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 2, 7/2018

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

In modern individual-difference studies, researchers often correlate performance on various tasks to uncover common latent processes. Yet, in some sense, the results have been disappointing as correlations among tasks that seemingly have processes in common are often low. 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 in different tasks 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 here, it greately attenuates measures of effect size and correlation. 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 hierarchical models, and this treatment rescues classical concepts of effect size, reliability, and correlation for studying individual differences with experimental tasks. Using recent data from Hedge et al. (2018) we show 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 confident about using scores from experimental tasks for the following reasons: First, 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). Second, 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 task-based measures 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 designed to measure the same latent concepts do not correlate heavily with one another. 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 measures are notoriously low; most do not exceed .2 in value. The Stroop and flanker measures, 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 measures of inhibition. Ito et al. (2015) considered several implicit attitude tasks used for measuring implicit bias. Here again, there is surprisingly little correlation among bias measures that purportedly measure the same concept.

Why correlations are so low among these task-based measures is a mystery (Hedge et al., 2018; Rey-Mermet et al., 2018). After all, the effects 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 substantive. The presence of low correlations in large-sampled studies may imply 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. The second explanation is that the lack of correlation is mostly methodological. High correlations may only be obtained if the tasks are highly reliable, and low correlations are not interpretable without high reliability. Hedge and colleagues 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.

Researchers typically use classical test theory, at least implicitly, to inform methodology 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. Theoretically important quantities, say effect size within a task and correlations across tasks, 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 correlation 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.*

The main problem with using classical test theory for experimental tasks, however, is that portability is violated. Concepts of 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 estimate different truths if they have different sample sizes.

In the next section, we show how dramatically portability is violated in practice, and how much these violations affect the interpretability of classical statistics. These failures serve as motivation for a call to use hierarchical linear models that model trial-by-trial variation as well as variation across individuals. We develop these models, and then apply them to address the low-correlation mystery between flanker and Stroop tasks.

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; another is the number 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 a 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.[1] Here, we use the Hedge et al. data set to assess how 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:

[1] Hedge ran 144 trials per block with 5 blocks per session for a total of 10 blocks. These 144 trials were divided into three conditions: congruent, neutral, and incongruent. We do not use the neutral condition in our analysis. As a result, there are 48 trials per individual per condition per block, or 480 trials per individual per condition in the experiment. Once we discarded errors, RTs that were too fast or slow, and warm-up trials, there were on average 42 trials per individual per condition per block. Our discard parameters are documented in the markdown version of this document.

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

Two quantities of interest can be calculated from these individual effects: The effect size and the test-retest reliability. To calculate the effect size, we first average the session-by-individual effects across sessions to obtain individuals' effects denoted $\bar{d}_i$. The effect size is just the ratio of the grand mean of the individual effects 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 session-by-individual effect scores for the first session ($j = 1$) to the second session ($j = 2$).

Figure 1 shows a lack of portability in these two 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.

The critical and outsized role of replicates is seemingly under-appreciated in practice. 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.

A Hierarchical Model

Gaining Portability in The Limit of Infinite Trials

In this paper, we strive for portability—the meaning of correlation and effect size should be independent of the number of trials in component tasks. Portability is largely solved in classical test theory by using standardized instruments, say intelligence tests of fixed length. Yet, we think it is unreasonable to expect the same of experimentalists. Experimentalists will invariably use different number of trials depending on many factors including their access to subject pools, their overall goals, and the number of tasks and conditions in their experiments. We suspect recommending a standard number of trials per experiment will be unsuccessful.

The solution to portability that we develop here is the implementation of hierarchical statistical models to “correct” measurements for trial-by-trial noise. In our case, where theoretical targets are sizes of effects within tasks and correlations of these effects across tasks, noise from finite trials serves as a nuisance. Researchers are most interested in measurement of individuals’ *true* scores, that is, the measurement of individuals’ true flanker-effect and Stroop-effect scores, and the true correlation between them. While individuals perform finite trials, we may use statistical models to query what may reasonably happen in the limit. And estimates of effect sizes and correlations in the large-sample limit become portable estimates of effect sizes and correlations.

Trial-noise-free estimates may be made with a hierarchical linear model that accounts for trial-by-trial variation and individual variation simultaneously. 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, and Wittmann (2007) and Voelkle, Brose, Schmiedek, and 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 address the low-correlation mystery and to make classical test-theory concepts portable in experimental settings is novel.

Practitioners of classical test theory sometimes account for trial-by-trial noise by treating it as measurement error. The goal is to calculate this measurement error and then subtract it out. A good example comes from Spearman (1904a), who describes how correlations may be corrected for attenuation from measurement error (see https://en.wikipedia.org/wiki/Correction_for_attenuation). We find that our hierarchical model estimates agree with these corrections, but the models offer a deeper understanding of the statistical and theoretical dynamics at play.

In the next section we provide a fairly formal exposition of the model. We use an equation-based rather than graphical presentation of the model because it is from the equations that the main results flow. Using the equations, we derive an expression for the sample effects and show how the values are attenuated by trial variability. In the following section, we do the same for sample correlation between two tasks, and again show how the values are attenuated by trial noise. We show how and why the hierarchical model provides for unattenuated estimates.

Model specification

Understanding what the model is and how it works relies on some basic notion. Let I be the number of people, J be the number of tasks, K be the number of conditions, and L be the number of trials per person per task per condition, or the number of replicates. Subscripts are used to denote individuals, tasks, conditions, and replicates. Let $Y_{ijk\ell}$ denote the ℓth observation, $\ell = 1, \ldots, L$ for the ith individual, $i = 1, \ldots, I$ in the jth task,

$j = 1, \ldots, J$ and kth condition, $k = 1, 2$. 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, 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 $x_1 = -1/2$ for the congruent condition and $x_2 = 1/2$ for the incongruent condition.

The goal then is to study θ_i, the ith person's Stroop effect. 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.

Hierarchical Regularization and the Portability of Effect Size

Even before applying the model, a bit of analysis shows the flaws in conventional analysis and the expected improvements from hierarchical modeling. 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. It is helpful to express the distribution of sample effects, d_i in model parameters:[2]

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

where L is the number of trials per individual per condition. The term $2\sigma^2/L$ corresponds to variability from trial-by-trial noise and σ_θ^2 corresponds to variability of individuals' effects. A classical effect size measure, mean(d)/sd(d), therefore estimates $\mu_\theta/\sqrt{2\sigma^2/L + \sigma_\theta^2}$. The problem is the inclusion of $2\sigma^2/L$, the nuisance trial-by-trial variation. This included nuisance trial variation results in individual effect estimates that are too variable and in effect size measures that are too small. More importantly, portability is violated as there is an explicit dependence on L, the number of trials. In the model-based approach, the critical parameter is θ_i, and its distribution is:

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

The effect size calculated from these individual effects is an estimate of μ_θ/σ_θ, a portable quantity.

It is now clear what aggregation and the hierarchical model do. Aggregation is a way of getting at the true values, but only in the large-trial limit (as $L \to \infty$). In this limit, the term $2\sigma^2/L$ becomes vanishingly small. The hierarchical model provides an estimate of the same quantity, but it estimates the correct quantity even with finite trials! By using it, results are portable to designs with varying numbers of individuals and trials per individual. These results are estimates of underlying, theoretically useful properties of tasks.

[2] To derive this distribution, note that $\bar{Y}_{i1}|\alpha_i, \theta_i \sim \text{Normal}(\alpha_i - \theta_i/2, \sigma^2/L)$ and $\bar{Y}_{i2}|\alpha_i, \theta_i \sim \text{Normal}(\alpha_i + \theta_i/2, \sigma^2/L)$. These random variables are *conditionally independent*. Hence, the mean and variance come about from summing. Therefore $(\bar{Y}_{i2} - \bar{Y}_{i1})|\alpha_i, \theta_i \sim \text{Normal}(\theta_i, 2\sigma^2/L)$. Marginalizing over θ_i yields the distribution.

How does d_i compare to model-based estimates of effects, θ_i? The hierarchical model is an example of a linear mixed model, and as such, may be implemented in a variety of packages including `SAS PROC MIXED`, `lmer`, `lavaan`, `Mplus`, `stan`, and `JAGS`. In the Appendix, we complete specification of the model as a hierarchical Bayesian model and use the `R`-package `BayesFactor` for analysis. We have also made our analysis code available at https://github.com/PerceptionAndCognitionLab/ctx-reliability.

Figure 2A and B shows the comparison for a single block of Stroop data in Hedge et al. (2018). In this block there are about $L = 42$ trials per individual per condition. There are two panels—one for sample effects (d_i), and one for model-based estimates of θ_i. Participants are ordered from the ones that show the smallest Stroop effect to the ones that show the largest. In the left panel, the solid line shows the sample effect for each participant. The shaded area is the extent of each individual's 95% confidence interval. In the right panel, the solid line shows the point estimate of θ_i and the shaded area is the analogous 95% credible interval. The three horizontal lines are the mean (solid) and one standard deviation markers (dashed).

Although the grand mean effect is about the same for model estimates and sample effects, individuals' effects are different. The sample effects are spread out further from the grand mean than are the model estimates. This increased spread is expected and, as discussed above, comes about because the variability of d_i is the sum of true across-individual variation (σ_θ^2) and trial-by-trial variation ($2\sigma^2/L$). The spread of model based estimates, in contrast, reflects true individual variation alone. This reduction of variability in hierarchical models is known as *shrinkage* or *regularization* (Efron & Morris, 1977). Regularization of estimates is a well-known property of hierarchical models, and regularized estimates are often more accurate than sample statistics (Lehmann & Casella, 1998).

Because the benefits of regularization remain opaque in many hierarchical applications in cognitive psychology, we explore them in a bit more depth here. The estimates in Figure 2B were from a single block. 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. The 10-block set consists of over 400 trials per person and condition. Figure 2C, the style of which is borrowed from Efron and Morris (1977), 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 brings the variability of the model-based estimates from one block in line with the variability 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 1.62 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).

A Two-Task Model for Reliability and Correlation

The hierarchical model may be expanded from accounting for only one task to account for two tasks or two sessions 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-retest 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 opposite 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. These individual common effects and oppositional effects are random effects, and we place a hierarchical constraint on them:

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

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

To gain an expression for the correlation among two sessions, 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' effects across two sessions:

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

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 quantity is portable as it does not include trial-by-trial variability.

For comparison, we may also express in multivariate distribution for sample effects for two individuals in two sessions:

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

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

Figure 3 provides a real-data comparison of model-based and sample correlations. 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 (C, D). Here, there is shrinkage, especially for the single block (C). 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.

Figure 3 highlights the dramatic difference between conventional and hierarchical-model analysis, especially for smaller numbers of trials per participant. If one uses the conventional correlation coefficient for one block of data, the resulting value is 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 -0.32 to 0.82). 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.

The above real-world demonstration that sample estimates of correlations are so badly attenuated should be alarming. The one-block case was based on a design with 48 trials per individual per condition. This is a typical number for individual-difference batteries with a large number of tasks. Yet, the attenuation is so severe that test-retest correlations are barely detectable. No wonder observed correlations across tasks are so low. The hierarchical model provides a more sensitive view of the structure in the data by allowing for shrinkage to a regression line.

The attenuation of correlation from measurement error is well known in classical test theory and one approach is to apply a correction formula. The basic idea underlies the Spearman-Brown prophecy formula and the Spearman formula for disattenuation of correlations (Spearman, 1904b). We used the latter to compute an adjusted test-retest correlations for both tasks and for both the one-block and full data sets. The results are in Table 1, and there is a high degree of concordance for the model correlations and the corrected correlation. This overall concordance is expected as the correction for attenuation and the hierarchical models share the same foundational assumption about noise.

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 participated 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 4. 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. 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 numbers of trials.

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 for model comparison.

Theoretical Positions and Model Implementation

When assessing the relationship between two tasks, the main target is the true latent correlation in the large-trial limit. There are two opposing theoretically important positions: 1. that there is no correlation; and 2. there is full correlation. A lack of true correlation indicates that the two tasks are measuring independent psychological processes or abilities. Likewise, if there is full correlation, then the two tasks are measuring the same psychological processes or abilities.

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 the two theoretical 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 per individual. In the full-correlation model, regularization is to a line with a slope of 1.0.

Bayes Factor Analysis

We use the Bayes factor (Edwards, Lindman, & Savage, 1963; Jeffreys, 1961) to measure the strength of evidence for the three models. The Bayes factor is the probability of the observed data under one model relative to the probability of the observed data under a competitor model.

Table 2 shows the Bayes factor results for the Stroop and flanker task 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 3 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. The main difficulty in classical analysis occurs when researchers aggregate across trials to form individual-by-task scores. We recommend that researchers avoid this aggregation. If aggregated scores are used as input, then conventional sample measures are not interpretable without correction because they estimate quantities contaminated by removable trial-by-trial variation. With this contamination, effect sizes, reliabilities, and correlations are too low, sometimes dramatically so. We advocate applying hierarchical models to the trial level to remove trial-by-trial variation. 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.

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). In the Appendix, we show how the `nWayAOV` function in the `BayesFactor` package may be used for implementation.

Models vs. Corrections

One of the promising results here is that existing corrections for measurement noise yield similar results to hierarchical modeling. This is not surprising as these corrections are based on the same assumptions about noise. Yet, we believe that researchers should invest in modeling even though correction formulas are much easier to implement in practice. Here is why: First, correction yields a point estimate of the true latent value without a sense of uncertainty. Model-based point estimates, like the ones presented here, often come with a measure of uncertainty. Second, models may be compared to answer questions about whether the correlation is zero, one, or intermediary. Corrections offer no such inferences. Third, models may be extended to represent more complex structure in data. We know of no comparable correction for measurement noise in say factor analysis. Fourth, and most importantly, modeling provides for a deeper and more nuanced understanding of structure in data. Knowing how models represent structure and how data influences model-based conclusions allows researchers to add value in addressing theoretically-important substantive issues in a way that correction procedures cannot.

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 evidence for a lack of correlation across Stroop and flanker tasks. That said, we should not understate the statistical considerations.

We suspect there is 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, we believe, 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 several 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 to study individual differences in these tasks may requires several hundred trials per individual per condition, and perhaps this design choice may prove more critical than collecting from hundreds of individuals.

Appendix

Model Specification

The One-Task Model. The model of data for a single task is given by:

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

where i, k, and ℓ respectively index individual, condition, and replicate, and and where $x_1 = -1/2$, and $x_2 = 1/2$. Models on α_i and θ_i are

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

We use a g-prior specification for variance components (Zellner & Siow, 1980,Zellner (1986)), These priors have become popular for linear models [Liang, Paulo, Molina, Clyde, and Berger (2008),Rouder:etal:2012a]. Let $g_\alpha = \sigma^2_\alpha/\sigma^2$ and $g_\theta = \sigma^2_\theta/\sigma^2$. Then:

$$\pi(\mu_\alpha,\sigma^2) \propto 1/\sigma^2,$$

and

$$\mu_\theta \sim \text{Normal}(0, g_{\mu_\theta}\sigma^2),$$
$$g_\alpha \sim \text{inverse-}\chi^2(1, r^2_\alpha),$$
$$g_\theta \sim \text{inverse-}\chi^2(1, r^2_\theta),$$
$$g_{\mu_\theta} \sim \text{inverse-}\chi^2(1, r^2_{\mu_\theta}),$$

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 inverse-χ^2 prior on g is a popular choice because it is computationally convenient and leads to posteriors with desirable objective Bayes properties [Berger (2006);@Liang:etal:2008;Rouder:etal:2012].

The Two-Task Model. The model of data for two task is given by:

$$Y_{ik\ell}|\alpha_{ij}, \theta_{ij}, \sigma^2 \overset{\text{ind}}{\sim} \text{Normal}(\alpha_{ij} + x_k\theta_{ij}, \sigma^2),$$

where j indexes the task. Models on α_{ij} and θ_{ij} are

$$\alpha_{ij} \overset{\text{iid}}{\sim} \text{Normal}(\mu_\alpha, g_\alpha\sigma^2),$$

and

$$\theta_{i1}|\nu_1, \omega_i, \gamma_i = \nu_1 + \omega_i - \gamma_i,$$

$$\theta_{i2}|\nu_2, \omega_i, \gamma_i = \nu_2 + \omega_i + \gamma_i.$$

Models on ν_k, ω_i and γ_i are

$$\nu_k|g_\nu, \sigma^2 \overset{\text{iid}}{\sim} \text{Normal}(0, g_\nu\sigma^2)$$

$$\omega_i|g_\omega\sigma^2 \overset{\text{iid}}{\sim} \text{Normal}(0, g_\omega\sigma^2),$$

$$\gamma_i|g_\gamma, \sigma^2 \overset{\text{iid}}{\sim} \text{Normal}(0, g_\gamma\sigma^2).$$

Priors on

$$g_\nu \sim \text{inverse-}\chi^2(1, r_\nu^2),$$

$$g_\omega \sim \text{inverse-}\chi^2(1, r_\omega^2),$$

$$g_\gamma \sim \text{inverse-}\chi^2(1, r_\gamma^2),$$

Prior settings

The above models are estimated in the Bayesian framework, and as such, priors are placed on parameters. The critical settings are for the scale parameters r_α, r_θ, r_{μ_theta}, r_ν, r_ω, and r_γ In this section, we justify our choices for these settings. In the next section, we discuss how the choices affect analysis.

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. In our experience people also vary about this much, that is, in ordinary response time experiments, that variation of people is on the order of a few hundred milliseconds. We set $r_\alpha = 1$ to reflect that the expected trial-to-trial variation and expected variation across people on over response time is about the same size.

A healthy Stroop or flanker effect here would be about 50 ms, or 0.25 of the trial-to-trial standard deviation. Hence, we set 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. This value reflects our belief that nobody has negative true Stroop of flanker effects, and therefore, the overall size of the effect, 50 ms, limits the possible size of the variability. 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 5A 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 5B 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.

How To Analyse The Model

We analyzed the above models in the `BayesFactor` package (Morey & Rouder, 2015) in the `R` environment. This package provides for the analysis of a wide variety of ANOVA and regression models. It has been a key infrastructure component of the increasing popularity of Bayes factor model comparison in psychological science, and serves as the back end computational engine in JASP (Love et al., 2015). One of the key advantages here is that `BayeFactor` implements the g-prior set up, and this fact makes using it simple. The critical function for analysis is the function `nWayAOV()` (documentation available at https://www.rdocumentation.org/packages/BayesFactor/versions/0.9.12-4.2/topics/nWayAOV) This function outputs posterior distributions for all parameters and a Bayes factor model comparison statistic. The inputs to it are the data, the models defined by a design matrix, and the g-prior scale settings.

As mentioned previously, this paper and all the code needed for analysis and graphs are available at https://github.com/PerceptionAndCognitionLab/ctx-reliability. The heart of our code is at share/lib.R. The `R` script contains functions for loading Hedge et al's data and cleaning it. The function `design1Session` sets up a design matrix for the single-task model for inputted data. The function `est1Session` then returns the model-based estimates for the specified prior settings. The functions for two tasks work similarly.

Sensitivity to Prior Specification

When performing parameter estimation, the above models are rather robust to the choice of prior form and prior settings. The reason is simple, the priors are rather diffuse and the sample sizes in the data sets are quite large. Changes in the scale settings has minimal effects on the posterior distributions of the parameters.

The influence of the prior settings is more relevant for model comparison, and the 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, 2014; Spiegelhalter, Best, Carlin, & Linde, 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, Morey, and Wagenmakers (2016) argue that the goal of analysis is to add value by searching for theoretically-meaningful structure in data. Vanpaemel (2010) and Vanpaemel and 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 3 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 comfortable 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 by these sensitivity analyses.

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*.

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., & 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 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.

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. (2014). *Doing bayesian data analysis, 2nd edition: A tutorial with r, jags, and stan*. Waltham, MA: Academic Press.

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

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.

Love, J., Selker, R., Verhagen, J., Smira, M., Wild, A., Marsman, M., … Wagenmakers, E.-J. (2015). Software to sharpen your stats. *APS Observer*, *28*, 27–29.

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

Morey, R. D., & Rouder, J. N. (2015). BayesFactor 0.9.12-2. Comprehensive R Archive Network. Retrieved from http://cran.r-project.org/web/packages/BayesFactor/index.html

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*, *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., 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

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.

Spearman, C. (1904a). 'General intelligence,' obectively determined and measured. *American Journal of Psychology*, *15*, 201–293.

Spearman, C. (1904b). 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

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.

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.

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
Test-retest correlations for Hedge et al (2018) data sets.

	Sample	Model	Corrected for Attenuation
Stroop, one block	0.10	0.31	0.31
Stroop, full set	0.55	0.72	0.74
Flanker, one block	0.23	0.49	0.52
Flanker, full set	0.50	0.68	0.71

Table 2
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 3

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	8.6	170
0.05	0.04	0.03	9.3	2.9
0.17	0.12	0.12	4.8	76
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.

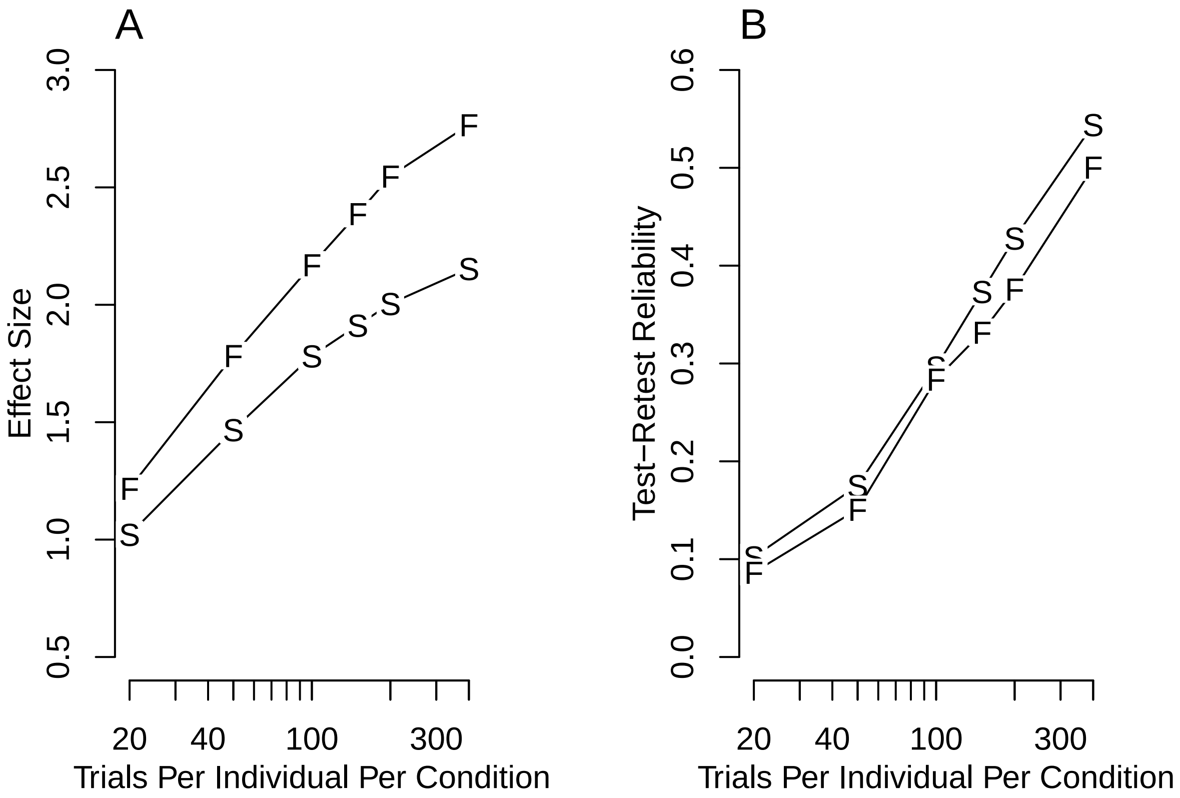

Figure 1. The effect of the number-of-trials-per-individual on sample effect sizes and sample reliability for Stroop and Flanker data from Hedge et al. (2018). Each point is the mean over 100 different samples at the indicated sample size.

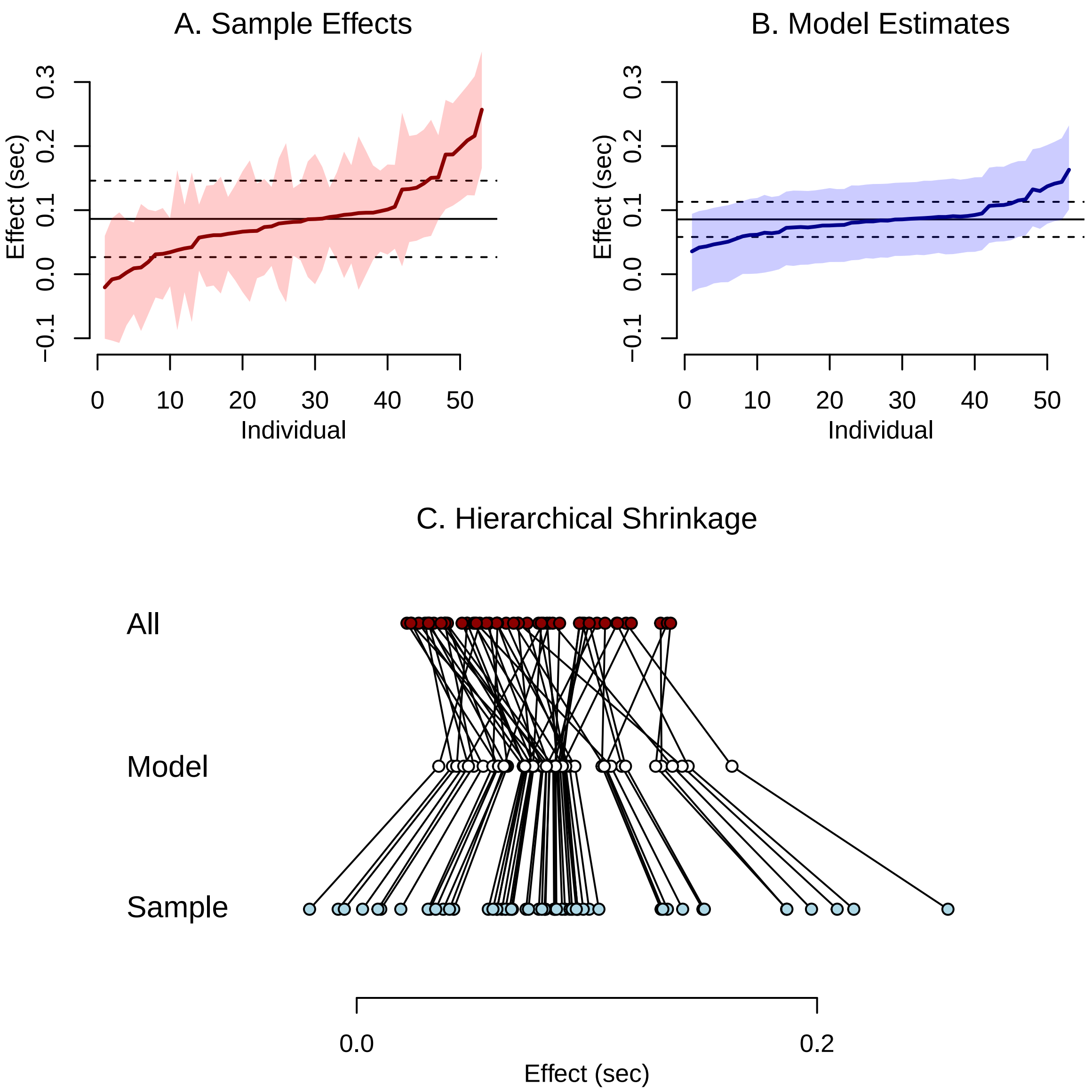

Figure 2. Analysis of a single A. Sample effects from one block of data for all individuals with 95% CIs. B. Model estimates from one block of data for all individuals with 95% credible intervals. C. Hierarchical shrinkage in practice. The bottom row shows sample estimates from one block; the middle 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.

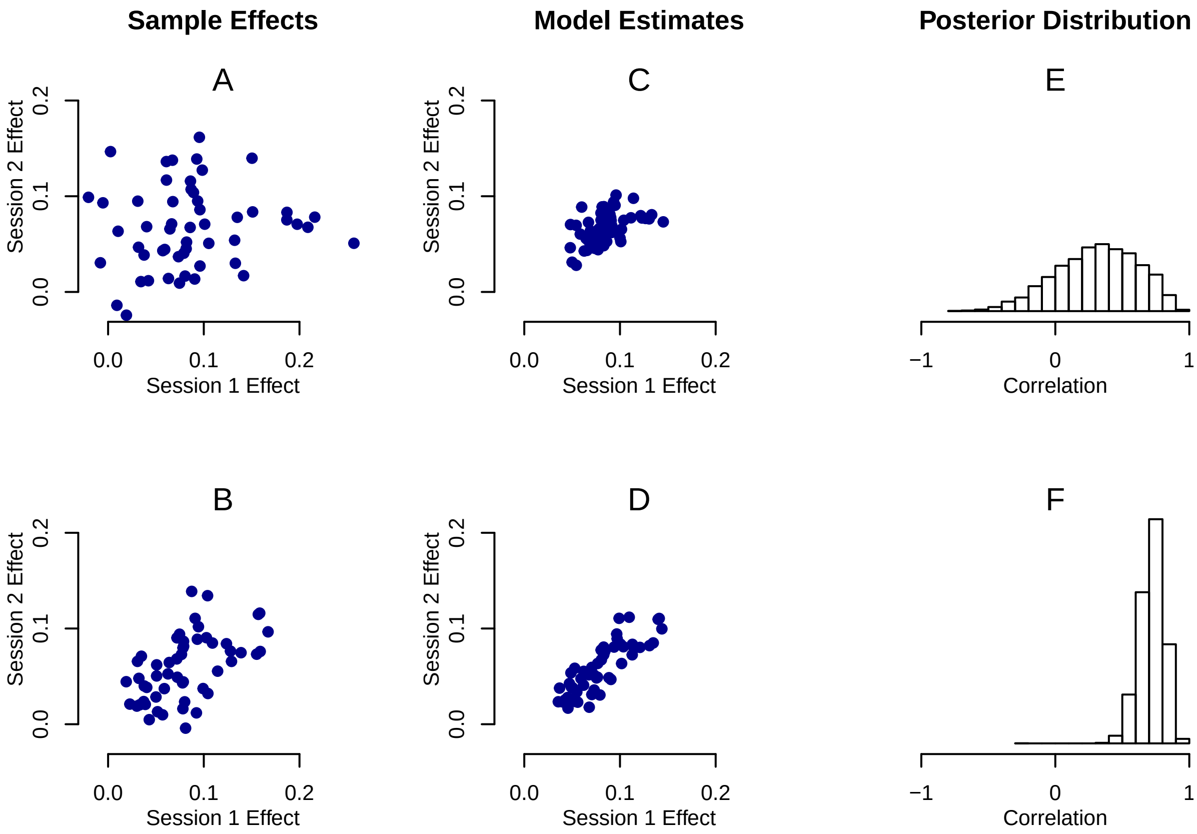

Figure 3. Test-retest reliability for Hedge et al.'s Stroop-task data set. The top row shows analysis for one block per session; the bottom row shows analysis for all blocks. The model analyses show substantial correlation even with one block. When all the data are considered the test-retest correlation is well localized for satisfactorily high values.

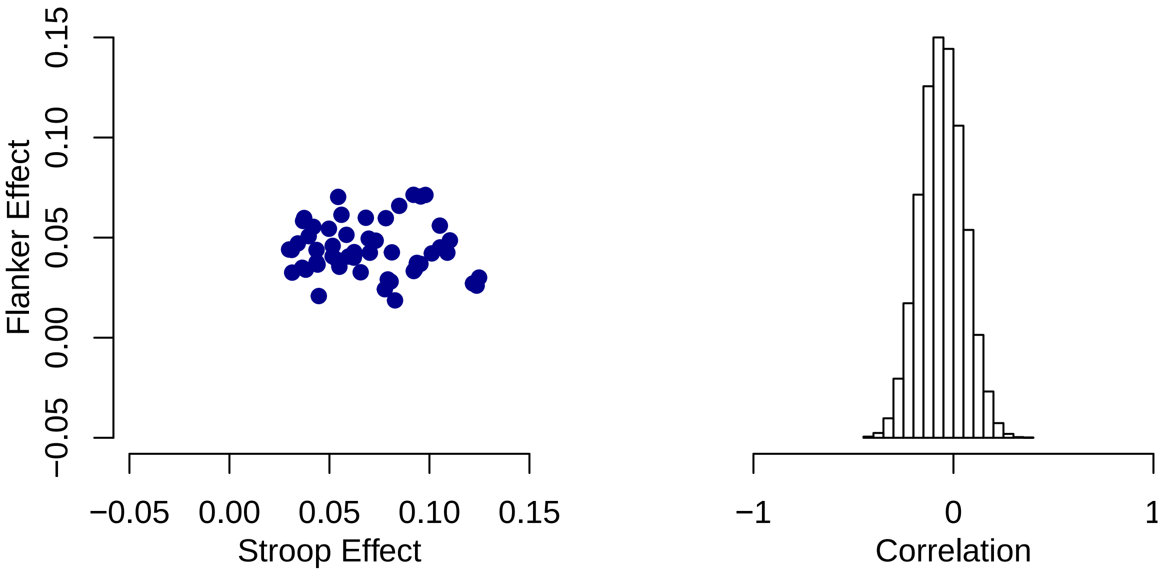

Figure 4. The lack of correlation between flanker task and Stroop task performance. Left: Scatter plot of individuals' model-based estimates. Right: Posterior distribution of the correlation coefficient.

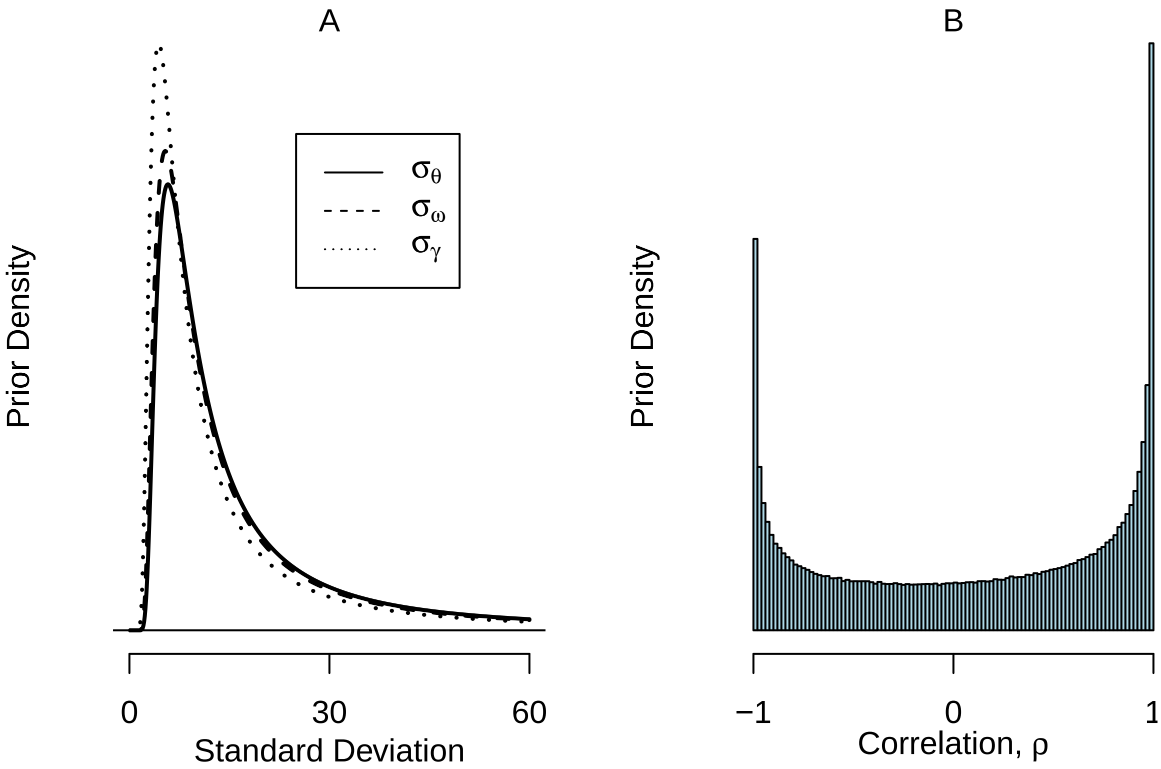

Figure 5. Prior specifications. A. The critical priors are on σ_θ^2, σ_ω^2, and σ_γ^2. Shown are the density functions for the standard deviations $(\sigma_\theta, \sigma\omega, \sigma_\gamma)$. B. The implied prior for the correlation coefficient ρ.