Estimating Correlations Across Tasks In Experimental Psychology

Shanglin Yang[1] & Jeffrey N. Rouder[1]

[1] University of California, Irvine

Author Note

Version 2, Sept/Oct, 2025

Correspondence concerning this article should be addressed to Jeffrey N. Rouder, Department of Cognitive Science, University of California, Irvine, CA, 92697. E-mail: jrouder@uci.edu

Abstract

Understanding how people covary in performance across experimental tasks is central to individual-difference psychology. The classic Pearson correlation has two strengths: (1) it is invariant to the scale of measurement, and (2) it is invariant to including additional variables in the analysis. However, it is susceptible to attenuation from measurement noise. Bayesian hierarchical models address this issue by modeling measurement error directly. Resulting estimates, however, depend on prior specifications and are not invariant to scale or variable inclusion. We compare three common priors—Inverse Wishart (IW), Scaled Inverse Wishart (SIW), and LKJ—to assess robustness to prior assumptions in hierarchical settings. Our main tools are visualizing the priors and evaluating their effects on posterior estimates through simulation. When prior settings match ground truth, all priors recover true correlations accurately in low-dimensional settings. When prior variance is misspecified, the IW shows strong bias: low-variance priors inflate correlations, and high-variance priors deflate them. The SIW shows the same pattern but less severely, while the LKJ remains largely unaffected by scale misspecification. When more variables are added, the IW is most stable, whereas the SIW and LKJ show slight shrinkage toward lower correlations. The main drawback of the LKJ is computational speed—models with it can take orders of magnitude longer than those using IW or SIW. Overall, the LKJ provides the most accurate estimates, while the SIW offers a practical compromise for large-scale models where computational speed is crucial.

Keywords: Individual differences, correlations, hierarchical models, prior selection, Bayesian analysis

Estimating Correlations Across Tasks In Experimental Psychology

It is common in experimental psychology to study how people covary across tasks. An example is assessing whether people who are highly susceptible to Stroop interference are also susceptible to Flanker interference. Researchers who study individual differences hope that the pattern of observed correlations may provide insight into the nature and dimension of processing. Here are two well-known examples where this hope was realized: First, in personality psychology, the pattern of correlations across various personality questionnaire items forms the evidence for the Big 5 theory of personality (McCrae & Costa Jr, 1997). Second, in cognitive psychology, the pattern of correlations across executive-function tasks forms the evidence for various theories of executive function including the Inhibition-Shifting-Updating theory of Miyake et al. (2000). Although each of these theories has been criticized, the general method of studying covariation across tasks remains timely and topical.

Estimating correlations at first glance seems straightforward: Point estimates come from Pearson's sample correlation formula (Pearson, 1900); confidence intervals (CIs) come from Fisher's z-transform method (Fisher, 1921). Each of these computations has the following beneficial property: The analyst can change the *location* and *scale* of the variables without changing the estimation of correlation. For example, suppose we are studying the association between heart rate and ambient temperature. One analyst chooses to measure temperature in Celsius while another in Fahrenheit even though the measure vary in location (where the zero point is) and scale (how big a degree is). Fortunately, the point estimate and associated CIs do not depend on this choice—the same values are obtained under linear transform of either variable. This property may be termed **location and scale invariance**. Even more impressively, each of the computations display **invariance to inclusion of other variables**. For example, suppose one researcher considered height and weight in isolation and another considered height, weight, shoe size

and waist circumference simultaneously. It is not obvious that the correlation coefficient between weight and height would be the same for both researchers. In the latter case, the resulting set of correlation in the correlation matrix must be positive semidefinite implying a common constraint. Yet, it may be shown that the correlation for the latter depends only on height and weight, and the point estimate and CIs are the same in both cases. Given these two advantages, as well as the ease of computation, estimation of correlations seems a solved problem.

A critical complication comes when variables are measured with substantial measurement error, as they almost always are in psychology. Let's take the following case where each of many individuals participates in two psychological tasks. Let Y_{i1} and Y_{i2} be the scores for the ith individual, $i = 1, \ldots, I$ respectively for Task 1 and Task 2. To incorporate measurement error, let

$$Y_{ij} \mid \theta_{ij} \sim \mathrm{N}(\theta_{ij}, \tau_j^2), \quad j = 1, 2,$$

where θ_{i1} and θ_{i2} are true scores for the ith individual for respectively Task 1 and Task 2. The terms τ_1 and τ_2 are the error variance for the two tasks. Figure 1A shows a scatter plot of hypothetical individual true scores θ_{i1} and θ_{i2} for $I = 200$ individuals. These are correlated; the true population value is .7 and the goal is to recover this value. The sample correlation for these 200 individuals is .71, which is quite close to the true population value. Figure 1B shows the scatter plot of observed scores Y_{i1} and Y_{i2}, and again, we wish to recover the true value of .7. These observed scores are perturbed by error in random directions and to random degrees. The net result is that the correlation is attenuated (Spearman, 1904). The sample correlation on observed scores is .41 in value, which is quite far from the true population value of .7.

More formal development is instructive in addressing the problem. First, a bivariate

normal is placed on true scores:

$$\begin{pmatrix}\theta_{i1}\\ \theta_{i2}\end{pmatrix} \sim \mathrm{N}_2\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}\right).$$

The key quantity here is population correlation ρ, and it is this quantity we hope to recover. Of course, we do not observe true scores, but the observations. The distribution on Y_{ij} is:

$$\begin{pmatrix}Y_{i1}\\ Y_{i2}\end{pmatrix} \sim \mathrm{N}_2\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2+\tau_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2+\tau_2^2\end{pmatrix}\right).$$

From this distribution, it is now clear that the sample correlation from observed scores, denoted $\hat{\rho}$ measures:

$$E(\hat{\rho}) \approx \left(\frac{\sigma_1\sigma_2}{\sqrt{\sigma_1^2+\tau_1^2}\sqrt{\sigma_2^2+\tau_2^2}}\right)\rho.$$

The coefficient here describes attenuation as the denominator necessarily larger than the numerator. And without any further information about measurement noise, we are unsure by how much.

Our application is to experimental psychology, and tasks tend to be comprised of many repeated trials. The usual course of analysis is averaging over replicates to form individual-by-task scores which are then correlated. There is noise in these trials, which means the averages also have noise, although certainly less so. This trial noise effect on the averages results in attenuation. Moreover, in this averaging approach, the only way to reduce attenuation is to increase the number of trials per individual. It cannot be reduced by adding more individuals as adding individuals simply adds more perturbed points to the scatter plot.

An attractive alternative to averaging for disattenuating correlations in analysis is through hierarchical models (Haines et al., 2025; Matzke et al., 2017; Rouder & Haaf, 2019). In a hierarchical model, multiple sources of noise are specified and all the data, not

just the averages, are used. For the above example, trial noise as well as covariability of people in tasks are included. Hierarchical models have become popular in experimental psychology because they lead to more accurate estimation and inference across a wealth of paradigms (Rouder & Lu, 2005; Rouder & Province, 2019). In this case, the separation of trial noise from variability across tasks in suitable hierarchical models leads to disattenuated estimates of correlation (Rouder, Chavez de la Peña, Mehrvarz, & Vandekerckhove, 2023; Rouder & Mehrvarz, 2024). As proof of concept, Figure 1C shows the posterior distribution of ρ when all the data are analyzed in a subsequently presented hierarchical model. As can be seen, they are centered close to the true value of .7 rather than the attenuated value of .41.

Estimation of correlation in hierarchical models is not without difficulties. The measurement invariances in sample correlations—invariance to location and scale and invariance to inclusion of additional variables—may not be attainable or even desirable in hierarchical settings.

Analysis of hierarchical models is convenient in the Bayesian framework (Gelman, Carlin, Stern, & Rubin, 2004). This paper addresses how to estimate a correlation matrix in both hierarchical and conventional normal models. A key issue is the specification of the prior. How does the choice of prior affect the posterior distribution of correlation coefficients? We frame our assessment in terms of measurement invariances highlighted above. Invariance to location occurs fairly naturally for reasons discussed subsequently and will not play a role. Invariance to scale and invariance to inclusion of additional variables is more difficult with the methods we present. For example, in all the prior specifications discussed here, the analyst must set a scale or expectation of the degree of variability beforehand. Robustness to scale specification occurs when the posterior of correlation coefficients are relatively unaffected by reasonable variation in this scale setting. Likewise, robustness to inclusion occurs when the posterior of correlations coefficients is relatively unaffected by the inclusion or exclusion of a handful of additional variables.

We study the following prior classes for correlations: The Inverse Wishart prior (O'Hagan & Forster, 2004), which offers advantages of conjugacy, computational convenience; the Scaled Inverse Wishart prior (Huang & Wand, 2013), which is a continuous mixture of Inverse Wishart components and provide more flexibility in modeling variances; the LKJ prior (Lewandowski, Kurowicka, & Joe, 2009), which is less informative than either Wishart-based alternatives (Tokuda, Goodrich, Mechelen, Gelman, & Tuerlinckx, 2025). This paper explores the conditions under which each of these choices is useful for estimating correlations. Our main approach to comparing these models is through simulation with known truths. In each simulation run, posterior distributions of population-level correlations are compared to these truths. This simulation approach is precedented for covariance matrices, and examples include Rouder et al. (2007), Schuurman, Grasman, and Hamaker (2016); Liu, Zhang, and Grimm (2016). To our knowledge, however, we are the first to do so for assessing correlations across tasks in experimental designs. Moreover, although the Inverse Wishart prior has been studied extensively, there is far less work with the Scaled Inverse Wishart and LKJ priors, perhaps because they are relatively new and only incorporated into Jags and Stan in the last decade.

Manifest and Hierarchical Models

We use the term *manifest* model for the conventional case where researchers place a multivariate model on observables, and manifest models are appropriate when there is an ignorable degree of measurement error. Example of variables which we may treat as free of measurement error are body metrics such as height, weight, and waist circumference. This case is easy in the conventional setup; one simply uses sample correlations along with Fisher confidence intervals. For the Bayesian case, the manifest model is given as follows: Let Y_{ij} denote the score for the ith individual, $i = 1, \ldots, I$ on the jth task, $j = 1, \ldots, J$,

and let $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iJ})'$ be a column vector of scores for the ith individual:

$$\text{Manifest Model}: \qquad \boldsymbol{Y}_i \sim \mathrm{N}_J(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_J)'$ is a column vector of means per task and $\boldsymbol{\Sigma}$ is a variance matrix:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \ldots & \rho_{1J}\sigma_1\sigma_J \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \ldots & \rho_{2J}\sigma_2\sigma_J \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1J}\sigma_1\sigma_J & \rho_{2J}\sigma_2\sigma_J & \ldots & \sigma_J^2 \end{bmatrix}.$$

The variance matrix can be expressed in terms of a correlation matrix as follows:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1 & 0 & \ldots & 0 \\ 0 & \sigma_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_J \end{bmatrix} \times \begin{bmatrix} 1 & \rho_{12} & \ldots & \rho_{1J} \\ \rho_{12} & 1 & \ldots & \rho_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1J} & \rho_{2J} & \ldots & 1 \end{bmatrix} \times \begin{bmatrix} \sigma_1 & 0 & \ldots & 0 \\ 0 & \sigma_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_J \end{bmatrix},$$

which may be written compactly as:

$$\boldsymbol{\Sigma} = D(\boldsymbol{\sigma})\boldsymbol{\rho}D(\boldsymbol{\sigma}), \tag{1}$$

where $D(\boldsymbol{\sigma})$ is a diagonal matrix with $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_J)$ on the diagonal and $\boldsymbol{\rho}$ is the correlation matrix. Priors on $\boldsymbol{\mu}$ may be set broadly making the model robust to changes in location.

A hierarchical model is appropriate when there are several repeated observations for each individual in each task. Each individual performs L_{ij} trials on each task, and the score on each trial is denoted $Y_{ij\ell}$, where, as before i indexes individual, $i = 1, \ldots, I$, j

indexes task, $j = 1, \ldots, J$, and ℓ indexes replicate, $\ell = 1, \ldots, L_{ij}$. The simplest model is

$$\begin{aligned} \text{Hierarchical Model:} \qquad & Y_{ij\ell} \mid \theta_{ij} \sim \mathrm{N}(\theta_{ij}, \tau_j^2), \\ & \boldsymbol{\theta}_i \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \end{aligned}$$

The above hierarchical model is comprised of two levels. The first a *data-level* specification that describes how scores on trials vary with people and tasks. Parameter θ_{ij} is the true score for the ith person in the jth task. The term τ_j^2 describes the trial-to-trial variability for the jth task. Although it is possible to expand this model to include conditions, covariates, and other distributional families, this simplified model is ideal for exploring the effects of priors in Bayesian analysis. The second level is the *individual-level* specification that describes how individual's performance covary across tasks. An individual's profile $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{iJ})'$ is a column vector of true scores, and the correlation among tasks, the target of study, is contained within $\boldsymbol{\Sigma}$ via Equation (1).

In Bayesian analysis, priors are needed on parameters. In the manifest model, priors are needed for mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$. In the hierarchical model, priors are needed for these two as well as for trial variances τ_j^2. In practice, prior on $\boldsymbol{\mu}$ and the collection of trial variances are not problematic as diffuse priors work well. The critical component is the variance matrix $\boldsymbol{\Sigma}$, especially in the hierarchical model. The goal then is to pick a good prior on $\boldsymbol{\Sigma}$ so that the collection of correlations may be faithfully recovered without undue influence.

In the next section, we review some choices of prior for $\boldsymbol{\Sigma}$. Before doing so, we note that in most multivariate studies of performance, researchers estimate correlations en route to further analysis such as factor modeling or structural-equation modeling. So, why study correlations themselves? Bollen (1989) notes that factor models and structural-equation models impose constraints on correlation matrices. In comparison, we are studying *unconstrained* priors—priors that do not restrict the pattern of covariance other than that

the matrix is a proper covariance matrix. Having good unconstrained priors allows researchers to explore possible patterns without unwarranted assumptions. These models then serve as a general model against which constrained models, that is specific factor model or structural-equation models, may be compared.

Priors on Covariance and Correlation

In this section, we provide three common choices of prior and explore how the constraints within them could affect invariance to scale and invariance to inclusion. We visualize the covariances using insights from Tokuda et al. (2025).

Inverse Wishart Prior. The Inverse Wishart distribution (IW) is a popular choice for a prior on covariance matrix $\boldsymbol{\Sigma}$ because it is conjugate for normally-distributed observations. When a covariance matrix follows an Inverse Wishart, we write:

$$\boldsymbol{\Sigma} \sim \text{Inverse Wishart}(\boldsymbol{S}, v),$$

where $\boldsymbol{S}$ and v are prior settings that must be chosen beforehand. Setting v is the degrees-of-freedom of the distribution. It may be set to $v = J + 1$ as a default, where J is the size of covariance matrix (the number of tasks or measures). The more critical setting is that of the scale matrix $\boldsymbol{S}$. This matrix may be diagonal, and the diagonal entries, set beforehand, are the expected value of variance. The question then is how these settings affect the posterior of correlation.

Figure 2 shows a series of visualizations of the prior for the default $v = J + 1$ and $\boldsymbol{S} = \boldsymbol{I}$, the identity matrix. Figure 2A shows the marginal prior on the pairwise correlations, that is, the off-diagonal elements in $\boldsymbol{\rho}$. These are uniformly distributed on the interval $[-1, 1]$, and this distribution holds regardless of the number of tasks J. Such a prior is highly desirable as no ranges of correlations are unduly favored. The top row addresses the invariance to scale. It shows the relationship between the prior on scale, or

standard deviation ($\sqrt{\Sigma_{jj}}$) and correlation. Figure 2B is a contour plot of prior joint distribution of a correlation coefficient and the standard deviation ($\sqrt{\Sigma_{jj}}$). Here we see an issue—there is a notable lack of independence. The two histograms on the top row (Figure 2C-D) highlight the dependencies. These plots are conditional prior densities on ρ. Figure 2C is conditional on small standard deviations and the overweighting of small-magnitude correlations is seen. Figure 2D is conditional on large standard deviations and the overweighting of large-magnitude correlations is seen. What do these dependencies mean? They imply that the prior is dependent on the specification of scale, and moreover, the analyst needs a reasonable sense of the variation. If the analyst sets the scale $\boldsymbol{S}$ too small, the data are concordant with relatively high variabilities and posteriors may be biased toward large-magnitude correlations. Conversely, if the analyst sets the scale $\boldsymbol{S}$ too large, the data are concordant with relatively low variabilities and posteriors may be biased toward low-magnitude correlations.

Figure 2E addresses invariance to inclusion. It shows the relationship between two correlation coefficients, or how the inclusion of one variable affects another. The contour plot reveals a lack of independence. Correlation coefficients tend to be similar in magnitude to each other, and there is less prior mass where one is large in magnitude and the other is small in magnitude. This relationship is visualized more concretely in the conditional prior distributions in Figure 2F-G. Analysts that include additional variables with large-magnitude correlations may bias posteriors toward larger correlations; those that include additional variables with small-magnitude correlation may bias posteriors in that direction.

The advantage of the Inverse Wishart prior is computational speed. The prior is conjugate with respect to the normal, hence conditional posterior distributions are also Inverse Wishart with updated parameters. Inverse Wishart prior is convenient to sample from, and MCMC chains run quickly and efficiently (as will be seen). The usual critique is that the scale specification is too informative and excludes small possible variance values

(Gelman & Hill, 2007; Huang & Wand, 2013). We are less concerned with this aspect here—our focus is squarely on the effect of the estimation of correlation.

Scaled Inverse Wishart Prior. The Scaled Inverse Wishart distribution (SIW), from Huang and Wand (2013), is relatively new prior for covariance. It is a bit misnamed for it is not a scaled version of the Inverse Wishart (which already has a scale setting). Instead, the distribution is formed by taking a continuous weighted mixture of Inverse Wishart distributions across all scales. The hope then is that this more diffuse form lessens any burden of specifying a single scale *a priori*. When a covariance matrix follows a Scaled Inverse Wishart, we write,

$$\boldsymbol{\Sigma} \sim \text{SIW}(v, \boldsymbol{s}),$$

where v is a degrees-of-freedom parameter and $\boldsymbol{s}$ is a vector of scales on standard deviations, $\Sigma_{jj}^{1/2}$. The Scaled Inverse Wishart is related to the Inverse Wishart as follows. If $\boldsymbol{\Sigma} \sim \text{SIW}(v, \boldsymbol{s})$, then,

$$\boldsymbol{\Sigma} \mid \alpha_1, \ldots, \alpha_J \sim \text{Inverse Wishart}\Big(v + J - 1,\, 2v\, D\Big(\tfrac{1}{\alpha_1}, \ldots, \tfrac{1}{\alpha_J}\Big)\Big), \quad \alpha_j \overset{\text{ind}}{\sim} \text{Inverse-Gamma}\Big(\tfrac{1}{2},\, \tfrac{1}{s_j^2}\Big),$$

where $D()$ denotes a diagonal matrix. The degrees-of-freedom parameter v plays a different role in this parameter than in the Inverse Wishart. In the Inverse Wishart, v references the number of variables, and $v = J + 1$ is a useful default corresponding to a uniform prior on all correlations. In the Scaled Inverse Wishart, the default value is $v = 2$ (see Huang and Wand, 2013). With this setting, the marginal priors on correlation coefficients are distributed uniformly on [-1,1], and this distribution holds for all J (see Figure 3A). The main question revolves around the role of $\boldsymbol{s}$, which are scale values on the standard. Figure 3B shows the joint distribution of variability and correlation for the Scaled Inverse Wishart prior for comparable settings to the Inverse Wishart. Here, there remains some dependencies, though to a lesser degree than with the Inverse Wishart. Figure 3C-D show the conditional prior distributions, and the presence of the dependence though to a lesser

degree may be seen. Figure 3E-G show, respectively, the joint and conditional distributions of correlation coefficients. The behavior here is the same as the Inverse Wishart—the prior tends to make correlation coefficients similar in magnitude and hence violate at least to some degree invariance to inclusion.

LKJ Prior. The LKJ prior (Lewandowski et al., 2009) is known as a least-informative prior. This prior makes use of Equation (1)—rather than placing priors directly on $\boldsymbol{\Sigma}$, priors are placed on $\boldsymbol{\sigma}$, the vector of standard deviations, and $\boldsymbol{\rho}$, the correlation matrix. These specifications are separate and decouples variance and correlation.

$$\sigma_j \sim \text{Half-t}(2, s_j), \quad j = 1, \ldots, J,$$

$$\boldsymbol{\rho} \sim \text{LKJ}(v),$$

The distribution labeled "Half-t" is the positive half of a scaled t distribution with 2 degrees of freedom and scale s_j. The collection $\boldsymbol{s} = (s_1, \ldots, s_J)$ serves as scale settings that must be set before analysis, and plays a similar role as that in the Scaled Inverse Wishart. The LKJ prior is a distribution specifically for correlation matrices, and the distribution depends on shape parameter v. The default setting is $v = 1$, and the marginal distribution on the correlation coefficients is shown in Figure 4A. Unlike the Inverse Wishart and Scaled Inverse Wishart, this distribution depends on the number of variables or tasks, J. When the distribution is bivariate, $J = 2$, the prior distribution on correlation is uniformly distributed on [-1,1]. However, as more tasks are added, the prior emphasizes smaller-magnitude correlations, which is a violation of invariance to inclusion.

Figures 4B-D show the relationship between variability and correlation, and there is no relationship by construction. This independence is visualized in the conditional histograms (Figure 4C-D). Likewise, there is no relationship between different pairwise correlations (Figure 4C-D). The LKJ violates invariance to inclusion by downweighting smaller values with more variables; the Inverse Wishart and Scaled inverse Wishart priors do so by making correlation values more similar to each other in magnitude.

Effect of Prior Specification For The Manifest Model

The discussed priors are broad defaults that are used repeatedly throughout the literature. Our expectation was that in the manifest case with little trial noise, they would each be excellent choices. To explore the effects of prior specification on the measurement of correlation, we used the anthropometric data from the U.S. Army (Army, 2014). In the data set, over 6000 soldiers provided about 100 different body measures. We examined the 4082 identified male soldiers, and correlated the height with weight for a random sample of $I = 200$ of them. To set the scale, we used the standard deviation of the full sample of 4082 as reference. For example, the standard deviation of height was 6.86 cm, and we used this number as a baseline. To assess robustness to misspecification, we took this value and manipulated it from 1/10th baseline to 10 times baseline, or in this case, from 0.69 cm to 68.6 cm. This is a huge range of variation for an empirical scientist. We estimated either the height or weight alone as a bivariate problem, or included another 8 anthropometric measurements as a 10-variable problem.

Table 1 shows the effect of prior for both the bivariate case ("Exclude" for exclude the other variables) or the 10-variate case ("Include" for include the other variables) for the range of scales. The entries show the point estimate and CIs of the correlation between height and weight. The Pearson sample correlation, as shown, does not vary for include and exclude cases. Overall, results from priors are fairly similar and match the sample-correlation results well. There are two relatively minor deviations. First, posterior estimates from the Inverse Wishart prior are a tad too low when the standard deviation scale is 10 times from baseline, and this underestimation reflects the prior dependencies show in Figure 2B-D. Second, the posterior estimates from the LKJ are a tad too low when additional variables are included. This underestimation reflects the dependence on the marginal prior on correlation with increasing numbers of variables as shown in Figure 4A.

The above application shows that all three prior classes do well in the manifest case.

There is sufficient prior mass in all cases that data with as few as 200 cases are more than sufficient to dominate the prior. It matters little here which of the three priors is used. Stated alternatively, there is little reason to use Bayesian analysis at all for the manifest case; the sample correlation along with Fisher's z-transform confidence intervals provides the same summary.

The manifest case is only appropriate when there is little measurement noise. Tolerably low measurement noise, such as in weight measures, is the exception and not the rule in experimental psychology. In fact, in all cases we know of, there is a high degree of measurement noise. And in all these cases, the Pearson correlation coefficient is inadvisable because of a large degree of attenuation. Hence, the critical question is how these priors perform when the data have measurement noise and this measurement noise is modeled in a hierarchical context.

Simulation Study 1: Two Tasks. Calibrated Scale Settings

Simulations in this paper follow a common form. First, to generate data, true individual parameters ($\boldsymbol{\theta}_i$) are sampled from a multivariate normal parent distribution. For two tasks each individual has two true values with one for each task, and there is a single true population correlation parameter ρ. The value of ρ is set as ground truth and the estimate of which serves as the target of analysis. Here are the settings for the parent distribution on $\boldsymbol{\theta}_i$:

$$\boldsymbol{\theta}_i \sim \mathrm{N}_2\left(\begin{pmatrix} .5 \\ .5 \end{pmatrix}, \begin{pmatrix} .1^2 & \rho \times .1^2 \\ \rho \times .1^2 & .1^2 \end{pmatrix}\right),$$

In all the simulations reported here, there were 200 synthetic individuals, that is, $I = 200$. Trial level data were sampled from true individual values $\boldsymbol{\theta}_i$ as $Y_{ij\ell} \sim \mathrm{N}(\theta_{ij}, \tau_j^2)$. To speed simulations and draw sharp contrasts between estimation approaches, we set a low number of trials, $L = 20$, throughout. True correlations were set to $\rho = .3$, $\rho = .5$, and

$\rho = .7$ to cover a variety of cases; trial variance τ_j^2 was set either to a low-noise value of $.2^2$ or to a high noise value of $.5^2$, and these values held for all tasks.

In each simulation, trial-level data were generated from the above steps and analyzed with hierarchical models that varied only in the prior specification on $\mathbf{\Sigma}$. One specification was the Inverse Wishart prior with shape set to the default setting of $v = 3$ and scale s^2 is set to $.1^2$, which matches the true between-individual variation. Another was the Scaled Inverse Wishart prior with default shape, $v = 2$, and scale $s = .1$, which matches true between-individual standard deviation. The final specification is the LKJ with prior shape the default of $v = 1$ and prior scale $s = .1$. The Inverse Wishart and Scaled Inverse Wishart prior specifications do not have an explicit parameter ρ, but it is straightforward to compute a value of ρ on each iteration of the MCMC chain using Equation (1). The LKJ has lower triangle matrix outputs on each iteration; the cross product of this matrix yields the posterior correlation matrix.

A simulation consisted of 100 replicate runs. New data were generated on each replicate, and the same data were submitted to analysis by each model. The Inverse Wishart model and the Scaled Inverse Wishart model were implemented in JAGS (Denwood, 2016; Plummer, 2003) using a Gibbs sampling algorithm. For each model, 1000 burn-in iterations were followed by 3000 iterations retained for posterior computation. Because JAGS does not support the LKJ prior, the LKJ model was implemented in RStan (Carpenter et al., 2017) using the NUTS algorithm. Again, there were 1000 burn-in iterations followed by 3,000 retained iterations. Mixing was satisfactory in all cases as assessed by visual inspection of the chains, inspection of autocorrelation functions, and computation of effective sample sizes (ESS, Gong & Flegal, 2016). In fact, ESS values in this report always exceeded 2000 effective samples from the 3000 retained samples indicating good mixing throughout. We evaluate each model's efficiency as ESS per second.

The results of the simulation are shown in Figure 5. Consider the five left-most box

plots colored in red, purple, green, yellow, and blue. The true correlation here is $\rho = .3$, as indicated above and as denoted by the dashed line. For these five box plots, there is a low degree of trial noise as indicated by $\tau_j = .2$. The red box plot is the distribution of sample correlations among the 200 individual true values across the 100 runs. The variation serves as what is expected from finite samples of 200 people if there were infinite numbers of trials. We use this as a best-case outcome as individual true-values are known. The remaining four boxplots are based on the trial-level data. The purple box plot is the distribution of correlations from the averaging method where averages are tabulated across $L = 20$ trials. As expected, this distribution is biased slightly low. This bias is the attenuation of correlation from measurement noise. The bias is not that severe because there is not a high degree of trial noise. The green, yellow, blue box plots are distributions of the posterior mean of the correlation parameter across the runs for the three different priors. These are similar to the correlations among true values indicating that the hierarchical model with any of the three priors does an excellent job of accounting for trial noise. The results for the high-trial-noise case follows in the next four bars with the subtitle $\tau_j = .5$. The following trends hold: (i) there is much attenuation for averaging (purple box plot); (ii) correlation recovery is somewhat more variable for the hierarchical models; and (iii) the performance does not vary appreciably across different priors. The remaining panels show the results with higher true correlations. The results are largely the same with the exception that attenuation from averaging becomes more pronounced as the true correlation increases. Overall, there is little to distinguish the performance of Inverse Wishart, Scaled Inverse Wishart, and LKJ priors in these simulations.

One way of characterizing the results is to compute the RMSE for each estimate across the 100 runs in the simulation. Table 2 shows these RMSE values and the pattern is clear. For low noise and true correlation of .3, the hierarchical model with any of the three prior is about as accurate as knowing the individuals' true values. There is loss in the high noise case, but the performance is about equivalent.

One of the most useful benefits of Bayesian hierarchical analysis is that posterior distributions themselves provide estimates on how well parameters are estimated. The relevant quantity is the *posterior credible interval* (Kruschke, 2014), which is somewhat analogous to a confidence interval (cf., Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2016). The credible intervals for the first 50 runs in the high-noise case is shown in Figure 6. Consider the upper left-hand plot which is for a true population correlation of $\rho = .3$. There are 50 horizontal lines for the 50 runs. In this case, each line is a posterior credible interval for the correlation coefficient for the run, and the dot is the posterior mean. For $\rho = 0.3$ and $\rho = 0.7$, for all four priors the vast majority (96%) of credible intervals contain the true value. For $\rho = 0.5$, this proportion was slightly lower (88%) but still reasonably high. Overall, the credible intervals appear to provide satisfactory coverage. The simulations show encouraging results with all three priors.

One of the primary advantages of the Inverse Wishart and Scaled Inverse Wishart priors is their computational convenience. The Inverse Wishart is conjugate with the normal distribution, which makes sampling relatively inexpensive. This computational efficiency enables the faster execution: In Simulation 1, LKJ prior's sampling efficiency was approximately ten times lower than Inverse Wishart's and Scaled Inverse Wishart prior's (see Figure 7). This multiple in efficiency only grows with the size of the covariance matrix.

Although this computational-time gap is substantial in relative terms, it should not be overstated. For a single dataset, analyses with any of these priors typically complete within a few minutes, and the difference is mainly relevant in simulation studies. More importantly, the LKJ prior offers critical advantages in robustness which motivates our recommendation to consider it for analyzing real-world data.

Simulation Study 2: Four Tasks with Calibrated Priors

The results in Simulation Study 1 show the benefits of hierarchical modeling and the relative equivalence of the three priors. The study, however, covered just two tasks and a single correlation coefficient. We performed Simulation Study 2 with 4 tasks. The four-task simulations follow almost all the same settings in Simulation Study 1 with the exception of the ground truth correlation matrices. The true correlations for the four-task versions are shown in Figure 8A. The four-task true correlation matrix followed a two-factor pattern with set values of $\rho_{12} = 0.3, \rho_{13} = 0.5, \rho_{14} = 0.7$.

We ran the same simulation study with these new correlations matrices, and assessed how well correlations may be recovered with the averaging method and three priors. The results for the four-task setup are shown in Figure 9 and Table 2. These results are presented in the same format as those for Simulation 1, and the results are quite similar. As before, correlations from trial averaging have noticeable attenuation; there is much improvement with the hierarchical models, and performance is similar with four priors.

Simulation Study 3: Robustness to scale settings

Simulation Study 1 and 2 are best-case scenarios in that the prior setting on variability matched the variance used in data generation. We cannot be so lucky in application because we do not know the true scale. As discussed previously, the Inverse Wishart exhibits dependency between correlation and variability meaning that poor settings of scale may bias the posterior of the correlation coefficients. And the Scaled Inverse Wishart has the same dependency though not to as large a degree. We saw only a minor effect of these posterior settings for the manifest model as the data for $I = 200$ observations were sufficient in number to dominate the prior. Restated, all three priors demonstrate a robustness to scale setting in that context. Yet, hierarchical models are more complex and more heavily parameterized. Consequently, the influence of the prior is

often much greater.

How does the setting of the scale of variance, s^2 affect posterior estimation of correlation in a hierarchical model with measurement noise? This question is addressed by Simulation Study 3 where s (or s^2) is manipulated across a few levels. To match the data-generation process in the previous simulations, we set $s = .1$ (or $s^2 = .01$). Here, we take values that are either double or half this value, as well as values that are 5 times or 1/5th this value, that is $s = .02, .05, .1, .2, .5$. We reran Simulation Study 2 with four tasks and with intermediate trial noise of $\tau_j^2 = .4^2$.

Figure 10 shows sets of smoothed densities. The densities are **not** posterior densities of the correlation coefficient. They are the distribution of the posterior mean across the 100 replicates. Consider the dotted line in the leftmost upper panel. This panel is for the Inverse Wishart for true correlations of .3. That dotted line is the sample correlation among true individual values, and it is not perturbed by trial noise. It corresponds to the red box plots in Figures 6 and 9, and it serves as a best case. The solid lines correspond to different settings of s for the Inverse Wishart. As can be seen, there is much dependence on s. This behavior is not desirable. The lower row shows the same for true correlations of .7, and the trend here is even more obvious. The second column shows plots for Scaled Inverse Wishart prior, it is less sensitive to prior settings compared with Inverse Wishart, however, when s is 5 times smaller than the true value, it has a tendency to over-estimate the correlation.

The third and fourth columns show similar plots for the LKJ prior. The third column is for changes in s, and the behavior here may be directly compared to that for the Inverse Wishart prior. Here, we see the desired robustness of posterior correlation to prior variance scale settings. Table 3 shows the RMSE values for these comparable prior settings—they favor the LKJ when the setting is far from the data-generating value. Because the LKJ fared favorably, we also explored the robustness to settings of v, the prior setting on correlations. The values of v were manipulated through $v = .25, .5, 1, 2, 4$, and the results

for true low correlation values exhibit the same robustness. The only scenario where there was an effect of prior settings is for higher true correlations as shown in the lower panel on the right.

Simulation Study 4: Robustness to Inclusion

None of the three priors are invariant to inclusion. For the Inverse Wishart and Scaled Inverse Wishart priors there is increased prior density for correlation coefficients that are similar in magnitude (see Figures 2E-G and 3E-G); for the LKJ prior there is decreasing density for extreme magnitude correlations as the number of variables is increased (see Figure 4A). Do these dependencies affect posterior estimates in reasonably-sized data?

To explore the issue, we constructed ground truths with correlations shown in Figure 8B. The key correlation is that between Tasks 1 and 2, which was set to the high value of .8. In one case, the data from Tasks 1 and 2 were submitted to a bivariate analysis much as in Simulation 1. In a second case, data form all 8 tasks were submitted to an 8-dimensional multivariate normal model, and the focus was on the estimate in the correlation between Task 1 and 2 with the six other tasks included and simultaneously analyzed. The trial noise was moderate at$\tau = .4$ and prior scales on variability were set to .1ˆ2 (Inverse Wishart prior) or .1 (Scaled Inverse Wishart and LKJ priors) as in Simulation 1.

Figure 11 shows the results as scatter plots. First, focus on the small black dots on the diagonal. These are the correlation between true individual values for Tasks 1 and Task 2, and these are the same whether 2 tasks or 8 tasks are analyzed as sample correlation is invariant to inclusion. There are 100 of these points corresponding to the 100 simulation runs. The colored larger points are posterior mean estimates for the correlation between Task 1 and Task 2. The x-axis is for the two tasks alone in a bivariate analysis; the y-axis is the same correlation when the other six tasks are included. For the Inverse Wishart, the estimates are well-centered whether the two tasks were considered in isolation or in the

8-task context. For the SIW and LKJ, the prior dependency is more clear. There is a tendency toward lower posterior means overall, and this effect is accentuated in the 8 task case as expected. As an aside, we had expected similar behavior for the Inverse Wishart and Scaled Inverse Wishart, so we reran the Inverse Wishart simulation with different seeds to check. We obtained the same patterns. We are not sure why the Inverse Wishart fared so well but are reasonably confident this finding is not a statistical fluke.

RMSE values for Simulation 4 are reported in Table 4. One critical comparison is between the the estimates of .8 true-valued correlation coefficient for the two-task vs. eight-task context. The Inverse Wishart prior is slightly better in the 8-task case than the 2-task case, and this result replicated on an additional simulation with different seeds. The RMSE is slightly worse in 8-task case than the 2-task case for the SIW and LKJ, and given the systematic pattern of underestimation in Figure 11, this result is more interpretable. Table 4 shows that while the estimate of the high correlation value between Task and Task 2 was accurate for the Inverse Wishart, the estimate of the low correlations between the other tasks was less accurate. This accuracy reflects an inflation of the low correlation values for this prior.

Simulation Study 5: Robustness To Distribution of Variability for LKJ Prior

In the Inverse Wishart and Scaled Inverse Wishart priors, the analyst places a prior on covariance which encompasses both the variance and correlation. The resulting marginal prior on variance is inverse gamma for the Inverse Wishart prior, and the marginal marginal prior on standard deviation is a half-t for the Scaled Inverse Wishart. The LKJ prior is qualitatively different in that separate, priors are placed on variability and correlation. The analyst is free to choose any form for the prior on standard deviation they wish. We chose a half-t distribution. Does this choice matter?

One reason to suspect this choice does not matter is that there is prior independence

between the variability and correlation. Hence, any dependency comes from the likelihood itself. Nonetheless, to explore the possibility we ran a brief simulation with seven different distributions on standard deviation in the LKJ prior. The distributions are shown in Figure 12A, and the center of these distribution were at a value around $s = .1$.[1] Although the priors differ in shape and tail behavior, they all share a common reference point—the center—allowing a reasonable comparison of their influence on the posterior correlation distribution.

The simulation setup was the four task setup in Simulation 2 with the correlations shown in Figure 8A which has correlation values of .3, .5, and .7. The upper right panel shows the distribution of posterior means as a smoothed density for the 100 replicates for each prior. The true value here is .3. As can be seen, the choice of prior form matters little perhaps with the exception of the lognormal distribution. The The lower plot shows the case for the true value of .7, and the same result holds. Overall, it seems to matter little which broad distribution is used for the standard deviation in the LKJ prior so long as it has broad coverage and tails at least as fat as an exponential.

Simulation 6: Contrasts

The preceding simulations are designed to provide insight into the performance of the three priors and to be computationally convenient. For these purposes, we used simplified models that do not account for experimental conditions and contrasts. The generalization

[1] Note that some distributions do not have means.

of the hierarchical model setup to is straightforward. Consider for simplicity tasks comprised of two contrasting conditions. Examples include Stroop or flanker tasks (congruent vs. incongruent), priming tasks (primed vs. not primed), or task-switching (repeated vs. switched). The key in all these tasks is that the measure of interest is a contrast, say the contrast between incongruent and congruent or between repeated and switched conditions. The following model centers this contrast. An observation is denoted $Y_{ijk\ell}$ where i denotes the individual, j denotes the task, $k = 1, 2$ denotes the condition, and ℓ denotes the replicate trial. The data model is

$$Y_ijk\ell \sim \mathrm{N}(\alpha_{ij} + x_k\theta_{ij}, \tau_j^2),$$

where α_{ij} is the overall mean or intercept for the ith person in the jth task, $x_k = -.5, .5$ is the contrast code for the kth condition, θ_{ij} is the contrast or slope for the ith person in the jth task, and τ_j^2 is the variability across replicate trials in the jth task. The target of interest is θ_{ij}, the contrast, and the previous individual latent multivariate normal model, $\boldsymbol{\theta}_i \sim \mathrm{N}_J(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is appropriate. And as before, priors are needed on $\boldsymbol{\Sigma}$.

The main difficulty with the Inverse Wishart prior is that the prior choice of scale on variability affects the posterior distribution of correlations (see Figure 10). This problem is less salient for contrast parameters because researchers often have appropriate *a priori* information. Here, we argue that researchers often know at least ot a rough approximation the variability of contrast parameters across individuals. The argument starts with on observation about the direction of individual effects: Rouder and Haaf (2021) argue that most experimental manipulations yield individual true contrasts that are only in one direction. Take, for example, the Stroop effect where color naming is quicker on average to congruent than incongruent stimuli. The argument is that each individual has a true Stroop effect in the same direction, and that nobody truly names the color of incongruent items faster than congruent ones. When all individuals have true effects in the same

direction, Haaf and Rouder (2017) call it a *everybody does* situation. This *everyone does* situation places substantial limits on the scale of variability. Suppose a contrast has a mean of 60 ms, which is typical of Stroop, flanker, and priming effects. The variability in this effect could not be too large else a sizable proportion of individuals would have true effects in the opposing direction. For example, if the mean effect was 60 ms, the standard deviation may not be larger than 30 ms as any more would imply more than 2.5% of the population have the opposite effect. Hence, just by knowing the mean effect, we have good *a priori* information about the variability across individuals. The question is whether posterior distributions of correlations vary appreciably with this knowledge in the Inverse Wishart setup.

To answer this question, we ran a simulated Stroop experiment with two conditions in each of two tasks—say a color Stroop task and a numerical Stroop task (MacLeod, 1991). The conditions of the simulations were designed to be typical of several Stroop data sets we have examined (see Haaf, Hoffstadt, & Lesche, 2024). As before there were 200 synthetic individuals. Each ran 150 trials in each task and each condition. Individuals varied in their true overall speed: $\alpha_{ij} \overset{iid}{\sim} \text{Normal}(600, 100^2)$. They also varied in their true Stroop effect as follows: These Stroop effects were drawn from a bivariate normal with a mean vector of 60 ms in each task. The standard deviation in each was 25 ms, and the correlation across the two tasks was .5. Trial noise, τ_j was to 175 ms.

To test how the prior choice of scale on variability affects the posterior distribution of correlations, we reran the analysis with several choices of s^2. Our guiding principle was that s^2 should reflect a range of reasonable prior opinion but no more. Here, for a 60 ms mean, we though value of (40 ms)^2 was very large as it implies 7% of individual truly have a reverse Stroop effect where they truly respond quicker to incongruent than congruent items. Likewise, a value of (15 ms)^2, the mean being 4-times the standard deviation, strikes us as very small. Hence, the question is how varying values of s^2 in this range affects posterior correlation estimates. The resulting posterior correlation distributions are plotted

as box plot in Figure 13, and the effect of prior scale across this reasonable range is modest at best. In conclusion, the variability issue is less salient for contrast parameters especially when researchers have a good idea what the mean effect may be.

Discussion

Hierarchical models provide a superior approach to estimating correlations over sample correlations computed from sample means. These models treat trial noise and variation across individuals separately. As a result, correlation estimates are disattenuated and uncertainty reflects both variation across individuals and trials. Sample correlations from sample means, in contrast, have neither of these sanguine properties—they are dramatically attenuated and the associated confidence intervals are too small as they do not reflect trial variability. **We recommend that hierarchical models be always used in estimating correlations from experimental data where trials are nested within individuals.** Analyses without these models run a substantial risk of misinterpretation.

The question then is what prior specification is best for estimating correlation coefficients? In this paper, we have compared the Inverse Wishart, Scaled Inverse Wishart, and LKJ prior specifications in manifest and hierarchical contexts. The recommendation is that the LKJ prior is better for general use with empirical data sets. The reason for this recommendation is that the LKJ prior is less sensitive to prior settings than the others. Our recommendation reflects a visualization of the priors as well as simulations of robustness.

A comparison of the visualization of the priors (Figures 2-4) shows that the Inverse Wishart and Scaled Inverse Wishart priors have at least some sensitivity to prior scale setting. Furthermore, each is sensitive to the inclusion of additional variables inasmuch as correlation coefficients are shrunk to be more similar in magnitude. The LKJ prior on correlation is robust to variance scale settings, but the marginal prior on correlation becomes more peaked at small magnitude values as more variables are included in the

analysis.

The performance of the priors in simulation reflect in part biases evident in the visualizations of the priors. The dependence in the Inverse Wishart of correlation on prior scale manifests itself in simulation—we draw readers' attention to the difficult results for the Inverse Wishart models in Figure 10. Although the LKJ prior is a safer choice, it is important not to overstate the costs. The scale of variance, the critical prior setting, is not particularly mysterious or unknown. For many applications, especially those with contrasts, the range of reasonable prior scale may indeed be small.

The choice of the LKJ prior is perhaps the safest. There are two considerations analysts should be aware of in making this choice: The first is that the LKJ marginal prior is a function of the number of variables included in an analysis. This makes the choice of how many variables to include salient. Including more variables lowers in magnitude the posterior estimate of each though the degree of this biasing effect is not overwhelming. The second is that the LKJ prior is computationally slow and does not scale well to higher dimensional models. For example, for eight variables, the Inverse Wishart and Scaled Inverse Wishart models ran at about 40 samples per second; the LKJ prior ran at 2 sample per second. That factor of 20 may be the difference between 1 day and 20 days of run time. Moreover, this factor increases with increasing numbers of variables, so it is entirely possible to wait weeks and months for an LKJ prior analysis. If the computational demands of the LKJ are difficult to meet, ee recommend the Scaled Inverse Wishart prior over the Inverse Wishart as a default alternative. The Scaled Inverse Wishart prior is far less sensitive to prior scale while providing for just as rapid computations.

Lastly, we remind readers that simulation results are just that. We have simulated a small sliver of possible cases that reflect our best judgment of the needs of the community. Perhaps the thorniest remaining issues involve when to include variables in analyses. It is our sense that there is more work to be done in understanding the statistical consequences

of such choices.

Declarations

Funding: JNR was supported by ONR N00014-23-1-2792.

Conflict of Interest: None.

Ethical Approval: NA

Consent to Participate NA

Consent for Publication: NA

Data Availability. No human data are collected in conjunction with this manuscript.

Code Availability. JAGS and Stan code for the models may be found at https://osf.io/qe9ts.

Author Contributions. YS developed models, ran the simulations, and constructed figures. JR conceived the project. Both authors jointly wrote the manuscript.

References

Army. (2014). *2012 Anthropometric Survey of U.S. Army Personnel: Methods and Summary Statistics* (No. ADA611869). Army Natick Soldier Research and Engineering Center. Retrieved from https://apps.dtic.mil/sti/citations/ADA611869

Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Bettencourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*.

Denwood, M. J. (2016). Runjags: An R package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS. *Journal of Statistical Software*, *71*, 1–25.

Fisher, R. A. (1921). On the "Probable Error" of a Coefficient of Correlation Deduced from a Small Sample. *Metron*, *1*, 3–32.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis (2nd edition)*. London: Chapman and Hall.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press. Retrieved from https://books.google.com/books?hl=en&lr=&id=lV3DIdV0F9AC&oi=fnd&pg=PR17&dq=gelman+hill+2007&ots=6nhGI8NyQ4&sig=c4k2j5JxOlu2VUHZEsFOV04wwY0

Gong, L., & Flegal, J. M. (2016). A Practical Sequential Stopping Rule for High-Dimensional Markov Chain Monte Carlo. *Journal of Computational and Graphical Statistics*, *25*(3), 684–700. doi:10.1080/10618600.2015.1044092

Haaf, J. M., Hoffstadt, M., & Lesche, S. (2024). Attentional Control Data Collection: A Resource for Efficient Data Reuse. doi:10.31234/osf.io/4evy6

Haaf, J. M., & Rouder, J. N. (2017). Developing Constraint in Bayesian Mixed Models. *Psychological Methods*, *22*(4), 779–798.

Haines, N., Kvam, P. D., Irving, L., Smith, C. T., Beauchaine, T. P., Pitt, M. A., ...

Turner, B. M. (2025). A tutorial on using generative models to advance psychological science: Lessons from the reliability paradox. *Psychological Methods*. doi:10.1037/met0000674

Huang, A., & Wand, M. P. (2013). Simple Marginally Noninformative Prior Distributions for Covariance Matrices. *Bayesian Analysis*, *8*(2), 439–452. doi:10.1214/13-BA815

Kruschke, J. K. (2014). *Doing Bayesian data analysis, 2nd edition: A tutorial with R, JAGS, and Stan*. Waltham, MA: Academic Press.

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*(9), 1989–2001.

Liu, H., Zhang, Z., & Grimm, K. J. (2016). Comparison of Inverse Wishart and Separation-Strategy Priors for Bayesian Estimation of Covariance Parameter Matrix in Growth Curve Analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(3), 354–367. doi:10.1080/10705511.2015.1057285

MacLeod, C. (1991). Half a Century of Research on the Stroop Effect: An Integrative Review. *Psychological Bulletin*, *109*, 163–203.

Matzke, D., Ly, A., Selker, R., Weeda, W. D., Scheibehenne, B., Lee, M. D., & Wagenmakers, E.-J. (2017). Bayesian inference for correlations in the presence of measurement error and estimation uncertainty. *Collabra: Psychology*, *3*(1).

McCrae, R. R., & Costa Jr, P. T. (1997). Personality trait structure as a human universal. *American Psychologist*, *52*(5), 509.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology*, *41*(1), 49–100.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*,

23(1), 103–123.

O'Hagan, A., & Forster, J. J. (2004). *Kendall's advanced theory of statistics, volume 2B: Bayesian inference* (Vol. 2). Arnold.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonable be supposed to have arisen from random sampling. *Philosophical Magazine*, *50*, 157–175.

Plummer, M. (2003). JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*.

Rouder, J. N., Chavez de la Peña, A. F., Mehrvarz, M., & Vandekerckhove, J. (2023, December 21). *On Cronbach's merger: Why experiments may not be suitable for measuring individual differences*. doi:10.31234/osf.io/8ktn6

Rouder, J. N., & Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic Bulletin and Review*, *26*(2), 452–467. Retrieved from https://doi.org/10.3758/s13423-018-1558-y

Rouder, J. N., & Haaf, J. M. (2021). Are There Reliable Qualitative Individual Difference in Cognition? *Journal of Cognition*, *4*(1).

Rouder, J. N., & Lu, J. (2005). An Introduction to Bayesian Hierarchical Models with an Application in the Theory of Signal Detection. *Psychonomic Bulletin and Review*, *12*, 573–604.

Rouder, J. N., Lu, J., Sun, D., Speckman, P. L., Morey, R. D., & Naveh-Benjamin, M. (2007). Signal Detection Models with Random Participant and Item Effects. *Psychometrika*, *72*, 621–642.

Rouder, J. N., & Mehrvarz, M. (2024). Hierarchical-Model Insights for Planning and Interpreting Individual-Difference Studies of Cognitive Abilities. *Current Directions in Psychological Science*, *33*(2), 128–135. doi:10.1177/09637214231220923

Rouder, J. N., & Province, J. M. (2019). Bayesian Hierarchical Models in Psychological

Science: A Tutorial. *New Methods in Cognitive Psychology*, 32–66.

Schuurman, N. K., Grasman, R. P. P. P., & Hamaker, E. L. (2016). A Comparison of Inverse-Wishart Prior Specifications for Covariance Matrices in Multilevel Autoregressive Models. *Multivariate Behavioral Research*, *51*(2–3), 185–206. doi:10.1080/00273171.2015.1065398

Spearman, C. (1904). 'General intelligence,' objectively determined and measured. *The American Journal of Psychology*, *15*(2), 201–293. doi:10.2307/1412107

Tokuda, T., Goodrich, B., Mechelen, I. V., Gelman, A., & Tuerlinckx, F. (2025). Visualizing distributions of covariance matrices. *Journal of Data Science, Statistics, and Visualisation*, *5*(7). doi:10.52933/jdssv.v5i7.132

Table 1
Estimates of correlation under Pearson, IW, SIW, and LKJ priors across different scale multipliers.

		Exclude		Include	
Scale	Method	Mean	95% CI	Mean	95% CI
0.1x	Pearson	0.50	[0.38, 0.59]	0.50	[0.38, 0.59]
	IW	0.50	[0.39, 0.59]	0.50	[0.39, 0.60]
	SIW	0.49	[0.39, 0.60]	0.50	[0.39, 0.59]
	LKJ	0.49	[0.38, 0.59]	0.45	[0.35, 0.54]
1x	Pearson	0.50	[0.38, 0.59]	0.50	[0.38, 0.59]
	IW	0.49	[0.38, 0.59]	0.50	[0.39, 0.59]
	SIW	0.49	[0.38, 0.59]	0.50	[0.39, 0.59]
	LKJ	0.49	[0.38, 0.59]	0.45	[0.36, 0.54]
10x	Pearson	0.50	[0.38, 0.59]	0.50	[0.38, 0.59]
	IW	0.47	[0.36, 0.58]	0.48	[0.36, 0.58]
	SIW	0.49	[0.38, 0.59]	0.50	[0.39, 0.59]
	LKJ	0.49	[0.38, 0.60]	0.46	[0.35, 0.55]

Table 2
RMSE with Calibrated Priors for Individual Variation

		$\tau_j = 0.2$			$\tau_j = 0.5$		
Case	True Correlation	IW	SIW	LKJ	IW	SIW	LKJ
2 Tasks							
	0.3	0.043	0.043	0.042	0.168	0.153	0.154
	0.5	0.042	0.041	0.041	0.162	0.164	0.17
	0.7	0.034	0.036	0.036	0.079	0.103	0.105
4 Tasks							
	0.3	0.047	0.044	0.048	0.192	0.148	0.151
	0.5	0.045	0.047	0.047	0.119	0.116	0.121
	0.7	0.037	0.048	0.044	0.081	0.121	0.107

Table 3
RMSE without Calibrated Priors for Individual Variation

		s				
True Correlation	Prior	0.02	0.05	0.10	0.20	0.50
0.3						
	IW	0.186	0.147	0.119	0.094	0.115
	SIW	0.152	0.121	0.108	0.105	0.105
	LKJ	0.117	0.115	0.116	0.115	0.116
0.5						
	IW	0.186	0.138	0.105	0.103	0.206
	SIW	0.139	0.108	0.103	0.105	0.105
	LKJ	0.104	0.101	0.101	0.101	0.1
0.7						
	IW	0.153	0.109	0.087	0.144	0.304
	SIW	0.113	0.093	0.109	0.117	0.121
	LKJ	0.09	0.091	0.089	0.092	0.094

Table 4
RMSE in 8-tasks inclusion case.

Case	True Correlation	IW	SIW	LKJ
2 Tasks				
	0.8	0.064	0.073	0.08
8 Tasks				
	0.2	0.157	0.137	0.102
	0.8	0.056	0.093	0.097

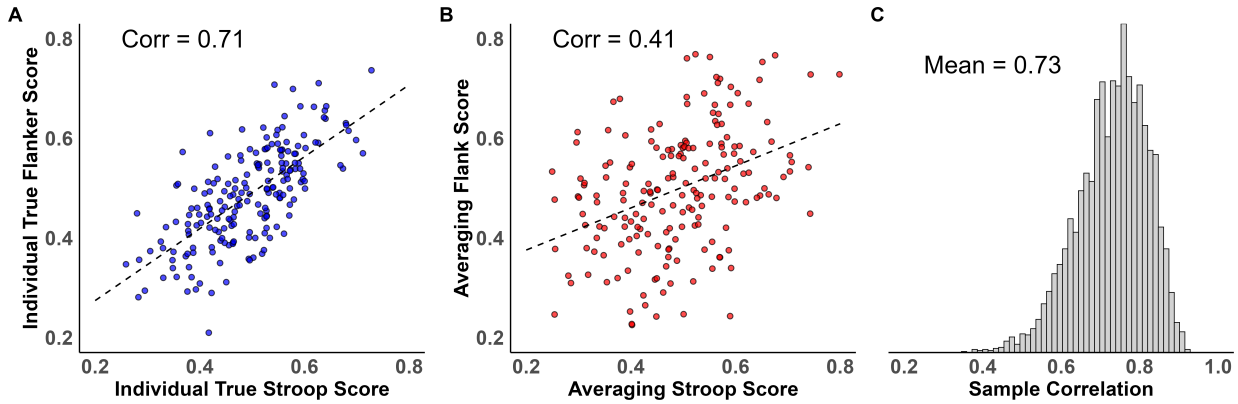

Figure 1. The effects of trial variability on correlation estimates. Synthetic individuals were sampled from a bivariate distribution with a true population correlation of .7. A. Scatter plot among individual true values for 200 individuals. the correlation is .73 in value which is close to the true value. B. True values from individuals are not known and must be estimated from trial data by averaging. If the trials are too few or the trials are excessively noise, then the averaged individuals' scores themselves are randomly perturbed from true values. This perturbation systematically attenuates the observed correlation among the tasks. C. Hierarchical models provide separate estimates of trial noise and covariance across individuals in tasks. By separately modeling different sources of variation, the posterior estimates of correlation are disattenuated; moreover, uncertainty in the posterior reflects limitations from the finite samples of participants and trials.

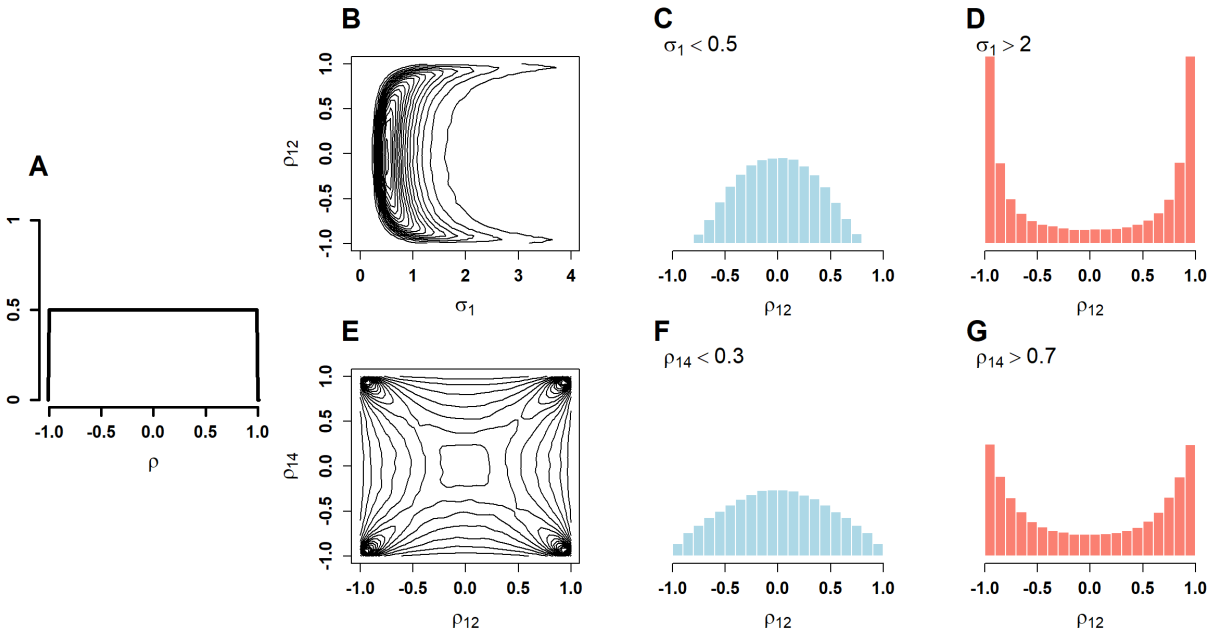

Figure 2. Visualization of the Inverse Wishart prior distribution. A. Marignal prior on correlation (pairwise correlation coefficients). B. Joint prior on standard deviation and correlation. C-D Priors on correlation conditioned on low and high variability, respectively. E. Joint prior on two correlation coefficients. F-G. Prior on one correlation coefficient conditional on low and high values of another, respectively.

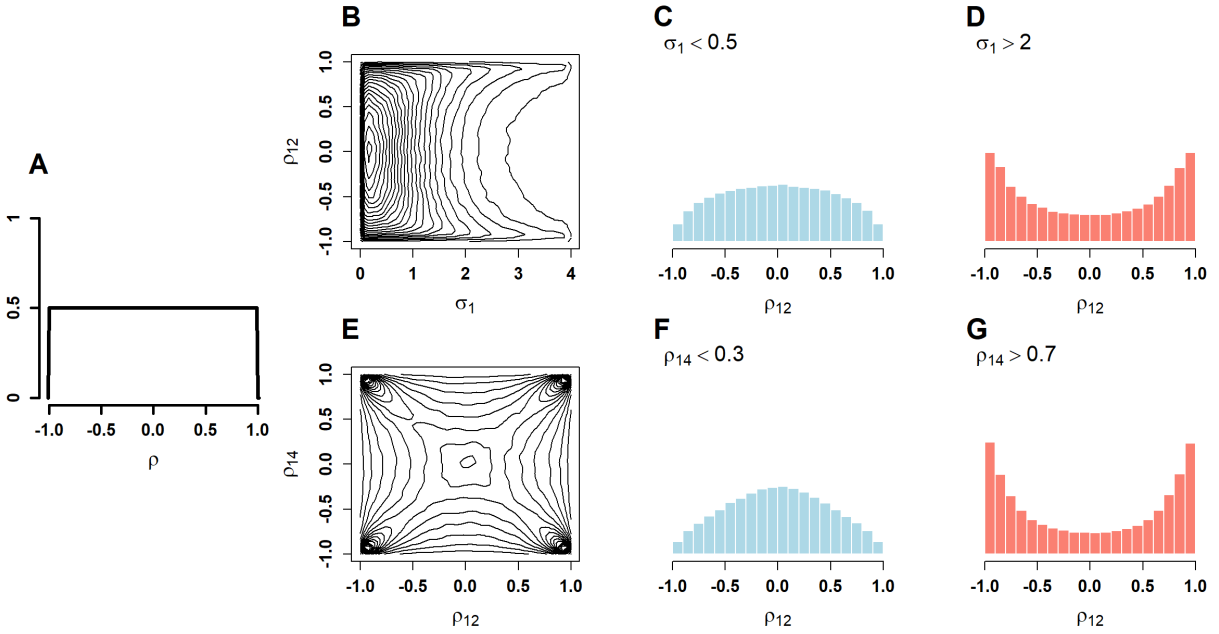

Figure 3. Visualization of the Scaled Inverse Wishart prior distribution. A. Marignal prior on correlation (pairwise correlation coefficients). B. Joint prior on standard deviation and correlation. C-D. Priors on correlation conditioned on low and high variability, respectively. E. Joint prior on two correlation coefficients. F-G. Prior on one correlation coefficient conditional on low and high values of another, respectively.

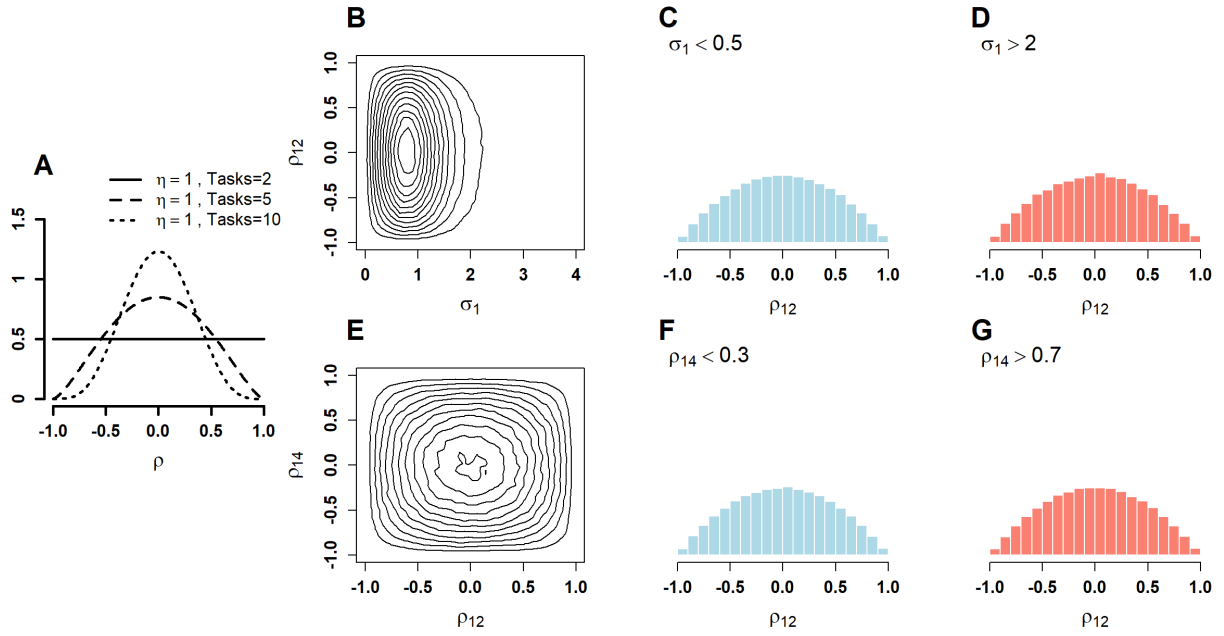

Figure 4. Visualization of the LKJ distribution. A. Marignal prior on correlation (pairwise correlation coefficients). B. Joint prior on standard deviation and correlation. C-D Priors on correlation conditioned on low and high variability, respectively. E. Joint prior on two correlation coefficients. F-G. Prior on one correlation coefficient conditional on low and high values of another, respectively.

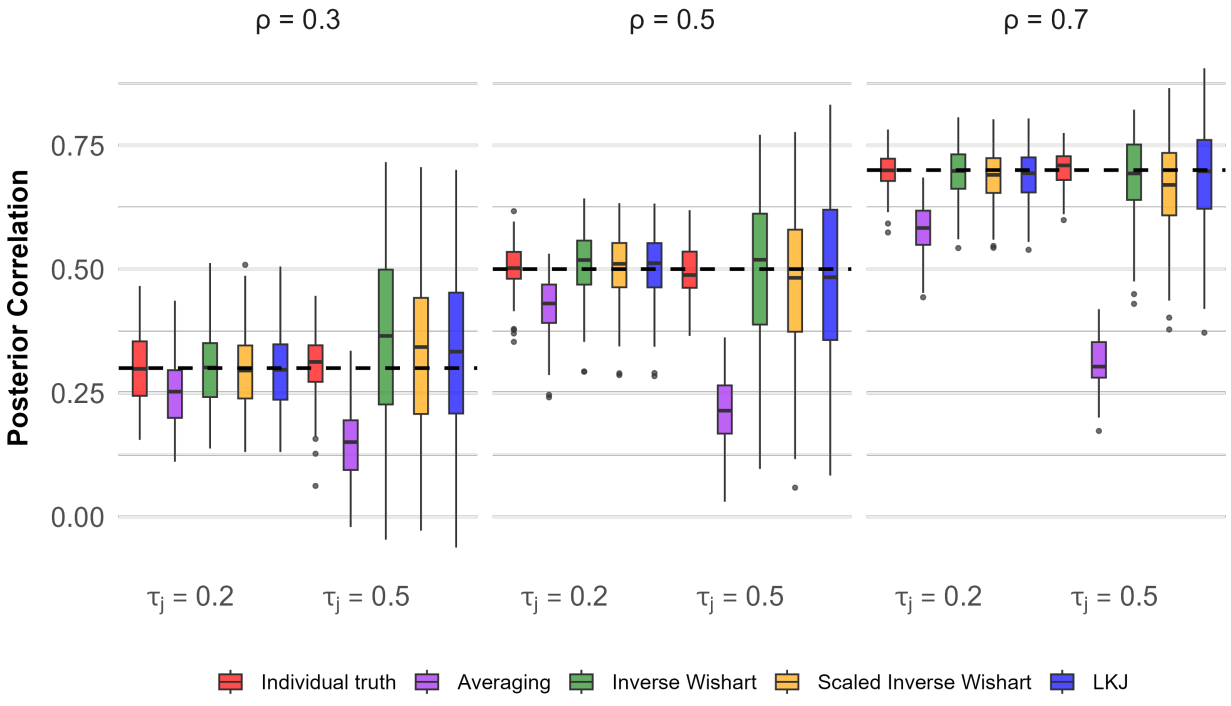

Figure 5. The recovery of correlation for two tasks. There is noticeable attenuation from averaging across trials. There are only slight difference among the four prior specifications.

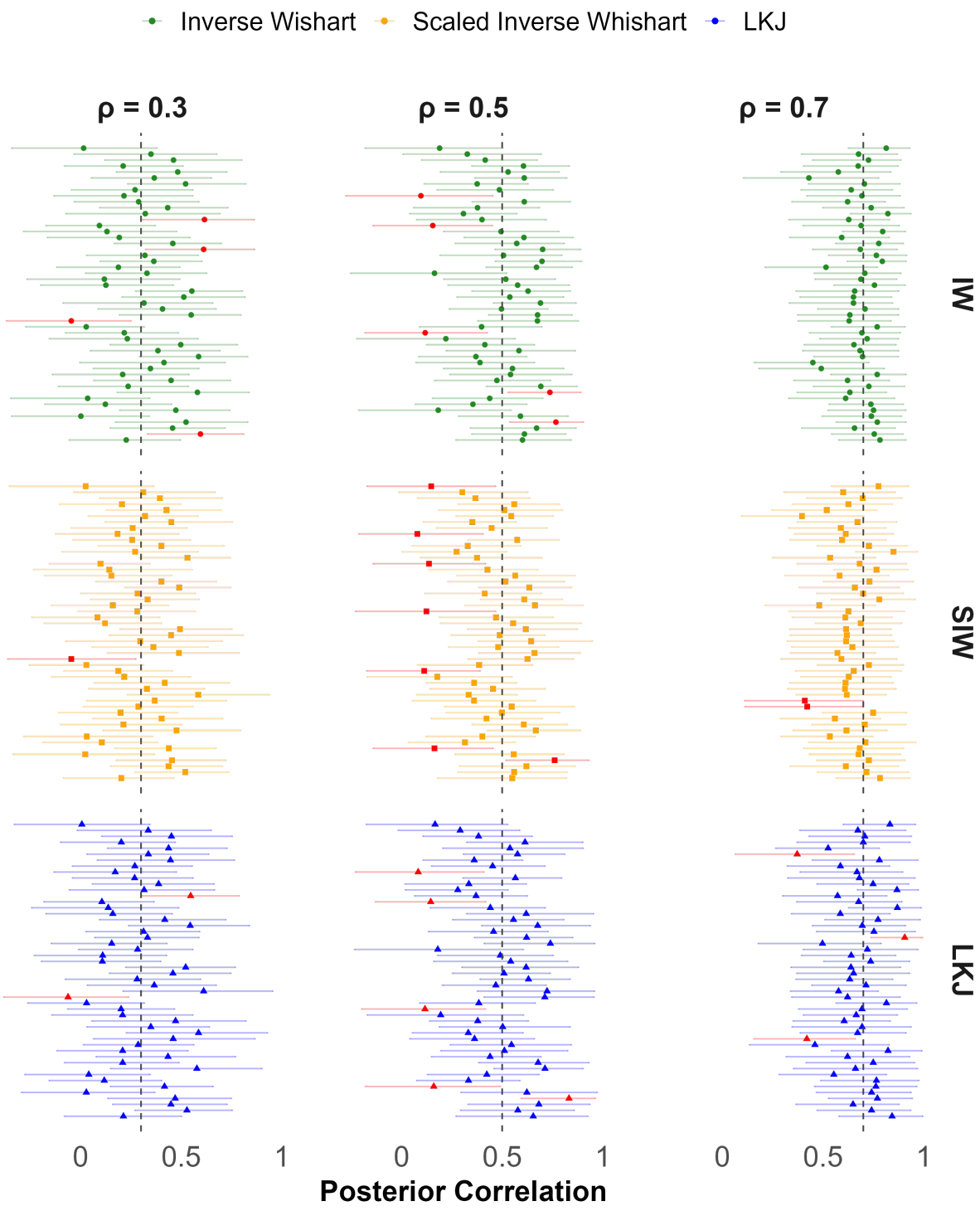

Figure 6. Means and 95 credible intervals for recovered correlation from four model across the first 50 runs. Intervals that fail to cover the true correlation are colored in red. These occur at an expected proportion for all prior specifications.

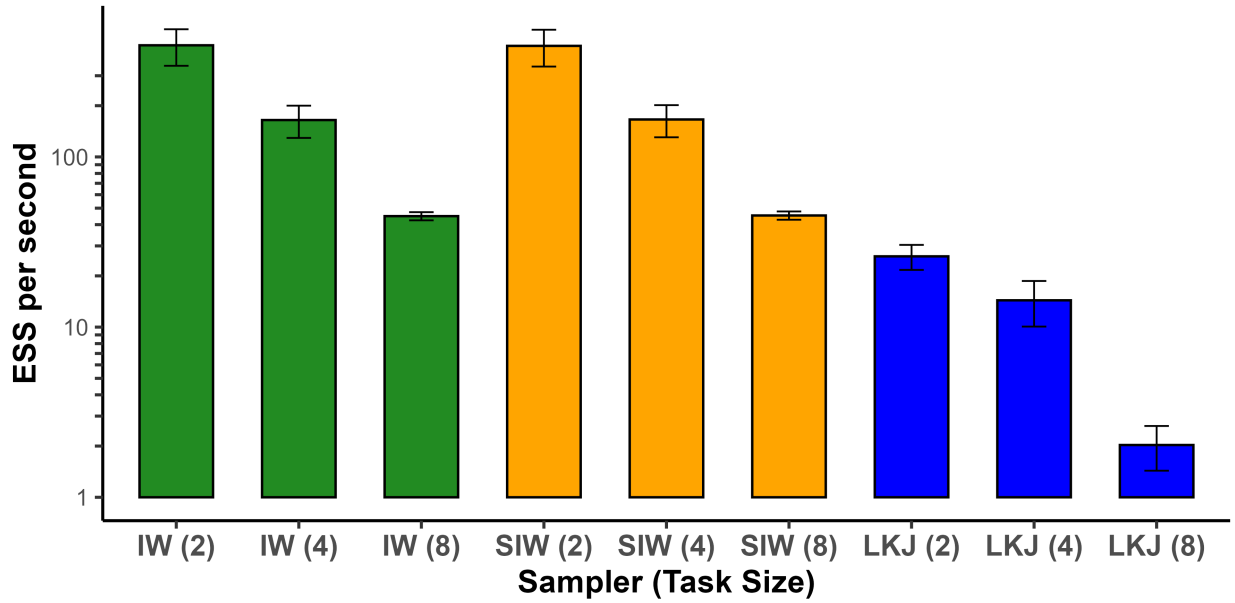

Figure 7. Effective sample size per second for four models in the 2-task and 4-task cases.

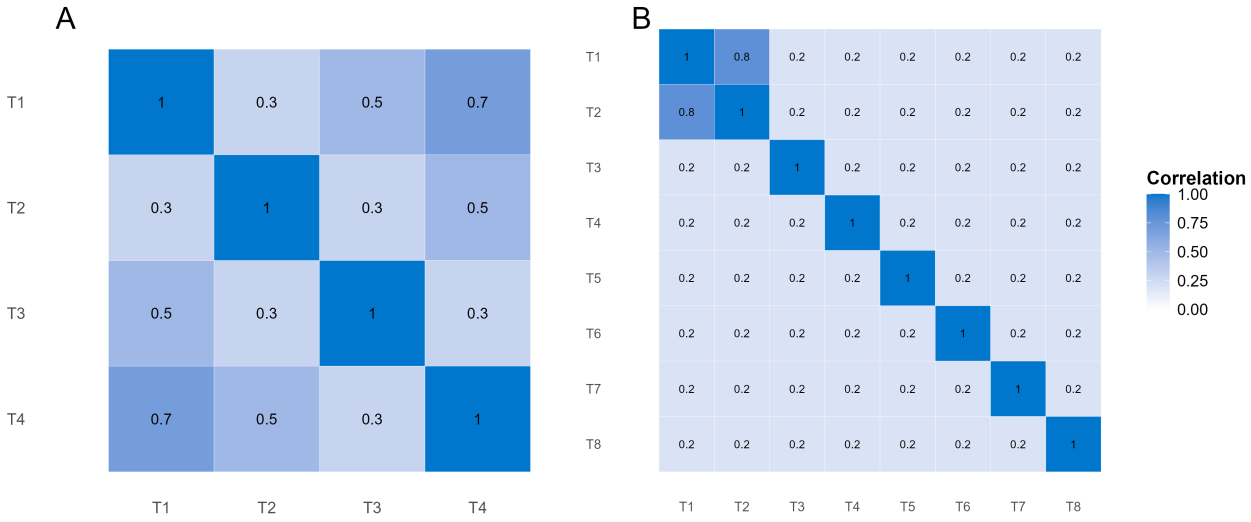

Figure 8. Groud truth correlations. A. Simulation 2 with 4 tasks. B. Simulation 4 with 8 tasks.

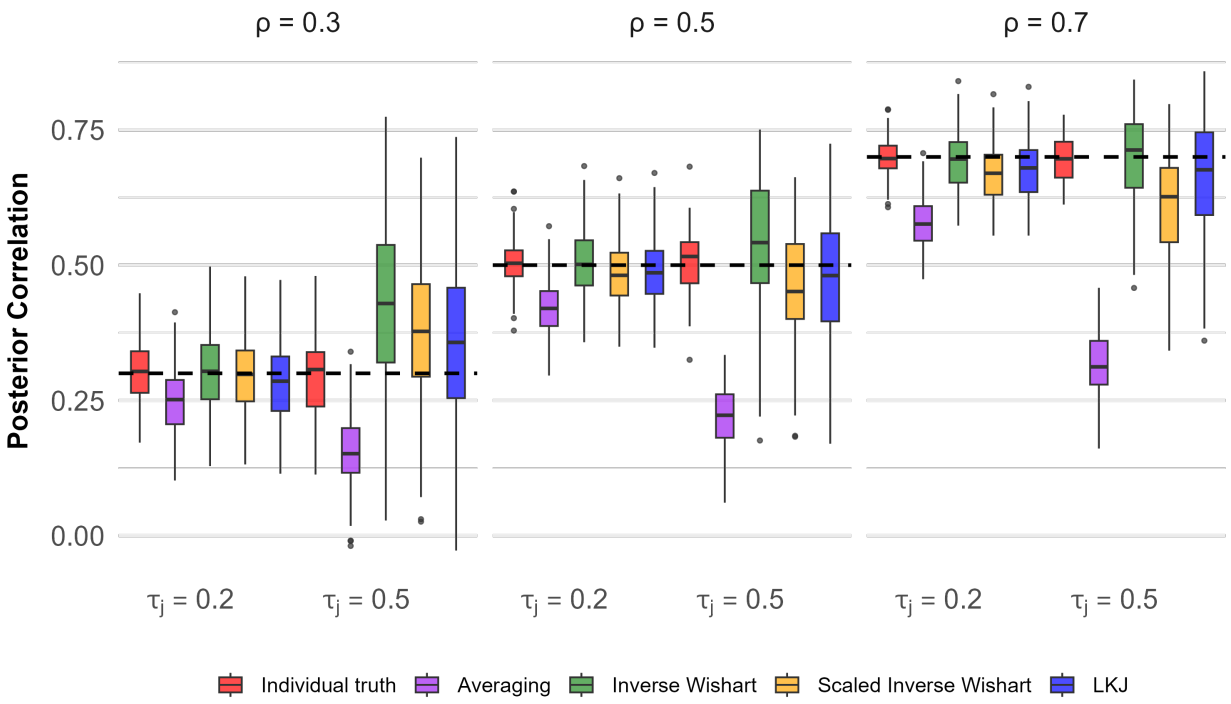

Figure 9. The recovery of correlation for four tasks. There is substantial attenuation from averaging across trials. There are moderate deviations from true values for all prior specifications across the 100 runs, except that under high trial noise and low true correlation, the Inverse Wishart prior tended to over-cover the correlation.

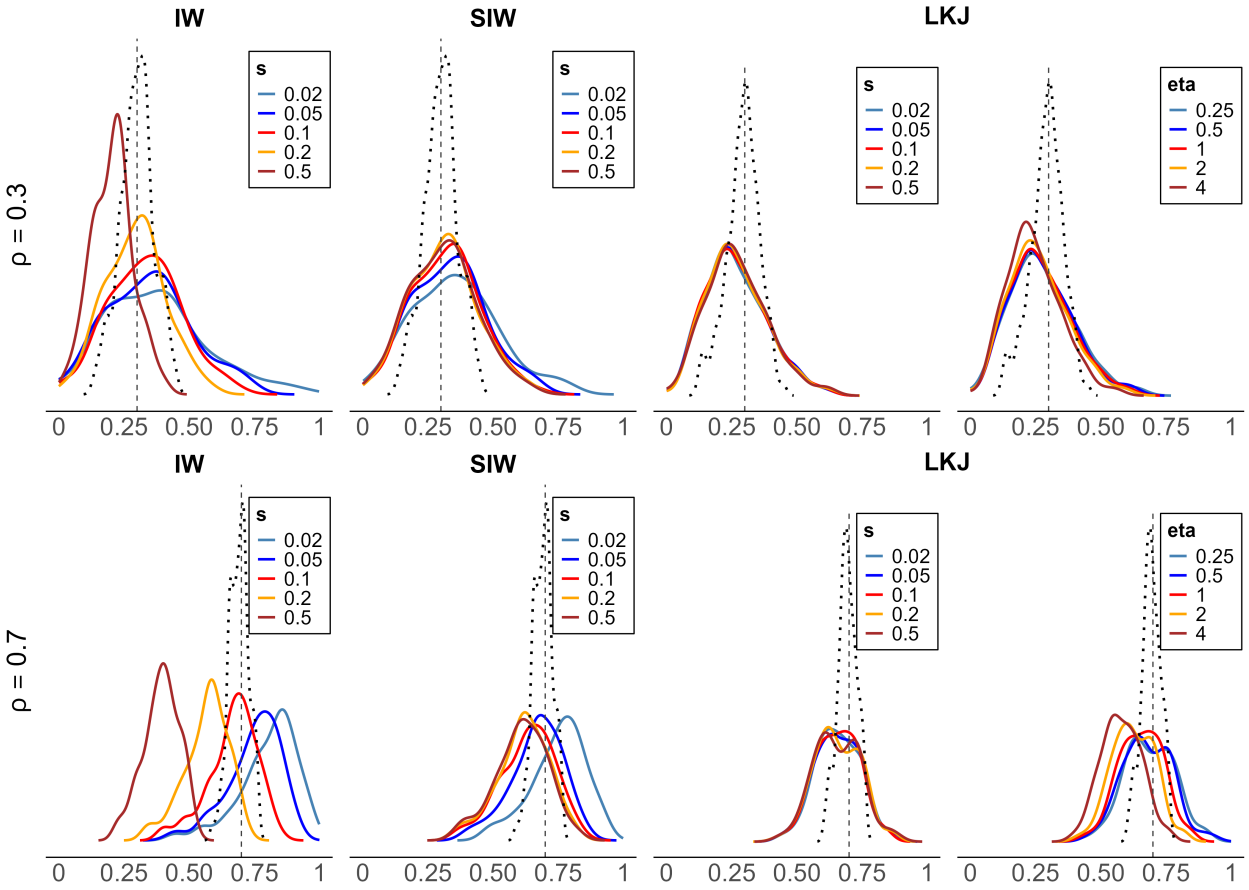

Figure 10. Posterior correlation estimates for four tasks under three models with varying prior settings. Each density curve represents the posterior correlation distributions across 100 runs per condition. The dotted curve shows the individual truth, while the vertical dashed line denotes the true correlation. Posterior means of correlations are highly sensitive to prior scale settings for the Inverse Wishart prior, and less sensitive to prior scale settings for the Scaled Inverse Whishart prior. In contrast, posterior means are much more stable to this prior variation with the LKJ prior. Posterior means are also stable in the LKJ prior with respect to variation in the prior shape parameter. These stabilities make the LKJ prior an attractive choice when there is little guiding prior knowledge.

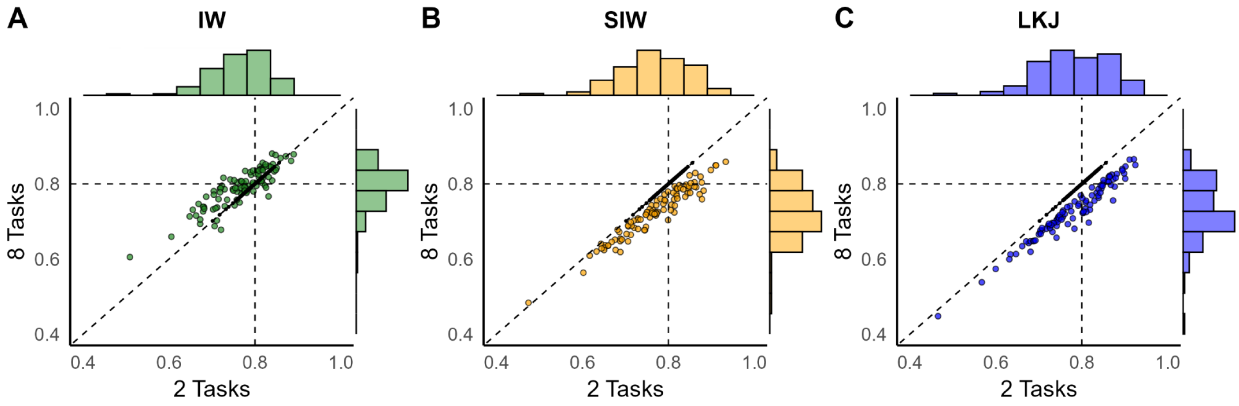

Figure 11. Robustness to inclusion of additional variables. The scatter plots show the relationship of posterior correlation estimates whether in isolation (x-axis) or in the context of 8 tasks (y-axis). There is some downward influence form the prior with inclusion for the Scaled Inverse Wishart prior and LKJ priors.

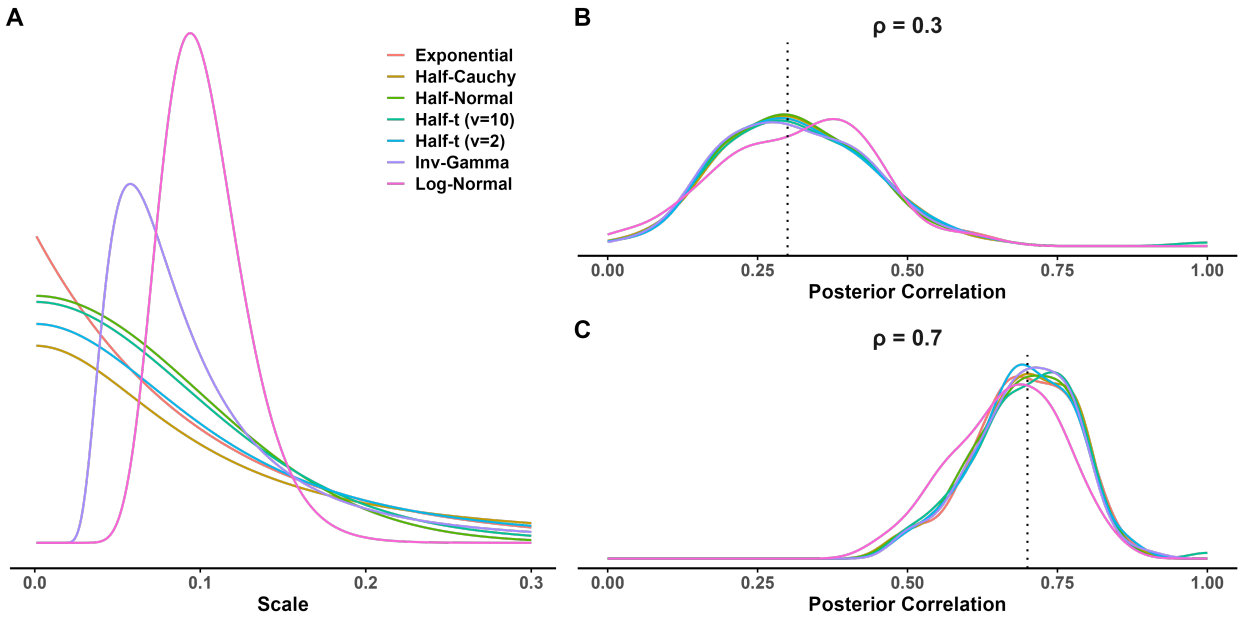

Figure 12. Different scale prior distributions' impact on correlation estimates in LKJ model. A. Seven selected scale prior distributions. All of them were setted to center around the true scale value of 0.1. B. The posterior correlation estimates for low and C. high correlation. All scale prior distributions resulted in similar posterior correlation estimates.

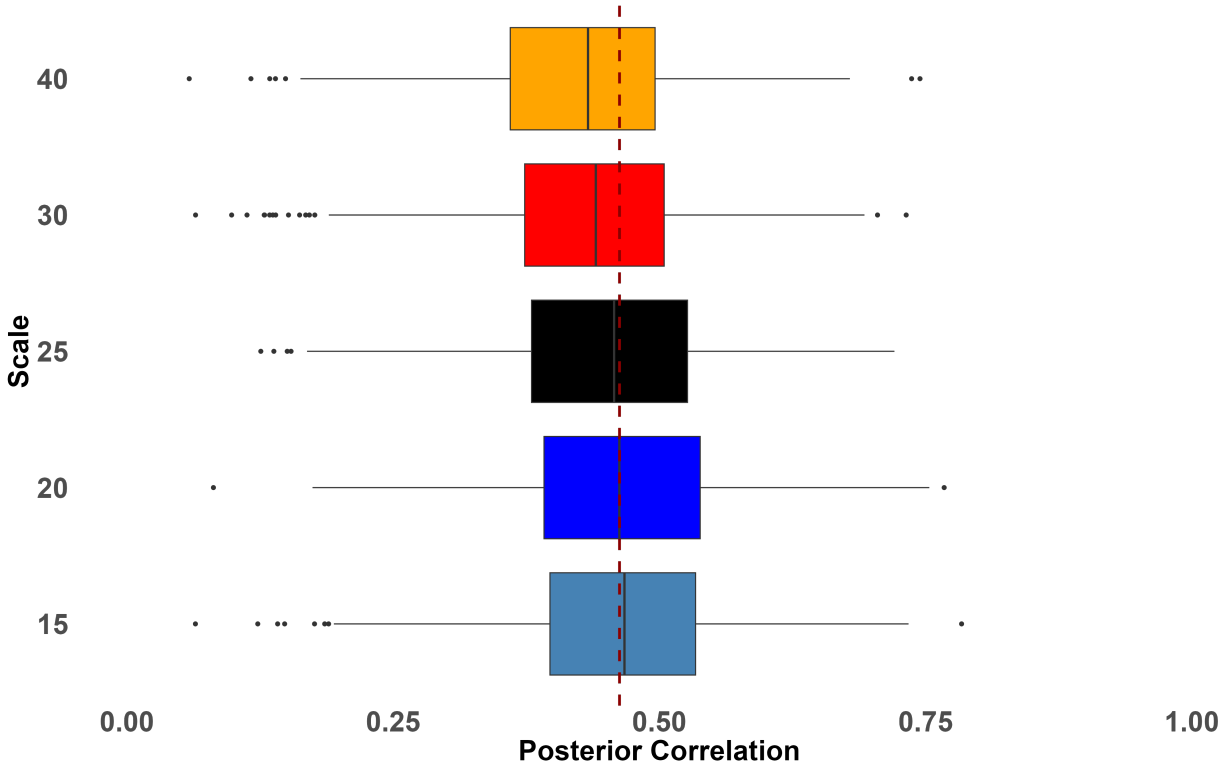

Figure 13. Posterior distributions of the correlation coefficient as a function of prior variance setting. Boxplots do not vary appreciable showing the stability of these posteriors across a reasonable range of settings. The true population correlation is 0.5; the true correlation among these 200 participant is slightly less and is the dashed vertical line.