PSYCHOMETRIKA
https://doi.org/10.1007/s11336-023-09902-z

A MODELING FRAMEWORK TO EXAMINE PSYCHOLOGICAL PROCESSES UNDERLYING ORDINAL RESPONSES AND RESPONSE TIMES OF PSYCHOMETRIC DATA

INHAN KANG

YONSEI UNIVERSITY

DYLAN MOLENAAR

UNIVERSITY OF AMSTERDAM

ROGER RATCLIFF

THE OHIO STATE UNIVERSITY

This article presents a joint modeling framework of ordinal responses and response times (RTs) for the measurement of latent traits. We integrate cognitive theories of decision-making and confidence judgments with psychometric theories to model individual-level measurement processes. The model development starts with the sequential sampling framework which assumes that when an item is presented, a respondent accumulates noisy evidence over time to respond to the item. Several cognitive and psychometric theories are reviewed and integrated, leading us to three psychometric process models with different representations of the cognitive processes underlying the measurement. We provide simulation studies that examine parameter recovery and show the relationships between latent variables and data distributions. We further test the proposed models with empirical data measuring three traits related to motivation. The results show that all three models provide reasonably good descriptions of observed response proportions and RT distributions. Also, different traits favor different process models, which implies that psychological measurement processes may have heterogeneous structures across traits. Our process of model building and examination illustrates how cognitive theories can be incorporated into psychometric model development to shed light on the measurement process, which has had little attention in traditional psychometric models.

Key words: response time, psychological process, measurement, psychometric process modeling, decision-making, confidence judgments.

With the advent of computerized measurement methods, it has become possible to collect response times (RTs) along with responses with relatively little additional effort. At the same time, joint modeling of responses and RTs has been gaining popularity in the field of psychometrics (see De Boeck & Jeon, 2019, for a review). A major benefit of the additional RT measures is an improvement in the measurement of latent abilities/traits in psychometrics (De Boeck & Jeon, 2019; Bolsinova & Tijmstra, 2018) . First, latent variables can be better estimated simply by the collateral information from RTs compared to when only item responses are available. Second, joint modeling of responses and RTs can help us account for the speed-accuracy trade-off (SAT; Wickelgren, 1977; Luce, 1986) by allowing the decomposition of ability and speed factors. Response-only models may produce the same ability estimates for respondents with the same data accuracy values but with different RT values. Incorporating RTs in modeling can disentangle the confounding effect of the latent ability and speed factors and produce more accurate ability

Correspondence should be made to Inhan Kang, Yonsei University, 403 Widang Hall, 50 Yonsei-ro, Seodaemun-gu, Seoul 03722, Republic of Korea. Email: qpsy@yonsei.ac.kr

Published online: 12 May 2023

estimates. Furthermore, modeling RTs can also help detect anomalous responses such as fast guessing and cheating (Ratcliff & Kang, 2021; Schnipke & Scrams, 1997; Wang & Xu, 2015; Wang et al., 2018) and identify latent response classes (e.g., fast vs slow; DiTrapani, Jeon, De Boeck, & Partchev, 2016; Molenaar, Oberski, Vermunt, & Boeck, 2016; Partchev & De Boeck, 2012).

Another advantage of joint modeling of responses and RTs that we particularly emphasize in this article is that it opens a path to a theory-based modeling approach (namely, *psychometric process modeling*). Earlier psychometric models of responses and RTs have been developed mainly based on extensions of traditional psychometric models (e.g., the hierarchical framework by van der Linden, 2007) such as factor analysis (FA) and item response theory (IRT) models. Although these models perform well in describing behavioral patterns of data, they do not shed light on the individual-level psychological processes underlying the measurement. In contrast, process modeling in perceptual and cognitive decision-making starts from a psychological theory of cognitive processes (e.g., sequential sampling framework, Ratcliff & Smith, 2004; Forstmann, Ratcliff, & Wagenmakers, 2016) and builds a mathematical model based on it. This approach provides a theoretical conceptualization of what latent variables refer to and how they generate responses and RTs. Psychometrics can also benefit from this modeling approach, and this can lead us to a theory-based study of the intra-individual processes of the measurement that explicitly describe latent variables as cognitive process components and their causal relationship with outcome variables. Also, variations in intra-individual processes provide primary sources of individual differences. In this regard, psychometric process modeling can furnish a new perspective on validity and measurement issues, in light of earlier discussions from Borsboom and colleagues (2003,2004).

To our knowledge, the first process model of responses and RTs in psychometrics is the diffusion IRT model (Molenaar et al., 2015b; Ranger et al., 2017; Tuerlinckx & De Boeck, 2005; Tuerlinckx et al., 2016; van der Maas et al., 2011). The model is based on the sequential sampling framework: when an item is presented, a respondent accumulates evidence for decision-making over time and eventually makes a response when sufficient information is accumulated. For binary responses (e.g., correct and incorrect responses), it is assumed that evidence accumulates toward one of the two decision boundaries, each of which corresponds to each of the binary response options. The accumulation process terminates when it hits a boundary and the corresponding response is predicted. RT is predicted as the sum of decision time and nondecision time where decision time refers to the time that the decision process (evidence accumulation) takes before the termination and nondecision time refers to the time for all the other cognitive processes not directly related to decision-making.

Two crucial components of the diffusion IRT model (and some other process models) are the mean rate of evidence accumulation (drift rate) and the amount of information required for decision-making (boundary separation). The diffusion IRT model assumes that these components can be decomposed into person and item parameters. This decomposition allows the model to account for the person and item effects confounded in measurement data. Furthermore, we interpret person parameters of cognitive components as process-based definitions of latent variables. Note that they span the same two-dimensional space as latent ability/trait and speed factors used in early psychometric models of responses and RTs. Importantly, with the definitions given by the cognitive components, it becomes clear what latent variables refer to (i.e., quality and quantity of information processing).

Also, the diffusion IRT model provides a process-based description of how within- and between-person variations are generated. The model (as most sequential sampling models do) assumes internal noise within a response process of a single person responding to a single item (i.e., evidence is noisy within the accumulation process). This internal process variability corresponds to intra-individual variation in response processes and it is one of the primary sources of the variations

of outcome variables, generating response and RT distributions for a single respondent and a single condition. Furthermore, individual differences in the mean rate of evidence accumulation and the necessary amount of information across respondents are other primary sources of variation, producing inter-individual differences in outcome variables. In this sense, the model provides us with an explanatory modeling approach in which we concretely define latent ability/trait and examine theoretical causality between psychological constructs and response outcomes, motivated by perceptual/cognitive decision-making theories.

However, there are only a few empirical applications of the diffusion IRT model, and process-based approaches have not yet been widely studied in the field. Also, the diffusion IRT model is only for binary responses and RTs. Latent abilities have been measured and modeled with binary item responses in the traditional IRT approach and so the diffusion IRT model can also be used to examine abilities. For latent traits (e.g., personality traits, attitude traits), it is more typical to use other types of responses such as ordinal responses (e.g., M-point Likert scale) rather than binary responses (but see Molenaar et al., 2015b; Kang, De Boeck, & Ratcliff, 2022b; Kang, De Boeck, & Partchev, 2022a; Tuerlinckx & De Boeck, 2005) and so the development of process models for other response types is required. For nonordinal multiple-choice items, van der Maas et al. (2011) proposed a version of the diffusion IRT model and Rouder et al. (2015) proposed the log-normal race model. However, these models are interested in modeling response accuracy and RTs and they are not applicable to ordinal response items to measure latent traits. For ordinal responses, Ranger and Kuhn (2018) proposed the first innovative psychometric process model. They successfully integrated the linear ballistic accumulator (LBA) model (Brown and Heathcote 2008) and the balance-of-evidence hypothesis (Vickers 1979) with person-wise and item-wise parameterization for psychometric data, which motivated our modeling approach presented in this article. However, unlike the diffusion model and many other sequential sampling models, the LBA model assumes no within-trial noise in the evidence accumulation and the fundamental source of probabilistic features of the model is from across-trial variability components (i.e., variability in model parameters across multiple trials in psychological experiments). Thus, given trial-wise parameters, this model predicts response and RT deterministically. In contrast, we argue that internal noise (corresponding to within-trial noise) is a fundamental source of noise in the measurement process and responses and RTs cannot be deterministically explained. We will discuss this difference further in Sect. 5.

In this article, we aim to build a framework of psychometric process modeling to study psychological processes underlying the measurement of latent personality/attitude traits that uses the Likert scale. To this end, we review early theories of confidence judgments in perceptual and cognitive decision-making fields and measurement with ordinal scales in psychometrics. Then, we integrate different theories and models to develop psychometric models that (1) decompose ordinal responses and RTs from personality/attitude measurement into person-wise cognitive components (i.e., 'cognitive' latent variables) and item parameters and (2) have their own theoretical representations of intra-individual processes of a respondent in a measurement procedure.

A model developed based on our framework should be capable of capturing important behavioral patterns in data response proportions and RT distributions. Thus, we test the model by contrasting data statistics and distributions against the corresponding model predictions, as done for mathematical models for perceptual and cognitive decision-making. A severe discrepancy between data and model prediction could signify that the model representation of measurement processes might be flawed. In this case, the model can be modified or rejected. Although a good absolute model fit cannot be a sufficient condition for a model representation to be the ground truth of measurement processes, it is certainly an important necessary condition. We do not claim that the best model out of our proposed models shows the ground truth of the cognitive processes underlying the personality/attitude measurement. However, our modeling approach can demon-

strate a theory-based way to model unobservable measurement processes through which latent variables generate outcome responses.

After discussing various early theories and models, we propose three psychometric process models. We start with a fundamental assumption that noisy evidence accumulation can provide an appropriate representation of the measurement processes, as in perceptual and cognitive decision-making models and the diffusion IRT model. Then, different cognitive theories of decision-making and confidence judgments and psychometric theories of measurement are introduced and integrated for the development of different models. As a result, each of our proposed models has a different representation and formulation of cognitive processes underlying the measurement of latent personality/attitude traits. We examine multiple models rather than just a single one (although a single model could be sufficient to describe our modeling approach) because there is no dominant theory on the intra-individual temporal dynamics of a respondent in a psychometric measurement scene. Also, even if one process model provides a good account of ordinal responses and RTs, there can be a model with a better account of data and a better theoretical explanation of cognitive processes of measurement. In this sense, comparing multiple models (and their relevant theories) is a strategic way to find a better representation of the psychological processes underlying personality/attitude measurement.

The article is organized as follows: In Sect. 1, we describe theories in cognitive psychology and psychometrics on which we base our modeling. This includes evidence accumulation, the latent response formulation, and some cognitive models of responses, RTs, and confidence judgments. In Sect. 2, we propose three psychometric process models that are based on cognitive and psychometric theories introduced in Sect. 1. The models differ in the number of evidence accumulators and how they represent cognitive processes underlying psychometric measurement. In Sect. 3, we conduct simulation studies to examine parameter recovery and cross recovery of the models. In Sect. 4, we fit the three models to empirical data to investigate which model provides the best illustration of behavioral patterns of the data. We will examine absolute model fits, and reject a model and its representation of the measurement processes if the model prediction shows a large discrepancy from the data pattern. Finally, we conclude this article with a theoretical discussion of our process models and potential future extensions. Throughout this article, we mainly consider ordinal responses with $M = 5$ response options (e.g., 5-point Likert scale). For the measurement of latent traits, the response options would be, for example, 1 = *strongly disagree*, 2 = *disagree*, 3 = *neutral*, 4 = *agree*, and 5 = *strongly agree* to a presented statement.

1. Cornerstone Theories

1.1. Single-boundary Wiener Process

A fundamental assumption in our model development is that the measurement processes of latent traits can be described by evidence accumulation. In the sequential sampling framework for perceptual and cognitive decision-making, a respondent accumulates evidence over time to make a decision when a task stimulus is presented. During the measurement processes of latent traits, for example, when responding to an item sentence in a personality questionnaire, respondents should find how much the presented sentence matches their personalities. This process requires the respondents to navigate through their individual experiences, growth history, or any other memories related to what is being asked. The respondents collect information from these, formulate their *perceived* personalities, and finally respond to the presented item based on it. If the perceived personality matches the item sentence, a response is likely to be positive, and if not, a negative response is more likely. Also, the degree of the match determines the response strength (e.g., agree vs strongly agree). This process is relatively simpler than the measurement processes

of latent abilities with complicated and difficult items (e.g., solving a calculus problem), and it can be adequately described by evidence accumulation as simple cognitive decision-making processes (see Sect. 5 for a further discussion of this assumption and cognitive modeling studies related to psychological processes of reading and inference).

For our modeling, we consider evidence accumulation with a single (upper) decision boundary (illustrated in, e.g., Panel B of Fig. 1 in Sect. 2.1). Evidence $E(t)$ accumulates over time t in the mean rate of v (i.e., drift rate), starting at $E(0) = 0$. The drift rate represents the quality of evidence and the efficiency of information processing. Evidence accumulated at each time point is noisy, and the noise is assumed to be normally distributed. Thus, the accumulation process can be expressed as $dE(t) = vdt + \sigma B(t)\sqrt{dt}$ where $dE(t)$ is the infinitesimal stochastic change in the accumulation process $E(t)$ during a small time interval dt, σ^2 is the diffusion coefficient that represents the variance of the noise within the accumulation process, and $B(t)$ is a Gaussian process with zero mean and unit variance (Cox & Miller, 1965; Smith, 2000) . The diffusion coefficient is a scaling factor and is typically fixed to some constant for identifiability (e.g., $\sigma^2 = 1$ as we do hereafter in this article).

The accumulation process terminates when the accumulated evidence reaches the decision boundary $\alpha > 0$. The boundary represents the (positive) amount or quantity of information required to make a decision. In perceptual and cognitive decision-making and psychometric testing, the decision boundary is related to the speed–accuracy trade-off (SAT) of a respondent in that a larger boundary is associated with more emphasis on response accuracy, while a smaller boundary is associated with more emphasis on speed. In the measurement of latent traits, a larger boundary would be associated with being more cautious in making a response to a presented item. During the response behavior, there are other cognitive processes that are not directly related to the decision-making, such as encoding item sentences and producing responses. The times taken for these nondecision processes are collectively modeled by the nondecision time parameter t_0. Then, RT is predicted as the sum of the decision time (the time taken by the accumulation process before termination) and the nondecision time.

The accumulation process we described is known as the single-boundary Wiener process, and it has been shown that its first passage time distribution (the distribution of the time at which the evidence accumulation process crosses the decision boundary, i.e., RT distribution) is the Wald or inverse Gaussian distribution (Cox & Miller, 1965; Luce, 1986; Ratcliff, 1978; Wald, 1947). With drift rate $v > 0$, decision boundary α, and nondecision time t_0, the first passage time density is given as:

$$f_T(t) = \frac{\alpha}{\sigma\sqrt{2\pi(t-t_0)^3}} \exp\left(-\frac{(\alpha - v(t-t_0))^2}{2\sigma^2(t-t_0)}\right) \tag{1}$$

It is worth noting again that a fundamental assumption in our modeling is that evidence accumulation of the cognitive processes underlying the measurement of latent traits includes within-process noise (i.e., evidence is noisy within the response process of a single person responding to a single item). The choice of the single-boundary Wiener process is also consistent with this assumption in that this process and its first passage time density in Eq. 1 include the diffusion coefficient σ^2. The noise in evidence accumulation is the primary source of the probabilistic characteristics of our process models, capturing within-person (and also within-item) variability and producing a response and RT distribution for each person-by-item pair. Throughout our model development, we integrate this evidence accumulation process with cognitive and psychometric theories for ordinal responses. In doing so, our models can jointly account for ordinal responses and RTs with different representations of the psychological processes of the measurement of latent traits.

1.2. Latent Response Formulation

A traditional way of modeling ordinal responses in psychometrics is to use a continuous latent response. This approach, called latent response formulation (Skrondal & Rabe-Hesketh, 2004), views an ordinal response as a thresholded realization of the underlying continuous response (illustrated in, e.g., Panel A of Fig. 1 in Sect. 2.1). The idea was first introduced by Pearson (1901) for dichotomous responses and later extended for ordinal responses with multiple response options (Bollen & Barb, 1981; Muthén, 1983, 1984; Olsson, 1979). Thurstone (1927a,b, 1928) also presented a similar idea to model one-dimensional discrimination of stimuli (e.g., signal vs noise) based on a normally distributed magnitude on a *psychological continuum*. This idea later contributed to the development of signal detection theory (Green & Swets, 1966; Macmillan & Creelman, 1966). and Thurstonian scaling (Bock & Jones, 1968; Torgenson, 1958)

Let y be an ordinal outcome variable and y^* be its underlying continuous latent response variable. For M response options, the outcome y is determined as follows:

$$y = \begin{cases} 1 & \text{if } y^* \leq \tau_1 \\ k & \text{if } \tau_{(k-1)} < y^* \leq \tau_k, \quad k = 2, \cdots, M-1 \\ M & \text{if } \tau_{(M-1)} < y^* \end{cases} \tag{2}$$

where τ_k ($k = 1, \cdots, M-1$) is a response threshold that maps continuous latent responses onto the ordinal scale. In factor analysis literature, y^* is assumed to follow a normal distribution as in a normal ogive model, which is equivalent to the graded response model from item response theory (Takane & De Leeuw, 1987). The measurement relation is described as $y^* = \nu + \lambda_1\xi_1 + \cdots + \lambda_L\xi_L + \epsilon$ where ν is an intercept, ϵ is a residual, Λ_l and ξ_l are factor loading and factor score corresponding to the lth factor ($l = 1, \ldots, L$), respectively, assuming L underlying factors.

1.3. Balance-of-Evidence Hypothesis

Along with response accuracy and RT, confidence judgments have been used to constrain mathematical models of perceptual and cognitive decision-making and to study the underlying psychological processes (Festinger, 1943a,b; Merkle & Van Zandt, 2006; Pleskac & Busemeyer, 2010; Ratcliff & Starns, 2009, 2013; Smith & Vickers, 1988; Van Zandt, 2000; Van Zandt & Maldonado-Molina, 2004; Vickers, 1979; Volkmann, 1934). Confidence is of particular importance for our modeling purpose because it is inherently an ordinal scale (from low confidence to high confidence). Thus, earlier process models of confidence judgments can be modified to fit ordinal responses for psychometric measurement of latent traits.

Confidence models built on the balance-of-evidence hypothesis (Merkle & Van Zandt, 2006; Smith & Vickers, 1988; Van Zandt & Maldonado-Molina, 2004; Vickers, 1979) have been applied to data from psychophysical discrimination tasks in which a subject is presented with a stimulus and is asked to compare it with an internal criterion value (e.g., greater or less in magnitude). These models conceive two competing evidence accumulators each of which represents one of the binary response options (e.g., one for 'greater' and the other for 'less'). The accumulators race toward a decision boundary, and the decision process terminates when one of the accumulators reaches the boundary (i.e., wins the race). Response is predicted as the response option corresponding to the winning accumulator, and RT is predicted as the sum of the termination time of the winning accumulator and nondecision time. The confidence level is predicted based on the difference in evidence between the two accumulators at the decision time. The amount of evidence from the winning accumulator is equal to the amount set by the decision boundary. When one response option is compelling, the evidence difference would be large and the model predicts a high

confidence level. Otherwise, the evidence difference would be small and the model predicts a low confidence level (This process is illustrated in, e.g., Panels B and C of Fig.2 in Sect. 2.2).

The balance-of-evidence hypothesis implicitly assumes that the amounts of evidence for both accumulators are directly accessible to calculate the difference at the time of the decision (Ratcliff & Starns, 2009). Also, there is a scaling issue in the difference in evidence in that its value can largely differ by the scale on which a subject is asked to make confidence judgments (e.g., confidence judgments with a decision boundary set on a 1–5 scale vs a 1–100 scale). To circumvent this issue, (Merkle and Van Zandt 2006) proposed to use a relative balance of evidence defined as the ratio of the winning accumulator's evidence to the sum of the two accumulators' evidence.

1.4. Response and Time Models of Confidence Judgments (RTCON)

Ratcliff and Starns (2009, 2013) developed *Response and Time Models of Confidence Judgments*, namely RTCON and RTCON2 models (hereafter denoted as RTCON models) based on an extension of the diffusion decision process for simple two-choice tasks (Ratcliff, 1978; Ratcliff & McKoon, 2008) to confidence judgments in perceptual and cognitive decision-making tasks. They based their model development on the observation that a subject in a confidence judgment task requires a separate response (e.g., a separate key on a keyboard) for each confidence category (Ratcliff & Starns, 2009). This led them to the assumption that the decision process for confidence judgments involves multiple evidence accumulators each of which corresponds to one of the confidence categories.

There are three key assumptions in the RTCON models (Fig. 3 in Sect. 2.3 illustrates a psychometric version of these models). First, for each stimulus, the model represents the information for decision-making from memory as a distribution (memory strength distribution) rather than a single value. Second, internal confidence criteria divide the memory strength distribution into several regions, each of which corresponds to one of the confidence levels on the scale. Also, the area of each region on the distribution is mapped onto the drift rate of the corresponding evidence accumulator. Lastly, evidence accumulators have their own decision boundaries and thus different confidence categories require more or less information for their accumulators to terminate.

The RTCON2 model (unlike the RTCON model) implements the constant summed evidence algorithm. This algorithm puts a constraint on evidence that different accumulators collect at each time point so that the sum of the evidence over confidence categories is fixed to zero. Specifically, at each time point, evidence for only one selected accumulator increases in its drift rate, and the other accumulators get equal amounts of decrements in evidence whose sum is equal to the increment of the selected accumulator. The race between multiple accumulators terminates when one of the accumulators reaches its decision boundary. Response is predicted as the confidence response corresponding to the winning accumulator and RT is predicted as the sum of the termination time of the winning accumulator and nondecision time.

The RTCON models were originally proposed for perceptual and cognitive multiple-choice tasks such as recognition memory tasks and motion discrimination tasks with confidence judgments. Therefore, they can also fit ordinal responses for the measurement of latent traits with slight modifications.

2. Models

In this section, we describe our psychometric process models of ordinal responses and RTs. We develop these models by integrating the theories described in Sect. 1. Table 1 provides a summary of the models. All three models assume evidence accumulation as an adequate representation

TABLE 1.
Summary of the Three Proposed Models. Section 2 provides the detail of the proposed models. The 'Source' column shows the primary motivations (early theories and models) of the proposed models, but note that other theories and models are also integrated to develop them.

	K	Source	n.par	Drift	Boundary
Model 1	1	LRF (Sect. 1.2)	$3P + 2I + (M-1)$	$v_{pi} = \lvert\theta_p - b_i\rvert$	$s \cdot \gamma_p / a_i$
Model 2	2	BoE (Sect. 1.3)	$3P + 2I + (M-1)$	$v_{1.pi} = \Phi(\theta_p - b_i)$ $v_{2.pi} = 1 - v_{1.pi}$	γ_p / a_i
Model 3	M	RTCON (Sect. 1.4)	$3P + 2I + 2M - 2$	$v_{k.pi} = \Phi(\tau_k - (\theta_p - b_i))$ $-\Phi(\tau_{k-1} - (\theta_p - b_i))$ $k = 1, \ldots, M$	$s_k \cdot \gamma_p / a_i$ $k = 1, \ldots, M$

EA: Evidence Accumulation, n.par: The number of parameters, P: The number of persons, I: The number of items, K: The number of accumulators, M: The number of response options.

of psychological processes underlying the measurement of latent traits. Also, as psychometric models, all three models should be able to decompose person and item effects that are entangled in the measurement data. To this end, we assume that decision boundary α_{pi} and drift rate v_{pi} for person p and item i are functions of the person and item parameters as follows:

$$\begin{aligned} \alpha_{pi} &= w(\gamma_p, a_i, \ldots) \\ v_{pi} &= u(\theta_p, b_i, \ldots) \end{aligned} \tag{3}$$

where γ_p and θ_p represent person-wise decision boundary and person-wise drift rate, respectively, and a_i and b_i represent item time-pressure (which can also be interpreted as the inverse of item discrimination in the IRT models) and item strength[1] parameters, respectively. In our models, a respondent with a larger person-wise decision boundary is more likely to produce longer RTs. Also by spending more time, response is likely to be the one predicted by the other model parameters (a 'dominant' response). A large positive (negative) person-wise drift rate is likely to produce positive (negative) responses to the item with short RTs, while a small (close to zero) value is likely to produce intermediate and neutral responses with longer RTs. Item parameters are assumed to have the opposite relationship with responses and RTs; a smaller item time pressure parameter is associated with longer RTs and more dominant responses and a large negative (positive) item strength parameter is associated with shorter RTs and more positive (negative) responses. This association will be further elaborated with model equations in Section 2.1–Section 2.3 and a simulated result in Section 2.5. A similar decomposition was used in earlier psychometric process models for binary responses and RTs such as the diffusion IRT model and its extensions (Kang et al., 2022a,b; Molenaar et al., 2015b; Ranger et al., 2017; Tuerlinckx & De Boeck, 2005; Tuerlinckx et al., 2016; van der Maas et al., 2011). For nondecision time, we include person-wise nondecision time parameter t_{0p} to capture individual differences in how much time each respondent spends on nondecision processes.[2] We also denote the random variables of response and RT for person

[1] We used the term 'item strength', in a similar sense as item difficulty in IRT analysis of test data in that a respondent with a higher value of latent trait than item strength has a higher probability of endorsing the personality/attitude measurement item. This can also be called 'item attractiveness' in that it represents the attractiveness of the item.

[2] An alternative to this choice could be item-wise nondecision time. However, modeling both person-wise and item-wise nondecision times is not feasible because their effects on outcome variables are confounded. In other words, person-wise (item-wise) nondecision time parameters can also account for inter-item (inter-person) differences. We consider person-wise nondecision time in this article because typically measurement data have more respondents (than items) and

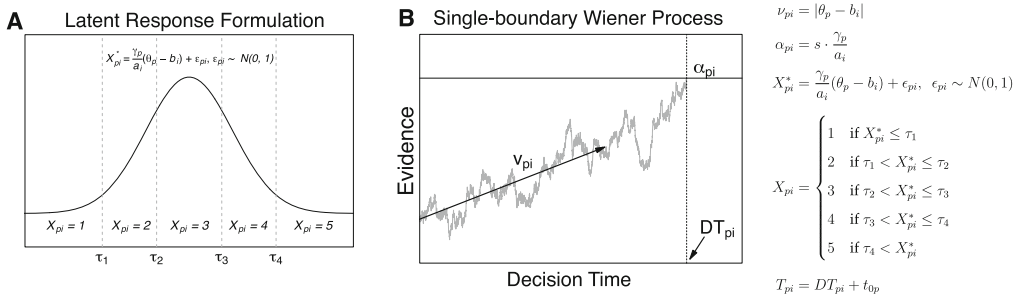

FIGURE 1.
Illustration of Model 1: Single-accumulator model. The model is a combination of the latent response formulation and the single-boundary Wiener process.

p and item i by X_{pi} and T_{pi}, their realizations by x_{pi} and t_{pi}, and the matrices of all responses and of all RTs by $\boldsymbol{X}$ and $\boldsymbol{T}$, respectively. In addition, we use P for the number of persons and I for the number of items.

2.1. Model 1: Single-accumulator Model

The first model we propose is the single-accumulator model in which a single evidence accumulation process represents a continuous psychological construct. The model, as shown in Fig. 1, is a combination of the latent response formulation (Sect. 1.2) and the single-boundary Wiener process (Sect. 1.1). The latent response formulation determines the response probabilities of different response options and the single-boundary Wiener process determines the RT. The model has only a single accumulator, but different responses on an ordinal scale can be predicted. The model representation of the cognitive processes underlying the measurement says that evidence accumulates only for the predicted response option. Winning probabilities of different accumulators are assumed to be determined by the latent response formulation, as will be described below.

The model assumes a continuous latent response X^*_{pi} that underlies measurement outcomes. It is assumed that X^*_{pi} is normally distributed as $X^*_{pi} = \frac{\gamma_p}{a_i}(\theta_p - b_i) + \epsilon_{pi}$ where $\epsilon_{pi} \sim N(0, 1)$. The combinations of person and item parameters, $\frac{\gamma_p}{a_i}$ and $\theta_p - b_i$, will also be related to the decision boundary α_{pi} and the drift rate ν_{pi} for RT predictions, producing dependency between responses and RTs. Compared to the latent response formulation in the factor analysis literature, the factor score is assumed to be the person drift rate θ_p with item threshold parameter b_i as its mean and the factor loading is assumed to be a function of person boundary and item time-pressure (or the inverse of item discrimination). This choice for the factor loading is motivated by the parameterization of the diffusion IRT model introduced in Tuerlinckx and De Boeck (2005), which showed that the discrimination parameter in the two-parameter logistic IRT model (which is corresponding to the factor loading in the linear factor analysis model) can be interpreted as the distance between the two decision boundaries in the diffusion model.

For M response options, the model applies $M - 1$ response thresholds $\tau_1, \cdots, \tau_M$ to predict ordinal responses X_{pi} as $X_{pi} = k$ if $\tau_{(k-1)} < X^*_{pi} \le \tau_k$ where $\tau_{(k-1)} < \tau_k$ $(k = 1, \ldots, M)$ and $\tau_0 = -\infty$ and $\tau_M = \infty$. Note that the response thresholds are fixed across items as in the modified graded response model (Muraki, 1990) and the rating scale model (Andrich, 1978a,b) which are appropriate for attitude questionnaires with the same response anchors (Embretson & Reise, 2000). However, they can be allowed to vary by item (i.e., $\tau_{i,k}$) given a sufficiently large number of respondents as in Muthén (1984) and Samejima (1969), Samejima (1997).

thus models with person-wise nondecision time parameters can better decompose RTs into decision and nondecision times.

For RTs, the model employs the single-boundary Wiener process (Eq. 1) with the decompositions $\nu_{pi} = |\theta_p - b_i|$ and $\alpha_{pi} = s \cdot \frac{\gamma_p}{a_i}$. These are similar to the decompositions used for the diffusion IRT model but with two differences. First, we take for the drift rate the absolute distance between person drift rate and item strength to ensure the termination of the single-boundary process. This also allows the model to account for the distance-difficulty hypothesis for traits which postulates that a person takes more time to respond to an item when the absolute distance between latent ability and item threshold is small and less time when the absolute distance is large (Ferrando & Lorenzo-Seva, 2007a,b; Kuiper, 1981; Kuncel, 1973; Molenaar et al., 2015a). The second one is the parameter s in the decision boundary decomposition. Without this parameter, the model assumes that the decision boundary is equal to the factor loading in the latent response formulation. This assumption is too restrictive because it is unlikely that the decision boundary (which scales RTs) can be measured on the same scale as the factor loading (which scales continuous latent responses). The scaling parameter fills the gap in scale between the continuous latent responses and the RTs while keeping the parsimony of the model (compared to a model with separate parameters for boundaries and loadings). Introducing this scaling parameter produces an interaction between person-wise decision boundary, item time-pressure, and the scaling parameter per se. Thus, for identifiability, we fixed one of the item-time pressure parameters to 1 (e.g., $a_1 = 1$). In total, the current model has $3P + 2I + (M - 1) + 1 - 1$ parameters (-1 for a constraint).

Given the parameterization above, the log-likelihood of Model 1 can be obtained as follows. For RTs, it is simply the log of the density in Eq. 1 summed over persons and items. For responses, the log-likelihood can be obtained by summing the log of the probability of observing a response x_{pi} over all persons and items, as in the graded response model (Samejima, 1969, 1997), but with a different formula for response probabilities. From the formula for X_{pi} given above, the response probability is $P(X_{pi} = x_{pi}|\gamma_p, a_i, \theta_p, b_i, \boldsymbol{\tau}) = \Delta\Phi(\tau_{x_{pi}} - \frac{\gamma_p}{a_i} \cdot (\theta_p - b_i)) = \Phi(\tau_{x_{pi}} - \frac{\gamma_p}{a_i} \cdot (\theta_p - b_i)) - \Phi(\tau_{(x_{pi}-1)} - \frac{\gamma_p}{a_i} \cdot (\theta_p - b_i))$ where $\boldsymbol{\tau}$ is the vector of response thresholds and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Hence, the full log-likelihood of Model 1 is:

$$
\begin{aligned}
l_1(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{t_0}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\tau}, s|\boldsymbol{X}, \boldsymbol{T}) &= \sum_{p=1}^{P}\sum_{i=1}^{I}\left[\log f_{T_{pi}}(t_{pi}; \alpha_{pi}, \nu_{pi}, t_{0p}) + \log \Delta\Phi(\tau_{x_{pi}} - \frac{\gamma_p}{a_i} \cdot (\theta_p - b_i))\right] \\
\nu_{pi} &= |\theta_p - b_i|, \quad \alpha_{pi} = s \cdot \frac{\gamma_p}{a_i}
\end{aligned}
\tag{4}
$$

2.2. *Model 2: Dual-accumulator Model*

Our second model, which we call the dual-accumulator model, has a similar structure as the balance-of-evidence models (Sect. 1.3): two competing accumulators and their balance in evidence mapped to confidence or strength of response (Fig. 2). The fundamental assumption of this model is that a psychological construct is represented by a continuum rather than a single value, as in the latent response formulation (e.g., continuum of extraversion–introversion, continuum of motivated–demotivated, etc.). The two competing accumulators assumed in the balance-of-evidence hypothesis represent the two extremes on the psychological continuum. Without loss of generality, we assume that the first accumulator represents the positive extreme (*agree* side) and the second accumulator represents the negative extreme (*disagree* side).

To determine the drift rates of the two accumulators, the model implements some key assumptions in the RTCON models (Sect. 1.4). First, we represent the item strength as a distribution, rather than a single value. We further assume that item strength is normally distributed with item strength parameter b_i as its mean and unit standard deviation for identifiability. Second, we assume that the person-wise drift rate θ_p plays a role as a confidence criterion in the RTCON models. For person p and item i, θ_p splits the item strength distribution into two regions and the area of

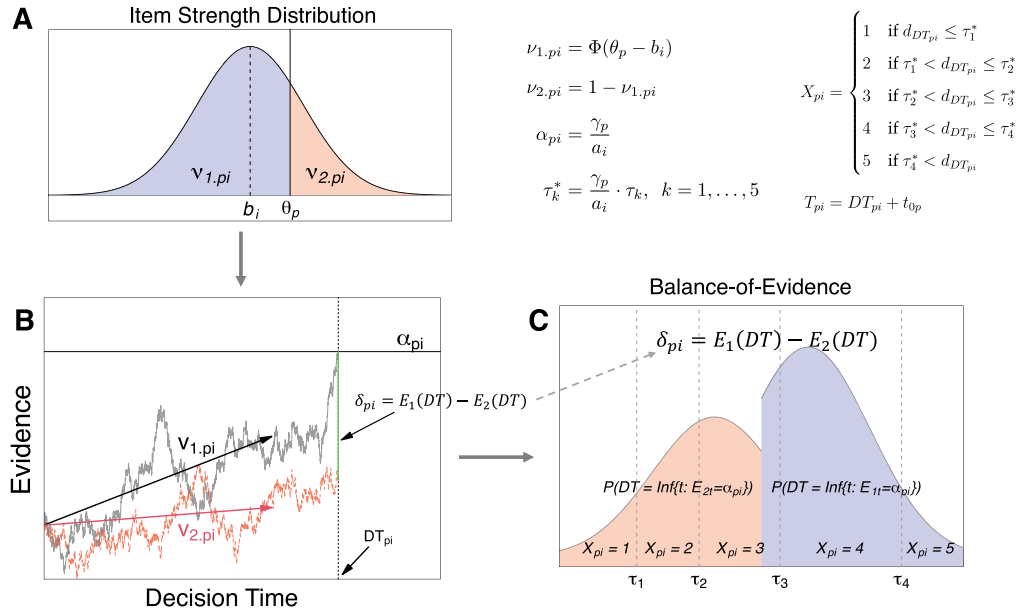

FIGURE 2.
Illustration of Model 2: Dual-accumulator Model. The model implements balance-of-evidence hypothesis to predict ordinal responses from the race between two evidence accumulators.

each region determines the drift rate of evidence accumulation as $\nu_{1.pi} = \Phi(\theta_p - b_i)$ for the first accumulator and $\nu_{2.pi} = 1 - \nu_{1.pi}$ for the second accumulator. Thus, a higher value of θ_p and a smaller value of b_i would be associated with a higher proportion of positive responses, which is similar to the relationship between person-wise latent ability, item difficulty parameter, and correct response proportion in the IRT models for binary responses in ability tests. For the decision boundary, we assume that the two accumulators race toward the same boundary determined as $\alpha_{p,i} = \frac{\gamma_p}{a_i}$.

As in the balance-of-evidence models, response and RT are predicted from the race of the two competing accumulators. Let $E_j(t)$ be the amount of evidence accumulated until time t by the j-th accumulator ($j = 1, 2$). For example, suppose that the first accumulator represents introversion, while the second accumulator represents extraversion. If the first accumulator wins the race, an introversion-oriented response is made. If the difference in accumulated information is large, an extreme response such as *strongly agree* to an introversion-oriented item is predicted. If the difference is small, a relatively weak response such as *agree* or *neutral* is predicted. For person p and item i, suppose that the jth accumulator terminates at time $DT_{j,pi} = \inf\{t : E_j(t) = \alpha_{pi}\}$. Then, the model predicts the decision time (DT_{pi}) and RT (T_{pi}) as:

$$\begin{aligned} DT_{pi} &= \min\{DT_{1,pi}, DT_{2,pi}\} \\ T_{pi} &= t_{0p} + DT_{pi} \end{aligned} \tag{5}$$

Also, we define $d(t) = E_1(t) - E_2(t)$, the difference in accumulated evidence between the two accumulators at time t. The difference at the decision time $d(DT_{pi})$ is positive (negative) if the first (second) accumulator wins the race. We apply response thresholds τ_k to $d(DT_{pi})$ to determine the ordinal response as $X_{pi} = k$ if $\tau^*_{(k-1)} < d(DT_{pi}) \leq \tau^*_k$. Here $\tau^*_k = \frac{\gamma_p}{a_i} \cdot \tau_k$ ($k = 1, \cdots, M$), that is, the response threshold scaled by the decision boundary $\alpha_{pi} = \frac{\gamma_p}{a_i}$. We

used this scaling because the accumulated evidence can be larger and smaller as a function of α_{pi}. We also impose an additional constraint, $\tau_2 < 0 < \tau_3$ (for $M = 5$) on the response threshold. Without such a constraint, for example if $0 < \tau_2 < \tau_3$, a response can be 2 (e.g., *disagree*) even if the first accumulator (that represents positive responses) wins the race. Thus, the constraint makes the two accumulators correctly represent the two extremes of the ordinal response scale. In total, Model 2 has $3P + 2I + (M - 1)$ parameters.

The first passage time distribution of each accumulator is obtained as a shifted Wald distribution, but with a different drift rate $\nu_{1,pi}$ or $\nu_{2,pi}$. For accumulators $j = 1, 2$, let f_{T_1} and f_{T_2} be their first passage time densities and F_{T_1} and F_{T_2} be their first passage time distribution functions, respectively. Then, the defective density of the first accumulator is obtained as $g_1(t; *) = f_{T_1}(t; *)(1 - F_{T_2}(t; *))$ and that of the second accumulator as $g_2(t; *) = f_{T_2}(t; *)(1 - F_{T_1}(t; *))$ where '$*$' represents the set of all model parameters related to the function. By 'defective', it means that these density functions do not integrate to 1; instead, the sum of two integrals over t is 1 and each integral produces the probability of the corresponding accumulator winning the race. We can also obtain the defective cumulative distribution functions $G_j(t; *) = \int_0^t g_1(u; *)du$, $j = 1, 2$ and the winning probability of accumulator j can be expressed as $P(\text{j wins}) = G_j(\infty; *) = \int_0^\infty g_1(u; *)du$. The RT density h_T of Model 2 is obtained as the sum of the two defective densities:

$$\begin{aligned} h_T(t; *) &= h_T(t, 1 \text{ wins}; *) + h_T(t, 2 \text{ wins}; *) = g_1(t; *) + g_2(t; *) \\ &= f_{T_1}(t; *)(1 - F_{T_2}(t; *)) + f_{T_2}(t; *)(1 - F_{T_1}(t; *)) \end{aligned}$$

The likelihood of the responses can be obtained using the formula for X_{pi} explained above and the cumulative distribution function Φ_2 of $d(DT_{pi})$ derived in Section S1 in the supplementary material:

$$\begin{aligned} \Phi_2(\delta) =& G_{T_1}(\infty; -) \cdot \left(1 - \Phi_{TR}\left(\frac{(\alpha_{pi} - \delta) - \nu_{2.pi} \cdot DT_{pi}}{\sqrt{(DT_{pi})}}\middle| -\infty, \frac{\alpha_{pi} - \nu_{2.pi} \cdot DT_{pi}}{\sqrt{(DT_{pi})}}\right)\right) \\ &+ G_{T_2}(\infty; -) \cdot \Phi_{TR}\left(\frac{(\alpha_{pi} + \delta) - \nu_{1.pi} \cdot DT_{pi}}{\sqrt{(DT_{pi})}}\middle| -\infty, \frac{\alpha_{pi} - \nu_{1.pi} \cdot DT_{pi}}{\sqrt{(DT_{pi})}}\right) \end{aligned}$$

where $\Phi_{TR}(\cdot|a, b)$ is the cumulative distribution function of the truncated standard normal distribution with a and b as its lower and upper bounds. Hence, the full log-likelihood of Model 2 can be obtained as:

$$\begin{aligned} l_2(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{t_0}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\tau}|\boldsymbol{X}, \boldsymbol{T}) &= \sum_{p=1}^{P}\sum_{i=1}^{I}\left[\log h_{T_{pi}}(t_{pi}; \alpha_{pi}, \nu_{pi}, t_{0p}) + \log \Delta\Phi_2(\alpha_{pi} \cdot \tau_{x_{pi}})\right] \\ \Delta\Phi_2(\alpha_{pi} \cdot \tau_{x_{pi}}) &= \Phi_2(\alpha_{pi} \cdot \tau_{x_{pi}}) - \Phi_2(\alpha_{pi} \cdot \tau_{(x_{pi}-1)}) \end{aligned} \quad (6)$$

The idea of implementing the balance-of-evidence hypothesis into a psychometric model is not new and was already used in Ranger and Kuhn (2018)'s joint model of ordinal responses and RTs based on the LBA model (Brown & Heathcote, 2008). The biggest difference between this model and our Model 2 is in the assumption of noise in evidence within the accumulation process; Model 2 (as the other proposed models in this article) assumes noisy evidence accumulation, while Ranger and Kuhn's model assumes that evidence accumulation is linear without noise (i.e., no diffusion coefficient σ^2 is defined). Despite some similarities in their structures, these two

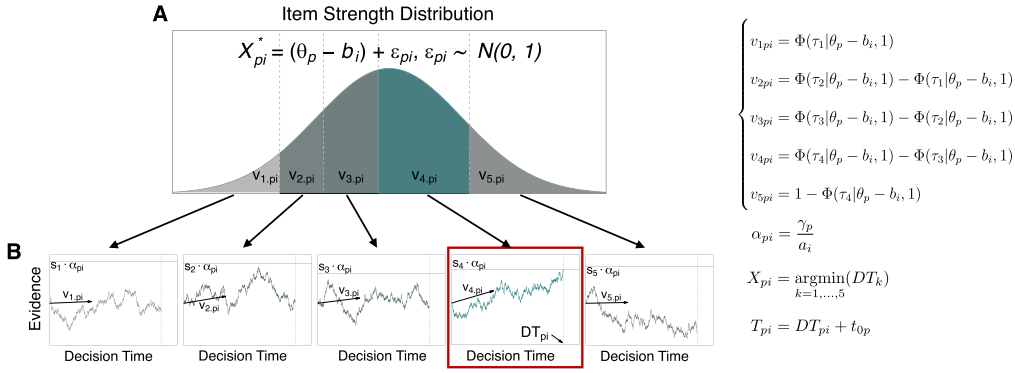

FIGURE 3.
Illustration of Model 3: Multi-accumulator Model. The model is a psychometric modification of the RTCON models.

models have different primary sources of their probabilistic characteristics due to this difference in assumption, which has an important implication for process models. This will be more thoroughly discussed in Sect. 5.

2.3. *Model 3: Multi-accumulator Model*

Model 3 (Fig. 3), which we call the multi-accumulator model or the M-accumulator model, is a psychometric modification of the RTCON models (Sect. 1.4). We modify the three key assumptions of the RTCON models to account for ordinal responses and RTs for the measurement of latent traits. For the first two key assumptions, the modifications are similar to those for Model 2, but with important differences to determine drift rates for M evidence accumulators. First, we assume that the item strength is normally distributed with mean $\theta_p - b_i$ and unit standard deviation for identifiability. Second, we assume that $M - 1$ response thresholds $\tau_1 < \cdots < \tau_{(M-1)}$ work as confidence criteria in the RTCON models. They divide the item strength distribution into M separate regions. Each of these regions corresponds to one of the ordinal response options and its area determines the drift rate $v_{k.pi}$ $(k = 1, \ldots, M)$ of the corresponding accumulator. For $M = 5$, we have:

$$\begin{cases} v_{1.pi} = \Phi(\tau_1 - (\theta_p - b_i)) \\ v_{k.pi} = \Phi(\tau_k - (\theta_p - b_i)) - \Phi(\tau_{k-1} - (\theta_p - b_i)), \quad k = 2, \ldots, M-1 \\ v_{5.pi} = 1 - \Phi(\tau_4 - (\theta_p - b_i)) \end{cases}$$

With the modifications stated above, Model 3 can determine M drift rates and M-point ordinal responses without additional parameters while it keeps some favorable psychometric properties. For example, a higher proportion of positive responses is positively correlated with the person-wise drift rate θ_p and is negatively correlated with the item strength parameter b_i as in Model 2 despite their differences in parameterizations. The mean of $\theta_p - b_i$ of the item strength distribution implies that the information that persons with different levels of cognitive processing (i.e., drift rate) extract from the same item i can differ in strength. An alternative interpretation for the item strength distribution in Model 3 is that we use the latent response formulation for the item strength distribution. In this view, a continuous latent response X^*_{pi} represents the item strength, but this does not directly determine the ordinal responses. Instead, it determines the processing efficiency of the evidence accumulators, which, in turn, predicts response proportions. The continuous latent response can be expressed as $X^*_{pi} = (\theta_p - b_i) + \epsilon_{pi}, \epsilon_{pi} \sim N(0, 1)$, which

is similar to the expression used for Model 1. However, the factor loading is fixed to 1 in Model 3 for identifiability (instead of setting it to $\frac{\gamma_p}{a_i}$ as done in Model 1).

The third key assumption of the RTCON models states that each evidence accumulator has its own decision boundary so that accumulators for extreme response options may require more or less information to terminate than accumulators for relatively intermediate options. The RTCON models are able to estimate all decision boundaries independently for each subject using data with multiple trials from the same subject. This is not plausible for psychometric data as each person responds to each item only once. Thus, we use $\alpha_{pi} = \frac{\gamma_p}{a_i}$ as a general decision boundary but introduce the boundary scaling parameters s_k $(k = 1, \cdots, M)$ to allow accumulators to have different boundaries. The decision boundary for the k-th accumulator is determined as $s_k \cdot \alpha_{pi}$. One of the scaling parameters should be fixed to 1 for identifiability. In our simulation study and empirical applications, we have $M = 5$, and thus, we fix $s_3 = 1$ so that the accumulator for the intermediate option has the decision boundary of α_{pi} and the other accumulators have their boundaries relative to this. In total, Model 3 has $3P + 2I + (M - 1) + M - 1$ parameters (-1 for a constraint).

With the M accumulators for M response options, Model 3 predicts RT as the sum of the termination time of the winning accumulator and the nondecision time. Also, the model predicts that the response option corresponding to the winning accumulator is chosen. The model likelihood is obtained from the defective densities of the accumulators. For accumulator k $(k = 1, \ldots, M)$, its defective density is $g_k(t; *) = f_{T_k}(t; *) \prod_{j \neq k} \left(1 - F_{T_j}(t; *)\right)$ where '$*$' represents the set of all model parameters related to the function. Then, the full log-likelihood is:

$$l_3(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{t_0}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\tau}, \boldsymbol{s} | \boldsymbol{X}, \boldsymbol{T}) = \sum_{p=1}^{P} \sum_{i=1}^{I} \left[\log g_{x_{pi}}(t_{pi}; *) \right] \tag{7}$$

If one wants to evaluate the likelihood of responses and that of RTs separately, the probability of choosing response option k can be obtained as the integral of the defective density g_k for the k-th accumulator from 0 to ∞, or equivalently, as the value of the defective cumulative distribution function G_k evaluated at $t \to \infty$. The RT density h_T of this model can be obtained as the sum of all defective densities:

$$\begin{aligned} P(X = k; *) &= G_k(\infty; *) = \int_0^{\infty} g_k(u; *) du \\ h_T(t; *) &= \sum_{k=1}^{M} \log g_k(t; *) \end{aligned} \tag{8}$$

2.4. An Overarching Population Distribution

All three models that we proposed above have person-wise decision boundary γ_p and drift rate θ_p. As done in earlier psychometric models of responses and RTs (e.g., Bolsinova et al., 2017; van der Linden, 2007), we assume a hierarchical population (prior) distribution for these person-wise parameters so that we can account for a potential across-person correlation between them. Specifically, we use a multivariate normal distribution as follows:

$$(\theta_p, log(\gamma_p)) \sim MVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \omega_\theta^2 & \rho_{\theta\gamma}\omega_\theta\omega_\gamma \\ \rho_{\theta\gamma}\omega_\theta\omega_\gamma & \omega_\gamma^2 \end{bmatrix} \right) \tag{9}$$

Note that because person-wise boundary separation is positive-valued, it is log-transformed in Eq. 9 to match the support (real-valued) of the distribution. It is similar to the choice of log-normal prior on γ_p used for the diffusion IRT model in van der Maas et al. (2011) although they did not assume a prior covariance/correlation between person-wise drift rate and boundary separation. The prior specification in Eq. 9 adds three more parameters (variances and covariance) to all three models. It is also possible to consider an item-domain distribution and describe the correlational structure of the item parameters. Although we do not follow this approach, we refer readers to van der Linden (2007) and Bolsinova et al. (2017) for examples of item-domain distributions.

2.5. Relations Between Latent Variables, Response Proportions, and RT Distributions

The models we proposed have different theories and structures of cognitive processes underlying personality/attitude measurement. It would be informative to illustrate how parameter values of the different models are related to outcome variables. From the model structures presented in Figs. 1, 2, and 3, it can be expected that:

- A larger person-wise decision boundary γ_p (a smaller item time-pressure parameter a_i) is associated with longer RTs and more dominant responses predicted by the other model parameters (e.g., in Model 3, if the drift rate for the fourth accumulator is the largest, the fourth response option would be the dominant response predicted by the model).
- A positively (negatively) large person-wise drift rate θ_p and a negatively (positively) large item strength parameter b_i are associated with short RTs and strong positive (negative) responses on an ordinal scale (e.g., *strongly agree* and *strongly disagree*). In fact, the distance between θ_p and b_i determines these effects.
- A large person-wise nondecision time t_{0p} increases RTs (equally for the entire RT ranges) as it is a shift parameter while it does not affect response proportions.
- Distributions of response thresholds τ_k produce large effects on response proportions. For example, more extreme responses (e.g., *strongly disagree* and *strongly agree*) are predicted when response thresholds are closer to the zero point and to each other. For Model 3, response thresholds also affect RTs as they determine the drift rates for accumulators; an accumulator with a large area determined by response thresholds would have a larger drift rate, and thus, more responses corresponding to this accumulator would be expected.
- A larger scaling parameter (s in Model 1 and s_k in Model 3) produces larger decision boundaries for accumulators and so it has a similar effect as a larger person-wise decision boundary and a larger item time-pressure parameter).

In Fig. 4, we generated response proportions and RT quantiles from all three proposed models as a function of log-transformed person-wise decision boundary $\log(\gamma_p)$ (Fig. 4A, the first two columns) and person-wise drift rate θ_p (Fig. 4B, the last two columns), which are parameters of our main interests. Effects of the other parameters can be similarly predicted, as described in the bulleted list above. For each panel in Fig. 4, each of the model parameters, except for the one on the x-axis, ($\log(\gamma_p)$ in Fig. 4A and θ_p in Fig. 4B) was fixed to a constant. For example, $b_i = 0$ for all three models but a_i and τ_k were given different values across the three models. In Panel A, θ_p was given positive and sufficiently large values for all three models so that the dominant response option is Response 5. Similarly in Panel B, $\log(\gamma_p)$ was given a different value for each of the three models. This choice was to generate similar data distributions despite the structural differences of the models.

In each of Fig. 4A and B, the panels in the first columns show the changes of RT quantiles and those in the second columns show the changes of response proportions, as a function of the parameter on the x-axis. The legend for each column is shown at the top panel in the same column. The figure confirms what is expected based on the model equations. As the (log-transformed) person-wise decision boundary $\log(\gamma_p)$ increases (Fig. 4A), RTs get longer and the proportion

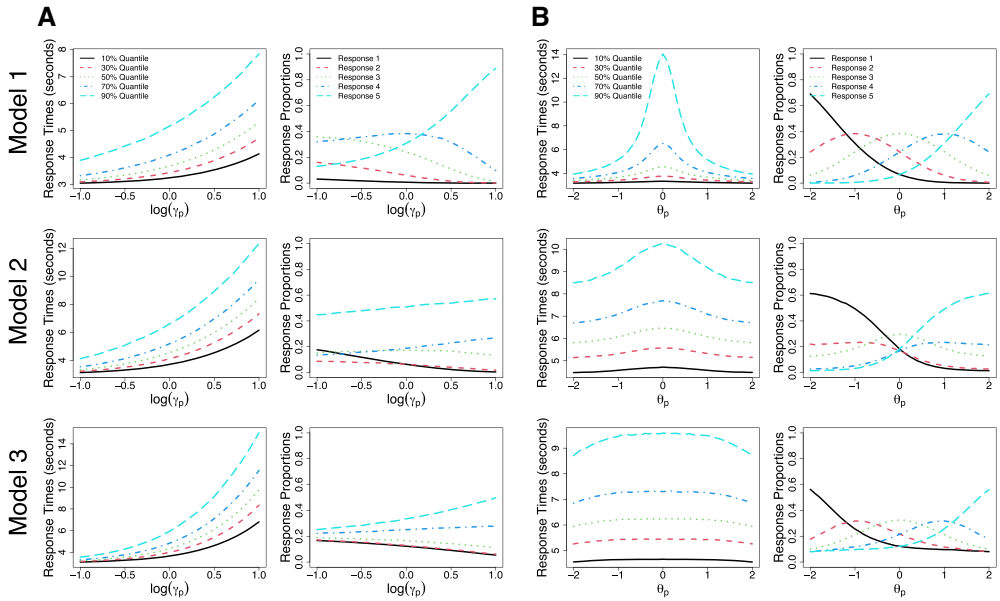

FIGURE 4.
Illustration of the Relations between Cognitive Latent Variables (Person-wise Drift Rate and Decision Boundary), Response Proportions, and RT Distributions. Each of the model parameters, except for the one on the x-axis, was fixed to a constant (Different constant values were used across models due to their structural differences). In Panel A, person-wise drift rate θ_p was given positive and sufficiently large values for all three models so that the dominant response option is Response 5.

of the dominant response option (Response 5 in this figure, for all models) increases. Which response option is dominant is determined by the other model parameters such as person-wise drift rates, item strength parameters, and response thresholds. For Fig. 4, we made Response 5 dominant to show that all three models can make similar predictions, but other response options can be dominant as well. The figure also shows that Models 2 and 3 show more gradual changes in both the RT quantiles and the response proportions than Model 1. This is because the models have differences in how they define drift rates for different accumulators and for Models 2 and 3, these drift rates are always smaller than 1.

When the person-wise drift rate θ_p is manipulated (Fig. 4B), changes in RT quantiles and response proportions are also consistent with what we expect from the model equations (the bulleted list above). When θ_p increases and diverges from $b_i = 0$, RTs tend to decrease. If θ_p moves toward the positive (negative) direction, more positively (negatively) stronger responses are predicted. These changes are more gradual in Models 2 and 3 because their drift rates for accumulators are smaller than 1. Also, a change in 90% RT quantile is rather drastic in Model 1 because there is only one accumulator for Model 1, whereas there are multiple accumulators and the minimum decision time over accumulators is used to determine RT in the other models.

Note that, although the result illustrated in Fig. 4 is based on our choice of model parameter values, similar patterns can be predicted by the model equations and structures when the parameters are given other values. For example, if person-wise drift rate θ_p is given negatively large values in all three models in Panel A, the dominant response option would be Response 1 (instead of 5) but RT would show similar patterns.

3. Simulation: Parameter Recovery

3.1. Data Generation and Model Estimation

This section provides simulation studies with which we aimed to show that the models can recover their parameters reasonably well. We conducted the simulations as follows. First, we generated true values of data-generating parameters. Given the parameter values, we simulated responses and RTs with $P = 200$ persons and $I = 10$ items. The data size was meant to be a minimum requirement for reasonable recovery of the model parameters. Each of the three models produced two $P \times I$ matrices, one for ordinal responses with $M = 5$ possible options and the other for RTs. For each model, we generated 50 synthetic datasets. Then, we fitted each model to the synthetic data generated from that model.

For the data-generating parameters, person-wise drift rates (θ_p) were sampled from the standard normal distribution and log-transformed person-wise boundary separations ($\log(\gamma_p)$) were sampled from a uniform distribution with mean 0 and standard deviation of $1/2$. Item strength (b_i) parameters were sampled from a uniform distribution with the range of $[-1, 1]$ for all three models. Item time-pressure (a_i) parameters were sampled from uniform distributions with different ranges for the three models due to their structural differences. Specifically, the ranges were [0.5, 1.5], [0.3, 1.0], and [0.2, 1.0], for Models 1, 2, and 3, respectively. Similarly, we used different values for response threshold parameters and boundary scaling parameters by model, as shown in Table 3. Our choice for true parameter values and their distributions was motivated to produce synthetic data with reasonably distributed response proportions and RTs that generally vary in the range of 2-12 seconds with median RTs of 4–6 seconds, which are similar to those in our target data with ordinal responses and RTs (Sect. 4).

Then, we fitted the models to the synthetic data using **Stan** (Stan Development Team, 2021). The repository of **Stan** codes to fit the proposed models are provided in the Code Availability section in the article. As we used a Bayesian sampling method, prior distributions of the model parameters should be specified. For person-wise drift rates and decision boundaries, we implemented a hierarchical distribution as stated in Sect. 2.4. For person-wise nondecision time, we used a uniform distribution $t_{0p} \sim U(0, \min_i(T_{pi}))$. The range was a natural choice because the RT measure cannot be negative, and it is the sum of decision and nondecision times and thus nondecision time of person p should be smaller than the minimum RT of the same person. For item parameters, we used weakly informative priors with wide ranges and dispersions: $log(a_i) \sim N(0, 5)$ and $b_i \sim N(0, 5)$ where $N(\mu, \sigma)$ is a normal distribution with mean μ and standard deviation σ. Response threshold parameters were also given a uniform prior, but they are constrained to be monotonically increasing. For $M = 5$, we used $\tau_1 \sim U(-5, \tau_2)$, $\tau_2 \sim U(\tau_1, \tau_3)$, $\tau_3 \sim U(\tau_2, \tau_4)$, and $\tau_4 \sim U(\tau_3, 5)$ where $U(a, b)$ is a uniform distribution with range (a, b), so that $\tau_1 < \tau_2 < \tau_3 < \tau_4$. The range of $(-5, 5)$ was wide enough to cover various possibilities. Models 1 and 3 have the boundary scaling parameters and we gave them a noninformative uniform prior with range (0, 10).

With the prior specifications above and the model likelihoods in Eqs. 4, 6, and 7, we obtained 2,500 posterior samples of the model parameters with 3 chains. We discarded the first 1,000 samples for each chain for burn-in. The remaining samples were used to approximate the joint posterior distributions of the model parameters. We assessed the convergence of Bayesian chains with the Gelman–Rubin convergence diagnostic ($\hat{R}$) by checking whether its values were smaller than 1.1 for all model parameters (Gelman, 1996; Gelman et al., 2013). We also obtained a posterior density from each chain and inspected if densities were consistent across chains.

TABLE 2.
Parameter recovery of the single-accumulator model (Top; Model 1), the dual-accumulator model (Middle; Model 2), and multi-accumulator model (Bottom; Model 3).

	Model 1				
	MSE	Bias	SE	BSE	Cor
θ_p	0.110	0.111	0.270	0.280	0.955 (0.008)
γ_p	0.069	0.101	0.208	0.246	0.872 (0.014)
t_{0p}	0.104	0.154	0.186	0.215	0.905 (0.020)
a_i	0.003	0.009	0.052	0.054	0.988 (0.005)
b_i	0.012	0.090	0.064	0.097	0.996 (0.002)
τ_k	0.003	0.006	0.055	0.054	
s or s_k	0.011	0.075	0.076	0.085	
ω_γ	0.001	0.023	0.022	0.032	
ω_θ	0.003	0.049	0.033	0.065	
$\rho_{\gamma\theta}$	0.013	0.105	0.042	0.085	
	Model 2				
	MSE	Bias	SE	BSE	Cor
θ_p	0.409	0.252	0.491	0.671	0.777 (0.030)
γ_p	0.057	0.071	0.204	0.239	0.891 (0.020)
t_{0p}	0.100	0.138	0.187	0.233	0.891 (0.027)
a_i	0.004	0.051	0.030	0.037	0.994 (0.003)
b_i	0.042	0.034	0.198	0.223	0.974 (0.015)
τ_k	0.002	0.013	0.040	0.043	
s or s_k					
ω_γ	0.000	0.004	0.014	0.030	
ω_θ	0.022	0.101	0.110	0.143	
$\rho_{\gamma\theta}$	0.015	0.080	0.092	0.104	
	Model 3				
	MSE	Bias	SE	BSE	Cor
θ_p	0.527	0.266	0.628	0.725	0.716 (0.035)
γ_p	0.050	0.069	0.158	0.178	0.935 (0.014)
t_{0p}	0.116	0.118	0.160	0.183	0.882 (0.050)
a_i	0.001	0.010	0.028	0.034	0.998 (0.001)
b_i	0.057	0.041	0.230	0.675	0.931 (0.044)
τ_k	0.047	0.079	0.192	0.657	
s or s_k	0.002	0.005	0.034	0.030	
ω_γ	0.000	0.015	0.014	0.028	
ω_θ	0.030	0.124	0.122	0.133	
$\rho_{\gamma\theta}$	0.016	0.088	0.091	0.111	

MSE: Mean-squared error, SE: Standard Error of Point Estimates (Maximum A Posteriori), BSE: Bayesian Standard Error (posterior standard deviations averaged over repetitions). Cor: The Pearson correlation (first computed across items for a_i and b_i, or across persons for τ_p, γ_p, and θ_p, and then averaged over repetitions) with its standard deviation across repetitions in the following parentheses. For person and item parameters, MSE, bias, SE, and BSE were calculated for each parameter and then averaged over persons or over items, respectively. Statistics for decision thresholds and boundary scaling parameters were also first calculated for each parameter and then averaged.

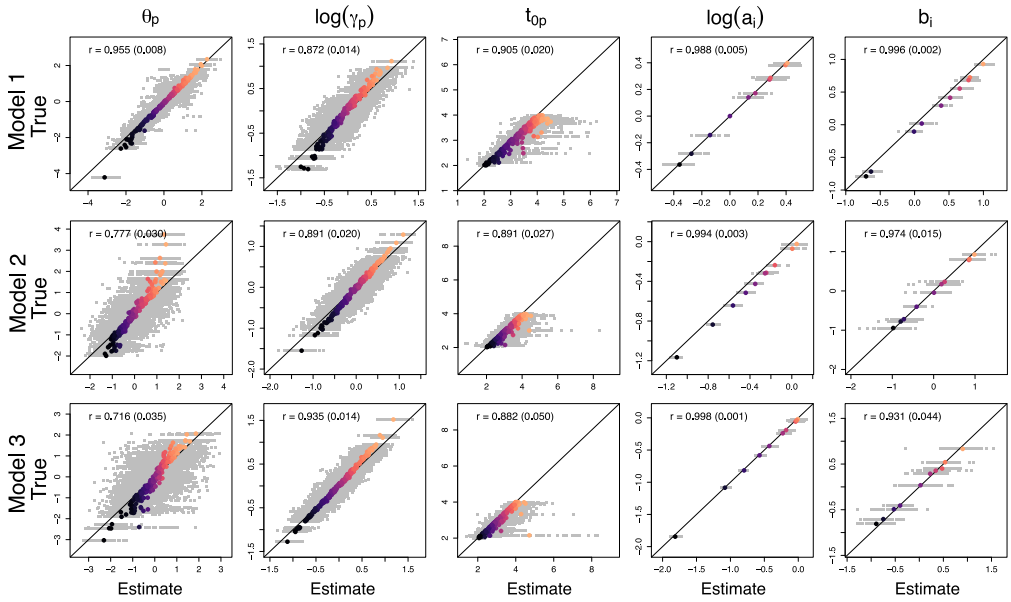

FIGURE 5.

Parameter Recovery of the Models (Person and Item Parameters). In each row, the five panels show the scatter plots of the Maximum A Posteriori (MAP) estimates of the parameters on the x-axis against the true parameter values on the y-axis. The parameter plotted in each panel is shown on the top: person-wise drift rate (θ_p), person-wise decision boundary (γ_p; log-transformed), person-wise nondecision time (τ_p), item time-pressure (a_i; log-transformed), and item strength (b_i), respectively, from left to right. Each row shows the recovery result of one of the three models as shown on the left. In each panel, gray squares indicate estimates and true parameter values for all 50 repetitions, while colored dots indicate their averages across repetitions (i.e., mean estimates). At the top-left side of each panel, the Pearson correlation (r) between the MAP estimates and true parameter values averaged across repetitions is shown, followed by its standard deviation across repetitions.

3.2. Recovery Results

With the posterior samples, we obtained the maximum A posterior (MAP) estimates of the parameters and compared them with the true parameter values to evaluate the parameter recovery. Table 2 presents the recovery results of Models 1–3 with statistics such as mean squared error (MSE), bias, standard error (SE) of the MAP estimates, Bayesian standard errors (BSE; posterior standard deviation) averaged over repetitions, and the Pearson correlation coefficients averaged over repetitions (with its standard error in the following parentheses). For person and item parameters, MSE, bias, SE, and BSE were calculated for each parameter and then averaged over persons or over items, respectively. Statistics for decision thresholds and boundary scaling parameters were also first calculated for each parameter and then averaged.

Figure 5 shows the recovery results of the three person-wise and two item-wise parameters. In each row of the figure, the five panels show the scatter plots of the MAP estimates of the parameters on the x-axis against the true parameter values on the y-axis. The parameter plotted in each panel is shown on the top: person-wise drift rate (θ_p), person-wise decision boundary (γ_p; log-transformed), person-wise nondecision time (τ_p), item time-pressure (a_i; log-transformed), and item strength (b_i), respectively, from left to right. Each row shows the recovery results of one of the three models as shown on the left. In each panel, gray squares indicate estimates and true parameter values for all 50 repetitions, while colored dots indicate their averages across repetitions (i.e., mean estimates). At the top-left side of each panel, the Pearson correlation (r) between the

MAP estimates and true parameter values averaged across repetitions is shown, followed by its standard deviation across repetitions (the same values as in Table 2).

Generally speaking, the results show the consistency between the estimates and the true values. Mostly, there was no noticeable bias in the estimation as scatter plots did not show large deviations from the diagonal line. Some exceptions include that person-wise drift rate estimates for Models 2 and 3 and (log-transformed) person-wise decision boundary estimates showed small shrinkage effects when their values were large. Also, although person-wise nondecision times were generally well estimated, they were overestimated in a couple of repetitions and for some specific simulated respondents (e.g., one respondent in Model 3 with $t_{0p} = 2.154$). Conceptually, the person-wise min RT is the upper bound of the corresponding person-wise nondecision time parameter. When a respondent has a very large boundary separation and there are only a few items, it is possible that RTs of this respondent could be much longer than the nondecision time. In this case, it could be hard to obtain a precise min RT value to sufficiently constrain the nondecision time parameter.[3]

The average Pearson correlations were high for most of the parameters for all three models. One exception was for the person-wise drift rate estimates in Model 3, which showed some spread with $r = 0.716$ on average across repetitions. Our interpretation is that the lower consistency in the person-wise drift rates occurs because the model has to determine $M = 5$ drift rates for accumulators with three sets of parameters (the person-wise drift rates θ_p, item strength parameters b_i, and response thresholds τ_k) and θ_p to be estimated with $I = 10$ item responses for each person. In contrast to θ_p, the estimation of b_i was accurate with little bias and fairly precise, which would be due to the larger number of persons $P = 200$ compared to the number of items assumed in the simulation study.

Table 3 shows the recovery results of the decision thresholds and boundary scaling parameters in the three models. The estimates of the response threshold parameters $\tau_1, ..., \tau_4$ were reasonably close to the true values for all three models. Boundary scaling parameters for Models 1 and 3 were all estimated well, close to their true values. Taken together, the results show that all three models can recover their parameters reasonably well even when a relatively small sample ($P = 200$ and $I = 10$) was used.

3.3. Cross Recovery

For a more thorough investigation of the models, we further conducted a cross recovery study. The main idea is to generate data from one model and fit all three models. In this way, we can obtain parameter estimates of all three models from the same data set and study the relationship among the parameters. Despite the structural differences among the models, it can be expected that the parameter estimates (e.g., estimates of the parameter b_i) would show some associations across the models because they play similar roles in all three models (e.g., item strength to define drift rates related to item i). Also, by comparing the model fits, we can check whether we can distinguish the models and whether any of the models tends to overfit data. If the best-fitting model is the data-generating model, it would be safe to claim that there is no overfitting model and all three models can be distinguished in terms of their fits to data. In contrast, if one model fits the data better than the data-generating model, it says the model with a better fit is likely to overfit data.

[3]The simulated respondent in the Model 3 result, mentioned in the text, has $\log(\gamma_p) = 1.520$, which is the largest in the simulation study. The mean and SD of data-generating $\log(\gamma_p)$ values for Model 3 were -0.002 and 0.498, respectively, and the second largest value was 1.162. Having more items can better constrain nondecision time parameters (Kang et al., 2022b). In particular, including an item with a highly strong inclination (i.e., the one with a positively or negatively highly large b_i) could be helpful as an RT for this item could be closer to the minimum RT of the respondent mostly spent for nondecision processes only.

TABLE 3.
Parameter recovery of the models (response threshold parameters and boundary scaling parameters).

	Parameters	τ_1	τ_2	τ_3	τ_4	s	s_1	s_2	s_3	s_4	s_5
Model 1	True	−1.500	−0.500	0.500	1.500	1.500					
	Estimate	−1.498	−0.493	0.502	1.514	1.425					
	SE	(0.063)	(0.047)	(0.050)	(0.061)	(0.076)					
	BSE	(0.059)	(0.049)	(0.049)	(0.060)	(0.085)					
	MSE	0.004	0.002	0.002	0.004	0.011					
Model 2	True	−1.200	−0.500	0.500	1.200						
	Estimate	−1.220	−0.521	0.493	1.197						
	SE	(0.029)	(0.046)	(0.052)	(0.032)						
	BSE	(0.035)	(0.051)	(0.052)	(0.035)						
	MSE	0.001	0.003	0.003	0.001						
Model 3	True	−1.500	−0.500	0.500	1.500		0.900	1.250	1*	1.250	0.900
	Estimate	−1.546	−0.503	0.601	1.667		0.908	1.248	1*	1.258	0.908
	SE	(0.186)	(0.176)	(0.161)	(0.244)		(0.036)	(0.054)	–	(0.048)	(0.031)
	BSE	(0.660)	(0.654)	(0.654)	(0.652)		(0.029)	(0.044)	–	(0.045)	(0.029)
	MSE	0.036	0.030	0.036	0.086		0.051	0.065	–	0.067	0.050

τ_1, ..., τ_4: response threshold parameters, s: boundary scaling parameter for Model 1, s_1, ..., s_5: boundary scaling parameters for Model 3. The third boundary scaling parameter s_3 in Model 3 was fixed to 1 to establish identifiability. Standard error of estimates is shown in the parentheses. SE: Standard error of point estimates (Maximum A Posteriori), BSE: Bayesian standard error (posterior standard deviations averaged over repetitions), MSE: Mean squared error.

To examine the cross recovery of the models, we fitted all three models to the first 25 synthetic datasets that were generated from each model and used in the parameter recovery simulations. The model fitting procedure was the same as that in the recovery study. Then, we obtained the MAP estimates for all model parameters and all models. For model evaluation, we used the modified Akaike information criterion (mAIC) and the modified Bayesian information criterion (mBIC) that Bolsinova and colleagues proposed (Bolsinova et al., 2017a; Bolsinova & Molenaar, 2018, 2019). This choice was motivated by earlier findings that some measures such as deviation information criterion (DIC; Spiegelhalter, Best, Carlin, & Van Der Linde, 2002) can perform suboptimally (i.e., favoring other models over the data-generating model)[4] for highly nonlinear models and mixture models (Molenaar & De Boeck, 2018), which led Bolsinova and colleagues to use the modified information criteria for joint models of responses and RTs estimated with a Bayesian method. These criteria are *modified* versions of the conventional AIC (Akaike, 1974) and BIC (Schwarz, 1978) in the sense that they are calculated with the -2 log-likelihood (-2LL) evaluated at the posterior means of the model parameters. In our model evaluation, we used the MAP estimates instead of the posterior means because some model parameters have skewed posterior distributions (e.g., person-wise nondecision time t_{0p}), and thus, the posterior means may not be the best summary statistics for them.

[4]This was also the case in our simulation study. For example, when Model 2 was the data-generating model, Model 3 beat Model 2 in 3 out of 25 repetitions when we judged based on DIC. Also, there are at least two ways to compute the effective number of parameters and DIC (Gelman et al., 2013) , which can produce different results. For example, when Model 3 was the data-generating model, one way of calculating the effective number of parameters and DIC (Equation 7.10 in Gelman et al.) predicted that Model 3 was the best-fitting model for all repetitions but another way (Equation 7.8 in Gelman et al.) predicted that Model 1 was the best-fitting model for all repetitions. Thus, we were not able to obtain consistent conclusions for the proposed models with DIC.

TABLE 4.
Cross recovery results.

	Model 1			Model 2			Model 3		
	−2LL	mAIC	mBIC	−2LL	mAIC	mBIC	−2LL	mAIC	mBIC
Model 1	**9067.3**	**10321.4**	**12389.4**	9977.3	11231.3	13299.3	10688.8	11942.8	14010.9
Model 2	10612.6	11866.6	13934.6	**9556.1**	**10810.1**	**12878.2**	9392.1	10646.1	12714.1
Model 3	10692.0	11954.1	14035.3	9770.0	11032.0	13113.3	**8776.4**	**10038.4**	**12119.7**

The three sections of the table present the model comparison results with the models denoted in the leading row of the table as the data-generating models. In each section, the three columns correspond to the model fit indices (−2LL, mAIC, and mBIC) as denoted in the second row and the three rows correspond to the model fitted to the synthetic data as denoted in the leading column of the table. Each cell shows the averaged fit index values across 25 repetitions. In all 25 repetitions, the data-generating model was the best-fitting model. −2LL: Log-likelihood multiplied by −2. mAIC: modified Akaike information criterion. mBIC: modified Bayesian information criterion.
Bold values indicate represents the best values.

Because wrong models (i.e., models not used in data generation) were intentionally fitted to data generated from a different model in a cross recovery simulation, there were more difficulties in model fitting such as convergence issues. In most cases, these were resolved by simply re-fitting the models with different initial values or adjusting some tuning parameters.[5] However, Model 2 was not able to achieve convergence when it was fitted to the synthetic datasets generated from Model 3. This could be attributed to the structural difference between the two models, implying that they could be distinguished (which is the purpose of the cross recovery study). To further validate the discriminability between the models for this case, we picked up one chain from the fitting result of Model 2 that produced the best -2LL value and used its MAP estimates in the model comparison. This approach favors Model 2 and challenges more the data-generating model, which is Model 3.

Table 4 shows the model fit comparison. The three sections of the table present the model comparison results with the models denoted in the leading row of the table as the data-generating models. In each section, the three columns correspond to the model fit indices (-2LL, mAIC, and mBIC) as denoted in the second row and the three rows correspond to the model fitted to the synthetic data as denoted in the leading column of the table. Each cell shows the averaged fit index values across 25 repetitions. Although not shown, all three criteria predicted that the data-generating model was the best-fitting model in all 25 repetitions. This shows that the models can be distinguished in terms of model fits and there is no overfitting model.

Figures S1–S3 in the supplementary material show the scatter plots of the person-wise and item-wise parameter estimates from the cross recovery study, across different models. These estimates were obtained from the model fits to the same (synthetic) data. In general, the figures show that the parameter estimates are positively associated across the three models as expected, although the strength of the association can be weaker or stronger depending on the data-generating models and parameters. A more detailed description of the results can be found in Section S2 in the supplementary material.

4. Empirical Applications

Provided the reasonable parameter recovery results, we examined and compared the proposed psychometric process models with empirical data. For this purpose, we fitted the three models

[5] For example, `adapt_delta` and `max_treedepth` in **Stan**.

to the response and RT data from the performance motivation questionnaire (Hermans, 1968; Hermans et al., 1972; Modick, 1977) adapted for chess players (van der Maas & Wagenmakers, 2005). The dataset is a part of a larger dataset, which includes more items on the chess performance, knowledge test, memory test, etc. The motivation questionnaire aims to measure three traits related to the performance motivation of the chess players, namely desire to win (DTW; the motive to achieve better performance), negative fear of failure (NFF; anxiety due to fear of failing a task that has a debilitating effect on performance), and positive fear of failure (PFF; anxiety due to fear of failing a task that has a facilitating effect on performance). The questionnaire consists of 30 items, and each trait is measured by 10 items. Items are short sentences about the trait being measured, for example, *"Overpowering my opponent makes me feel good"* (DTW), *"In a difficult game against a very strong player, I often feel discouraged"* (NFF), and *"When I notice that I am worried about my game, I stimulate myself to concentrate better"* (PFF). Each item response was measured by a 5-point Likert scale and the response options were *fully disagree*, *disagree*, *neutral*, *agree*, and *fully agree*. Each person was presented with each item at a time and responded to the item by choosing one of the five response options, within the item-wise time limit of 10 seconds. The full dataset and item sentences are available online at van der Maas' webpage.

The dataset includes 259 respondents, but eight respondents do not have responses to the motivation items. In addition, one respondent has RTs shorter than 0.5 s for half of the 30 items and RTs shorter than 1.0 s for 70% of the items, implying that the respondents made fast random responses for most of the items. Thus, we excluded these 9 respondents in our main analysis, leaving $P = 250$ persons. Considering the length of the item sentences, we also assumed that RTs shorter than 2 s would be too short to reasonably process item sentences. Accordingly, we excluded responses with such short RTs, removing about 0.4% of data responses and RTs.

We fitted the models with the same sampling methods and prior specifications used in our simulation study. When fitting the models, we treated item sets measuring different traits separately so that we can examine the three models on the basis of three datasets. Thus, for each trait (DTW, NFF, or PFF), we fitted the three models to the dataset with $P = 250$ persons and $I = 10$ items and we repeated this for three item sets. We assessed convergence using the Gelman–Rubin convergence diagnostic and consistency of posterior densities across chains, and found no issue in convergence (Section S3 in the supplementary material).

We evaluated the models with relative model fit indices such as mAIC and mBIC to see which model provides a better account for the data. However, we put more emphasis on the absolute model fit (whether the model predictions match the data well) in the study of the psychological measurement processes. The central idea is to falsify a model with a severe misfit because a theory of the measurement processes assumed in such a model cannot be an adequate account for the true processes. This has been a conventional way to test mathematical models and compare different theories of cognitive processes in perceptual and cognitive decision-making and we apply the same strategy to psychometrics data. For binary choice data, the absolute model fit can be evaluated by contrasting the data-based response proportions and RT distributions (usually summarized by RT quantiles) to the model predictions (Brown & Heathcote, 2008; Kang & Ratcliff, 2020; Ratcliff, 2002; Ratcliff et al., 2003; Ratcliff & McKoon, 2008).

We conducted a similar absolute fit analysis using posterior predictive samples (Gelman et al., 2013). For each of the model fits to each item set, we first randomly selected 5000 samples of model parameters from the joint posterior samples (i.e., Bayesian samples obtained with Stan). Then, with each of the parameter samples, we generated a single sample of response and RT, resulting in 5,000 posterior predictive samples of responses and RTs. To account for the effect of the item-wise time limit of 10 seconds, posterior predictive samples with RTs longer than this time limit were not accepted and resampled. We repeated this procedure for all three models and for all three item sets and obtained the model predictions of overall response proportions (the proportion of each response option across all persons and items), overall RT distributions, item-

TABLE 5.
Relative model fits.

	Desire to Win			Negative Fear of Failure			Positive Fear of Failure		
	−2LL	mAIC	mBIC	−2LL	mAIC	mBIC	−2LL	mAIC	mBIC
Model 1	14858.3	16410.3	19143.0	**13784.7**	**15336.7**	**18069.4**	14300.5	15852.5	18585.2
Model 2	**14668.7**	**16220.7**	**18953.4**	14256.3	15808.3	18541.0	14512.7	16064.7	18797.4
Model 3	14687.9	16245.9	18989.1	13820.9	15378.9	18122.1	**14253.6**	**15811.6**	**18554.8**

−2LL: Log-likelihood multiplied by −2. mAIC: modified Akaike information criterion. mBIC: modified Bayesian information criterion.
Bold values indicate represents the best values.

wise response proportions (the proportion of each response option across all persons, but obtained separately for each item), and item-wise RT distributions. We contrasted these predictions with the data counterpart and checked whether the model predictions well cover behavioral patterns in the data.

In addition, the empirical fitting results can be used to study how the main model parameters relate to the descriptive measures such as (person-wise or item-wise) median RTs and mean responses. This analysis could show the nature of individual differences in responses and RTs, for example, whether the observed individual differences in responses and RTs are due to information processing rates (person-wise drift rates), amount of evidence required for response (person-wise decision boundaries), or differences in nondecision processes (nondecision time parameters). Also, it can be studied how similar sets of parameters (e.g., person-wise drift rates) are related across the three models, as done in the cross recovery simulation in Sect. 3.3 and Section S2 in the supplementary material. A related analysis was conducted and is presented in Section S4 in our supplementary material.

4.1. Relative Model Fits

Table 5 presents -2LL (evaluated with the MAP estimates of the model parameters), mAIC, and mBIC values of the three models for model comparison. The results show that different item sets prefer different models. The best-fitting model was Model 2 for the DTW item set, Model 1 for the NFF item set, and Model 3 for the PFF item set. Therefore, the relative model fits did not provide decisive evidence for a single best-fitting model and theory of the measurement processes.

The model comparison results may be interpreted as that different latent traits are associated with different measurement processes that are represented by their corresponding best-fitting models. An alternative possibility is that, because the three latent traits have something in common in that they are associated with motivation, it could be that their differences in underlying psychological processes are not sufficient to distinguish the proposed models in terms of relative model fits. This would correspond to the similar absolute fit results of the three models, which we present in the next subsection.

4.2. Absolute Model Fits: Posterior Predictive Checking

For the next step of our model evaluation, we examined the absolute model fits of the three process models with their posterior predictive samples. Table 6 shows the overall response proportions obtained by aggregating responses from all person-item pairs. The table presents the results for all three item sets in the order of DTW, NFF, and PFF. For each item set, the top row (bolded) shows the data response proportions (for the five response options shown in the second row of the table), while the last three rows present the differences between the data response proportions

TABLE 6.
Response proportions over all persons and items

Response options	Desire to Win				
	1	2	3	4	5
Data	**0.092**	**0.241**	**0.142**	**0.351**	**0.174**
Model 1	0.001	−0.001	−0.002	0.000	0.002
Model 2	−0.003	0.006	0.005	0.001	−0.009
Model 3	−0.012	0.013	0.013	−0.001	−0.013
	Negative Fear of Failure				
	1	2	3	4	5
Data	**0.030**	**0.141**	**0.190**	**0.482**	**0.157**
Model 1	0.000	0.000	0.000	−0.001	0.001
Model 2	−0.001	0.001	0.008	−0.003	−0.005
Model 3	−0.006	0.008	0.010	−0.001	−0.011
	Positive Fear of Failure				
Response Options	1	2	3	4	5
Data	**0.035**	**0.229**	**0.255**	**0.383**	**0.097**
Model 1	−0.001	0.001	−0.001	−0.001	0.002
Model 2	−0.002	−0.003	0.008	0.000	−0.003
Model 3	−0.009	0.010	0.008	0.004	−0.013

For each of the three item sets examined, the top row (bolded) shows the data response proportions (for the five response options shown in the second row of the table) while the last three rows present the differences between the data response proportions and the corresponding model predictions from the three proposed models. Predictions were produced based on posterior predictive samples of responses from all persons and items.

and the corresponding model predictions from the three proposed models. The response option $(1, \ldots, 5)$ is shown in the leading row of the table.

For the first item set measuring DTW, predictions from Model 2 showed the best consistency with the data-based response proportions. However, for the other two item sets, Model 1 performed the best in reproducing the overall response proportions. Model 3 did not perform better than the other models in terms of the overall response proportions, but the differences were not that large.

Figure 6 shows the overall RT distributions obtained by aggregating RTs from all person-item pairs. The leftmost, middle, and rightmost panels show the RT distributions for the three item sets measuring DTW, NFF, and PFF, respectively. In each panel, the gray histogram shows the data-based distribution while green solid, yellow dashed, and red dotted curves show the predicted densities from Models 1–3, respectively. Overall, model predictions match the data RT distributions well without a large misfit. An interesting finding is that Models 2 and 3 produced very similar (but not perfectly the same) predictions for the overall RT distributions. Model 1 also produced similar density predictions, but with slight differences.

For a more thorough investigation of the model performance, Figs. 7 and 8 present item-wise data summary statistics and the corresponding model predictions. These were obtained by aggregating responses over persons but separately by item. Thus, these figures show data variations across persons in response proportions and RT distributions per item and how well the models can account for these. Figure 7 shows item-wise response proportions. Because we have $M \times I$ values, i.e., $5 \times 10 = 50$ values for each item set, we present these proportions as scatter plots in Fig. 7 instead of listing their values. The three panels in Fig. 7 show the results for DTW, NFF, and PFF, respectively. In each panel, the data-based item-wise response proportions are plotted on the x-axis against the model predictions on the y-axis. The integers, $1, \ldots, 5$, represent item-

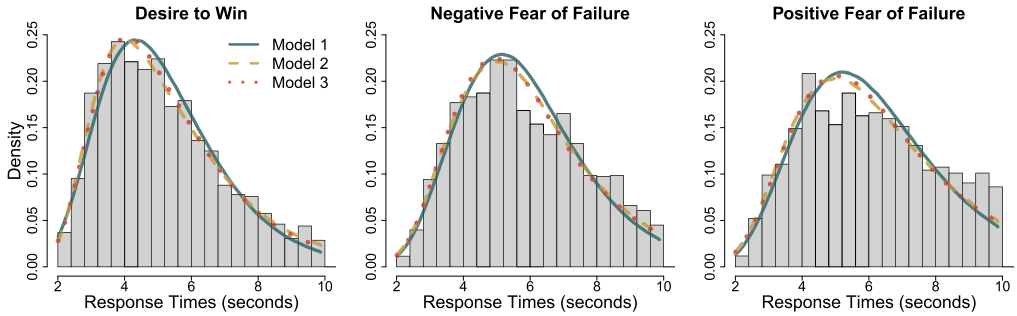

FIGURE 6.
RT Distributions over All Persons and Items. In each panel corresponding to one of the three item sets examined (as shown on top of each panel), the histogram shows the data RT distribution while the green-solid, yellow-dashed, and red-dotted lines the predicted RT distributions from Models 1, 2, and 3, respectively. Predictions were produced based on posterior predictive samples of RTs from all persons and items (Color figure online).

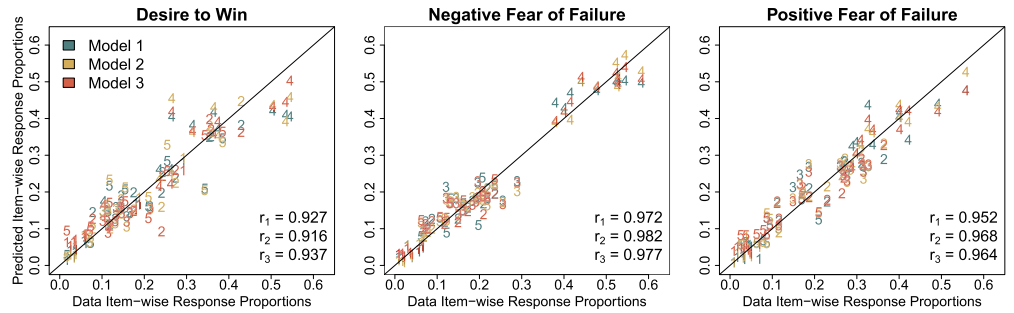

FIGURE 7.
Scatter Plots of Item-wise Response Proportions. In each panel corresponding to one of the three item sets examined (as shown on top of each panel), data item-wise response proportions (i.e., response proportions for the five response options computed separately for each item) are plotted on the x-axis against their corresponding predicted item-wise response proportions on the y-axis. The numbers, 1, ..., 5, indicate the response proportions of the corresponding response options, for all items. The green, yellow, and red numbers indicate results from Models 1, 2, and 3, as shown by the legend on the top-left side of the left panel. Predictions were produced based on posterior predictive samples of responses from all persons, but separately for each item (Color figure online).

wise response proportions of the response options $1, \ldots, 5$. Each panel shows the predictions from all three models simultaneously: green, yellow, and red numbers correspond to Models 1, 2, and 3, respectively. For all three item sets and all three models, model predictions are generally consistent with the data. The Pearson correlations between data-based and predicted item-wise response proportions are shown at the bottom-right side of each panel, r_1, r_2, and r_3 for Models 1, 2, and 3, respectively. Careful attention should be paid to the correlations for the NFF item sets because only response option 4 (*agree*) has high item-wise response proportions while the other response options have low proportions, making the correlation value inappropriately higher. However, there is no systematic bias for all model predictions for all item sets, assuring us that the models perform well in capturing item-wise proportions of the ordinal responses.

Figure 8 shows the item-wise RT distributions (i.e., RT distributions over respondents for each of the items). There are three column-like panels for three models as denoted on the top of the figure. Each panel presents data-based and predicted item-wise RT distributions of all 30 items in the three item sets vertically. Item numbers are shown on the left, and the item sets are shown on the right. The x-axis of each panel represents RTs in quantiles and for each item, there are five

x's indicating 10% (black), 30% (red), 50% (green), 70% (blue), and 90% (skyblue) RT quantiles from left to right. The gray circles indicate the model-predicted RT quantiles. Each circle has a colored vertical bar at its center and circles with different colors indicate different (predicted) RT quantiles. Match/mismatch between x's and their corresponding circles (i.e., those with the same color bar as x's) show the absolute fit of the models in terms of item-wise RT quantiles. For example, a gray circle with a black vertical bar is a model-predicted 10% quantile and so, if a model performs well, it should be close to the black 'x' for the same item. Item-wise RT distributions are presented by these item-wise RT quantiles from which the whole distributions can be reproduced. For example, the gray histogram at the top-left side of the figure shows the data RT distribution for the first item in the item set measuring DTW. Thus, the figure shows RT distributions and model predictions for all items in our consideration. In addition to this figure, scatter plots of item-wise RT quantiles (similar to Fig. 7 but generated with item-wise RT quantiles) are provided in Section S5 in the supplementary material (Figure S8). Also, the Pearson correlation between data-based item-wise RT quantiles and the corresponding model predictions were higher than 0.99 for all models and item sets.

Visual inspection of Fig. 8 led us to conclude that all three models are capable of accounting for item-level RT distributions. For most of the items, predictions (circles) match the corresponding data points (x's) well without a large discrepancy. There was no big difference across the three models. Instead, the performance showed some differences across item sets in that consistency between data and prediction was the best in the NFF item set for all three models. Not all predicted RT quantiles perfectly overlap the data. For example, predicted 90% RT quantiles are generally slightly shorter than their data-based counterparts. This can be attributed to our resampling protocol used in posterior predictive checking to account for the item-wise time limit of 10 s. It seems that resampling a posterior sample of response and RT when a sampled RT is larger than 10 s is too restrictive to accurately describe implicit (respondents attempt to finish each item more quickly than when no time limit is imposed) and explicit (all RTs are recorded as less than 10 s) effects of the time limit. However, posterior predictive samples without any handling of the imposed time limit produced 90% RT quantiles much longer than 10 seconds due to sampling variability and the right-skewness of predicted RT distributions. Despite this limitation, the misfits at the RT tails are not that large. For some items such as items 1 and 3 for PFF, predictions from all three models show large misfits as predicted RTs are generally shorter than their data-based counterparts. However, this is not a dominant pattern over all items and the generally good consistency between data and predictions provides sufficient evidence to conclude that all three models produce good descriptions of the observed RT trends.

5. Discussion

In this article, we proposed a modeling framework to build psychometric process models for the measurement of latent personality/attitude traits. We considered ordinal responses on the Likert scale, as practically done in the field, but jointly with RTs, which has not been widely done yet (particularly for ordinal responses). We showed how substantive cognitive theories on decision-making processes can be integrated with psychometric modeling and we produced three process models based on different theories of decision-making. The resulting models were equipped with different theoretical representations of measurement processes and different empirical predictions of responses and RTs. We examined these models with three different (but all related to motivation) latent traits and were to reject a model with a severe discrepancy between the data and the model prediction. However, the results showed that all three models performed similarly well in our absolute fit test. Therefore, we tentatively conclude that all three models have the potential to be a reasonable account of the cognitive measurement processes underlying ordinal responses and

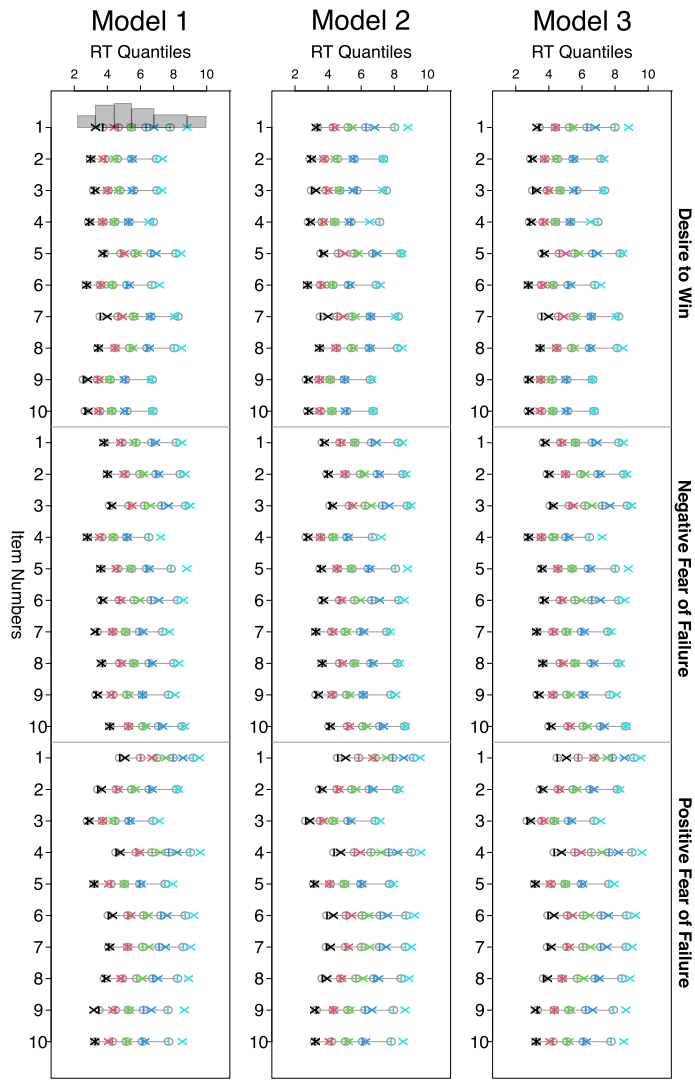

FIGURE 8.
Item-Quantile Plot. Each column shows data-prediction comparison of item-wise RT quantiles for all items investigated, for the model denoted on top of the column. Item numbers are shown on the left and the item sets are shown on the right. The x-axis of each panel represents RTs in quantile and for each item and there are five x's indicating 10% (black), 30% (red), 50% (green), 70% (blue), and 90% (skyblue) RT quantiles from left to right. The gray circles indicate the model-predicted RT quantiles. Each circle has a colored vertical bar at its center and circles with different colors indicate different (predicted) RT quantiles. Match/mismatch between x's and their corresponding circles show the absolute fit of the models in terms of item-wise RT quantiles (Color figure online).

RTs. At the same time, relative model fit indices showed that item sets measuring different latent traits favored different process models. A potential implication of this result is that measurement of different latent traits may be driven by cognitive processes with different structures. However, this conclusion is not yet settled, particularly because all three latent traits examined are related to motivation. Thus, there could be only minor differences in their underlying processes, which could be an explanation for why the absolute fit results are similar across traits. At this point, we

see the models that we presented as a first step in the process of model development and testing, and more comprehensive evaluations and further model development are needed.

Process models must be evaluated not only based on empirical data but also from the theoretical perspective. The central assumption we relied on in our modeling is that the sequential sampling framework and evidence accumulation can account for cognitive processes underlying the measurement of latent traits. It is hard to conceive evidence accumulation as a completely appropriate description of complex problem-solving processes driven by latent abilities (although it may provide an approximate account; Kang et al., 2022a) because complex tests tend to require multiple heterogeneous processes while evidence accumulation is a simple single decision process. In contrast, we claim that the psychological processes related to the measurement of latent traits by ordinal responses (e.g., Likert scale) and RTs with short item sentences are relatively simple and evidence accumulation is well applicable. When responding to an item sentence describing, for example, a personality trait, respondents need to collect information to determine whether the sentence appropriately characterizes their personalities. In this process, the information may refer to a piece of memory from the respondents' individual history related to the described personality and a match/mismatch between the item sentence and the respondents' self-identified personality. Although this assumption is new and needs to be further investigated, earlier studies on text/sentence reading skills and related inferences have shown that sequential sampling models such as the Ratcliff diffusion model (Ratcliff, 1978; Ratcliff & McKoon, 2008) can provide excellent accounts for data from reading tasks that are as complex as or even more complex than reading psychometric items to measure latent traits (McKoon & Ratcliff, 2016, 2017, 2018). These findings provide evidence supporting our assumption that reading and processing psychometric item sentences and inferring the degree of match between the sentences and personality/attitude traits can be accounted for by the sequential sampling framework and evidence accumulation.

Another fundamental assumption we made, by using the diffusion process in our modeling, is that within-process noise is the primary source of probabilistic and stochastic properties of the proposed models. This was shown by wiggly trajectories of evidence accumulation in Figs. 1, 2, and 3. Thus, the models account for distributions of responses and RTs by (1) person and item effects captured by differences in person-wise and item-wise parameters and (2) stochastic variability introduced by within-process noise. Thanks to the second component, the models can predict distributions of responses and RTs for each person-by-item pair. This within-process variability is a fundamental source of noise that has perceptual and cognitive plausibility. It explains why there is randomness and moment-to-moment fluctuation in our information processing. It also has neural plausibility because neurons produce random spikes even when there is no stimulus/item being processed.

In contrast, an earlier psychometric process model for ordinal responses and RTs proposed by Ranger and Kuhn (2018) assumes that there is no internal noise within the evidence accumulation process for each person-item pair. This model predicts that the trend of evidence accumulation is linear so that, at every time point in a single response process, exactly the same amount of evidence is accumulated during the same amount of time (as it is developed based on the LBA model by Brown & Heathcote 2008). Instead, the model assumes random variability across persons (but per item) in cognitive components, which corresponds to across-trial variability parameters in cognitive models (i.e., variability of cognitive components across multiple trials done by a single subject in a psychological experiment; e.g., Brown & Heathcote, 2008; Ratcliff & McKoon, 2008). For person p and item i, Ranger and Kuhn's model assumes two accumulators (as our Model 2) and determines a drift rate for accumulator j ($j = 1, 2$) via the log-linear model: $\log(v_{pij}) = c_{1ij} + c_{2ij} \cdot \theta_{pj} + \eta_{ij} \cdot e_{pij}$ where c_{1ij} and c_{2ij} are item-wise intercept and slope with respect to the person-wise latent trait θ_{pj}, respectively, η_{ij} is a standard deviation of random variation in drift rate, and e_{pij} is a realization of a standard normal random variable. The variation

in drift rate $\eta_{ij} \cdot e_{pij}$ is the only source of variability of the model. Given the parameters for person p and item i such as drift rate v_{pij} and a common decision boundary α_{pi}, the Ranger and Kuhn's model predicts RT *deterministically* as $\min_{j=1,2} \frac{\alpha_{pi}}{v_{pij}}$ [6].

Although drift rate v_{pij} in Ranger and Kuhn's model has a random component, it leaves a theoretical question if this variability can fully account for the variation in process and outcome variables; for a single pair of person p and item i, its response and RT are probabilistic (rather than deterministic) realizations with variability. This can only be explained by the internal noise of cognitive processes (which corresponds to within-trial noise in cognitive models). Also, it is important to note that, while across-trial variability components are important sources of noise in perceptual and cognitive decision-making data (e.g., Kang, Ratcliff, & Voskuilen, 2020; Ratcliff, Voskuilen, & McKoon, 2018), its counterpart in psychometric models (random variability across persons and items, or across persons but per item as in Ranger and Kuhn) captures *residual or conditional* dependency between responses and RTs in psychometric data (i.e., remaining associations between responses and RTs after controlling for person and item effects; Bolsinova, Tijmstra, Molenaar, & De Boeck, 2017b; Kang, De Boeck,& Partchev, Kang et al. (2022a,b)). Although it is important to find a model-based account for the unexplained dependency between responses and RTs to better understand response processes, it is hard to conceive this residual dependency as a fundamental source of noise in cognitive processes of psychometric measurement. Therefore, we chose to build our psychometric process models based on the diffusion process with internal noise in evidence accumulation rather than a simplified account by the LBA model.

The three models we examined have different theoretical characteristics and limitations. Model 1 has the simplest and most parsimonious structure among the three proposed models and was built based on established theories and models in psychometrics and cognitive psychology. However, it has a theoretical limitation in that accumulated evidence is not directly involved with the determination of responses and response proportions. Given person and item parameters, a response is predicted by the distribution of continuous latent response without the accumulated evidence. Instead, drift rate and decision boundary of the evidence accumulation process account for response proportions and capture the association between responses and RTs. Thus, Model 1 can be said to be an approximate or a 'pseudo' description of cognitive processes.

An important theoretical consideration for Model 2 is that the balance-of-evidence hypothesis was proposed for tasks in which a subject is asked to make a binary decision first and then make a confidence judgment (e.g., Baranski & Petrusic, 1998; Vickers, 1979). For example, in a word recognition task, a subject decides whether a presented word stimulus is in the previously studied word list or not, and then is asked to determine the confidence level of this binary decision. Although the absolute fit of Model 2 was reasonably good, this hypothesis may not be an adequate description of response behavior for the psychometric measurement in which persons respond to an item directly on the ordinal scale. A possible interpretation that can reconcile this gap is that respondents first make an internal decision of which side of the Likert scale (*agree* side or *disagree* side) is more appropriate for their personalities and then determine the degree of appropriateness. In this case, the validity of the assumption remains to be studied. Another consideration is that the balance-of-evidence models do not explain internal processes that make a confidence judgment after a binary decision as they only use a threshold method. These models also assume that a cognitive system such as a respondent in the psychometric measurement has direct access to the amount of accumulated evidence (Ratcliff & Starns, 2009) while they do not explain the time course of the processes that read out evidence difference exactly at the decision time. It has been shown that neural activity immediately begins to decay after the decision threshold is reached

[6] Ranger and Kuhn's model does not consider nondecision time, unlike our models.

(Ratcliff et al., 2007a) and it is unclear how the balance-of-evidence hypothesis can account for this finding.

The primary assumption of the multi-accumulator model (Model 3) is that cognitive processes employ as many accumulators as the number of response options given in the questionnaire. This implies that the latent structure of cognitive processes can largely differ as a function of the number of response options given. The assumption is an appropriate account for memory tasks with confidence judgments in which a subject has to press one of the M separate keys on a keyboard to express their decision on the M-point confidence scale. The assumption can be questioned in psychometric measurement particularly when the measurement is done with pencil and paper or with mouse-clicking on one of the response options. Some behavioral properties of neurons correspond to the assumption on the number of evidence accumulators. In general, neurons have their own response preferences and respond to the maximal degree when a presented stimulus matches their preference (Beck et al., 2008; Cowell et al., 2006; Jazayeri & Movshon, 2006). The level of activation decreases as the discrepancy between the stimulus and preference. Also, it has been shown that build-up neurons in Superior Colliculus (SC) show systematic changes in their activity as a function of the experimenter-imposed number of targets in a saccadic movement task (Basso & Wurtz, 1998). Thus, it makes sense to assume that neurons respond to different degrees to provided item response options and their summed activity defines the behavior of competing evidence accumulators (Ratcliff & Starns, 2013).

The proposed models were designed to provide parsimonious accounts, instead of complete and comprehensive descriptions, for psychological processes. To this end, we based our modeling on cognitive and psychometric theories introduced in Section 1. One can achieve a more thorough process model by introducing other psychometric factors, cognitive theories, and neurally plausible components. For example, response is subject to aberrant behavior such as rapid guessing and cheating and there could be hidden latent classes. Variation in response due to these factors can be accounted for by mixture modeling (DiTrapani et al., 2016; Lu et al., 2021; Wang & Xu, 2015; Wang et al., 2018). Also, random variability in cognitive components can be introduced as done in some previous psychometric process models for binary responses and RTs (Kang et al., 2022a,b), which allows a study of conditional dependency between the two outcome variables. Accumulated evidence may decay or leak as a function of the currently accumulated amount (as in the Ornstein–Uhlenbeck process; Smith, 2000; Usher & McClelland, 2001). We may also allow accumulators to inhibit their competitors (Usher & McClelland, 2001). When evidence in the accumulation process represents firing rates of neurons (which cannot be negative), we may need to constrain accumulated evidence to be nonnegative (Ratcliff et al., 2007b; Ratcliff & Starns, 2009). A spatially continuous evidence (in contrast to scalar-valued evidence) can be used, as in the spatially continuous diffusion model (SCDM; 2018), which can potentially provide a theoretically better account for the measurement processes with a single accumulator than our Model 1. Models with these additional components may provide a more adequate account of the cognitive processes of measurement. However, these extensions are rather challenging because resulting models do not have closed-form likelihoods (as the leaky competing accumulator model by Usher & McClelland, 2001). Fitting these models requires likelihood-free estimation methods such as approximate Bayesian computation (Turner & Sederberg, 2012; Turner & Van Zandt, 2012, 2014), probability density approximation (Turner & Sederberg, 2014), or likelihood approximation networks (Fengler et al., 2021).

A primary function of joint modeling of responses and RTs in psychometrics has been improving the precision of the measurement of latent attributes (abilities and traits) by combining information from both outcome variables (Bolsinova & Tijmstra, 2018; De Boeck & Jeon, 2019). We attempted to show that further theoretical improvement can be achieved by modeling individual-level measurement processes. In our modeling, latent variables were redefined as components in cognitive processes of measurement. Unlike traditional latent ability and speed factors, these

components have their own meaningfully interpretable referents; drift rate refers to the mean rate of evidence accumulation or the quality of information processing for decision-making and decision boundary refers to the amount of quantity of information required to respond to experimental stimuli or measurement items. Our model development also showed how a causal relationship between the variations in latent attributes and the variations in measurement outcomes can be explicitly described, with noise in evidence accumulation and individual differences in cognitive components. This modeling approach was in line with Borsboom and colleagues' suggested solution to the validity issue they raised (Borsboom et al., 2003, 2004). We hope that this psychometric process modeling approach can illustrate how cognitive theories of decision-making can be integrated with psychometric theories to provide principled accounts for individual-level measurement processes and further stimulate other novel attempts.

Code Availability The **Stan** codes to fit the proposed models can be found online at https://osf.io/76jb4/.

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

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Andrich, D. (1978). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement, 2*(4), 581–594. https://doi.org/10.1177/014662167800200413

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561–573. https://doi.org/10.1007/BF02293814

Baranski, J., & Petrusic, W. (1998). Probing the locus of confidence judgments: experiments on the time to determine confidence. *Journal of Experimental Psychology. Human Perception and Performance, 24*(3), 929–945.

Basso, M. A., & Wurtz, R. H. (1998). Modulation of neuronal activity in superior colliculus by changes in target probability. *Journal of Neuroscience, 18*(18), 7519–7534. https://doi.org/10.1523/JNEUROSCI.18-18-07519.1998

Beck, J. M., Ma, W. J., Kiani, R., Hanks, T., Churchland, A. K., Roitman, J., Shadlen, M. N., Latham, P. E., & Pouget, A. (2008). Probabilistic population codes for Bayesian decision making. *Neuron, 60*(6), 1142–1152. https://doi.org/10.1016/j.neuron.2008.09.021

Bock, R. D., & Jones, L. V. (1968). *The measurement and prediction of judgment and choice*. Holden-Day.

Bollen, K., & Barb, K. H. (1981). Pearson's r and coarsely categorized measures. *American Sociological Review, 46*(2), 232–239.

Bolsinova, M., De Boeck, P., & Tijmstra, J. (2017). Modelling conditional dependence between response and accuracy. *Psychometrika, 82*(4), 1126–1148. https://doi.org/10.1007/s11336-016-9537-6

Bolsinova, M., & Molenaar, D. (2018). Modeling nonlinear conditional dependence between response time and accuracy. *Frontiers in Psychology, 9*(1525), 1–12. https://doi.org/10.3389/fpsyg.2018.01525

Bolsinova, M., & Molenaar, D. (2019). Nonlinear indicator-level moderation in latent variable models. *Multivariate Behavioral Research, 54*(1), 62–84. https://doi.org/10.1080/00273171.2018.1486174

Bolsinova, M., & Tijmstra, J. (2018). Improving precision of ability estimation: Getting more from response times. *British Journal of Mathematical and Statistical Psychology, 71*(1), 13–38. https://doi.org/10.1111/bmsp.12104

Bolsinova, M., Tijmstra, J., & Molenaar, D. (2017). Response moderation models for conditional dependence between response time and response accuracy. *British Journal of Mathematical and Statistical Psychology, 70*, 257–279. https://doi.org/10.1111/bmsp.12076

Bolsinova, M., Tijmstra, J., Molenaar, D., & De Boeck, P. (2017). Conditional dependence between response time and accuracy: An overview of its possible sources and directions for distinguishing between them. *Frontiers in Psychology, 8*, 202. https://doi.org/10.3389/fpsyg.2017.00202

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110*(2), 203–219. https://doi.org/10.1037/0033-295X.110.2.203

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 1114*, 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

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

Cowell, R. A., Bussey, T. J., & Saksida, L. M. (2006). Why does brain damage impair memory? a connectionist model of object recognition memory in perirhinal cortex. *Journal of Neuroscience, 26*(47), 12186–12197. https://doi.org/10.1523/JNEUROSCI.2818-06.2006

Cox, D., & Miller, H. D. (1965). *The theory of stochastic processes*. Methuen.

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

DiTrapani, J., Jeon, M., De Boeck, P., & Partchev, I. (2016). Attempting to differentiate fast and slow intelligence: Using generalized item response trees to examine the role of speed on intelligence tests. *Intelligence, 56*, 82–92. https://doi.org/10.1016/j.intell.2016.02.012

Embretson, S., & Reise, S. (2000). *Item response theory for psychologists*. L. Erlbaum Associates.

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

Ferrando, P. J., & Lorenzo-Seva, U. (2007). An item response theory model for incorporating response time data in binary personality items. *Applied Psychological Measurement, 316*, 525–543. https://doi.org/10.1177/0146621606295197

Ferrando, P. J., & Lorenzo-Seva, U. (2007). A measurement model for likert responses that incorporates response time. *Multivariate Behavioral Research, 424*, 675–706. https://doi.org/10.1080/00273170701710247

Festinger, L. (1943). Studies in decision: I. decision-time, relative frequency of judgment and subjective confidence as related to physical stimulus difference. *Journal of Experimental Psychology, 32*(4), 291–306.

Festinger, L. (1943). Studies in decision. ii. an empirical test of a quantitative theory of decision. *Journal of Experimental Psychology, 32*(5), 411–423.

Forstmann, B., Ratcliff, R., & Wagenmakers, E.-J. (2016). Sequential sampling models in cognitive neuroscience: Advantages, applications, and extensions. *Annual Review of Psychology, 67*(1), 641–666. https://doi.org/10.1146/annurev-psych-122414-033645

Gelman, A. (1996). Inference and monitoring convergence. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain monte Carlo in practice* (pp. 131–143). CRC Press.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). CRC Press.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Wiley.

Hermans, H. J. M. (1968). *Handleiding bij de prestatie motivatie test [manual of the performance motivation test]*. Harcourt Assessment B.V.

Hermans, H. J. M., Ter Laak, J. J. F., & Maes, P. C. J. M. (1972). Achievement motivation and fear of failure in family and school. *Developmental Psychology, 6*, 520–528.

Jazayeri, M., & Movshon, J. (2006). Optimal representation of sensory information by neural populations. *Nature Neuroscience, 9*, 690–696. https://doi.org/10.1038/nn1691

Kang, I., De Boeck, P., & Partchev, I. (2022). A randomness perspective on intelligence processes. *Intelligence, 91*, 101632. https://doi.org/10.1016/j.intell.2022.101632

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, Advance Online Publication.* https://doi.org/10.1007/s11336-021-09819-5

Kang, I., & Ratcliff, R. (2020). Modeling the interaction of numerosity and perceptual variables with the diffusion model. *Cognitive Psychology, 120*, 1–42. https://doi.org/10.1016/j.cogpsych.2020.101288

Kang, I., Ratcliff, R., & Voskuilen, C. (2020). A note on decomposition of sources of variability in perceptual decision-making. *Journal of Mathematical Psychology, 98*, 102431. https://doi.org/10.1016/j.jmp.2020.102431

Kuiper, N. A. (1981). Convergent evidence for the self as a prototype: The "inverted-u rt effect" for self and other judgments. *Personality and Social Psychology Bulletin, 7*(3), 438–443. https://doi.org/10.1177/014616728173012

Kuncel, R. B. (1973). Response processes and relative location of subject and item. *Educational and Psychological Measurement, 333*, 545–563. https://doi.org/10.1177/001316447303300302

Lu, J., Wang, C., & Shi, N. (2021). A mixture response time process model for aberrant behaviors and item nonresponses. *Multivariate Behavioral Research, Advance Online Publication.* https://doi.org/10.1080/00273171.2021.1948815.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. Oxford University Publication.

Macmillan, N. A., & Creelman, C. D. (1966). *Detection theory: A user's guide*. Taylor & Francis.

McKoon, G., & Ratcliff, R. (2016). Adults with poor reading skills: How lexical knowledge interacts with scores on standardized reading comprehension tests. *Cognition, 146*, 453–469. https://doi.org/10.1016/j.cognition.2015.10.009

McKoon, G., & Ratcliff, R. (2017). Adults with poor reading skills and the inferences they make during reading. *Scientific Studies of Reading, 21*(4), 292–309. https://doi.org/10.1080/10888438.2017.1287188

McKoon, G., & Ratcliff, R. (2018). Adults with poor reading skills, older adults, and college students: The meanings they understand during reading using a diffusion model analysis. *Journal of Memory and Language, 102*, 115–129. https://doi.org/10.1016/j.jml.2018.05.005

Merkle, E., & Van Zandt, T. (2006). An application of the poisson race model to confidence calibration. *Journal of Experimental Psychology. General, 135*, 391–408.

Modick, H. E. (1977). A 3-scale measure of achievement motivation: Report on a German extension of the prestatie motivatie test. *Diagnostica, 23*(4), 298–321.

Molenaar, D., & De Boeck, P. (2018). Response mixture modeling: Accounting for heterogeneity in item characteristics across response times. *Psychometrika, 83*(2), 279–297.

Molenaar, D., Oberski, D., Vermunt, J., & Boeck, P. D. (2016). Hidden Markov item response theory models for responses and response times. *Multivariate Behavioral Research, 51*(5), 606–626. https://doi.org/10.1080/00273171.2016.

1192983
Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. J. (2015). A bivariate generalized linear item response theory modeling framework to the analysis of responses and response times. *Multivariate Behavioral Research, 50*(1), 56–74. https://doi.org/10.1080/00273171.2014.962684
Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. J. (2015). Fitting diffusion item response theory models for responses and response times using the r package diffirt. *Journal of Statistical Software, 66*(4), 1–34. https://doi.org/10.18637/jss.v066.i04
Muraki, E. (1990). Fitting a polytomous item response model to likert-type data. *Applied Psychological Measurement, 14*(1), 59–71. https://doi.org/10.1177/014662169001400106
Muthén, B. (1983). Latent variable structural equation modeling with categorical data. *Journal of Econometrics, 22*(1), 43–65. https://doi.org/10.1016/0304-4076(83)90093-3
Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49*, 115–132. https://doi.org/10.1007/BF02294210
Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika, 44*(4), 443–460.
Partchev, I., & De Boeck, P. (2012). Can fast and slow intelligence be differentiated? *Intelligence, 40*(1), 23–32. https://doi.org/10.1016/j.intell.2011.11.002
Pearson, K. (1901). Mathematical contributions to the theory of evolution. viii. on the inheritance of characters not capable of exact quantitative measurement. *Philosophical Transactions of the Royal Society of London A, 195*, 79–150. https://doi.org/10.1098/rsta.1900.0024
Pleskac, T. J., & Busemeyer, J. (2010). Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. *Psychological Review, 117*(3), 864–901.
Ranger, J., & Kuhn, J.-T. (2018). Modeling responses and response times in rating scales with the linear ballistic accumulator. *Methodology, 14*(3), 119–132. https://doi.org/10.1027/1614-2241/a000152
Ranger, J., Kuhn, J.-T., & Szardenings, C. (2017). Analysing model fit of psychometric process models: An overview, a new test and an application to the diffusion model. *British Journal of Mathematical and Statistical Psychology, 70*(2), 209–224. https://doi.org/10.1111/bmsp.12082
Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*(2), 59–108.
Ratcliff, R. (2002). A diffusion model account of response time and accuracy in a brightness discrimination task: Fitting real data and failing to fit fake but plausible data. *Psychological Science, 9*(2), 278–291.
Ratcliff, R. (2018). Decision making on spatially continuous scales. *Psychological Review, 125*, 888–935. https://doi.org/10.1037/rev0000117
Ratcliff, R., Gomez, P., & McKoon, G. (2003). A diffusion model account of the lexical decision task. *Psychological Review, 111*(1), 159–182. https://doi.org/10.1037/0033-295X.111.1.159
Ratcliff, R., Hasegawa, Y., Hasegawa, R., Smith, P., & Segraves, M. (2007). Dual diffusion model for single-cell recording data from the superior colliculus in a brightness-discrimination task. *Journal of Neurophysiology, 97*, 1756–74. https://doi.org/10.1152/jn.00393.2006
Ratcliff, R., Hasegawa, Y. T., Hasegawa, R. P., Smith, P. L., & Segraves, M. A. (2007). Dual diffusion model for single-cell recording data from the superior colliculus in a brightness-discrimination task. *Journal of Neurophysiology, 97*(2), 1756–1774. https://doi.org/10.1152/jn.00393.2006
Ratcliff, R., & Kang, I. (2021). Qualitative speed-accuracy tradeoff effects can be explained by a diffusion/fast-guess mixture model. *Scientific Reports, 11*, 15169. https://doi.org/10.1038/s41598-021-94451-7
Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation, 20*(4), 873–922.
Ratcliff, R., & Smith, P. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review, 111*, 333–67. https://doi.org/10.1037/0033-295X.111.2.333
Ratcliff, R., & Starns, J. J. (2009). Modeling confidence and response time in recognition memory. *Psychological Review, 116*(1), 59–83.
Ratcliff, R., & Starns, J. J. (2013). Modeling confidence judgments, response times, and multiple choices in decision making: Recognition memory and motion discrimination. *Psychological Review, 120*(3), 697–719. https://doi.org/10.1037/a0033152
Ratcliff, R., Voskuilen, C., & McKoon, G. (2018). Internal and external sources of variability in perceptual decision-making. *Psychological Review, 125*(1), 33–46. https://doi.org/10.1037/rev0000080
Rouder, J., Province, J., Morey, R., Gómez, P., & Heathcote, A. (2015). The lognormal race: A cognitive-process model of choice and latency with desirable psychometric properties. *Psychometrika, 80*(2), 491–513. https://doi.org/10.1007/s11336-013-9396-3
Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34*(4, Pt. 2), 100.
Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). Springer. https://doi.org/10.1007/978-1-4757-2691-6_5
Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement, 34*(3), 213–232. https://doi.org/10.1111/j.1745-3984.1997.tb00516.x
Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464. https://doi.org/10.2307/2958889

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Chapman & HallCRC.
Smith, P. L. (2000). Stochastic dynamic models of response time and accuracy: A foundational primer. *Journal of Mathematical Psychology, 44*(3), 408–463. https://doi.org/10.1006/jmps.1999.1260
Smith, P. L., & Vickers, D. (1988). The accumulator model of two-choice discrimination. *Journal of Mathematical Psychology, 32*(2), 135–168. https://doi.org/10.1016/0022-2496(88)90043-0
Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64*(4), 583–639. https://doi.org/10.1111/1467-9868.00353
Stan Development Team. (2021). Stan modeling language user's guide and reference manual stan modeling language user's guide and reference manual. Retrieved from http://mc-stan.org/.
Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*, 393–408. https://doi.org/10.1007/BF02294363
Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*(4), 273–286. https://doi.org/10.1037/h0070288
Thurstone, L. L. (1927). Psychophysical analysis. *The American Journal of Psychology, 38*(3), 368–389.
Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology, 33*, 529–554. https://doi.org/10.1086/214483
Torgenson, W. S. (1958). *Theory and methods of scaling*. Wiley.
Tuerlinckx, F., & De Boeck, P. (2005). Two interpretations of the discrimination parameter. *Psychometrika, 70*(4), 629–650. https://doi.org/10.1007/s11336-000-0810-3
Tuerlinckx, F., Molenaar, D., & van der Maas, H. L. J. (2016). Diffusion-based response-time models. In W. J. van der Linden (Ed.), *Handbook of item response theory* (pp. 283–300). Chapman and Hall/CRC.
Turner, B. M., & Sederberg, P. B. (2012). Approximate Bayesian computation with differential evolution. *Journal of Mathematical Psychology, 56*(5), 375–385. https://doi.org/10.1016/j.jmp.2012.06.004
Turner, B. M., & Sederberg, P. B. (2014). A generalized, likelihood-free method for posterior estimation. *Psychonomic Bulletin and Review, 21*, 227–250. https://doi.org/10.3758/s13423-013-0530-0
Turner, B. M., & Van Zandt, T. (2012). A tutorial on approximate Bayesian computation. *Journal of Mathematical Psychology, 56*(2), 69–85. https://doi.org/10.1016/j.jmp.2012.02.005
Turner, B. M., & Van Zandt, T. (2014). Hierarchical approximate Bayesian computation. *Psychometrika, 79*, 185–209. https://doi.org/10.1007/s11336-013-9381-x
Usher, M., & McClelland, J. (2001). The time course of perceptual choice: the leaky, competing accumulator model. *Psychological Review, 108*, 550–92. https://doi.org/10.1037//0033-295X.108.3.550
van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72*(3), 287–308. https://doi.org/10.1007/s11336-006-1478-z
van der Maas, H. L. J., Molenaar, D., Maris, G., Kievit, R. A., & Borsboom, D. (2011). Cognitive psychology meets psychometric theory: On the relation between process models for decision making and latent variable models for individual differences. *Psychological Review, 118*(2), 339–356. https://doi.org/10.1080/20445911.2011.454498
van der Maas, H. L. J., & Wagenmakers, E.-J. (2005). A psychometric analysis of chess expertise. *The American Journal of Psychology, 118*, 29–60.
Van Zandt, T. (2000). Roc curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(3), 582–600. https://doi.org/10.1037/0278-7393.26.3.582
Van Zandt, T., & Maldonado-Molina, M. (2004). Response reversals in recognition memory. Journal of experimental psychology. *Learning, Memory, and Cognition, 30*, 1147–1166. https://doi.org/10.1037/0278-7393.30.6.1147
Vickers, D. (1979). *Decision processes in visual perception*. Academic Press.
Volkmann, J. (1934). The relation of time of judgment to certainty of judgment. *Psychological Bulletin, 31*, 672–673.
Wald, A. (1947). *Sequential analysis*. Wiley.
Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology, 68*(3), 456–477. https://doi.org/10.1111/bmsp.12054
Wang, C., Xu, G., & Shang, Z. (2018). A two-stage approach to differentiating normal and aberrant behavior in computer based testing. *Psychometrika, 83*(1), 223–254. https://doi.org/10.1007/s11336-016-9525-x
Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica, 41*(1), 67–85. https://doi.org/10.1016/0001-6918(77)90012-9

Manuscript Received: 1 MAR 2022
Final Version Received: 25 OCT 2022
Accepted: 3 JAN 2023