A Psychometrics of Individual Differences in Experimental Tasks

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

[1] University of California, Irvine

[2] University of Missouri

Version 1, 3/2018

Author Note

We are indebted to Craig Hedge for making data available, working with us to insure we understand them, and providing comments on this work. The data and analysis code for this project may be found at https://github.com/PerceptionAndCognitionLab/ctx-reliability

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

Abstract

One vexing problem in cognition is the dramatic attenuation of correlations among inhibition tasks that purportedly tap similar underlying constructs. Individuals that show large Stroop effects, for example, do not show any larger flanker or Simon effects than average (Rey-Mermet et al., in press). A pressing question then is whether these attenuated correlations reflect statistical considerations, such as a lack of individual variability on tasks, or substantive considerations, such as that inhibition is not a unified concept. One problem in addressing this question is that researchers aggregate performance across trials to tally individual-by-task scores, and the covariation of these scores is subsequently studied much as it would be with classical test theory. It is tempting to think that aggregation here is fine and everything comes out in the wash, but as shown, it renders classical-test-theory concepts of reliability, effect size and correlation deeply flawed. We propose an alternative psychometrics of task performance that is based on accounting for trial-by-trial variability along with the covariation of individuals' performance across tasks. The implementation is through common Bayesian hierarchical models, and this treatment rescues classical concepts of effect size, reliability, and correlation for studying individual differences with experimental tasks. We show using recent data from Hedge et al. (2018) that there is Bayes-factor support for a lack of correlation between the Stroop and flanker task. This support for a lack of correlation indicates a psychologically relevant result—Stroop and flanker inhibition are seemingly unrelated, contradicting unified concepts of inhibition.

Keywords: Individual Differences, Inhibition, Reliability, Hierarchical Models, Bayesian Inference

A Psychometrics of Individual Differences in Experimental Tasks

In individual-differences studies, a number of variables are measured for each individual. The goal is to decompose the covariation among these variables into a lower-dimensional, theoretically-relevant structure (Bollen, 1989; Skrondal & Rabe-Hesketh, 2004). Critical in this endeavor is understanding the psychometric properties of the measurements. Broadly speaking, variables used in individual-difference studies come from the following three classes: The first is the class of rather natural and easy-to-measure variables such age, weight, and gender. The second is the class of *instruments* such as personality and psychopathology instruments. Instruments have a fixed battery of questions and a fixed scoring algorithm. Most instruments have been benchmarked, and their reliability has been well established. The final class of variables is performance on experimental tasks. These experimental tasks are often used to assess cognitive abilities in memory, attention, and perception.

On the face of it, individual-difference researchers should be sanguine about using scores from experimental tasks. For one, many of these tasks are robust in that the effects are easy to obtain in a variety of circumstances. Take, for example, the Stroop task, which may be used as a measure of inhibition. The Stroop effect is so robust, it is considered universal (Haaf & Rouder, 2017; MacLeod, 1991). Another advantage is that many of these tasks have been designed to isolate specific cognitive processes. The Stroop task, for example, requires participants to inhibit the prepotent process of reading. 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. It is no wonder they have become popular in the study of individual differences.

Yet, there has been a wrinkle in the setup. As it turns out, these task-based measures correlate with one another far less than one might think *a priori.* The best example of these wrinkles is perhaps in the individual-differences study of inhibition tasks (Friedman & Miyake, 2004; Ito et al., 2015; Pettigrew & Martin, 2014; Rey-Mermet, Gade, & Oberauer,

2018; Stahl et al., 2014). In these large-scale individual-differences studies, researchers correlated scores in a variety of tasks that require inhibitory control. An example of two such tasks are the Stroop task (Stroop, 1935) and the flanker task (Eriksen & Eriksen, 1974). Correlations among inhibition tasks are notoriously low; most do not exceed .2 in value. The Stroop and flanker tasks, in particular, seemingly do not correlate at all (Hedge, Powell, & Sumner, 2018; Rey-Mermet et al., 2018; Stahl et al., 2014). These 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.

Why correlations are so low among these tasks is a mystery (Hedge et al., 2018; Rey-Mermet et al., 2018). After all, these tasks are robust, easy-to-obtain, and seemingly measure the same basic concepts. There are perhaps two leading explanations: One is that the problem is mostly statistical. High correlations may only be obtained if the tasks are highly reliable, and low correlations are not intepretable without high reliability. Hedge et al. document a range of test-retest reliabilities across tasks, with most tasks having reliabilities below .7 in value, and some tasks having reliabilities below .4 in value. In effect, the problem is that people simply don't vary enough in some tasks given the resolution of the data to document correlations. The second explanation is that the lack of correlation is substantive. The presence of low correlations in large-sampled studies means that the tasks are truly measuring different sources of variation. In the inhibition case, the ubiquity of low correlations in large-sample studies has led Rey-Mermet et al. (2018) to make a substantive claim that the psychological concept of inhibition is not unified at all.

Researchers typically use classical test theory, at least implicitly, when studying individual differences. They make this choice when they tally up an individual's data into a performance score. For example, in the Stroop task, we may score each person's performance by the mean difference in response times between the incongruent and congruent trials. Because there is a single performance score per person per task, tasks are very much treated

as instruments. When trials are aggregated into a single score, the instrument (or experiment) is a *test* and classical test theory serves as the analytic framework.

The unit of analysis in classical test theory is the test, or in our case, the experimental task. Reliability and effect size, for example, are assigned to the task itself and not to a particular sample or a particular sample size. For instance, when a test developer states that an instrument has test-retest reliability of .80, that value holds as a population truth for all samples. We may see more noise if we re-estimate that value in smaller samples, but, on whole, the underlying population value of an instrument does not depend on the researcher's sample size. We call this desirable property *portability.*

We may contrast classical test theory to item-response theory (Lord & Novick, 1968). In item-response theory, the main unit of analysis is the item rather than the test. Properties of the test, say reliability and correlation, are replaced with properties of items, say item difficulty and item disciminability. Yet, for tasks, where items are trials, item difficulty and discriminability are not as helpful as concepts of reliability and correlation. It is no surprise, therefore, that researchers tend to use classical test theory, with its focus on the task rather than the trial, to conceptualize individual differences in tasks.

The main problem with using classical test theory, however, is that portability is grossly violated. Concepts of reliability and effect sizes within a task and correlations across tasks are critical functions of sample sizes (Green et al., 2016). As a result, different researchers will be estimating different truths when they invariably have different sample sizes. We show how dramatically portability is violated in practice, and how much these violations affect the interpretability of classical statistics.

In this paper, we gain portability in experimental tasks by using a straightforward hierarchical model to account simultaneously for trial-by-trial variability, session variability, and individuals' variability. The outputs are the estimates of individuals' effects, variability in these effects, and shared covariation of these effects across tasks. These outputs are portable across different sample sizes including different numbers of trials per participant

and different numbers of participants. Hence, they preserve their meaning across different studies and are broadly interpretable. This enhanced interpretability licenses their use in addressing the low-correlation mystery.

The application of hierarchical models to trial-by-trial performance data in experimental tasks is not new. Previous applications in individual-differences research include Schmiedek, Oberauer, Wilhelm, Süß, & Wittmann (2007) and Voelkle, Brose, Schmiedek, & Lindenberger (2014). Moreover, trial-by-trial hierarchical modeling is well known in cognitive psychology (Lee & Webb, 2005; Rouder & Lu, 2005) and linguistics (Baayen, Tweedie, & Schreuder, 2002). That said, to our knowledge, using hierarchical models to make classical test-theory concepts portable in experimental settings is novel. It is tasty, low-hanging fruit on the methodological tree.

The Dramatic Failure of Portability

In classical test theory, an *instrument*, say a depression inventory, is a fixed unit. It has a fixed number of questions that are often given in a specific order. When we speak of portability, we speak of porting characteristics of this fixed unit to other settings, usually to other sample sizes in similar populations. In experimental tasks, we have several different sample sizes. One is the number of individuals in the sample; others are the numbers of trials each individual completed within the various conditions in the task. We wish to gain portability across both of these sample-size dimensions. For example, suppose one team runs 50 people each observing 50 trials in each condition and another team runs 100 people each observing 25 trials in each condition. If measures of reliability and effect size are to be considered a property of a task, then these properties should be relatively stable across both teams. Moreover, if performance correlated with other variables, portability would mean that the correlation is stable as well.

But standard psychometric measures that are portable for instruments fail dramatically on tasks. To show this failure, we reanalyze data from a few tasks from Hedge

et al. (2018). Hedge et al. compiled an impressively large data set. They asked individuals to perform an exceptionally large number of trials on several tasks. For example, in their Stroop task, participants completed a total of 1440 trials each. The usual number for individual-differences studies is on the order of 10s or 100s of trials each. Moreover, Hedge et al. explicitly set up their design to measure test-retest reliability. Individuals performed 720 trials in the first session, and then, three-weeks later, performed the remaining 720 trials in a second session. Here, we use the Hedge et al. data set to assess conventional measures of effect size and reliability are dependent on the number of trials per condition.

The following notation is helpful for specifying the conventional approach to effect size and reliability. Let $\bar{Y}_{ijk}$ be the mean response time for the ith individual in the jth session (either the first or second) in the kth condition. For the Stroop task, the conditions are congruent ($k = 1$) or incongruent ($k = 2$). Effects in a session are denoted d_{ij}. These are the differences between mean incongruent and congruent response times:

$$d_{ij} = \bar{Y}_{ij2} - \bar{Y}_{ij1}. \tag{1}$$

The effect per individual can be averaged across sessions and denoted $\bar{d}_i$. The grand mean effect is just the mean of these individual effects; the effect size is just the ratio of this grand mean to the standard deviation of these effects, i.e., es=mean($\bar{d}_i$)/sd($\bar{d}_i$). The test-retest reliability is the correlation of individuals' effect scores for the first session ($j = 1$) to the second session ($j = 2$).

Figure 1 shows a lack of portability in these measures. We randomly selected different subsets of the Hedge et al.'s data and computed effect sizes and test-retest reliabilities. We varied the number of trials per condition per individual from 20 to 400, and for each level, we collected 100 different randomly-chosen subsets. For each subset, we computed effect size and reliability, and plotted are the means across these 100 subsets. The line denoted "S" shows the case for the Stroop task, and both effect size and reliability measures increase appreciably with the number of trials per condition per individual. The line denoted "F" is from Hedge

et al.'s flanker task, and the results are similar. In fact, these increases with the number of replicates are a general property of tasks. Measures that are treated as properties of the task are critically tied to sample sizes. In summary, classical measures are portable when applied to instruments because the numbers of items are fixed. They are importable when applied to task performance where the number of trials per individual will vary across studies.

We think the above results may be surprising. The critical and outsized role of replicates is seemingly under appreciated. Many researchers are quick to highlight the numbers of individuals in studies. You may find this number repeatedly in tables, abstracts, method sections, and sometimes in general discussions. Researchers do not, however, highlight the number of replicates per condition per individual. These numbers rarely appear in abstracts or tables; in fact, it usually takes careful hunting to find them in the method section if they are there at all. And researchers are far less likely to discuss the numbers of replicates in interpreting their results. And yet, as shown above, this number of replicates is absolutely critical in understanding classical results.

One approach is to ask, "what is an appropriate number of replicates?" (Hedge et al., 2018). This approach, however, strikes us as unfortunate for studying individual differences because whatever the number, the meaning of reliability, correlation, and effect size is still tied to that number. Instead, a better approach is to strive for portability—the meaning of reliability, correlation and effect size should be independent of the number of replicates. We use simple hierarchical statistical models to make these quantities portable. Our approach is to consider the effects sizes, reliabilities, and correlations in the large-sample limit of trials per individual. While in reality individuals perform finite trials, we may use statistical models to query what may reasonably happen in the limit. And estimates of these values in the limit become portable estimates of effect sizes, reliabilities, and correlations. Low numbers of replicates certainly affect the quality of the estimates, but the true value—the value being estimated—does not change with the number of replicates.

A Hierarchical Model

Portability may be gained by using rather ordinary hierarchical models to account for trial-by-trial variation and individual variation simultaneously.

Model specification

The following univariate notation is used: Subscripts are used to denote individuals, tasks, conditions, and replicates. Let $Y_{ijk\ell}$ denote the ℓth observation for the ith individual in the jth task and kth condition. Observations are usually performance variables, and in our case, and for concreteness, they can be response times on trials. For now, we model response times in just one task, and in this case, the subscript j may be omitted. Consider a trial-level base model:

$$Y_{ik\ell} \sim \text{Normal}(\mu_{ik}, \sigma^2),$$

where μ_{ik} is the true mean response time of the ith person in the kth condition and σ^2 is the true trial-by-trial variability. It is important to differentiate between true parameter values like μ_{ik} and their sample estimates.

We develop this model for the Stroop task. In this task and tasks like it, the key contrast is between the congruent ($k = 1$) and incongruent ($k = 2$) conditions. This contrast is embedded in a model where each individual has an average speed effect, denoted α_i, and a Stroop effect, denoted θ_i:

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

where $x_1 = -1/2$ and $x_2 = 1/2$.

The goal then is to study θ_i, the ith person's Stroop effect size. In modern mixed models, individual's parameters α_i and θ_i are considered latent traits for the ith person, and

are modeled as random effects:

$$\alpha_i \sim \text{Normal}(\mu_\alpha, \sigma^2_\alpha,$$
$$\theta_i \sim \text{Normal}(\mu_\theta, \sigma^2_\theta),$$

where μ_α and μ_θ are population means and σ^2_α and σ^2_θ are population variances.

The above mixed model is simple and it may be implemented in frequentist or Bayesian frameworks. We use the Bayesian framework here because we are more comfortable with the intellectual commitments made in Bayesian probability than in frequentist probability (see, for example, Edwards, Lindman, & Savage, 1963). For estimation, the choice of framework hardly matters, and the resulting analyses are by-and-large the same regardless of how one implements them. The key feature here is the hierarchical or mixed nature of the model. This feature is critical for gaining portability, and the portability from the hierarchical specification holds in both Bayesian and frequentist contexts.[1]

Hierarchical Regularization and the Portability of Effect Size

The conventional analysis of the Stroop task centers on sample means aggregated over trials. The critical quantity is $d_i = \bar{Y}_{i2} - \bar{Y}_{i1}$, the observed Stroop effect for the ith individual. How does d_i compare to model-based estimates of effects, θ_i? Figure 2 shows the comparison for a subset of Stroop data in Hedge et al. (2018). We see that the Bayesian estimates are smoother and far less variable than sample means. This is an example of the benefits of

[1]The full Bayesian specification of the model comes from Haaf & Rouder (2017) and is as follows. Let $g_\alpha = \sigma^2_\alpha/\sigma^2$ and $g_\theta = \sigma^2_\theta/\sigma^2$. Then:

$$\begin{aligned}\pi(\mu_\alpha, \sigma^2) &\propto 1/\sigma^2,\\ \mu_\theta &\sim \text{Normal}(0, g_{\mu_\theta}\sigma^2),\\ g_\alpha &\sim \text{inverse-}\chi^2(1, r^2_\alpha),\\ g_\theta &\sim \text{inverse-}\chi^2(1, r^2_\theta),\\ g_{\mu_\theta} &\sim \text{inverse-}\chi^2(1, r^2_{\mu_\theta}),\end{aligned}$$

where inverse-χ^2(a,b) is a scaled inverse chi-squared distribution with a degrees-of-freedom and a scale of b (see Gelman, Carlin, Stern, & Rubin, 2004). The following settings are used $r_\alpha = 1$, $r_{\mu_\theta} = 0.25$, and $r_\theta = 0.17$. The prior settings have only marginal effects on parameter estimates as the priors are relatively diffuse and the data are numerous.

regularization from hierarchical models (Efron & Morris, 1977), and they hold broadly (Gelman et al., 2004; Lehmann & Casella, 1998).

Just because the model-based estimates are smoother and there is less individual variation doesn't mean they are better. The better estimates are the ones closer to the true latent values! Here, we follow the demonstration of Efron & Morris (1977), a delightful paper that illustrated the gain in accuracy with regularization from hierarchical models. Efron & Morris (1977) is written for a general audience, and the goal was to predict the baseball hitting averages of select players using their first 50 at bats. In one case, Efron and Morris used the observed proportion of hits after 50 at bats, in the other they used a hierarchical model similar to ours above. They found that the hierarchical model substantially outperformed the observed-proportion approach in predicting the end-of-season batting average for each player.

We can do the same with Hedge et al.'s Stroop data. The estimates in Figure 2 were from a single block of 144 trials per participant. Hedge et al. ran 10 such blocks distributed across two sessions, and we may ask how well the estimates from one of the blocks predict the whole 10-block set. Figure 3, the style of which is borrowed from Efron and Morris, shows the sample estimates from one block (bottom row), the regularized model estimates from the same block (middle row), and the sample estimates from all ten blocks (top row). The shrinkage toward the grand mean in hierarchical models is evident, and it is about right to bring the variability of the model-based estimates from one block in line with that from all ten blocks.

The model-based one-block estimators are better predictors of the overall data than are the sample one-block estimators. This improvement may be formalized by a root-mean-squared error measure. The error is about 1.60 times larger for the sample means than for the hierarchical model estimates. This benefit is general—hierarchical models provide for more accurate estimates of individuals latent abilities (James & Stein, 1961).

Figure 3 provides insight into the importability of sample effect size measures. The

problem here is that the variability of individual sample effects are too large. That means the standard deviation estimate is too large, and the resulting effect-size estimates are too small. The model based estimates are regularized so that the scale of individual effects is about right. Hence, the effect size estimate is not systematically biased.

We can formalize the dynamics in play. It is helpful to express the distribution of sample effects, d_i in model parameters.

$$d_i \sim \text{Normal}(\mu_\theta, 2\sigma^2/L + \sigma_\theta^2).$$

Classical effect size measure, $\text{mean}(d)/\text{sd}(d)$, are measuring $\mu_\theta/\sqrt{2\sigma^2/L + \sigma_\theta^2}$. The problem is the inclusion of $2\sigma^2/L$, which is nuisance trial variation. Not only does this included nuisance trial variation result in individual effect estimates that are too variable and in effect size measures that are too small, it violates portability as there is an explicit dependence on L, a sample-size measure. In the model-based approach, the distribution of θ_i is:

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

and the effect size calculated from these individual effects are the correct, portable, quantity.

It is now clear what the hierarchical model does—it accounts for trial-by-trial variation. By doing so, results are portable to designs with varying numbers of individuals and trials per individual. These results are estimates of underlying properties of tasks.

A Two-Task Model for Reliability and Correlation

The hierarchical model may be expanded to account for two tasks or two session simultaneously. For two sessions, the goal is to estimate a portable measure of test-retest reliability; for two tasks, the goal is to measure a portable estimate of correlation. Let's take reliability first. The Hedge et al. data set, which we highlight here, has a novel feature. These researchers sought to measure the test re-test reliability of several cognitive tasks.

They had individuals perform 720 trials of a task one day, and three weeks later the individuals returned and performed another 720 trials. We let the subscript $j = 1, 2$ index the session. The trial-level model is expanded to:

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

Here, we simply expand the parameters to hold for individual-by-session combinations.

The parameters of interest are θ_{ij} and there are several specifications that may be made here. We start with the most general of these, an additive-components decomposition into common and oppositional components:

$$\theta_{i1} = \nu_1 + \omega_i - \gamma_i,$$
$$\theta_{i2} = \nu_2 + \omega_i + \gamma_i.$$

The parameter ν_j is the main effect of the jth session, and by having separate main effect parameters for each session, the model captures the possibility of a systematic effect of session on the Stroop effect. The parameter ω_i is the common effect of the ith individual; individuals that have large Stroop effects on both sessions have high values of ω. The parameter, γ_i, is the oppositional component. It captures idiosyncratic deviations where one individual may have a large Stroop effect in one session and a smaller one in another. Whereas these individual common effects and oppositional effects are random effects, we place a hierarchical constraint on them:

$$\omega_i \sim \text{Normal}(0, \sigma^2_\omega),$$
$$\gamma_i \sim \text{Normal}(0, \sigma^2_\gamma).$$

In the previous section, we constructed portable measures by focusing on the distribution of θ, individuals' true effects. Here we do the same. To gain an expression for the reliability of a task, it is helpful to express the multivariate distribution of the θs. The

easiest way to do this is to write out the distribution for two individuals' performance across two sessions:

$$\begin{bmatrix} \theta_{11} \\ \theta_{12} \\ \theta_{21} \\ \theta_{22} \end{bmatrix} \sim \mathrm{N}_4 \left(\begin{bmatrix} \nu_1 \\ \nu_2 \\ \nu_1 \\ \nu_2 \end{bmatrix}, \begin{bmatrix} \sigma^2_\omega + \sigma^2_\gamma & \sigma^2_\omega - \sigma^2_\gamma & 0 & 0 \\ \sigma^2_\omega - \sigma^2_\gamma & \sigma^2_\omega + \sigma^2_\gamma & 0 & 0 \\ 0 & 0 & \sigma^2_\omega + \sigma^2_\gamma & \sigma^2_\omega - \sigma^2_\gamma \\ 0 & 0 & \sigma^2_\omega - \sigma^2_\gamma & \sigma^2_\omega + \sigma^2_\gamma \end{bmatrix} \right).$$

From these distributions, it follows that the test-retest reliability, the correlation across θ_{i1} and θ_{i2}, is

$$\rho = \frac{\sigma^2_\omega - \sigma^2_\gamma}{\sigma^2_\omega + \sigma^2_\gamma}$$

Importantly, this reliability is portable as it does not include trial-by-trial variability.

To see where the importability of sample reliability comes from, we start with the equivalent multivariate distribution for sample effects:

$$\begin{bmatrix} d_{11} \\ d_{12} \\ d_{21} \\ d_{22} \end{bmatrix} \sim \mathrm{N}_4 \left(\begin{bmatrix} \nu_1 \\ \nu_2 \\ \nu_1 \\ \nu_2 \end{bmatrix}, \begin{bmatrix} \sigma^2_\omega + \sigma^2_\gamma + 2\sigma^2/L & \sigma^2_\omega - \sigma^2_\gamma & 0 & 0 \\ \sigma^2_\omega - \sigma^2_\gamma & \sigma^2_\omega + \sigma^2_\gamma + 2\sigma^2/L & 0 & 0 \\ 0 & 0 & \sigma^2_\omega + \sigma^2_\gamma + 2\sigma^2/L & \sigma^2_\omega - \sigma^2_\gamma \\ 0 & 0 & \sigma^2_\omega - \sigma^2_\gamma & \sigma^2_\omega + \sigma^2_\gamma + 2\sigma^2/L \end{bmatrix} \right).$$

From this distribution, it is clear that the sample correlation between sample effects is estimating $(\sigma^2_\omega - \sigma_\gamma)/(\sigma^2_\omega + \sigma^2_\gamma + 2\sigma^2/L)$. It is the added sample noise in the denominator, σ^2/L, that renders the sample test-retest reliability importable.

Figure 4 provides a real-data comparison of model-based and sample test-retest reliabilities. The top row is for the single block of Hedge et al.'s Stroop task; the bottom row is for all ten blocks of the same data set. The first column are scatter plots of the sample effects of the first session vs. the second session. There is almost no test-retest correlation using a single block of data (A); but when all data are considered there is a moderate

correlation (B). This pattern is the same as in Figure 1 where reliability was diminished for smaller numbers of trials. The model estimates of individual effects are shown in the middle column. Here, there is shrinkage, especially for the single block (C,D). This shrinkage, however, is to a line rather than to a point in the center. This pattern reflects the model specification where positive and negative correlations are explicitly modeled. The last column shows the posterior distributions of the test-retest reliability coefficient. Here, for the single-block data there is much uncertainty (E). The uncertainty shrinks as the sample sizes grow, as indicated by the same plot for all the data (F). The posterior mean, a point-estimate for the test-retest reliability of the Stroop task, is 0.72. We also analyzed the flanker task, and the test-retest reliability of this this task is 0.68. With these relatively reasonable levels of reliability, there is sufficient resolution to explore the correlation across the two tasks.

Figure 4 highlights the dramatic difference between conventional and hierarchical-model analysis, especially for smaller numbers of trials per participant. If one uses the conventional model for the reliability for one block of data, then the correlation coefficient appears small, 0.10, and somewhat well localized (the 95% confidence interval is from-0.17 to 0.36). We emphasize this result is misguided. When all the data are used, the conventional correlation coefficient is 0.55, which is well outside this 95% confidence interval. Contrast this result to that from the hierarchical model. Here, not only is the correlation coefficient larger, 0.31), but the uncertainty is quite large as well (the 95% credible interval is from -0.31 to 0.80). Moreover, the value of the correlation from the hierarchical model with all of the data, 0.72, is within this credible interval. In summary, using the conventional analysis results in over confidence in a wrong answer. The hierarchical model, however, tempers this overconfidence by accounting for trial-by-trial uncertainty as well as uncertainty across individuals.

Correlation between tasks may be handled with the same machinery. Here we compare Stroop task performance to flanker task performance. Each of Hedge et al's participants ran in both tasks. To correlate tasks, we combined data across sessions and fit the model where j indexes task rather than session. The results are shown in Figure 5. As can be seen, there

appears to be no correlation. Inference about the lack of correlation will be made in the next section where model comparison is discussed.

Model Comparison

The above analyses were focused on parameter estimation. With estimation, the targets are the effect sizes and reliability of tasks, and the correlation among tasks. Model-based estimation provided here are portable analogs to sample-based measures of effect size, reliability, and correlation. The difference is that they account for variation at the trial level, and consequently, may be ported to designs with varying sample sizes.

Researchers, however, are often interested in stating evidence for theoretically meaningful propositions. In the next section, we describe a set of theoretically meaningful propositions and their model implementation. Following this, we present a Bayes factor method of model comparison.

Theoretical Positions and Model Implementation

When assessing the relationship between two tasks, the main target is the correlation. The above estimation model is helpful for estimating the correlation, but often researchers are interested in the following statements: The first one is about a lack of correlation among two tasks. A lack of correlation is a necessary condition for independence, and if there is evidence for a lack of correlation, then independence is plausible. The state is full correlation among the tasks. If there is full correlation, then there is evidence that both tasks are measuring a single dimension or ability.

In the preceding section, we presented an estimation model, which we now call the *general model*. The critical specification is that of θ_{ij}, the individual-by-task effect. We

modeled these as:

$$\begin{aligned}\mathcal{M}_g: \quad \theta_{ij} &= \nu_j + \omega_i + u_j\gamma_i,\\ \omega_i &\sim \text{Normal}(0, \sigma^2_\omega),\\ \gamma_i &\sim \text{Normal}(0, \sigma^2_\gamma),\end{aligned}$$

where $u = (-1, 1)$ for the two tasks. In this model, the correlation among an individuals reflects the variability of ω and γ. All values of correlation on the open interval (-1,1) are possible. Full correlation is not possible, and there is no special credence given to no correlation. To represent those positions we develop alternative models on θij.

A no-correlation model is given by putting uncorrelated noise on θ_{ij}:

$$\mathcal{M}_0: \quad \theta_{ij} \sim \text{Normal}(\nu_j, \sigma^2_\theta).$$

The no-correlation and the general models provide for different constraints. The general model has regularization to a regression line reflected by the balance of the variabilities of ω and γ. The no-correlation has regularization to the point (ν_1, ν_2).

A full correlation model is given by simply omitting the γ parameters in the general model.

$$\begin{aligned}\mathcal{M}_1: \quad \theta_{ij} &= \nu_j + \omega_i,\\ \omega_i &\sim \text{Normal}(0, \sigma^2_\theta)\end{aligned}$$

Here, there is a single random parameter, ω_i for both tasks.

Before continuing to model comparison, a more in-depth consideration of priors on parameters is needed. In Bayesian analysis, priors are needed for all parameters. In the preceding sections, we de-emphasized the role of the prior. For estimating parameters, the influence of the priors quickly diminishes with increasing sample sizes (Lee, 1997). In our context, we had a large number of observations, and the reported estimates are very robust to large changes in prior settings. For model comparison, however, the story changes markedly. Priors play a critical role and must be set judiciously. In the Appendix, under the

section on Prior Specification, we discuss the role of prior settings and justify our choices. We encourage all readers to read the Appendix though it is not necessary to continue.

Bayes Factor

In Bayesian analysis, it is possible to use Bayes' rule to update beliefs about the plausibility of models themselves. Let $\mathcal{M}_a$ and $\mathcal{M}_b$ be two models under consideration and let $\boldsymbol{Y}$ denote a collection of observations. The relevant equation is:

$$\frac{P(\mathcal{M}_a|\boldsymbol{Y})}{P(\mathcal{M}_b|\boldsymbol{Y})} = \frac{P(\boldsymbol{Y}|\mathcal{M}_a)}{P(\boldsymbol{Y}|\mathcal{M}_b)} \times \frac{P(\mathcal{M}_a)}{P(\mathcal{M}_b)}.$$

Here $P(\mathcal{M}_a|\boldsymbol{Y})/P(\mathcal{M}_b|\boldsymbol{Y})$ are the posterior odds on the models; $P(\mathcal{M}_a)/P(\mathcal{M}_b)$ are the prior odds, and $P(\boldsymbol{Y}|\mathcal{M}_a)/P(\boldsymbol{Y}|\mathcal{M}_b)$ is the Bayes factor. This factor describes how odds should be updated in light of data. A growing chorus of psychologists and statisticians argue that the Bayes factor is ideal for scientific communication because it is the updating factor regardless of whatever prior odds are held (Edwards et al., 1963; Jeffreys, 1961; Myung & Pitt, 1997; Wagenmakers, 2007). The Bayes factor may also be thought of as describing how well the model predicts the observed data (Morey, Romeijn, & Rouder, 2016; Rouder, Haaf, & Aust, 2018; Rouder, Morey, & Wagenmakers, 2016). Models in the Bayesian context provide a predictive probability distribution across the data, that is, they describe where the data should lie before seeing them. Then, we assess how well each model predicted the observed data.

Table 1 shows the Bayes factor results for a few data sets from Hedge et al. (2018). The top two rows are for the Stroop and flanker data, and the correlation being tested is the test-retest reliability. The posterior mean of the correlation coefficients are 0.72 and 0.68, for the Stroop and flanker tasks, respectively. The Bayes factors confirm that there is ample evidence that the correlation is neither null nor full. Hence, we may conclude that there is indeed some though not a lot of added variability between the first and second sessions in these tasks. The next row shows the correlation between the two tasks. Here, the posterior

mean of the correlation coefficient is -0.06 and the Bayes factors confirm that the no-correlation model is preferred. The final row is a demonstration of the utility of the approach for finding dimension reductions. Here, we split the flanker task data in half by odd and even trials rather than by sessions. We then submitted these two sets to the model, and calculated the correlation. It was quite high of course, and the posterior mean of the correlation was 0.82. The Bayes factor analysis concurred, and the full-correlation model was favored by 31-to-1 over the general model, the nearest competitor.

The Appendix provides the prior settings for the above analyses. It also provides a series of alternative settings for assessing how sensitive Bayes factors are to reasonable variation in priors. With these alternative settings, the Bayes factors attain different values. Table 2 in the Appendix shows the range of Bayes factors corresponding to these alternative settings. This table provides context for understanding the limits of the data and the diversity of opinion they support.

General Discussion

In this paper we examined classical test theory analysis of experimental tasks. In the classical-test-theory framework, conventional sample measures are not portable because they are estimating quantities contaminated by removable trial-by-trial variation. With this contamination, effect sizes reliabilities, and correlations are dramatically too low. The main innovation here is modeling trial-by-trial variation along with individuals' covariation across tasks. Concepts such as effect size, reliability, and correlation are portable when defined in the asymptotic limit of unbounded trials per individual. In the current models, performance in these asymptotic limits are explicit model parameters.

With this development, it is possible to assess whether observed low correlations across tasks reflect low reliability or a true lack of association. We examine this problem for the Stroop and flanker task data reported in Hedge et al. (2018). We find that there is relatively high test-retest reliability for both tasks. This high reliability allows for the interpretation of

the correlation between Stroop and flanker tasks. There is direct evidence for a null correlation.

The main difficulty in classical analysis occurs when researchers aggregate across trials to form individual-by-task scores. We recommend that researcher avoid this aggregation. Instead, they should extend hierarchical models to the trial level, and by doing so, classical concepts of reliability, effect size, and correlation remain meaningful and portable. Individual-difference researchers are intimately familiar with mixed linear models, and these are used regularly to decompose variability. Adding one additional level, the trial level, is conceptually trivial and computationally straightforward. Indeed, modeling at this level is common in high-stakes testing (IRT, Lord & Novick, 1968), cognition (Lee & Webb, 2005; Rouder & Lu, 2005), and linguistics (Baayen et al., 2002).

Bayesian Modeling

The models presented here are simple hierarchical models, and consequently, they may be analyzed in conventional or Bayesian frameworks. For parameter estimation, there is often little difference between Bayesian approaches and more classical ones. Bayesian approaches make latent variable modeling a bit simpler and more conceptually straightforward (Gelman et al., 2004), but the regularized estimates from hierarchical models behave similarly in both frameworks (Lehmann & Casella, 1998).

The big difference between conventional and Bayesian latent-variable modeling, at least for the methods we ascribe to, is in model comparison. There are several different Bayesian approaches to model comparison. We advocate comparison by Bayes factor, but other approaches include using the coverage of credible intervals (Kruschke & Liddell, 2017), comparison by deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & Linde, 2002), and comparison through cross validation (Vehtari, Gelman, & Gabry, 2017). The reason we prefer Bayes factor is that it follows directly from Bayes' rule (Laplace, 1986) and uniquely guarantees rational updating of the plausibility of models in light of data (de

Finetti, 1974). Other methods necessarily do not update rationally. Instead, they meet other desiderata, mostly focused either on out-of-sample considerations or minimizing the influence of prior settings. The debate among Bayesians about model comparison is certainly vigorous and well litigated. As a matter of principle, we advocate Bayes factor as a rational updater. Simultaneously, we respect that others may invoke alternative desiderata, and we appreciate their critiques. Those who use Bayes factor should be prepared to admit a diversity of opinion that reflects the range of reasonable prior settings, as we do in the Appendix.

More Advanced Task Models

The field of individual differences has moved far beyond the consideration of two tasks or instruments. The field is dominated by multivariate, latent-variable models including factor models, structural equation models, state-space models, and growth models (e.g., Bollen, 1989). When task scores are used with these advanced models, the results are importable and, consequently, difficult to interpret unless trial-by-trial variation is modeled. A critical question is whether the hierarchical models specified here extend well beyond two tasks. The generalization, at least for estimation, is straightforward. Bayesian development of latent variable models is a burgeoning field (Lee, 2007). Moreover, because Bayesian analysis explicitly allows for conditional rather than marginal expressions, adding covariance models is conceptually straightforward. Computational issues have been well explored, and general-purpose packages such as Stan (Carpenter et al., 2017) and JAGS (Plummer, 2003) are well suited to developing advanced latent variable models that account for trial-by-trial variation.

Although some forms of analysis are straightforward in Bayesian latent-variable modeling, Bayes-factor model comparison is not among these. Developing Bayes factors for complicated mixed models is certainly timely and topical, but the computational issues may be difficult. As a result, there is much work to be done if one wishes to state the evidence for theoretically motivated constraints on covariation across several tasks.

A Caveat

In the beginning of this paper, we asked if low correlations reflect statistical considerations or substantive considerations. The answer here, with a large data set, is that there is a substantive claim. We state positive evidence for a lack of correlation across Stroop and flanker tasks. That said, the better answer to our initial question is "Yes." Yes, the attenuated correlations have substantive meaning, and, yes, there are difficult statistical considerations.

We suspect there is far less true individual variability in many tasks than has been realized, and apparent variability comes more from the trial level than from true individual differences. Consider a typical priming task. Trials in these tasks take about 500 ms to complete, and a large effect is about 50 ms. If the average is 50 ms, how much could individuals truly vary? The answer is "not that much," especially if we assume that no individual has true negative effects (Haaf & Rouder, 2017; Rouder & Haaf, 2018). Indeed, we have analyzed the observed and true (latent) variation across individuals in many tasks (Haaf & Rouder, 2017, 2018). Observed variation is usually on the order of 100s of milliseconds. Yet, once trial-by-trial variation is modeled, individuals' true values vary from 10 ms to 40 ms. For such a narrow range, accurate understanding of individuals would require estimating each individual's effect to within a few milliseconds. Even the hierarchical models we advocate cannot mitigate this fact; if there is too little resolution then the posterior reliabilities and correlations will not be well localized. Obtaining the requisite level of resolution requires several hundred trials per individual per condition. It may be that these types of small-effect tasks have too small a range of individual differences to be useful in individual-difference research without large-scale data collection efforts.

Appendix

Prior settings

Priors are needed on models, and the specification of these priors plays a critical role in Bayes factor model comparison. Some of the parameters are common to all the models under consideration. For Models $\mathcal{M}_g$, $\mathcal{M}_0$, and $\mathcal{M}_1$, the common parameters are σ^2, the collection of all α_{ij}, and the two ν_j. For these parameters, the priors have a marginal effect on model comparison so long as they are not overly constraining. Other parameters—the ones that define the differences between the models—are critical, and their priors must be set judiciously. The three models are defined by different specifications of θ_{ij}, and the priors on the related parameters need careful consideration. In the general model, these parameters are σ^2_ω and σ^2_γ. In the zero-correlation and full correlation model, the parameters are σ^2_θ and σ^2_ω, respectively. It is popular to express these variabilities as multiples of the trial-by-trial variability (Zellner, 1986; Zellner & Siow, 1980). Let $g_\omega = \sigma^2_\omega/\sigma^2$, $g_\gamma = \sigma^2_\gamma/\sigma^2$, and $g_\theta = \sigma^2_\theta/\sigma^2$. Then, priors on the g parameters are often scaled inverse-χ^2 distributions (Liang, Paulo, Molina, Clyde, & Berger, 2008; Overstall & Forster, 2010; Rouder, Morey, Speckman, & Province, 2012):

$$
\begin{aligned}
g_\omega &\sim \text{inverse-}\chi^2(1, r^2_\omega),\\
g_\gamma &\sim \text{inverse-}\chi^2(1, r^2_\gamma),\\
g_\theta &\sim \text{inverse-}\chi^2(1, r^2_\theta),
\end{aligned}
$$

where there is a single degree of freedom for each distribution, and scales of r^2_ω, r^2_γ, and r^2_θ. The scaled inverse-χ^2 is chosen because it makes the ensuing Bayes factor computations convenient and, additionally, it leads to inference that has strong objective Bayes properties (Berger, 2006; Liang et al., 2008). For example, it may be shown that with this choice the Bayes factor goes to appropriate extremes as sample sizes increase or as the data becomes less and less variable.

The critical issue is the setting of scale constants: r_ω^2, r_γ^2, and r_θ^2. At first glance, setting these constants seems arbitrary. More alarming, the Bayes factor model comparisons will depend on these settings. It is for this reason—that seemingly arbitrary settings have big impacts on inference— that Gelman & Carlin (2017) are skeptical about the usefulness of Bayes factors. Here we justify our settings. Subsequently, we will explore the sensitivity of inference to reasonable variations in these settings.

As substantive scientist we have background knowledge of how variable effects tend to be in comparable experiments. In a typical Stroop or flanker experiment, the trial-to-trial standard deviation is somewhere around 200 ms to 300 ms. Hedge et al. (2018) had particularly fast and error prone responses in their data set, and as a result, we take the 200 ms value to help set prior scale constants. A healthy Stroop or flanker effect here would be about 50 ms, or 0.25 of the trial-to-trial standard deviation. The next question is how much do we think people could possible vary in each task. We chose 33.33 ms or $r_\theta = 0.17$ of the trial-to-trial standard deviation, and this value was used in both the full correlation and no correlation models. To set r_ω and r_γ, we equated the bivariate variances between the general model and no correlation model and chose r_γ to be 0.67 of r_ω. The reason we chose this ratio is that we *a priori* expect some positive covariation across the tasks. With these choices, the value of r_ω=0.14 and the value of r_γ=0.10.

Figure 6A shows the prior on standard deviations σ_θ, σ_ω and σ_γ based on these choices. As can be seen, although the priors have scales, their flexibility comes from their slow, fat right tails.

The next panel, Figure 6B shows how these choices specify a prior over correlation, ρ, in the general model. The distribution is u-shaped which is reasonable for priors on bounded spaces (Jaynes, 1986). The slight weight toward positive correlations reflects the choice that $r_\omega > r_\gamma$. Had these two settings been equal, then there would be no slight weight toward positive and negative values. If $r_\omega < r_\gamma$, the weight is toward negative values.

Sensitivity to Prior Specification

One issue is that Bayes factor values are dependent on prior specification. A few points of context are helpful in understanding this dependence. It seems reasonable to expect that if two researchers run the same experiment and obtain the same data, then they should reach similar conclusions. To meet this expectation, many Bayesian analysts actively seek to minimize this dependence by picking likelihoods, prior parametric forms, and heuristic methods of inference so that variation in prior settings have minimal influence (Aitkin, 1991; Gelman & Shalizi, 2013; Kruschke, 2012; Spiegelhalter et al., 2002). In the context of these views, the dependence of prior settings on inference is viewed negatively; not only is it something to be avoided, it is a threat to the validity of Bayesian analysis.

We reject this expectation that minimization of prior effects is necessary or even laudable. Rouder et al. (2016) argue that the goal of analysis is to add value by searching for theoretically-meaningful structure in data. Vanpaemel (2010) and Vanpaemel & Lee (2012) provide a particularly appealing view of the prior in this light. Accordingly, the prior is where theoretically important constraint is encoded in the model. When different researchers use different priors, they are testing different theoretical constraints, and not surprisingly, they will reach different opinions about the data. Rouder et al. (2016) argue that this variation is not problematic in fact it should be expected and seen positively. Methods that are insensitive to different theoretical commitments are not very useful. Rouder et al. (2016) recommend that so long as various prior settings are justifiable, the variation in results should be embraced as the legitimate diversity of opinion. When reasonable prior settings result in conflicting conclusions, we realize the data do not afford the precision to adjudicate among the positions.

The critical prior specifications in the models are the settings r_θ, r_ω, and r_γ. How does the Bayes factor change if we make other reasonable choices? Let's start by noting that about a 50 ms effect is reasonably expected. We cannot imagine individual variation being so large as to have a standard deviation much greater than this value. If say the true value of σ_θ was

50 ms, it would imply that 14% of individuals have true negative Stroop or flanker effects, which seems implausible. Likewise, we cant imagine variation being smaller than say 10 ms across people. These values, 10 ms and 50 ms, inform a bracket of reasonable settings for r_θ.

The same approach works for finding reasonable ranges for the ratio of r_γ/r_ω. The chosen value was 0.67 meaning that mildly positive correlations were expected. We think a reasonable range for this ratio is from a high value of 1.0 to a low value of 1/3. The value of 1.0 corresponds to as great a chance of negative correlation as positive, which is the bottom limit on the possible reasonable ranges of correlation for tasks that purportedly tap the same underlying construct. The value of 1/3 sets a lofty expectation of positive correlation. We cannot imagine settings greater than or less than these values would be reasonable for the general model.

Table 2 shows how variation in prior settings resulted variation of the Bayes factors. The first three columns show the settings for r_θ, r_ω, and r_γ. The next two columns show Bayes factors for the winning model for two cases. The first is the correlation between the Stroop and flanker task, and the Bayes factor is by how much the null-correlation model is preferred over the general model. The second is the correlation among even and odd trials in the flanker task, and the Bayes factor is by how much the full-correlation model is preferred over the general model. As can be see, the prior settings do matter. Take the correlation between the Stroop and flanker tasks. The null correlation model wins in all cases, but the value ranges from about 5-to-1 to about 11-to-1 over the general model. These are relatively stable results bolstering the claim that there is evidence for a lack of correlation. For the high reliability case, the full correlation model wins in all cases, but the Bayes factor values are quite variable, from 3-to-1 to 176-to-1. Here we are less sanguine because the evidence is more dependent on prior assumptions. Had we taken a larger bracket, the model preferences may have even reversed. While we may favor a full correlation model, that preference should be tempered and made with caution.

References

Aitkin, M. (1991). Posterior Bayes factors. *Journal of the Royal Statistical Society. Series B (Methodological)*, *53*(1), 111–142. Retrieved from http://www.jstor.org/stable/2345730

Baayen, R. H., Tweedie, F. J., & Schreuder, R. (2002). The subjects as a simple random effect fallacy: Subject variability and morphological family effects in the mental lexicon. *Brain and Language*, *81*, 55–65.

Berger, J. O. (2006). The case for objective Bayesian analysis. *Bayesian Analysis*, *1*, 385–402.

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*.

de Finetti, B. (1974). *Theory of probability* (Vol. 1). New York: John Wiley; Sons.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193–242. Retrieved from http://dx.doi.org/10.1037/h0044139

Efron, B., & Morris, C. (1977). Stein's paradox in statistics. *Scientific American*, *236*, 119–127.

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.

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.

Gelman, A., & Carlin, J. (2017). *Some natural solutions to the p-value communication problem—and why they won't work*.

Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics.

British Journal of Mathematical and Statistical Psychology, *66*, 57–64.

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

Green, S. B., Yang, Y., Alt, M., Brinkley, S., Gray, S., Hogan, T., & Cowan, N. (2016). Use of internal consistency coefficients for estimating reliability of experimental task scores. *Psychonomic Bulletin & Review*, *23*(3), 750–763.

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. (2018). *Some do and some don't? Accounting for varieties of individual difference structures.* 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.

James, W., & Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the fourth berkeley symposium on mathematical statistics and probability* (pp. 361–379 ,).

Jaynes, E. (1986). Bayesian methods: General background. In J. Justice (Ed.), *Maximum-entropy and bayesian methods in applied statistics*. Cambridge: Cambridge University Press.

Jeffreys, H. (1961). *Theory of probability (3rd edition)*. New York: Oxford University Press.

Kruschke, J. K. (2012). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*.

Kruschke, J. K., & Liddell, T. M. (2017). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*. Retrieved from

http://link.springer.com/article/10.3758/s13423-016-1221-4

Laplace, P. S. (1986). Memoir on the probability of the causes of events. *Statistical Science*, *1*(3), 364–378. Retrieved from http://www.jstor.org/stable/2245476

Lee, M. D., & Webb, M. R. (2005). Modeling individual differences in cognition. *Psychonomic Bulletin & Review*, *12*(4), 605–621.

Lee, P. M. (1997). *Bayesian statistics: An introduction*. New York: Wiley.

Lee, S.-Y. (2007). *Structural equation modelling: A bayesian approach*. New York: Wiley.

Lehmann, E. L., & Casella, G. (1998). *Theory of point estimation, 2nd edition*. New York: Springer.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g-priors for Bayesian variable selection. *Journal of the American Statistical Association*, *103*, 410–423. Retrieved from http://pubs.amstat.org/doi/pdf/10.1198/016214507000001337

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

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

Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, –. Retrieved from http://www.sciencedirect.com/science/article/pii/S0022249615000723

Myung, I.-J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin and Review*, *4*, 79–95.

Overstall, A. M., & Forster, J. J. (2010). Default bayesian model determination methods for generalised linear mixed models. *Computational Statistics & Data Analysis*, *54*(12), 3269–3288.

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

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

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*.

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*. 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., Haaf, J. M., & Aust, F. (2018). From theories to models to predictions: A Bayesian model comparison approach. *Communication Monographs*, *85*, 41–56. Retrieved from https://doi.org/10.1080/03637751.2017.1394581

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., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*, 356–374. Retrieved from http://dx.doi.org/10.1016/j.jmp.2012.08.001

Schmiedek, F., Oberauer, K., Wilhelm, O., Süß, H.-M., & Wittmann, W. W. (2007). Individual differences in components of reaction time distributions and their relations to working memory and intelligence. *Journal of Experimental Psychology: General*, *136*(3), 414.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling:*

Multilevel, longitudinal, and structural equation models. Boca Raton: CRC Press.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Linde, A. van der. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, *64*, 583–639.

Stahl, C., Voss, A., Schmitz, F., Nuszbaum, M., Tüscher, O., Lieb, K., & Klauer, K. C. (2014). Behavioral components of impulsivity. *Journal of Experimental Psychology: General*, *143*(2), 850.

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

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, *54*, 491–498.

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

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and wAIC. *Statistics and Computing*, *27*(5), 1413–1432.

Voelkle, M. C., Brose, A., Schmiedek, F., & Lindenberger, U. (2014). Toward a unified framework for the study of between-person and within-person structures: Building a bridge between two research paradigms. *Multivariate Behavioral Research*, *49*(3), 193–213.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problem of p values. *Psychonomic Bulletin and Review*, *14*, 779–804. Retrieved from https://doi.org/10.3758/BF03194105

Zellner, A. (1986). On assessing prior distirbutions and Bayesian regression analysis with g-prior distribution. In P. K. Goel & A. Zellner (Eds.), *Bayesian inference and decision techniques: Essays in honour of Bruno de Finetti* (pp. 233–243). Amsterdam: North Holland.

Zellner, A., & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses.

In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian statistics: Proceedings of the First International Meeting held in Valencia (Spain)* (pp. 585–603). University of Valencia.

Table 1
Bayes Factor Values for Competing Models of Correlation.

	General	No Correlation	Full Correlation
Stroop	1-to-1	1-to-2330	1-to-50
Flanker	1-to-1	1-to-469	1-to-31
Stroop v. Flanker	1-to-5.4	1-to-1	1-to-1.41×10^{37}
High Correlation	1-to-31	1-to-1.03×10^{5}	1-to-1

Note. Bayes factors are relative to the preferred model. These by convention have Bayes factors of 1-to-1. The remaining factors describe how much worse a model fares.

Table 2
Sensitivity to Prior Settings

r_θ	r_ω	r_γ	Stroop v. Flanker, B_{0g}	High Correlation, B_{1g}
0.17	0.14	0.1	5.4	31
0.25	0.22	0.14	8.5	176
0.05	0.04	0.03	9.3	3
0.17	0.12	0.12	4.9	73
0.17	0.2	0.07	11	12

Note. B_{0g} = Support for Null Model over General Model; B_{1g} = Support for Full Model over General Model.

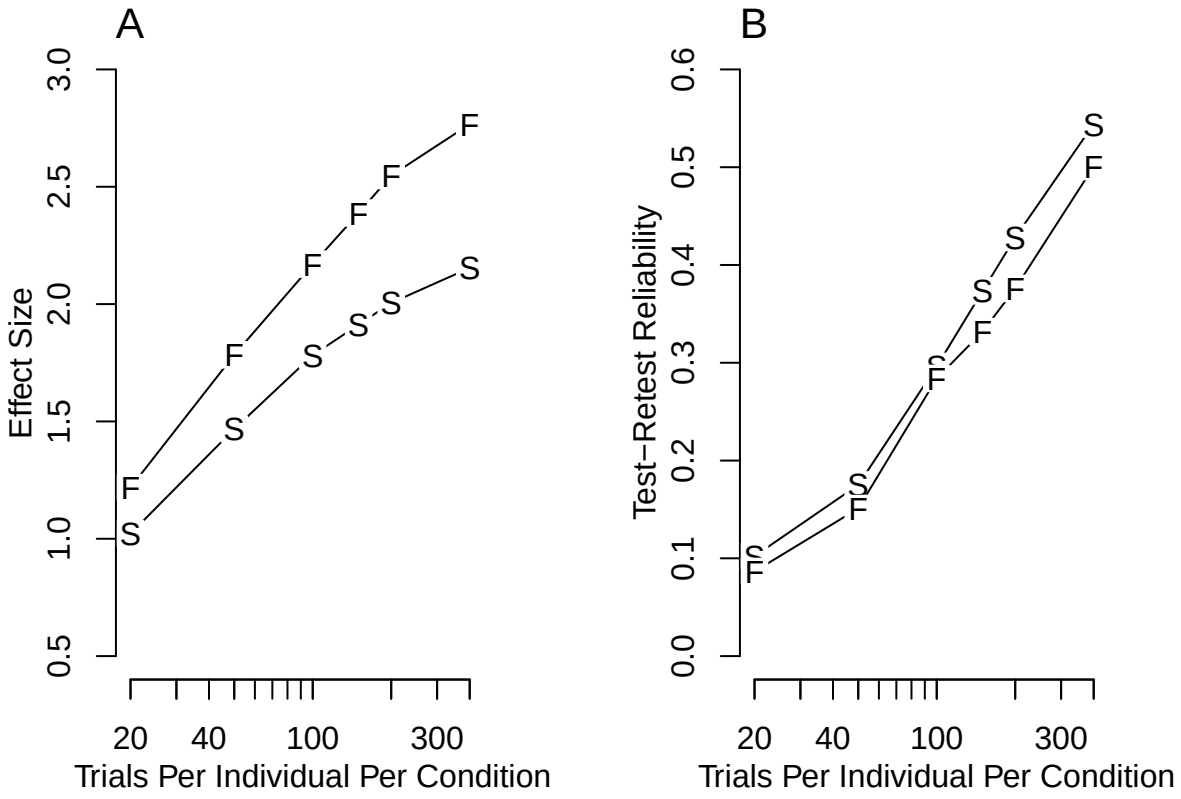

Figure 1. The effect of the number-of-trials-per-individual on sample effect sizes and sample reliability for Stroop and Flanker data from Hedge et al. (2018). Each point is the mean over 100 different samples at the indicated sample size.

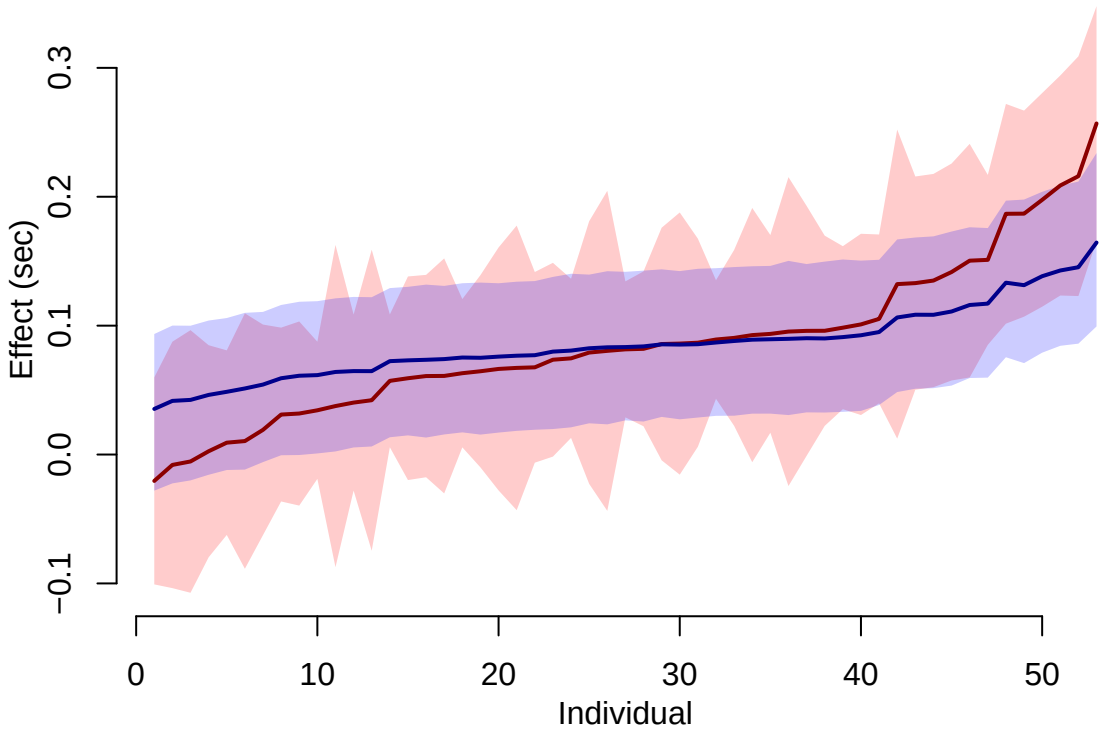

Figure 2. Estimates of individual's Stroop effect from one block of data. The estimates that have greater variability are sample effects; the smoothed estimates are from the hierarchical model.

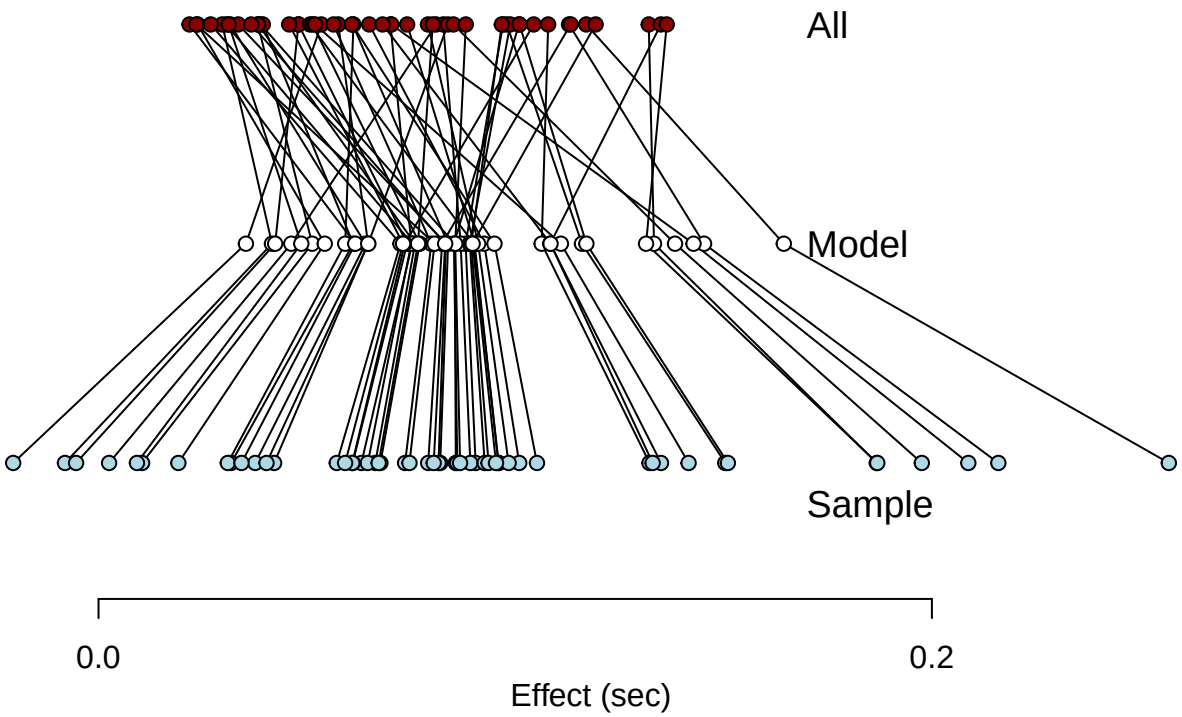

Figure 3. A comparison of estimates from one block to the those from all blocks shows the benefits of hierarchical regularization. The bottom row shows sample estimates from one block; the middle row shows the same from the hierarchical model; the top row shows sample estimates from all the blocks. The shrinkage from the hierarchical model attenuates the variability so that it matches that from much larger samples.

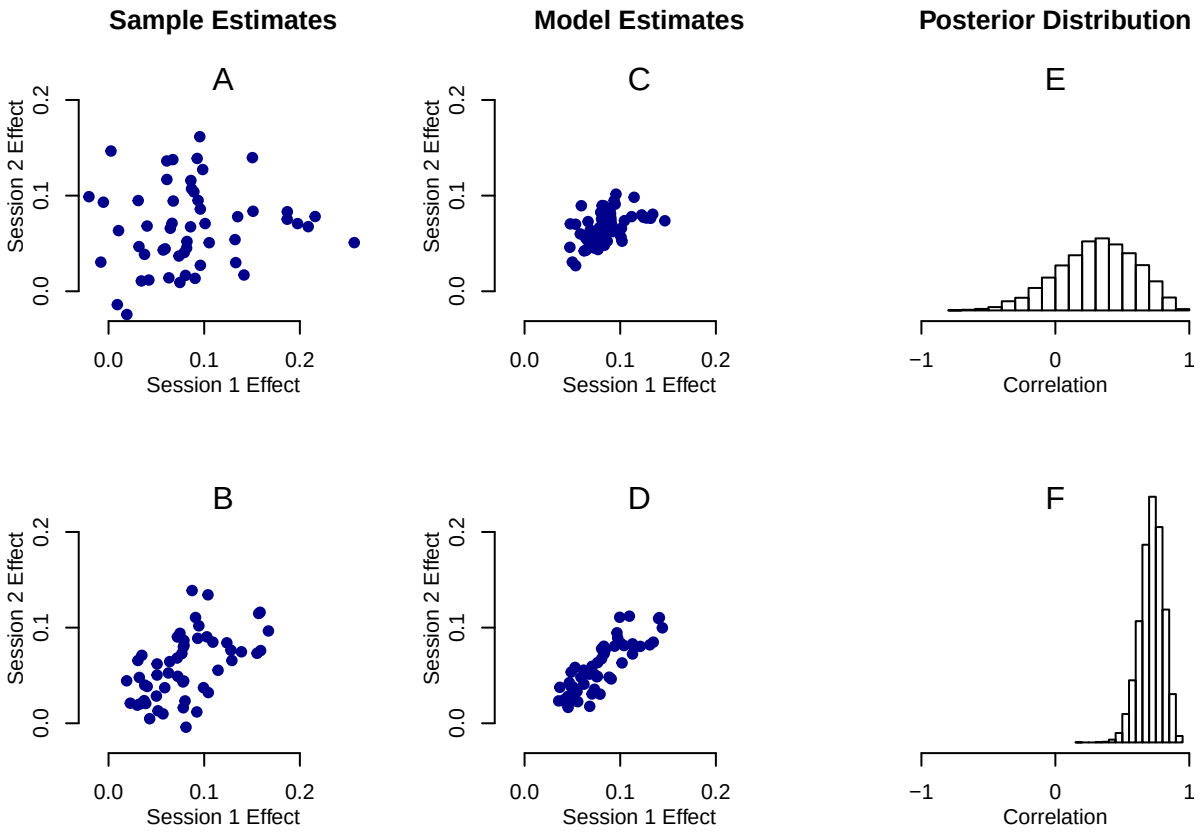

Figure 4. Test-retest reliability for Hedge et al.'s Stroop-task data set. The top row shows analysis for one block per session; the bottom row shows analysis for all blocks. The model analyses show substantial correlation even with one block. When all the data are considered the test-retest correlation is well localized for satisfactorily high values.

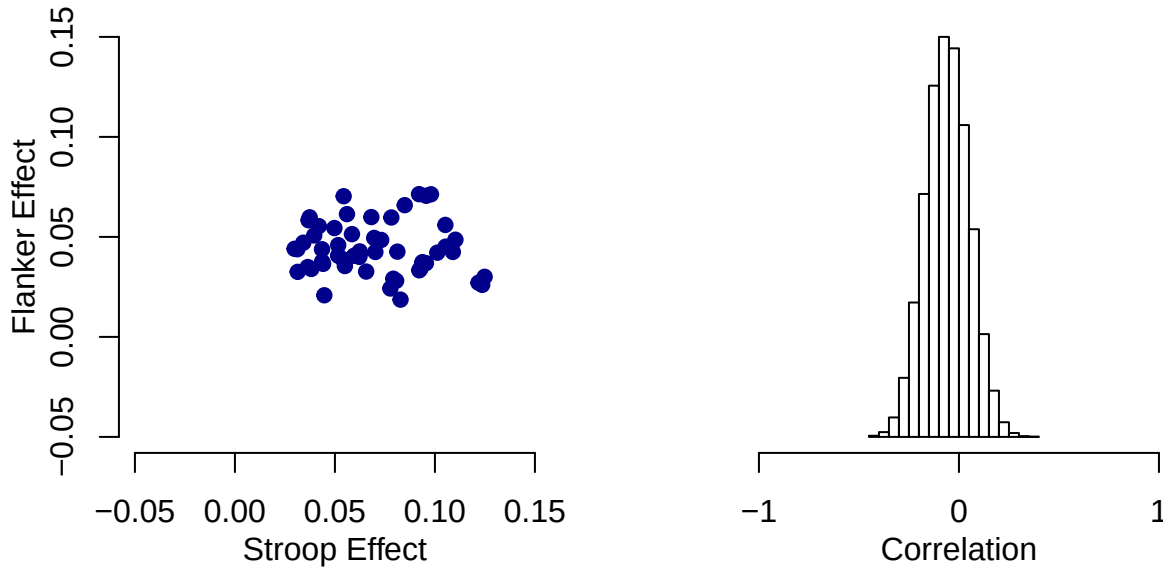

Figure 5. The lack of correlation between flanker task and Stroop task performance. Left: Scatter plot of individuals. Right: Posterior distribution of the correlation coefficient.

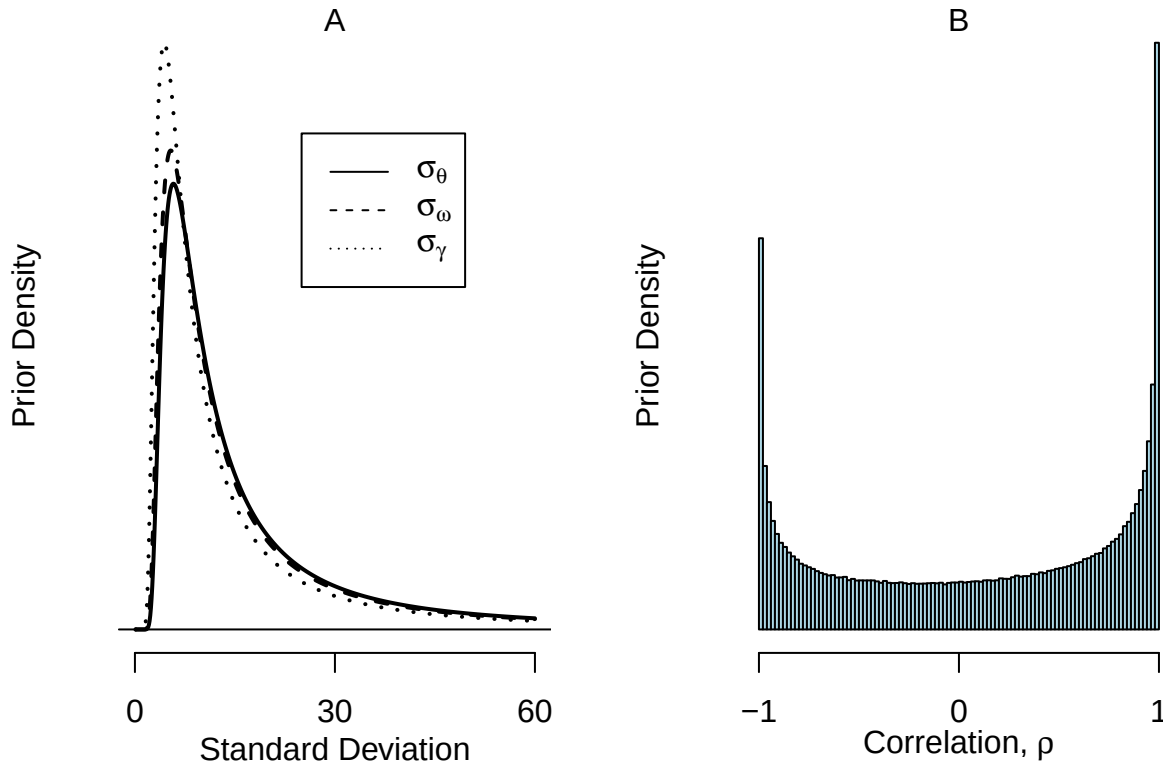

Figure 6. Prior specifications. A. The critical priors are on σ_θ^2, σ_ω^2, and σ_γ^2. Shown are the density functions for the standard deviations $(\sigma_\theta, \sigma\omega, \sigma_\gamma)$. B. The implied prior for the correlation coefficient ρ.