

Weber, D., Krieger, F., Spinath, F. M., Greiff, S., Hissbach, J., & Becker, N. (2023). A log file analysis on the validity of partial solutions in figural matrices tests. *Journal of Intelligence*, *11*, 37. https://doi.org/10.3390/jintelligence11020037

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*, 1–25. https://doi.org/10.2307/1912526

Wigfield, A., & Eccles, J. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology*, *25*, 68–81. https://doi.org/10.1006/ceps.1999.1015

Wise, S. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement*, *36*, 52–61. https://doi.org/10.1111/emip.12165

Wise, S., & DeMars, C. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, *43*, 19–38. https://doi.org/10.1111/j.1745-3984.2006.00002.x

Wise, S., & Smith, L. (2011). A model of examinee test-taking effort. In J. Bovaird, K. Geisinger, & C. Buckendal (Eds.), *High-stakes testing in education: Science and practice in K-12 settings* (pp. 139–153). American Psychological Association. https://doi.org/10.1037/12330-009

Wolgast, A., Schmidt, N., & Ranger, J. (2020). Test-taking motivation in education students: Task battery order affected within-test-taker effort and importance. *Frontiers in Psychology*, *11*, 559683. https://doi.org/10.3389/fpsyg.2020.559683

Cite this article: Ranger, J., Much, S., Neek, N., Wolgast, A., Mutak, A., Gaviria Soto, J. L. and Pohl, S. (2026). Extending the Diffusion Model by a Process Accounting for the Persistence of Test Takers. *Psychometrika*, 1232–1254. https://doi.org/10.1017/psy.2026.10112

https://doi.org/10.1017/psy.2026.10112 Published online by Cambridge University Press