Modeling Conditional Dependence of Response Accuracy and Response Time with the Diffusion Item Response Theory Model – Supplementary Material

This supplementary material presents a simulation study conducted to test if the diffusion IRT model with random variability (DIRT-RV) can recover its parameters (Section 4 of the article). In our first simulation, 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 randomly generated or set true values of data-generating parameters as follows (Table S1): $\log(a_i) \sim U(-1,0)$, $b_i \sim U(-1.5, 0.5)$, $\log(\gamma_p) \sim N(0, 0.5)$, $\theta_p \sim N(0,1)$, $\tau_p \sim N(0.3, 1)$, $\eta = 1$, and $s_{zr} = 0.5$ where $U(a,b)$ indicates a uniform distribution with the range of (a,b) and $N(\mu_1, \sigma_1)$ indicates a normal distribution with mean μ_1 and standard deviation σ_1.

To estimate the joint posterior distributions of the model parameters, prior distributions should be specified (Table S1). For person boundary separation and drift rate, we used $\log(\gamma_p) \sim N(0,1)$ and $\theta_p \sim N(0,1)$. We set the standard deviations in these prior distributions to 1 as done in Van der Mass et al. (2011). For person-wise nondecision time, a natural choice was to put $\tau_p \sim U\left(0, \min_i(T_{pi})\right)$ since the nondecision time of person p should be smaller than the minimum RT of the same person. For item time-pressure and difficulty parameters, we set $a_i \sim U(0, 1.5)$ and $b_i \sim N(0,2)$, respectively, which have sufficiently wide range and large variance to cover typical ranges of the parameter estimates (Molenaar, Tuerlinckx, & Van der Mass, 2015). For the two variability parameters, we chose to use $\eta \sim TN(0,3;0,\infty)$ and $s_{zr} \sim Beta(3,2)$ where $TN(\mu_2, \sigma_2; c, d)$ indicates a truncated normal distribution with μ_2 and σ_2 as the mean and standard deviation of a normal random variable, respectively, and (c,d) as the range of the distribution, and $Beta(m,n)$ indicates a beta distribution with two shape parameters m and n. The truncated normal prior for η is noninformative or weakly informative with a sufficiently large variance but with a constraint to make the variance component positive. The beta prior for s_{zr} is weakly informative in that it favors intermediate values with mean of 0.6 and puts a little weight on extreme values close to 0 or 1.

Parameters	a_i	b_i	τ_p	$log(\gamma_p)$	θ_p	η	s_{zr}
Data-generating	$\exp(U(-1,0))$	$U(-1.5, 1.5)$	$U(0.3, 1)$	$N(0, 0.5)$	$N(0,1)$	1	0.5
Prior	$U(0, 1.5)$	$N(0,2)$	$\tau_p \sim U\left(0, \min_i(T_{pi})\right)$	$N(0,1)$	$N(0,1)$	$TN(0,3;0,$	$Beta(3,2)$

Table S1. **Data-generating and Prior Distributions for the Simulation Study**.

With the prior specification above and the model likelihood computed using Equation 5, we implemented a Bayesian estimation method to fit the model to the generated data. We used the Differential Evolution Markov Chain Monte Carlo (DE-MCMC; Ter Braak, 2006; Turner et al., 2013) sampling method in which different chains interact with each other to efficiently approximate the joint posterior distribution. The key idea is to generate a new proposal for one chain based on the difference between the current states of some other chains. This sampling method is known to perform better than a conventional MCMC method for models with a correlated parameter space (Turner et al., 2013), as is the case with the model we examine in this article. We ran the sampling algorithm for 10,000 iterations with 8 chains to obtain the joint posterior distribution of the model parameters. We discarded the first 2,000 samples as burn-in. For the first half of the burn-in iterations, we implemented the migration step with the migration probability of 0.2. Finally, we assessed convergence by visually inspecting if the posterior densities of each parameter obtained by chain matched with each other and the Gelman-Rubin convergence diagnostic ($\hat{R}$) which was estimated as smaller than 1.1 for all model parameters (Gelman, 1996; Gelman et al., 2013).

To evaluate the parameter recovery, we obtained the Maximum A Posteriori (MAP) estimates of the item and person parameters and plotted them against the true parameter values (top panels in Figure S1). The parameter plotted in each panel is labeled on top: item-pressure (a_i; log-transformed), item difficulty (b_i), person-wise nondecision time (τ_p), person boundary separation (γ_p; log-transformed), and person drift rate (θ_p), respectively, from the top-left panel to the bottom-right panel. At the top-left side of each panel, the Pearson correlations (r) between the MAP estimates and true parameter values are shown as 0.972 ($\log(a_i)$), 0.988 (b_i), 0.954 (τ_p), 0.941 ($\log(\gamma_p)$), and 0.868 (θ_p), respectively. For all parameters, the MAP estimates are consistent with their parameter values without any noticeable bias. In general, the result shows that the model can recover the item and person parameters reasonably well.

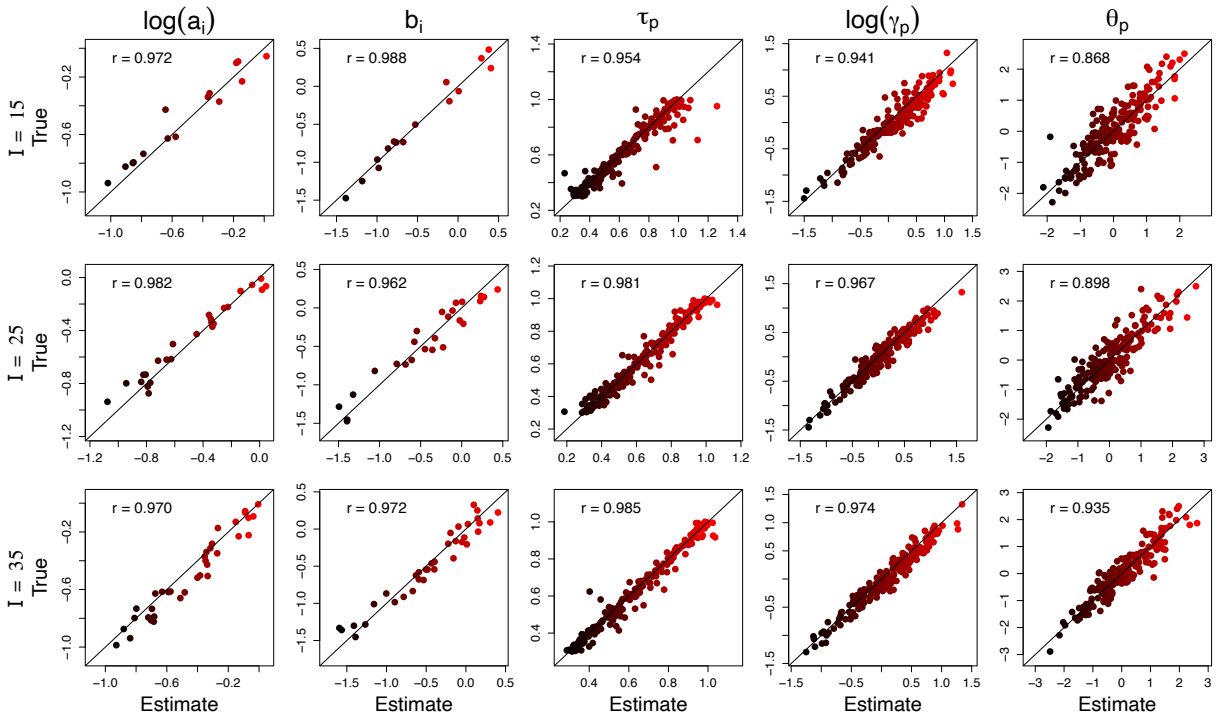

Figure S1. **Parameter Recovery of the Diffusion IRT Model with Random Variability: Item and Person Parameters.** Each row presents recovery of item and person parameters with a different number of items I (shown on the left). In the five panels, the Maximum A Posteriori (MAP) estimates of the parameters are plotted on the x-axis against the true parameter values on the y-axis. The parameter plotted in each panel is denoted on the top: item-pressure (a_i; log-transformed), item difficulty (b_i), person-wise nondecision time (τ_p), person boundary separation (γ_p; log-transformed), and person drift rate (θ_p), respectively, from top-left to bottom-right. In each panel, the Pearson correlation (r) between the MAP estimate and true parameter values is presented at the top-left side.

Figure S2 (top panels) shows the posterior distributions of random variability in drift rate η (left) and those of random variability in starting point s_{zr} (right). Each curve in a panel shows a posterior density of the parameter obtained from each chain (i.e., eight density estimates from the eight Bayesian chains). The consistency among the chains shows that there is no convergence issue in the estimation and the model can reach the same solution regardless of how chains are initialized. The gray vertical line in each panel represents the true value of the random variability parameter ($\eta = 1$ and $s_{zr} = 0.5$). The MAP estimates of the two random variability parameters are $\hat{\eta}^{MAP} = 1.204$ and $\hat{s}_{zr}^{MAP} = 0.467$, which are fairly close to the true values but show slight bias. Note that the general central tendency and spread of the bivariate RT distributions are predicted by the main person and item parameters and the variability parameters provide additional improvement in model fit over and above this prediction, by producing the best balance between the correct and error RT distributions. Capturing the relative speed of correct and error RTs is difficult, particularly when there are fewer error observations because error RTs

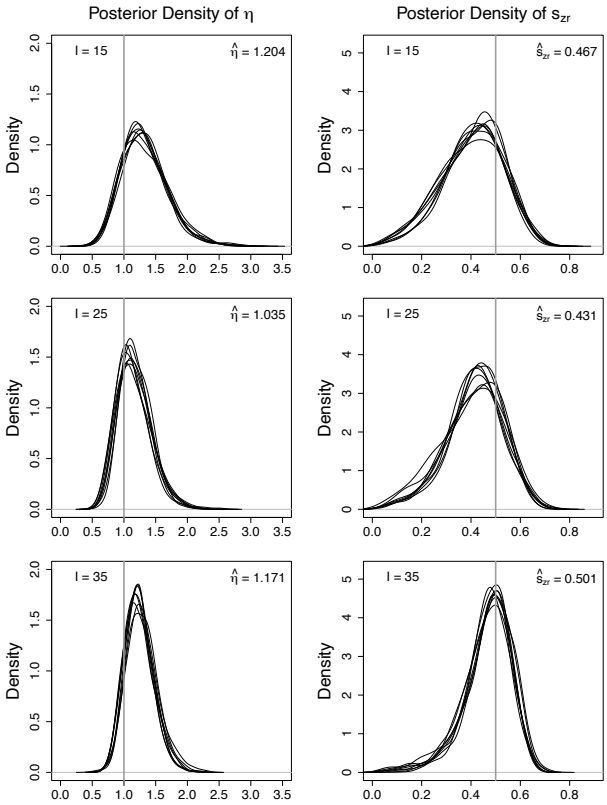

Figure S2. **Parameter Recovery of the Diffusion IRT Model with Random Variability: Variability Parameters.** Each row presents the posterior distributions of random variability in drift rate η (left) and those of random variability in starting point s_{zr} (right) obtained with a different number of items I (shown on the top-left side of each panel). Different curves in the same panel show posterior densities of the parameter obtained from different chains. The gray vertical line in each panel represents the true value of the random variability parameter ($\eta = 1$ and $s_{zr} = 0.5$).

can be much more variable in this case (Ratcliff & Tuerlinckx, 2002; Ratcliff & McKoon, 2008). This makes the variability parameters harder to precisely estimate compared to the person and item parameters. Considering this model mechanism, it can be concluded that the recovery of the variability parameters is reasonably good.

There would be potential ways to improve recovery of the model parameters (particularly, the variability parameters). For example, increasing sample sizes can be beneficial. As we focus on the joint estimation of person and item parameters, more persons (items) in the data require more person (item) parameters to estimate. As our first simulation result showed that the item parameters can be recovered well with $P = 200$, we examined the recovery with more items. To this end, we simulated data with $I = 25$ and $I = 35$ and fitted the model. Due to the computational cost, we conducted the recovery simulation once per condition without repetition.

The middle and bottom panels in Figure S1 show recovery of the item and person parameters and those in Figure S2 show recovery of the variability parameters, with $I = 25$ and $I = 35$, respectively. In addition, Table S2 ('Non-hierarchical Prior' Columns) shows the MAP estimates and 95% credible interval (in parentheses) of the variability parameters and Table S3 ('Non-hierarchical Prior' Columns) shows the Pearson correlations between true parameter values and their estimates and the mean squared errors (MSEs) of estimation obtained with three different number of items conditions.

Generally, an increase in the number of items improved the estimation of person parameters as shown by increases in the Pearson correlation and decreases in MSE. Also, the variability parameters were more precisely estimated (narrower posterior densities) with more items. It can be inferred that, as the person parameters are more stably estimated, the related variability parameters can also be more reliably estimated. The point estimates of the variability parameters were not necessarily closer to the true value when the dataset was larger. This would be due to the sampling variability as we conducted a simulation only once per condition.

Another estimation method that can probably improve the estimation quality is to implement a hierarchical structure. For example, when the model includes a hierarchical population distribution on person-wise drift rates and boundary separations (e.g., a bivariate normal distribution), an individual person parameter can be better estimated due to additional information from other persons. Also, explicitly modeling the correlation between the two person-wise parameters across persons can help improve the estimation. Similarly, an over-arching item distribution can be implemented in the model. In this study, we only examined the population distribution because item parameters were generally more accurately and precisely estimated in our earlier simulation. As done with non-hierarchical priors, we fitted the model to the simulated data with different numbers of items ($I = 15$, 25, and 35) and the correlation between person drift rates and boundary separations of 0.2.

	Non-hierarchical Prior		**Hierarchical Prior**	
$\boldsymbol{I}$	$\boldsymbol{\eta}$	$\boldsymbol{s_{zr}}$	$\boldsymbol{\eta}$	$\boldsymbol{s_{zr}}$
15	1.204 (0.650, 2.018)	0.467 (0.159, 0.641)	0.859 (0.419, 1.717)	0.379 (0.097, 0.616)
25	1.035 (0.692, 1.676)	0.457 (0.174, 0.623)	1.038 (0.631, 1.766)	0.393 (0.109, 0.585)
35	1.164 (0.825, 1.730)	0.501 (0.266, 0.627)	1.058 (0.737, 1.755)	0.453 (0.184, 0.626)

Table S2. **Maximum A Posterior (MAP) Estimates and 95% Credible Intervals (in Parentheses) of the Variability Parameters.** The left two columns show the values obtained from the model fit with non-hierarchical priors on the person parameters and the right two columns show those with a hierarchical population prior.

		Non-hierarchical Prior							Hierarchical Prior						
	I	$log(a_i)$	b_i	τ_p	$log(\gamma_p)$	θ_p	η	s_{zr}	$log(a_i)$	b_i	τ_p	$log(\gamma_p)$	θ_p	η	s_{zr}
r	**15**	0.972	0.988	0.954	0.941	0.868			0.985	0.971	0.960	0.954	0.841		
	25	0.991	0.963	0.981	0.968	0.899			0.993	0.957	0.971	0.967	0.909		
	35	0.972	0.971	0.985	0.971	0.935			0.992	0.956	0.988	0.975	0.937		
MSE	**15**	0.085	0.112	0.010	0.140	0.542	0.170	0.017	0.041	0.210	0.008	0.083	0.657	0.150	0.033
	25	0.075	0.171	0.004	0.105	0.443	0.068	0.015	0.035	0.219	0.005	0.064	0.487	0.095	0.027
	35	0.105	0.176	0.003	0.112	0.364	0.082	0.009	0.033	0.243	0.003	0.057	0.438	0.076	0.015

Table S3. **Pearson Correlations between True Parameter Values and Their Estimates (r) and Mean Squared Errors (MSEs).** The left seven columns show the values obtained from the model fit with non-hierarchical priors on the person parameters and the right seven columns show those with a hierarchical population prior.

In general, the estimation results with a hierarchical distribution were very similar to our earlier results with non-hierarchical priors ('Hierarchical Prior' columns in Tables S2 and S3). One thing to note is that when the number of items is small (e.g., $I = 15$), there would not be sufficient information in the data to appropriately constrain person-wise parameters and so their estimates can shrink toward the group mean (which is fixed to 0 for identifiability). The underestimation bias can also propagate to the variability parameter estimation. This would not happen when the number of items is not too small, with which the estimation results are similar to those obtained with non-hierarchical priors. Different parameters can be better or worse estimated in this case. However, it should be noticed that this simulation study is limited in that the model fitting was conducted only once per condition and per prior specification due to the computational cost.

Although we focus on the joint estimation of person and item parameters, a marginalized method such as the Marginal Maximum Likelihood (MML) can also be applied to fit the proposed model. With all person parameters (latent traits/abilities) marginalized out, this method can potentially improve the estimation of the variability parameters. However, the improvement would be at the expense of losing the uncertainty information of the person-wise parameters, which are of primary importance in the diffusion modeling.

In the current simulation study, we tried to generate data with a reasonable choice of parameter values. Our choice was motivated by the data-generating distributions implemented in simdiff() function in the diffIRT package in R (Molenaar et al., 2015). For a diffusion IRT model with drift rate defined as a difference between person drift rate and item difficulty (threshold) parameter, this function by default generates main item and person parameter values with $\log(a_i) \sim U(-1,0)$, $b_i \sim U(-1.5, 0.5)$, $\gamma_p \sim LN(0, 0.3)$, and $\theta_p \sim N(0, 0.5)$. For the item parameters, we used the same distributions and for the person parameters, we used similar distributions but with wider ranges. However, it can be questioned if the model can recover its parameters from different distributions. For example, one of our reviewers pointed out that our prior choice for (log-transformed) boundary separation $log(\gamma_p)$ may favor the

true data-generating values. To address this issue, we conducted another simulation study with different data-generating distributions for $log(\gamma_p)$: $N(0, 0.25)$ and $N(0, 1)$. Our original choice was $N(0, 0.5)$ while the prior distribution is $N(0, 1)$. Thus, this simulation can show how the estimation quality changes as a function of a prior choice.

Figures S3 and S4 show the results. The top, middle, and bottom panels in these figures present the recovery results when the true data-generating values for $log(\gamma_p)$ are sampled with standard deviations of 0.25, 0.5 (the original result presented in Figures S1 and S2), and 1, respectively. In general, the results show that the model performed well in parameter recovery across the conditions examined. When the prior distribution for $log(\gamma_p)$ was much wider than its true underlying distribution (top panels), the estimation of $log(\gamma_p)$ got worse but still produced a reasonable recovery.

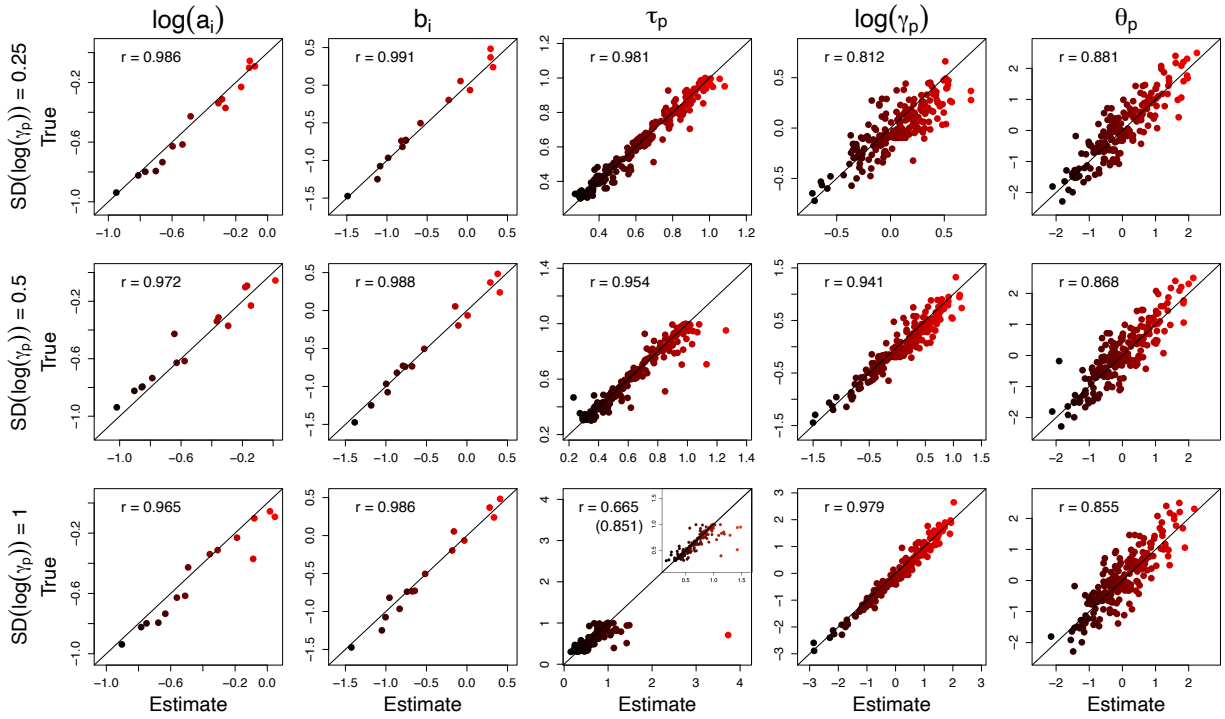

Figure S3. **Parameter Recovery of the Diffusion IRT Model with Random Variability: Item and Person Parameters.** Each row presents recovery of item and person parameters with a different choice of the standard deviation in the data-generating distribution for $log(\gamma_p)$ (shown on the left). In the five panels, the Maximum A Posteriori (MAP) estimates of the parameters are plotted on the x-axis against the true parameter values on the y-axis. The parameter plotted in each panel is denoted on the top: item-pressure (a_i; log-transformed), item difficulty (b_i), person-wise nondecision time (τ_p), person boundary separation (γ_p; log-transformed), and person drift rate (θ_p), respectively, from top-left to bottom-right. In each panel, the Pearson correlation (r) between the MAP estimate and true parameter values is presented at the top-left side.

Also, changes in the standard deviations of the data-generating distribution of $log(\gamma_p)$ affected the estimation of nondecision time τ_p. This is because the estimates of τ_p are highly associated with the minimum RT of the subject p (across items). When $log(\gamma_p)$ values are generally small, persons are more likely to have their minimum RTs close to the nondecision time. In contrast, when the distribution of $log(\gamma_p)$ is wide and some persons have large boundary separations, their RTs for all item responses can largely deviate from their nondecision time. For these persons, τ_p would be overestimated. Consistent with this relation, the result shows that the estimation of τ_p got worse when the distribution for $log(\gamma_p)$ is wider. Particularly, when the standard deviation of $log(\gamma_p)$ was 1, there was one simulated person with a too large γ_p; $\gamma_p = 14.142$ and $log(\gamma_p) = 2.649$ while the second largest values were $\gamma_p = 7.180$ and $log(\gamma_p) = 1.971$[1]. The nondecision time estimate for this person largely deviated from its true value (the right-most red dot in the bottom middle panel in Figure S3). However, excluding this person, the recovery was reasonably good with $r = 0.851$ (a scatter plot is presented in the inset on the top-right side of the bottom middle panel in Figure S3). In practice, a good estimation of nondecision time can be secured when there are more items. Alternatively, the item set can intentionally include some items/trials to explicitly measure nondecision time only (e.g., an item asking for a response without actual problem-solving).

[1] It would be informative to provide empirical values of person-wise boundary separation to evaluate the appropriateness of the distribution examined in this study. In our empirical applications, most person-wise boundary separation (log-transformed) estimates are between -1 and 1 (except for two persons in the extraversion data and two persons in the rotation data). The maximum estimate of $\log(\gamma_p)$ was 1.149 (γ_p of 3.156) in the extraversion data and 1.184 (γ_p of 3.267) in the rotation data. In another independent project, the model was applied to analogies and matrix reasoning data that have much longer RTs. The maximum estimate of $\log(\gamma_p)$ was 0.519 (γ_p of 1.680) in the analogies data and 0.384 (γ_p of 1.468) in the matrix reasoning data and so generally person-wise boundary separations and individual differences were smaller. Instead, longer data RTs were accounted for by smaller item-wise time-pressure parameters (a_i).

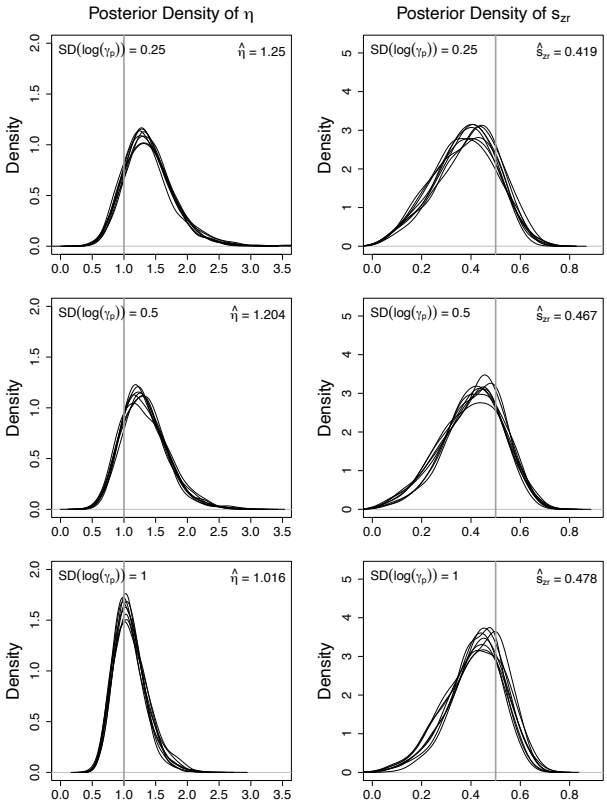

Figure S4. **Parameter Recovery of the Diffusion IRT Model with Random Variability: Variability Parameters.** Each row presents the posterior distributions of random variability in drift rate η (left) and those of random variability in starting point s_{zr} (right) obtained a different choice of the standard deviation in the data-generating distribution for $log(\gamma_p)$ (shown on the top-left side of each panel). Different curves in the same panel show posterior densities of the parameter obtained from different chains. The gray vertical line in each panel represents the true value of the random variability parameter ($\eta = 1$ and $s_{zr} = 0.5$).

Code Availability

The R scripts to fit the diffusion IRT model with random variability is available online at

https://osf.io/vg2nf/

References

Gelman, A. (1996). Inference and monitoring convergence. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain monte carlo in practice* (p. 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.

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. doi: 10.18637/jss.v066.i04

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.

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 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. doi: 10.1080/20445911.2011.454498