Why Most Studies of Individual Differences With Inhibition Tasks Are Bound To Fail

Jeffrey N. Rouder[1], Aakriti Kumar[1], & Julia M. Haaf[2]

[1] University of California, Irvine

[2] University of Amsterdam

Version 1, 3/2019

Author Note

We are indebted to Craig Hedge, Claudia von Bastian, and Alodie Rey-Mermet who allowed us to reuse their individual-differences data sets. The Rmarkdown source code for this paper is available at https://github.com/PerceptionAndCognitionLab/ctx-inhibition/papers/problem. This source code contains links to all data sets, all analyzes, and code for drawing the figures and typesetting the paper.

Correspondence concerning this article should be addressed to Jeffrey N. Rouder, . E-mail: jrouder@uci.edu

Abstract

Establishing correlations among common inhibition tasks such as Stroop or flanker tasks has been proven quite difficult despite many attempts. It remains unknown whether this difficulty occurs because inhibition is a disparate set of phenomena or whether the analytical techiques to uncover a unified inhibition phenomenon fail in real-world contexts. In this paper, we explore the field-wide inability to assess whether inhibition is unified or disparate. We do so by showing that ordinary methods of correlating performance including those with latent variable models are doomed to fail because of trial noise (or, as it is sometimes called, measurement error). We then develop hierarchical models that account for variation across trials, variation across individuals, and covariation across individuals and tasks. These hierarchical models also fail to uncover correlations in typical designs for the same reasons. While we can charaterize the degree of trial noise, we cannot recover correlations in typical designs that enroll hundreds of people. We discuss possible improvements to study designs to help uncovering correlations, though we are not sure how feasible they are.

Keywords: Individual Differences, Cognitive Tasks, Hierarchical Models, Bayesian Inference

Why Most Studies of Individual Differences With Inhibition Tasks Are Bound To Fail

In the past two decades, it has become popular to include experimental tasks in studies of individual differences. This is particularly salient in the study of individual differences in inhibition where studies often include experimental tasks such as the Stroop task (Stroop, 1935), the Simon task (Simon, 1968), and the Flanker task (Eriksen & Eriksen, 1974). On the face of it, individual-difference researchers should be sanguine about using such tasks for the following five reasons: First, many of these tasks are designed to isolate a specific cognitive process such as inhibition by contrasting specific conditions. For example, in the Stroop task, the score is the contrast between incongruent and congruent items. The subtraction inherent in the contrast controls for unrelated sources of variation such as overall speed. Second, many of these tasks are robust in that the effects are easy to obtain in a variety of circumstances. Take again, for example, the Stroop task. The Stroop effect is so robust that it is considered universal (MacLeod, 1991). Third, because these tasks are laboratory based and center on experimenter-controlled manipulations, they often have a high degree of internal validity. Fourth, because these tasks are used so often, there is usually a large literature about them to guide implementation and interpretation. Fifth, task scores are relatively easy to collect and analyze with latent-variable models.

Figure 1 shows the usual course of analysis in individual-difference research with cognitive tasks. There are raw data (Panel A), which are quite numerous, often on the order of hundreds of thousands of observations. These are cleaned, and to start the analysis, a task scores for each participant are tabulated (Panel B). For example, if Task 1 is a Stroop task, then the task scores would be each individual's Stroop effect, that is, the difference between the mean RT for incongruent and congruent conditions. A typical task score is a difference of conditions, and might be in the 10s of milliseconds range. The table of individual task scores is treated as a multivariate distribution, and the covariation of this distribution (Panel C) is decomposed into meaningful sources of variation through latent variable models (Panel

D; e.g., Bollen, 1989; Skrondal & Rabe-Hesketh, 2004).

There is a wrench, however, in the setup. Unfortunately, scores from experimental tasks correlate with one another far less than one might think *a priori.* An example is the lack of correlation among the Stroop task and the flanker task. While Friedman and Miyake (2004) found a healthy correlation of .18 between the tasks; subsequent large-scale studies from Hedge, Powell, and Sumner (2018), Pettigrew and Martin (2014), Rey-Mermet, Gade, and Oberauer (2018) and Von Bastian, Souza, and Gade (2015) have found correlations that range from -.09 to .03, and average -.03 in value. The near-zero value of correlation between these two tasks is not an outlier. As a rule, effects in inhibition tasks show surprisingly low correlations (Rey-Mermet et al., 2018). And the low correlations are not limited to inhibition tasks. Ito et al. (2015) considered several implicit attitude tasks used for measuring implicit bias. Here again, there is surprisingly little correlation among tasks that purportedly measure the same concept. This lack of correlation may also be seen in latent variable analyses. Factor loadings from latent variables to tasks are often dominated by a single task indicating that there is little covariation to decompose (MacKillop et al., 2016).

The main question is, "why are these correlations so low?" On one hand, they could reflect underlying true task performance that is uncorrelated or weakly correlated. In this case, the low correlations indicate that performance on the tasks do not largely overlap, and that the tasks are indexing different mental processes. Indeed, this substantive interpretation is taken by Rey-Mermet et al. (2018), who argue that inhibition should be viewed as a disparate rather than unified concept. By extension, different tasks rely on different and disparate inhibition processes.

On the other hand, the true correlations could be large but masked by measurement error. A realistic example is provided in Figure 2. Shown in Panel A are *true difference scores* (or true effects) for 200 individuals on two tasks. The plot is a scatter plot—each point is for an individual; the x-axis value is the true score on one task, the y-axis value is

the true score on the other task. As can be seen, there is a large correlation, in this case it is 0.82. Researchers do not observe these true scores; instead they analyze difference scores from noisy trial data with the tabulation shown in Figure 1. Figure 2B shows the scatterplot of these observed difference scores (or observed effects). Because these observed effects reflect trial-by-trial noise, the correlation is attenuated. In this case it is 0.29. While this correlation is statistically detectable, the observed value is dramatically lower than the true one. Moreover, in simulations with less pronounced true correlations, observed correlations are often undetectable and sometimes reversed.

The amount of attenuation of the correlation is dependent on critical inputs such as the number of trials and the degree of trial-to-trial variability. Therefore, to get a realistic picture of the effects of measurement error it is critical to obtain realistic values for these inputs. In this paper, we survey 15 fairly large inhibition studies. From this survey, presented here subsequently, we derive typical values for the number of trials and the degree of trial-to-trial variability. These typical values are used in Figure 2, and the amount of attenuation of the correlation therefore represents a typical rather than a worst-case scenario.

Measurement Error and Latent Variable Analysis

The amount of attenuation shown in Figure 2, from 0.82 to 0.29, is shocking. No wonder it has been so hard to find correlations. We worry that the common approach of using latent variables to decompose correlations has been more of a distraction than a fruitful approach if only because there is very little covariation to decompose without attention to measurement error.

Most studies follow the analytic workflow in Figure 1, and in the process of forming participant-by-task-score tables (Panel B), ignore the detrimental effects of trial-to-trial variability. It may seem like not much can be done. After all, we computed the correlation

among sample scores. The sample scores are the difference between sample effects, and it may seem hard to improve on the well-known formulas for mean and correlation.

Yet, data from tasks have a hierarchical structure. Trials are nested in conditions and participants who are crossed with task. Sample effects by themselves do not capture trial-to-trial variability, which, in our experience, is the largest source of variability in these designs. One solution is to extend latent variable models down to the trial level (Rouder & Haaf, 2019). By proposing hierarchical models, trial-by-trial variation may be modeled and, perhaps, removed in estimating correlations. The latent-variable approach then becomes simultaneously a means of decomposing variability among tasks and of disattenuating the effects of measurement noise. Figure 2C shows cause for optimism. A hierarchical model discussed subsequently was applied to the data in 2B, and the resulting posterior estimates of participants' inhibition in the tasks shows the strong correlation unmitigated by measurement noise. While the demonstration in 2C certainly breeds confidence, it will not be sufficient. We show here that the hierarchical model recovers only high true correlations well. When the true correlation is lower, the estimates from hierarchical models are too imprecise to be useful in typical study designs.

To foreshadow, we suspect that in general hierarchical models are useful for characterizing the overall degree of measurement noise but are not nearly as useful in recovering the latent correlations. To our knowledge, there is no means of recovering these latent correlations to any degree of acceptable precision. The sad implication is that most individual-difference research with experimental tasks is doomed to fail as one cannot describe adequately the underlying structure of covariation across tasks. Through simulation and model analysis, we broach this difficult question—despite the vast investment of time and money in such studies—are typical individual difference studies with experimental tasks doomed to fail?

Before continuing, we note that we had previously promoted the benefits of extending

hierarchical models down to the trial-by-trial level (Rouder & Haaf, 2019). We had applied a hierarchical model to a single study, from Hedge et al. (2018), which had over 1,400 trials per task. We were able to show the value of the hierarchical approach in characterizing measurement noise, but we never simulated with known true values. Hence, we could not comment on the model's ability to accurately recover correlations. More to the point, we did not imagine then that the model would perform so poorly in simulation. We remain shocked and saddened at the depressing results we present here. This paper is not at all the story we had hoped for, but it is still a critical story for the community of individual-differences scholars to digest.

Spearman's Correction for Attenuation

Before addressing the main question about recovery, we consider the Spearman (1904) correction for the attenuation of correlation from measurement noise. In this brief detour, we assess whether Spearman's correction leads to the recovery of latent correlations among tasks in typical designs. The assessment provides guidance because the data generation in simulations match well with the assumptions in Spearman's correction. If Spearman's correction cannot recover the latent correlations, these correlations may indeed be unrecoverable.

Spearman's derivation comes from decomposing observed variation into true variation and measurement noise. When reliabilities are low, correlations may be upweighted to account for them. In Spearman's classic formula, the disattenuated correlation, denoted r'_{xy} between two variables x and y is

$$r'_{xy} = \frac{r_{xy}}{\sqrt{r_{xx} r_{yy}}},$$

where r_{xy} is the sample correlation and r_{xx} and r_{yy} are the sample reliabilities.[1]

[1] Sample reliability for a task is defined as follows. Let $\bar{Y}_{ik}$ and $s_{\bar{y}_{ik}}$ be the sample mean and sample standard error for the ith individual in the kth condition, $k = 1, 2$. Let $d_i = \bar{Y}_{i2} - Y_{i1}$ be the effect for the

Spearman's correction, while well known, is not used often. The problem is that it is unstable. Panel C of Figure 2 shows the results of a small simulation based on realistic values from inhibition tasks discussed subsequently. The true correlation is .80. The Spearman-corrected correlations, however, are not only variable ranging from 0.18 to 1.99, but not restricted to valid ranges. In fact, 15.90% of the simulated values are greater than 1.0. We should take these problems with Spearman's correction seriously. The poor results in Figure 2 may indicate that in low-reliability environments, true correlations among tasks may not be recoverable. And this lack of recoverability may be fundamental—measurement noise may destroy the correlation signatures.

In the next section, we analyze existing data sets to find appropriate settings for simulations. These settings include sample sizes and estimates of the amount of variability we may reasonably expect across trials and across individuals. With these settings established, we simulate data and assess whether correlations are recoverable. The hierarchical latent correlation estimators, while far from perfect, are better than Spearman-corrected correlation estimators. Subsequently, we apply the same analysis to a large data set from Rey-Mermet et al. (2018) spanning four inhibition tasks to assess whether the observed low correlations reflect independent task performance or attenuation from trial noise. As many researchers before us, even with hierarchical modeling we have a difficult time answering this question.

Variability in Experimental Tasks

To explore whether it is possible to recover correlations in typical designs, it is important to understand not only typical sample sizes, but typical ranges of variability. To estimate within-trial and across-individual variabilities, we us an ordinary variance-components hierarchical model. To truly appreciate how variation can be assessed,

---

ith individual, and let V_d be the sample variance of these effects. Then, r, the reliability, is $r = (V_d - V_s)/V_d$, where V_s is subtractable variability from the cells given by $V_s = \sum_I \sum_K s^2_{\bar{y}_{ik}}/IK$.

the models need to be fully specified rather than left to short-hand. Let $Y_{ijk\ell}$ be the ℓth response for the ith individual in the jth task and kth condition. In this section we analyze each task independently, so we may safely ignore j, the task subscript (we will use it subsequently, however). The model for one task is:

$$Y_{ik\ell} \sim \text{Normal}(\alpha_i + x_k\theta_i, \sigma^2),$$

where α_i is the ith individual's true response time in the congruent condition, $x_k = 0, 1$ codes for the incongruent condition, θ_i is the ith individual's true effect, and σ^2 is the trial noise within an individual-by-condition cell. The critical target are the θ_is, and these are modeled as random effects:

$$\theta_i \sim \text{Normal}(\mu_\theta, \sigma_\theta^2),$$

where μ_θ describes the overall mean effect and σ_θ^2 is the between-person variation in individuals' true effects. Our targets then are within cell variance, σ^2, and between-individual variance, σ_θ^2.

To analyze the model priors are needed for all parameters. Our strategy is to chose scientifically-informed priors (Dienes & Mclatchie, 2018; Etz, Haaf, Rouder, & Vandekerckhove, 2018; Rouder, Morey, & Wagenmakers, 2016; Vanpaemel & Lee, 2012) that anticipate the overall scale of the data. The parameters on baseline response times, in seconds, are $\alpha_i \sim \text{Normal}(.8, 1)$. These priors are quite broad and place no substantive constraints on the data other than baselines are somewhere around 800 ms plus or minus 2000 ms. The prior on variability is $\sigma^2 \sim \text{Inverse Gamma}(.1, .1)$, where the inverse gamma is parameterized with shape and scale parameters (Rouder & Lu, 2005). This prior, too, is broad and places no substantive constraint on data. Priors for μ_θ and σ_θ^2 were informed by the empirical observation that typical inhibition effects are in the range of 10 ms to 100 ms. They were $\mu_\theta \sim \text{Normal}(50, 100^2)$ and $\sigma_\theta^2 \sim \text{Inverse Gamma}(2, 30^2)$, where the values are in milliseconds rather than seconds. A graph of these prior settings for μ and $\sigma_\theta = \sqrt{\sigma_\theta^2}$ is

shown in Figure 3. These priors make the substantive assumption that effects are relatively small and are not arbitrarily variable across people. The scale setting on σ_θ^2 is important as it controls the amount of regularization in the model, and the choice of 30 (on the ms scale) is scientifically informed (see Haaf & Rouder, 2017).

We applied this model to 15 different experimental tasks from a variety of authors. Brief descriptions of the tasks are provided in the Appendix. The results are shown in Table 1, and the specific values inform our subsequent simulations. The first three columns describe the sample sizes: The first column is the total number of observations across the two conditions after cleaning (see Appendix), the second column is the total number of individuals, and the third column is the average number of replicates per individual per condition. The fourth column is the reliability (see footnote 1), and following that is the mean observed effect. The sixth column is model-based estimates of σ, the standard deviation across replicate trials. The last three columns are estimates of the variability across individuals. The first of these columns provides the sample standard deviation of effects, and this value reflects the combination of measurement noise and variability across people. The second is the standard deviation of all θ_i (denoted s_θ), which is a model-based estimate from the first latent level of the hierarchical model. The last column is the model-based estimate of σ_θ, which is an estimate of the same quantity as s_θ, but σ_θ is the estimate from the hyperprior, or the second latent level of the model. The value of $\hat{\sigma}_\theta$ is typically a bit larger in value than s_θ.

From the table, we derive the following critical values as typical. We set the number of individuals to $I = 200$ and the number of trials per condition to $L = 100$. We set the trial-by-trial variation to $\sigma = 180$ ms, and the variation of individuals' true effects to $\sigma_\theta = 20$ ms. Figure 2 is based on these values, as are the following simulations.

Let's examine these choices in more detail. The choice of $I = 200$ people and $L = 100$ replicates per condition is made to emulate designs where many people are going to run in

several inhibition tasks. For tasks with two conditions, there are 40,000 observations per task. In a typical battery with $J = 10$ tasks, the total number of observations is 400,000, which is quite large. Hence, our choices seem appropriate to typical large-scale individual-difference studies with experimental tasks.

Next, let's examine the choices of variabilities: $\sigma = 180$ ms and $\sigma_\theta = 20$ ms. The critical choice is the latter, and a reader may question its small size. Does it make sense, and why is the larger value s_d, the empirically observed standard deviation of individuals effect scores not used. The values s_d are larger because they necessarily include contributions from trial noise and variability across individuals. The second column, s_θ reflects the model's partition of variance, that is, what is left over after trial noise, given by σ, is accounted for. Given the assumptions of the model, it reflects only the variability across individuals. Hence, it is the far better value for simulation.

We provide a second argument that may be more intuitive for understanding the 20 ms value. Consider the possibility that all people truly respond faster in the congruent than in the incongruent condition. Or, restated, nobody has a negative true effect. This condition is called *dominance* in Rouder and Haaf (2018), and is explored extensively in Haaf and Rouder (2017) and Haaf and Rouder (in press). The results from these studies is that dominance broadly holds. In the Stroop case, everyone Stroops, that is, in the large trial limit, everyone has truly faster scores for congruent than incongruent stimuli. If dominance holds, and the true mean effect is small across the population, say 50 ms, then the variance between individuals cannot be too high. For if it were large, then some proportion of people must have negative true effects. Dominance—which is natural and seems to hold in almost all sets we have examined—provides a limit on the size of variability. Figure 3 provides a graph of true values with a spread of 20 ms. As can be seen, there is only minimal mass for negative true values, and the spread of true values to us seems appropriate for a true 50 ms effect.

Estimating Correlations Among Tasks

The critical question is then whether accurate estimation of correlation is possible. The small simulation in the introduction, which was based on the above typical settings for two tasks and a true population correlation of .80, showed that naive correlations among sample effects were greatly attenuated and Spearman's correction was unstable. We now assess the recoverability of true latent correlations with the hierarchical models used to simulate data and for several values of true correlations.

A Hierarchical Model for Correlation

Here we develop a hierarchical trial-level model for many tasks that explicitly models the covariation in performance among them. A precursor to this model is provided in Rouder and Haaf (2019). At the top level, the model is:

$$Y_{ijk\ell} \sim \text{Normal}(\alpha_{ij} + x_k\theta_{ij}, \sigma^2).$$

The target of inquiry is θ_{ij} the effect for the ith participant in the jth task. The specification is made easier with a bit of vector and matrix notation. Let $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{iJ})'$ be a column vector of the ith individual's true effects. This vector comes from a group-level multivariate distribution. The following is the case for three tasks:

$$\boldsymbol{\theta}_i = \begin{bmatrix} \theta_{i1} \\ \theta_{i2} \\ \theta_{i3} \end{bmatrix} \sim \text{N}_3 \left(\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1\sigma_3 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 \\ \rho_{13}\sigma_1\sigma_3 & \rho_{23}\sigma_2\sigma_3 & \sigma_3^2 \end{bmatrix} \right).$$

More generally, for J tasks,

$$\boldsymbol{\theta}_i \sim \mathrm{N}_J(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \quad (1)$$

Priors are needed for $\boldsymbol{\mu}$, the vector of task means, and $\boldsymbol{\Sigma}$, the covariance across the tasks. We take the same strategy of using scientifically-informed priors. For $\boldsymbol{\mu}$, we place the normal in Figure 3A on each element. For $\boldsymbol{\Sigma}$, we have a number of choices as there has been much recent development in the Bayesian literature. Perhaps the state-of-the-art is the *LKJ prior* named after the initials of its developers (Lewandowski, Kurowicka, & Joe, 2009). This prior is based on decomposing the covariation into a set of variances and a correlation matrix. The LKJ prior is placed on the correlation matrix and priors on the variance terms may be set independently. This prior has been popularized by McElreath (2016), and implementation is convenient in the `R`-package `rstan` (Stan Development Team, 2018). We implemented this prior as well as the more classic inverse Wishart prior (O'Hagan & Forster, 2004). The inverse Wishart was formerly popular because it is conjugate and convenient in practice. But there is a drawback to the inverse Wishart. Unlike the LKJ prior, the inverse Wishart is a prior on the covariation matrix $\boldsymbol{\Sigma}$, and prior settings on variability will affect the posterior values of correlation. In contrast, settings on the variance in the LKJ setup have a much smaller effect on posterior values of correlation. In the following we used an LKJ(1) prior, where the value 1 refers to the shape of the distribution, and the priors on σ_θ^2 as discussed previously are used on the variances terms in $\boldsymbol{\Sigma}$. For two tasks, a shape of 1 implies a uniform marginal prior on the correlation, ρ.

Two Tasks

The first simulation is for two tasks. Using the typical sample sizes discussed above, each hypothetical data set consisted of 80,000 observations (200 people $\times$ 2 tasks $\times$ 2 conditions $\times$ 100 replicates per condition). One might hope that with such a large sample

size and with a goal of estimating a single correlation, the true population correlation, ρ, might be recoverable. Supporting this hope is the success of the single run in Figure 2C. On the other hand, given the large degree of measurement noise and the instability of Spearman's correction (Figure 2B), it seems plausible that ρ may not be unrecoverable. For the simulations, true correlation values across the two tasks were varied on three levels with values of .2, .5, and .8. For each of these levels, 100 data sets were simulated and analyzed.

Figure 4A shows the results as boxplots. Naive correlations from participant-by-task sample means are shown in red. As expected, these correlations suffer a large degree of attenuation from trial noise. Correlation estimates from Spearman's correction are shown in green. These values are better centered though some of the corrected values are greater than 1.0. The correlation estimates from the hierarchical model with the LKJ(1) prior are shown in blue.

Overall, the correlation estimates from Spearman's correction and the hierarchical model are dramatically better than the naive sample-effect correlations. Yet, the estimates are quite variable. For example, consider correlations when the population value is .2. The model estimates range from -0.38 to 0.73 and miss the target with a RMSE of 0.20. Spearman corrected estimates are a smidge better and have an RMSE of 0.19. Overall this variability is too high to warrant confidence that correlations may be faithfully recovered.

Are there risks in using model-based recovery? We see in simulation that the model and Spearman-corrected recovery is exceedingly variable. One potential problem is that in any one study, researchers using the model inflate the values of correlations. The attenuation in the naive correlations is conservative in that recovered correlations are never inflated, rather, they are dramatically deflated. In this regard, we can think of naive-correlations as having a fail-safe quality where high-value correlation estimates are avoided at the draconian expense of not detecting true high correlations. Spearman-corrected correlations do not share this fail-safe orientation. The variability in estimation results in values that are both

inflated and deflated.

The critical question is about model-based recovery. Figure 4A shows only posterior mean estimates. Yet, in the Bayesian approach, the target is not just the posterior mean, but the entirety of the posterior distribution. Figure 4B-D shows the posterior 95% credible intervals for all runs with true correlations of .2, .5, and .8, respectively. As can be seen, the 95% credible intervals are fairly wide, and as a result, the analyst knows that correlations have not been well localized. This lack of localization provides the needed hedge for overinterpreting inflated values. With the Bayesian model-based estimates, at least we know how little we can say about the true correlations.

Six Tasks

We explored correlations across six tasks. Each hypothetical data set consisted of 240,000 observations. To generate a wide range of correlations, we used a one-factor model to simulate individuals' true scores. This factor represents the individual's inhibition ability. This ability, denoted z_i, is distributed as a standard normal. Tasks may require more or of the individuals' inhibition ability. Therefore, task loadings onto this factor z_i are variable and, as a result, a wide range of correlations occur. The following task loading values work well in producing a diversity of correlations: 0.5 ms, 5.4 ms, 10.3 ms, 15.2 ms, 20.1 ms, and 25 ms. Following the one-factor structure we may generate true scores, θ_{ij}, for each task and participant:

$$\theta_{ij} \sim \text{Normal}(\mu_j + z_i w_j, \eta^2),$$

where z_i is the true ability, w_j is the task loading, μ_j is the task overall mean, and η^2 is residual variability in addition to that from the factors. In simulation we set $\eta = 10$ ms, and this setting yields standard deviations across θ_{ij} between 10 ms and 30 ms, which is similar to the 20 ms value used previously. The true population variance for the one-factor model is

$\boldsymbol{\Sigma} = \boldsymbol{w}\boldsymbol{w}' + \boldsymbol{I}\eta^2$, where $\boldsymbol{w}\boldsymbol{w}'$ is the matrix formed by the outer product of the task loadings. The true correlation matrix from the variance-covariance matrix $\boldsymbol{\Sigma}$ is shown in Figure 5A, and the values subtend a large range from near zero to 0.83.

The recovery of correlations is shown for a single simulation run in Figure 5B-D. The attenuation for the naive correlations is evident, as is variability in model-based and Spearman corrected estimates. Figure 6 shows the performance of the methods across the 10 simulation runs. As can be seen, there remains the dramatic attenuation for the naive correlation of sample effects and the excessive variability for the Spearman-corrected and model-based correlation estimates. The RMS error across the whole range of true values for the Spearman corrected estimates and model-based estimates are 0.25 and 0.18, respectively. The improvement of the LKJ(1) prior over the Spearman correction results with larger numbers of tasks is heartening.

Analysis of Rey-Mermet, Gade, and Oberauer (2018)

To demonstrate the real-world difficulties of correlation recovery, we re-examined the flanker and Stroop tasks in Rey-Mermet et al.'s battery of inhibition tasks. The authors included two different types of Stroop tasks (a number Stroop and a color Stroop task, see the Appendix for details) and two different types of flanker tasks (a letter flanker and an arrow flanker task, see the Appendix for details). The question then is about the correlation across the tasks. A reasonable expectation is that all of these tasks correlate positively.

The top three rows of Figure 7 show the estimated correlations from sample effects, Spearman's correction, and the hierarchical model. Given the previous simulations results, it is hard to know how much credence to give these estimated correlations. In particular, it is hard to know how to interpret the negative correlation between the arrow flanker and color Stroop task.

To better understand what may be concluded about the range of correlations, we plot the posterior distribution of the correlation (Figure 8A). These distributions are unsettling. The variation in most of these posteriors is so wide that firm conclusions are not possible. The exception is the null correlation between number and color Stroop which seems to be somewhat well localized. The surprisingly negative correlation between color Stroop and arrow flanker comes from a posterior so broad that the 95% credible interval is [-0.70,0.77]. Here, all we can say is that very extreme correlations are not feasible. We suspect this limited result is not news.

Analysis of Rey-Mermet et al. (2018) provides an opportunity to examine how hierarchical models account for variation across trials as well as variation across people. Figure 8B shows sample effects across individuals for the color Stroop and arrow flanker tasks, the two tasks that were most negatively correlated. There is a far greater degree of variation in individual's effects for the color Stroop task than for the arrow flanker task. The model estimates (Figure 8C) reflect this difference in variation. The variation in arrow flanker is so small that it can be accounted for with trial variation alone. As a result, the hierarchical model shows almost no individual variability. In contrast, the variability in the color Stroop is large and the main contributor is true variation across individuals rather than trial variation. Hence, there is relatively little shrinkage in model estimates. The lack of variation in the arrow flanker task gives rise to the uncertainty in the recovered correlation between the two tasks.

Discussion

A basic question facing researchers in cognitive control is whether inhibition is a unified phenomenon or a disparate set of phenomena. A natural way of addressing this question is to study the pattern of individual differences across several inhibition tasks. In this paper, we have explored whether correlations across inhibition tasks may be recovered.

We consider typically large studies that enroll hundreds of participants. The answer is negative—correlations are difficult to recover with anywhere near the accuracy that would allow for a definitive answer to this basic question. This statement of poor recovery holds even for hierarchical models that are extended to the trial level.

Why this depressing state-of-affairs occurs is fairly straightforward. Relative to trial noise, there is little true individual variation in inhibition tasks. To see why this is so, consider an average effect, say one that is 50 ms. In inhibition tasks like Stroop and flanker, we can safely make a *dominance assumption*—nobody truly has a negative effect (Haaf & Rouder, 2017). That is to say nobody truly identifies incongruent stimuli faster than congruent ones. Under this assumption, where all true scores are positive, a small mean necessarily implies a small variance. For example, if true Stroop effects are reasonably normally shaped and the mean is 50 ms and there can be no mass below zero, then an upper bound on variability across true scores is a standard deviation of 20 ms. This is a small amount of variation compared to trial variability, which is typically 10 times larger. This small degree of variation necessarily implies a small degree of covariation across tasks. This small degree is beyond the resolution of our experimental designs, and that is why our studies are doomed to fail.

Solutions

There are several possible solutions to the problem of small individual variation, though none are easy or straightforward in practice. We take them in turn:

More Trials: Perhaps the simplest solution is to run more trials per person per condition. The usual 50 or 100 trials per task per condition is clearly not enough. To calculate a good number, researchers should decide in advance how well they need to estimate an individuals' true effect. We recommend a 5 ms standard error on individual

effects, which seems reasonable if true individual variability is around 20 ms. With this value, we can calculate the number of needed trials. If people have 180 ms of noise per trial, and we are computing a difference score, then the standard error is $180\sqrt{2/n}$, where n is the number of trials per condition per task. Setting this standard error to 5 ms yields $n = 2591$. Such a large number of trials per individual per task per condition is outside the practical constraints of most research agendas.

Better Tasks: Perhaps the most obvious solution is to search for inhibition tasks with greater individual variation. In practice, this means engineering tasks that have large overall effects. Yet, as far as we know, there is no magic bullet to increase effect sizes. Take, for example, manual Stroop tasks. Outside of increasing the number of responses, we do not know otherwise how to increase the size of the effects. And when the number of responses is increased, the trial variability is increased as well. In summary, it may not be known at this time how to increase the size of effects over what is typically seen.

One alternative to increasing individual variability, from Engle and colleagues (Kane & Engle, 2003; Unsworth, Schrock, & Engle, 2004), is to dispense with contrasts altogether. In this approach, task scores reflect the average rather than the difference among conditions. For example, the Stroop effect could be defined as the average speed in congruent and incongruent conditions. Indeed, there is far more individual variation in condition averages than in condition differences. The downside of this approach, however, is interpretability. It is not clear that such condition averages can be interpreted as inhibition measures as they reflect the contribution of a host of processes and are likely dominated by a general speed component (Salthouse, 1996). Given these difficulties in interpretation, we are hesitant to recommend this approach.

More Constrained Models: One future direction is the development of highly constrained hierarchical models. The population-level variance matrix in our model, $\mathbf{\Sigma}$ in Equation 1, is modeled with the unconstrained LKJ(1) prior of Lewandowski et al. (2009).

Yet, perhaps better recovery is possible with more *a priori* constraints on correlations built into the model. One obvious approach is to build reduced-factor models, say a principle-components decomposition of the covariation (Bishop, 1999). The effectiveness of such constraints is a matter of how much variation there is to decompose. The risk is that with impoverished data, the recovered structure may simply reflect the *a priori* constraints. Whether latent variable decompositions offer true gains in such impoverished contexts remains an open question.

It is unlikely that any of the three possible solutions, increased trial sizes, better-engineered tasks, or analysis with *a priori* constraints, will be sufficient to understand the structure of inhibition. Instead, we suspect, it will be advances on all three fronts in concert that may hold the possibility. The take-home point is that this problem of understanding inhibition from individual difference is hard, and answering it requires an inordinate degree of care, judiciousness, and humility.

Trial-Level Models Are Necessary

The key innovation in this paper is the use of hierarchical linear models that extend down to the trial level. By modeling variation at the trial level and across people simultaneously, we are able to avoid the dramatic attenuation of correlation from trial noise. In this paper, we have emphasized the shortcomings of the whole endeavor to recover correlations. We urge readers not to misread this emphasis as one against such models.

On the contrary, we feel strongly that a precondition for recovering correlations is the use of hierarchical models that extend to the trial level. While it may be that in typical designs, recovery will be difficult with these hierarchical models, it will be impossible without them. We cannot stress this point enough—there is no loss in using models that account for trial variation. Conversely, the loss from aggregating the data in forming sample effects is

dramatic as the resulting degree of measurement error is large.

The good news here is that individual difference researchers are well acquainted with hierarchical and latent variable models. The ones we use are run-of-the-mill linear mixed models with normally-distributed errors. Although we fit them in the Bayesian framework using `R` (R Core Team, 2018) and `stan` (Carpenter et al., 2017), there is nothing to prevent classical analysis. Classical model analysis should be convenient in a wide variety of packages including `lme4`, `Mplus`, and `AMOS`. Given the field's familiarity with mixed linear models and the accumulated expertise in application, the wide-spread use of trial-level hierarchical models is feasible. Not using such models in this context strikes us as analytic malpractice.

Appendix

Data Set 1, Von Bastian et al. (2015): The task was a number Stroop task. Participants were presented a string of digits. In each string, the digits were always replicates, say *22* or *444*, and the lengths varied from one digit to four digits. The participants identied the length of the string, for example, the correct report for *444* is 3. In the congruent condition, the length and the digits matched; e.g., *22* and *4444*. In the incongruent condition, the length and digits mismatched, e.g., *44* and *2222*. We used somewhat different data cleaning steps than the original authors. Ours are described in Haaf and Rouder (2017).

Data Set 2, Pratte, Rouder, and Morey (2010), Experiment 1: The task was a color Stroop task. Participants identified the color of the color words, e.g. the word *RED* presented in blue. In the congruent condition, presentation color and word meaning matched, e.g. *BLUE* presented in blue. In the incongruent condition, they did not match, e.g. *RED* presented in blue. We used the original authors' cleaning steps.

Data Set 3, Pratte et al. (2010), Experiment 2: The task was a sidedness judgment Stroop task. Participants were presented the words *LEFT* and *RIGHT*, and these were presented to the left or right of fixation. Participants identified the position of the word while ignoring the meaning of the word. A congruent trial occurred when position of the word and word meaning corresponded; an incongruent trial emerged when position and word meaning did not correspond. We used the original authors' cleaning steps.

Data Set 4, Rey-Mermet et al. (2018): The task was a number Stroop task. Participants identified the length of digit strings much like in Data Set 1. Cleaning proceeded as follows. First, note that in the original, trials ended at 2.0 seconds even if the participant did not respond. We call these trials *too slow*. 1. We discarded the five participants discarded by the original authors; 2. We discarded too-slow trials, error trials, and trials with RTs below .275 seconds (*too-fast* trials). 3. We discared all participants who had more

than 10% errors, who had more than 2% too-slow trials, or more than 1% too fast trials.

Data Set 5, Rey-Mermet et al. (2018): The task was a color Stroop task. Participants identified the color of the presented words (red, blue, green, or yellow). The presentation color and word meaning matched in the congruent condition and did not match in the incongruent condition. Cleaning steps were the same for Data Set 4.

Data Set 6, Hedge et al. (2018): The task was a color Stroop task. Participants identified the color of a centrally presented word (red, blue, green, or yellow). In the congruent condition, presentation color and word meaning matched. In the incongruent condition, they did not match. Following Hedge et al. (2018), we combined data from their Experiments 1 and 2. Our cleaning steps differed from Hedge et al. (2018) and are described in Rouder and Haaf (2019).

Data Set 7, Von Bastian et al. (2015): The task was a Simon task. Participants were presented either a green or red circle to the left or right of fixation. They identifed the color, green or red color by pressing buttons with their left or right hand, respectively. The spatial location of the circle and of the response could be either congruent (e.g., a green circle appearing on the left) or incongruent (e.g., a green circle appearing on the right). Cleaning steps are described in Haaf and Rouder (2017).

Data Set 8, Pratte et al. (2010), Experiment 1: The task was a Simon task almost identical to that in Data Set 7. Participants identified the color of a square presented to the left or right of fixation by making a lateralized key response. A congruent trial occurred when position of the square was ipsilateral correct key response.; an incongruent trial occurred when the position of the square was contralateral to the correct key response. We used the original authors cleaning steps.

Data Set 9: Pratte et al. (2010), Experiment 2: The task was a *lateral-words* Simon task. Participants were presented the words *LEFT* and *RIGHT* to the left or right of

fixation. Participants identified the meaning of the word while ignoring the location of the word. A congruent trial occurred when position of the word and word meaning corresponded; an incongruent trial occurred when position of the word and word meaning did not match. We used the original authors cleaning steps.

Data Set 10, Von Bastian et al. (2015): The task was a letter-flanker task. Participants were presented strings of seven letters and judged whether the center letter was a vowel (A, E) or consonant (S, T). The congruent condition was when the surrounding letters came from the same category as the target (e.g. $AAAEAAA$); the incongruent condition was when the surrounding letters came from the opposite category of the target (e.g., $TTTETTT$). Cleaning steps are described in Haaf and Rouder (2017).

Data Set 11, Rey-Mermet et al. (2018): The task was an arrow flanker task. Participants identified the direction of the central arrow (left/right) while ignoring four flanking arrows. Congruency and incongruency occured when the center arrow matched and mismatched the direction of the flanker arrows, respectively. Cleaning steps were the same for Data Set 4.

Data Set 12, Rey-Mermet et al. (2018): The task was a letter-flanker task almost identical to Data Set 10. Cleaning steps were the same for Data Set 4.

Data Set 13: Hedge et al. (2018): The task was an arrow flanker task almost identical to Data Set 11. Following Hedge et al. (2018), we combined data from their Experiments 1 and 2. Our cleaning steps differed from Hedge et al. (2018) and are described in Rouder and Haaf (2019).

Data Set 14, Rouder, Lu, Speckman, Sun, and Jiang (2005): The task was a digit-distance task. Participants were presented digits 2, 3, 4, 6, 7, 8, and had judged whether the presented digit was less-than or greater-than five. Digits further from five are identified faster than those close to 5. Responses to digits 2 and 8 conprised the *far*

condition; responses to digits 4 and 6 comprised the *close* condition. The difference in conditions comprised a *distance-from-five* effect. We used the original authors' cleaning steps.

Data Set 15, Rouder, Yue, Speckman, Pratte, and Province (2010): The task was a grating-orientation discrimination. Participants were presented nearly-vertical Gabor patches that were very slightly displaced to the left or right; they indicated whether the displacement was left or right. Displacements were $\pm 1.5°$, $\pm 2.0°$, and $\pm 4.0°$ from vertical. Responses from the $\pm 1.5°$ comprised the *hard* condition; responses from the $\pm 4.0°$ comprised the *easy* condition; the difference comprised a *orientation-strength* effect. We used the original authors' cleaning steps.

References

Bishop, C. M. (1999). Bayesian pca. In *Advances in neural information processing systems* (pp. 382–388).

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*.

Dienes, Z., & Mclatchie, N. (2018). Four reasons to prefer bayesian analyses over significance testing. *Psychonomic Bulletin & Review*, *25*(1), 207–218.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*, 143–149.

Etz, A., Haaf, J. M., Rouder, J. N., & Vandekerckhove, J. (2018). Bayesian inference and testing any hypothesis you can specify. *Advances in Methods and Practices in Psychological Science*.

Friedman, N. P., & Miyake, A. (2004). The relations among inhibition and interference control functions: A latent-variable analysis. *Journal of Experimental Psychology: General*, *133*, 101–135.

Haaf, J. M., & Rouder, J. N. (2017). Developing constraint in Bayesian mixed models. *Psychological Methods*, *22*(4), 779–798.

Haaf, J. M., & Rouder, J. N. (in press). Some do and some don't? Accounting for variability of individual difference structures. *Psychonomic Bulletin and Review*. Retrieved

from https://psyarxiv.com/zwjtp/

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavioral Research Methods*.

Ito, T. A., Friedman, N. P., Bartholow, B. D., Correll, J., Loersch, C., Altamirano, L. J., & Miyake, A. (2015). Toward a comprehensive understanding of executive cognitive function in implicit racial bias. *Journal of Personality and Social Psychology*, *108*(2), 187.

Kane, M. J., & Engle, R. W. (2003). Working-memory capacity and the control of attention: The contributions of goal neglect, response competition, and task set to stroop interference. *Journal of Experimental Psychology: General*, *132*(1), 47.

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.

MacKillop, J., Weafer, J., Gray, J. C., Oshri, A., Palmer, A., & Wit, H. de. (2016). The latent structure of impulsivity: Impulsive choice, impulsive action, and impulsive personality traits. *Psychopharmacology*, *233*(18), 3361–3370.

MacLeod, C. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, *109*, 163–203.

McElreath, R. (2016). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Boca Raton, FL: Chapman & Hall/CRC.

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

Pettigrew, C., & Martin, R. C. (2014). Cognitive declines in healthy aging: Evidence

from multiple aspects of interference resolution. *Psychology and Aging*, *29*(2), 187.

Pratte, M. S., Rouder, J. N., & Morey, R. D. (2010). Separating mnemonic process from participant and item effects in the assessment of ROC asymmetries. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 224–232.

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Rey-Mermet, A., Gade, M., & Oberauer, K. (2018). Should we stop thinking about inhibition? Searching for individual and age differences in inhibition ability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Retrieved from http://dx.doi.org/10.1037/xlm0000450

Rouder, J. N., & Haaf, J. M. (2018). Power, dominance, and constraint: A note on the appeal of different design traditions. *Advances in Methods and Practices in Psychological Science*, *1*, 19–26. Retrieved from https://doi.org/10.1177/2515245917745058

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., Speckman, P. L., Sun, D., & Jiang, Y. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin and Review*, *12*, 195–223.

Rouder, J. N., Morey, R. D., & Wagenmakers, E.-J. (2016). The interplay between subjectivity, statistical practice, and psychological science. *Collabra*, *2*, 6. Retrieved from

http://doi.org/10.1525/collabra.28

Rouder, J. N., Yue, Y., Speckman, P. L., Pratte, M. S., & Province, J. M. (2010). Gradual growth vs. shape invariance in perceptual decision making. *Psychological Review*, *117*, 1267–1274.

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

Salthouse, T. A. (1996). The processing speed theory of adult age differences in cognition. *Psychological Review*, *103*, 403–428.

Simon, J. R. (1968). Effect of ear stimulated on reaction time and movement time. *Journal of Experimental Psychology*, *78*, 344–346.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Boca Raton: CRC Press.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, *15*, 72–101. Retrieved from https://www.jstor.org/stable/pdf/1412159.pdf?refreqid=excelsior%3Af2a400c0643864ecfb26464f09f022ce

Stan Development Team. (2018). RStan: The R interface to Stan. Retrieved from http://mc-stan.org/

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*, 643–662.

Unsworth, N., Schrock, J. C., & Engle, R. W. (2004). Working memory capacity and the antisaccade task: Individual differences in voluntary saccade control. *Journal of*

Experimental Psychology: Learning, Memory, and Cognition, *30*, 1302–1321.

Vanpaemel, W., & Lee, M. D. (2012). Using priors to formalize theory: Optimal attention and the generalized context model. *Psychonomic Bulletin & Review*, *19*, 1047–1056.

Von Bastian, C. C., Souza, A. S., & Gade, M. (2015). No evidence for bilingual cognitive advantages: A test of four hypotheses. *Journal of Experimental Psychology: General*, *145*(2), 246–258.

Table 1

	Obs	Individuals	Replicates	Reliability	Effect	$\hat{\sigma}$	s_d	s_θ	$\hat{\sigma}_\theta$
Stroop									
1. von Bastian	11,245	121	46.47	0	64	198	47	11	23
2. Pratte i	11,114	38	146.24	1	91	264	50	28	36
3. Pratte ii	12,565	38	165.33	0	12	160	20	8	15
4. Rey-Mermet i	48,937	264	92.68	0	54	155	30	12	19
5. Rey-Mermet ii	48,966	261	93.80	1	59	175	69	59	64
6. Hedge	43,408	53	409.51	1	69	188	32	27	29
Simon									
7. von Bastian	23,453	121	96.91	1	79	128	36	22	28
8. Pratte i	17,343	38	228.20	0	17	186	24	12	18
9. Pratte ii	12,266	38	161.39	1	30	175	30	16	22
Flanker									
10. von Bastian	11,215	121	46.34	0	2	152	32	6	15
11. Rey-Mermet i	49,300	265	93.02	0	30	147	24	6	13
12. Rey-Mermet ii	39,275	207	94.87	1	36	107	43	37	40
13. Hedge	43,384	53	409.28	1	44	100	16	13	15
Other									
14. Rouder i	11,346	52	109.10	0	50	165	28	12	19
15. Rouder ii	16,859	58	145.34	1	142	351	72	42	52
Mean	26,712	115	155.90	1	52	177	37	21	27
Median	17,343	58	109.10	1	50	165	32	13	22

Note. All sample sizes and estimates reflect cleaned data. See the Appendix for our cleaning steps which differ from those of the original authors.

A.

Participant	Trial	Condition	RT	Accuracy
1	1	1	1203	0
1	2	1	578	1
1	3	2	817	1
…	…	…	…	…

B.

Participant	Task 1 Score	Task 2 Score	Task 3 Score	Task 4 Score
1	75	120	60	70
2	100	150	30	20
3	20	10	40	50
…	…	…	…	…

C.

	Task 1	Task 2	Task 3	Task4
Task 1	1	.3	.2	.1
Task 2		1	.6	-.1
Task 3			1	.3
Task 4				1

D.

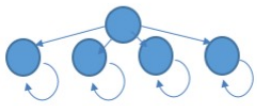

Figure 1. In the usual course of analysis, the raw data (A) are used to tabulate sample effects (B). The covariation among these task-by-person sample effects (C) then serve as input to latent variable modeling (D).

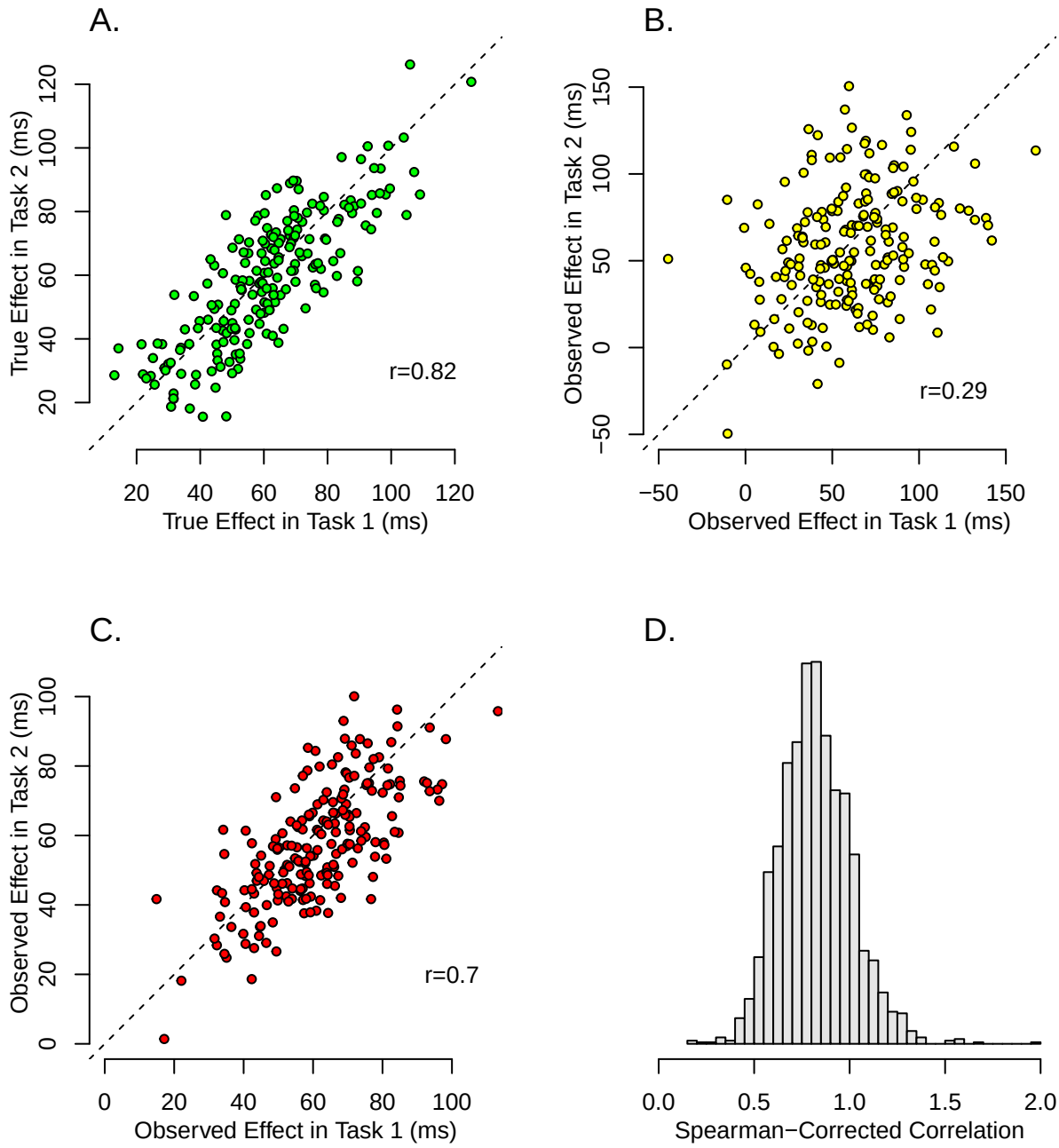

Figure 2. The effects of trial variability on the assessment of correlations among tasks. A: Hypothetical true individual effects show a large degree of correlation across two tasks. B: Observed effects are so perturbed by trial variability that the correlation is greatly attenuated. C: Hierarchical model recovery for the data in A. D: Spearman correction-for-attenuation in a small simulation with realistic settings.

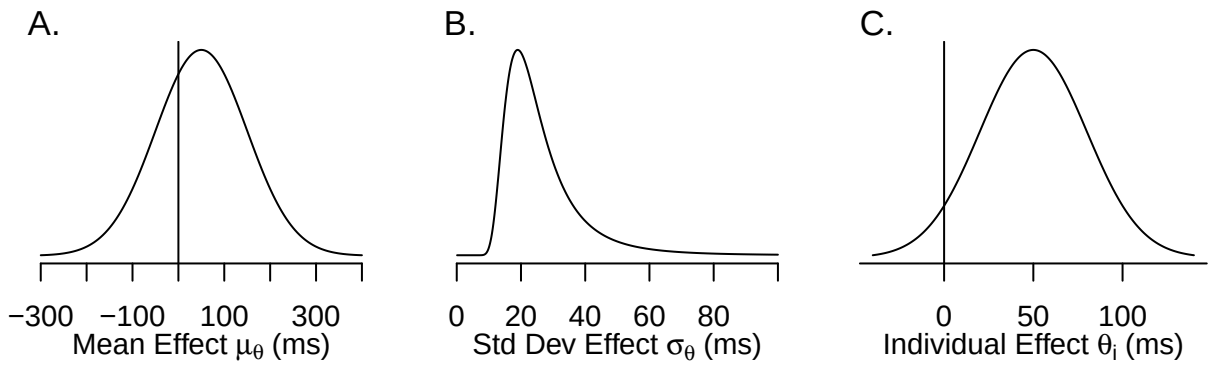

Figure 3. A, B: Prior distributions of μ_θ and σ_θ, respectively. C: Prior distribution of θ_i for $\mu_\theta = 50$ ms and $\sigma_\theta = 30$ ms.

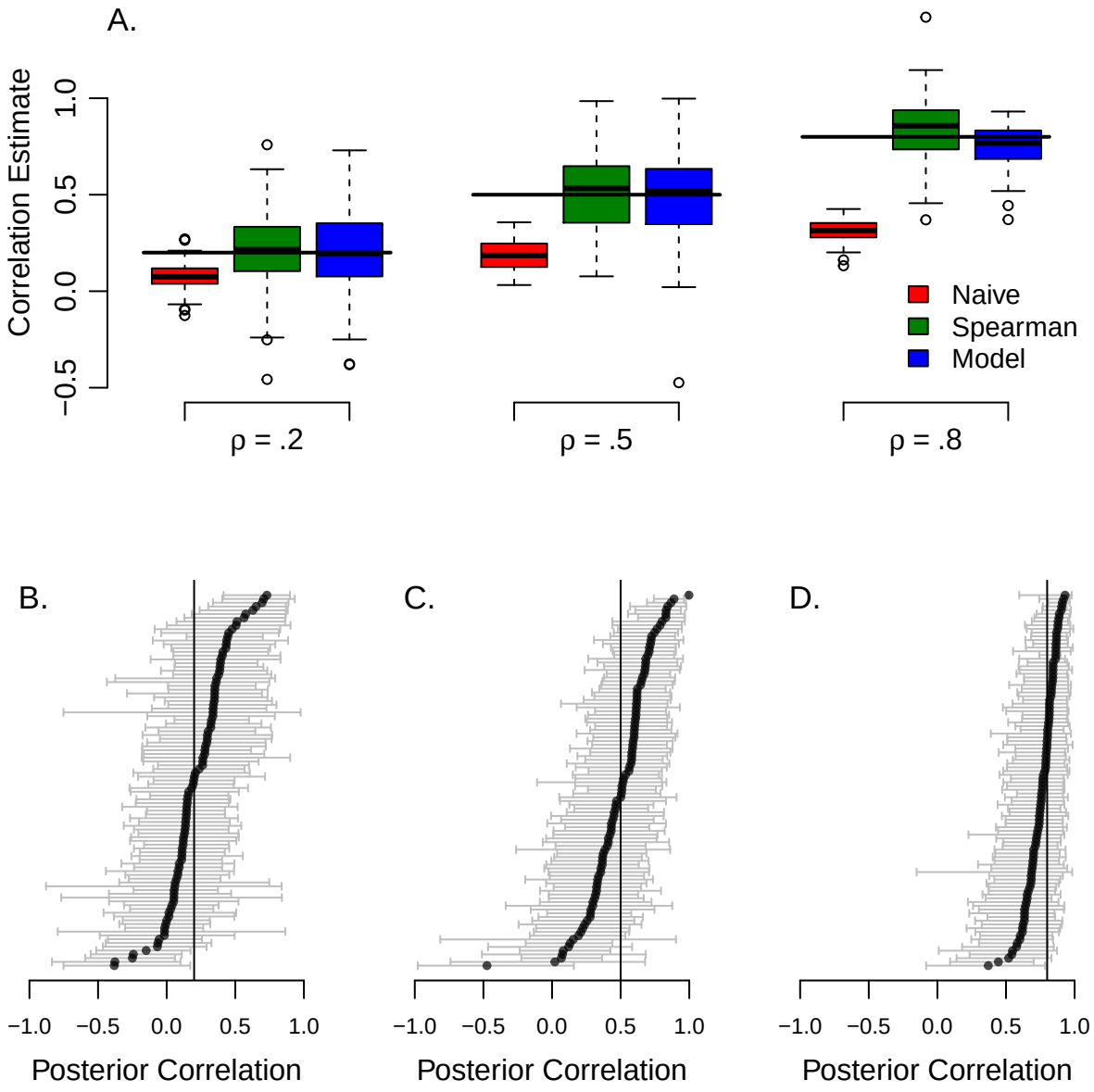

Figure 4. Revocery of correlations from two tasks. A: Boxplots of recovered correlations from naive sample correlations, Spearman's correction, and the hierarchical model. B-D: Posterior 95% credible intervals for the model-recovered correlations for true correlations of .2, .5, and .8, respectively.

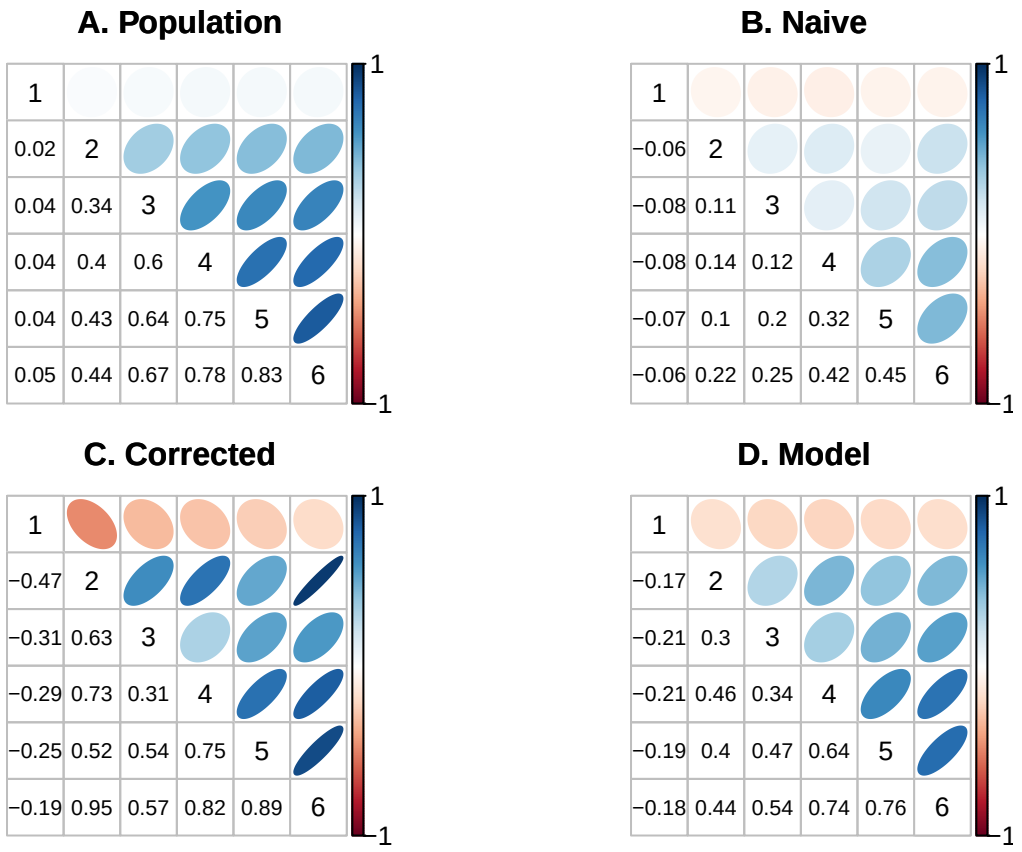

Figure 5. True and recovered correlation matrices for six tasks. A: True population correlations. B-D: Correlation estimates from a single run.

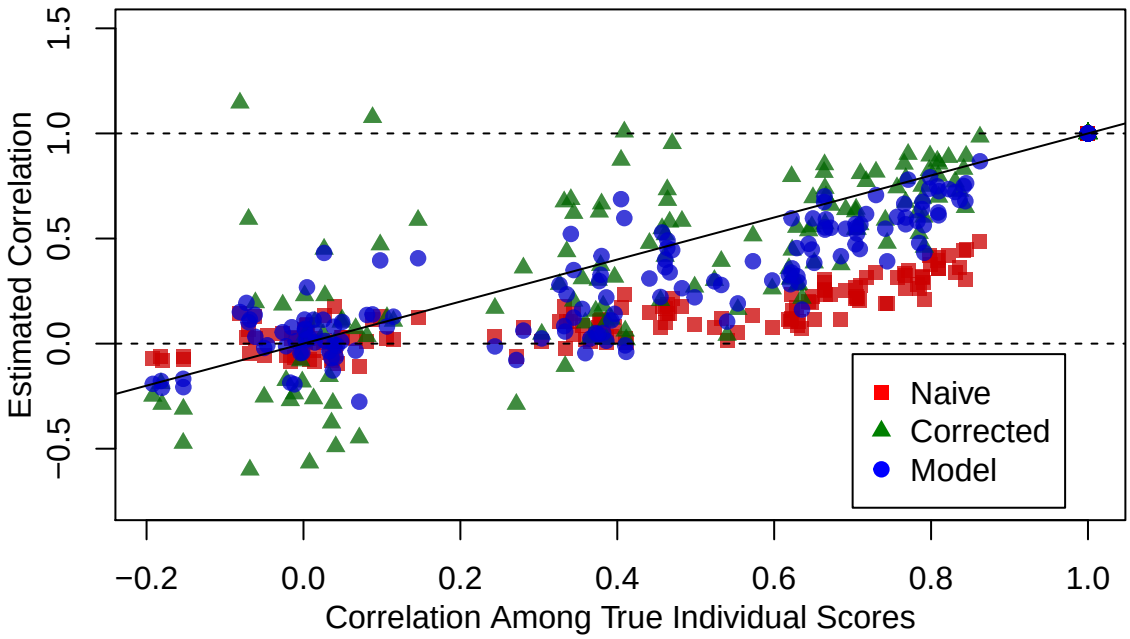

Figure 6. Recovery of correlations from six tasks. True correlations are derived from a one-factor model and are displayed in Figure 5.

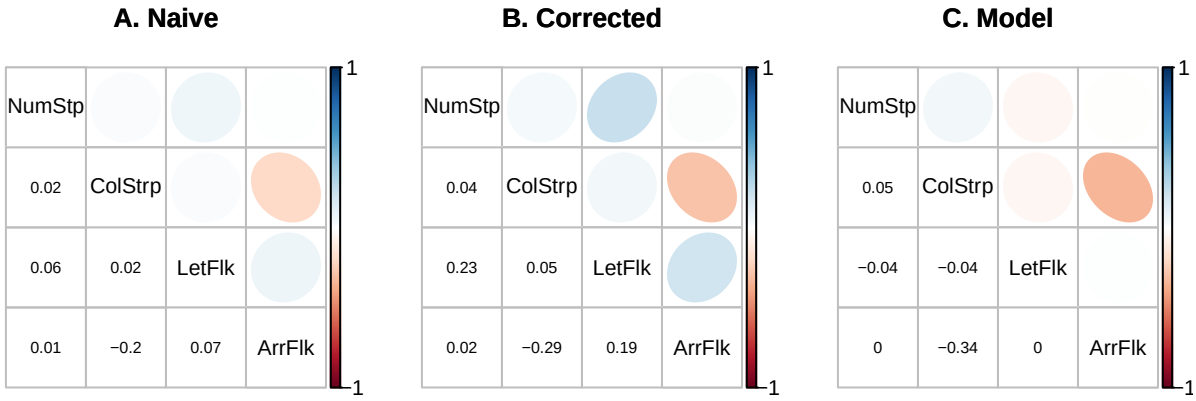

Figure 7. Correlations among select tasks in the Rey-Mermet data set. Tasks are a number Stroop task, a color Stroop task, a letter flanker task, and an arrow flanker task. Details of the tasks are provided in the Appendix.

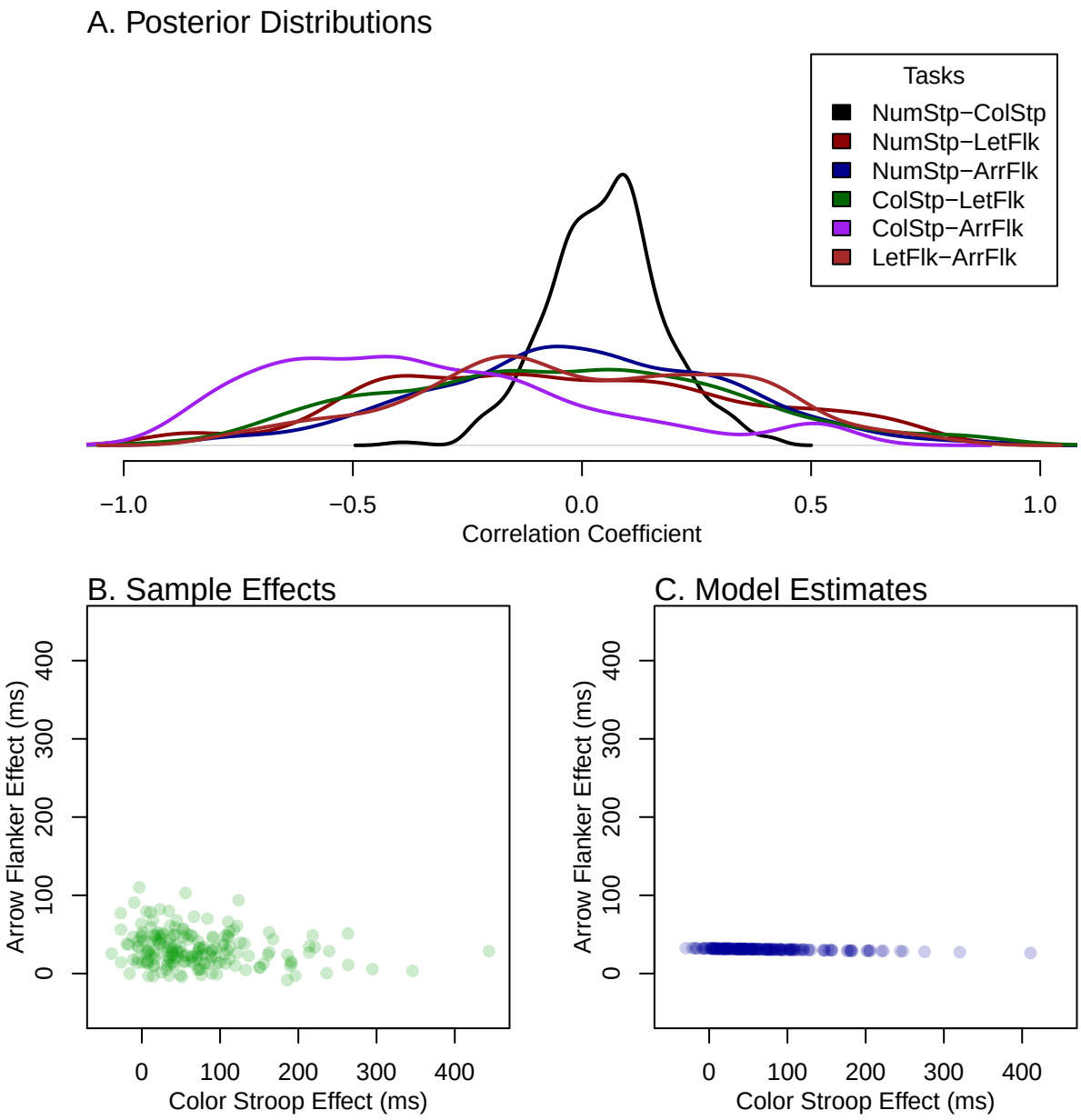

Figure 8. A. Model-based posterior distributions of population correlations among tasks. The large variance shows the difficulty of recovery. B. Individuals' sample effects for color Stroop and arrow flanker tasks show. C. Hierarchical model estimates show a large degree of shrinakge for arrow flankers but not for color Stroop reflecting the increased range of color Stroop effects.