PSYCHOMETRIKA—VOL. 87, NO. 2, 725–748
JUNE 2022
https://doi.org/10.1007/s11336-021-09819-5

MODELING CONDITIONAL DEPENDENCE OF RESPONSE ACCURACY AND RESPONSE TIME WITH THE DIFFUSION ITEM RESPONSE THEORY MODEL

INHAN KANG, PAUL DE BOECK AND ROGER RATCLIFF

THE OHIO STATE UNIVERSITY

In this paper, we propose a model-based method to study conditional dependence between response accuracy and response time (RT) with the diffusion IRT model (Tuerlinckx and De Boeck in Psychometrika 70(4):629–650, 2005, https://doi.org/10.1007/s11336-000-0810-3; van der Maas et al. in Psychol Rev 118(2):339–356, 2011, https://doi.org/10.1080/20445911.2011.454498). We extend the earlier diffusion IRT model by introducing variability across persons and items in cognitive capacity (drift rate in the evidence accumulation process) and variability in the starting point of the decision processes. We show that the extended model can explain the behavioral patterns of conditional dependency found in the previous studies in psychometrics. Variability in cognitive capacity can predict positive and negative conditional dependency and their interaction with the item difficulty. Variability in starting point can account for the early changes in the response accuracy as a function of RT given the person and item effects. By the combination of the two variability components, the extended model can produce the curvilinear conditional accuracy functions that have been observed in psychometric data. We also provide a simulation study to validate the parameter recovery of the proposed model and present two empirical applications to show how to implement the model to study conditional dependency underlying data response accuracy and RTs.

Key words: diffusion IRT model, response time, psychological process, conditional dependency, process modeling.

In this article, we present a model-based analysis of conditional dependence between response accuracy and response time (RT) using a psychological decision-making process model. The process model describes the psychological processes that give rise to the measurement outcomes. Understanding psychological processes underlying cognitive tests and psychometric inventories is important because it provides a different view of intra-individual differences and how they bring about inter-individual differences than do descriptive models of response accuracy and RT. Therefore, process-based modeling can help to properly conceptualize latent variables and their existence, facilitate improvement in their measurement and study of validity, and, ultimately, bridge the gap between inter-individual-level processes and intra-individual causal inferences (Borsboom, Mellenbergh, & van Heerden, 2003, 2004). Furthermore, modeling based on a theory of psychological processes provides a coherent account of the complicated behavioral patterns of responses and RTs and allows identification of cognitive sources of important aspects of data such as residual dependency between response accuracy and RT, controlling for the confounded person and item effects.

Although psychometric inventories and tests measure latent traits and abilities, they rarely shed light on psychological processes when they are based only on outcome performance such as response accuracy. This is because different models, with or without theories relevant to the underlying processes, can produce very similar model predictions, and thus, it is difficult to discriminate between models with different theories of response processes. Extending the modeling

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11336-021-09819-5.

Correspondence should be made to Inhan Kang, The Ohio State University, 291 Psychology Building, 1835 Neil Avenue, Columbus, OH 43210, USA. Email: kang.985@osu.edu

to account for the joint behavior of different variables is necessary to gain insight into how we process information and give item responses.

Among the outcomes, RT is a straightforward and easily accessible outcome along with response accuracy. Accuracy and RT have different measurement properties and scales; response accuracy is a binomial random variable divided by the total number of responses, while RT is a random variable with continuous and right-skewed distributions. Thus, jointly modeling accuracy and RT more strongly constrains mathematical models of psychological processes and provides a finer measurement of latent traits and abilities (Bolsinova & Tijmstra, 2018).

There has been an increasing interest in RT modeling in psychometrics, and one of the seminal works is the hierarchical framework (van der Linden, 2007). In this approach, response accuracy is modeled by the three-parameter normal ogive model and RT is modeled by a log-normal model. As a link between these two model parts, it is assumed that latent variables in both models follow a multivariate normal distribution (population model), and item parameters in both models follow another multivariate normal distribution (item domain model). Given the latent variables and item parameters, the framework assumes that there is no further dependency between response accuracy and RT. In other words, all the associations between the two outcome variables can be captured by the latent variables and item parameters.

Although the assumption of conditional independence has facilitated joint modeling of response accuracy and RT, this has often not been justified by psychometric data. In fact, violations of this assumption have become a robust finding across various test types in psychometrics (Bolsinova, De Boeck, & Tijmstra, 2017; Bolsinova & Maris, 2016; Bolsinova & Molenaar, 2018; Bolsinova, Tijmstra, & Molenaar, 2017; Chen, De Boeck, Grady, Yang, & Waldschmidt, 2018a; De Boeck, Chen, & Davison, 2017; Goldhammer et al., 2014; Goldhammer, Naumann, & Greiff, 2015; Meng, Tao, & Chang, 2015; Partchev & De Boeck, 2012; van Rijn & Ali, 2017; van der Linden & Glas, 2010; Wang & Xu, 2015). Such violations show that RTs provide information about the corresponding response over and above what is captured by latent variables and item effects. Therefore, it could be that models with the conditional independence assumption fail to appropriately measure latent traits and abilities and to represent the latent structure underlying response behavior.

There have been a number of relatively recent studies examining condition dependency between accuracy and RT (Bolsinova, De Boeck, & Tijmstra, 2017; Bolsinova & Molenaar, 2018; Chen et al., 2018a; Goldhammer et al., 2014, 2015). Results have shown negative, positive, and curvilinear dependency. Negative conditional dependency is shown by accuracy decreasing as a function of RT when the person and item effects are controlled for. Positive conditional dependency occurs when accuracy increases as a function of RT given latent variables and item parameters. It has been found that different items have different patterns of conditional dependency and these inter-item differences are associated with item difficulty (Bolsinova, De Boeck, & Tijmstra, 2017; Goldhammer et al., 2014, 2015). Typically, responses and RTs for easy items show negative conditional dependency, but this trend tends to get weaker for more difficult items and can even flip to positive dependency for highly difficult items.

A recent study by Chen et al. (2018a) provides evidence for curvilinear conditional dependency in which accuracy first increases relatively steeply until it asymptotes and then decreases over RTs. Chen et al. divided RTs into several bins and plotted response accuracy against bins of log-transformed RTs that were double-centered based on person-wise and item-wise means to eliminate person and item effects. For all five tasks that were examined (three achievement tests and two cognitive ability tests), the plots of accuracy showed the curvilinear pattern. From the same datasets, a follow-up study by Chen, De Boeck, Grady, Yang, & Waldschmidt (2018b) found that the curvilinear dependency can differ by item difficulty because easy items showed the first-increasing and later-decreasing curvilinear pattern, while difficult items showed the mirror image, the first-decreasing and later-increasing curvilinear pattern with much lower accuracy. Bolsinova

and Molenaar (2018) also found a similar relationship between response accuracy and RT. They modeled the intercept of the item characteristic curve (ICC; a curve of response probability as a function of latent ability/trait) of the IRT model as a function of log-transformed RTs that were centered by the mean RT predicted by the log-normal RT model and scaled by the item-wise residual standard deviation. They examined different nonlinear models and found a curvilinear relationship between the intercept and the standardized residual log-transformed RT. Because the intercept parameter is positively related to response accuracy, this result implies curvilinear conditional dependency of response accuracy and RT. The generalized speed–accuracy response model for dichotomous items which van Rijn and Ali (2017, 2018) developed based on the scoring rule by Maris and van der Maas (2012) also implies a curvilinear relationship between response probability and response time.

For the various conditional dependency patterns found in the literature, it is important to find their potential sources (Bolsinova, Tijmstra, Molenaar, & DeBoeck, 2017). Negative dependency has been interpreted as a consequence of within-person variability in cognitive capacity (Chen et al., 2018a; De Boeck et al., 2017; DeBoeck & Jeon, 2019), heterogeneity in item difficulty across persons (i.e., the same item can be more or less difficult to different persons), etc. Positive dependency has been associated with a total time limit of a test, within-person variability in response caution, etc. Fast aberrant responses such as fast guessing and cheating (Wang & Xu, 2015; Wang, Xu, & Shang, 2018) can produce positive residual dependency particularly at the early RT period because these responses usually have fast RTs with lower accuracy (Bolsinova, Tijmstra, Molenaar, & De Boeck, 2017). Some factors such as ability-based guessing (San Martín, del Pino, & De Boeck, 2006), attractive distractors in multi-choice items, and different types of processing (e.g., fast vs slow processes; DiTrapani, Jeon, De Boeck, & Partchev, 2016; Goldhammer et al., 2014; Molenaar, Bolsinova, Rozsa, & De Boeck, 2016; Partchev & De Boeck, 2012) can produce either negative or positive dependency. Curvilinear dependency can probably be explained by the sources for negative and positive dependency listed above because a combination of positive and negative dependency produces a curvilinear trend. However, this explanation requires that, for easy items, positive dependency appears only when RT is shorter than expected (i.e., short residual RT) and negative dependency appears only when RT is longer than expected (i.e., long residual RT). As there is no evidence for this relationship between positive/negative dependency and residual RTs, probable sources of curvilinear dependency require further investigation. We will present and test a more general and integrated explanation.

Although the explanations above are promising, it is hard to corroborate these explanations using data exploration and descriptive models. These methods are useful to discover and describe unrevealed relationships, but they cannot throw light on why the relationships occur. This is where we bring our attention back to process-based models because these are explanatory models that aim to explain psychological processes through theories represented by the model structure and components. For a study of conditional dependency, we mainly focus on a process model called the diffusion item response theory (diffusion IRT) model (Molenaar et al., 2016; Ranger & Kuhn, 2018; Ranger, Kuhn, & Szardenings, 2016, 2017, 2020; Tuerlinckx & De Boeck, 2005; Tuerlinckx, Molenaar, & van der Maas, 2016; van der Maas et al., 2011) that explains the response process using the sequential sampling framework. The sequential sampling framework assumes that, when an item is presented, a respondent accumulates information/evidence until the accrued information becomes sufficient to make a response. The model parameters represent different cognitive aspects of the response processes, and we aim to study the sources of conditional dependency of response accuracy and RT by identifying cognitive components associated with the dependency. To this end, we extend the diffusion IRT model so that the model can account for overall conditional dependency and heterogeneity in dependency across persons and items.

The paper is organized as follows. We first introduce the diffusion IRT model, which serves as our modeling framework (Sect. 1). We extend this model with random variability in cognitive

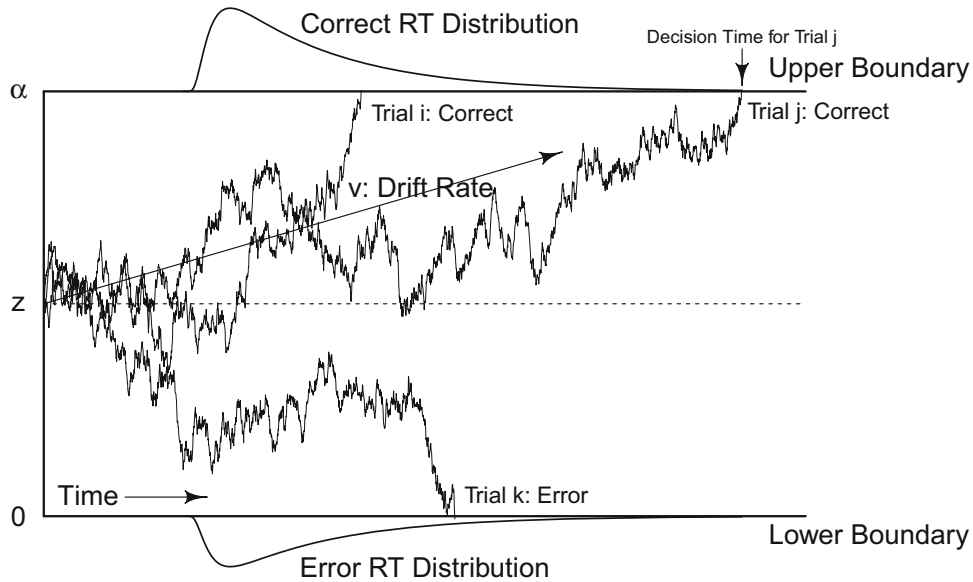

FIGURE 1.
Illustration of the evidence accumulation in the Wiener/Ratcliff diffusion model.

components to account for conditional dependency (Sect. 2). Next, we provide a simulation study to validate the extended model by examining its parameter recovery (Sect. 3) and illustrate how one can use the model to study conditional dependency between response accuracy and RT with empirical data (Sect. 4). We conclude with a discussion on further modeling issues of conditional dependency (Sect. 5).

1. Diffusion Item Response Theory Model

Process models take into account the cognitive processes that simultaneously produce responses and RTs during response procedures. This modeling approach has been dominant in perceptual and cognitive psychology, but little attention has been paid to this approach in psychometrics. A notable example of a process model in psychometrics is the diffusion IRT model proposed by Tuerlinckx and De Boeck (2005). The model is a combination of the Wiener diffusion model from cognitive psychology and the IRT model from psychometrics. The Wiener diffusion model is one of the sequential sampling models (SSMs) assuming that when a stimulus for a psychological task is presented, information or evidence extracted from the stimulus is accumulated over time to determine which response option is appropriate for the trial (Fig. 1). In the Wiener diffusion model for a binary-response task, evidence accumulation begins at a starting point z toward one of the two (upper and lower) boundaries at a mean rate (drift rate) of ν. The distance between the two boundaries is called the boundary separation α and typically the upper boundary is mapped to α and the lower boundary is mapped to 0. These two boundaries represent two response options. The boundary at which the process terminates determines a response made and the time that the process takes to reach a boundary determines a decision time (DT). There also are processes unrelated to evidence accumulation, and the time for these processes is collectively modeled as nondecision time t_0. The model predicts an RT as the sum of decision and nondecision times (RT $= t_0 +$ DT). Evidence accumulation is noisy within each trial, and thus, the same set of model parameter values (α, z, t_0, and ν) can give rise to different responses and RTs (trials

i, j, and k in Fig. 1), producing bivariate RT distributions for two response options. One of the important features of the diffusion IRT model (in fact, of most of the SSMs) is that the model parameters correspond to components of the cognitive processes assumed by the model. Drift rate v represents the mean rate (quality and efficiency) of evidence accumulation, boundary separation α represents the amount (quantity) of information required to make a response, starting point z represents an initial bias between the two response options, and nondecision time t_0 represents time consumed for nondecision processes such as stimulus encoding and response production.

The diffusion IRT model is an extension of the unbiased Wiener diffusion model and can be used for psychometric measurement data in which person p ($p = 1, \ldots, P$) responds to item i ($i = 1, \ldots, I$) once, resulting in two ($P \times I$) matrices of responses and RTs. The unbiasedness of the model means that there is no initial bias assumed ($z = \alpha/2$). Furthermore, for person p's response to item i, the following decompositions of the drift rate (Molenaar, Tuerlinckx, & van der Maas, 2015; Tuerlinckx & De Boeck, 2005; Tuerlinckx et al., 2016) and the boundary separation (Molenaar et al., 2015; Tuerlinckx et al., 2016; van der Maas et al., 2011)[1] were implemented to account for both person and item effects:

$$\begin{aligned} v_{pi} &= \theta_p - b_i \\ \alpha_{pi} &= \gamma_p / a_i \end{aligned} \tag{1}$$

where γ_p and θ_p represent person-wise decision criterion (or cautiousness) and person-wise drift rate, respectively, and a_i and b_i represent item time-pressure (or the inverse of item discrimination) parameter and item difficulty parameter, respectively. In perceptual and cognitive psychology, experimental conditions defined by stimulus manipulations typically affect the difficulty of tasks (e.g., Brown & Steyvers, 2005; Kang & Ratcliff, 2020; McKoon & Ratcliff, 2016; Ratcliff, 2002; Ratcliff, Gomez, & McKoon, 2003; Ratcliff & Rouder, 1998; Ratcliff & McKoon, 2018) but not the amount of information required to make a choice. Accordingly, for a single person, drift rate is allowed to vary by condition, but boundary separation is fixed across conditions unless the experimental conditions are defined by the experimenter's instruction on the speed–accuracy trade-off (stressing either speed or accuracy; Ratcliff & McKoon, 2008). In contrast, in psychometrics, items have different characteristics such as difficulty and discrimination. Therefore, both drift rate and boundary separation are modeled to vary by persons and items (Molenaar et al., 2015; Tuerlinckx et al., 2016; van der Maas et al., 2011). Changes in boundary separations across items can be attributed to the factors related to the amount of time required by the item (van der Linden, 2007), item time-intensity (Bolsinova, De Boeck, & Tijmstra, 2017), item-wise time pressure due to the item context (e.g., test instructions, item positions, etc.; Molenaar et al., 2015; Tuerlinckx et al., 2016), time pressure for the whole test (van der Maas et al., 2011), and item discrimination (Tuerlinckx & De Boeck, 2005).

Further assuming person-wise nondecision time τ_p (i.e., $t_0 = \tau_p$), the first passage time density function of the diffusion IRT model is given as (Cox & Miller, 1970; Tuerlinckx & De Boeck, 2005):

[1] van der Maas et al. (2011) used a different parameterization for the drift rate, and the difference will be explained in the Discussion section.

$$f_{X_{pi},T_{pi}}(x,t|*) = \frac{\pi s^2}{\alpha_{pi}^2}\exp\left(\frac{\alpha_{pi}(x-\frac{1}{2})\nu_{pi}}{s^2} - \frac{\nu_{pi}^2}{2s^2}(t-\tau_p)\right) \times \sum_{k=1}^{\infty} k\sin\left(\frac{\pi k}{2}\right)\exp\left(-\frac{\pi^2 k^2 s^2}{2\alpha_{pi}^2}(t-\tau_p)\right) \tag{2}$$

where '*' denotes all the model parameters, X_{pi} and T_{pi} (with $T_{pi} > \tau_p$) are the random variables of binary response and RT for person p and item i, respectively, and s is the diffusion coefficient which is the standard deviation of the within-trial noise of the evidence accumulation process. The parameter s is typically fixed to a constant (for example, 1, as we do in this article hereafter) as a scaling coefficient. Letting $X_{pi} = 1$, Eq. 2 gives the first passage time density function of the process hitting the upper boundary first (before it reaches the lower boundary). Analogously, the first passage time density function corresponding to the lower boundary can be obtained by letting $X_{pi} = 0$. Note that the density function given in Eq. 2 is bivariate and defective: Neither $f_{1,T_{pi}}(1,t)$ nor $f_{0,T_{pi}}(0,t)$ integrates to 1. Instead, the sum of the two integrals is 1 as $f_{1,T_{pi}}(1,t)$ integrates to $Pr(X_{pi}=1)$, the probability of the process terminating at the upper boundary, while $f_{0,T_{pi}}(0,t)$ integrates to $Pr(X_{pi}=0)$, the probability of the process terminating at the lower boundary. The probability of choosing the response option corresponding to the upper boundary $Pr(X_{pi}=1)$ is derived as (Cox & Miller, 1970; Luce, 1986; Tuerlinckx & De Boeck, 2005):

$$Pr(\text{hitting the upper boundary}) = Pr(X_{pi}=1) = \frac{\exp(\frac{\gamma_p}{a_i}(\theta_p - b_i))}{1+\exp(\frac{\gamma_p}{a_i}(\theta_p - b_i))}. \tag{3}$$

Equation 3 is equal to the response probability of the two-parameter logistic IRT model (2PLM; Birnbaum, 1969) with $\alpha_{pi} = \frac{\gamma_p}{a_i}$ as the discrimination parameter, b_i as the item difficulty parameter, and θ_p as the person latent ability in the IRT models. Therefore, the diffusion IRT model predicts the response probability just as the 2PLM does, but it also predicts RT distributions with Eq. 2.

The diffusion IRT model is not able to predict conditional dependence between response accuracy and RT. The first passage time density of the model satisfies the following relationship (Laming, 1968; Stone, 1960; Tuerlinckx & De Boeck, 2005):

$$f_{T_{pi}|X_{pi}}(t|x=0,*) = f_{T_{pi}|X_{pi}}(t|x=1,*) = f_{T_{pi}}(t,*) \tag{4}$$

This implies $f_{X_{pi},T_{pi}}(x,t|*) = f_{X_{pi}}(x|*) \times f_{T_{pi}}(t|*)$, and thus, the model assumes conditional independence between choice response and RT given the model parameters. As responses do not provide information on the RT distributions, the model predicts that correct and error RT distributions have the same mean RTs and RT quantiles (hereafter called 'symmetric' correct and error RT distributions). As we discussed in the previous section, often psychometric measurement data have not supported this assumption and thus the model should be modified to appropriately account for the psychological process underlying responses and RTs. One possibility is to extend the model by introducing additional model parameters that represent some unexplained aspects of the cognitive process. An interesting idea that we will examine in this article is to introduce variability in some cognitive components of the model. This idea, previously used by Ratcliff (Ratcliff, 1978, 2002; Ratcliff & Rouder, 1998), gave birth to the Ratcliff diffusion model, one of the most compelling models of responses and RTs in perceptual and cognitive decision making (Forstmann, Ratcliff, & Wagenmakers, 2016; Ratcliff & McKoon, 2008; Ratcliff, Smith, Brown, & McKoon, 2016). For psychometric measurement data, Tuerlinckx De Boeck (2005) have already proposed a related idea, which we will review with our additional proposal in the next section.

2. Random Variability and Conditional Dependence of the Diffusion IRT Model

The conditional independence assumption is often not supported by data in perceptual and cognitive decision making. It has been shown that, in the same experimental condition, an error RT distribution is typically slower than the correct RT distribution. Also, there are fast errors that make the leading edge of the error RT distribution relatively faster than that of the correct RT distribution. In general, error RTs are slower than correct RTs when a task is difficult and accuracy is stressed, while error RTs are faster when a task is easy and speed is stressed (Luce, 1986; Ratcliff & McKoon, 2008; Ratcliff & Rouder, 1998; Ratcliff, Van Zandt, & McKoon, 1999; Smith & Vickers, 1988; Swensson, 1972; see also Luce, 1986, pp. 233–236 for a more detailed review of faster and slower errors).

Accounting for the asymmetry between correct and error RT distributions has been a primary interest in perceptual and cognitive decision making, and one of the successful approaches is to add across-trial variability in some model parameters. A single set of Wiener diffusion model parameters (boundary separation, starting point, drift rate, and nondecision time) is too restrictive and is not able to account for the imbalance between correct and error RTs across multiple trials in a psychological experiment. As a modification to this model, Ratcliff added across-trial variability in drift rate, in starting point, and in nondecision time (Ratcliff, 1978, 2002; Ratcliff & McKoon, 2008; Ratcliff & Rouder, 1998), claiming that it is unlikely that a subject is able to accumulate evidence at the same rate (Kang, Ratcliff, & Voskuilen, 2020; Ratcliff, Voskuilen, & McKoon, 2018), to start the evidence accumulation at exactly the same starting point, and to spend an equal amount of time for nondecision processes across all trials. It has been shown that the Ratcliff diffusion model can account for various behavioral benchmarks of response proportions and RT distributions over a variety of psychological experiments, including slow and fast errors and thus the asymmetry between correct and error RT distributions (Ratcliff, 1978, 2002; Ratcliff & McKoon, 2008; Ratcliff & Rouder, 1998). These extensions may help to explain conditional dependencies between response accuracy and RT and have not been used much for diffusion IRT models.

Conditional accuracy function (CAF), a function of accuracy conditioned on RT, provides a good way to study the conditional (in)dependence assumption of different models, particularly, how the across-trial variability parameters in the Ratcliff diffusion model produce conditional dependency. For any model with joint probability density function $f(x, t|*)$ of binary response and RT, the CAF is obtained as $P(x = 1|t, *) = \frac{f(x=1,t|*)}{f(t|*)} = \frac{f(x=1,t|*)}{f(x=1,t|*)+f(x=0,t|*)}$ (Luce, 1986). Models with the conditional independence assumption (i.e., $f(x, t|*) = f(x|*)f(t|*)$) satisfy $P(x = 1|t, *) = \frac{f(x=1|*)}{f(x=1|*)+f(x=0|*)}$. Thus, CAFs of these models (including the Wiener diffusion model without variability extensions) have a flat shape and do not vary as a function of RT. The three black solid lines (one at the top, another in the middle, and the other at the bottom) in Fig. 2 represent the CAFs of the Wiener diffusion model with parameter values $\alpha = 1.1$, $z = \alpha/2$, $t_0 = 0.2$, and $\nu = 2$ for the top one (with accuracy of about 0.9), $\nu = 0$ for the middle one (with accuracy of 0.5), and $\nu = -2$ for the bottom one (with accuracy of about 0.1), showing that the model predicts flat CAFs.

The Ratcliff diffusion model predicts asymmetric RT distributions and nonflat CAFs, unlike the Wiener diffusion model. The model prediction is based on two variability parameters: across-trial variability in drift rates η (trial-wise drift rate $\sim N(\nu, \eta^2)$) and across-trial variability in starting points s_z (trial-wise starting point $\sim U(z-\frac{s_z}{2}, z+\frac{s_z}{2})$). The effects of across-trial variability on the CAFs differ by the values of the other model parameters, particularly the signs of the drift rate. Suppose that responses corresponding to the upper boundary are 'correct' responses and those corresponding to the lower boundary are 'error' responses. Given the other diffusion model parameters are fixed and with a positive drift rate ($\nu > 0$), the model predicts accuracy higher than

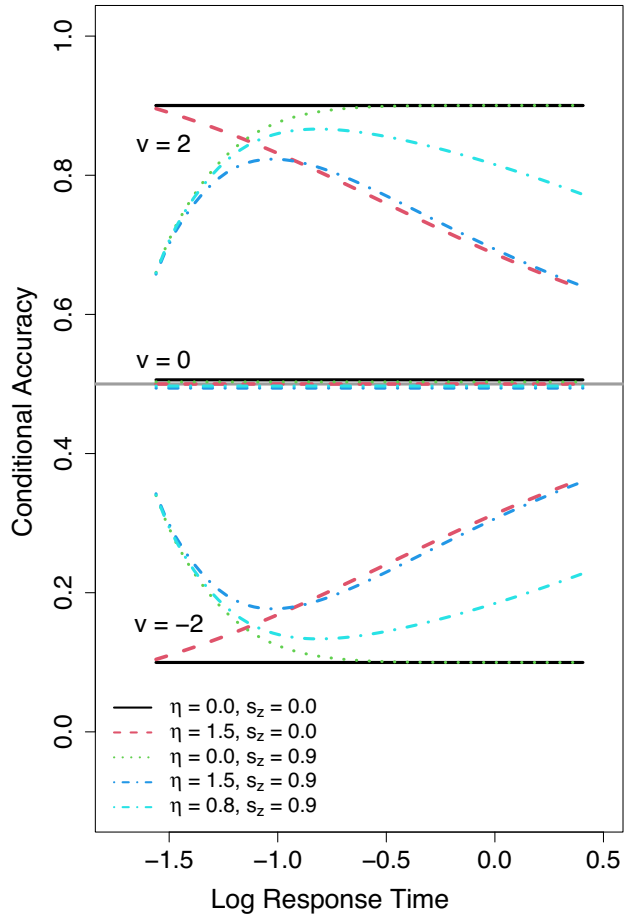

FIGURE 2.
Conditional accuracy functions (CAFs) of the Ratcliff diffusion model. The predicted CAFs are generated with varying values of across-trial variability in drift rates (η) and in starting points (s_z) shown at the bottom-left side of the figure. The other parameters used are boundary separation $\alpha = 1.1$, starting point $z = \alpha/2$, nondecision time $t_0 = 0.2$, and positive drift rate of $\nu = 2$ for the top four curves (above the gray horizontal line indicating accuracy of 0.5), negative drift rate of $\nu = -2$ for the bottom four curves (below the gray curve), and zero drift rate $\nu = 0$ for the four curves in the middle (Color figure online).

chance (0.5). In this case, nonzero η produces slow errors, while nonzero s_z generates fast errors. With the interaction of these two parameters, the Ratcliff diffusion model can capture various asymmetries between correct and error RT distributions in the data. Figure 2 shows different CAFs predicted by the Ratcliff diffusion model with different values of the variability parameters as shown at the bottom-left. The predicted CAFs are plotted on the log scale of RTs as done in Bolsinova and Molenaar (2018) and Chen et al. (2018a). The main four parameters are set to the same values used for the solid black curves (the flat CAFs of the Wiener diffusion model). Across-trial variability in nondecision time is set to 0 as it does not affect the shape of the CAFs. We do not discuss this kind of variability because it does not help to explain the local dependencies we are interested in. The four curves on the top side (above the gray line indicating accuracy of 0.5) show the CAF predictions with a positive drift rate ($\nu = 2$). The model with nonzero η predicts a decreasing CAF (e.g., red dashed line): Due to slow errors, there are likely more errors in slower RT ranges and thus accuracy in these RT ranges gets lower. In contrast, the model with nonzero s_z

predicts a CAF with an early steep increase (green dotted line): Due to early fast errors, it is likely that there are more errors in the RT leading edge, but the relative amount of fast errors decreases up to some RT point. Thus, accuracy keeps increasing until that RT point at which accuracy achieves its asymptote. The model with nonzero variability parameters for both drift rate and starting point predicts a curvilinear CAF (dot-dashed lines) which is the result of combining the first-increasing and later-decreasing patterns. Among the two dot-dashed lines in Fig. 2, the model with a larger value of η predicts the CAF with a lower peak (as more errors are predicted), while the model with a smaller value predicts the one with a higher peak. Note that heights, slopes of change, and asymptotes (if any) of the curves can differ by the parameter values.

The four curves on the bottom side of Fig. 2 show the CAF predictions with a negative drift rate ($\nu = -2$). Note that CAFs predicted by the model with negative drift rates are the mirror images of those predicted with positive drift rates (reflected over the horizontal line of accuracy = 0.5). For psychometric tests, a negative drift rate corresponds to the case where an item is too difficult for a person so that the predicted accuracy is lower than chance (c.f., stimuli with conflicting features in perceptual/cognitive tasks can also produce below-chance accuracy, Kang & Ratcliff, 2020). In this case, the model with nonzero η predicts an increasing CAF (e.g., the red dashed line below the gray horizontal line), but accuracy cannot reach 0.5. Similarly, the model with nonzero s_z predicts a CAF with an early steep decrease until its asymptotic lower bound, and a combination of the nonzero across-trial variability parameters causes the model to produce curvilinear CAFs with the first-decreasing and later-increasing pattern. Importantly, the opposite curvilinear CAFs and their association with item difficulty predicted from the Ratcliff diffusion model are consistent with the empirical findings in Chen et al. (2018a, 2018b).

When the drift rate is zero ($\nu = 0$), the diffusion process accumulates noisy evidence without any mean trend toward either boundary and thus responses are randomly determined. In this case, the model predicts chance accuracy (i.e., 0.5) regardless of the values of the variability parameters. As a result, CAFs predicted by the model are always flat with accuracy of 0.5, as shown by the four horizontal lines in the middle of Fig. 2. In general, the CAFs become flatter as the absolute value of the drift rate is smaller.

The model predictions shown in Fig. 2 correspond to the interpretations of conditional dependency in previous studies. Conditional dependency has been found to correlate with item difficulty. For easier items, the dependency is negative, and for more difficult items, it is weaker and can be positive (Bolsinova, Tijmstra, Molenaar, & De Boeck, 2017; Chen et al., 2018a; De Boeck et al., 2017; De Boeck & Jeon, 2019), which is also in line with Embretson's (2021) results for within-person correlations. These findings are consistent with the effects of variability in mean rate of information processing (i.e., drift rate) in the Ratcliff diffusion model. De Boeck et al. (2017) and De Boeck and Jeon (2019) have interpreted the local dependency findings referring to the Ratcliff diffusion model. When the cognitive capacity (i.e., drift rate) varies, the probability of the dominant response increases (decreases) and the response time decreases (increase). The dominant response for easy items is the correct response, while it is an incorrect response for difficult items. This can explain the correlation of local dependency with item difficulty and the switch from a negative dependency to a positive dependency.

Bolsinova, Tijmstra, Molenaar, and De Boeck (2017) proposed another interpretation of positive conditional dependency. They claimed that this pattern can appear due to within-person variation in the speed–accuracy balance and response caution during a test. Conceptually, this variation is directly related to the variation in boundary separation in the diffusion model, but across-trial variability in starting point can also account for this variation as both variability components can change the amount of information required for the process to reach either boundary. In this sense, we can interpret random variability in starting point as random variability in the

amount of information or response caution.[2] Because the starting point of the diffusion process represents an initial bias of response process or a starting point of problem-solving and it is more effective in the early RT period, the variability in starting point corresponds to the early increasing (decreasing) pattern of the CAF predicted by the diffusion model when the drift rate is positive (negative).

Tuerlinckx and De Boeck (2005) proposed to extend the variability idea to the diffusion IRT model. In particular, they adopted across-trial variability in drift rates, but only a single random variability parameter that works for all persons and items. This is because conceptually there is no 'multiple-trial experiment' in psychometric measurements as a single subject responds only once to a single item. With the random variability in drift rate η, the drift rate corresponding to person p and item i is modeled as $\nu_{pi} = \theta_p - \beta_i + \epsilon_{pi}$ where $\epsilon_{pi} \sim N(0, \eta)$. As shown in Fig. 2, the resulting model produces a decreasing CAF with a positive drift rate and an increasing CAF with a negative drift rate. Thus, given the other diffusion model parameters, it predicts that accuracy of person p's response to item i decreases (increases) as a function of RT when the drift rate is positive (negative). The response probability and the first passage time density function of this model can be obtained by adding ϵ_{pi} to drift rate in Eqs. 2 and 3 and integrating ϵ_{pi} out over its normal distribution. It also has been shown that there is a closed-form solution to this integration (Blurton, Kesselmeier, & Gondan, 2017; Tuerlinckx, 2004; Tuerlinckx & De Boeck, 2005).

Although the diffusion IRT model with random variability in drift rate can produce a nonflat CAF, accuracy as a monotone (either decreasing or increasing) function of time is not justified by psychometric data. As reviewed above, Bolsinova and Molenaar (2018) and Chen et al. (2018a) provided evidence for a curvilinear CAF that implies accuracy grows as a function of RT in the short RT range and then accuracy decreases after it reaches the peak. This trend can be predicted by a combination of random variability in drift rate and in starting point as shown in Fig. 2. From this observation, we propose to further extend the diffusion IRT model by introducing random variability in starting point, expecting that the starting point variability can introduce the positive (negative) dependency for easy (difficult) items that occurs primarily at the beginning of the response process and produce a curvilinear CAF. As done for random variability in drift rate, we implement a single parameter that governs random variability in starting point for all persons and items. Additionally, instead of the s_z parameter in the Ratcliff diffusion model which is the absolute range of the uniform distribution of trial-wise starting points, we will use $s_{zr} \in (0, 1)$ such that $s_{z,pi} = \alpha_{pi} \cdot s_{zr}$, which represents the same range but as a ratio relative to the boundary separation. This is because boundary separation is the maximum possible value of s_z for the unbiased process and it does vary by person and item in the diffusion IRT model. Thus, s_z can be severely underestimated if there are large individual differences in the boundary separation α_{pi} and $\min_{p,i}(\alpha_{pi})$ is too small. The relative range parameter s_{zr} does not incur this problem, and thus, we will use this parameterization in the current modeling.

Random variability in starting point in the diffusion IRT model represents that a level of initial bias or a starting point of problem-solving can differ by person and by item in psychometric measurement data. There are several potential sources of this variability. For example, responses to earlier items can affect response to the current item, producing a bias at the beginning of the response processes. The cognitive processes of problem-solving are thought to be induction-based multiple-trial processes that starts closer to or farther away from the correct response depending on the item and based on a partly random starting process. Starting points of the problem-solving processes may vary across pairs of persons and items and can be represented by variability in starting point rather than a change in the mean starting point (i.e., z). Given random variability in

[2]Random variability in decision boundaries is computationally more expensive because it requires two integrations (one for the upper boundary and the other for the lower boundary). Thus, we only consider random variability in starting point in our modeling.

starting point s_{zr}, starting point of person p's response to item i is modeled as $z_{pi} = \frac{\alpha_{pi}}{2} + \delta_{pi}$ where $\delta_{pi} \sim U(-\frac{\alpha_{pi} \cdot s_{zr}}{2}, \frac{\alpha_{pi} \cdot s_{zr}}{2})$.

The first passage time density function of the proposed model can be obtained by integrating out the two variability components as

$$g_{X_{pi},T_{pi}}(x,t|*) = \int_{-\frac{\alpha_{pi} \cdot s_{zr}}{2}}^{\frac{\alpha_{pi} \cdot s_{zr}}{2}} \int_{-\infty}^{\infty} f_{X_{pi},T_{pi}}(x,t|*) N(0,\eta^2) U\left(-\frac{\alpha_{pi} \cdot s_{zr}}{2}, \frac{\alpha_{pi} \cdot s_{zr}}{2}\right) \mathrm{d}\epsilon_{pi} \mathrm{d}\delta_{pi}, \tag{5}$$

and the response probability of the model can be obtained by integrating this density function with respect to RT over (τ_p, ∞)[3]

$$Pr(X_{pi} = 1) = \lim_{t \to \infty} \int_{\tau_p}^{t} g_{X_{pi},T_{pi}}(x = 1, u|*) \mathrm{d}u \tag{6}$$

In sum, conditional dependency between response accuracy and RT of psychometric data can be examined and explained via the CAF predicted by the model. Importantly, the model can help to identify the dominant source of conditional dependence by examining whether either the variation in the quality of evidence accumulation or the variation in the starting point of the response process (or both) can account for the dependence.

3. Simulation: Parameter Recovery

We conducted a simulation study to test if the diffusion IRT model with random variability can recover its parameters. A detailed description of the simulation setting and results can be found in our supplementary material, and here, we briefly summarize the results.

- We generated data of RTs and binary responses with $P = 200$ persons and $I = 15$ items from the diffusion IRT model with random variability in drift rate and in starting point.
- We used the differential evolution Markov Chain Monte Carlo (DE-MCMC; Ter Braak, 2006; Turner, Sederberg, Brown, & Steyvers, 2013) sampling method to fit the model to the simulated data and compared the MAP estimates to the true data-generating values.
- For the person and item parameters, the Pearson correlation between the MAP estimates and true parameter values is 0.972 ($\log(a_i)$), 0.988 (b_i), 0.954 (τ_p), 0.941 ($\log(\gamma_p)$), and 0.868 (θ_p), and there was no noticeable bias (Figure S1 and Table S3 in the supplementry material).
- Although variability in drift rate was slightly overestimated and the variability in starting point was slightly underestimated, the recovery was reasonably good. Their posterior distributions are provided in the supplementary material.
- Having more items can improve the estimation of the variability parameters (Figure S1 and Tables S3 in the supplementry material).

[3]The joint (cumulative) distribution function of response and RT of the diffusion IRT model can be obtained as $F_{X_{pi},T_{pi}}(x,t) = \int_t^{\tau_p} f_{X_{pi},T_{pi}}(x,u|*)\mathrm{d}u$, and thus, the response probability can also be computed as $Pr(X_{pi} = 1) = \int_{-\frac{\alpha_{pi} \cdot s_{zr}}{2}}^{\frac{\alpha_{pi} \cdot s_{zr}}{2}} \int_{-\infty}^{\infty} \lim_{t \to \infty} F_{X_{pi},T_{pi}}(x = 1, t|*) N(0,\eta^2) U(0,1) \mathrm{d}\epsilon_{pi} \mathrm{d}\delta_{pi}$ (see Ratcliff & Childers, 2015; Tuerlinckx, 2004, for related materials).

4. Empirical Applications

In this section, we provide two empirical applications of the diffusion IRT model with random variability parameters to two real datasets: extraversion and rotation (Molenaar et al., 2015). Particularly, we focus on explaining how to use different diffusion IRT models to study sources of conditional dependence between response and RT and how to produce model predictions of the CAFs. The extraversion data are from 143 respondents who were presented with 10 words related to introversion and extraversion (e.g., 'active') asking them to respond with 'Yes' or 'No' to describe their personality. Response accuracy is not defined for the extraversion data as we deal with a latent trait. Thus, the diffusion IRT model and its CAF predictions jointly describe response proportions and RT distributions. The rotation data have responses from 121 respondents to 10 binary mental rotation items with varying rotation angles (Borst, Kievit, Thompson, & Kosslyn, 2011; van der Maas et al., 2011). Each item presented two three-dimensional objects and the respondents indicated whether the second object was a rotated version of the first object. The response is coded 1 (correct) or 0 (incorrect). There was a time limit of 7, 500 ms in the rotation data. This may have made the data RT distribution less right-skewed than usual RT distributions, which can produce some misfits. Despite this limitation, we analyzed the rotation data (along with the extraversion data) to provide a descriptive example of how to produce the CAF predictions, not to test a general theory about the mental rotation processes. Both datasets are available from the **diffIRT** package in R (Molenaar et al., 2015).

To study sources of conditional dependence, we fitted four variants of the diffusion IRT models, with and without each of the random variability parameters. The first model was the diffusion IRT model (hereafter denoted as DIRT) without any random variability. The other three models are diffusion IRT models with random variability components (hereafter denoted as DIRT-RV). The second model in our comparison was the model with random variability in drift rate but without random variability in starting point which we denote as DIRT-RV(η). Similarly, the third model denoted as DIRT-RV(s_{zr}) was defined as the model with random variability in starting point but not in drift rate. The last model was the full model with both variability parameters denoted as DIRT-RV(η, s_{zr}). We fitted the models to the data with the same prior specification and sampling method as used in the simulation study (see supplementary material). We also assessed convergence in the same way as done in the simulation study (by visually inspecting the posterior densities and with the Gelman–Rubin convergence diagnostic; Gelman, 1996; Gelman, Carlin, Stern, Dunson, & Vehtari, 2013), and there was no convergence issue. When either a response or an RT was not recorded for person p and item i, we considered the data point a missing value and did not include it in the analysis. There was only one missing value (less than 0.1%) in the extraversion data, and there were 32 missing values (about 2.6%) in the rotation data. Also, we excluded one person in the rotation data from the analysis due to overly fast responding (9 out of 10 items were responded to in about a second, while 5% quantile of the overall RT distribution is 1.2 seconds, and accuracy of this person was 0.2 which is much lower than chance accuracy 0.5, implying that the person responded with wrong buttons), leaving 120 persons in the final analysis.

Sources of conditional dependence and their magnitudes can be studied by comparing the four models. The DIRT model plays a role as a reference model as it assumes conditional independence. If any of the other three models improves the general model fits, it implies the existence of conditional dependence. Furthermore, if the DIRT-RV(η, s_{zr}) model performs the best, it implies that both variability in drift rate and variability in starting point are sources of conditional dependence. If either the DIRT-RV(η) model or the DIRT-RV(s_{zr}) model shows the best model fit, it implies that the dominant source of conditional dependence is the random variability assumed in the model and the other variability has little contribution to behavioral patterns of the data. We conducted this comparison based on the modified Akaike information criterion (mAIC) and

the modified Bayesian information criterion (mBIC). These information criteria were proposed to use for joint models of responses and RTs estimated with a Bayesian method (Bolsinova, Tijmstra, & Molenaar, 2017; Bolsinova & Molenaar, 2018, 2019). These were calculated with -2 log-likelihood (-2LL) evaluated at the posterior means of the model parameters, but we used the MAP estimates for model evaluation instead. Because we jointly estimated all person and item parameters (not marginalizing person parameters or latent traits/abilities), we included the number of person parameters in the penalty term calculation. For example, the total number of parameters of the full DIRT-RV model is $3P + 2I + 2$.

The model comparison can be made at different levels. For each data set, the models are fit to two $(P \times I)$ matrices of responses and RTs, respectively, and produce a $(P \times I)$ matrix of log-likelihood values. Overall model fits can be evaluated with the sum of the log-likelihood values, and the models can be compared with these sum values to identify general sources of conditional dependence. However, there might be heterogeneity in conditional dependence between persons, between items, or even between person-item combinations (i.e., between responses). For example, a respondent may have more variability in starting point across items than other respondents and an item may incur more variability in drift rate than other items. Conditional dependence on different levels can be evaluated based on P row-wise sums, I column-wise sums, and $P \times I$ points of the log-likelihood matrix, for between-person, between-item, and between-combination conditional dependence, respectively. Then, heterogeneity in different conditional dependence levels can be studied by comparing these values across the four models.

Table 1 summarizes the model fitting results. Relative model fit indices such as -2LL of the four models evaluated at the MAP estimates, mAIC, and mBIC are shown in *Overall Model Fits* section of the table in which the bold values indicate the best-fitting model. For both of the datasets examined, the full model with the two random variability parameters performed the best, indicating that there was a certain amount of random variability in drift rate and in starting point that would be a dominant source of conditional dependence between response accuracy/proportion and RT. Judging by the MAP estimates of the variability parameters (in *Random Variability Estimates* section of Table 1), there was large random variability in drift rate and small random variability in starting point in the extraversion data, while both variability components were fairly large in the rotation data. *Individual-Level Comparisons* section presents the number of persons, the number of items, and the number of person-item combinations (i.e., responses) that prefer the model in the corresponding column. The numbers show that there was heterogeneity in conditional dependence at different levels. Thus, even though the full model was supported by the overall model fit indices, persons, items, and their combinations may have conditional dependence from different sources and some may satisfy conditional independence between response accuracy/proportion and RT.

The model comparison result above only shows the relative model fits and does not guarantee that the best model indeed accounts for the behavioral patterns of the data such as response accuracy/proportions and the shape of the RT distributions. The absolute model fit is particularly important for our analysis as it is necessary for the model-predicted CAFs to correctly represent conditional dependence underlying response accuracy/proportion and RT; if a model fails to explain behavioral patterns in the data, obviously it also fails to produce good CAF predictions. The absolute model fit to data can be examined by comparing data and model predictions. One can conclude that the model fits the data well and has a good absolute fit if the model predictions match the data, capturing important behavioral patterns.

Figure 3 shows the posterior predictive checking results of the extraversion data. In the left-most panel at the top row, the predicted proportions of the positive responses ('Yes' to given extraversion-related words) computed by item are plotted against the data-based response proportions. For all 10 items, the predicted item-wise response proportions are consistent with the data proportions. At the top-left side of the panel, the Pearson correlation between the data and predicted item-wise response proportions is shown as $r = 0.993$. Below the correlation estimates,

TABLE 1.
Empirical fitting results.

Data	Extraversion				Rotation			
Model	DIRT	DIRT-RV(η)	DIRT-RV(s_{zr})	DIRT-RV(η, s_{zr})	DIRT	DIRT-RV(η)	DIRT-RV(s_{zr})	DIRT-RV(η, s_{zr})
Overall model fits								
-2LL	1519.5	1471.2	1508.0	**1465.6**	3456.7	3449.3	3427.3	**3387.3**
mAIC	2417.5	2371.2	2408.0	**2367.6**	4222.7	4217.3	4195.3	**4157.3**
mBIC	3747.8	3704.5	3741.3	**3703.9**	5293.5	5290.9	5268.9	**5233.7**
Random variability estimates								
η	–	2.097 (1.546, 2.715)	–	2.085 (1.570, 2.867)	–	0.714 (0.400, 1.204)	–	1.006 (0.625, 1.550)
s_{zr}	–	–	0.108 (0.029, 0.271)	0.267 (0.084, 0.432)	–	–	0.473 (0.347, 0.597)	0.605 (0.476, 0.738)
Individual-level comparisons								
Person	33	33	40	37	25	35	34	27
Item	2	2	2	4	3	2	0	5
Combination	418	320	344	347	323	275	301	279

(1) Overall model fits: -2 log-likelihood, the modified Akaike information criterion (mAIC), and the modified Bayesian information criterion (mBIC) of the four diffusion IRT models fitted to extraversion and rotation data. The bold values indicate the best-fitting model. (2) Random variability estimates: MAP estimates of the random variability in drift rates (η) and random variability in starting points (s_{zr}) from different models, followed by their 95% credible intervals in parentheses. (3) Individual-Level Comparisons: The number of persons ('Person'), the number of items ('Item'), the number of the person-item combinations (i.e., responses; 'Combination') that prefer the model in the corresponding column. DIRT: Diffusion IRT Model, DIRT-RV: Diffusion IRT Model with Random Variability.

the overall data proportion of the positive responses is shown with the predicted response proportion in parentheses. The predicted proportion of 'Yes' responses is 0.806, which is consistent with the data proportion. The result shows that the best-fitting model is able to predict the overall and item-wise response proportions very well. In the top-middle panel, overall RT distributions (in seconds) obtained by aggregating all persons and items are shown. The histograms show the positive ('Yes' response; black) and negative ('No' response; red) RT distributions of the data. The RTs for the negative responses are coded negative (multiplied by -1) and plotted correspondingly for visual clarity. The black and red curves overlaid on the histograms show the predicted densities of the RT distributions for positive and negative responses, respectively. The consistency between the histograms and densities shows that the model predictions match the data very well. For an additional inspection of the RT distributions, the rightmost panel at the top row plots the data RT quantiles obtained by item on the x-axis against the predicted RT quantiles on the y-axis. The numbers 1, 3, 5, 7, and 9 represent the 10%, 30%, 50%, 70%, and 90% quantiles, respectively, and the black and red numbers represent RT quantiles for positive and negative responses, respectively. The predicted RT quantiles match the data well with the Pearson correlation of $r_{\text{Yes}} = 0.965$ and $r_{\text{No}} = 0.819$, for positive and negative responses, respectively. There is some misfit particularly at the tail of the negative RT distributions (red 9's). This is typical in the RT data from binary choice tasks as the RT distributions are right-skewed so that there is large variability in the tail. Also, the number of negative responses is much smaller than that of the positive responses (as shown in the panels in the bottom two rows of Fig. 3), which makes it harder to obtain precise data RT quantiles for the negative responses. Furthermore, there are some items with relatively 'balanced' data RT distributions, while the model predicts that negative responses are generally slower than positive responses unless the estimated drift rate is close to 0. This also can make the model overestimate some of the 90% quantiles of the item-wise negative RT distributions.

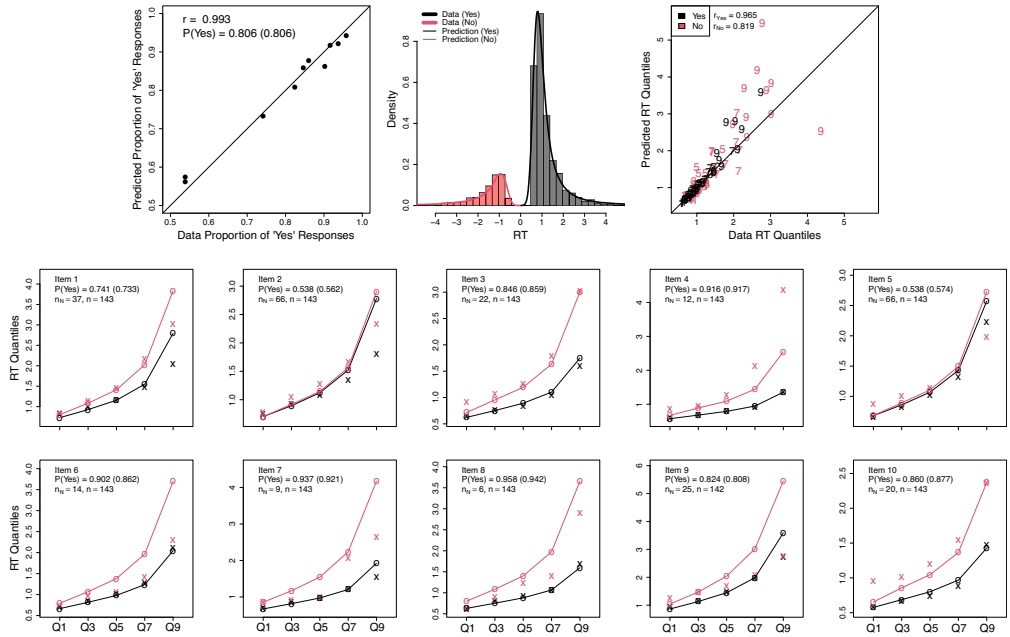

FIGURE 3.
Absolute model fits of the DIRT-RV(η,s_{zr}) model to extraversion data. In the top-left panel, item-wise data response proportions are plotted on the x-axis against the model predictions on the y-axis. In the top-middle panel, overall RT distributions (in seconds) obtained by aggregating all persons and items are shown. The histograms show the data, while the densities show the model predictions and the positive ('Yes') RT distribution is colored black, while the negative ('No') RT distribution is colored red. The RTs for negative responses are coded negative and plotted for visual clarity. In the top-right panel, item-wise data RT quantiles are plotted on the x-axis against the model predictions on the y-axis. The numbers 1, 3, 5, 7, and 9 represent the 10%, 30%, 50%, 70%, and 90% quantiles, respectively, and the black and red numbers represent RT quantiles for positive and negative responses, respectively. In the bottom two rows, the item-wise response proportions and RT distributions are presented. Within each panel, the data proportion of positive responses is shown at the top-left corner with the model prediction shown in the following parentheses. The number of negative responses (n_{N}) and the total number of responses (n) for each item are shown under the proportion of 'Yes' responses. For the item-wise RT predictions, RT quantiles in the top-right panel are plotted again, but now separately per item. In each panel, 'x's indicate data, while 'o's with the line connecting them indicate the model predictions. The five quantile points are plotted over the x-axis against their RT values on the y-axis. Positive responses are color-coded in black, while negative responses are in red (Color figure online).

In the bottom two rows of Fig. 3, the item-wise response proportions and RT distributions are presented. Within each panel, the data proportion of positive responses is shown at the top-left corner with the model prediction shown in parentheses. These values correspond to the black dots in the top-left panel. For the item-wise RT predictions, RT quantiles in the top-right panel are plotted again, but now separately per item. In each panel, 'x's indicate data, while 'o's with the line connecting them indicate the model predictions. The five quantile points are plotted over the x-axis against their RT values on the y-axis. Positive responses are color-coded in black, while negative responses are in red. The model predictions for the positive RT distributions match the data well for most of the items, while there are some misfits in the negative RT distributions for some items. The misfits in the negative RT distributions can be attributed to fewer negative response observations. The number of negative responses (n_{N}) and the total number of responses (n; the number of persons minus the number of missing values) for each item are shown at the top-left corner of each panel, under the proportion of 'Yes' responses. For example, the proportion of 'Yes' responses to item 6 is 0.902 and there are only 14 ($\approx 143 \times 0.098$) 'No' responses. These are obviously insufficient to obtain precise values of five (data) RT quantiles for the negative responses,

particularly considering that different persons responded to this item and thus the person effects are intermixed within those responses. The parameter estimation is more influenced by the more frequent responses (such as positive responses when there are only a few negative responses, and also, negative responses when there are many as for items 1–3) and less so by the less frequent responses (such as the negative RT distribution for item 6) to capture the dominant patterns of the data. Hence, it can be concluded that the misfits at the tails of the negative RT distributions are mostly due to the right-skewness, the corresponding larger sampling variability, and the small number of observations as pointed out above.

The model accounts for the asymmetry between the item-wise positive and negative RT distributions. The asymmetry is captured well for some items (e.g., item 3) but not well for other items (e.g., item 6). The misfit is related to the fewer negative response observations as described above, which makes it unlikely to obtain precise data RT quantiles. Furthermore, some items (e.g., items 2, 5, 6, and 9) have relatively symmetric data RT distributions, while the negative responses are slower in the overall RT distributions. Thus, there are across-item differences in the balance between positive and negative RT distributions (i.e., item-specific conditional dependency), which cannot be fully captured by the current model with a single random variability in drift rate η and a single random variability in starting point s_{zr}. Although the actual effects of a single variability parameter would differ by person and by item (for example, as a function of drift rates v_{pi} as shown in Fig. 2), a single variability parameter for all persons and items might be too restrictive to fully capture the asymmetry for all item-wise RT distributions. The model predicts rather flat CAFs when the drift rate is close to 0, which makes it capture symmetric RT distributions when item response proportion is near chance as for items 2 and 5. However, items 6 and 9 have high proportions of positive responses, and thus, the model predicts slower negative RT distributions which do not match the data. Inter-item heterogeneity in the RT balance can be better accounted for if the random variability is allowed to vary by item. However, this requires a large number of persons to obtain precise estimates of the item-wise variability parameters which is why we do not further investigate. Given the limited sample size ($P = 143$) and the parsimoniousness of our modeling, the model predictions of the item-wise response proportions and RT distributions are generally consistent with data, and thus, the absolute model fit with single variability parameters is reasonably good.

Figure 4 shows the posterior predictive checking results of the rotation data. The top-left panel shows that the data-based response accuracy is quite high for all items and the model produces good predictions for accuracy with the Pearson correlation of $r = 0.974$ and no noticeable bias. The overall data accuracy is 0.872, and the model prediction (0.893) is close to the data. The top-middle panel shows the data RT distributions obtained by aggregating all persons and items (histograms) and the corresponding model predictions (densities). Due to the time limit (7500 ms) of the task, posterior predictive samples with RTs greater than the time limit were excluded and the model predictions were generated with the remaining samples. In general, the densities match the histograms well, showing that the model performs well in predicting the overall RT distributions. The top-right panel shows the good consistency between the data and predicted RT quantiles per item, with $r_{\mathrm{C}} = 0.982$ for correct responses and $r_{\mathrm{E}} = 0.929$ for error responses. The panels in the bottom two rows provide a more thorough inspection of the item-wise RT distributions. The correct RT quantiles match the data well for all items, while there are some mismatches in the error RT quantiles. The discrepancies can be explained by the fewer numbers of error observations (as shown as n_{E} in each panel) and the effect of the time limit of the task. Despite these limitations in the data, the model predictions of the 10–70% error RT quantiles match the data reasonably well. Taken together, the model shows a good absolute fit for the rotation data.

Having demonstrated good absolute fits of the DIRT-RV(η, s_{zr}) model to the extraversion and rotation data, we generated the model predictions of the CAFs underlying responses and RTs. As described in Sect. 2, the diffusion IRT model produces a single CAF prediction per set

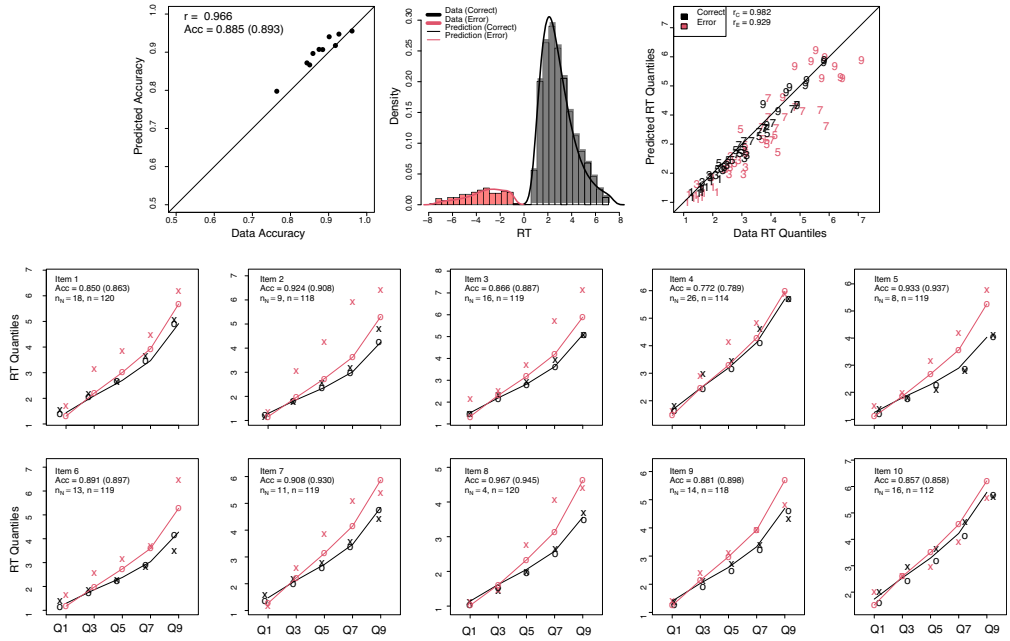

FIGURE 4.
Absolute model fits of the DIRT-RV(η,s_{zr}) model to rotation data. See the text or the caption of Fig. 3 for the detailed description of this figure (Color figure online).

of diffusion parameters such as drift rate and boundary separation. Thus, with P persons and I items, the model produces $P \times I$ different CAFs for the $P \times I$ pairs of persons and items. This makes the model capable of capturing heterogeneity in conditional dependence across responses although the capability of the model is limited to the differences in the person and item parameters estimated from the model fit (as the model has only a single η and a single s_{zr}, while drift rate and boundary separation are functions of the person and item parameters). As the absolute fit of the model was particularly good for the positive ('Yes') responses in the extraversion data and the correct responses in the rotation data, the CAF predictions were generated only for the person-item combinations with these responses. When generating the CAFs, the nondecision time parameters were fixed to 0 as they do not affect the trend of the CAFs. Thus, the predicted CAFs are based on the decision times, not on the whole RTs including nondecision times. Also, the CAFs are displayed on the logarithmic scale of the time as in Bolsinova and Molenaar (2018) and Chen et al. (2018a).

Figure 5 shows the predicted CAFs for the extraversion data (left) and the rotation data (right). In each panel, each black curve shows a CAF corresponding to one of the $P \times I$ person-item combinations (i.e., responses). For the extraversion data, most of the CAFs show a decreasing trend but with different slopes. The decrease is due to the large estimate of the random variability in drift rate ($\hat{\eta} = 2.085$). Although this single random variability parameter determines the overall decrease in the predicted CAFs, the slopes differ by person-item combination as different persons and items are associated with different boundary separations and drift rates. There are also some CAFs that grow from very low positive response proportions to about 0.5. Person-item pairs corresponding to these CAFs have higher probabilities of 'No' responses to given extraversion items. Both decreasing and increasing trends are from random variability in drift rate, which shows that conditional dependency is correlated with predicted response proportions. The estimate of the random variability in starting point was $\hat{s}_{zr} = 0.267$. It turns out that this estimate is too

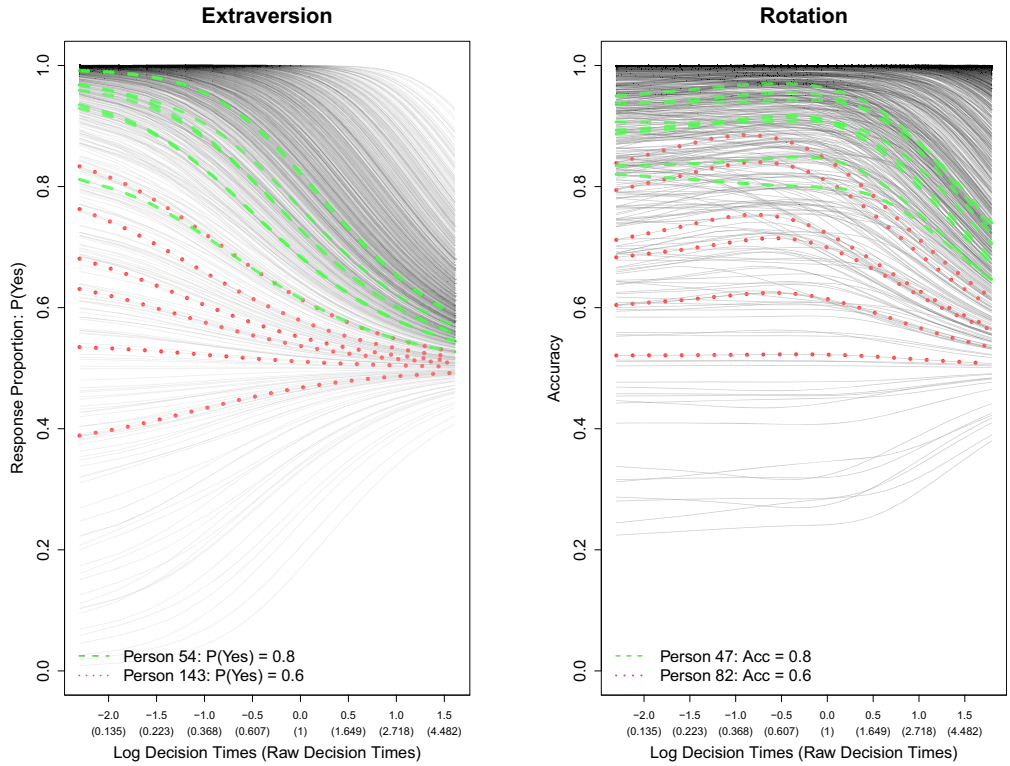

FIGURE 5.
The predicted conditional accuracy functions (CAFs) for the extraversion data (left) and the rotation data (right): In each panel, each black curve shows a CAF corresponding to one of the $P \times I$ person-item combinations (i.e., responses). The green dashed curves and the red dotted curves represent the CAFs of selected persons with different response proportions as shown at the bottom-left side of each panel. A single person has multiple CAFs each of which corresponds to an item. Acc: Accuracy (Color figure online).

small to make the first-increasing trend clearly appear in the CAFs, but it makes the slope of the decreasing trend less steep in the early RT period.

For the rotation data, the dominant pattern of the predicted CAFs is not monotone: accuracy increases first and then decreases as a function of the decision time. The random variability in starting point that determines the overall degree of the first-increasing trend is estimated as $\hat{s}_{zr} = 0.605$, which is fairly large. The estimate of the random variability in drift rate is also sufficiently large ($\hat{\eta} = 1.006$) so that it allows to predict the later-decreasing trend. Like those of the extraversion data, the predicted CAFs of the rotation data also show heterogeneity as the CAFs start to increase at different rates, reach the peak at different decision times, and also decrease at different rates. The curvilinear pattern appears when accuracy is generally high. There are also some CAFs whose accuracy is generally low and flat in the early RT period and later increases over time. As in the extraversion data, these predictions are from random variability in drift rate when the predicted drift rates for person-item combinations are negative. That is, the same source of conditional dependency predicts different patterns as a function of item difficulty.

Heterogeneity in the CAFs over all persons and items (i.e., differences across all the gray curves in Fig. 5) might be too ambiguous to provide meaningful information. This can be more thoroughly studied by taking some persons as examples and looking at their CAFs. In each panel of Fig. 5, the green dashed curves represent the CAFs of a selected person with the response

proportion (of positive responses in the extraversion data and of correct responses in the rotation data) of 0.8 and the red dotted curves represent the CAFs of another selected person with the response proportion of 0.6. For each person, different curves represent the CAFs corresponding to different items to which the person made positive or correct responses. These curves show that, even for a single person, conditional dependency can vary by item (i.e., item-specific dependency). Differences in the height and slope of the curves show that the response proportion and its change as a function of RT can differ due to item effects. In particular, when items are rather difficult for a person, the model predicts that the general trend of the CAF can entirely change. For example, the red dotted curves in the right panel show that a single person (Person 82) has curvilinear CAFs for most of the items, while the same person has rather flat CAFs with low accuracy for other items. A similar analysis can be done with some selected items as examples, which can show heterogeneity due to person effects.

Our results show that the best-fitting variant of the diffusion IRT model predicts previously observed trends of conditional dependency. In the extraversion data, negative dependency is dominant although there are some cases with positive dependency with low positive response proportions. These predictions are mainly from random variability in drift rate and its interaction with predicted item response proportions. In the rotation data, the prevalent pattern in the CAFs is curvilinear with a first-increasing and then decreasing trend. As in the Ratcliff diffusion model, this trend results from a combination of random variability in drift rate and in starting point. The former produces the decreasing (increasing) trend over time for easy (difficult) items, while the latter produces changes in CAFs in the early RT period. As in the extraversion data, predictions from random variability in drift rate show that conditional dependency can be positive or negative depending on item difficulty, which is consistent with previous findings (Bolsinova, De Boeck, & Tijmstra, 2017; Bolsinova, Tijmstra, Molenaar, & De Boeck, 2017; De Boeck & Jeon, 2019).

5. Discussion

In this paper, we extended the diffusion IRT model by implementing random variability in drift rate and in starting point. With this extension, the model can account for previously found conditional dependence between response accuracy and RT given latent variables and item parameters. Random variability in drift rate is a source of negative conditional dependency when the drift rate is positive (i.e., when an item is relatively easy for a person), while the same variability causes positive conditional dependency when the drift rate is negative (i.e., when an item is difficult for a person). Thus, drift rate variability explains both trends of dependency and also their correlation with item difficulty. Random variability in starting point produces positive conditional dependency in the early residual RT period when the drift rate is positive and also early negative conditional dependency when the drift rate is negative. By combining both the variability in drift rate and variability in starting point, the extended model can account for various conditional dependency patterns including the curvilinear trend. Although our extension included a single parameter for random variability in drift rate and another for random variability in starting point, the model can capture the heterogeneity of conditional dependency across persons and items. This is because the effects of variability parameters vary as a function of the other cognitive components of the model.

The model-based explanation of conditional dependency that we provided with the random variability extension is consistent with the interpretations proposed in earlier studies. Positive and negative conditional dependency and their interaction with item difficulty can be interpreted as outcomes of the variation in cognitive capacity (Chen et al., 2018a; De Boeck & Jeon, 2019), which is in accord with the model prediction with random variability in drift rate. This natural variation in the available capacity across time can occur due to random changes in the level of attention,

effort, or motivation, as well as due to person-by-item specificities of the tasks or expressions of a trait (e.g., extraversion). Although the size of the variation may differ by person or by item, it is plausible that the variation is a general phenomenon that applies to all persons and items. The emergence of positive conditional dependency, primarily at the beginning of the response process, was predicted by random variability in starting point. The variation in starting point is a natural process in that, even with an equal level of capacity, the starting point of a cognitive process may be closer or farther away from the correct response. This is most likely the case for inductive processes because these are necessarily based on repeated hypothesis testing, for knowledge-based processes because not all knowledge is equally available at a constant level of accessibility, and for complex processes with different possible approaches and trials. In a general test setting, responses to earlier items can induce an initial bias for response to a new item. A possible interpretation for the curvilinear conditional dependency between response accuracy and RTs is that it reflects the solution process of a respondent working on an item after individual difference parameters and item difference parameters are controlled for.

We also illustrated how to study sources and trends of conditional dependency with the diffusion IRT model. By comparing models with different assumptions, we can identify the presence of conditional dependency and its dominant sources. The plain diffusion IRT model without any variability component serves as a reference model. If a model with conditional dependency works better for data, this provides evidence for the dependency of response accuracy and RT unexplained by person and item effects. Provided that the conditional dependency model accounts for behavioral patterns of data, we can interpret sources of dependency assumed in the model as sources of dependency underlying response and RT data. Then, trends of conditional dependency can be visualized by CAFs predicted by the model.

We provided two empirical examples to describe the procedure illustrated above. With the extraversion data, we showed how to apply the procedure to personality measures. In this case, CAFs describe the pattern of response proportion, not accuracy. The full model with both variabilities in drift rate and in starting point was the best-fitting model, and its absolute model fit was also good. However, judging from the predicted CAFs, the dominant source of conditional dependency was random variability in drift rate as the functions of response proportion showed large decreasing trends for extraversion-oriented person-item pairs and large increasing trends for introversion-oriented person-item pairs. The variability of drift rate suggests that there are idiosyncratic aspects to the nature of extraversion depending on the individual person, with item-specific aspects of extraversion per person. The estimate of random variability in starting point was relatively small, and thus, the predicted CAFs did not clearly show the early positive/negative conditional dependency induced by variation in the initial bias. Our CAF analysis of the extraversion data also showed that RT is generally faster when a response is strongly in favor of either option (positive or negative) and slower when there is no preferred option. That is, more cognitive effort (i.e., evidence accumulation according to the diffusion model account) is required when it is unclear to a respondent if a presented extraversion-related word applies to the respondent's personality. A behavioral trend of residual dependency of other latent traits can be studied in the same way.

The rotation data provide an empirical application of the diffusion IRT model with random variability parameters to a cognitive ability test. The full model was also the best-fitting model for the rotation data and the model predictions accounted for item-wise accuracy and item-wise RT distributions. The predicted CAFs showed the curvilinear pattern, implying that both variability in the efficiency in information processing and variability in starting point of the problem-solving process were dominant sources of conditional dependence underlying the rotation data. Variability in starting point also suggests a partly random repeated trial strategy to solve the rotation problems. There is one caveat we should mention regarding our analysis result of the rotation data. There was a time limit imposed in this task which potentially affected characteristics of the data RT

distributions, and in turn, conditional dependency with response accuracy. Therefore, the result we obtained should be interpreted as a behavioral feature of the rotation task under time pressure rather than a general feature.

The curvilinear conditional dependency we obtained from the rotation data is consistent with the findings in Chen et al. (2018a) and Bolsinova and Molenaar (2018). Chen et al. (2018a) obtained a single CAF by aggregating all the data points and applying double-centering. Although it is not guaranteed that centering by person-wise and item-wise mean is sufficient to control for person and item effects, the curvilinear pattern shown in Chen et al.'s work corresponds to conditional dependency at an 'aggregate' level (i.e., across all person-item pairs or across all responses) Bolsinova and Molenaar (2018) found the curvilinear relationship between an item intercept (which has a positive relationship with accuracy) and standardized log residual RTs, providing evidence for the curvilinear relationship at the 'item' level. In contrast, the model predictions from the diffusion IRT model illustrate trends of conditional dependency for a person-item pair (the 'response' level). Although there are some differences in approach, the two earlier studies and our current work provide evidence for the curvilinear pattern of conditional dependency, and at the different levels: aggregate, item, and response levels.

In our modeling approach, we defined a drift rate of evidence accumulation as the difference between a person-wise drift rate and an item difficulty parameter, in line with the diffusion IRT model as defined by Tuerlinckx and De Boeck (2005). van der Maas et al. (2011) have proposed an alternative parametrization of the drift rate as the quotient of a person-wise drift rate and an item difficulty, with both parameters constrained to be positive (i.e., their 'Q-diffusion IRT' model as an alternative for the 'D-diffusion IRT' model by Tuerlinckx and De Boeck). The authors explained that, unlike the practice of most other IRT and factor-analytic models, abilities need to be parameterized in the positive range of numerical values. Although this is a valuable alternative, we have not followed the Q-diffusion IRT approach for the following reasons. First, the drift rate parameter of the diffusion model as conceived and used in cognitive psychology has a range that comprises the real line and thus negative values as well. In the context of testing, a negative drift rate corresponds to the tendency to move to the opposite response option and it is a conceptual issue whether the reasoning of van der Maas et al. (2011) regarding ability can be applied to the drift rate. Second, as far as drift rate can be interpreted in terms of ability, we side with a large majority of psychometric models and work with positive and negative values for latent abilities. Third, the positive ability assumption implies that the lowest response accuracy that the quotient parameterization can predict is 0.5 for binary-choice items unless the model is further adjusted based on assumptions for the multiple-choice format (which does not apply to our data). Thus, without a modification, the Q-diffusion IRT model cannot account for difficult items with accuracy less than 0.5. Nonetheless, the random variation extension can also be applied to the Q-diffusion IRT model. However, this would be more restricted than our current extension because a negative drift rate is not allowed and some patterns of dependency cannot be explained. For example, Chen et al. (2018b) showed that, for difficult items, the conditional accuracy function first decreases and then increases. This is consistent with the prediction from the D-diffusion IRT model with a negative drift rate (the bottom-half of Fig. 2), but the positive quotient drift rate cannot account for this.

Although we only implemented a single random variability parameter that works for all persons and items, it is also possible to extend the model with multiple variability parameters. For example, item-wise variability in drift rate and in starting point can be implemented given responses from a large number of persons to each item. In fact, De Boeck et al. (2017) provided evidence for the item-specific nature of conditional dependency. This implies the potential of modeling item-wise variability in cognitive components in the study of conditional dependency. The estimation of a person-wise variability parameter is rather unrealistic since it requires responses from a single person to hundreds of items, which is not typically done in psychometrics. Also,

as Bolsinova, Tijmstra , Molenaar, and De Boeck (2017) stated, variation of person components across items cannot easily be distinguished from the variation of item components across persons because both of them represent the same interaction effect of persons and items.

Conditional dependence is a complicated feature of responses and RTs, and its characteristics can vary by different tests and inventories. For example, cognitive tests may involve multiple heterogeneous processes and conditional dependency underlying these tests can be much different from dependency underlying tests involving a single process. The diffusion IRT model assumes a single process, namely the evidence accumulation process, and thus, the current model may not provide the best account for psychometric measurements with multiple heterogeneous processes. However, variability in cognitive components can potentially capture the unexplained heterogeneity in psychological processes and the diffusion model with variability extension can provide a reasonable approximation to complex psychological processes. For example, it has been under debate whether the mental rotation entails a single process or multiple heterogeneous processes (Cooper & Shepard, 1973; Shepard & Cooper, 1982; Shepard & Metzler, 1971). If the latter is the case, it might be possible to find a multi-process model to provide a better description of conditional dependency underlying the mental rotation. However, the diffusion IRT model showed a good absolute model fit to the rotation data we examined and this implies that the model with the variability extension can provide a good account for or a good approximation to the mental rotation processes. Similarly, different features of psychometric measurement such as response modalities (e.g., binary response vs multiple-option response) and time pressure can influence how responses affect RTs and vice versa. We believe identifying and modeling this potential heterogeneity of conditional dependence will provide us with insightful information for more accurate and precise measurement of psychological constructs and a better understanding of cognitive processes underlying item response behavior.

Declarations

Code Availability The R program to fit the diffusion IRT model with random variability can be found online at https://osf.io/vg2nf/.

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

References

Birnbaum, A. (1969). Statistical theory for logistic mental test models with a prior distribution of ability. *Journal of Mathematical Psychology, 6*(2), 258–276. https://doi.org/10.1016/0022-2496(69)90005-4

Blurton, S. P., Kesselmeier, M., & Gondan, M. (2017). The first-passage time distribution for the diffusion model with variable drift. *Journal of Mathematical Psychology, 76*, 7–12. https://doi.org/10.1016/j.jmp.2016.11.003

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., & Maris, G. (2016). A test for conditional independence between response time and accuracy. *British Journal of Mathematical and Statistical Psychology, 69*(1), 62–79. https://doi.org/10.1111/bmsp.12059

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, 111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061
Borst, G., Kievit, R. A., Thompson, W. L., & Kosslyn, S. M. (2011). Mental rotation is not easily cognitively penetrable. *Journal of Cognitive Psychology, 23*(1), 60–75. https://doi.org/10.1080/20445911.2011.454498
Brown, S., & Steyvers, M. (2005). The dynamics of experimentally induced criterion shifts. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(4), 587–599. https://doi.org/10.1037/0278-7393.31.4.587
Chen, H., De Boeck, P., Grady, M., Yang, C.-L., & Waldschmidt, D. (2018a). Curvilinear dependency of response accuracy on response time in cognitive tests. *Intelligence, 69*, 16–23. https://doi.org/10.1016/j.intell.2018.04.001
Chen, H., De Boeck, P., Grady, M., Yang, C.-L., & Waldschmidt, D. (2018b). A bifactor approach to modeling dependencies between response time and accuracy. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME).
Cooper, L. A., & Shepard, R. N. (1973). Chronometric studies of the rotation of mental images. In W. G. Chase (Ed.), *Visual information processing* (pp. 75–176). Academic Press. https://doi.org/10.1016/B978-0-12-170150-5.50009-3.
Cox, D., & Miller, H. D. (1970). *The theory of stochastic processes*. London: Methuen.
De Boeck, P., Chen, H., & Davison, M. (2017). Spontaneous and imposed speed of cognitive test responses. *British Journal of Mathematical and Statistical Psychology, 70*(2), 225–237. https://doi.org/10.1111/bmsp.12094
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. (2021). Response Time relationships within examinees: Implications for item response time models. In M. Wiberg, D. Molenaar, J. González, U. Böckenholt, & J. S. Kim (Eds.), *Quantitative psychology*. Springer Proceedings in Mathematics & Statistics (Vol. 353). Springer, Cham. https://doi.org/10.1007/978-3-030-74772-5_5
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., Dunson, D. B., & A. Vehtari, D. B. R. (2013). Bayesian data analysis (3rd ed.). CRC Press.
Goldhammer, F., Naumann, J., & Greiff, S. (2015). More is not always better: The relation between item response and item response time in raven's matrices. *Journal of Intelligence, 3*(1), 21–40. https://doi.org/10.3390/jintelligence3010021.
Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology, 106*(3), 608–626. https://doi.org/10.1037/a0034716
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
Laming, D. R. J. (1968). *Information theory of choice-reaction times*. Academic Press.
Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. Oxford University Publication.
Maris, G., & van der Maas, H. L. J. (2012). Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika, 4*, 615–633. https://doi.org/10.1007/s11336-012-9288-y
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
Meng, X.-B., Tao, J., & Chang, H.-H. (2015). A conditional joint modeling approach for locally dependent item responses and response times. *Journal of Educational Measurement, 52*(1), 1–27. https://doi.org/10.1111/jedm.12060
Molenaar, D., Bolsinova, M., Rozsa, S., & De Boeck, P. (2016). Response mixture modeling of intraindividual differences in responses and response times to the Hungarian wisc-iv block design test. *Journal of Intelligence*. https://doi.org/10.3390/jintelligence4030010
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
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
Ranger, J., & Kuhn, J.-T. (2018). Estimating diffusion-based item response theory models: Exploring the robustness of three old and two new estimators. *Journal of Educational and Behavioral Statistics, 43*(6), 635–662. https://doi.org/10.3102/1076998618787791
Ranger, J., Kuhn, J.-T., & Szardenings, C. (2016). Limited information estimation of the diffusion-based item response theory model for responses and response times. *British Journal of Mathematical and Statistical Psychology, 69*(2), 122–138. https://doi.org/10.1111/bmsp.12064
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
Ranger, J., Kuhn, J.-T., & Szardenings, C. (2020). Minimum distance estimation of multidimensional diffusion-based item response theory models. *Multivariate Behavioral Research, 55*(6), 941–957. https://doi.org/10.1080/00273171.2019.1704676
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., & Childers, R. (2015). Individual differences and fitting methods for the two-choice diffusion model of decision making. *Decision, 2*, 237–279.
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., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation, 20*(4), 873–922.
Ratcliff, R., & McKoon, G. (2018). Modeling numerosity representation with an integrated diffusion model. *Psychological Review, 125*(2), 183–217. https://doi.org/10.1037/rev0000085
Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science, 9*(5), 347–356. https://doi.org/10.1111/1467-9280.00067
Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences, 20*(4), 260–281. https://doi.org/10.1016/j.tics.2016.01.007
Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review, 106*(2), 261–300. https://doi.org/10.1037/0033-295x.106.2.261
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
San Martín, E., del Pino, G., & De Boeck, P. (2006). IRT models for ability-based guessing. *Applied Psychological Measurement, 30*(3), 183–203. https://doi.org/10.1177/0146621605282773
Shepard, R. N., & Cooper, L. N. (1982). *Mental images and their transformations*. MIT Press.
Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science, 171*(3972), 701–703. https://doi.org/10.1126/science.171.3972.701
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
Stone, M. (1960). Models for choice-reaction time. *Psychometrika, 25*, 251–260. https://doi.org/10.1007/BF02289729
Swensson, R. G. (1972). The elusive tradeoff: Speed vs accuracy in visual discrimination tasks. *Perception & Psychophysics, 12*, 16–32. https://doi.org/10.3758/BF03212837
Ter Braak, C. J. F. (2006). A Markov Chain Monte Carlo version of the genetic algorithm differential evolution: Easy Bayesian computing for real parameter spaces. *Statistics and Computing, 16*, 239–249.
Tuerlinckx, F. (2004). The efficient computation of the cumulative distribution and probability density functions in the diffusion model. *Behavior Research Methods, Instruments, and Computers, 36*(4), 702–716. https://doi.org/10.3758/BF03206552
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., Brown, S. D., & Steyvers, M. (2013). A method for efficiently sampling from distributions with correlated dimensions. *Psychological Methods, 18*(3), 368–384.
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 Linden, W. J., & Glas, C. A. W. (2010). Statistical tests of conditional independence between responses and/or response times on test items. *Psychometrika, 75*, 120–139. https://doi.org/10.1007/s11336-009-9129-9
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 Rijn, P. W., & Ali, U. S. (2017). A comparison of item response models for accuracy and speed of item responses with applications to adaptive testing. *British Journal of Mathematical and Statistical Psychology, 70*(2), 317–345. https://doi.org/10.1111/bmsp.12101
van Rijn, P. W., & Ali, U. S. (2018). A generalized speed-accuracy response model for dichotomous items. *Psychometrika, 83*(1), 109–131. https://doi.org/10.1007/s11336-017-9590-9
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

Manuscript Received: 30 MAR 2021
Final Version Received: 5 SEP 2021
Published Online Date: 6 JAN 2022