# Modelling the covariance structure of complex data sets using cognitive models: An application to individual differences and the heritability of cognitive ability

Nathan J. Evans[a], Mark Steyvers[b] and Scott D. Brown[c]
[a] Department of Psychology, University of Amsterdam, The Netherlands
[b] Department of Cognitive Sciences, University of California, Irvine, USA
[c] School of Psychology, University of Newcastle, Australia

*Keywords: Covariance Structure — Complex Data — Cognitive Models — Individual Differences — Heritability*

Abstract

Understanding individual differences in cognitive performance is an important part of understanding how variations in underlying cognitive processes can result in variations in task performance. However, the exploration of individual differences in the components of the decision process – such as cognitive processing speed, response caution, and motor execution speed – in previous research has been limited. Here, we assess the heritability of the components of the decision process, with heritability having been a common aspect of individual differences research within other areas of cognition. Importantly, a limitation of previous work on cognitive heritability is the underlying assumption that variability in response times solely reflects variability in the speed of cognitive processing. This assumption has been problematic in other domains, due to the confounding effects of caution and motor execution speed on observed response times. We extend a cognitive model of decision-making to account for relatedness structure in a twin study paradigm. This approach can separately quantify different contributions to the heritability of response time. Using data from the Human Connectome Project, we find strong evidence for the heritability of response caution, and more ambiguous evidence for the heritability of cognitive processing speed and motor execution speed. Our study suggests that the assumption made in previous studies – that the heritability of cognitive ability is based on cognitive processing speed – may be incorrect. More generally, our methodology provides a useful

Correspondence concerning this article may be addressed to: Nathan Evans, Department of Psychology, University of Amsterdam, The Netherlands; Email: nathan.j.evans@uon.edu.au

avenue for future research in complex data that aims to analyze cognitive traits across different sources of related data, whether the relation is between people, tasks, experimental phases, or methods of measurement.

## Introduction

Understanding why individuals differ from one another has been a common question across many areas of psychology, including learning (Reber, Walkenfeld, & Hernstadt, 1991), age-related decline (Ratcliff, Thapar, & McKoon, 2010), personality (Humphreys & Revelle, 1984), and memory (Just & Carpenter, 1992). The investigation of the latent cognitive attributes that underlie task performance, and how individual differences in those attributes can lead to differences in observed performance, allows for a better understanding of what specific variations in the cognitive process can cause specific variations in the observed task performance.

For decision-making in particular, it is well known that multiple components of the decision-making process contribute to the response times measured in simple choice tasks, and that people differ substantially in their mental processing speed, response caution, and motor processing speed. The few studies that have focused on individual differences in these attributes have linked white matter tract strength (Forstmann et al., 2010) and personality traits (Evans, Rae, Bushmakin, Rubin, & Brown, 2017) with the level of response caution that people display, and aging with processing speed – independent of IQ (Ratcliff et al., 2010).

The study of genetic inheritance has expanded over recent decades to include investigations of the heritability of psychological traits. Some of this interdisciplinary work has focused particularly on the heritability of cognitive abilities, such as intelligence and memory (Bouchard, 2004; DeFries & Fulker, 1985; Bouchard, McGue, et al., 1981; Erlenmeyer-Kimling & Jarvik, 1963). General intellectual abilities appear to have strong heritability, with some evidence suggesting that more than 50% of the variability in intelligence is explained by genetic variations (Vernon, 1989). Twin studies are the most commonly-used paradigm for investigating cognitive heritability. These studies compare identical twins, who share identical genetic material at birth with non-identical twins, who have the same amount of genetic overlap as regular siblings. Heritability in some measure (such as IQ) is assessed by comparing the strength of association in that measure between identical (monozygotic, or MZ) and non-identical (dizygotic, or DZ) twins. Higher association for MZ than DZ twins indicates genetic heritability of the measure.

Analysis of data from twins has supported the notion that there are heritable components to both general cognitive processes (e.g. intelligence) and the cognitive processes that underpin them (e.g. processing speed: Luciano et al., 2001; Finkel & Pedersen, 2004; Vernon, 1989; Beaujean, 2005; Ogata, Kato, Honda, & Hayakawa, 2014; Kochunov et al., 2016; Posthuma, Mulder, Boomsma, & De Geus, 2002). However, a limitation of the work so far is an untested assumption that response time is a pure reflection of underlying cognitive processing speed. This assumption has proven problematic in other research domains, because the observed response times for cognitive tests can be heavily influenced by factors other than cognitive processing speed. For example, the observed response time with which people complete a standard memory task is undoubtedly influenced by how

efficiently they can retrieve memories, but is also influenced by how quickly they execute motor responses, and by how cautiously they respond. Previous cognitive heritability research has also suggested a strong genetic component of response caution, with Engelhardt et al. (2016) suggesting that the entire heritable component seen in general intelligence might be explained through the heritability in executive functions.

Our study investigates whether the heritability of response time in a simple cognitive test is really attributable to inherited cognitive processing speed, or to some other inherited factor, such as caution or motor speed. This necessitates addressing both response speed and accuracy simultaneously, to disentangle the sometimes-complex relationships between them. Studies of inheritance using the twins paradigm have most frequently focused on response time and response accuracy, separately. When they have considered both measures, the statistical approach has been to use an "off the shelf" model, such as structural equation modeling (SEM). Such approaches do not take into account the detailed knowledge about the micro-structure of the relationship between response speed and accuracy which has been gathered over the past 50 years (Luce, 1986). Instead, our approach is based on a cognitive model which allows the analysis to leverage that knowledge. We replace the statistical model of SEM with a well-validated and statistically tractable cognitive model. This approach has proven successful in a wide range of psychological paradigms, investigating the underlying cognitive effects of fatigue, alcohol consumption, depression, and many others (for a review, see Donkin & Brown, 2017).

The most commonly applied cognitive models for disentangling the effects of caution, processing speed, and motor speed are evidence accumulation models of simple decision-making (Ratcliff, Thapar, & McKoon, 2001, 2011; Evans, Rae, et al., 2017; Evans & Brown, 2017b; Evans, Hawkins, Boehm, Wagenmakers, & Brown, 2017). These models posit that decision-making is the result of evidence accumulating in favor of each response alternative until a threshold amount is reached for one of the alternatives, at which time a response is triggered. In these models, "evidence" is defined very generally, as whatever task-relevant information is used to discriminate the different choice options. In some specific fields, detailed models have been developed of how this evidence is produced, and what defines it (e.g.: in perceptual decision-making, by Lu & Dosher, 2008; in consumer choice, by Trueblood, Brown, & Heathcote, 2014 and Busemeyer & Townsend, 1992; in confidence judgments by Ratcliff & Starns, 2013 and Pleskac & Busemeyer, 2010; and in absolute identification by Brown, Marley, Donkin, & Heathcote, 2008).

We extend the evidence accumulation framework to take into account genetic relatedness, which allows direct assessment of the heritability of the different components of decision-making. We focus on three key components that contribute to overall performance speed: the rate of evidence accumulation, which is a measure of processing speed; the threshold amount of evidence required to trigger a decision, which is a measure of the balance between caution and urgency; and the amount of time taken by non-decision components of processing, especially motor execution time. We apply our methods to data from the Human Connectome Project (HCP: Van Essen et al., 2013). These data include

behavioral measures from a range of cognitive tasks, genetic information including twin status, and a variety of other measures not analyzed here, such as brain structure and medical history. From the HCP, we focus on data from the 2-back task, which measures working memory, which is a good measure of an important cognitive function (working memory).

## Method

### *Participants*

We analyzed data from the March 2017 release of the HCP (Van Essen et al., 2013). Of the 1,200 participants in that release, 1,092 had data for the working memory 2-back task. These consisted of 298 MZ twins (149 pairs), 168 DZ twins (84 pairs), and 434 non-twin siblings. In total, we analyzed 450 pairs of participants. Not every non-twin sibling pair was unique, as the siblings of some twins were also members of other pairs.

### *Task*

We analyzed decision-by-decision data from the 2-back task, which is a variant of a commonly-used and well-validated measure of working memory, known as the $n$-back task. During the 2-back task (see Figure 1), participants view a sequence of images, and decide – for each image – whether the current image matches that presented two images previously. A matching image is called a "target", and a non-matching image is called a "non-target". In addition, some non-target images match images which were recent, but not exactly two images previous (e.g. the last, or third-last image). These images are known as "lures", because they are non-targets which will nevertheless feel very familiar to the participant, encouraging an incorrect response. There were 80 trials in the task, split into four different types of stimuli (separated by blocks): faces, bodies, spaces, and tools. We do not investigate differences between the stimulus types. Each image was presented for two seconds in which the participants were allowed to respond, or else a non-response was recorded, followed by a 500ms inter-trial-interval. Full details of this task, and all tasks, can be found within the HCP documentation (link: `https://www.humanconnectome.org/study/hcp-young-adult/document/1200-subjects-data-release`).

### *Standard analysis*

We report two kinds of analysis based upon the data of the HCP. Firstly, we use a standard analysis method from the heritability literature, and calculate $h^2$ values using linear Pearson correlations on surface-level manifest variables (mean RT, decision accuracy, the proportion of non-responses, the variance in RT, and the minimum RT). The quantity $h^2$ is a standard measure of heritability based on the difference between correlations for MZ and DZ twins: $h^2 = (r_{MZ} - r_{DZ}) \times 2$. The logic behind this measures is based on the amount of genetic overlap between MZ and DZ twins (Bouchard, 2004; Vernon, 1989; Kochunov et al., 2016; Visscher, Hill, & Wray, 2008; DeFries & Fulker, 1985). We also use another

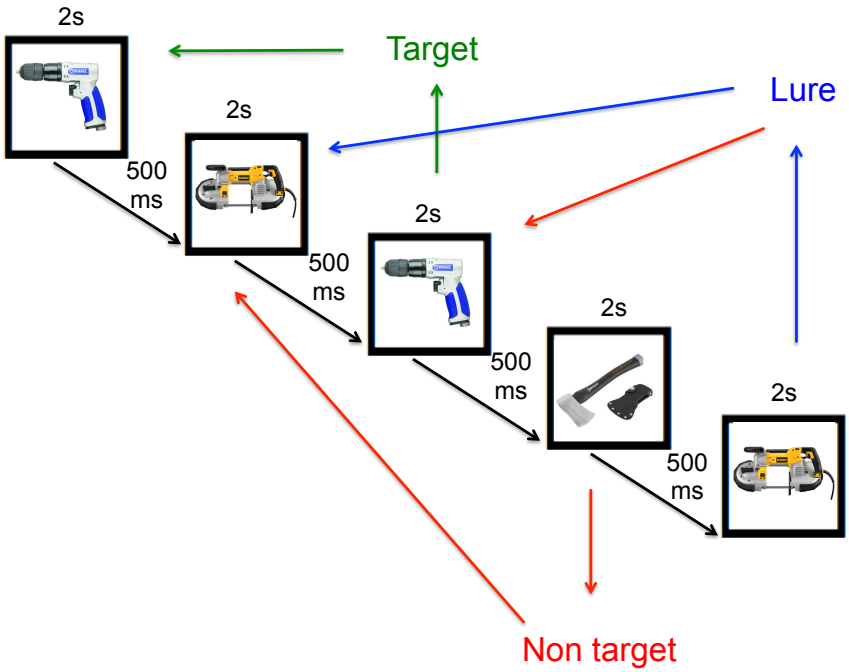


*Figure 1.* An example of the procedure for the 2-back task from the HCP. This figure shows five stimulus presentations, which results in three testing trials (because the first two trials have no stimulus to be matched to in a 2-back task). The third stimulus is a "target" because it matches the stimulus presented two images earlier (another blue power drill). The fourth stimulus does not match the image presented two images earlier, and so is a "non-target". The fifth image is also a "non-target" (as it does not match the image presented two images earlier), but is also a "lure", because it does match the image presented *three* images earlier.

standard analysis method from the heritability literature, ACE modelling (Zyphur, Zhang, Barsky, & Li, 2013), which extends upon $h^2$ by breaking the overall variability into three components: additive genetic variance ($a^2$), shared environmental variance between twins ($c^2$), and unshared environment variance ($e^2$). We performed ACE modelling using a least squares minimization routine between the expected covariances and the actual covariances, with a differential evolution optimizer that ran for 3,000 iterations with 50 particles. The estimated covariance were calculated as follows:

$$COV_{MZ} = a^2 + c^2$$
$$COV_{DZ} = (0.5 \times a^2) + c^2$$

*Cognitive model analysis*

The second analysis uses a cognitive model, the linear ballistic accumulator model (LBA: Brown & Heathcote, 2008), to make inferences about the heritability of the latent variables that underlie decision-making. The LBA is a well-validated and statistically tractable evidence accumulation model (for a review of its use, and the general topic, see Donkin & Brown, 2017), with Donkin, Brown, and Heathcote (2009) finding that it generally comes to the same conclusions as the diffusion model (Ratcliff, 1978). The LBA proposes that the decision process is made up of evidence being accumulated for each of the different decision alternatives over the course of the decision. For example, in the case of the 2-back task, these responses alternatives would be to say that the current stimulus matches the one presented 2 stimuli ago (i.e., responding "target") or to say that the current stimulus does not match the one presented 2 stimuli ago (i.e., responding "non-target"). The alternatives continue to accumulate evidence until the amount of evidence for one of these alternatives reaches some threshold, which triggers a decision for that alternative (see Figure 2 for an example).

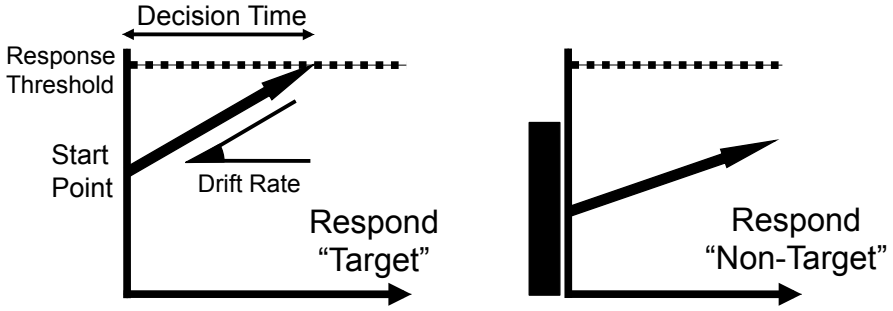


*Figure 2.* The linear ballistic accumulator model (LBA). Evidence accumulates for each alternative, in this case "target" vs. "non-target", until one reaches a threshold level of evidence, which triggers a decision. Rather than starting at zero, each alternative contains some starting amount of evidence, with this being randomly drawn for each accumulator and each trial from a uniform distribution. The rate of accumulation also differs from trial to trial, according a normal distribution (truncated to positive values).

The LBA makes quantitative predictions for response time and accuracy on the basis of five key parameters, which can be interpreted as latent cognitive variables: the drift rate ($v$), which determines how quickly evidence accumulates; the decision threshold ($b$), which determines how much evidence is needed to trigger a decision; the non-decision time ($t0$), which determines the amount of time dedicated to non-decision processes such

as perceptual and motor processes; and the start point, which determines the amount of evidence in favor of each response before the decision process is started. The model also allows for variability in drift rate between trials, governed by a normal distribution with standard deviation $s$, and in starting points between trials, with $A$ giving the height of the uniform starting point distribution.

The following equations provide a precise mathematical specification of the model that we used for these data, including of the prior distributions we assumed over all parameters. In these equations, subscript $i = [1, 2, ...]$ indexes the pairs of participants, subscript $j = [1, 2]$ indexes the individuals in each pair, subscript $k = [1, 2, 3]$ indexes the experiment condition, being target (1), lure (2), and non-target (3), and $l = [1, 2]$ indexes whether the accumulator was for the correct (1) or incorrect (2) response. When a subscript is not included for a parameter, it indicates that this parameter was constrained to take the same value across the values of this subscript. Beginning at the level of individual response times and choices:

$$(RT_{i,j}, response_{i,j}) \sim LBA(A_{i,j}, b_{i,j}, t0_{i,j}, v_{i,j,k,l}, s_{i,j,l})$$

The drift rate distributions for the accumulator corresponding to the correct response in each of the three conditions have their standard deviation (i.e., $s_{i,j,1}$) set to 1, to constrain a scaling property of the model (Donkin et al., 2009), whereas the standard deviations corresponding to the incorrect response (i.e., $s_{i,j,2}$) was estimated. The means for the drift rate distributions where allowed to vary over both experimental condition and correct/incorrect responses, meaning that each condition-response combination contained a different mean drift rate (6 in total). All other parameters were constrained to take on the same values across all experimental conditions, and correct/incorrect response accumulators. This model has a total of 10 free parameters per participant: 6 $v$ parameters, and a single $A$, $b$, $t0$, and $s$ parameter. To model missed responses in the HCP's 2-back task, we evaluated the survivor function at the cutoff point of two seconds, which is a standard approach for censored distributions (Ulrich & Miller, 1993).

We fit the LBA using a hierarchical Bayesian approach. This allows for individual differences in a constrained manner – each participant is allowed their own parameter estimates, but these estimates are constrained to follow group-level distributions. This approach allows an analysis of the entire group of participants, rather than just each participant individually, without suffering from well-known issue associated with averaging distributions over participants (Estes, 1956). The key in our hierarchical approach is modeling the covariance structure, which we implemented by constraining the logarithms of these individual-level parameters to follow bivariate normal distributions at the group level. The mean (denoted by $\mu$) and standard deviation (denoted by $\sigma$) of the group-level distributions were forced to be equal across the different relatedness groups, meaning that our model contained 10 $\mu$ parameters, and 10 $\sigma$ parameters (i.e., equal to the number of free parameters per participant). Note that the $\mu$ and $\sigma$ parameters, although important for

the hierarchical structure, were of little theoretical interest to us for assessing our questions regarding the heritability of the components of the decision-making process. Instead, we were interested in the parameters that formed the key extension of our approach: the correlation parameters, which were allowed to be different for MZ twins, DZ twins, and non-twin siblings to assess the heritability. To simplify the model, we constrained the correlation between pairs to be the same value for all drift rate parameters (i.e., the 6 mean drift rates across all experimental conditions and correct/incorrect responses, and the single standard deviation for the incorrect responses). We also constrained the correlation between pairs to be the same value for the start-point and threshold parameters ($b$ and $A$), as these parameters are highly correlated. This means that we estimated 9 correlation parameters in total: the correlations for the 3 different parameter types (drift rate, threshold, and non-decision time) for each of the 3 different relatedness groups (MZ twins, DZ twins, and non-twin siblings). We denote these group-level correlation parameters by $\rho_{p,g}$, where the subscript $p = [1, 2, 3]$ indexes which individual-level parameters the correlation applies (1 for thresholds, 2 for drift rates, and 3 for non-decision time), and the subscript $g(i) = [1, 2, 3]$ indexes the genetic relatedness group (1 for MZ, 2 for DZ, and 3 for sibling pairs) for the $i^{th}$ pair of participants.

$$
\begin{aligned}
\log(A_{i,j}) &\sim \mathrm{N}\left(\begin{bmatrix} \mu_A \\ \mu_A \end{bmatrix}, \sigma_A^2 \begin{bmatrix} 1 & \rho_{1,g(i)} \\ \rho_{1,g(i)} & 1 \end{bmatrix}\right) \\
\log(b_{i,j} - A_{i,j}) &\sim \mathrm{N}\left(\begin{bmatrix} \mu_b \\ \mu_b \end{bmatrix}, \sigma_b^2 \begin{bmatrix} 1 & \rho_{1,g(i)} \\ \rho_{1,g(i)} & 1 \end{bmatrix}\right) \\
\log(t0_{i,j}) &\sim \mathrm{N}\left(\begin{bmatrix} \mu_{t0} \\ \mu_{t0} \end{bmatrix}, \sigma_{t0}^2 \begin{bmatrix} 1 & \rho_{3,g(i)} \\ \rho_{3,g(i)} & 1 \end{bmatrix}\right) \\
\log(v_{i,j,k,l}) &\sim \mathrm{N}\left(\begin{bmatrix} \mu_{v\{k,l\}} \\ \mu_{v\{k,l\}} \end{bmatrix}, \sigma_{v\{k,l\}}^2 \begin{bmatrix} 1 & \rho_{2,g(i)} \\ \rho_{2,g(i)} & 1 \end{bmatrix}\right) \\
\log(s_{i,j,2}) &\sim \mathrm{N}\left(\begin{bmatrix} \mu_s \\ \mu_s \end{bmatrix}, \sigma_s^2 \begin{bmatrix} 1 & \rho_{2,g(i)} \\ \rho_{2,g(i)} & 1 \end{bmatrix}\right)
\end{aligned}
$$

Finally, the priors we specified were relatively uninformative. Most importantly, the prior distributions were identical for parameters pertaining to different experimental conditions (lure, target, non-target) and different relatedness groups (MZ, DZ, and non-twin siblings). This ensures that any observed differences in the posterior distributions were driven by the data:

$$\begin{aligned}
\log\left(\mu_A, \mu_b, \mu_{v\{k,l\}}, \mu_s\right) &\sim \mathrm{N}(0, 2) \\
\log\left(\mu_{t0}\right) &\sim \mathrm{N}(-2, 2) \\
\log\left(\sigma_A^2, \sigma_b^2, \sigma_{v\{k,l\}}^2, \sigma_s^2, \sigma_{t0}^2\right) &\sim \mathrm{N}(-3, 2) \\
\rho_{p,g} &\sim \mathrm{N}(0, 0.3)
\end{aligned}$$

We sampled from the posterior distribution over parameters using Markov Chain Monte Carlo with proposals generated by differential evolution method (DE-MCMC; Ter Braak, 2006; Turner, Sederberg, Brown, & Steyvers, 2013). Parameter updating was blocked by subject, meaning that the highest dimension of the updating process was 10 parameters. DE-MCMC has been found to be highly efficient at sampling from correlated dimensions in dimensionality of this size (Ter Braak, 2006; Turner et al., 2013). We ran 30 parallel chains for 4,000 iterations of burn-in, initiated with broad starting points, and then drew 2,000 samples from each chain for the posterior. The correlation posterior distributions were checked for convergence and mixing using the $\hat{R}$ statistics, with all $\hat{R}$ values being below 1.2 (the value recommended by Ter Braak, 2006; ter Braak & Vrugt, 2008), and all except one ($t0$ MZ) below 1.1.

In order to make inferences on whether the correlations for each parameter differed between the groups (e.g., whether the correlation parameter for threshold differed between MZ twins and DZ twins), we used the Savage-Dickey ratio (Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010; Wetzels, Grasman, & Wagenmakers, 2010). The Savage-Dickey ratio provides an estimate of the Bayes factor (see Evans & Brown, 2017a for a discussion of different ways to estimate the Bayes factor) between a null model (i.e., no difference between groups) and an alternative model (i.e., a difference between groups) by taking the ratio of the density of the prior distribution and the posterior distribution at the point of 0 (i.e., no difference). Specifically, if the posterior is more likely than the prior at 0, then this shows evidence for the null model, and if the prior is more likely than the posterior at 0, then this shows evidence for the alternate model. Bayes factors of around 1 indicate approximately equal evidence for each model, whereas Bayes factors of above 3 or below $\frac{1}{3}$ indicate moderate evidence for the alternative model or the null model, respectively. Bayes factors of above 10 or below $\frac{1}{10}$ indicate strong evidence, with the data being 10 times more likely under one model than the other. We obtained the "difference posterior" by taking the difference between the estimated correlation distributions of the different relatedness groups for every posterior sample taken, and obtained the "difference prior" by randomly generating 1,000,000 samples from the prior of each relatedness group's correlation distribution, and taking the difference between these samples. We then estimated the density at 0 for all difference distributions through a Gaussian density kernel.

For example, to make an inference on whether the correlation between MZ twins differed from the correlation between DZ twins for thresholds (top-left panel of Figure 6),

we obtained the difference posterior by subtracting the joint posterior samples of the DZ twin correlation from the MZ twin correlation. We then obtained the difference prior by subtracting the 1,000,000 samples from the DZ twin correlation prior from the MZ twin correlation prior. Lastly, we divided a density estimate of the difference posterior at 0 by an estimate of the difference prior at 0, giving the Bayes factor in favour of there being a difference between the distributions.

*Parameter recovery analysis*

To test the robustness of our inferences on the estimated group-level correlations within our joint model of heritability, we performed a parameter recovery simulation. Within this recovery, we specifically focus on the recovery of these group-level correlations (shown in **bold** in the formal model definition below), as they 1) are the only parameter values that we perform inferences on within our assessment of heritability, and 2) form the novel extension of our approach. Towards this goal, we generated a single group of synthetic participant pairs from the following simplified model:

$$
\begin{aligned}
\text{A} &\sim \text{N}\left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}, 0.04 \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix}\right) \\
\text{b} &\sim \text{N}\left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}, 0.04 \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix}\right) \\
\text{vc} &\sim \text{N}\left(\begin{bmatrix} 3 \\ 3 \end{bmatrix}, 0.04 \begin{bmatrix} 1 & \mathbf{0.7} \\ \mathbf{0.7} & 1 \end{bmatrix}\right) \\
\text{ve} &\sim \text{N}\left(\begin{bmatrix} 2 \\ 2 \end{bmatrix}, 0.04 \begin{bmatrix} 1 & \mathbf{0.5} \\ \mathbf{0.5} & 1 \end{bmatrix}\right) \\
\text{t0} &\sim \text{N}\left(\begin{bmatrix} 0.2 \\ 0.2 \end{bmatrix}, 0.04 \begin{bmatrix} 1 & \mathbf{0.2} \\ \mathbf{0.2} & 1 \end{bmatrix}\right)
\end{aligned}
$$

where *vc* refers to the correct drift rate, and *ve* refers to the incorrect drift rate.

For this simulation, we generated 150 pairs of synthetic participants, with 150 trials per participant. Figure 3 displays the recovered posterior distributions of the correlations parameters, with a red line at the generating value. All estimated posteriors are centered on approximately their true generating value, showing an good overall recovery performance. Further, the parameters that were generated with high correlations (*vc* and *ve*) seem to show near-perfect recovery, with the distributions centered on the generating value, and the posteriors being very narrow. Thirdly, the parameters that were generated with no correlation (*A* and *b*) were recovered without any systematic bias, though the precision was lower than when generated with a higher correlation, with the variance of the posterior being much wider. Lastly, the *t*0 parameter seems to break the "rule" suggested by the previous two trends, showing posteriors that are about as wide, if not wider, than

the parameters generated with 0 correlation, despite $t0$ being generated with a non-zero correlation (0.2). Since the $t0$ parameter is not tightly constrained in the LBA, as opposed to the diffusion model where $t0$ must take a value very close to the minimum time of the response time distribution, this poorer recovery may make some sense (Tillman & Logan, 2017).

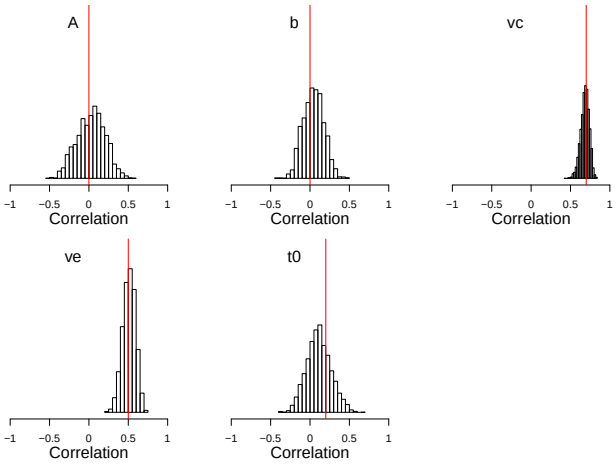


*Figure 3.* Estimated posteriors for the group-level correlations between pairs for each of the five parameters. The x-axis displays the correlation values, and the y-axis displays the frequency at which this correlation was sampled. The red vertical line displays the true generating value for that parameter.

## Results

### *Standard analysis results*

We calculated the $h^2$ index of heritability, as well as the parameters of the ACE model, using five different dependent variables that might reasonably be extracted from the data: response time mean, response time variance, response time minimum, decision accuracy, and the proportion of non-responses. Of course, many other summary statistics could be analyzed, which again highlights the strength of the model-based analysis, which entails all such choices, by considering the full structure of the response time distribution.

The correlations for each relatedness group, and the corresponding $h^2$ values can be seen in Table 1. For all five variables, the correlations were largest for MZ twins (ranging from $r = 0.05$ to $r = 0.62$), though there was a split between DZ twins ($-0.1 < r < 0.41$) and non-twin siblings ($-0.05 < r < 0.37$) for next largest across the five variables. These correlations lead to heritability estimates ranging from just $h^2 = 0.3$ (for minimum

Table 1: Correlations between the different relatedness groups (columns) for four summary statistics (rows). The four right-most columns shows heritability ($h^2$), additive genetic component ($a^2$), shared environmental component ($c^2$), and unshared environmental component ($e^2$).

| | MZ | DZ | Sib | $h^2$ | $a^2$ | $c^2$ | $e^2$ |
|---|---|---|---|---|---|---|---|
| RT Mean | 0.44 | 0.26 | 0.17 | 0.37 | 0.37 | 0.08 | 0.56 |
| Proportion Correct | 0.62 | 0.41 | 0.37 | 0.43 | 0.43 | 0.19 | 0.38 |
| Proportion Missed | 0.43 | 0.09 | 0.15 | 0.68 | 0.38 | 0 | 0.62 |
| RT Variance | 0.42 | 0.02 | 0.08 | 0.8 | 0.34 | 0 | 0.66 |
| RT Minimum | 0.05 | -0.1 | -0.05 | 0.3 | 0 | 0 | 1 |

RT) to $h^2 = 0.68$ (for the proportion of missed responses). This raises the question of what inferences should be drawn about the heritability of cognitive performance, given the different results from the four different summary statistics. It might be argued that RT mean is more important than the other variables, but it is difficult to imagine principled reasons on which to prefer one of these measures over the others, or a coherent basis on which to weight the different outcomes.

Table 2: The goodness-of-fit for the ACE models for each variable. Displays the actual covariances (Act Cov) and estimated covariances (Est Cov) in the data for MZ and DZ twins, as well as the estimated covariances of the models

| | Act Cov | | Est Cov | |
|---|---|---|---|---|
| | MZ | DZ | MZ | DZ |
| RT Mean | 0.00758 | 0.00445 | 0.00758 | 0.00445 |
| Proportion Correct | 0.00311 | 0.00203 | 0.00311 | 0.00203 |
| Proportion Missed | 0.0012 | 0.00025 | 0.00106 | 0.00053 |
| RT Variance | 0.0004 | 0.00002 | 0.00033 | 0.00016 |
| RT Minimum | 0.0008 | -0.00159 | 0 | 0 |

Table 1 shows the results of the ACE modelling for each of the five variables. For mean RT and accuracy, the estimates for the additive genetic component, $a^2$, were identical to that of the basic heritability measure, $h^2$. The ACE model also provided a good fit to the data for these variables, perfectly predicting the covariance in MZ and DZ twins, as seen in Table 2. For the variance in RT, minimum RT, and the proportion of misses, there were discrepancies between the $a^2$ estimates and the $h^2$ calculation, with the $a^2$ estimates being substantially lower for all three variables (going as low as 0 for minimum RT). However, the ACE model also did a much poorer job predicting the covariances for these variables, as can be seen in Table 2. This seems to be due to the inability of the ACE model to predict covariance for the MZ twins that is more than double that for DZ twins, or a negative

covariance, resulting in smaller $a^2$ estimates and poorer fits when this is the case within the data. As the $a^2$ estimates were identical to $h^2$ when the ACE model provided a good fit to the data, we limit discussion below to the simpler $h^2$ values.

*Model-based heritability results*

Figure 4 displays how our novel LBA model extension accounts for the data, through group-averaged cumulative distribution function (CDF) plots. The green and red dots display the data, and the green and red lines display the predictions of the model, across the distribution of RTs. Green lines and dots indicate distributions for correct responses, and red lines and dots indicate distributions for incorrect responses. The bottom-left corner of each panel also includes the the proportion of missed trials (non-responses), using the y-axis scale. The model accounts for the data quite well, closely matching the RT in accuracy in all conditions, apart from some minor misfit in the lure and target conditions, where the model slightly over-estimates and under-estimates the variance in RT, respectively. It is natural that the model fits less well in the target and lure conditions, as those two conditions have extremely small numbers of trials (16 in each, per participant), as conditions with larger number of trials will have a greater impact on the likelihood, and the exact empirical trends for each individual are likely quite noisy.

Next, we move on to the assessment of the correlations of the latent parameters of each of the different groups. Firstly, it is interesting to assess the estimated posterior distributions for each of the correlation parameters, separately for each relatedness group and for the three correlation parameters (drift rates, thresholds, and non-decision times). Figure 5 displays these posterior distributions, with different colors for the different relatedness groups, and different panels for the different parameters. As with the $h^2$ estimates, the correlations for MZ twins are the highest for all parameters, though DZ twins and non-twin siblings are difficult to distinguish from one another. In terms of the widths of the correlation distributions, the drift rate distributions appear to have the lowest variance (SD: MZ = 0.058, DZ = 0.076, SIB = 0.048), with thresholds being slightly more variable (SD: MZ = 0.075, DZ = 0.111, SIB = 0.064), and the non-decision time distributions being highly variable (SD: MZ = 0.195, DZ = 0.223, SIB = 0.166). These results make sense given the parameter recovery analysis above, which found the drift rate distributions to be estimated with low variance, and the non-decision time distribution to be estimated with much higher variance. However, despite the differences in posterior width, there appears to be a great deal of overlap between every group's distribution for each parameter, apart from the MZ twins for threshold, which barely overlap with DZ twins or non-twin siblings.

Figure 6 displays the distributions of the differences between the correlations of the relatedness groups for the different latent parameters. These allow the assessment of the heritability and environmental influences on the different latent parameters, and include both the posterior distributions estimated and the prior distributions required for the Savage-Dickey ratio, as well as the Savage-Dickey ratio estimate of the Bayes factor. The left column displays the difference distribution for the MZ twins and DZ twins correlations,

which indicate the influence of heritability on the parameters. Only threshold displays decisive evidence on whether or not an effect is present, showing strong evidence for the MZ correlation being higher than the DZ correlation, and therefore, suggesting genetic heritability in threshold (95% HDI = [0.11, 0.63]). Both drift rate (95% HDI = [-0.02, 0.37]) and non-decision time (95% HDI = [-0.35, 0.79]) yielded Bayes factors very close to 1, suggesting that the evidence is ambiguous for whether or not the MZ correlation is higher than the DZ correlation. The right column displays the difference distribution for the DZ twins and non-twins siblings correlations, which indicate the influence of environment on the parameters. Both threshold and drift rate display moderate evidence in favor of no difference between the DZ and non-twin sibling correlations, suggesting that there is little evidence that there are systematic environmental influences that are shared between family members on threshold or drift rate. Non-decision time again showed a Bayes factors very close to 1, suggesting that the evidence is quite ambiguous for whether or not the DZ correlation is higher than the sibling correlation.

## Discussion

We aimed to investigate different components of the heritability of cognitive performance in a standard working memory task, using the twin study paradigm. Using data from the Human Connectome Project (Van Essen et al., 2013), we developed and tested a new approach to jointly model response times, response accuracy, and twin-pair correlations. Our method employed a Bayesian hierarchical statistical approach, and used a cognitive decision-making model to separate out different sources of variability in data. While previous studies made the simple assumption that observed heritability in response time is attributable solely to heritability in the speed of mental processing (Luciano et al., 2001; Finkel & Pedersen, 2004; Vernon, 1989; Beaujean, 2005; Ogata et al., 2014; Kochunov et al., 2016; Posthuma et al., 2002), we found greater genetic contribution to response caution than to underlying cognitive speed within the memory task assessed. Our results fall in line with the previous findings of Engelhardt et al. (2016), who used psychometric testing to show that a strong genetic component exists within executive functions.

Previous twin-based studies have investigated the heritability of decision-related processes, such as processing speed, through psychometric tests, linear correlations, and structural equation models (Vernon, 1989). However, such approaches do not naturally differentiate between the specific components of mental processing speed and physical processing speed, or account for the well-known speed-accuracy tradeoff. This means that previous conclusions about the heritability of processing speed might actually be better interpreted as being about the heritability of caution, in much the same manner as age-related slowing was originally mis-attributed to processing slowdown instead of increased caution. Our approach differentiates between the heritability of these different latent variables by using a cognitive model.

Our analysis resolves another limitation of standard methods, involving the choice of

summary statistic. When assessing the heritability of cognitive speed, traditional methods require the observed data to be reduced to summary statistics (nearly always the just the mean). The decision as to which discrete variables should be taken, and how these variables should be combined, can be unclear. An unfortunate choice of summary statistic could lead to either over-estimation of heritability (through a process of multiple comparisons, much like p-hacking) or to an under-estimation of heritability (for example, if twins co-vary more in the variance of their data than the mean). Analysis using a cognitive model of decision-making coherently includes the entire data distributions, resolving this problem.

Our study is the first to assess the heritability in the latent components of the decision-making process, and as such, the results should be interpreted with caution. Although there was clear evidence that response caution can be heritable, and that it shows greater evidence of heritability in this dataset than mental processing speed, it may be the case that response caution in other datasets is less heritable than mental processing speed. Therefore, we do not believe that our results are the basis for a strong claim that response caution is generally more heritable than mental processing speed, but instead that response caution can be heritable and needs to be considered in future research, and that within working memory tasks it appears to be more heritable than mental processing speed.

Our method provides an interesting avenue for future research with complex data, as a method for using and interpreting the covariance structure of the data set to provide better estimation and better understanding of the task. For example, some large data sets may contain multiple sessions of experimental data from the same task for the same person, but the person has completed these different sessions on different devices, such as a computer and a mobile phone. As both sessions of data were collected from the same person for the same task, one would expect them to be highly related. However, based on the different device, which has a different viewing size, response mechanism, etc., one would expect that performances should differ in some unknown, but systematic, way. Our method provides a sensible way of using this relationship to better constrain data estimation. Not only does this provide a sensible way of informing parameter estimation based on the relationship between different sources of data, but it also allows for a measurement of how related the different parameters are between tasks. For example, one could potentially learn that threshold is fairly consistent (i.e., high covariance) across devices, but that drift rate is highly variable (i.e., low covariance). Therefore, we believe that in addition to our interesting results regarding the heritability of decision-making components, our study provides a useful method for future research investigating complex data within cognitive science.

Additionally, our method suggests two new avenues for future research in cognitive heritability. Like the study of Ratcliff et al. (2011), not all tasks share the same differences and similarities of the underlying components (e.g. young and old people in associative and item recognition), meaning that the heritability of other cognitive tasks might be explored through cognitive models. Secondly, as genetic research moves towards a focus on DNA sequences over twin studies, future research could focus on combining cognitive models

with detailed genetic sequences.

Lastly, although our study has found response caution to be a highly heritable trait, this does not imply that response caution is not under cognitive control. Many previous studies have shown that people are able to change their level of response caution to better account for task demands (Wagenmakers, Ratcliff, Gomez, & McKoon, 2008; Rae, Heathcote, Donkin, Averell, & Brown, 2014; Ratcliff & Rouder, 1998). Our findings instead suggest that the manner in which people set their response caution by default have a large genetic component. Future research could explore if the way in which people adapt their response caution to different task demands, such as those in the papers cited above, also has a large heritable component, or if only the default settings appear to show such a finding.

# References


Beaujean, A. A. (2005). Heritability of cognitive abilities as measured by mental chronometric tasks: A meta-analysis. *Intelligence*, *33*(2), 187–201.

Bouchard, T. J. (2004). Genetic influence on human psychological traits a survey. *Current Directions in Psychological Science*, *13*(4), 148–151.

Bouchard, T. J., McGue, M., et al. (1981). Familial studies of intelligence: A review. *Science*, *212*(4498), 1055–1059.

Brown, S. D., & Heathcote, A. J. (2008). The simplest complete model of choice reaction time: Linear ballistic accumulation. *Cognitive Psychology*, *57*, 153–178.

Brown, S. D., Marley, A. A. J., Donkin, C., & Heathcote, A. (2008). An integrated model of choices and response times in absolute identification. *Psychological review*, *115*(2), 396.

Busemeyer, J. R., & Townsend, J. T. (1992). Fundamental derivations from decision field theory. *Mathematical Social Sciences*, *23*(3), 255–282.

DeFries, J. C., & Fulker, D. W. (1985). Multiple regression analysis of twin data. *Behavior genetics*, *15*(5), 467–473.

Donkin, C., & Brown, S. D. (2017). Response time modeling. In J. T. Wixted & E.-J. Wagenmakers (Eds.), *The stevens' handbook of experimental psychology and cognitive neuroscience, volume 5, fourth edition. the stevens' handbook of experimental psychology and cognitive neuroscience* (Vol. 5).

Donkin, C., Brown, S. D., & Heathcote, A. (2009). The overconstraint of response time models: Rethinking the scaling problem. *Psychonomic Bulletin & Review*, *16*(6), 1129–1135.

Engelhardt, L. E., Mann, F. D., Briley, D. A., Church, J. A., Harden, K. P., & Tucker-Drob, E. M. (2016). Strong genetic overlap between executive functions and intelligence. *Journal of Experimental Psychology: General*.

Erlenmeyer-Kimling, L., & Jarvik, L. F. (1963). Genetics and intelligence: A review. *Science*, *142*(3598), 1477–1479.

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological bulletin*, *53*(2), 134.

Evans, N. J., & Brown, S. D. (2017a). Bayes factors for the linear ballistic accumulator model of decision-making. *Behavior Research Methods*, 1–15.

Evans, N. J., & Brown, S. D. (2017b). People adopt optimal policies in simple decision-making, after practice and guidance. *Psychonomic Bulletin & Review*, *24*(2), 597–606.

Evans, N. J., Hawkins, G. E., Boehm, U., Wagenmakers, E.-J., & Brown, S. D. (2017). The computations that support simple decision-making: A comparison between the diffusion and urgency-gating models. *Scientific Reports*, *7*, 16433.

Evans, N. J., Rae, B., Bushmakin, M., Rubin, M., & Brown, S. D. (2017). Need for closure is associated with urgency in perceptual decision-making. *Memory & Cognition*, *45*, 1193–1205.

Finkel, D., & Pedersen, N. L. (2004). Processing speed and longitudinal trajectories of change for cognitive abilities: The swedish adoption/twin study of aging. *Aging Neuropsychology and Cognition*, *11*(2-3), 325–345.

Forstmann, B. U., Schafer, A., Anwander, A., Neumann, J., Brown, S. D., Wagenmakers, E.-J., ... Turner, R. (2010). Cortico-striatal connections predict control over speed and accuracy in perceptual decision making. *Proceedings of the National Academy of Sciences*, *107*, 15916–15920.

Humphreys, M. S., & Revelle, W. (1984). Personality, motivation, and performance: a theory of the relationship between individual differences and information processing. *Psychological review*, *91*(2), 153.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: individual differences in working memory. *Psychological review*, *99*(1), 122.

Kochunov, P., Thompson, P. M., Winkler, A., Morrissey, M., Fu, M., Coyle, T. R., ... others (2016). The common genetic influence over processing speed and white matter microstructure: Evidence from the old order amish and human connectome projects. *Neuroimage*, *125*, 189–197.

Lu, Z.-L., & Dosher, B. A. (2008). Characterizing observers using external noise and observer models: assessing internal representations with external noise. *Psychological review*, *115*(1), 44.

Luce, R. D. (1986). *Response times*. New York: Oxford University Press.

Luciano, M., Wright, M., Smith, G., Geffen, G., Geffen, L., & Martin, N. (2001). Genetic covariance among measures of information processing speed, working memory, and iq. *Behavior genetics*, *31*(6), 581–592.

Ogata, S., Kato, K., Honda, C., & Hayakawa, K. (2014). Common genetic factors influence hand strength, processing speed, and working memory. *Journal of epidemiology*, *24*(1), 31–38.

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. *Psychological review*, *117*(3), 864.

Posthuma, D., Mulder, E., Boomsma, D., & De Geus, E. (2002). Genetic analysis of iq, processing speed and stimulus-response incongruency effects. *Biological Psychology*, *61*(1), 157–182.

Rae, B., Heathcote, A., Donkin, C., Averell, L., & Brown, S. (2014). The hare and the tortoise: Emphasizing speed can change the evidence used to make decisions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(5), 1226.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological review*, *85*(2), 59.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*(5), 347–356.

Ratcliff, R., & Starns, J. J. (2013). Modeling confidence judgments, response times, and multiple choices in decision making: Recognition memory and motion discrimination. *Psychological review*, *120*(3), 697.

Ratcliff, R., Thapar, A., & McKoon, G. (2001). The effects of aging on reaction time in a signal detection task. *Psychology and aging*, *16*(2), 323.

Ratcliff, R., Thapar, A., & McKoon, G. (2010). Individual differences, aging, and iq in two-choice tasks. *Cognitive psychology*, *60*(3), 127–157.

Ratcliff, R., Thapar, A., & McKoon, G. (2011). Effects of aging and iq on item and associative memory. *Journal of Experimental Psychology: General*, *140*(3), 464.

Reber, A. S., Walkenfeld, F. F., & Hernstadt, R. (1991). Implicit and explicit learning: Individual differences and iq. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(5), 888.

Ter Braak, C. J. (2006). A markov chain monte carlo version of the genetic algorithm differential evolution: easy bayesian computing for real parameter spaces. *Statistics and Computing*, *16*(3), 239–249.

ter Braak, C. J., & Vrugt, J. A. (2008). Differential evolution markov chain with snooker updater and fewer chains. *Statistics and Computing*, *18*(4), 435–446.

Tillman, G., & Logan, G. D. (2017). The racing diffusion model of speeded decision making.

Trueblood, J. S., Brown, S. D., & Heathcote, A. (2014). The multiattribute linear ballistic accumulator model of context effects in multialternative choice. *Psychological review*, *121*(2), 179.

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.

Ulrich, R., & Miller, J. (1993). Information processing models generating lognormally distributed reaction times. *Journal of Mathematical Psychology*, *37*(4), 513–525.

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., ... others (2013). The wu-minn human connectome project: an overview. *Neuroimage*, *80*, 62–79.

Vernon, P. A. (1989). The heritability of measures of speed of information-processing. *Personality and Individual Differences*, *10*(5), 573–576.

Visscher, P. M., Hill, W. G., & Wray, N. R. (2008). Heritability in the genomics era—concepts and misconceptions. *Nature Reviews Genetics*, *9*(4), 255–266.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the savage–dickey method. *Cognitive psychology*, *60*(3), 158–189.

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & McKoon, G. (2008). A diffusion model account of criterion shifts in the lexical decision task. *Journal of Memory and Language*, *58*(1), 140–159.

Wetzels, R., Grasman, R. P., & Wagenmakers, E.-J. (2010). An encompassing prior generalization of the savage–dickey density ratio. *Computational Statistics & Data Analysis*, *54*(9), 2094–2102.

Zyphur, M. J., Zhang, Z., Barsky, A. P., & Li, W.-D. (2013). An ace in the hole: Twin family models for applied behavioral genetics research. *The Leadership Quarterly*, *24*(4), 572–594.

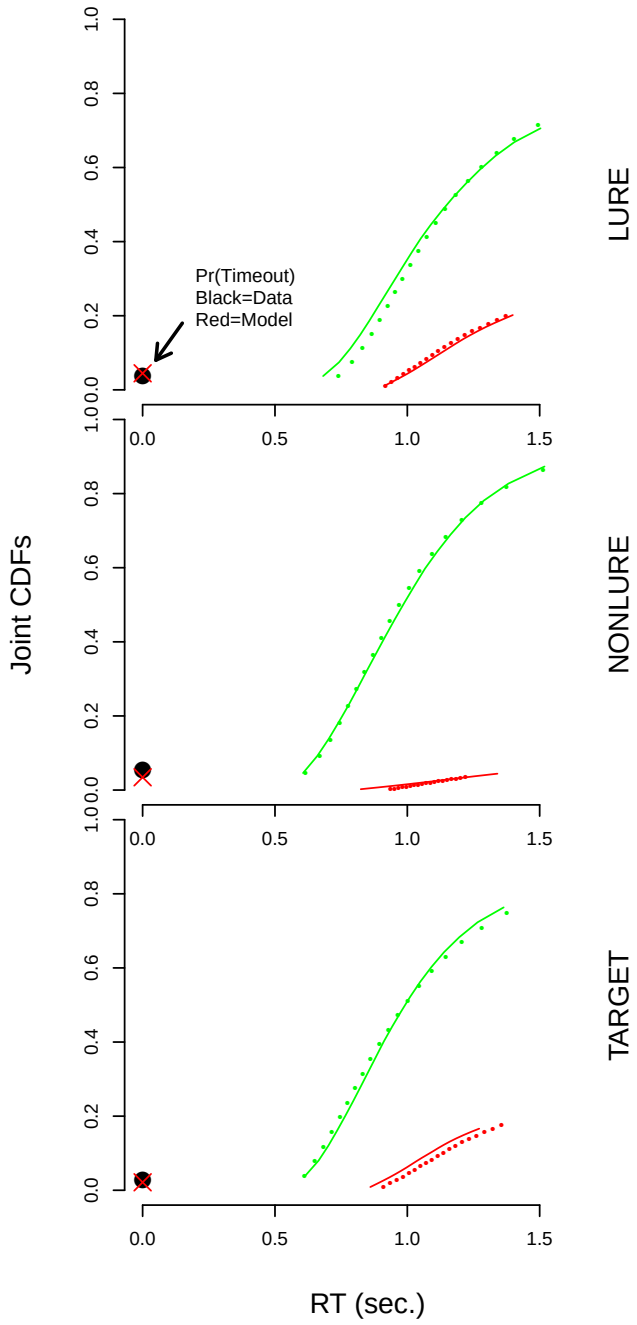


*Figure 4.* Joint cumulative distribution function plots, calculated through averaging percentiles over participants. The three panels display the actual data and model predictions for the three conditions: lure, non-lure, and target. The x-axis displays the RT, with the 19 points being the $5^{th}, 10^{th}, ..., 95^{th}$ percentiles of the data. The y-axis displays the proportion of responses observed faster than each percentile, in the appropriate response class (correct vs. incorrect responses). The dots show the data and the lines show the model predictions, with green for correct responses and red for error responses. The proportion of missed responses (non-responses) observed in each condition is shown by the height of the black dot in the lower left corner, with the corresponding model prediction given by the red cross.

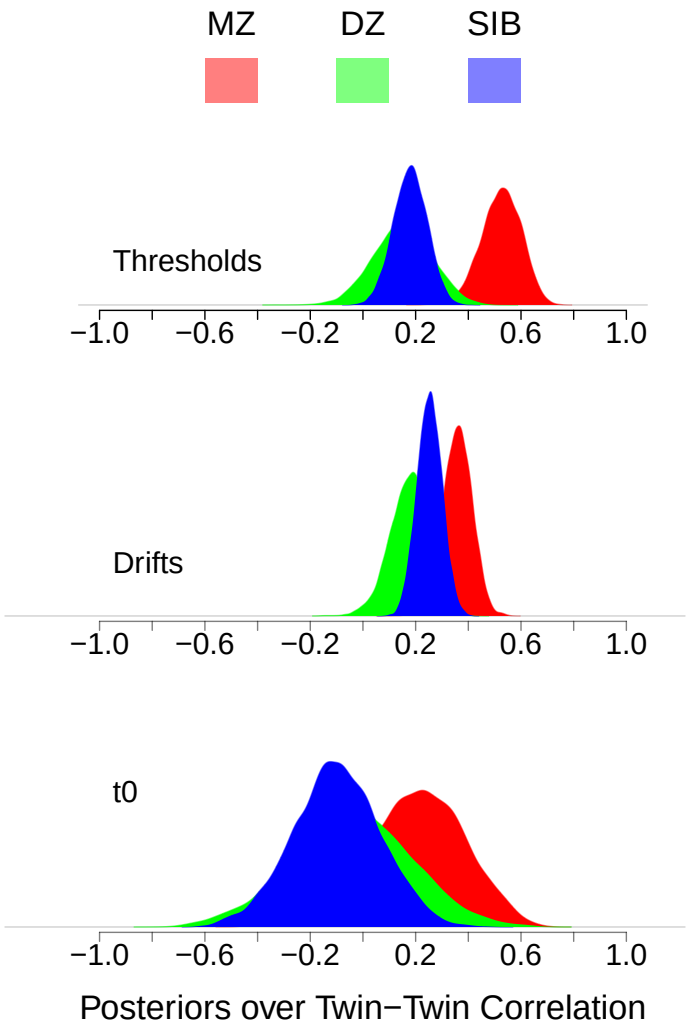


*Figure 5.* Marginal posterior distributions over correlation parameters (i.e., the $\rho$ parameters in the model definition) of the joint model. Each row gives a different parameter of the model: thresholds (response caution), drift rates (processing speed), and non-decision time (physical processing speed). The different colored distributions give the different levels of relatedness, with monozygotic twins in red, dizygotic twins in green, and siblings in blue.

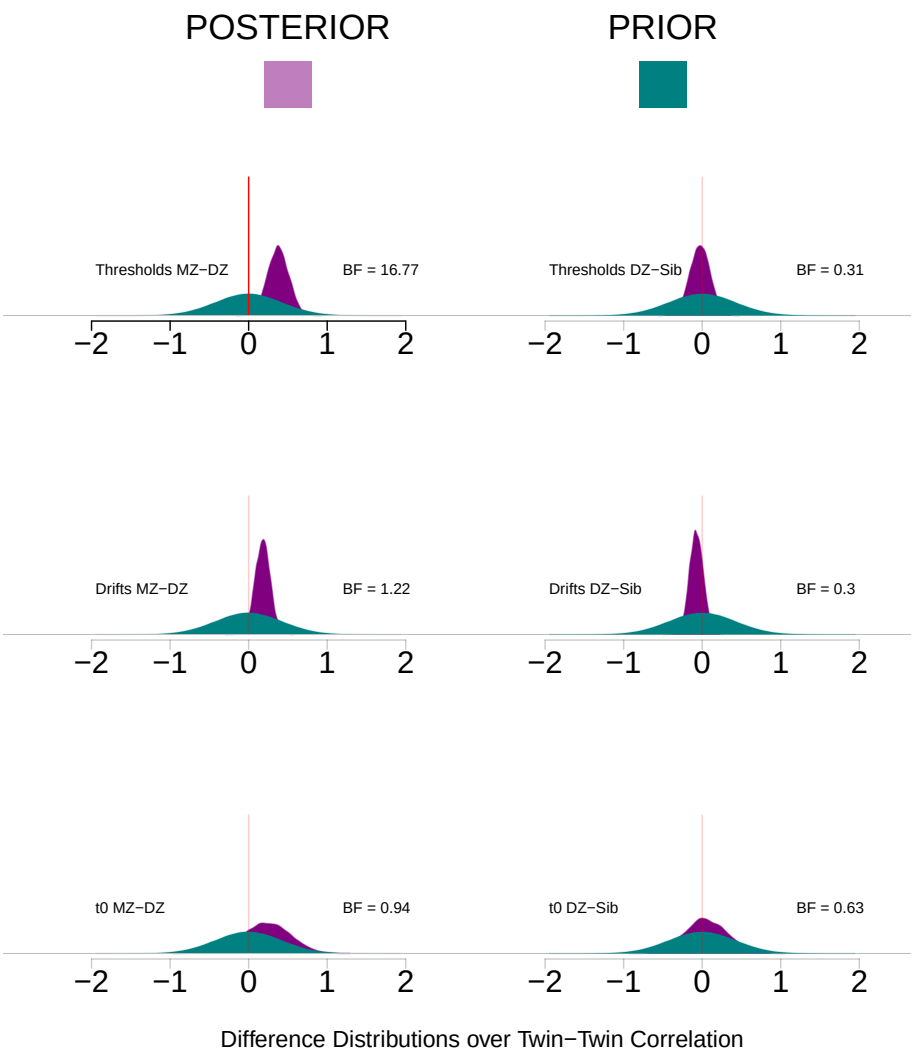


*Figure 6.* Displays the posterior (purple) and prior (light blue) difference distributions for each type of parameter (threshold, drift rate, non-decision time) for the correlation differences that are meaningful to interpreting heritability and environmental influences. The first of these difference distributions, MZ twin correlation minus DZ twin correlation, is displayed in the left column. These difference distributions give an indication of the heritability in each of the parameters. The second of these difference distributions, DZ twin correlation minus non-twin sibling correlation, is displayed in the right column. These difference distributions give an indication of the shared family environmental influences in each of the parameters. Each panel displayed a red vertical line at 0, the point whether the Savage-Dickey ratio between the posterior and prior densities is applied, with the Savage-Dickey ratio estimate of the Bayes factor being displayed as "BF", which is expressed in favor of the alternative. There is only strong evidence for heritable influences in thresholds, with drift rate and non-decision time showing no strong evidence either for or against an effect. For environmental influences, there is moderate evidence in favor of no effect for threshold and drift rate, though non-decision time is again ambiguous.