Beneath the Surface: Unearthing Within-Person Variability and Mean Relations with Bayesian Mixed Models

Donald R. Williams
University of California, Davis

Jeff N. Rouder
University of California, Irvine

Philippe Rast
University of California, Davis

Author Note

Research reported in this publication was supported by funding from the National Science Foundation Graduate Research Fellowship to DRW and the National Institute On Aging of the National Institutes of Health under Award Number R01AG050720 to PR. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

Abstract

Mixed-effects models are becoming common in psychological science. Although they have many desirable features, there is still untapped potential that has not yet been fully realized. It is customary to view homogeneous variance as an assumption to satisfy. We argue to move beyond that perspective, and to view modeling within-person variance ("noise") as an opportunity to gain a richer understanding of psychological processes. This can provide important insights into behavioral (in)stability. The technique to do so is termed mixed-effects location scale model. The formulation can simultaneously estimate mixed-effects sub-models to both the mean (location) and within-person variance (scale) for repeated measures data common to psychology. We develop a framework that goes beyond assessing the sub-models in isolation of one another, and allows for testing structural relations between the mean and within-person variance with the Bayes factor. We first present a motivating example, which makes clear how the model can characterize mean–variance relations. We then apply the method to reaction times gathered from two cognitive inference tasks. We find there are more individual differences in the within-person variance than the mean structure, as well as well as a complex web of structural mean–variance relations in the random effects. This stands in opposition to the dominant view of within-person variance–i.e., measurement "error" or "noise." The results also point towards paradoxical within-person, as opposed to between-person, effects. That is, in both tasks, several people had *slower* and *less* variable incongruent responses. This contradicts the typical pattern, wherein *larger* means are expected to be *more* variable. We conclude with future directions. These span from methodological to theoretical inquires that can be answered with the presented methodology.

Keywords: Bayesian, Mixed-effects location scale models, Bayes factor, Within-person variance

Beneath the Surface: Unearthing Within-Person Variability and Mean Relations with Bayesian Mixed Models

Repeated measurement designs are common to the social-behavioral sciences. Their use spans from observational inquires that track individuals over an extended period of time, to controlled settings that can include hundreds of experimental trials for each person. Modeling these kinds of data requires techniques that are able to partition and account for different sources of variation. This includes variation due to experimental effects (Aarts, Dolan, Verhage, & van der Sluis, 2015), time-varying predictors (Hoffman & Stawski, 2009), type of stimulus (Wolsiefer, Westfall, & Judd, 2017), as well as between-group differences (Lazic & Essioux, 2013). Adequately accounting for these sources of variability leads to the desired inference by ensuring nominal error rates are maintained (Aarts, Verhage, Veenvliet, Dolan, & van der Sluis, 2014; Barr, Levy, Scheepers, & Tily, 2013; Judd, Westfall, & Kenny, 2012; Williams, Carlsson, & Bürkner, 2017). Modeling these sources of variance can also provide valuable insight into psychological processes. For example, by capturing individual differences in temporal changes (Liu, Rovine, & Molenaar, 2012), interference (Haaf & Rouder, 2017), and learning trajectories (Williams & Rast, 2018).

In psychological inquires, it is customary to focus on the mean structure. By this we are referring to explaining the outcome of interest with one, or perhaps several, independent variables. For example, in experimental settings predicting reaction times, with an experimental condition (e.g., congruent vs. incongruent; King, Colla, Brass, Heuser, & Cramon, 2007; Parris, 2014). In these situations, it is common to aggregate repeated measures data at the individual level (Davidson, Zacks, & Williams, 2003; Wright, 2017). That is, for a given condition, each person contributes only their respective mean score. This allows for using relatively simple, and well-known, statistical methods such as the dependent samples t-test. Despite its popularity, this approach is not without criticism (Leppink, 2019; Leppink & Merriënboer, 2015). These limitations typically do not apply to

inferences about the average effect (or mean structure), for example the false positive rate (Williams et al., 2017), but that it cannot provide information about individual variability (Bauer, 2011). This is not only important for the study of individual differences (Kliegl, Wei, Dambacher, Yan, & Zhou, 2010), but also for the success of future inquiries. Power calculations can be more accurate when sources of variability are incorporated into the calculation (Judd et al., 2012; Westfall, Kenny, & Judd, 2014)

A variety of techniques have been proposed to model sources of variance. These include the repeated measures ANOVA and mixed-effects models (Liu et al., 2012). We focus on the latter because of its flexibility and less restrictive assumptions (Krueger & Tian, 2004). We refer interested readers to Boisgontier and Cheval (2016), where both approaches are thoroughly discussed. Mixed-models have been used extensively to study individual differences in psychology (Cudeck, 1996; Rast, Hofer, & Sparks, 2012; Rouder & Jun, 2005). They strike a middle ground between aggregating data and estimating person-specific models (Gelman & Pardoe, 2012). That is, they can simultaneously provide the average effect across individuals and the person-specific estimates (i.e., random effects). The latter allows for investigating, for example, how many people had an effect in the predicted direction or whether certain people had an effect in the opposite direction (Haaf & Rouder, 2017; Williams, Liu, Martin, & Rast, 2019). Both of these inferences provide important insight and much needed nuance for understanding psychological processes.

In this work, we argue that, although mixed-effects models have many desirable features, there is still untapped potential that has not yet been *fully* realized. An important limitation is that the focus has mostly remained on the mean structure. For example, in the case of computing intraclass correlations, the variance is partitioned into between- and within-person components (Bartko, 1976; McGraw & Wong, 1996). The latter provides a measure of relation between observations from a given individual (or group), which can then be used to compute reliability (Bartko, 1966). Importantly, the focus in ICC remains on the mean structure, that is, the similarity between observations of

the dependent variable. Further, it is common to account for between-person variance with so-called level-two predictors such as gender, ethnicity, or socioeconomic status (McNeish & Stapleton, 2016). The unexplained variance, that is the within-person variance at level one, then goes into the residual component that is typically considered a fixed, non-varying constant in mixed-effects models. Herein lies the untapped potential: rather than viewing homogeneous variance as an assumption to satisfy, we can seek to explain it just like the dependent variable.

Outside of psychology, for example in economics, there is a tradition of modeling heteroskedasticity (Engle, 2001). The variance is commonly referred to as *volatility* (Fernández-Villaverde, Guerrón-Quintana, Rubio-Ramírez, & Uribe, 2011). In fact, there are methods that exclusively model the variance and ignore the mean structure altogether (Bauwens, Laurent, & Rombouts, 2006; Ledoit, Santa-Clara, & Wolf, 2003)–which is quite opposite to the dominant approach in psychology.

In applied settings variability is sometimes studied in the social-behavioral sciences, where it is referred to as *intraindividual* variability (IIV; Christ, Combrinck, & Thomas, 2018; Fagot et al., 2018; Röcke & Brose, 2013), and thought to reflect behavioral consistency (Rast et al., 2012). This conceptualization builds upon a central idea that within-person, or IIV, is not regarded as reflecting mere measurement error but conveys systematic information (Cattell, Cattell, & Rhymer, 1947; Fiske & Rice, 1955; Horn, 1972; Ram & Gerstorf, 2009; Woodrow, 1932). Here the research question is specifically about the relation between variance and the outcome of interest. As an example, inconsistency in cognitive abilities has been proposed as an indicator of Alzheimer's disease in aging populations (Kalin et al., 2014). And (in)consistency was even suggested to predict death (MacDonald, Hultsch, & Dixon, 2008). That is, those that were more volatile tended to die before those that were relatively stable.

The most common statistical approaches for studying IIV rely on estimating the individual means (iM's) and individual standard deviations (iSD's) using a two-stage

approach: in the first stage, the iM's are computed, for example in a mixed-effects model or from individual regressions, and the residuals are recorded. In the second stage, individual SD's are obtained from the residuals which are used in a separate model as either a predictor or as the outcome (MacDonald et al., 2008; Ram & Gerstorf, 2009). While these approaches were recognized in the cognitive modeling literature, for example in Wagenmakers and Brown (2007) and Wagenmakers, Grasman, and Molenaar (2005), they have several limitations to consider. They can result in unreliable estimates that are particularly sensitive to the number of trials or repeated measurement occasions (Estabrook, Grimm, & Bowles, 2012; Wang & Grimm, 2012) and the underlying assumption of normality (Mestdagh et al., 2018; Wang, Hamaker, & Bergeman, 2012). Moreover, separating iM's from iSD's assumes independence of means and variances, which seems unlikely in most applications (e.g., the "law" proposed in: Wagenmakers & Brown, 2007), and it results in biased variance estimates (Leckie, French, Charlton, & Browne, 2014; Rast & Ferrer, 2018).

These limitations partially motivate this work. We describe an extension to the traditional mixed-effects approach, which allows for partitioning this unexplained variance, or within-person variance, and explain it as a function of covariates. The technique to do so is termed mixed-effects *location scale* model (MELSM Hedeker, Mermelstein, & Demirtas, 2008; Rast & Ferrer, 2018; Williams & Rast, 2018). The location refers to the mean structure and the scale refers to the (within-person) variance. The MELSM *simultaneously* estimates sub-models to both structures. Additionally, it accounts for the underlying co-variances among the individual difference parameters as well–i.e., the mean–variance relations. In this work, we build upon this foundation and introduce an approach that goes beyond assessing the mean and variance structures in isolation of one another. We present a general method with the goal of testing structural relations between the mean and within-person variance. This not only overcomes limitations of the two-stage approach, but as we show below, opens the door for answering novel research questions

about the interplay between the mean and within-person variability in psychology.

This work is organized as follows. In the first section we introduce a generic MELSM, where it is made clear how it captures the mean–variance relationship. Additionally, we formulate the hypothesis testing strategy for the random effects correlations. We then apply the method to two well-known cognitive interference tasks. Here, general advantages of the MELSM are highlighted, and the specific approach for testing hypotheses is employed. Note that some may be concerned about "overfitting" the data. Thus, in this section, we also compare the fully parameterized model to those that are relatively simpler. We conclude by discussing limitations of the MELSM in general and specific future directions for psychological applications.

The Mixed-Effects Location Scale Model

Random Intercepts Only Model

We introduce the MESLM by first fitting random intercepts to both the location and scale, and then progress to the fully parameterized model (Section Illustrative Examples). For the following we use data from an interference task that investigated the so-called "Stroop Effect." These data were first reported in von Bastian, Souza, and Gade (2016). They consist of 121 participants, each of which completed approximately 90 trials in total. About half of the trials were in the congruent condition, wherein the number of characters matched the displayed numbers–e.g., 22. The remaining trials were in the incongruent condition–e.g., 222. Further details are provided below (Section Data Set 1: The Stroop Task). The outcome is reaction time for correctly identifying the number of characters. For illustrative purposes, separate models are estimated for each condition (congruent and incongruent). Substantively these models answer the questions of whether slower individuals are also more variable in their responses (e.g., Figure 1 in: Wagenmakers, 2007). This simple example highlights how the MELSM can be used to investigate mean–variance relations hierarchically.

For the ith person and jth trial in the congruent condition, the mean structure is defined as

$$y_{ij} = \beta_0 + u_{0i} + \epsilon_{ij}, \tag{1}$$

where β_0 is the fixed effect and u_{0i} the individual deviation. More specifically, β_0 is the average of the individual means and for, say, the first subject ($i = 1$), the respective mean response time is $\beta_0 + u_{01}$. This is not equivalent to estimating the empirical means, because of the hierarchical structure of the model. Before describing this aspect of the model, we must account for the "errors". While they are typically assumed to be normally distributed with a constant variance, this is not the case for the MELSM–i.e.,

$$\sigma^2_{\epsilon_{ij}} = \exp[\eta_0 + u_{1i}]. \tag{2}$$

The subscripts denote the residual for the ith person and jth trial. These parameters are analogous to those in (2), in that η_0 is the average of the individual variances. Again for the first subject ($i = 1$), $\eta_0 + u_{11}$ is the variability of their respective response time distribution. Note the exponent is used to ensure the variance is restricted to positive values, and thus, is lognormally distributed (Hedeker et al., 2008).

This work focuses on the relations between the mean and within-person variance. As such, we need to estimate the random effects correlations by assuming the individual effects are drawn from the same multivariate normal distribution–i.e.,

$$\begin{bmatrix} u_{0i} \\ u_{1i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{bmatrix} \right). \tag{3}$$

Here σ_0^2 is the variance of location intercepts $var(u_{0i})$, σ_1^2 is the variance of the scale intercepts $var(u_{1i})$, and σ_{01} is the respective covariance $cov(u_{0i}, u_{1i})$. In order to ease computation and the definition of priors, we re-express the covariance matrix as $\boldsymbol{\tau}\ \boldsymbol{\Omega}\boldsymbol{\tau}'$, where $\boldsymbol{\Omega}$ is a 2×2 correlation matrix, and τ is the 2×2 diagonal matrix of random effects standard deviations $\text{diag}(\boldsymbol{\tau}) = \boldsymbol{\sigma} = (\sigma_0, \sigma_1)$. The assumed prior distribution for the correlations is

$$\mathbf{\Omega} \sim \text{LKJ}(\nu = 1), \tag{4}$$

where LJK is the Lewandowski, Kurowicka, and Joe prior (Lewandowski, Kurowicka, & Joe, 2009). This distribution is governed by a single parameter ν. A value of one places a uniform prior over all correlation matrices. This results in a uniform (marginal) prior for each correlation that is between -1 and 1, assuming a 2×2 matrix. Although this formulation extends to any size correlation matrix, it is important to note the LKJ prior is not invariant to the dimension. In other words, as the dimensions grows, even with $\nu = 1$, the density concentrates around zero. We return to this topic in the discussion. For simplicity, we assume *weakly* informative priors for the fixed effect intercepts and random effects standard deviations–i.e.,

$$\begin{aligned} \beta_0, \eta_0, &\sim N(0, 5) \\ \tau_0, \tau_1 &\sim N^+(0, 1). \end{aligned} \tag{5}$$

This parameterization has several interesting features. As seen in (3), it captures the correlation between the means and variances of each response time distribution. This is made possible by assuming the random effects, for both the location and scale components, come from the same multivariate normal distribution. Additionally, due to the hierarchical formulation, these estimates will not often be equivalent to estimating the empirical means and variances. Indeed, in this model, the parameters share information (i.e., partial pooling of information) which can lead to improved parameter estimates due to shrinkage towards the fixed effect average (Efron & Morris, 1977; Stein, 1956). This is a defining feature of mixed-effects estimation, and also applies to location scale models.

We fitted this model for both the congruent and incongruent conditions with the R package `brms` (Bürkner, 2017). The parameter estimates are displayed in Figure 1. Here

"Hierarchical" (blue) refers to those obtained from the model, whereas "Empirical" (orange) refers to the sample based estimates.[1] The results highlight the central idea behind shrinkage, in that the model based estimates are often closer to the fixed effect average. This gravitation towards the average was especially apparent for the larger values. Of course, it could be that modeling individual variances resulted in overfitting the data. We thus made comparisons to models that assumed the customary constant variance with WAIC (Watanabe, 2010), which is analogous to AIC, but accounts for posterior uncertainty (Vehtari, Gelman, & Gabry, 2017). The differences in WAIC were at least 5.5 standard errors away from zero, which indicates that the MELSM was preferred over a traditional mixed-effects model.

Note that intraclass correlation coefficients (ICC) are computed from a random intercepts model that assumes a common variance for each person (or group). As such, for these models, the estimate of reliability would not fully capture the variability and (at best) indicate the average ICC.[2] This insight is made possible with the MELSM, and it also provides a solution for computing ICCs with heterogeneous variances. We return to this topic in the discussion (Section Future Directions).

We now discuss the mean-variance relations displayed in the scatter plots. The shrinkage can be inferred from these plots–i.e., the sample base estimates are more widely dispersed around the fitted line. For both outcomes, the correlation was larger for the models based estimates (incongruent: 0.677 vs. 0.612; congruent: 0.719 vs. 0.663). It is possible that, while the point estimate is larger, there is more uncertainty. We thus used a nonparametric bootstrap to estimate the standard error of the empirical correlations. The methods were very similar in this respect (incongruent: 0.061 vs. 0.059; congruent: 0.054 vs. 0.057).

---

[1] R-code: `mean(.)` and `sd(.)`

[2] We also fit a random intercepts model with both outcomes, which is how reliability would be assessed in practice. The MELSM was again preferred.

Together, this simple example illustrated several benefits of this novel approach for characterizing mean–variance relations. For example, possible concerns of overfitting were quelled, along the way it became apparent that an alternative method was warranted for computing reliability, the hierarchical formulation reduced variability in the estimates, and nothing was lost in terms of the estimating the correlation. In fact, the model based estimates also appeared to offer some advantages compared to the sample estimates. In our experiences, we have found that it is difficult to detect random effects correlations with WAIC (and LOO; Vehtari et al., 2017). Further, these Bayesian information criteria address the question of expected predictive accuracy on new data (Gelman, Hwang, & Vehtari, 2014; Vehtari & Ojanen, 2012), and do not explicitly *test* relations. Although it would be possible check the credible interval for zero, we present a flexible approach that employs Bayesian hypothesis testing.

Hypothesis Formulation

Bayesian hypothesis testing is synonymous with model comparison. In contrast to classical testing (i.e., using p-values), the Bayesian approach provides a measure of *relative* evidence for which model is most supported by the data. Thus there must be at least two models under consideration. In the case of two models, the prior distribution is commonly referred to as the alternative hypothesis (i.e., $\mathcal{H}_1$ or $\mathcal{M}_1$), which is then compared to the null hypothesis (i.e., $\mathcal{H}_0$ or $\mathcal{M}_0$). The Bayes factor then provides a measure of relative evidence for the competing models, for example, with $BF_{10} = 10$, this would indicate $\mathcal{H}_1$ is 10 times more likely than $\mathcal{H}_0$. Further information about Bayesian inference can be found in the many introductions specifically for psychology (Quintana & Williams, 2018; Rouder, Speckman, Sun, Morey, & Iverson, 2009; Wagenmakers, Love, et al., 2018; Wagenmakers, Marsman, et al., 2018).

In this work, we take the encompassing prior (EP) approach for computing Bayes factors (Klugkist, Kato, & Hoijtink, 2005a, 2005b). This simplifies model comparison,

because the prior distribution only needs to be specified for the unconstrained (encompassing) model ($\mathcal{M}_u$). For two competing models, $\mathcal{H}_1$ and $\mathcal{M}_u$ can be used interchangeably because they denote the predicted effect. The EP approach was originally developed for order restrictions, for example a hypothesized order for the magnitude of several correlation coefficients. In (Wetzels, Grasman, & Wagenmakers, 2010), it was extended to exact equality constraints such as the null hypothesis–i.e.,

$$\begin{aligned} \mathcal{H}_0 &: \rho_{ij} = 0 \\ \mathcal{H}_u &: \rho_{ij} \neq 0,\ 1 \leq j < i \leq k. \end{aligned} \tag{6}$$

Here $1 \leq j < i \leq k$ denotes the lower-triangular elements of $\mathbf{\Omega}$ (4), j denotes the column, and i the row. Note there are $\frac{1}{2}k(k-1)$ correlations in total. In (6), $\mathcal{H}_0$ is an equality constrained null hypothesis and $\mathcal{H}_u$ is the unconstrained hypothesis. These hypotheses are nested in that $\mathcal{M}_0 \subseteq \mathcal{M}_u$. Indeed, in a classical testing framework, the likelihood ratio test is one approach for testing elements of covariance matrices (e.g., pp 87 - 88; Højsgaard, Edwards, & Lauritzen, 2012) and random effects correlations (Baayen, Davidson, & Bates, 2008). It is common to use *convenience* priors to simplify the computation of the Bayes factor. In particular, those that allow for converting R^2 (Rouder & Morey, 2012) and certain test statistics to the corresponding Bayes factor (Johnson, 2005). These are restricted to relatively simple models, and to our knowledge, alternative techniques are necessary for mixed-effects locations scale models. We thus use the Savage-Dickey ratio (Dickey, 1971; Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010), which is also computationally convenient and reasonably accurate. It has been used for mixed-effects models (De la Cruz, Meza, Arribas-Gil, & Carroll, 2016), correlations (Nuijten, Wetzels, Matzke, Dolan, & Wagenmakers, 2015), and partial correlations (Williams & Mulder, 2019). Because we work directly with the correlations, the hypothesis test in favor of the null hypothesis can be formulated as

$$BF_{0u_{ij}} = \frac{p(\mathbf{Y}|\mathcal{H}_0)}{p(\mathbf{Y}|\mathcal{H}_u)} = \frac{p(\rho_{ij} = 0|\mathbf{Y}, \mathcal{H}_u)}{p(\rho_{ij} = 0|\mathcal{H}_u)}. \tag{7}$$

In words, by only considering $\mathcal{H}_u$ with respect to ρ_{ij}, the Bayes factor can be computed as the unconstrained (marginal) posterior density of ρ_{ij} evaluated at zero divided by the prior density also evaluated at zero (Mulder, Hoijtink, & de Leeuw, 2012). Consequently, the necessary ingredients for computing (7) are posterior samples drawn from the full model with Markov chain Monte Carlo simulation. Note that we have restricted our focus to the test-relevant parameters (i.e. ρ_{ij}), because the others (e.g.., the fixed effects) are considered nuisance parameters in this formulation–i.e., we are explicitly interested in testing the mean-variance relations captured in $\mathbf{\Omega}$.

What remains is computing the density at the test value of zero. While this could be computed with the `logspline` function in R (Deng & Wickham, 2011), we instead further simplify the computation by employing the Fisher Z transformation. This approach for achieving (approximate) normality has been used recently in the context of testing non-zero correlations (Mulder, 2016; Williams & Mulder, 2019; Williams, Rast, Pericchi, & Mulder, 2019). In this case, it is used because it allows for generalizing this formulation to hypotheses with several test-relevant parameters (see: Mulder, 2016; Williams & Mulder, 2019), it is also known to stabilize variance (Konishi, 1985), and it overcomes one limitation of the Savage-Dickey ratio. Namely, the accuracy of density estimation is sensitive to the number of samples near the test value (Wagenmakers et al., 2010).

The transformation is applied to the posterior and prior samples. Denote the latter $\xi^{(s)}$ and then transformed posterior is obtained as

$$\xi_{ij}^{(s)} = F(\rho_{ij}^{(s)}) = \frac{1}{2}\log\left(\frac{1+\rho_{ij}^{(s)}}{1-\rho_{ij}^{(s)}}\right), \text{ for } s = 1, ..., S. \tag{8}$$

To make clear the computation, first let $\zeta^{(s)}$ denote the prior samples (from the LKJ distribution for the correlation among the random effects) and assume that they are transformed with (8). The approximate Bayes factor, assuming normality, is then

$$
\begin{aligned}
BF_{0u_{ij}} &= \frac{f(0|\mu_\xi, \sigma_\xi)}{f(0|0, \sigma_\zeta)} \\
&= \frac{\sigma_\zeta}{\sigma_\xi} \exp[-\frac{-\mu_\xi^2}{2\sigma_\xi^2}] \\
&= \frac{\frac{1}{\sigma_\xi\sqrt{2\pi}} \exp[-\frac{(0-\mu_\xi)^2}{2\sigma_\xi^2}]}{\frac{1}{\sigma_\zeta\sqrt{2\pi}}}.
\end{aligned} \tag{9}
$$

In (9), μ_ξ and σ_ξ denote the posterior mean and standard deviation, respectively, whereas σ_ζ is the prior standard deviation. The prior mean is fixed to the test value of zero. This can be solved with `dnorm(.)` in R.

In the psychological literature, there are theoretical predictions for the relation between the mean and variance. For example, in RT tasks in particular, it has been noted that slower times are accompanied by more variability (Wagenmakers & Brown, 2007). Thus, in this case, a one-sided hypothesis can be formulated which follows naturally because the Bayes factor is transitive. First, we define the directional hypothesis as $\mathcal{H}_1 : \rho_{ij} > 0$. The Bayes factor is then

$$
BF_{1u} = \frac{\int_0^\infty p(\xi_{ij}|\mathbf{Y}, \mathcal{H}_u) d\xi_{ij}}{\int_0^\infty p(\xi_{ij}|\mathcal{H}_u) d\xi_{ij}} = \frac{\Pr(\xi_{ij} > 0|\mathbf{Y}, \mathcal{H}_u)}{\Pr(\xi_{ij} > 0|\mathcal{H}_u)}, \tag{10}
$$

where the numerator and denominator denote the posterior and prior probability that ξ_{ij} is greater than zero under the unrestricted model $\mathcal{H}_u$. This is again computed assuming normality. The relative evidence for $\mathcal{H}_1$ against $\mathcal{H}_0$ is $BF_{u0} \cdot BF_{1u}$.

Exhaustive Testing. A defining feature of Bayesian hypothesis testing is the ability to assess which theoretical model best predicts the data at hand (Kass & Raftery, 1995; Lee, 2018). However, in the absence of guiding theory, it is likely that a more exploratory approach is warranted, such as in research areas that have not yet formalized

the mean–variance relation. We thus present an exhaustive approach–i.e.,

$$\begin{aligned} \mathcal{H}_0 &: \rho_{ij} = 0 \\ \mathcal{H}_1 &: \rho_{ij} > 0 \\ \mathcal{H}_2 &: \rho_{ij} < 0, \end{aligned} \tag{11}$$

which covers the entire parameter space. The evidence for each can then be computed from the inverse of the Bayes factor given in (9)–i.e., BF_{u0} (evidence for the unrestricted model). The computation of each one-sided Bayes factor for $\mathcal{H}_1$ and $\mathcal{H}_2$ (BF_{10}) is given in (10). With the the three Bayes factors in hand $BF_{tu}, t = 0, .., 2$, each compared to the unrestricted model $\mathcal{M}_u$, the posterior hypothesis probabilities can then be computed. For example, the probability of $\mathcal{H}_1$ is

$$p(\mathcal{H}_1|\mathbf{Y}) = \frac{p(\mathcal{H}_1)BF_{1u}}{p(\mathcal{H}_0)BF_{0u} + p(\mathcal{H}_1)BF_{1u} + p(\mathcal{H}_2)BF_{2u}}, \tag{12}$$

where $p(.)$ is the prior probability for each hypothesis. For this work, we assume equal prior probabilities. This results in the posterior probability for each hypothesis. In other words, this provides an *exhaustive* assessment of the previously stated hypotheses. The ratio of these probabilities, for example $p(\mathcal{H}_1|\mathbf{Y})/p(\mathcal{H}_2|\mathbf{Y})$, then provide the Bayes factor for which direction is most supported by the observed data. Importantly, we emphasize that these probabilities are not *unconditional* but depend on the models under consideration and the prior distribution ($\mathcal{H}_u$).

What remains is the choice of $\mathcal{H}_u$, or the prior distribution, which is of critical importance when computing the Bayes factor (Bartlett, 1957; Jeffreys, 1961; Lindley, 1957). While it is customary to assume diffuse priors even for Bayesian model selection, such as a uniform distribution for correlations between -1 and 1 (Marsman & Wagenmakers, 2017), we do not follow this approach. This is because testing exact equality constraints is often accompanied by asymmetric information (Gu, Hoijtink, & Mulder, 2016)–i.e., too wide of a prior, that places too much density in implausible regions, can result in incorrectly accepting the null hypothesis when the effect is small or even moderate in size.

To better understand the LKJ marginal distribution, and thus $\mathcal{H}_u$, we drew samples from it. We set the dimensions to $k = 4$ (the matrix size in the full model) and varied $\nu \in \{1, 2, 3 \text{ and } 4\}$. These densities are plotted in Figure 2. Here the non-invariance to k is apparent. Even with $\nu = 1$ (50% HDI $\approx [-0.35, 0.35]$)[3], the marginal distribution is far from uniform. Further, the density concentrates around zero as $\nu \to \infty$, with $\nu = 4$ stating there is approximately a 50% probability that $\rho_{ij} > |0.20|$. As a baseline, we chose $v = 2$ because it follows the convention for what is commonly interpreted as a medium size effect. Specifically, there is approximately a 50 % probability that $\rho_{ij} > |0.30|$. This prior cannot be said to reflect our beliefs (Morey, Romeijn, & Rouder, 2016), as is sometimes suggested in the psychological literature, and it is not purely objective (Berger, 2006), but instead it is somewhere in between: given the data at hand, does a model that predicts (at least) a medium effect ($\mathcal{M}_u$) provide a better (relative) fit than a model that predicts no effect ($\mathcal{M}_0$) at all? We do believe, however, that comparing these models provides useful information. We include a sensitivity analysis below.

Summary

To this point, we have presented a random intercepts model to highlight the central idea behind the MELSM (Section Random Intercepts Only Model), and in particular, how it can hierarchically model mean-variance relations (Figure 1). This parameterization lends itself naturally to testing random effects correlations with the Bayes factor (Section Hypothesis Formulation), for example those capturing the associations between the location and scale individual effects (i.e., 3). In the following, we apply this framework to two well-known cognitive paradigms–i.e. Stroop and Flanker tasks. Importantly, we carefully stay true to each literature by fitting the customary model to the mean structures, and then expand this to also include a sub-model to the variance. We emphasize the random effects correlations and also highlight differences from traditional mixed-effects models.

[3] HDI: Highest Density Interval

Illustrative Examples

We first turn to the Stroop and then proceed to the flanker interference task. These data were first used in von Bastian et al. (2016), and both have also been used in methodological inquires (e.g., Haaf & Rouder, 2017). For the mean structure, we fit the model that was described in the later as the least constrained. There were random effects for the intercept and slope. That is, the mean reaction time for the congruent condition was allowed to vary among the individuals (the intercepts), in addition to the incongruent effect that is captured by individual deviations from the average "Stroop/flanker Effect" (the slopes). Note that we are answering a much different question than Haaf and Rouder (2017), where the focus was on the direction of individual effect. The covariance between the location and scale was not tested.

Data Set 1: The Stroop Task

In this task, participants were asked to count the number of characters displayed with key strikes. This is the *number* Stroop task. For the congruent condition, the number of characters matched the digits displayed–e.g., 3 characters shown as 333. For the incongruent condition, there was a mismatch between the number of characters and the digits displayed–e.g., 2 characters presented as 44. There were 121 participants in total. Each completed 48 trials for the two conditions. We chose these data because there is a large effect, with a mean difference of 65 milliseconds (ms), and also individual differences in the "Stroop effect" $\chi^2(3) = 10.94, p = 0.012$.[4]

Data Set 2: The Flanker Task

In this task, the goal was to identify a vowel (e.g., A or E) or consonant (e.g., B or C). The target was located in the middle, and was "flanked" by two characters on either side. The congruent flankers surrounded the target with letters from the same

[4] We compared a random intercept model to a random slope model with the R package lme4 `lmer(.)`

category–e.g., UUAUU. The incongruent flankers were surrounded by mismatched letters–e.g., CCACC. There was also a neutral condition (##A##), but we only used the congruent and incongruent trials. There were 121 participants in total. Each completed 48 trials for the two conditions. We explicitly chose these data to contrast the Stoop data. The mean difference was small (2 ms) and there were no significant individual differences $\chi^2(3) = 2.03, p = 0.566$. This may seem paradoxical on the "surface", but demonstrates the central idea of this work. That is, assessing within-person variance can provide useful information for understanding psychological processes.

Model Parameterization

Mean Structure. We fit the same model to both data sets. The outcome is reaction time for correct responses, and the one predictor is the experimental condition. For each outcome y, the location sub-model of the response times for the ith person and jth trial is given as:

$$\begin{aligned} y_{ij} \sim \beta_0 + \beta_1(\text{congruency}_{ij}) \\ + u_{0i} + u_{1i}(\text{congruency}_{ij}) + \epsilon_{ij} \end{aligned} \tag{13}$$

Here β_0 is the fixed effect intercept, which in this case, is the average reaction time for the congruent condition (the reference category). The predictor is congruency that indexes the respective experimental condition (congruent vs. incongruent). The fixed effect, β_1, is the average difference from the congruent condition–i.e., the experimental effect. This model also includes random intercepts u_{0i} that provide each person's deviation from β_0, as well as random slopes u_{1i} that capture variability in the "Stroop/flanker effect" (i.e., β_1).

Variance Structure. The above is a traditional individual differences model, in that the "errors" or not modeled. This is not the case for the MELSM. We extend the model to also consider the residual, or within-person variance, structure–i.e.,

$$\sigma^2_{\epsilon_{ij}} \sim \exp[\eta_0 + \eta_1(\text{congruency}_{ij}) + u_{2i} + u_{3i}(\text{congruency}_{ij})]. \tag{14}$$

The subscripts i and j denote residuals for the ith person and jth trial, respectively. These parameters are analogous to those in (13), but are inherently within-person effects. They differ from those in Figure 1, because the "errors" are with respect to the predictor in (13). That is, this model seeks to capture systematic patterns in the residual variance. Therefore, η_0 and η_1 denote the average within-person variance for the congruent condition and the "Stroop/flanker effect," respectively. The random effects, u_{2i} and u_{3i}, are the hierarchical, individual deviations, from those averages. Substantively, they can be interpreted as within-person instability, fluctuations, or conversely, as stability in their responses. To be clear, this formulation allows for modeling individual differences in the congruency effect (Equation 13), and the effect of congruency on within-person variance (Equation 14).

Random Effects Distribution. We assume the random effects, for both the location and scale, are drawn from a common multivariate normal distribution–i.e.,

$$[u_{0i}, u_{1i}, u_{2i}, u_{3i}]' \sim N(\mathbf{0}, \mathbf{\Sigma}). \tag{15}$$

This allows for testing mean–variance relations. We again re-express the covariance matrix as $\mathbf{\Sigma} = \boldsymbol{\tau}\mathbf{\Omega}\boldsymbol{\tau}'$. Note that τ_l, for $l = 0, 1, 2, 3$, corresponds to the random effects standard deviation $SD(u_{li})$. Individual differences in the congruency effect are captured in $SD(u_{1i})$. This can be compared to, for example, a model that fixes it to zero $SD(u_{1i}) = 0$. This can help determine whether a "common effect" model is supported by the data (Haaf & Rouder, 2017). In classical testing, on the other hand, this is analogous to testing for random (varying) slopes with the likelihood ratio test (Baayen et al., 2008). The same logic applies to the scale effects. For example, $SD(u_{2i})$ captures individual, within-person variance differences, in the congruent condition (η_0). Substantively, this is (in)consistency

in congruent responses. Further, $SD(u_{3i})$ is the spread of individual differences in the effect of incongruency on within-person variance. In our experience, we have found it conceptually difficult to test random effects variances, in particular for the scale, and instead prefer a more descriptive approach that relies on visualization.

The lower-triangular of the correlation matrix for the location and scale random effects, is then

$$\mathbf{\Omega} = \begin{bmatrix} 1 & & & \\ \rho_{01} & 1 & & \\ \rho_{02} & \rho_{12} & 1 & \\ \rho_{03} & \rho_{13} & \rho_{23} & 1 \end{bmatrix}. \tag{16}$$

The off-diagonal element entry at $\mathbf{\Omega}_{1,2}$ is the correlation (ρ_{01}) between reaction times in the congruent condition and the effect of congruency. This is also provided by a traditional mixed-effects model. $\mathbf{\Omega}_{4,3}$ is the relation (ρ_{23}) between within-person variance in the congruent condition attributable to the "Stroop/flanker effect." The remaining elements, $\mathbf{\Omega}_{3:4,1:2}$, correspond to the correlations among the location and scale random effects. Because they are the primary focus in this work, and lead to novel inferences, we explain them in detail here:

1. $\mathbf{\Omega}_{3,1}(\rho_{02})$: The correlation between reactions times and within-person variability for the congruent condition $cor(u_{0i}, u_{2i})$–i.e., the word "Red" displayed in the color red. This relation is displayed in Figure 2, but importantly, in this case, it captures within-person fluctuations, or conversely stability, in response times.

2. $\mathbf{\Omega}_{4,1}(\rho_{03})$: The correlation between reactions times for the congruent condition and within-person variability that can be attributed to incongruent condition $cor(u_{0i}, u_{3i})$. Interestingly, this allows for testing whether faster (or slower) individuals in the congruent condition were relatively more (or less) variable in the incongruent condition. That is, perhaps faster individuals in ideal conditions are more (or less) sensitive, with respect to stability, to less than ideal conditions.

3. $\boldsymbol{\Omega}_{3,2}(\rho_{12})$: The correlation between the "Stroop/flanker effect" and within-person variance in the congruent condition $cor(u_{1i}, u_{2i})$. This is similar to the previous, but the question asked is slightly different. In this case, it captures whether those with the largest (or smallest) effects were also more (or less) variable in the congruent condition.

4. $\boldsymbol{\Omega}_{4,2}(\rho_{13})$: The correlation between the "Stroop/flanker effect" on reaction times and variability $cor(u_{1i}, u_{3i})$. This is perhaps the most interesting relation for the reaction time literature in particular, because slower individuals are predicted to be more variable. In this case, it is not whether slower people are more variable in general, but whether those with the largest "Stroop/flanker effects" were relatively, in reference to the congruent condition, more variable in their responses.

Prior Specification. We assumed the following priors for the mean and variance fixed effects

$$\beta_0, \beta_1, \eta_0, \eta_1 \sim N(0, 5). \tag{17}$$

These priors do not express our beliefs, but rather are used to simplify the model formulation. We pay more attention to specifying the random effects prior distributions. Note, from (15), that the random effects share the same multivariate normal distribution with the covariance matrix $\boldsymbol{\Sigma}$. The prior distributions are defined as

$$\begin{aligned} \boldsymbol{\Sigma} &= \boldsymbol{\tau}_l \boldsymbol{\Omega} \boldsymbol{\tau}_l \\ \boldsymbol{\tau}_{1:2} &\sim N^+(0, 0.25) \\ \boldsymbol{\tau}_{3:4} &\sim N^+(0, 1) \\ \boldsymbol{\Omega} &\sim \text{LKJ}(\nu = 2). \end{aligned} \tag{18}$$

Here $\boldsymbol{\tau}_{1:2}$ are the location random effects SD's. These capture the *spread* of response times for the congruent condition (the intercepts) and the "Stroop/flanker effect" (the slopes). In this case, this prior states that 95 % the individual effects will be within 300 milliseconds of the averages across individuals.[5] For the scale random effects SD's (i.e., $\boldsymbol{\tau}_{3:4}$), on the other hand, we chose *weakly* informative prior distributions. The (marginal) LKJ prior for the correlations is displayed in Figure 2, and is used to test the mean–variance relations (Section Hypothesis Formulation).

Software and Estimation. All computations were done in R version 3.5.2 (R Core Team, 2017). The models were fitted with the the package `brms` (Bürkner, 2017), which serves as a front-end to the probabilistic programming language Stan (Stan Development Team, 2016). There are several advantages of the package `brms`. The model specification follows that of `lme4` (Bates, Mächler, Bolker, & Walker, 2015), although `brms` allows for fitting a much wider range of models. Additionally, there are several post-processing features for model checking and Bayesian hypothesis testing. Each fitted model included four chains of 2,500 samples each, excluding a warm-up period of 1,000 samples. This resulted in a total of 10,000 draws from the posterior distribution. Each parameter is summarized with the posterior mean and standard deviations, as well as 90 % credible intervals (CrI). This number of samples provided a good quality of the parameter estimates in which the models converged with potential scale reduction factors $\hat{R}$ smaller than 1.1 (Gelman, 2006). The R code for this work is provided online (X)

Results

Comparison to Mixed-Effects Models

We first compared the MELSM to a traditional, individual differences, mixed-effects model (MEM). The models are the same (including the priors), but for the MEM, only the mean structure was specified (Equation 13). Further, it is important to note the MEM can

[5] The standard deviation of the half-normal distribution is $\sqrt{\sigma^2(1-\frac{2}{\pi})}$.

be understood as predicting the residual variance with an intercept (η_0) and the remaining scale effects are implicitly set to zero. We first compared the models with WAIC. This was again done to explicitly address the possibility of overfitting. The WAIC difference was 9.85 standard errors from zero for the Stroop task, whereas the difference was 10.54 standard errors from zero for the flanker task. Both provided substantial support for the MELSM, such that the MEM is expected to have *worse* out-of-sample predictive performance.

Figure 3 (panel A) includes the hierarchical estimates–i.e., $\beta_1 + u_{1i}$. These capture the effect of congruency for each person. There are clear differences between the MEM and MELSM. In reference to the empirical estimates (grey line), there was more shrinkage towards β_1 for the MEM. This was non-trivial, in that each would lead to a different conclusion. On the one hand, because only two individual differed from the average, there do not appear to be notable individual differences in the Stroop task. However, this assume that the residual variance is a fixed, non-varying constant. On the other hand, with the MELSM, 24 % of the individual effects differed from the average (β_1). These deviation were perhaps small, but nonetheless important to consider. The MELSM estimates were also more variable in terms of uncertainty (credible interval width), whereas the MEM based estimates appeared have the same width. These differences can be understood in reference to Best Linear Unbiased Prediction (BLUP) of random effects, which are computed assuming a common residual variance. We emphasize this is not the case for the MELSM, although it reduces to a MEM when the (non-intercept) scale fixed and random effects are actually zero. In other words, when the implicit scale model of the MEM (i.e., $\sigma^2_{\epsilon_{ij}} \sim \eta_0$) is warranted for the data.

A similar pattern was revealed for the flanker task, but the differences between models were even more pronounced. Note that, on average (β_1), there did not appear to be an effect of congruency (Table 1). Whereas this seemed to hold at the individual level for the MEM, there were individual effects in reference to zero for the MELSM. Indeed, $\approx 10\%$ had a negative effect that indicated faster response times in the incongruent condition.

Further $\approx 15\%$ had an effect in the expected direction. There were also individual differences, in that 24% of the sample differed from β_1. We further explored the models with respect to shrinkage. This is plotted in Figure 3 (panel B), which reveals the *severe* shrinkage that can occur when treating the residual variance as a fixed, non-varying constant, when this assumption is not warranted (as indicated by WAIC). Note that the MELSM provided shrinkage that appeared more reasonable, in that the individual effects gravitated towards the averages for $\beta_0 = 0.56$ and $\beta_1 = 0.00$. On the other hand, the MEM provided shrinkage more so towards the average slope (β_1). Most importantly, the two different model specifications would again lead to different inferences. The MEM could be suggested to reflect measurement error (i.e., "noise"), but from our perspective, a necessary condition of "noise" is that systematic patterns are absent. As we describe below (Table 1), there is actually more individual variation in the "noise," that is in the within-person variance, than in the mean structure.

Case 1: The Stroop Task

Mean Structure. The fixed and random effects are reported in Table 1. The effect of congruency was obtained as the difference from the *in*congruent to the congruent condition and resulted in 65 ms (Incongruent Δ (β_1) in Table 1), which reproduced the result in Haaf and Rouder (2017). In this case, because we are also fitting a mixed-effects model to the residual variance, this demonstrates the MELSM can accurately estimate the mean. This perhaps eases some concerns about modeling the residuals. Importantly, this similarity is for the average effect across individuals, but there were notable differences for the individual effects (Figure 3).

Variance Structure. Modeling individual differences in the within-person variance structure is the central idea of this work, and this is unique information provided by the MELSM. Figure 4 (panel A) displays the individual, within-person variance, estimates for the congruent condition. There were clear individual differences. Specifically, the standard

deviations ranged from 0.06–0.31. This indicates a five-fold increase from the least to the most variable individual in the congruent condition. To put this in perspective, for the congruent response times (i.e., the mean structure), the minimum–maximum range was 0.49–0.95 (ms), which corresponds to a two-fold increase. Panel A (bottom row) includes differences, in within-person variance, compared to the congruent condition. There was again individual variation to consider. The fixed effect (η_1) was 0.16 (Table 1), which indicates a 17 % increase in variability on average ($[\exp(0.16) - 1] \times 100$). This does not tell the full story, however, because 24 % of the sample actually became *less* variable (the opposite sign as the average). Said another way, for some participants, the responses were *more* consistent in the incongruent condition. Further, the change in variability ranged from a decrease of 28 % to an increase of 115 % ! At the same time, the "stroop effect" was positive across all participants (Figure 3). Overall, all participants were slower in the incongruent condition and some were also less variable. This runs contrary to the pattern revealed in Figure 1 where slower reaction times were accompanied by more variability. This apparently does not hold when the focus is on within-person effects *between* conditions. The relation between the two–the individual level effects of congruency on the mean and variance structures–is captured by a random effects correlation.

Mean–Variance Relations. The random effect correlations are displayed in Figure 5. The probabilities associated to each hypothesis are provided in Table 2. Panel A corresponds to $\mathbf{\Omega}_{3,1}$ and captures the relation between reaction times and within-person variability for the congruent condition. There was a large correlation ($r = .68$), and overwhelming evidence for $\mathcal{H}_u$ ($BF_{10} \approx 1.6 \times 10^{19}$). Panel C ($\mathbf{\Omega}_{4,1}$) relates the average reaction time of each individual in the congruent condition (y-axis) to average within-person SD in the incongruent condition (x-axis), reflecting the "Stroop effect" in the scale. The negative correlation ($r = -0.40, BF_{10} \approx 400$) suggests that participants who were generally slower in their congruent responses, tended to be *less* variable in their incongruent condition responses. Said another way, being fast in the congruent condition

seems to be associated with relatively *more* variability in the incongruent condition. Panel B ($\mathbf{\Omega}_{3,2}$) captures the relation between the "Stroop effect" on the location, and within-person variance (scale) in the congruent condition. While the correlation was negative ($r = -0.19$), the (relative) evidence did not point towards either the Null or the alternative model. Indeed, as can be seen in Table 5, $p(\mathcal{H}_0) = 0.42$ and $p(\mathcal{H}_2) = 0.53$. The correlation reflected in Panel D is perhaps the most interesting, as it captures the association between the effect of the "Stroop effect" on reaction time and within-person variability ($\mathbf{\Omega}_{4,2}$). This relation was the largest ($r = 0.80, BF_{10} = 2.7 \times 10^4$), although the posterior SD was larger than for $\mathbf{\Omega}_{3,1}$ (panel A), which resulted in a smaller Bayes factor. That said, according to Jeffery's categorization scheme (Jeffreys, 1961), the evidence was decisive. Substantively, those with the largest "Stroop effects" on reaction time also had the largest "Stroop effects" on within-person variability, in that the congruent condition slowed participants down and also increased their variability. As such, there is evidence for a positive "Stroop effect" in both–i.e., the location and the scale.

Case 2: The Flanker Task

Mean Structure. The fixed and random effects are reported in Table 1. The "flanker effect" in the location was practically zero ($\approx$ 2 ms), and the 90 % credible interval included zero. A non-significant finding has traditionally been considered uninteresting and not worth publishing. Bayes factors have recently been suggested to overcome this perspective (Dienes, 2014) by allowing one to gain evidence for the null hypothesis. However, while in most research endeavours the focus remains on the mean structure, we characterize the within-person variance structure.

Variance Structure. In contrast to the Stroop data, the effect of the congruency condition on variability (Incongruent Δ (η_1)) was small and the interval included zero. However, there were individual differences according to WAIC, as seen in Figure 4. The reaction times for the congruent condition were even more variable than those for the

Stroop task. The standard deviations ranged from 0.05–0.43. This is an eight-fold increase from the least to most variable individual (see panel A). On the other hand, the reaction times ranged between 0.43–0.80, a two-fold increase. Further, while on average there was a negligible effect (η_1), this does not readily generalize to each person. The 90 % CrI's excluded zero for $\approx 38\%$ of the sample. Interestingly, whereas for the Stroop task only two participants showed a "signficant" reduction in variability, $\approx 16\%$ of the participants were less variable in the incongruent condition for the flanker task (as indicated by 90 % CrI exclusion of zero). Substantively, this can be understood as behavioral (in)stability attributable to the flanker manipulation.

Mean–Variance Relations. The random effects correlation are displayed in Figure 6. The probabilities associated to each hypothesis are provided in Table 2. Panel A corresponds to $\mathbf{\Omega}_{3,1}$ and captures the relation between reaction times and within-person variance for the congruent condition. There was again a large correlation ($r = .81$), and overwhelming evidence for $\mathcal{H}_u$ ($BF_{10} \approx 1.6 \times 10^{19}$) indicating that slower individuals tend to be more variable in the congruent condition. Panel C corresponds to the correlation among congruent RT and and "flanker effect" on the within-person variance ($\mathbf{\Omega}_{3,2}$). The correlation was negative ($r = -0.12$), which reproduces the result obtained in the Stroop task, but the Bayes factor was inconclusive ($BF_{01} = 2.16$). Panel B ($\mathbf{\Omega}_{3,2}$), on the other hand, shows the relation between the "flanker effect" in the location and the within-person variance in the congruent condition. While the correlation was negative ($r = -0.21$), the (relative) evidence did not point towards either the null or the alternative model. The posterior hypothesis probabilities (Table 2) indicated that $p(\mathcal{H}_0) = 0.42$ and $p(\mathcal{H}_2) = 0.52$. The final correlation captures the association among the effect of reaction times and within-person variance for the incongruent condition (Panel D; $\mathbf{\Omega}_{4,2}$). Note that, on average, there was a negligible effect, on average, for both the mean and variance structures (Table 1). At the individual level, however, the result nonetheless paralleled the Stroop task. Those with the largest "flanker effects" on reaction time also had the largest

"flanker effects" on within-person variability. In fact, there was a nearly perfect correlation among the location and scale for the flanker task ($r = 0.87, BF_{10} = 9.2 \times 10^3$).

Sensitivity Analyses

Given the novelty of the approach we performed sensitivity analyses to chech the robustnes of the results. In this work, there are two primary sources that can influence the reported Bayes factors. The first is not specific to this model, but is in the nature of the Savage-Dickey ratio. That is, the accuracy diminishes with stronger support for $\mathcal{H}_1$. This is because there are fewer posterior samples around the test value of zero (Wagenmakers et al., 2010). We checked this by refitting the models, each with 10,000 saved posterior samples, and evaluating the stability of the Bayes factors. We found that they were stable up to around 400 in favor of $\mathcal{H}_1$. When following the customary guidelines (Jeffreys, 1961; Kass & Raftery, 1995; Lee, 2018), this indicates that the proposed method is "well within the accuracy required for drawing conclusion" (p. 777; Kass & Raftery, 1995).

The second source is $\mathcal{H}_u$, because Bayes factors can be sensitive to the choice of the prior distribution. We thus adjusted $\nu \in \{0.1, 1, 2, 3 \text{ and } 4\}$ of the LKJ prior. The value less than one resulted in an approximate uniform distribution. For larger values, the density increasingly concentrated around zero as $\nu \to \infty$ (Figure 2). The results are provided in Figure 7, wherein the grey regions denotes what is considered "positive" evidence for a given model (i.e., $BF_{10} > 3$ and $BF_{10} < 0.33$; Kass & Raftery, 1995). The largest Bayes factors in favor of $\mathcal{H}_1$ were robust to the choice of $\mathcal{H}_u$. The Bayes factors that pointed in the direction of $\mathcal{H}_0$, on the other hand, did cross the threshold of 3. In particular, for the flanker task and when assuming a uniform distribution, the relation between reaction times in the congruent condition and the "flanker effect" on variability $cor(u_{0i}, u_{3i})$ had a Bayes factor of $BF_{01} = 4.16$. Of course, this is considered only "positive" evidence that is not entirely convincing.

Summary

These illustrative examples lead to several interesting findings, that to our knowledge, are novel for these interference tasks. They were made possible by looking beneath the surface of the mean structure and to the within-person variance. We summarize the general points here:

1. There were individual differences in the within-person variance structure (Table 1). The degree of individual variation was often more pronounced than for the mean structure. For example, a four-fold (Stroop) and eight-fold (flanker) difference between the least and most variable response (Figure 4).

2. Despite a negligible "flanker effect" ($\approx$ 2 ms), where the 90 % CrI included zero, the individual estimates showed the same pattern in both tasks. Namely, those with the largest effects on reaction time also had the largest effects on within-person variance. (Figure 5 and 6; panel D).

3. There is a complex web of structural mean–variance relations (Figure 5 and 6). Some of these were expected (panel D), whereas others were perhaps paradoxical (panel B and C). In particular, the negative relation between reaction times in the congruent condition and the "Stroop effect" on within-person variance. Here the slowest individuals in the congruent condition were often more consistent (variability decreased) in the incongruent condition.

4. Slower responses were not always associated with more variability, which contradicts a well-known pattern in reaction time data (Wagenmakers et al., 2005). In both tasks, at the individual level, a substantial proportion of participants simultaneously became *slower* and also *less* variable in the incongruent condition (compared to the congruent condition).

There were also several interesting features of the mixed-effect location scale model, and in particular, compared to a customary individual differences model–i.e.,

1. The MELSM accommodates individual differences in variability. For example, the eight-fold difference between the least and most volatile response times (Figure 3; Panel A). This manifests in the individual random effects. Whereas the mixed-effects model suggests that each person has essentially the same uncertainty, the estimates for the MELSM reflect the respective within-person variances.

2. Both methods provided shrinkage towards the average. The mixed-effects model assumes the within-person variance is a fixed, non-varying constant. When this assumption is not warranted, as for both tasks (according to WAIC), this can result in *substantial* shrinkage (Figure 3; panel B), masking individual differences (Figure 3; panel A).

3. Even when the fixed effect of congruency was not "significant," the MELSM still provided important insights into within-person processes. That is, while on average the data appeared uninteresting (for the flanker task; Table 1), there were individual differences in the variance structure, that is in behavioral (in)stability, as well as mean–variance relations to consider (Figure 5). These additional levels of analysis are only available in the MELSM.

4. The customary mixed-effects model has an implicit scale model, where the within-person variance is predicted by an intercept–i.e., $\sigma^2_{\epsilon_{ij}} \sim \eta_0$. This would be overly restrictive for these data (Figure 4; panel B), and in our experience, we have found the MELSM is *always* preferable to a mixed-effects model. Concerns of overfitting can be addressed with WAIC as it directly answers the question of predicting unseen data.

Discussion

In this work, we introduced a novel framework for testing mean–variance relations. This was accomplished by extending the traditional individual differences model to include a sub-model for predicting the within-person variance. The model was conceptualized with the explicit goal of assessing the mean and within-person variance structures. This approach can simultaneously model both the mean (locations) and variance structure (scale) for repeated measures data that are common to psychology. The MELSM also provides random effects for both the location and the scale components, thus allowing for characterizing the uncertainty of individual effects. As a result, there are correlations between location and scale random effects, both across and within each structures. This opens the unique opportunity to assess mean–variance relations with Bayesian hypothesis testing.

Throughout this work, we emphasized that the residual variance can provide insight into psychological processes. That is, it can be modeled just like customary dependent variables such as observed reaction times. This includes the ability to characterize individual differences in the stability (or instbility) of behavioral responses, such as performance in RT tasks. The scale effects and mean–variance relations were framed in the context of substantive applications. However, another valid perspective is that they provide insight into reliability. For example, if larger effects are also associated with more variability, this suggests that those effects are less reliable, under certain conditions or for some individuals. From this perspective, it is certainly problematic that the control, or congruent condition, had such a large range of within-person variance. Indeed, to improve reliability, a natural target would be to investigate procedures that did not invoke such large individual differences. This would shift the average residual variance towards smaller values, and necessarily increase reliability, assuming that between-individual variance remains constant. Thus the MELSM can also be used to investigate possible targets for reducing measurement error, although we emphasize, in our opinion, modeling the residuals

allows for rich inferences.

Limitations

There are notable limitations of this work in particular, as well as the MELSM in general. For the latter, when the variances are of interest, it should be noted that their magnitude is also defined by the location of the average response. In other words, with bounded variables such as reaction times, the variance will be a function of the person's mean (Baird, Le, & Lucas, 2006; Eid & Diener, 1999; Kalmijn & Veenhoven, 2005; Rouder, Tuerlinckx, Speckman, Lu, & Gomez, 2008). In the present work, this particular correlation $cor(u_{0i}, u_{u_2 i})$ was large for both tasks. This can also be seen in Figure 1, although this is the actual variance and not the within-person variance (no predictors were included in the model). This could be an effect of substantive interest, or dictated by aspects of the study design. We refer interested readers to Mestdagh et al. (2018), where a "mean-corrected" measure of variability was introduced. Importantly, this is not necessarily a "problem" for the MELSM but it should be considered when making inference about the mean–variance relation. This is less of a concern for the the other correlations (Figure 5; panels B-D); for example, the "Stroop/flanker effect" on reaction times and within-person variance (Figure 5; panel D).

Moreover, while these reaction times were right skewed, we nonetheless assumed normality. This choice is not without precedent, as Rouder, Kumar, and Haaf (2019) and Rouder et al. (2019) also made this assumption for these same data. Further, Schramm and Rouder (2019) recently noted that transformations did not seem to offer advantages compared to assuming normality. In our view, the research question should determine the modeling strategy (Rousselet & Wilcox, 2019). We were specifically interested in the mean and variance. In the case of skewed data, both can be estimated accurately when assuming normality (e.g., Figure 1). While there are several distributions that accommodated skew (Table 1 in: Wagenmakers & Brown, 2007), the mean and variance are typically a function

of the distributional parameters. This is also applies to the log-normal distribution, although the location and scale retain their interpretability as the mean and variance. There is a caveat, however, in that the outcome is then on the log scale–i.e., the mean and variance of $\log(y)$. Thus, the log-normal distribution seems ideal for those not comfortable assuming normality. In order to assess the impact of the assumed distribution, we also fitted log-normal models for each task. The results were very similar to those reported here, and importantly, the same pattern emerged in the random effects correlations. Furthermore, because the correlations were so striking, we also examined the empirical means and standard deviations. Here we subtracted the individual means and SD's, which provided a naive contrast for both. We then correlated differences, which mimics Figures 5 and 6 (panel D). The same pattern emerged, in that there were *large* correlations for both tasks. Together, this suggests that our results are not an artifact of the model, but can also be seen in the empirical estimates.

Second, although there are several interference data sets publicly available, we looked at only these two. They stand in contrast to Haaf and Rouder (2017) and Rouder et al. (2019). At some point along the way, we determined that including any more information would be overwhelming. However, because some of our findings were perhaps surprising and counterintuitive, we briefly looked at some of those other data sets. At a glance, the random effects correlations were similar in their direction to those revealed in Figure 5. We address this further in Future Directions.

Third, the LKJ prior distribution is not invariant to the dimensions of the correlation matrix. This can be overly restrictive for Bayesian hypothesis testing in particular, in that the lower (or upper) bound for the Bayes factor will depend on the dimension (Williams & Mulder, 2019). We refer interested readers to the matrix-F prior distribution which scales better as the dimension of the matrix increases. Technical details can be found in Mulder and Raúl Pericchi (2018), with psychological applications provided in Williams and Mulder (2019), Williams, Rast, et al. (2019), and Williams, Liu, et al. (2019). This did not present

problems in this work, but for larger matrices we would use the matrix-F distribution. This can be implemented directly in Stan (Stan Development Team, 2015).

Future Directions

This work points towards several interesting future directions. These span from methodological to theoretical inquires. Our intention was not to consider reliability, or intraclass correlation coefficients (ICC), of these measures. In the motivating example (Section Random Intercepts Only Model), we saw that treating the variance as a fixed, non-varying constant was problematic for a relatively simple model that included only random intercepts. This is the same model that would be used to compute ICC's in practice. When scaling to the full model, it also became apparent that a customary mixed-effects model can result in *substantial* shrinkage. This could be interpreted as reflecting measurement error ("noise;" Haaf & Rouder, 2017), or more generally, that the ratio of between-individual to within-person variance approaches zero (e.g., Figure 4 in: Gelman, Hill, & Yajima, 2012). Importantly, the hierarchical estimates (and their variance) are weighted by a "shrinkage factor" that assumes a common residual variance (Equations 3 and 4 in: Gelman & Pardoe, 2012). As such, large individual differences in the within-person variance will unduly penalize (some) individual estimates. In comparison, the shrinkage was less pronounced for the MELSM as it can accommodate systematic patterns in the distribution of residual, underscoring that residual variance is not necessarily on white noise. In terms of reliability, this points towards moving beyond considering reliability as a fixed quantity and allowing it to vary over conditions and individuals. This possibility was briefly described in Hedeker et al. (2008), and this work motivates fully characterizing the idea of varying ICC's.

These results also have theoretical implications for reaction time modeling. That is, there is much to gain from looking beneath the surface of the mean and to within-person variance. The reaction time literature is in a unique position because the mean–variance

relation is predicted from theoretical models, such as the classical diffusion model, and several papers have explicitly discussed the interplay between the two (Wagenmakers & Brown, 2007; Wagenmakers et al., 2005). Indeed, the positive relation between slower reaction times and increased variability was termed a "law" (Wagenmakers, 2007). There has also been substantial interest in tasks that do not show this relation (e.g., the Simon task; Haaf & Rouder, 2017; Pratte, Rouder, Morey, & Feng, 2010). This work provides a framework for testing theoretical predictions about mean–variance relations; for example, hypotheses that go beyond what has been described in the extant literature (e.g., Figure 1). However, the "law" should first be clarified for within-person effects. For example, if slower reaction times are predicted to be more variable, several people in both tasks *violate* the "law" (Figure 5; panel D). On the other hand, if the slowest individuals are predicted to be the most variable, then the "law" would be *confirmed* in both tasks (Figure 5; panel D). The MELSM not only raised this question, but it can also be used to answer this important theoretical question.

Conclusion

In the social-behavioral sciences, it is customary to view homogeneous variance as an assumption to satisfy. The central message of this work is to move beyond that perspective, and to view within-person variance ("noise") as an opportunity to gain a richer understanding of psychological processes. We introduced a framework to facilitate this transition. The proposed model is not only suited for interference tasks, as presented in this work, but can also be used more generally in repeated measures designs. By focusing on the within-person variance, that is response time (in)stability, this approach opened up possibilities for modeling a component that is often disregarded as "noise" or measurement "error." The illustrative examples highlighted such possibilities and demonstrated that the residual variance may show systematic patterns, individual differences, and relations with the mean structure.

References

Aarts, E., Dolan, C. V., Verhage, M., & van der Sluis, S. (2015). Multilevel analysis quantifies variation in the experimental effect while optimizing power and preventing false positives. *BMC Neurosci*, *16*(1), 94. doi: 10.1186/s12868-015-0228-5

Aarts, E., Verhage, M., Veenvliet, J. V., Dolan, C. V., & van der Sluis, S. (2014, 4). A solution to dependency: using multilevel analysis to accommodate nested data. *Nature Neuroscience*, *17*(4), 491–496. doi: 10.1038/nn.3648

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412. doi: 10.1016/j.jml.2007.12.005

Baird, B. M., Le, K., & Lucas, R. E. (2006). On the nature of intraindividual personality variability: Reliability, validity, and associations with well-being. *Journal of personality and social psychology*, *90*(3), 512.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. doi: 10.1016/j.jml.2012.11.001

Bartko, J. J. (1966, 8). The Intraclass Correlation Coefficient as a Measure of Reliability. *Psychological Reports*, *19*(1), 3–11. doi: 10.2466/pr0.1966.19.1.3

Bartko, J. J. (1976). On various intraclass correlation reliability coefficients. *Psychological Bulletin*, *83*(5), 762–765. doi: 10.1037/0033-2909.83.5.762

Bartlett, M. S. (1957, 12). A comment on D. V. Lindley's statistical paradox. *Biometrika*, *44*(3-4), 533–534. doi: 10.1093/biomet/44.3-4.533

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using {lme4}. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Bauer, D. J. (2011, 4). Evaluating Individual Differences in Psychological Processes:. *Current Directions in Psychological Science*, *20*(2). doi: 10.1177/0963721411402670

Bauwens, L., Laurent, S., & Rombouts, J. V. (2006). Multivariate GARCH models: A survey. *Journal of Applied Econometrics*, *21*(1), 79–109. doi: 10.1002/jae.842

Berger, J. (2006). The case for objective Bayesian analysis. *Bayesian Analysis*, *1*(3), 385–402. doi: 10.1214/06-BA115

Boisgontier, M. P., & Cheval, B. (2016). The anova to mixed model transition. *Neuroscience and Biobehavioral Reviews*, *68*, 1004–1005. doi: 10.1016/j.neubiorev.2016.05.034

Bürkner, P.-C. (2017). brms : An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*(1). doi: 10.18637/jss.v080.i01

Cattell, R. B., Cattell, A. K. S., & Rhymer, R. M. (1947). P-technique demonstrated in determining psychophysiological source traits in a normal individual. *Psychometrika*, *12*(4), 267–288.

Christ, B. U., Combrinck, M. I., & Thomas, K. G. F. (2018). Both Reaction Time and Accuracy Measures of Intraindividual Variability Predict Cognitive Performance in Alzheimer's Disease. *Frontiers in human neuroscience*, *12*, 124. doi: 10.3389/fnhum.2018.00124

Cudeck, R. (1996). Mixed-effects Models in the Study of Individual Differences with Repeated Measures Data. *Multivariate Behavioral Research*, *31*(3). doi: 10.1207/S15327906MBR3103{_}6

Davidson, D. J., Zacks, R. T., & Williams, C. C. (2003, 6). Stroop Interference, Practice, and Aging. *Aging, Neuropsychology, and Cognition*, *10*(2), 85–98. doi: 10.1076/anec.10.2.85.14463

De la Cruz, R., Meza, C., Arribas-Gil, A., & Carroll, R. J. (2016, 1). Bayesian regression analysis of data with random effects covariates from nonlinear longitudinal measurements. *Journal of multivariate analysis*, *143*, 94–106. doi: 10.1016/j.jmva.2015.08.020

Deng, H., & Wickham, H. (2011). Density Estimation In R. *useR! 2011*(September), 17.

doi: 10.1016/S0040-4020(98)00814-X

Dickey, J. M. (1971, 2). The Weighted Likelihood Ratio, Linear Hypotheses on Normal Location Parameters. *The Annals of Mathematical Statistics*, *42*(1), 204–223. doi: 10.1214/aoms/1177693507

Dienes, Z. (2014, 7). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, 781. doi: 10.3389/fpsyg.2014.00781

Efron, B., & Morris, C. (1977). Stein's paradox in statistics. *Scientific American*, *236*(5), 119–127.

Eid, M., & Diener, E. (1999). Intraindividual variability in affect: Reliability, validity, and personality correlates. *Journal of Personality and Social Psychology*, *76*(4), 662.

Engle, R. (2001, 11). GARCH 101: The Use of ARCH/GARCH Models in Applied Econometrics. *Journal of Economic Perspectives*, *15*(4), 157–168. doi: 10.1257/jep.15.4.157

Estabrook, R., Grimm, K. J., & Bowles, R. P. (2012, 9). A Monte Carlo simulation study of the reliability of intraindividual variability. *Psychology and Aging*, *27*(3), 560–576. doi: 10.1037/a0026669

Fagot, D., Mella, N., Borella, E., Ghisletta, P., Lecerf, T., & De Ribaupierre, A. (2018). Intra-Individual Variability from a Lifespan Perspective: A Comparison of Latency and Accuracy Measures. *Journal of Intelligence*, *6*(1), 16.

Fernández-Villaverde, J., Guerrón-Quintana, P., Rubio-Ramírez, J. F., & Uribe, M. (2011). Risk Matters: The Real Effects of Volatility Shocks. *American Economic Review*, *101*(6), 2530–61. doi: 10.1257/AER.101.6.2530

Fiske, D. W., & Rice, L. (1955). Intra-individual response variability. *Psychological Bulletin*, *52*(3), 217.

Gelman, A. (2006). Prior distribution for variance parameters in hierarchical models. *Bayesian Analysis*, *1*(3), 515–533. doi: 10.1214/06-BA117A

Gelman, A., Hill, J., & Yajima, M. (2012, 4). Why We (Usually) Don't Have to Worry

About Multiple Comparisons. *Journal of Research on Educational Effectiveness*, *5*(2), 189–211. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/19345747.2011.618213 doi: 10.1080/19345747.2011.618213

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*(6), 997–1016. doi: 10.1007/s11222-013-9416-2

Gelman, A., & Pardoe, I. (2012). Bayesian Measures of Explained Variance and Pooling in Multilevel (Hierarchical) Models. *Technometrics*, *48*(2). doi: 10.1198/004017005000000517

Gu, X., Hoijtink, H., & Mulder, J. (2016). Error probabilities in default Bayesian hypothesis testing. *Journal of Mathematical Psychology*, *72*(April 2018), 130–143. doi: 10.1016/j.jmp.2015.09.001

Haaf, J. M., & Rouder, J. N. (2017). Developing Constraint in Bayesian Mixed Models. *Psychological Methods*, *22*(4), 779–798. doi: 10.1037/met0000156

Hedeker, D., Mermelstein, R. J., & Demirtas, H. (2008). An application of a mixed-effects location scale model for analysis of ecological momentary assessment (EMA) data. *Biometrics*, *64*(2), 627–634. doi: 10.1111/j.1541-0420.2007.00924.x

Hoffman, L., & Stawski, R. S. (2009, 6). Persons as Contexts: Evaluating Between-Person and Within-Person Effects in Longitudinal Analysis. *Research in Human Development*, *6*(2-3), 97–120. doi: 10.1080/15427600902911189

Højsgaard, S., Edwards, D., & Lauritzen, S. (2012). *Graphical Models with R*. doi: 10.1007/978-1-4614-2299-0

Horn, J. L. (1972). State, trait and change dimensions of intelligence. *British Journal of Educational Psychology*, *42*(2), 159–185.

Jeffreys, H. (1961). *The theory of probability*. Oxford: Oxford University Press.

Johnson, V. E. (2005, 11). Bayes factors based on test statistics. *Journal of the Royal*

Statistical Society: Series B (Statistical Methodology), *67*(5), 689–701. doi: 10.1111/j.1467-9868.2005.00521.x

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54–69. doi: 10.1037/a0028347

Kalin, A. M., Plfuger, M., Gietl, A. F., Riese, F., JÃncke, L., Nitsch, R. M., & Hock, C. (2014, 7). Intraindividual variability across cognitive tasks as a potential marker for prodromal Alzheimers disease. *Frontiers in Aging Neuroscience*, *6*, 147. doi: 10.3389/fnagi.2014.00147

Kalmijn, W., & Veenhoven, R. (2005). Measuring inequality of happiness in nations: In search for proper statistics. *Journal of Happiness Studies*, *6*(4), 357–396.

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

King, J. A., Colla, M., Brass, M., Heuser, I., & Cramon, D. v. (2007, 8). Inefficient cognitive control in adult ADHD: evidence from trial-by-trial Stroop test and cued task switching performance. *Behavioral and brain functions : BBF*, *3*, 42. doi: 10.1186/1744-9081-3-42

Kliegl, R., Wei, P., Dambacher, M., Yan, M., & Zhou, X. (2010). Experimental Effects and Individual Differences in Linear Mixed Models: Estimating the Relationship between Spatial, Object, and Attraction Effects in Visual Attention. *Frontiers in psychology*, *1*, 238. doi: 10.3389/fpsyg.2010.00238

Klugkist, I., Kato, B., & Hoijtink, H. (2005a). Bayesian model selection using encompassing priors. *Statistica Neerlandica*, *59*(1), 57–69. doi: 10.1111/j.1467-9574.2005.00279.x

Klugkist, I., Kato, B., & Hoijtink, H. (2005b, 2). Bayesian model selection using encompassing priors. *Statistica Neerlandica*, *59*(1), 57–69. doi:

10.1111/j.1467-9574.2005.00279.x

Konishi, S. (1985). Normalizing and variance stabilizing transformations for intraclass correlations. *Annals of the Institute of Statistical Mathematics*, *37*(1), 87–94. doi: 10.1007/BF02481082

Krueger, C., & Tian, L. (2004, 10). A Comparison of the General Linear Mixed Model and Repeated Measures ANOVA Using a Dataset with Multiple Missing Data Points. *Biological Research For Nursing*, *6*(2), 151–157. doi: 10.1177/1099800404267682

Lazic, S. E., & Essioux, L. (2013). Improving basic and translational science by accounting for litter-to-litter variation in animal models. *BMC Neuroscience*, *14*(1), 37. doi: 10.1186/1471-2202-14-37

Leckie, G., French, R., Charlton, C., & Browne, W. (2014). Modeling heterogeneous variance–covariance components in two-level models. *Journal of Educational and Behavioral Statistics*, *39*(5), 307–332.

Ledoit, O., Santa-Clara, P., & Wolf, M. (2003, 8). Flexible Multivariate GARCH Modeling with an Application to International Stock Markets. *Review of Economics and Statistics*, *85*(3), 735–747. doi: 10.1162/003465303322369858

Lee, M. D. (2018, 3). Bayesian Methods in Cognitive Modeling. In *Stevens' handbook of experimental psychology and cognitive neuroscience* (pp. 1–48). Hoboken, NJ, USA: John Wiley & Sons, Inc. doi: 10.1002/9781119170174.epcn502

Leppink, J. (2019, 3). When Negative Turns Positive and Vice Versa: The Case of Repeated Measurements. *Health Professions Education*, *5*(1), 76–81. doi: 10.1016/J.HPE.2017.03.004

Leppink, J., & Merriënboer, J. J. G. V. (2015). The Beast of Aggregating Cognitive Load Measures in Technology-Based Learning. *Educational Technology & Society*, *18*(4), 230–245.

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

Analysis, *100*(9), 1989–2001. doi: 10.1016/j.jmva.2009.04.008

Lindley, D. V. (1957, 6). A STATISTICAL PARADOX. *Biometrika*, *44*(1-2), 187–192. doi: 10.1093/biomet/44.1-2.187

Liu, S., Rovine, M. J., & Molenaar, P. C. M. (2012). Selecting a linear mixed model for longitudinal data: Repeated measures analysis of variance, covariance pattern model, and growth curve approaches. *Psychological Methods*, *17*(1), 15–30. doi: 10.1037/a0026971

MacDonald, S. W. S., Hultsch, D. F., & Dixon, R. A. (2008). Predicting impending death: inconsistency in speed is a selective and early marker. *Psychology and aging*, *23*(3), 595.

Marsman, M., & Wagenmakers, E. J. (2017). Bayesian benefits with JASP. *European Journal of Developmental Psychology*, *14*(5), 545–555. doi: 10.1080/17405629.2016.1259614

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46. doi: 10.1037/1082-989X.1.1.30

McNeish, D. M., & Stapleton, L. M. (2016, 6). The Effect of Small Sample Size on Two-Level Model Estimates: A Review and Illustration. *Educational Psychology Review*, *28*(2), 295–314. doi: 10.1007/s10648-014-9287-x

Mestdagh, M., Pe, M., Pestman, W., Verdonck, S., Kuppens, P., & Tuerlinckx, F. (2018, 12). Sidelining the mean: The relative variability index as a generic mean-corrected variability measure for bounded variables. *Psychological Methods*, *23*(4), 690–707. doi: 10.1037/met0000153

Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2016, 6). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, *72*, 6–18. doi: 10.1016/J.JMP.2015.11.001

Mulder, J. (2016). Bayes factors for testing order-constrained hypotheses on correlations.

Journal of Mathematical Psychology, *72*, 104–115. doi: 10.1016/j.jmp.2014.09.004

Mulder, J., Hoijtink, H., & de Leeuw, C. (2012). BIEMS: A Fortran 90 program for calculating {B}ayes factors for inequality and equality constrained model. *Journal of Statistical Software*, *46*.

Mulder, J., & Raúl Pericchi, L. (2018). The Matrix-F Prior for Estimating and Testing Covariance Matrices. *Bayesian Analysis*(4), 1–22. doi: 10.1214/17-BA1092

Nuijten, M. B., Wetzels, R., Matzke, D., Dolan, C. V., & Wagenmakers, E.-J. (2015, 3). A default Bayesian hypothesis test for mediation. *Behavior Research Methods*, *47*(1), 85–97. doi: 10.3758/s13428-014-0470-2

Parris, B. A. (2014). Task conflict in the Stroop task: When Stroop interference decreases as Stroop facilitation increases in a low task conflict context. *Frontiers in Psychology*, *5*. doi: 10.3389/FPSYG.2014.01182

Pratte, M. S., Rouder, J. N., Morey, R. D., & Feng, C. (2010, 10). Exploring the differences in distributional properties between Stroop and Simon effects using delta plots. *Attention, Perception & Psychophysics*, *72*(7), 2013–2025. doi: 10.3758/APP.72.7.2013

Quintana, D. S., & Williams, D. R. (2018). Bayesian alternatives for common null-hypothesis significance tests in psychiatry: A non-technical guide using JASP. *BMC Psychiatry*, *18*(1). doi: 10.1186/s12888-018-1761-4

R Core Team. (2017). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/

Ram, N., & Gerstorf, D. (2009). Time-structured and net intraindividual variability: Tools for examining the development of dynamic characteristics and processes. *Psychology and aging*, *24*(4), 778.

Rast, P., & Ferrer, E. (2018). A Mixed-Effects Location Scale Model for Dyadic Interactions. , 1–63. doi: 10.1080/00273171.2018.1477577

Rast, P., Hofer, S. M., & Sparks, C. (2012). Modeling Individual Differences in Within-Person Variation of Negative and Positive Affect in a Mixed Effects Location Scale Model Using BUGS/JAGS. *Multivariate Behavioral Research*, *47*(2), 177–200. doi: 10.1080/00273171.2012.658328

Röcke, C., & Brose, A. (2013). Intraindividual Variability and Stability of Affect and Well-Being. *GeroPsych*, *26*(3), 185–199. doi: 10.1024/1662-9647/a000094

Rouder, J. N., & Jun, L. U. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin and Review*, *12*(4), 573–604. doi: 10.3758/BF03196750

Rouder, J. N., Kumar, A., & Haaf, J. M. (2019). Why most studies of individual differences with inhibition tasks are bound to fail. , 1–37.

Rouder, J. N., & Morey, R. D. (2012). Default Bayes Factors for Model Selection in Regression. *Multivariate Behavioral Research*, *47*(6), 877–903. doi: 10.1080/00273171.2012.734737

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin and Review*, *16*(2), 225–237. doi: 10.3758/PBR.16.2.225

Rouder, J. N., Tuerlinckx, F., Speckman, P., Lu, J., & Gomez, P. (2008). A hierarchical approach for fitting curves to response time measurements. *Psychonomic Bulletin and Review*, *15*(6), 1201–1208. doi: 10.3758/PBR.15.6.1201

Rousselet, G. A., & Wilcox, R. R. (2019). Reaction times and other skewed distributions: problems with the mean and the median. *bioRxiv*, 383935. Retrieved from https://www.biorxiv.org/content/early/2019/01/16/383935.abstract?%3Fcollection= doi: 10.1101/383935

Schramm, P., & Rouder, J. (2019). Are Reaction Time Transformations Really Beneficial? Retrieved from https://psyarxiv.com/9ksa6/ doi: 10.31234/OSF.IO/9KSA6

Stan Development Team. (2015). *$\$pkg{{}Stan{}}: A $\$proglang{{}C++{}} Library*

for Probability and Sampling, Version 2.8.0. Retrieved from http://mc-stan.org/

Stan Development Team. (2016). *Rstan: the R interface to Stan.* Retrieved from http://mc-stan.org/

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the {t}hird {b}erkeley {s}ymposium on {m}athematical {s}tatistics and {p}robability, 1954–1955, vol. {i}* (pp. 197–206). University of California Press, Berkeley and Los Angeles.

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. doi: 10.1007/s11222-016-9696-4

Vehtari, A., & Ojanen, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, *6*(0), 142–228. Retrieved from http://projecteuclid.org/euclid.ssu/1356628931 doi: 10.1214/12-SS102

von Bastian, C. C., Souza, A. S., & Gade, M. (2016, 2). No evidence for bilingual cognitive advantages: A test of four hypotheses. *Journal of experimental psychology. General*, *145*(2), 246–258. doi: 10.1037/xge0000120

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. doi: 10.3758/BF03194105

Wagenmakers, E. J., & Brown, S. (2007). On the Linear Relation Between the Mean and the Standard Deviation of a Response Time Distribution. *Psychological Review*, *114*(3), 830–841. doi: 10.1037/0033-295X.114.3.830

Wagenmakers, E. J., Grasman, R. P., & Molenaar, P. C. (2005). On the relation between the mean and the variance of a diffusion model response time distribution. *Journal of Mathematical Psychology*, *49*(3), 195–204. doi: 10.1016/j.jmp.2005.02.003

Wagenmakers, E. J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey method.

Cognitive Psychology, *60*(3), 158–189. doi: 10.1016/j.cogpsych.2009.12.001

Wagenmakers, E. J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., ... Morey, R. D. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin and Review*, *25*(1), 58–76. doi: 10.3758/s13423-017-1323-7

Wagenmakers, E. J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., ... Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin and Review*, *25*(1), 35–57. doi: 10.3758/s13423-017-1343-3

Wang, L. P., & Grimm, K. J. (2012, 9). Investigating Reliabilities of Intraindividual Variability Indicators. *Multivariate Behavioral Research*, *47*(5), 771–802. doi: 10.1080/00273171.2012.715842

Wang, L. P., Hamaker, E., & Bergeman, C. S. (2012, 12). Investigating inter-individual differences in short-term intra-individual variability. *Psychological Methods*, *17*(4), 567–581. doi: 10.1037/a0029317

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. J. *ournal of Machine Learning Research*, *11*, 3571–3594.

Westfall, J., Kenny, D. A., & Judd, C. M. (2014, 10). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of experimental psychology. General*, *143*(5), 2020–45. doi: 10.1037/xge0000014

Wetzels, R., Grasman, R. P., & Wagenmakers, E. J. (2010). An encompassing prior generalization of the SavageDickey density ratio. *Computational Statistics and Data Analysis*, *54*(9), 2094–2102. doi: 10.1016/j.csda.2010.03.016

Williams, D. R., Carlsson, R., & Bürkner, P.-C. (2017). Between-litter variation in developmental studies of hormones and behavior: Inflated false positives and

diminished power. *Frontiers in Neuroendocrinology*(August), 0–1. doi: 10.1016/j.yfrne.2017.08.003

Williams, D. R., Liu, S., Martin, S. R., & Rast, P. (2019). Bayesian Multivariate Mixed-Effects Location Scale Modeling of Longitudinal Relations among Affective Traits, States, and Physical Activity.
doi: 10.31234/OSF.IO/4KFJP

Williams, D. R., & Mulder, J. (2019). Bayesian Hypothesis Testing for Gaussian Graphical Models: Conditional Independence and Order Constraints.

Williams, D. R., & Rast, P. (2018). A Bayesian Nonlinear Mixed-Effects Location Scale Model for Learning. , 1–18.

Williams, D. R., Rast, P., Pericchi, L. R., & Mulder, J. (2019). Comparing Gaussian Graphical Models with the Posterior Predictive Distribution and Bayesian Model Selection.
doi: https://doi.org/10.31234/osf.io/yt386

Wolsiefer, K., Westfall, J., & Judd, C. M. (2017, 8). Modeling stimulus variation in three common implicit attitude tasks. *Behavior Research Methods*, *49*(4), 1193–1209. doi: 10.3758/s13428-016-0779-0

Woodrow, H. (1932). Quotidian variability. *Psychological Review*, *39*(3), 245.

Wright, B. C. (2017, 8). What Stroop tasks can tell us about selective attention from childhood to adulthood. *British journal of psychology (London, England : 1953)*, *108*(3), 583–607. doi: 10.1111/bjop.12230

Table 1

Parameter estimates

	Mean Structure					
	Stroop			Flanker		
Parameter	M	SD	90 % CrI	M	SD	90 % CrI
Congruent (β_0)	0.71	0.01	[0.69,0.72]	0.56	0.03	[-1.84,-1.73]
Incongruent Δ (β_1)	0.07	0.00	[0.06,0.07]	0.00	0.00	[-0.00,0.01]
$SD(u_{0i})$	0.10	0.01	[0.09,0.11]	0.07	0.01	[0.06,0.08]
$SD(u_{1i})$	0.03	0.00	[0.02,0.03]	0.02	0.00	[0.01,0.02]
	Variance Structure					
	Stroop			Flanker		
Parameter	M	SD	90 % CrI	M	SD	90 % CrI
Congruent (η_0)	-1.78	0.03	[-1.84,-1.73]	-2.04	0.03	[-2.10,-1.99]
Incongruent Δ (η_1)	0.16	0.03	[0.11,0.20]	0.02	0.03	[-0.03,0.07]
$SD(u_{2i})$	0.34	0.02	[0.30,0.38]	0.37	0.03	[0.33,0.41]
$SD(u_{3i})$	0.24	0.02	[0.21,0.27]	0.28	0.02	[0.24,0.32]

Note. Posterior mean (M) and standard deviation (*SD*) in seconds. β_0 is the mean for the congruent condition. β_1 is the dummy-coded difference from the *in*congruent to the congruent condition. u_{0i} captures the *SD* for the congruent condition while u_{1i} is the *SD* of the difference from the incongruent to the congruent condition. η_0 is the *SD* on the log-scale for the congruent condition and η_1 is the difference in *SD* due to the incongruent condition. u_{2i} and u_{3i} capture the *SD*'s of η_0 and η_1, respectively.

Table 2

Exhaustive hypothesis testing results

	Stroop				
Parameter	M	SD	$p(\mathcal{H}_0\|\mathbf{Y})$	$p(\mathcal{H}_1\|\mathbf{Y})$	$p(\mathcal{H}_2\|\mathbf{Y})$
$cor(u_{0i}, u_{2i})$	0.68	0.06	0.00	1.00	0.00
$cor(u_{0i}, u_{3i})$	-0.40	0.09	0.00	0.00	1.00
$cor(u_{1i}, u_{2i})$	-0.19	0.14	0.42	0.05	0.53
$cor(u_{1i}, u_{3i})$	0.80	0.08	0.00	1.00	0.00
	Flanker				
Parameter	M	SD	$p(\mathcal{H}_0\|\mathbf{Y})$	$p(\mathcal{H}_1\|\mathbf{Y})$	$p(\mathcal{H}_2\|\mathbf{Y})$
$cor(u_{0i}, u_{2i})$	0.81	0.04	0.00	1.00	0.00
$cor(u_{0i}, u_{3i})$	-0.11	0.10	0.70	0.04	0.26
$cor(u_{1i}, u_{2i})$	-0.21	0.16	0.41	0.06	0.53
$cor(u_{1i}, u_{3i})$	0.87	0.07	0.00	1.00	0.00

Note. Posterior mean (M) and standard deviation (SD).

$\mathcal{H}_0 : \rho_{ij} = 0$ vs. $\mathcal{H}_1 : \rho_{ij} > 0$ vs. $\mathcal{H}_2 : \rho_{ij} < 0$.

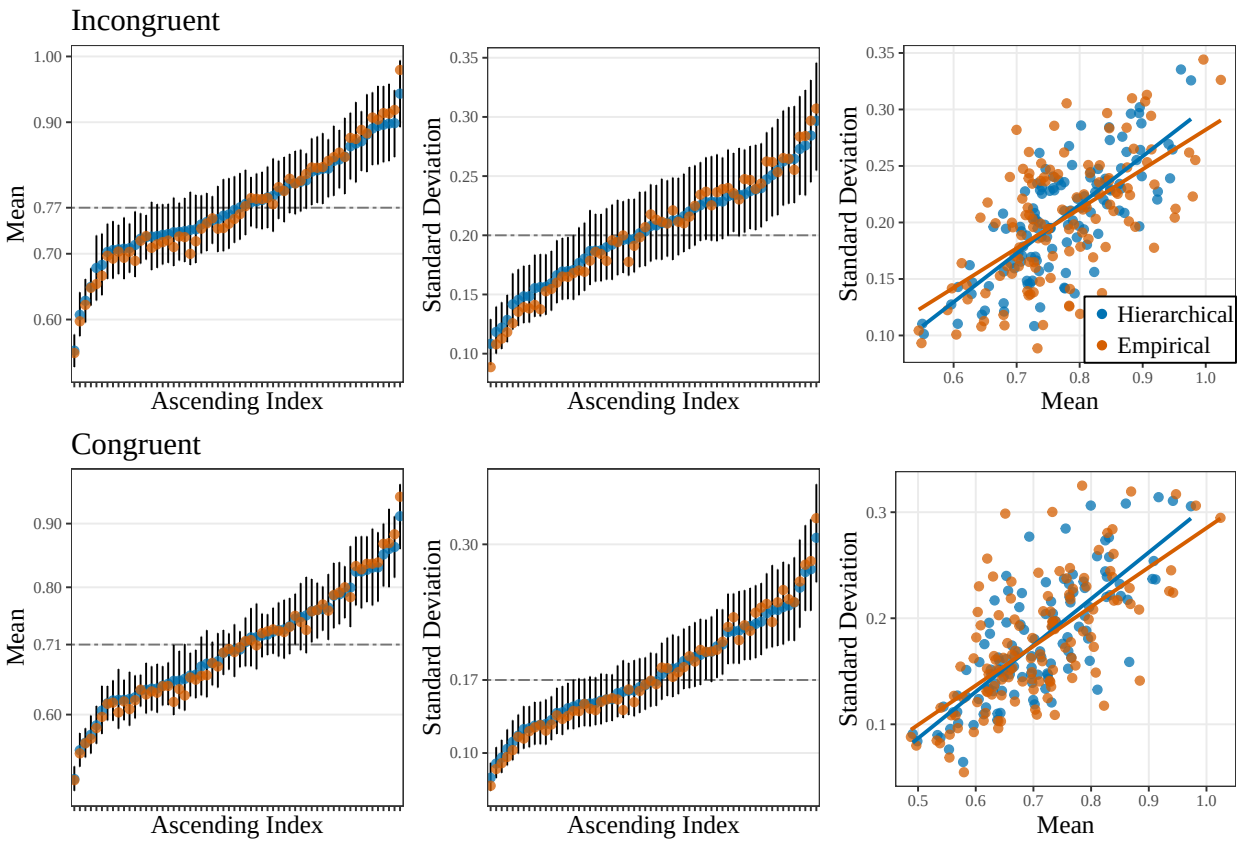

Figure 1. Results from the random intercepts only model. Separate models were fit to each outcome. The dotted line denotes the respective average across individuals (the fixed effect: β_0 and η_0). The "Empirical" estimates were obtained by computing the means and standard deviations for each person. Bars represent the 90% CrI for the hierarchical estimates.

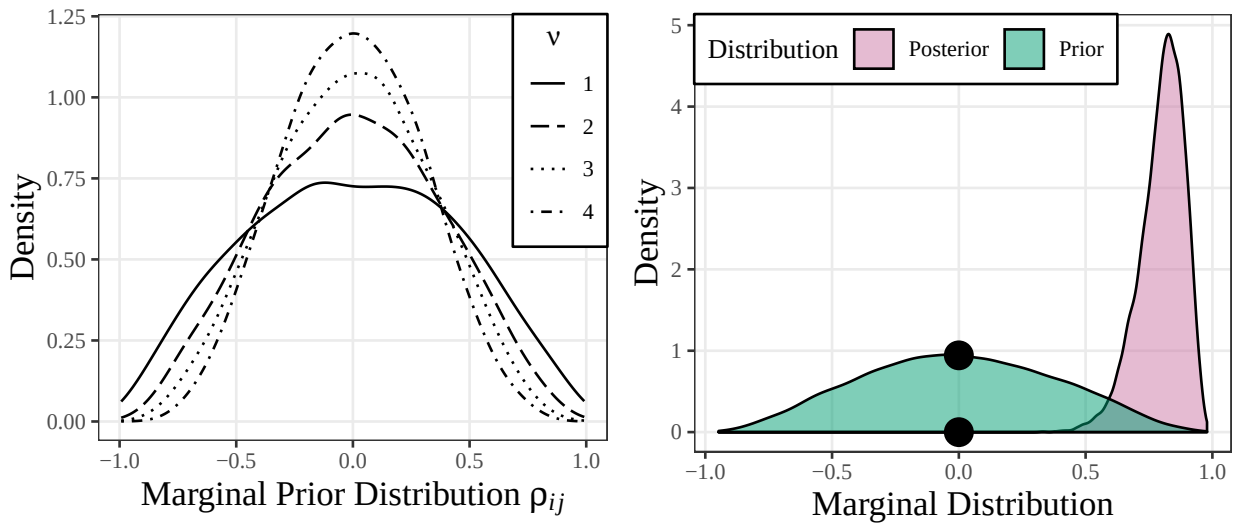

Figure 2. Marginal prior densities for the correlations. ν is the single parameter that governs the LKJ prior distribution. The samples were drawn from a 4×4 correlation matrix. This was the same dimension as the estimated matrices in the illustrative models.

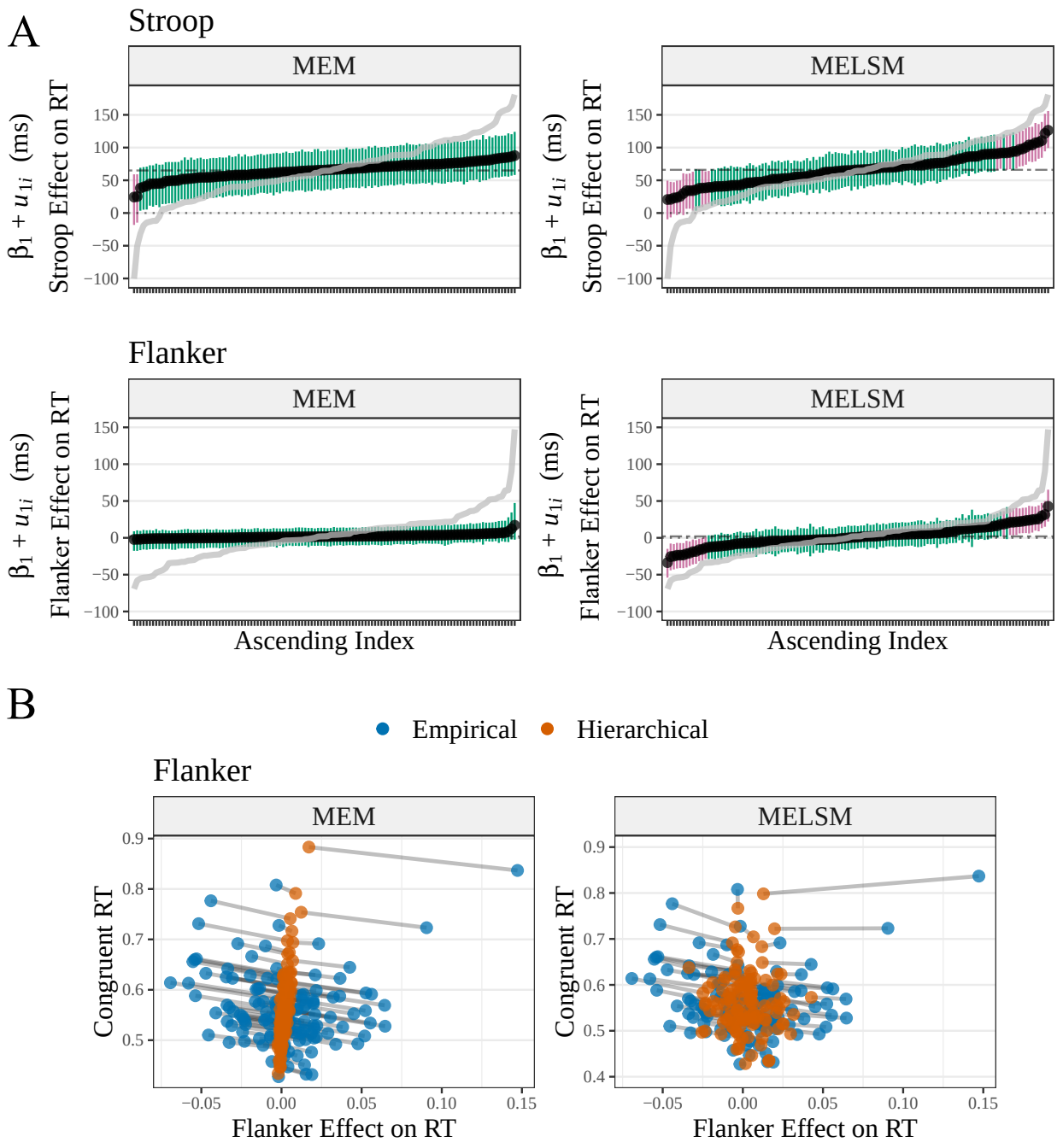

Figure 3. Comparison of posterior estimates from a standard mixed effects model (MEM) a mixed effects location scale model (MELSM). Panel A illustrates the greater shrinkage in the point estimates and the greater shrinkage due to the assumption of homoskedasticity in the MEM models compared to the MELSM for the same tasks. The gray line shows the empirical estimate for each individual. Red CrI's exclude the fixed effect estimate for the Strop and Flanker tasks, represented by the dotted line. Panel B explicitly highlights the difference in shrinkage among MEM and MELSM in the Flanker experiment.

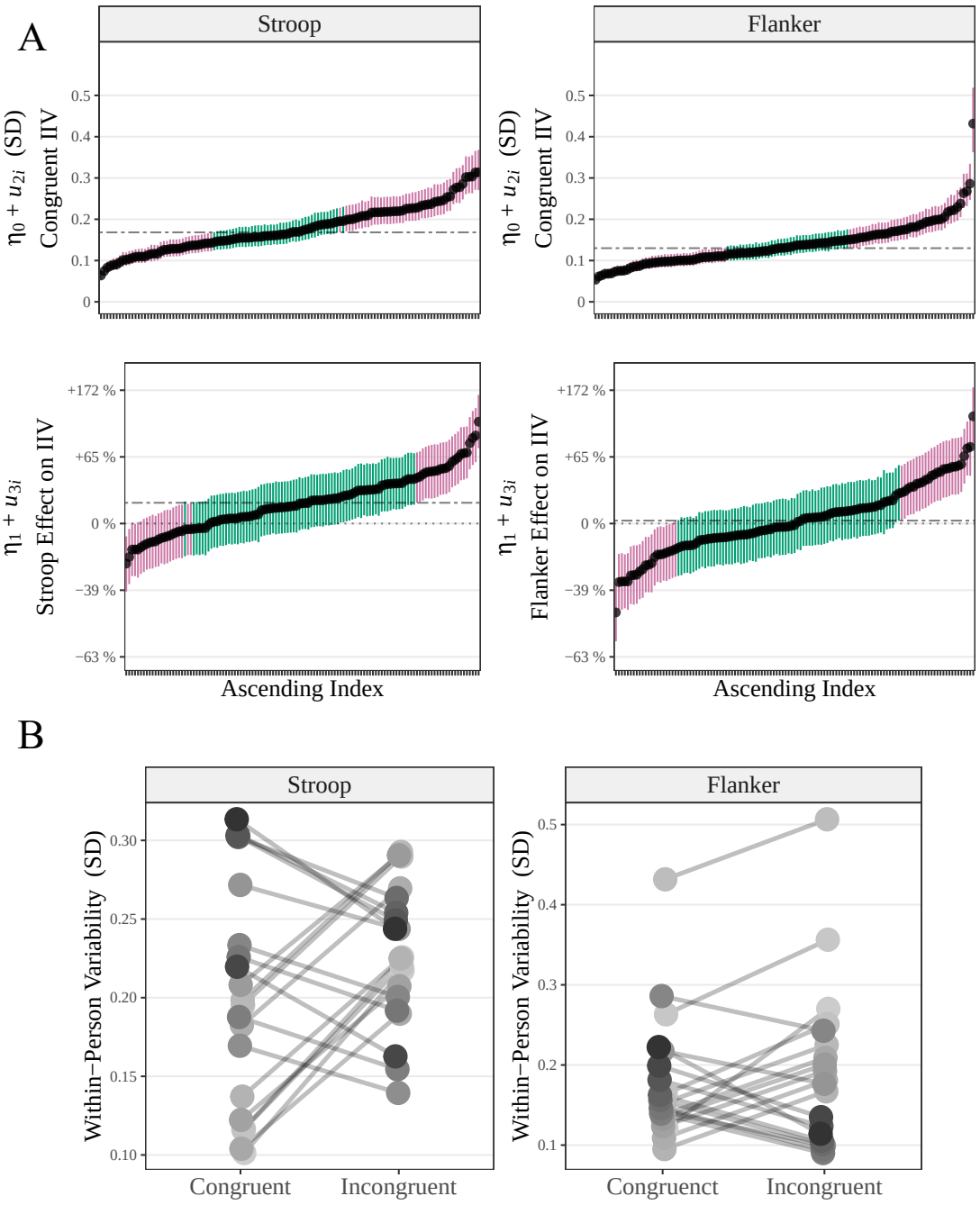

Figure 4. Effect of the Stroop and Flanker experimental conditions on intraindividual variability (IIV). The top row of Panel A captures the baseline SD for the congruent conditions. The second row shows % changes in SD due to the Stroop of Flanker condition. While, on average, change is positive or practically zero, there are considerable individual differences with some individuals showing more consistency and other showing less consistency in their reaction times. Red CrI's exclude the fixed effect estimate (dotted line). Panel B shows the posterior *SD*'s for 20 selected individuals in both the congruent and incongruent condition across the Stroop and Flanker task. This panel serves to illustrate that changes in the magnitude and direction in IIV were subject to substantial individual differences.

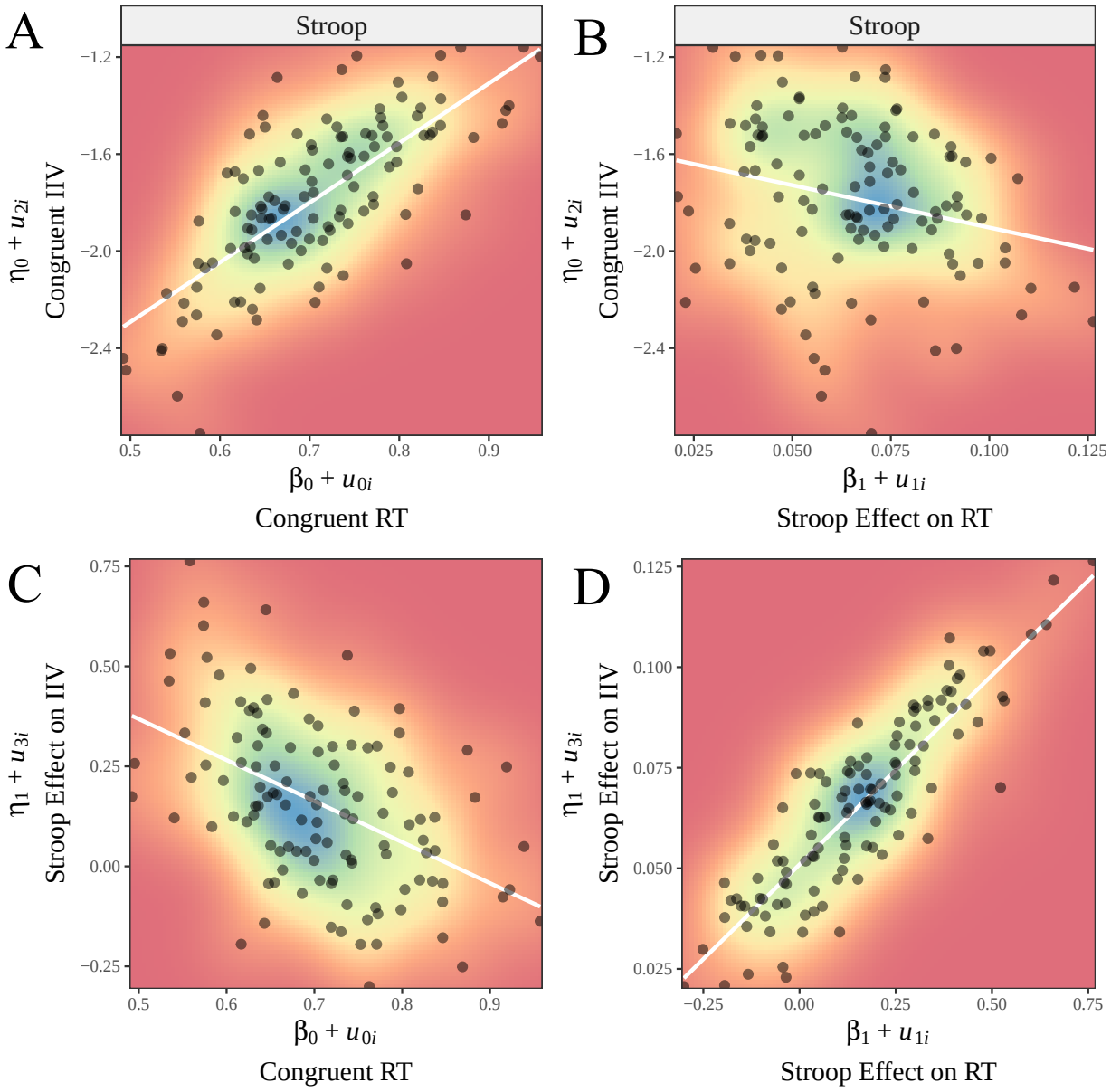

Figure 5. Predicted location and scale estimates for each individual (black dots) across different congruency conditions. The background color indicates multivariate density ranging from red (low) to blue (high). The top left panel shows a positive relation among the reaction times (RT) for the location and the SD of the RT's for the congruent conditions. Similarly, the bottom right panel shows the relation among the location and scale effects for the incongruent, or the Stroop effect, condition. In these two panels slower RT's were associated to larger within-person variance. The top right and bottom left panels show a negative relation among the location and the scale. Specifically, the bottom left panel depicts the relation among the congruent location RT's and the predicted RT variability on the scale, for the same participants. Participants with slower RT's in the congruent condition tend to be *less* variable in their RT's in the Stroop condition. The same pattern can be seen in the top right panel.

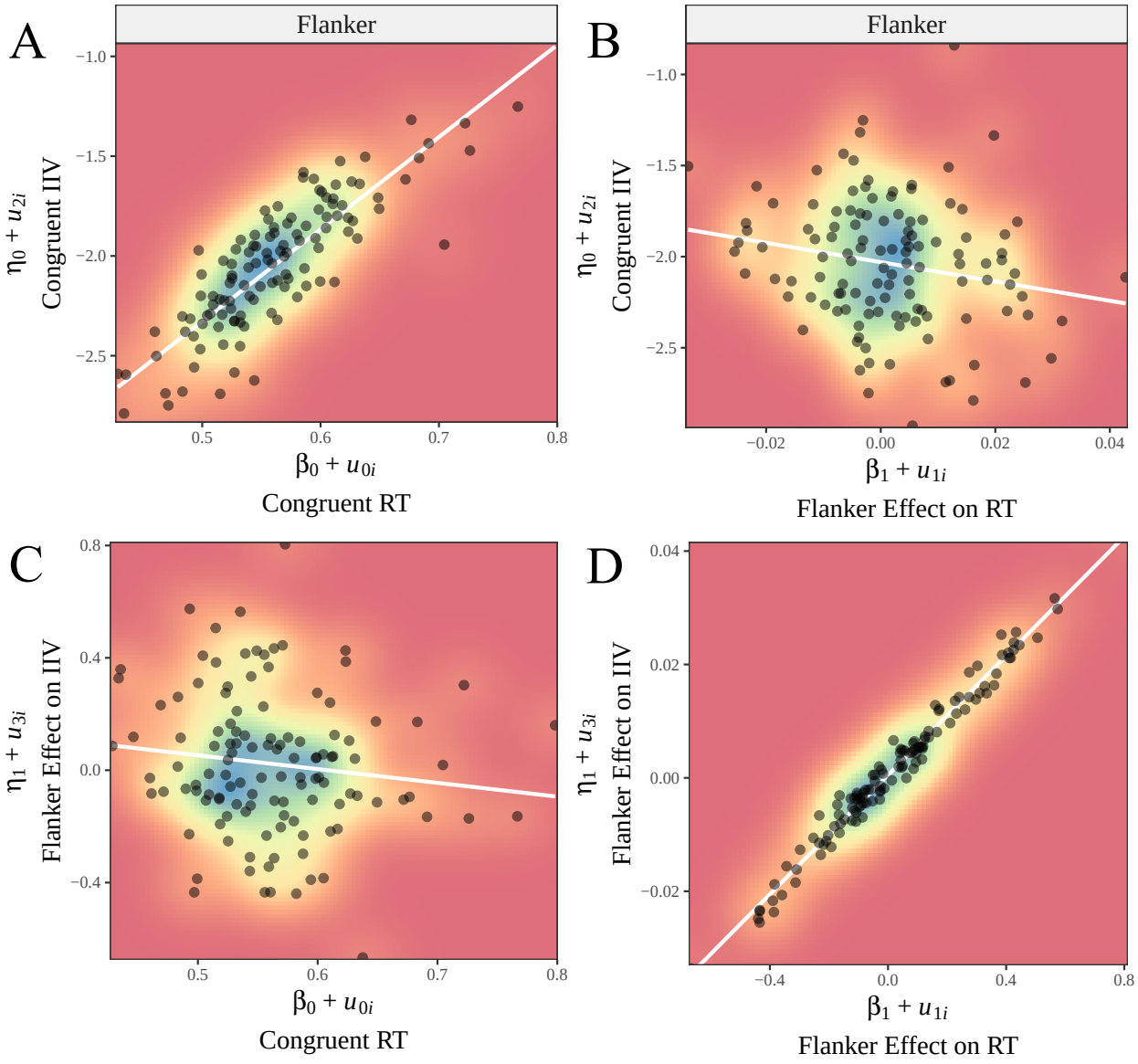

Figure 6. Predicted location and scale estimates for each individual overlayed on the color coded multivariate density ranging from red (low) to blue (high). The panels for the flanker test largely reproduce the findings from the Stroop test. The top left and lower right panel shows a positive relation among the reaction times location and scale, while the top right and lower left panel show no, or a negatvie relation among the locatino and and scale. It is noteworthy that for the flanker effect in both location and scale, in the lower right panel, approximately 24 % of the participant's RT's decreased (β_1 effect below zero).

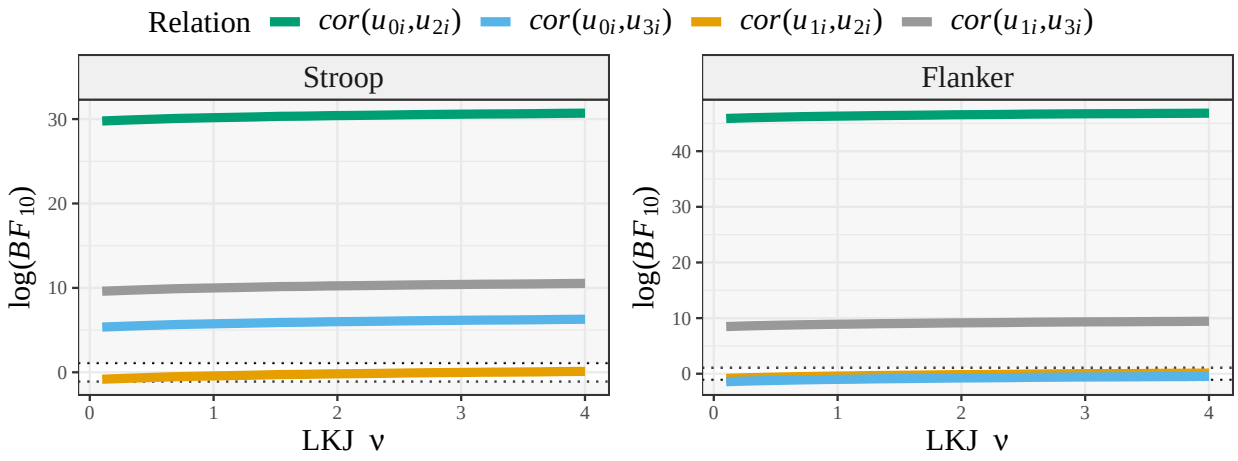

Figure 7. Sensitivity analyses. ν is the single parameter that governs the LKJ prior distribution. Larger values increasingly concentrate around zero.