A Latent Space Diffusion Item Response Theory Model to Explore Conditional Dependence between Responses and Response Times - Supplementary Material

S1. Stan code for the LSDIRT model

```
data {
  int<lower = 1> P;                 // number of persons
  int<lower = 1> I;                 // number of items
  int<lower = 1> N;                 // number of person-item pairs
  int<lower = 1, upper = P> pp[N];  // person index for the n-th obs
  int<lower = 1, upper = I> ii[N];  // item index for the n-th obs
  int<lower=0, upper=1> resp[N];    // response in the long form
  real<lower=0> rt[N];              // RT in the long form
  vector[P] minRT;                  // subject-wise min RTs
  real mu[2];       // two prior means for lambda: (-5, 0.5) in this article
  real kappa[2];    // two prior SDs for lambda: (1, 1) in this article
}
parameters {
  // person and item parameters
  vector[P] std_theta;
  vector[P] std_log_gamma;
  vector<lower = 0, upper = 1>[P] tp_rel;
  vector[I] log_a;
  vector[I] b;

  // coordinates in the latent space
  vector[P] xi1
  vector[P] xi2;
  vector[I] zt1;
  vector[I] zt2;

  // person-distribution parameters
  vector<lower = 0>[2] sigma_sq;
  cholesky_factor_corr[2] Lcorr;

  // tuning parameter for the latent space
```

```
  real log_lambda;
  real<lower = 0, upper = 1> pind;  // omega in the article
}
transformed parameters{
  vector[P] theta;
  vector[P] log_gamma;
  vector[P] gamma;
  vector[P] theta;
  vector[I] a;
  vector[N] alpha;
  vector[N] nu;
  vector[N] tau;
  vector[N] dist;
  real lambda;
  real rho_cor;
  real<lower=0> sigma_theta;
  real<lower=0> sigma_gamma;
  corr_matrix[2] Sigma_cor;

  Sigma_cor   = multiply_lower_tri_self_transpose(Lcorr);
  sigma_gamma = sqrt(sigma_sq[1]);
  sigma_theta = sqrt(sigma_sq[2]);
  rho_cor     = Sigma_cor[1,2];

  // non-centered parameterization (Stan user's guide 25.7. Reparameterization)
  log_gamma = std_log_gamma * sigma_gamma;
  theta     = std_theta * sqrt((1 - sigma_cor^2) * sigma_theta^2) + sigma_cor *
      sigma_theta / sigma_gamma * log_gamma;
  tp        = minRT .* tp_rel;
  gamma     = exp(log_gamma);
  a         = exp(log_a);
  lambda    = exp(log_lambda);  // the distance effect tuning parameter

  // parameters corresponding to each response (person-item pair) in the long form
  for (n in 1:N){
```

```
    dist[n]  = sqrt((xi1[pp[n]] - zt1[ii[n]])^2 + (xi2[pp[n]] - zt2[ii[n]])^2);
    alpha[n] = gamma[pp[n]] / a[ii[n]];
    tau[n]   = tp[pp[n]];
    if(resp[n] == 1){
      nu[n]      = (theta[pp[n]] - b[ii[n]] - lambda * dist[n]);
    } else {
      nu[n]      = - (theta[pp[n]] - b[ii[n]] - lambda * dist[n]);
    }
  }
}
model {
  vector[2] lps;
  lps[1] = log(1-pind);
  lps[2] = log(pind);

  // prior distributions
  sigma_sq ~ cauchy(0, 2.5);
  Lcorr    ~ lkj_corr_cholesky(1);
  std_theta     ~ std_normal();
  std_log_gamma ~ std_normal();
  log_a  ~ normal(0, 5);
  b      ~ normal(0, 5);
  tp_rel ~ uniform(0, 1);
  xi1 ~ std_normal();
  xi2 ~ std_normal();
  zt1 ~ std_normal();
  zt2 ~ std_normal();
  pind ~ beta(1,1);

  // target distribution
  for(s in 1:2){ lps[s] += normal_lpdf(log_lambda | mu[s], kappa[s]); }
  target += log_sum_exp(lps) + wiener_lpdf(rt | alpha, tau, 0.5, nu);
}
```

S2. Simulation Study

In this section, we provide details of the simulation study to examine 1) parameter recovery of the proposed latent space diffusion item response theory (LSDIRT) model (*Studies 1 and 2*) and 2) accuracy of the model selection property based on the slab-and-spike prior imposed on $\log(\lambda)$ (*Study 3*).

S2.1. Study 1: Parameter Recovery

To simulate data for the parameter recovery simulation, we used the following two sets of the main data-generating parameter values, with the number of persons $P = 200$ and the number of items $I = 15$.

Set 1

$$
\begin{aligned}
&[\theta_p, \log(\gamma_p)]^T \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \;\; p = 1, \ldots, P, \\
&\boldsymbol{\mu} = [0, 0]^T, \;\; \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_\theta^2 & \rho_{\theta\gamma}\sigma_\theta\sigma_\gamma \\ \rho_{\theta\gamma}\sigma_\theta\sigma_\gamma & \sigma_\gamma^2 \end{bmatrix}, \\
&\sigma_\theta = 1, \;\; \sigma_\gamma = 0.5, \;\; \rho_{\theta\gamma} = 0, \\
&\tau_p \sim TN(3, 1; 2, 4), \;\; p = 1, \ldots, P, \\
&\mathbf{a} = [0.3, 0.3, 0.3, 0.3, 0.3, 0.2, 0.2, 0.2, 0.2, 0.2, 0.1, 0.1, 0.1, 0.1, 0.1] \\
&\mathbf{b} = [0.0, -0.5, -1.0, -1.5, -2.0, 0.0, -0.5, -1.0, -1.5, -2.0, 0.0, -0.5, -1.0, -1.5, -2.0]
\end{aligned}
\tag{S1}
$$

Set 2

$$
\begin{aligned}
&[\theta_p, \log(\gamma_p)]^T \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \;\; p = 1, \ldots, P, \\
&\boldsymbol{\mu} = [0, 0]^T, \;\; \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_\theta^2 & \rho_{\theta\gamma}\sigma_\theta\sigma_\gamma \\ \rho_{\theta\gamma}\sigma_\theta\sigma_\gamma & \sigma_\gamma^2 \end{bmatrix}, \\
&\sigma_\theta = 1, \;\; \sigma_\gamma = 0.25, \;\; \rho_{\theta\gamma} = 0, \\
&\tau_p \sim TN(5, 2; 1, \infty), \;\; p = 1, \ldots, P, \\
&\mathbf{a} = [0.025, 0.025, 0.025, 0.025, 0.025, 0.050, 0.050, 0.050, 0.050, 0.050, \\
&\quad 0.075, 0.075, 0.075, 0.075, 0.075] \\
&\mathbf{b} = [0.0, -0.5, -1.0, -1.5, -2.0, 0.0, -0.5, -1.0, -1.5, -2.0, 0.0, -0.5, -1.0, -1.5, -2.0]
\end{aligned}
\tag{S2}
$$

The difference between the two sets is in generated response times (RTs). The first set has smaller nondecision times and larger item-wise time pressure parameters, producing relatively short RTs (mean: 7.787, median = 5.081, and SD: 11.601 in seconds). This would correspond to simple cognitive tasks and personality/item measurement based on words or short sentences. In contrast, the second set produces longer RTs (mean: 29.406, median = 15.432, and SD: 56.524 in seconds), which would correspond to relatively complex cognitive and psychometric tests. Also, the same values were repeatedly used for item parameters a_i and b_i to examine recovery when they have different combinations of values (e.g., $a_i = 0.3$ and $b_i = 0.0$ vs $a_i = 0.3$ and $b_i = -2.0$).

For both sets, parameters for a latent space were generated as follows:

$$
\begin{aligned}
&\lambda = 1 \\
&\boldsymbol{\xi}_p \sim MVN_k(\mathbf{0}, \boldsymbol{I}_k), \;\; p = 1, \cdots, P, \\
&\boldsymbol{\zeta}_i \sim MVN_k(\mathbf{0}, \boldsymbol{I}_k), \;\; i = 1, \cdots, I,
\end{aligned}
\tag{S3}
$$

For each set, we simulated 25 datasets. Then, we fitted the LSDIRT model to each of the simulated datasets with the HMC method and prior specifications described in Section 2.3 in the main manuscript. We ran three Bayesian chains for 2,500 iterations and discarded the first 1,000 posterior samples for burn-in. Convergence was assessed with $\hat{R}$ and we did not find any issue (Section S3). For the latent positions, we applied the Procrustes matching to their samples with the data-generating positions as a reference set, in order to see their recovery.

Result

To evaluate parameter recovery, we computed the Maximum A Posterior (MAP) estimates of the parameters with the posterior samples and compare them with the true data-generating parameter values. The recovery result is presented in Table 1 in terms of Mean Squared Error (MSE), Bias, Standard Error (SE) of the MAP estimates, Bayesian Standard Errors (BSE; posterior standard deviation) averaged over 25 repetitions, and the Pearson correlation coefficients averaged over 25 repetitions (with its standard error in the following parentheses). For person (item) parameters, we first computed MSE, Bias, SE, and BSE by person (by item), and then take the average over persons (over items). Also, Figure S1 and Figure S2 show the recovery results of the person and item parameters as scatter plot, when the data-generating parameter set was Set 1 and Set 2, respectively. In each panel, the MAP estimates of the parameters are plotted on the x-axis against the true parameter values on the y-axis. The gray squares represent the MAP estimates and their true parameter values for all 25 repetitions while the colored dots represent their averages across repetitions (i.e., mean estimates). In addition, the Pearson correlation (r) between the estimates and the corresponding true parameter values, averaged across repetition, is shown at the top-left side of each panel, with its SD across repetitions in the following parentheses (the same values as in Table 1).

In general, the result shows that the LSDIRT model can recover its parameters under the current condition. As the figures show, the MAP estimates are consistent with the true parameter values without large bias. The Pearson correlations are higher than 0.9 for most of the parameters with small SD. The statistics in Table 1 also did not imply any problem in parameter recovery. Note that, τ_p in Set 2 has relatively large values of the statistics just because τ_p has the same scale as RTs (seconds), which were much longer in this condition, and the statistics are scale-dependent.

	Set 1					Set 2				
	MSE	Bias	SE	BSE	Cor	MSE	Bias	SE	BSE	Cor
θ_p	0.139	0.184	0.225	0.336	0.931 (0.007)	0.066	0.127	0.138	0.215	0.971 (0.003)
$\log(\gamma_p)$	0.033	0.061	0.164	0.174	0.937 (0.008)	0.017	0.058	0.107	0.127	0.862 (0.015)
τ_p	0.100	0.115	0.226	0.245	0.893 (0.021)	0.870	0.349	0.729	0.770	0.904 (0.014)
$\log(a_i)$	0.003	0.031	0.040	0.052	0.997 (0.001)	0.001	0.019	0.031	0.037	0.999 (0.001)
b_i	0.021	0.076	0.113	0.167	0.985 (0.008)	0.008	0.053	0.062	0.103	0.996 (0.003)
ξ_{p1}	0.200	0.200	0.307	0.413	0.890 (0.014)	0.058	0.098	0.163	0.223	0.970 (0.005)
ξ_{p2}	0.120	0.126	0.278	0.362	0.929 (0.011)	0.034	0.064	0.147	0.201	0.981 (0.002)
ζ_{i1}	0.029	0.105	0.318	0.170	0.993 (0.003)	0.008	0.048	0.158	0.107	0.998 (0.001)
ζ_{i2}	0.021	0.052	0.326	0.178	0.994 (0.003)	0.005	0.030	0.168	0.109	0.999 (0.001)
λ	0.004	0.047	0.037	0.053		0.002	0.045	0.020	0.044	
σ_θ	0.002	0.023	0.040	0.062		0.002	0.036	0.026	0.055	
σ_γ	0.001	0.020	0.014	0.029		0.000	0.007	0.011	0.017	
$\rho_{\theta\gamma}$	0.003	0.045	0.035	0.083		0.004	0.035	0.052	0.090	

Table 1: **Parameter Recovery**. Data were simulated with $P = 200$ respondents and $I = 15$ items. MSE: Mean Squared Error, SE: Standard Error of Point Estimates (Maximum A Posteriori), BSE: Bayesian Standard Error (posterior standard deviations averaged over repetitions). Cor: The Pearson correlation across persons or items, with its SD across 25 repetitions. For person (item) parameters, MSE, Bias, SE, and BSE were computed by person (by item), and then averaged over persons (over items).

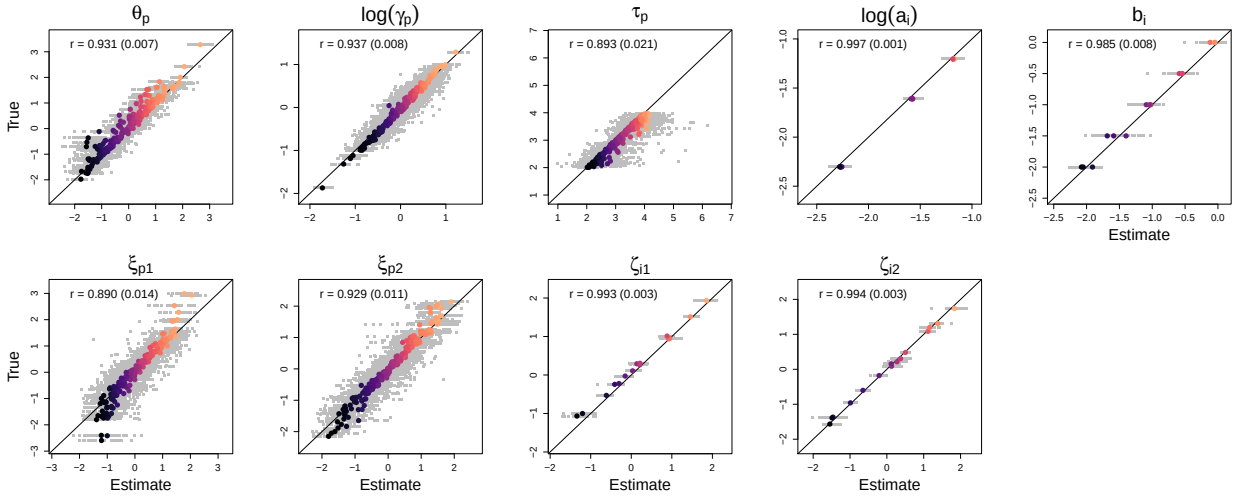

Figure S1: **Parameter Recovery: When Set 1 was used to simulate data**. In each panel, the Maximum A Posteriori (MAP) estimates of the parameters are plotted on the x-axis against the true parameter values on the y-axis. The gray squares represent the MAP estimates and their true parameter values for all 25 repetitions while the colored dots represent their averages across repetitions (i.e., mean estimates). In addition, the Pearson correlation (r) between the estimates and the corresponding true parameter values, averaged across repetition, is shown at the top-left side of each panel, with its SD across repetitions in the following parentheses.

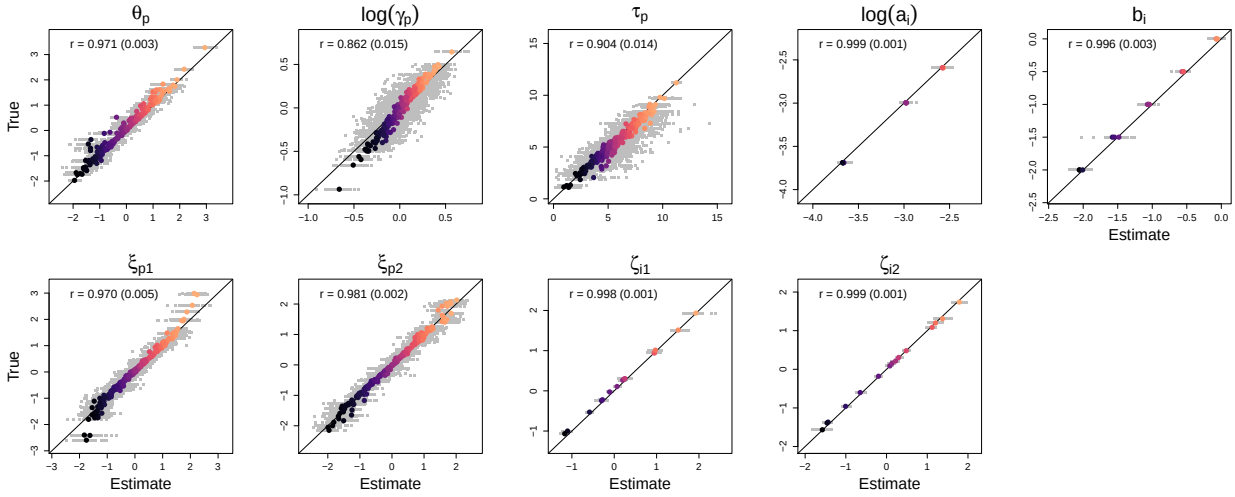

Figure S2: **Parameter Recovery: When Set 2 was used to simulate data**. In each panel, the Maximum A Posteriori (MAP) estimates of the parameters are plotted on the x-axis against the true parameter values on the y-axis. The gray squares represent the MAP estimates and their true parameter values for all 25 repetitions while the colored dots represent their averages across repetitions (i.e., mean estimates). In addition, the Pearson correlation R'(r) between the estimates and the corresponding true parameter values, averaged across repetition, is shown at the top-left side of each panel, with its SD across repetitions in the following parentheses.

S2.2. Study 2: Additional Recovery Study

Upon a request during the review process, we extended our parameter recovery study with two variations. In the first variation, the same simulation as done in *Study 1* was repeated but with $P = 500$ instead of $P = 200$. The previous choice was meant to be our suggestion for the minimum requirement to achieve precise item parameter estimation, which also corresponds to the choice made in previous articles based on the DIRT model and its extensions. Here, we aim to study how the accuracy of estimation changes with a larger number of respondents. The second variation was about our choice of the prior distributions in *Section 2.3.* in which item parameters were imposed some weakly-informative priors ($N(0, 5^2)$. The $Half\text{-}Cauchy(2.5)$ prior for the variance terms in the hierarchical person distribution can also be considered weakly-informative. Thus, it would be informative to investigate the effect of other choices such as highly diffuse (i.e., non-informative) priors. For this, we conducted another simulation study with $N(0, 100^2)$ as a prior for the item parameters $\log(a_i)$ and b_i and $Half\text{-}Cauchy(100)$ for σ^2_θ and σ^2_γ. Note that the other prior choices we made are non-informative.

The two variations described above were conducted with the two-parameter sets (Sets S1 and S2). Table 2 shows the recovery result of the simulation with the first variation (with $P = 500$). The estimation of the item parameters was improved as expected. But as more outlier respondents were likely to be sampled, the estimation of the person parameters was slightly worsened. However, the overall quality of parameter recovery was similar to the previous result.

Table 3 shows the recovery result of the simulation with the second variation in which the proposed model was fit to the same data as used in *Study 1* but with non-informative priors. The result shows that the recovery result was almost not affected by the change in the prior distributions. All statistics were consistent with those in Table 1, with only a little difference.

	Set 1					Set 2				
	MSE	Bias	SE	BSE	Cor	MSE	Bias	SE	BSE	Cor
θ_p	0.170	0.214	0.251	0.347	0.921 (0.006)	0.078	0.145	0.153	0.214	0.968 (0.003)
γ_p	0.034	0.067	0.161	0.174	0.933 (0.005)	0.018	0.062	0.108	0.132	0.846 (0.012)
τ_p	0.089	0.098	0.204	0.222	0.888 (0.019)	1.051	0.356	0.757	0.819	0.883 (0.014)
$\log(a_i)$	0.002	0.041	0.024	0.032	0.999 (0.001)	0.001	0.018	0.019	0.027	0.999 (0.000)
b_i	0.011	0.045	0.084	0.120	0.993 (0.003)	0.012	0.081	0.049	0.116	0.996 (0.005)
ξ_{p1}	0.240	0.205	0.346	0.446	0.865 (0.012)	0.078	0.098	0.197	0.245	0.959 (0.011)
ξ_{p2}	0.147	0.147	0.301	0.376	0.916 (0.011)	0.039	0.065	0.156	0.200	0.978 (0.002)
ζ_{i1}	0.016	0.061	0.374	0.117	0.993 (0.005)	0.006	0.054	0.186	0.104	0.999 (0.001)
ζ_{i2}	0.019	0.083	0.327	0.132	0.997 (0.002)	0.015	0.079	0.199	0.109	0.999 (0.001)
λ	0.008	0.083	0.027	0.036		0.004	0.056	0.020	0.031	
σ_θ	0.001	0.009	0.023	0.041		0.000	0.002	0.007	0.012	
σ_γ	0.000	0.002	0.009	0.018		0.000	0.014	0.013	0.036	
$\rho_{\theta\gamma}$	0.001	0.007	0.027	0.053		0.005	0.054	0.045	0.058	

Table 2: **Parameter Recovery**. Data were simulated with $P = 500$ respondents and $I = 15$ items. MSE: Mean Squared Error, SE: Standard Error of Point Estimates (Maximum A Posteriori), BSE: Bayesian Standard Error (posterior standard deviations averaged over repetitions). Cor: The Pearson correlation across persons or items, with its SD across 25 repetitions. For person (item) parameters, MSE, Bias, SE, and BSE were computed by person (by item), and then averaged over persons (over items).

	Set 1					Set 2				
	MSE	Bias	SE	BSE	Cor	MSE	Bias	SE	BSE	Cor
θ_p	0.140	0.185	0.225	0.336	0.931 (0.006)	0.066	0.129	0.138	0.215	0.971 (0.003)
γ_p	0.033	0.060	0.165	0.174	0.937 (0.008)	0.017	0.057	0.108	0.127	0.862 (0.014)
τ_p	0.099	0.115	0.226	0.245	0.893 (0.021)	0.875	0.350	0.732	0.770	0.903 (0.015)
$\log(a_i)$	0.002	0.029	0.040	0.051	0.997 (0.001)	0.001	0.017	0.031	0.037	0.999 (0.001)
b_i	0.022	0.081	0.113	0.168	0.985 (0.007)	0.008	0.056	0.060	0.103	0.996 (0.002)
ξ_{p1}	0.199	0.198	0.308	0.413	0.891 (0.014)	0.058	0.098	0.162	0.222	0.970 (0.005)
ξ_{p2}	0.119	0.126	0.278	0.362	0.929 (0.010)	0.033	0.064	0.146	0.201	0.981 (0.002)
ζ_{i1}	0.029	0.104	0.321	0.171	0.993 (0.003)	0.008	0.046	0.157	0.107	0.998 (0.001)
ζ_{i2}	0.020	0.053	0.330	0.178	0.994 (0.003)	0.006	0.030	0.165	0.109	0.999 (0.001)
λ	0.003	0.046	0.038	0.053		0.002	0.044	0.021	0.044	
σ_θ	0.001	0.019	0.013	0.030		0.000	0.007	0.010	0.017	
σ_γ	0.002	0.020	0.039	0.062		0.002	0.036	0.026	0.056	
$\rho_{\theta\gamma}$	0.003	0.045	0.034	0.083		0.004	0.035	0.053	0.091	

Table 3: **Parameter Recovery: Diffuse Priors**. Data were simulated with $P = 200$ respondents and $I = 15$ items. MSE: Mean Squared Error, SE: Standard Error of Point Estimates (Maximum A Posteriori), BSE: Bayesian Standard Error (posterior standard deviations averaged over repetitions). Cor: The Pearson correlation across persons or items, with its SD across 25 repetitions. For person (item) parameters, MSE, Bias, SE, and BSE were computed by person (by item), and then averaged over persons (over items).

S2.3. Study 3: Model Selection Property

This section provides details of the data-generating procedure for Figure 1 and the simulation study in Section 4 in the main manuscript. The purpose of *Study 3* is to see if the model selection property of the proposed model, based on the slab-and-spike prior imposed on λ, can 1) detect conditional dependence if data imply substantial residual dependence between responses and RTs due to interactions between persons and items that cannot be explained by the main model parameters and 2) reject the distance effect on the latent space when data imply no conditional dependence. For this study, we used Set 1 in Study 1 as data-generating values for the main model parameters. For a latent space, we assumed (as described in Section 2.1 in the main manuscript, regarding Figure 1) that the first 100 respondents have strong residual dependence with the first 8 items while the other 100 respondents have strong residual dependence with the last 7 items. This was done by sampling ϵ_{pi} from $N(2, 0.25^2)$ for persons and items with dependence (i.e., first for

$p = 1, \ldots, 100$ and $i = 1, \ldots, 8$, and second for $p = 101, \ldots, 200$ and $i = 9, \ldots, 15$). For the other person-item pairs with conditional independence, ϵ_{pi} was sampled from $N(0, 0.25^2)$. Note that we sampled ϵ_{pi} instead of latent positions $\boldsymbol{\xi}_p$ and $\boldsymbol{\zeta}_i$ to manipulate the latent space as shown in Figure 1.

As described in Section 4 of the main manuscript, We simulated 100 datasets with $\lambda = 0$ (i.e., the diffusion item response theory model with the conditional independence assumption) and the other 100 datasets with $\lambda = 1$. Then, we fitted the LSDIRT model to each of the simulated datasets with the HMC method and prior specifications used in Section 2.3. The model selection result is described in Section 4. Basically the result shows that the model can estimate a significant latent space when there is substantial residual dependence underlying the data and choose $\lambda = 0$ when there is no conditional dependence, reducing the model to the vanilla diffusion item response theory (DIRT) model.

Study 3 was conducted with assumed clusters of persons and items in the latent space (see Figure 1B). One may be curious if the model can enjoy the same quality of the model selection property even if there is no clear cluster in a latent space. To address this issue, we computed the posterior inclusion probabilities (PIPs) in *Study 1* where latent positions were just randomly sampled from the standard normal distribution (Equation S3). Mean, standard deviation, and range of PIPs were 0.911, 0.010, and $[0.889, 0.928]$, respectively for Set 1 and 0.904, 0.028, and $[0.822, 0.929]$, respectively for Set 2. Therefore, it is safe to conclude that the model can detect conditional dependence regardless of underlying structures in a latent space.

S3. Convergence of Bayesian Chains

To assure convergence of Bayesian chains in the application examples (Section 3 of the main manuscript), we computed the Potential Scale Reduction Statistics, $\hat{R}$, at every 200 iterations. In Figure S3, these values are plotted on the y-axis against the number of iterations on the x-axis so that changes in $\hat{R}$ over iterations can be investigated. Each (colored) dashed line represents $\hat{R}$ of each model parameter. The black dotted line shows the average $\hat{R}$ over all parameters. The black dashed horizontal line shows the cutoff for $\hat{R}$ (< 1.1). In general, the figure shows that Bayesian chains quickly achieve convergence. $\hat{R}$ for all parameters gets smaller than the cutoff in

$1,000$ iterations. Convergence was relatively slower in the extraversion data because of the smaller sample size ($P = 143$ persons and $I = 10$ items), but there was no issue in achieving convergence.

To further confirm convergence, we generated trace plots of some selected parameters. We chose the extraversion data for this investigation because convergence was the slowest in this example. In Figure S4, black, red, and green traces show different Bayesian chains. The figure shows that all the chains for the model parameters examined are well mixed without any convergence issue.

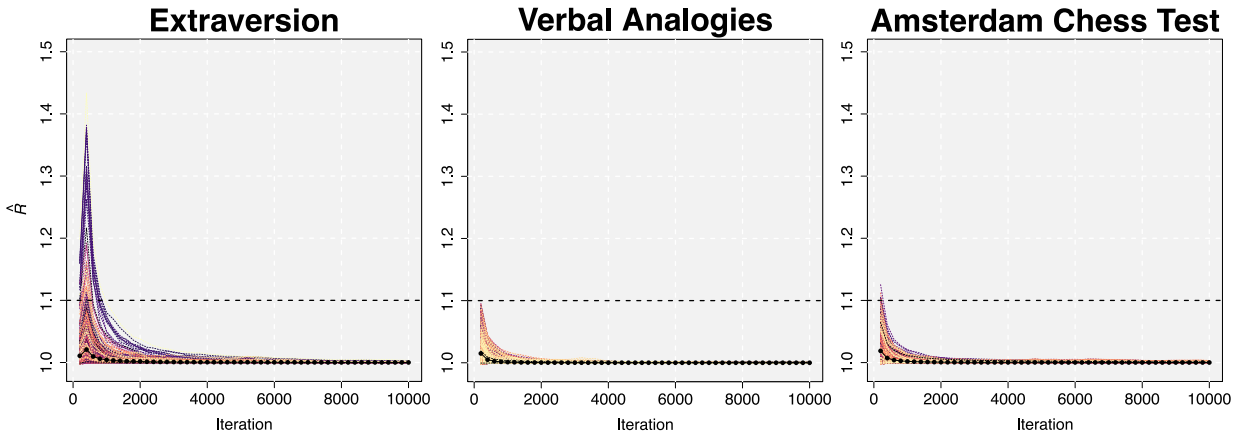

Figure S3: **Potential Scale Reduction Statistics ($\hat{R}$)**. $\hat{R}$ is computed at every 200 iteration and plotted on the y-axis against the number of iterations on the x-axis. Each dashed line shows the changes in $\hat{R}$ of each model parameter. The black dotted line shows the average $\hat{R}$ over all parameters. The black dashed horizontal line shows the cutoff for $\hat{R}$ (< 1.1).

With these results, we concluded that the LSDIRT model, estimated with the Hamiltonian Monte Carlo method implemented in **stan** (Stan Development Team, 2021), yielded no convergence issue in our application examples.

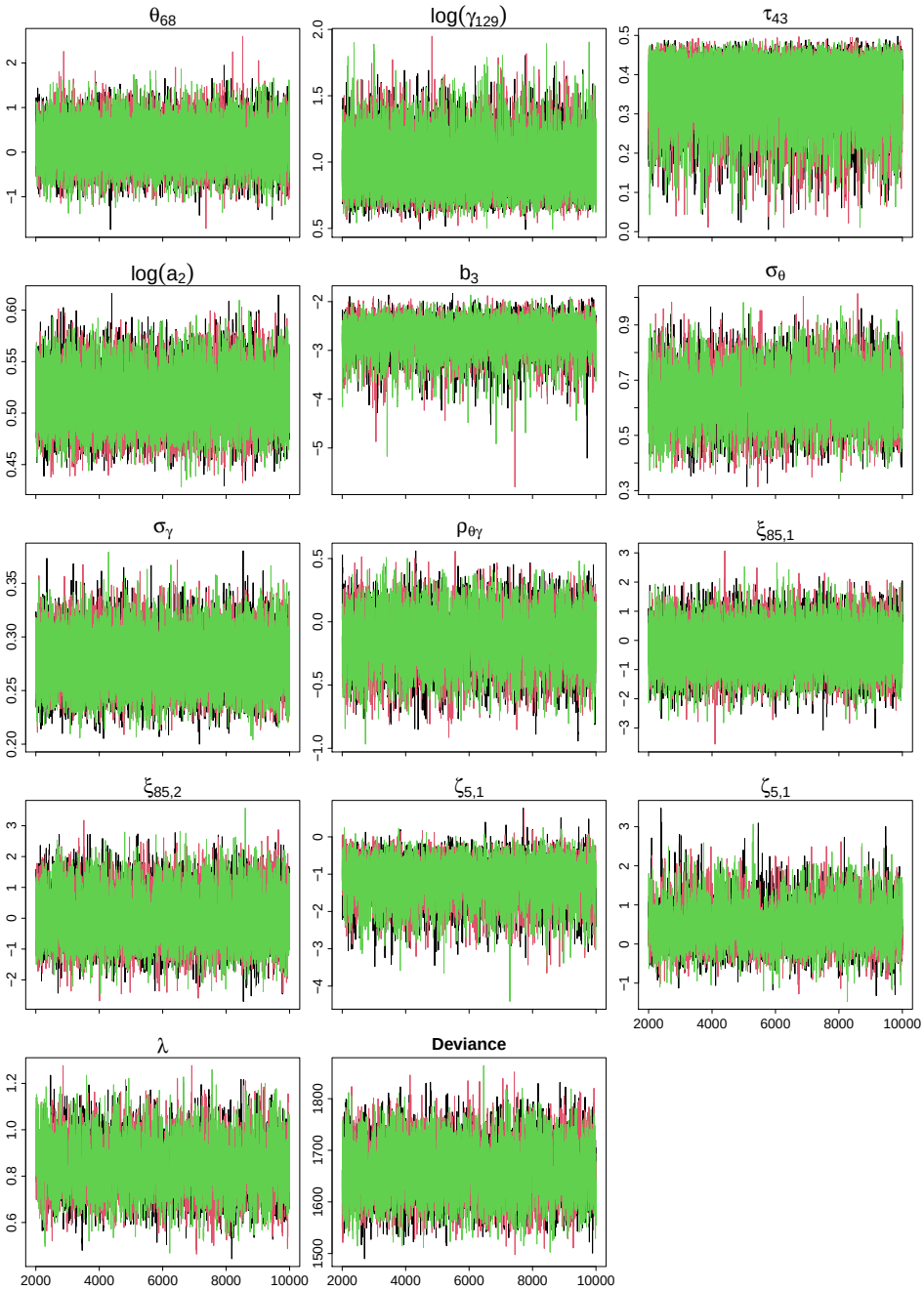

Figure S4: **Trace Plots of Parameters for Randomly Selected Persons and Items in the Extraversion Data**.

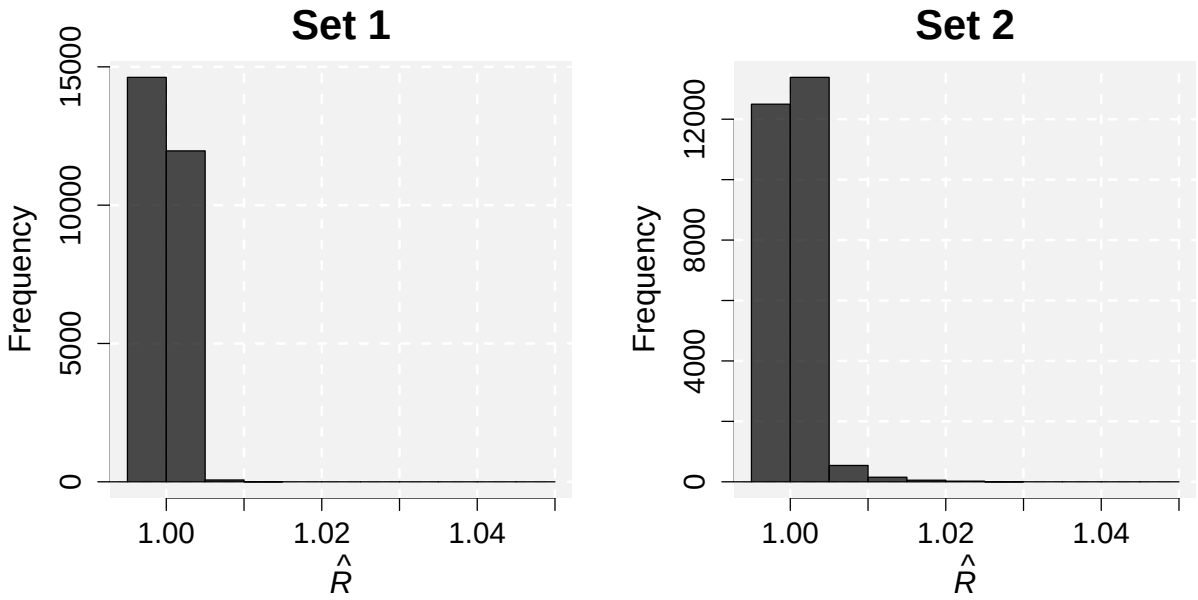

Figure S5: **Simulation Result: Histogram of Potential Scale Reduction Statistics ($\hat{R}$), from *Study 1*.**

Lastly, Figure S5 and Figure S6 present histograms of the computed $\hat{R}$ values in Simulation Studies 1 and 3, respectively. In Figure S5, the left and right panels show the histograms when Set 1 (Equation S2) and Set 2 (Equation S2) were used to simulate data, respectively. Also, Figure S6 shows the histograms when the datasets were generated with the conditional independence assumption (left) and when the data-generating model assumed conditional dependence across persons and items (right). The histograms include the $\hat{R}$ values for all the model parameters and those across repetitions (25 in *Study 1* and 100 in *Study 3*). The result does not show any convergence issue as all the $\hat{R}$ values are less than its cutoff of 1.1; In fact, more than 99.9% of the $\hat{R}$ values are less than 1.01 for the recovery simulations with Sets 1 and 2, respectively, and in *Study 3*, 99.8% and 98.3% are less than 1.01 for the conditional independence case and for the conditional dependence case, respectively.

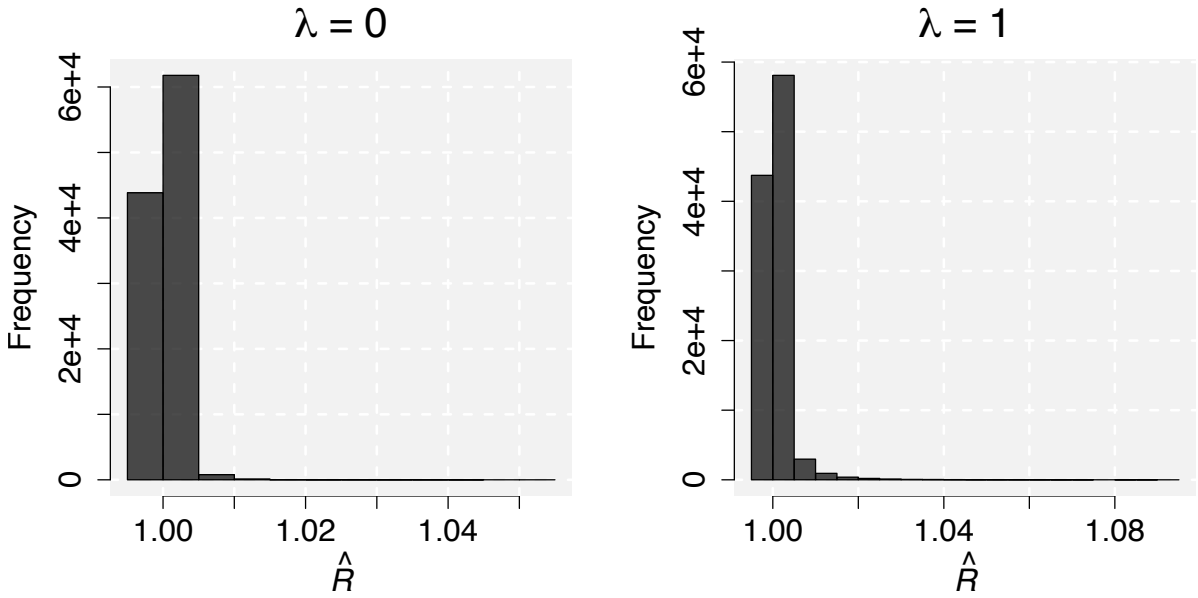

Figure S6: **Simulation Result: Histogram of Potential Scale Reduction Statistics ($\hat{R}$) from *Study 3*.**

S4. Posterior Predictive Checking for Absolute Model Fits

To assess the absolute fit of the LSDIRT model to the three empirical data used in Section 3 of the main manuscript, we conducted posterior predictive checking (Gelman, Carlin, Stern, Dunson, & A. Vehtari, 2013). We randomly selected $2,000$ posterior samples of the model parameters and then generated a random sample of response and RT from each of these posterior samples. This produced $2,000$ posterior predictive samples of responses and RTs, which we used to compute model predictions of response proportions and RT distributions (quantiles). In Figure S7, we contrasted these model predictions against the data to check if the LSDIRT model provided adequate accounts for the behavioral pattern observed from the data.

The top, middle, and bottom rows of Figure S7 present the posterior predictive checking results for the extraversion, verbal analogies, and Amsterdam chess test (ACT) data, respectively. In each row, the model absolute fit was assessed in three different ways: 1) response proportions, 2) overall RT distribution over persons and items, and 3) item-wise RT distributions (across persons, but separately by item). In the left column, the data response proportions (proportions of positive responses in the extraversion data and response accuracy in the other two data) were obtained

by item and plotted on the x-axis against the corresponding model prediction on the y-axis. The vertical interval for each dot indicates 95% credible intervals of the predicted item-wise accuracy. Also, at the top-left side of the scatter plot, the Pearson correlation (r) between the data-based item-wise accuracy and the corresponding predictions is shown. Below the correlation estimate, the overall response proportion (over all persons and items) is shown with the predicted proportion in parentheses.

The histograms in the middle column show the overall RT distributions obtained from data while the densities show the corresponding model predictions, but separately for positive/correct (black) and negative/incorrect (red) responses. For visual clarity, the negative/incorrect RTs were negatively coded (i.e., multiplied by -1) and plotted.

The RT distributions can be further investigated by the scatter plots in the right column in which data-based item-wise RT quantiles were obtained (across persons, but separately by item) and plotted on the x-axis against the corresponding model predictions on the y-axis. The numbers, 1, 3, 5, 7, and 9, indicate 10%, 30%, 50%, 70%, and 90% RT quantiles, and positive/correct RTs are colored black while negative/error RTs are colored red as in the histograms in the middle column. The Pearson correlations between data and prediction were also shown at the top side of the scatter plots, separately for positive/correct (r_{Yes} and r_C) and negative/error responses (r_{No} and r_E).

In general, the posterior predictive checking results show that the LSDIRT model provides adequate accounts for the data pattern with good absolute fits. Thus, we focus on delineating some minor misfits here. In the extraversion data, a few items show misfits at the tail of the negative RT distributions. Most RT distributions are right-skewed with large variations at the tail, so this misfit is quite common. This is also related to the high positive response proportions of these items, which produce a too small number of observations to precisely calculate data-based negative RT quantiles. For example, the 90% quantile at the far-right side of the scatter plots of item-wise quantiles is for item 4. The positive response proportion of this item is 0.916 and there were only 12 negative responses, which are not sufficient to obtain data RT quantiles precisely. The item-wise error RT distributions for some items in the verbal analogies data showed similar misfits. The overall response accuracy of the data was 0.731, which was not too high, but there

were several items with high accuracy. They formed a cluster on the top-right side of the left panel. Among 24 items, 12 items had accuracy higher than 0.8, 5 of which had accuracy higher than 0.9. For these items, there were relatively few error observations and thus misfit was larger for the error RT distributions, particularly at the tails. See Appendix B in Kang, De Boeck, and Partchev (2022) for a related analysis, which formally showed that the misfit in the error RT tails is indeed due to too few error observations.

For the ACT data, overall accuracy was 0.521 and so there were sufficient observations for both correct and error responses. However, some items had accuracy close to 1 (or 0) and thus their error (correct) RT distributions may have similar minor misfits at the tails, as in the extraversion and verbal analogies data. Another notable finding for the ACT data was that the overall error RT distribution (shown as a histogram in the middle column) had a bump at the tail, making the shape of the distribution abnormal. The bimodality could be due to the strong effect of the item-wise time limit (30 seconds) on the response behavior of chess players. While a respondent was solving each chess item, a small clock on the screen displayed the remaining time (van der Maas & Wagenmakers, 2005). If the time elapsed, the response was recorded as an error and no point was given. Presumably, this could introduce late guessing responses at 25-30 seconds when a respondent failed to solve the presented item in time. This corresponds to the finding that only the error RT distribution had the bump at the late RT period while the correct RT distribution did not. However, the LSDIRT model did not implement a component to account for this time limit effect and potential guessing behavior (or non-reached items) resulting from it, and thus, the model prediction for the error RT distribution was a common right-skewed distribution. Still, the predicted item-wise RT quantiles were generally consistent with the data, as shown by the scatter plot on the right column. Thus, it could be concluded that the model was able to provide a reasonably good fit to the data despite the fit in the error RT tails.

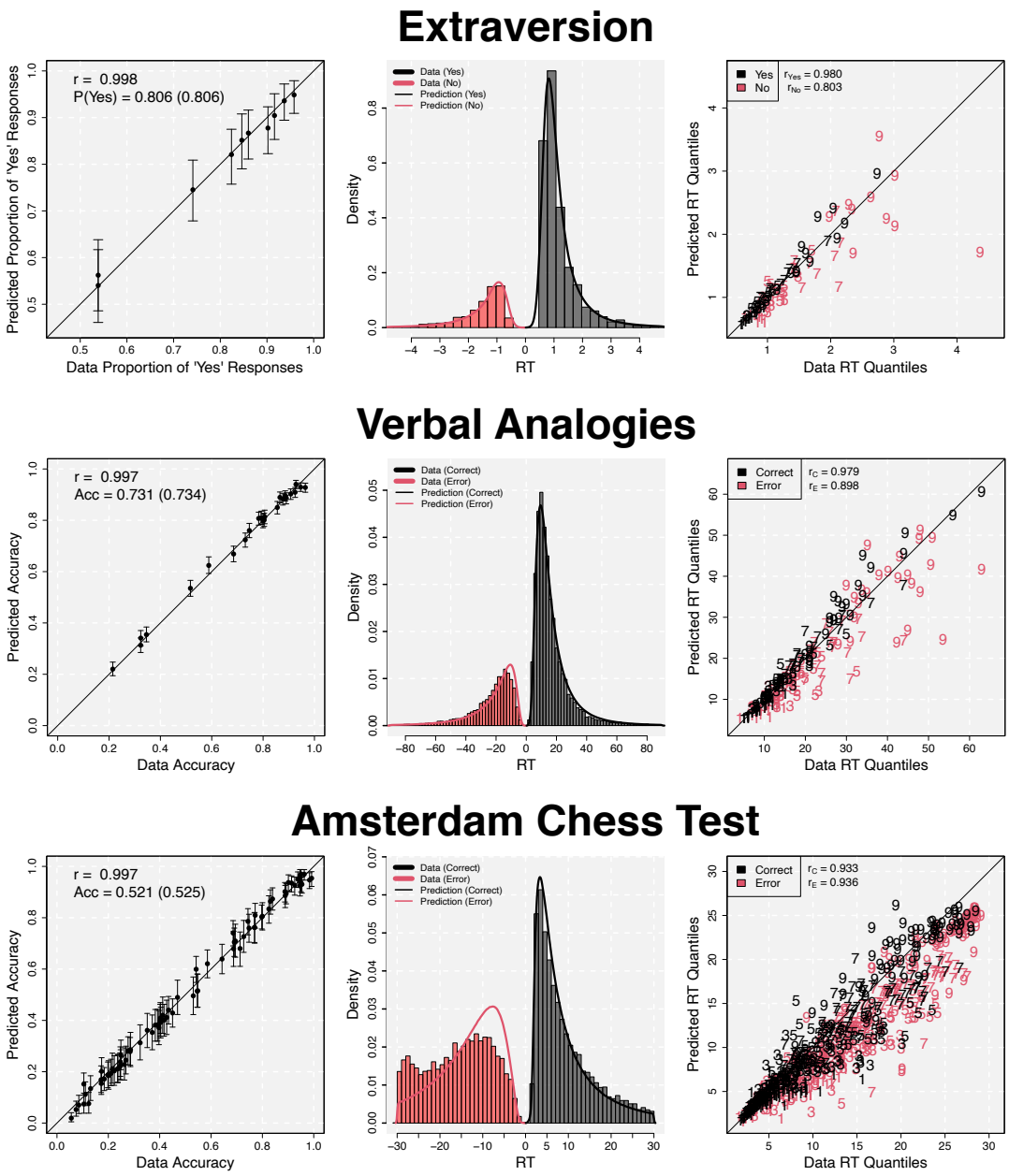

Figure S7: **Posterior Predictive Checking**. Left: Scatter plot of data and predicted response proportions. Their Pearson correlation (r), and overall positive/correct response proportion across persons and items (along with the corresponding model prediction in the following parentheses) are shown on the top-left side. The intervals of the dots show the 95% credible intervals. Middle: Overall RT distribution across persons and items. The histograms show the data distributions while the densities show the corresponding model predictions. Error RTs are coded negative and colored red for visual clarity. Right: Scatter plot of item-wise RT quantiles obtained across persons, but separately by item. The numbers, 1, 3, 5, 7, and 9, represent 10%, 30%, 50%, 70%, and 90% quantiles. Error RT quantiles are colored red. The Pearson correlations between data and predicted quantiles are shown on the top-left side.

References

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

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

Stan Development Team. (2021). *Stan Modeling Language User's Guide and Reference Manual*. Retrieved from `http://mc-stan.org/`

van der Maas, H. L. J., & Wagenmakers, E.-J. (2005). A psychometric analysis of chess expertise. *The American journal of psychology*, *118*, 29-60.