Bayesian Analysis for Systems Factorial Technology

Jonathan E. Thiele, Julia M. Haaf, and Jeffrey N. Rouder

University of Missouri

Author Note

Jeff Rouder, Department of Psychological Sciences, 210 McAlester Hall, University of Missouri, Columbia, MO 65211, rouderj@missouri.edu. This research was supported by National Science Foundation grants BCS-1240359 and SES-102408.

The data in this paper was collected under a *Born Open Data* protocol (Rouder, 2016) in which they were automatically logged, uploaded, and made freely available as they were created (https://github.com/PerceptionCognitionLab/data1/tree/master/sysfactorial). This paper was prepared in LaTeX with R code for data analysis knitted into the document. The LaTeX and R source are freely available at https://github.com/PerceptionAndCognitionLab/sysfac.

Abstract

Systems factorial technology (Townsend & Nozawa, 1995) is a leading methodology for assessing the processing of multiple-feature items. By using certain experimental designs and analyses, researchers can assess wether features are processed in serial, in parallel, or coactively. Current practice is to categorize each individual as displaying one of these three architectures. We argue this approach implicitly assumes heterogeneity of processing strategies across participants. A more scientifically meaningful approach may be to first ask whether all people are serial or parallel or coactive before assuming heterogeneity. We develop a series of Bayesian hierarchical models that captures both situations where everyone follows a common architecture and, alternatively, where there is heterogeneity in architecture. These models use g-prior structures that make computation of Bayes factors convenient. We report an application to investigate Miller's (1956) notion of chunking. We asked participants to compare object that are composed of separable features simultaneously, a perception task, and sequentially, a memory task. We assessed whether processing changed across the perception and memory tasks with the notion that participants might have to chunk features to store them, and that this chunking might make processing more efficient. The answer is "no." We find a serial architecture for processing for highly separable features (size of circle and the orientation of its diameter) in both the perception and memory tasks. We also find parallel processing for less separable features (first and second digit in a two-digit number) in both perception and memory tasks. Taken together, while processing may depend on the separability of features, it does not vary across perception and memory. As importantly, we find that all people had the same processing strategy; that is models that stated no heterogeneity outperformed those with heterogeneity. This result indicates that architecture may be universal in this setting and not under strategic control.

Bayesian Analysis for Systems Factorial Technology

The goal of this paper is to describe a new approach to evaluate evidence for equality and order constraints in psychological data. We use this new approach for inference in systems factorial technology, a method used to determine the mental architecture underlying processing in experimental tasks. In turn, we use a variant of systems factorial technology to address the question of whether the architecture of perception is the same as that of working memory. This paper, therefore, sits at the interface of three stories: a statistical story about how evidence should be evaluated, a methodological story about how architecture is determined, and a substantive story about perception and working memory. We take each in turn.

We think the most important contribution here is the statistical story. Here is the background: Often, researchers are concerned with average effects. For example, if a researcher thinks an interaction between two variables is theoretically important, they may compute the appropriate t-test value, which is a measure of the significance of the average interaction contrast. An improvement on this approach comes from the psychometrics tradition where individuals provide so much data that they are effectively experiments unto themselves. For example, if we are interested in the sign of the interaction of two variables, as we will be with systems factorial technology, we may manipulate both variables in a within-subject design where each individual observes many trials in each cell. We then can compute a t-test value for each individual and classify each as significantly negative, nonsignificant, or significantly positive. Examples of classifying people this way include Little, Nosofsky, and Denton (2011). The same basic logic has been enhanced by using explicit Bayesian mixture models where individuals are classified into psychologically distinct modes of processing (Houpt & Fific, 2017; Kary, Taylor, & Donkin, 2016; Rouder, Morey, Speckman, & Pratte, 2007). This Bayesian approach, though more intellectually defensible, shares the basic property of being an approach for classifying individuals.

We think classifying individuals is not necessarily the best way to proceed. Let's start

with a focus on the sign of an interaction term. We assume that this sign, whether positive, zero, or negative has theoretical importance. In this paper, where we use systems factorial technology, the sign of the interaction will be an indicator of the architecture. The details are provided subsequently, but parallel, serial, and coactive processing implies true interaction contrast terms that are negative, zero, and positive, respectively (Fific, Nosofsky, & Townsend, 2008). A search for lawfulness here takes the form of asking whether there is a common architecture in a task for all individuals. If all individuals approach a task with the same architecture, then we might view this architecture as more of a primitive—perhaps that it is biological or automatic, and not under volitional control. If not, that is if there is true variation in architecture across people, then perhaps the choice of architecture is under strategic control. Such a result leads to follow-up questions about why certain people with certain characteristics chose certain architectures.

The above emphasis on lawfulness leads to questions like, “what is the strength of evidence from the data for the proposition that all true values are positive (or zero or negative)?" These questions cannot be answered by classifying individuals. Instead, they are questions about global patterns, particularly about the possibility of multiple order and equality constraints holding simultaneously. They are most deftly answered by comparing models that impose varying constraints. Traditionally, comparing the fit of models with multiple order-constraints has proven difficult because calculations of the sampling distributions of relevant test statistics is not theoretically or computationally convenient (Robertson, Wright, & Dykstra, 1988). Heuristic approaches such as AIC and BIC are also difficult as the penalty terms depend on the number of parameters but not restrictions of the space (Klugkist & Hoijtink, 2007). To address these difficulties, we develop a Bayes factor approach to assess the strength of evidence for models with multiple simultaneous order restrictions. This development is broadly applicable and provides the answer to the question, "*Does everyone?*", For example, it may be used for questions like, "does everyone identify bright flashes faster than dim ones," or "does everyone show a Simon interference

effect?"

The second story, about methodology, goes as follows: One of the key questions across cognitive psychology is the nature of latent processing that underlies various information-processing tasks. Consider the perception of objects that can be described by their features. How these features are combined into coherent wholes remains timely and topical. This question has generated a long and fruitful mathematical-psychology literature on formal methods for understanding and querying processing architecture. A selective list includes Garner and Felfoldy (1970), Liu (1996), Schweikert and Townsend (1989), Sternberg (1969), Townsend (1990), and Townsend and Ashby (1982).

To make the situation concrete, consider the stimuli presented in Figure 1. We call these stimuli *screwheads* because they resemble the top view of a flathead screw. The stimuli are defined by two features: the size of the screwhead and the orientation of the slot. The question is how these two features are processed. Perhaps the most common approach is to consider three different architectures: 1. *Serial processing*, where features are processed one-at-a-time in sequence; 2. *Parallel processing*, where features are processed independently and simultaneously and with unlimited capacity; and 3. *Coactive processing*, where the processing of one feature facilitates the processing of the others.

The approach we take to assess architecture is Townsend and Nozawa's (1995) *Systems Factorial Technology*. Systems factorial technology refers to a collection of approaches developed by Townsend and his students (see Townsend & Wenger, 2004, for a review). The specific one used here is the logical-rules variant (Fific et al., 2008). Using this approach, Fific, Little, and colleagues have found the following two results: First, simple objects with separable features, such as the screwheads, are seemingly mediated by serial processing for most people (Fific, Little, & Nosofsky, 2010; Little et al., 2011). Second, objects with integral features such color patches comprised of hue and saturation are seemingly mediated by coactive processing (Little, Nosofsky, Donkin, & Denton, 2013).

The third story we consider is the substantive one. We ask whether the architecture

mediating perception is the same as that mediating working memory. We use simple objects with separable features for both perception and working memory tasks. Following Fific et al. (2010) and Little et al. (2011), we expect serial processing for these types of stimuli. The main question is about the effect of holding these objects in working memory. On one hand, one can think that storing, maintaining and recalling stimuli from working memory does not change processing much, and there is a tradition of thinking of memory as reexperiencing the object, albeit as a noisier and perhaps systematically distorted copy (Estes, 1997; Hebb & Foord, 1945). A modern version of this view is that memory is the reactivation of brain states during the original perception event (Danker & Anderson, 2010), and it has become popular in fMRI research of memory with pattern-recognition techniques (Rissman & Wagner, 2012). Alternatively, working memory is often thought as the process of binding disparate items and features together into one coherent chunk that can be store (Miller, 1956). There are several influential accounts of the role of working memory in chunking features together (Atkinson & Shiffrin, 1968; Cowan, 1995; Mandler, 1980). We ask whether this chunking changes the architecture of processing. It may be that before chunking, items are processed serially, but afterwards, they are processed in parallel or even coactively.

In the next section we review systems factorial technology. Included in this review is a discussion of the methodological limitations that motivate our Bayesian development. Thereafter, we develop a set of hierarchical models and derive Bayes factor computations for comparison among these models. Then, we present two experiments to assess whether recalling information from working memory changes the architecture of processing. Each experiment consists of a perception condition and a memory condition, and the critical question is whether processing changes across these conditions. The answer, perhaps surprisingly, is negative. We observed the same architecture across perception and memory tasks.

System Factorial Technology

Systems factorial technology refers to a broad collection of paradigms and analyses used to uncover the underlying architecture of processing (Townsend & Wenger, 2004). These approaches vary in detail in that some are based on mean responses while others are based on distributional characteristics. Approaches based on means have the potential of determining whether processing is serial, parallel, or coactive conditional on technical assumptions discussed subsequently. Distributional approaches provide for greater diagnosticity—they have the potential to assess the role of channel capacity and interdependence in processing multiple features and whether serial processes are self-terminating or exhaustive.

We use one specific variant of systems factorial technology, the logical rules approach based on mean response times in a conjunction task (Fific et al., 2010). This approach allows for the assessment of whether features are processed in serial, parallel, or coactively. In what follows we describe the approach in detail, first starting with the stimuli, task, and finishing with the statistical analysis. This statistical analysis is based on classification, and we expand upon the critique of classification we presented earlier.

Stimuli. The first step in applying systems factorial technology is operationalizing the features of a stimulus. Consider the screwheads in Figure 1. These stimuli have been used in several studies in categorization (e.g., Maddox & Ashby, 1993; McKinley & Nosofsky, 1995) because the features, the size of the screw and the orientation of the slot, may be manipulated factorially.

Task. We asked participants to compare two screwheads (Figure 1A) and respond positive if the stimuli were different on both dimensions and negative otherwise. There are three levels of difference for each feature. Consider orientation: The two stimuli could have the same orientation, a small orientation difference, or a large orientation difference, and we denote these three levels as 0, 1, and 2, respectively. The same holds for size: The two stimuli could have the same size, a small size difference, or a large size difference, again

denoted by 0, 1, and 2, respectively. Crossing these levels yield nine possibilities, and each possibility may be denoted by an ordered pair. For example, the ordered pair $(0, 2)$ denotes no change in orientation and a large change in size across the pair of screwheads. The task maps $(1, 1)$, $(1, 2)$, $(2, 1)$, and $(2, 2)$ into the positive response and the remaining 5 combinations into the negative response. This mapping is shown in Table 1.

Analysis. The relevant data in the systems factorial method are the affirmative responses. Let Y_{11}, Y_{12}, Y_{21}, Y_{22} be the response-time distributions for the conditions $(1, 1)$, $(1, 2)$, $(2, 1)$, and $(2, 2)$, respectively, and let $\mathrm{E}(Y_{11})$, $\mathrm{E}(Y_{12})$, $\mathrm{E}(Y_{21})$, $\mathrm{E}(Y_{22})$ be the respective expectation value of these distributions. Then the true mean interaction contrast (MIC), denoted M, is[1]

$$M = \frac{[\mathrm{E}(Y_{11}) + \mathrm{E}(Y_{22})] - [\mathrm{E}(Y_{12}) + E(Y_{21})]}{4}.$$

The observed MIC is

$$\hat{M} = \frac{(\bar{Y}_{11} + \bar{Y}_{22}) - (\bar{Y}_{12} + \bar{Y}_{21})}{4},$$

where $\bar{Y}_{11}$, $\bar{Y}_{12}$, $\bar{Y}_{21}$, and $\bar{Y}_{22}$ denote the observed cell means for conditions $(1, 1)$, $(1, 2)$, $(2, 1)$, and $(2, 2)$, respectively. The structure of these observed cell means and of the contrast is also shown in Table 1. The observed MIC is the best, unbiased estimator of the true MIC when the numbers of observations per cell are equal.

Perhaps Sternberg (1969) first popularized this interaction as a means of assessing architecture. Schweikert (1978) and Schweikert and Townsend (1989) provide more formal developments. Townsend and Nozawa (1995) showed that the sign of M is diagnostic of the nature of processing under certain technical conditions. The key results we leverage here are: **1.** If processing is parallel, the M is negative. **2.** If processing is coactive, then M is positive. **3.** If processing is serial, then $M = 0$. The technical conditions are that RT

[1]Usually, the true interaction contrast is defined as $M = \mathrm{E}(Y_{11}) + \mathrm{E}(Y_{22})] - [\mathrm{E}(Y_{12}) + E(Y_{21})]$. We prefer the scaled version, $M = ([\mathrm{E}(Y_{11}) + \mathrm{E}(Y_{22})] - [\mathrm{E}(Y_{12}) + E(Y_{21})])/4$ because then M is the interaction parameter in the usual linear model (see Eq. 2).

distribution order with strength: $Y_{22} \leq Y_{21}$, $Y_{22} \leq Y_{12}$, $Y_{21} \leq Y_{11}$, and $Y_{12} \leq Y_{11}$.[2] In our experience RT distribution order with strength variables over ecologically valid ranges (see also Luce, 1986; Rouder, Yue, Speckman, Pratte, & Province, 2010; Wagenmakers & Brown, 2007). We know of no instances where this ordering has been violated with strength variables such as those used here, and we accept these technical conditions as assumptions without further confirmation.

The Methodological Critique

Perhaps the least sophisticated approach is to assess the global interaction across all people. Figure 2A provides the overall cell means from our first experiment, to be presented subsequently. As can be seen, there is not even a whiff of an interaction. Yet, global contrasts tell us little about individuals and might be misleading. Practitioners of system factorial technology have instead estimated the MIC separately for each individual. Figure 2B provides an example from our to-be-presented experiment. Here we have plotted the contrast with 80% confidence intervals for each individual's MIC. As can be seen, 22 of the 32 of the confidence intervals contain 0, 5 of the 32 are localized above zero, and 5 of the 32 are localized below zero. One interpretation is that 22, 5, and 5 of the participants provide support for serial, parallel, and coactive architectures, respectively.

One of the weaknesses in the confidence interval approach is a difficult asymmetry where the serial signature is a point hypothesis while the parallel and coactive signatures are hypotheses across respective halves of the real line. The usual significance test approach allows rejection of the point but not acceptance of it. The usual approach of holding the nominal Type I error rate here is not appropriate here because it privileges serial processing. Even more problematic, this privilege varies with sample size, and it is almost complete with small sample sizes where Type II errors are common. To address this issue, we chose 80% threshold on the confidence interval. Yet, such a choice plays an

[2]The inequality of random variables refers to a stochastic ordering. The statement $Y_{22} \leq Y_{21}$ is equivalent to the statement that $Pr(Y_{22} < t) \geq Pr(Y_{21} < t)$ for all t.

outsized role in classifying individuals' architectures.

There are recent approaches using Bayesian mixture models to classify people. Houpt and Fific (2017) approaches systems factorial assessment this way. Each individual's MIC is treated as a parameter, and the prior has some mass below zero, some point mass at zero, and some mass above zero. By placing prior point mass at zero, and using a rational approach to updating the probability that the MIC is below, at, and above zero, these methods obviate the above concern. In this regard, they are needed advances.

But there is a deeper problem with confidence interval approach and Houpt and Fific's improvement. They are based on making categorizations of individuals, and, as we argued earlier, such an approach is ill-suited for assessing lawful regularities. In particular, they are poorly calibrated for assessing whether all people use a common processing architecture. We think these "does everybody" statements are foundational, theoretically important, and worthy of consideration. Our claim here is not so much that individual-categorization approaches preclude these types of lawful regularities, but that they do not provide a principled, calibrated approach to measuring the strength of evidence for them. One issue is prior probability of lawfulness. In the individual-categorization approach it may be quite low. For example, the prior probability of all individuals exhibiting serial architecture might be $(1/3)^I$, where 1/3 is the prior probability that a given individual exhibits serial processing and I is the number of individuals. Having such small prior probabilities for lawfulness is an undesirable bias against simplicity. Prior probabilities of finding at least some variation in processing across people may be quite high. Hence, the act of categorizing individuals seems biased toward finding variation of processing architectures across individuals at the expense of highlighting lawful regularity.

The solution to the problem is conceptually straightforward. To assess whether these lawful regularities hold, that is, whether all individuals exhibit serial processing, or all people exhibit parallel processing, or all people exhibit coactive processing, we compare models with and without these regularity constraints. In the simplest incarnation, we may

propose four models: the 'everyone-is-serial model where each individual's true MIC is constrained to be zero; the everyone-is-parallel model where each individual's true MIC is constrained to be negative; the everyone-is-coactive model where each individual's true MIC is constrained to be positive; and the unconstrained model where no such restrictions are placed on true MIC. Then, we can explicitly assess the strength of evidence for universal serial, parallel, or coactive processing or for variability in processing by assessing the relative fit of all four models.

As discussed above, stating strength of evidence for models with multiple simultaneous order and inequality constraints is relatively novel. The frequentist literature on the problem is large because the problem is considered very hard (Robertson et al., 1988). The Bayesian solution is far more elegant, and Klugkist, Hoijtink and colleagues have shown how an encompassing approach leads to conceptually straightforward and computationally convenient solutions (Klugkist, Kato, & Hoijtink, 2005; Klugkist & Hoijtink, 2007; Mulder, Klugkist, van de Schoot, Meeus, & Hoijtink, 2009). In Haaf and Rouder (2017), we show that the encompassing approach is feasible in mixed linear models, and in the course of doing so, we assess whether everyone has true Stroop and Simon effects in the usual direction where congruent stimuli are identified more quickly than incongruent ones. Here, we extend the Haaf and Rouder development for systems factorial technology.

Model Specification

We develop six hierarchical Bayesian mixed linear models to instantiate the theoretically relevant positions that all people exhibit serial processing, all people exhibit parallel processing, all people exhibit coactive processing, or that there is variation in processing architecture across people.

Let $Y_{ijk\ell}$ be the ℓth response time for the ith participant in the jth level of Factor A

and the kth level of Factor B, $i = 1, \ldots, I$, $j = 1, 2$, $k = 1, 2$, and $\ell = 1, \ldots, L_{ijk}$:

$$Y_{ijk\ell} \sim \text{Normal}\left(\tau_{ijk}, \sigma^2\right), \tag{1}$$

The cell mean parameters τ are additively decomposed as

$$\tau_{ijk} = \mu + \eta_i + \alpha_i s(j) + \beta_i s(k) + \gamma_i s(j)s(k), \tag{2}$$

where $s(m) = (-1)^m$ for $m = 1, 2$. The parameter μ is the grand mean, η_i, α_i, β_i, γ_i describe the ith participant's overall deviation from grand mean, main effect for Factor A, main effect for factor B, and interaction, respectively. The function s is a compact means of imposing the usual sums-to-zero balance constraints in linear models. This parameterization is similar to classic ANOVA parameterizations though it holds for each participant rather than across participants.

With this specification, M_i, the true interaction contrast is γ_i.

The serial architecture is straightforwardly implemented by placing the following constraints on γ_i:

$$\text{Serial:} \qquad \gamma_i = 0. \tag{3}$$

One way of understanding this specification and the ones that follow is to plot the interaction term for one individual as a function of another. The first column in Figure 3, labeled "Conditional Models," illustrates the model specifications for two individuals' interaction terms, γ_1 and γ_2. The restriction of the serial model, that all $\gamma_i = 0$, is shown in the first row of Figure 3.

Parallel models imply that each $\gamma_i < 0$. One way of instantiating this implication is

$$\text{Parallel-1} \qquad \gamma_i = \nu_\gamma, \;\; \nu_\gamma < 0. \tag{4}$$

Here, all participants share a common value, ν_γ, that is constrained to be negative. The constraint that all γ_i are the same and negative is shown in the second row of Figure 3 (first column). Here, the equality constraint is represented by the diagonal line. The line is darker the closer it is to (0, 0), denoting that smaller values of ν_γ are more likely than larger ones. We will make a prior assumption subsequently to this effect.

A second instantiation of the parallel model is

$$\text{Parallel-2:} \qquad \gamma_i \sim \text{Normal}_-(\nu_\gamma, g_\gamma \sigma^2), \tag{5}$$

where Normal$_-$ denotes a normal distribution truncated above at zero. In this specification, each participant has their own unique interaction parameter, and all of these parameters are constrained to be negative. The variance in parallel-2 is the product of g_γ and σ^2. This approach of placing priors on a parameter g was introduced by Zellner and Siow (1980) and is popular for analysis of linear models (Bayarri & Garcia-Donato, 2007; Liang, Paulo, Molina, Clyde, & Berger, 2008; Rouder & Morey, 2012; Rouder, Morey, Speckman, & Province, 2012). The parameter g may be viewed as a variance of γ_i/σ, each individual's true MIC measured in effect-size units rather than in time units.

An example of a parallel-2 model is shown in the first column and third row of Figure 3, where there is mass throughout the relevant lower-left quadrant. To draw the figure, we set $\nu_\gamma = 0$ and $g_\gamma \sigma^2 = .07^2$. For these settings, there is more mass close to the origin, (0, 0), denoting that smaller magnitudes for both, γ_1 and γ_2, are more likely. The parallel-2 specification is far more flexible than parallel-1 as it allows for individual variation. Parallel-1, in contrast, is far more compact with a model dimensionality of a single parameter rather than I parameters. Parallel-1 may be a preferred description when there are not sufficient observations per participant to resolve true participant variability. Moreover, the parallel-1 model is comparable in parsimony to the serial model in that there is no individual variability in either specification.

Coactive models imply that each $\gamma_i > 0$, and we use the same two specifications.

$$\text{Coactive-1:} \quad \gamma_i = \nu_\gamma, \;\; \nu_\gamma > 0, \tag{6}$$

$$\text{Coactive-2:} \quad \gamma_i \sim \text{Normal}_+(\nu_\gamma, g_\gamma \sigma^2). \tag{7}$$

As with the parallel models, the second specification captures individual differences in the interaction while the first one does not. Both specifications are shown in the first column of Figure 3 (fourth and fifth row), and for coactive-2, the settings are $\nu_\gamma = 0$ and $g_\gamma \sigma^2 = .07^2$.

In these models, all participants share a common architecture; that is, either everyone displays serial processing, everyone displays parallel processing, or everyone displays coactive processing. These models are parsimonious in that they do not specify differences across people allowing for a straightforward interpretation and easy generalization to larger populations. But it might be that people truly vary. Some might truly perform the task in a serial fashion; others might truly perform the task in a parallel fashion, and still others might truly perform the task in a coactive fashion. To account for this possibility, we included a general model:

$$\text{General:} \quad \gamma_i \sim \text{Normal}(\nu_\gamma, g_\gamma \sigma^2). \tag{8}$$

There are no constraints on γ_i other than the parametric shape specification. A graphical representation is shown in the last row, first column of Figure 3.

Misspecification

The advantage of the normal specification in (1) are two-fold: (i) the normal is computationally convenient in this application leading to rapid model development and quickly converging chains, and (ii) the MIC is easily parameterized and the placement of constraint, say that MIC must be positive, is straightforward to implement. These advantages are substantial, yet researchers may be concerned about the misspecification of

the normal. RT is skewed rather than symmetric, and the standard deviation tends to increase with the mean (Luce, 1986; Rouder et al., 2010; Wagenmakers & Brown, 2007).

Yet, we think any concern is misplaced. The main reason is that the logical-rules variant of systems factorial technology that we employ here uses expected values of cell means in drawing inferences about architecture. This point is critical—if we knew the true values of the cell means, we would not need to know the true shapes or true variances. The inference here has all the robustness of ANOVA or regression, which is highly robust for skewed distributions, so long as the left tail is thin. Indeed, RTs tend to have thin left tails that fall off no slower than an exponential (Burbeck & Luce, 1982; Van Zandt, 2000; Wenger & Gibson, 2004). Hence, we are not worried about the skewness of RT.

More pressing is the known positive relation between mean and variance, which violates the homogeneity assumption made above. The usual course is to propose a variance-stabilizing transform, say take the logarithm of RT after subtracting a minimum residual (e.g., Rouder, Province, Morey, Gomez, & Heathcote, 2015). Yet, this approach cannot be used here because it may affect the sign of the interaction term. If MIC $= 0$ in the untransformed space, then it will be negative in the log-transformed one. Hence, we cannot simply apply such a transformation. Yet, we think the heterogeneity is not problematic. The issue here is that heterogeneity is marginally violated given the relatively small experimental effects. Variance is most heterogeneous in Experiment 2b reported subsequently. But even here, averaged variances within the four cells differ by about 30%. This is a small degree of heterogeneity compared to those usually studied for robustness. We had to search to find studies that used this small degree as most use variances that differ by a factor of 2 (100%) or more. Perhaps the closest match is Rogan and Keselman (1977), who simulated data from variances that differ by as little as 50%. They found for the 50% variance case, the Type I error rates were just slightly inflated—they were at .053 for nominal .05 levels. Such a result indicates a high degree of stability of the sampling distribution of F in the face of the degree of heterogeneity present in our data.

After discussing model comparison and analyzing the data, we provide a small simulation study as a final check of the usefulness of our normal homogeneous models with skewed heterogeneous RT data.

Prior Specification

Bayesian analysis proceeds with specification of priors on all parameters. Consider the most richly parameterized model, the general model. We follow Rouder et al.'s multiple g-prior approach for factorial designs (2012). In the g-prior setup, parameters μ and σ are treated as parameters that locate and scale the models. They are common in all models, and as such, noninformative reference priors may be placed upon them, $\pi(\mu, \sigma^2) \propto 1/\sigma^2$ (Jeffreys, 1961). The remaining parameters have g-prior specifications are as follows:

$$\begin{aligned}
\eta_i &\sim \text{Normal}(0, g_\eta \sigma^2), \\
\alpha_i &\sim \text{Normal}(\nu_\alpha, g_\alpha \sigma^2), \\
\beta_i &\sim \text{Normal}(\nu_\beta, g_\beta \sigma^2), \\
\gamma_i &\sim \text{Normal}(\nu_\gamma, g_\gamma \sigma^2).
\end{aligned}$$

Specification is needed for mean parameters $(\nu_\alpha, \nu_\beta, \nu_\gamma)$ as well as variance multipliers $(g_\eta, g_\alpha, g_\beta, g_\beta)$. The usual course is to use normal and inverse-gamma distributions for mean and variance multipliers respectively:

$$\begin{aligned}
\nu_\alpha &\sim \text{Normal}(0, g_{\nu_\alpha} \sigma^2), \\
\nu_\beta &\sim \text{Normal}(0, g_{\nu_\beta} \sigma^2), \\
\nu_\gamma &\sim \text{Normal}(0, g_{\nu_\gamma} \sigma^2).
\end{aligned}$$

and

$$
\begin{aligned}
g_{\eta} &\sim \text{Inverse Gamma}\left(\frac{1}{2}, \frac{r_{\eta}^2}{2}\right), \\
g_{\alpha} &\sim \text{Inverse Gamma}\left(\frac{1}{2}, \frac{r_{\alpha}^2}{2}\right), \\
g_{\beta} &\sim \text{Inverse Gamma}\left(\frac{1}{2}, \frac{r_{\beta}^2}{2}\right), \\
g_{\gamma} &\sim \text{Inverse Gamma}\left(\frac{1}{2}, \frac{r_{\gamma}^2}{2}\right), \\
g_{\nu_{\alpha}} &\sim \text{Inverse Gamma}\left(\frac{1}{2}, \frac{r_{\nu_{\alpha}}^2}{2}\right), \\
g_{\nu_{\beta}} &\sim \text{Inverse Gamma}\left(\frac{1}{2}, \frac{r_{\nu_{\beta}}^2}{2}\right), \\
g_{\nu_{\gamma}} &\sim \text{Inverse Gamma}\left(\frac{1}{2}, \frac{r_{\nu_{\gamma}}^2}{2}\right).
\end{aligned}
$$

The vector $\boldsymbol{r} = (r_{\eta}, r_{\alpha}, r_{\beta}, r_{\gamma}, r_{\nu_{\alpha}}, r_{\nu_{\beta}}, r_{\nu_{\gamma}})$ are prior settings for the models, and values must be chosen before hand. The values set the *scale* on the variabilities of the relevant parameters. We are not directly setting how much variability there is in a mean effect or the variability of effects, but, coarsely, how variable the prior on variability is. These settings must be reasonable. They are scales on variabilities, and as such, represent approximate expectations of these variabilities. We justify the settings below and address how the choices affect inference in a separate section following data analysis.

These scale settings may be set by using known regularities in RT. The latencies analyzed here are fairly slow for perception and working memory tasks with RTs averaging over 1 second. For latencies in this range, the standard deviation for repeated observations (same person, same condition), σ, is around 600 ms or so (Wagenmakers & Brown, 2007). The scale settings are made relative to σ. Consider r_{η}, the variability of individual's overall speed effect, for example. In our experience across many tasks, we find the mean of individuals to be about as variable as repeated observations. For example, suppose one

individual is relatively fast and has RTs between say 300ms and 600 ms, a relatively slow individual would have responses between 600 ms and 900 ms. In this sense the variability across the means of people is on the order of the variability within a person. Hence, we set $r_\eta = 1$ to convey this knowledge. Mean effects, on the other hand, might be around 90 ms, or 15% of σ. This calculation provides for the setting that $r_{\nu_\alpha} = r_{\nu_\beta} = r_{\nu_\gamma} = .15$ The settings of $(r_\alpha, r_\beta, r_\gamma)$ reflect the scale of *a priori* variability across individuals in these effects. We figure that under a general model without constraint, people would be spread no more than 100 ms in their true effect, and quite possibly less. As a middle-of-the-road position, we use 60 ms, 10% of σ. Therefore, $r_{\nu_\alpha} = r_{\nu_\beta} = r_{\nu_\gamma} = .1$.

The remaining priors are on ν_γ in parallel-1 and coactive-1 models. We use the respective half normals. For parallel-1, $\nu_\gamma \sim \text{Normal}_-(0, r_\gamma\sigma^2)$; for coactive-1, $\nu_\gamma \sim \text{Normal}_+(0, r_\gamma\sigma^2)$. The setting of r_γ is the same as above.

Hierarchical model structure

One of the key features of the parallel-2, coactive-2, and general models is that they specify individual variability. Individual interactions, γ_i, are not free to vary arbitrarily. Instead, they are constrained to follow a hierarchical structure through a common group mean, ν_γ. Variability in γ_i comes either from the individual variability of the interaction term, $g_\gamma\sigma^2$, or from the variability in the mean effect, ν_γ. The latter variability is shared between all individuals, leading to correlation among γ_i. The second column in Figure 3, labeled "Marginal Models", shows these correlations for parallel-2 (third row), coactive-2 (fifth row), and general models (sixth row). In comparison to the first column, where the specification for set ν_γ and $g_\gamma\sigma^2$ are shown, the second column illustrates the model specifications integrated over all possible settings of these parameters. In this regard, the first column is a conditional specification and the second column is a marginal specification.

This correlation is impactful. If it is small, then the model has a large dimensionality, and the effective number of interaction contrasts apporaches I. If it is large, then the model

dimensionality is closer to that of parallel-1 or co-active-1 with a common parameter for all individuals' interactions. This fact, that correlation sets model dimensionality, means that prior settings that affect this correlation are going to have a large influence on inference. The key prior settings are r_{ν_γ}, the scale of the prior variability of the common component, and r_γ the scale of the prior variability of the individual component. As $r_{\nu_\gamma} \to 0$, there is decreasing shared variability of the individuals' interaction terms, and the marginal model approaches the conditional model in Figure 3. For I individuals, the interaction component of the model would be an I-dimensional ball. In contrast, as $r_\gamma \to 0$, there is no individual variability and the dimensionality of the interaction components reduces to a single parameter, like the parallel-1 and coactive-1 models. Since the prior settings define the dimensionality of the model, it is reasonable to expect that model comparison will be dependent on these settings. We will address this dependency after the main analysis.

Model Comparison

Bayes Factors

We use the Bayes factor model-comparison approach (Jeffreys, 1961) to state evidence for the six models. The Bayes factor is the direct consequence of using Bayes rule for computing the plausibility of competing models in light of data. The key equation for comparing two models, $\mathcal{M}_A$ and $\mathcal{M}_B$ is

$$\frac{P(\mathcal{M}_A|\boldsymbol{Y})}{P(\mathcal{M}_B|\boldsymbol{Y})} = \frac{P(\boldsymbol{Y}|\mathcal{M}_A)}{P(\boldsymbol{Y}|\mathcal{M}_B)} \times \frac{P(\mathcal{M}_A)}{P(\mathcal{M}_B)}, \tag{9}$$

where the term on the left-hand side is the posterior odds for the models, the term on the far right, $\frac{P(\mathcal{M}_A)}{P(\mathcal{M}_B)}$ is the prior odds, and the term $\frac{P(\boldsymbol{Y}|\mathcal{M}_A)}{P(\boldsymbol{Y}|\mathcal{M}_B)}$ is the Bayes factor. The Bayes factor is the key quantity; it denotes how analysts should update their beliefs about the models in the light of the data, $\boldsymbol{Y}$. Rouder, Morey, and Wagenmakers (2016) make the point that because the Bayes factor describes not beliefs but how data affect beliefs, it may

be viewed as the strength of evidence from the data.

There is a second equally important intepretation of the Bayes factor as predictive accuracy. $P(\boldsymbol{Y}|\mathcal{M}_A)$, the numerator of the Bayes factor, and $P(\boldsymbol{Y}|\mathcal{M}_B)$, the denominator, are the (joint) probability density of the data under the models. These densities are uniquely Bayesian constructs and may be viewed as predictions for a given model (Morey, Romeijn, & Rouder, 2016; Rouder et al., 2016). They are predictions in that a probability may be assigned to each possible outcome. Some possible outcomes will have a high probabilities under the given model while other outcomes will have low probabilities. In this sense, the model is predicting how probable an outcome is. The term $P(\boldsymbol{Y}|\mathcal{M}_A)$ is exactly how probable the observed data are under model $\mathcal{M}_A$. The Bayes factor therefore is how well one model predicts the data relative to the other model, that is, the relative predictive accuracy of two models. If the Bayes factor is 10, for example, the data are ten times more likely under Model $\mathcal{M}_A$ than under Model $\mathcal{M}_B$.

The third column of Figure 3 shows the predictions for the data from the six models. These predictions are smeared versions of the models specification. The darkness of a point predicts how probable it is. Model comparison may be done by comparing the density (or darkness of points) of any two models for observed data. The smearing of the model specification accounts for sampling noise. For example, the serial model predicts non-zero values of observed interaction contrast through smearing.

The dimensionality of the models affect their predicitve accuracy. Consider the parallel-1 and parallel-2 models. Both models may predict observations at $\hat{M}_1 = -.1$ sec and $\hat{M}_2 = -.095$ sec fairly well. Still, the Bayes factor is 1.9-to-1 in favor of the parallel-1 model. The reason for this preference is the parsimony of the model: While the Parallel-2 model predicts all combinations of negative interactions fairly well, the parallel-1 model predicts a smaller number of combinations of observed interaction terms that are closer to the diagonal.

Computation

Although the concept of Bayes factors is relatively simple, application is often times inconvenient and computationally difficult. The difficulty resides in calculating the probability of data given a model, $P(\boldsymbol{Y}|\mathcal{M})$. This probability may be expressed in conditional form as $P(\boldsymbol{Y}|\mathcal{M}) = \int_{\boldsymbol{\theta}\in\Theta} P(\boldsymbol{Y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ is a vector of parameters from parameter space Θ with prior $\pi(\boldsymbol{\theta})$. The integration is not at all trivial. In most cases, the integral is highly multidimensional and does not admit closed-formed solutions. In our case, for example, the integral for the general model in Experiment 1A comprises 136 dimensions. More alarmingly, the integrand is often highly peaked so much so that integration is more like finding a needle in a haystack. Numerical methods often fail, and finding computationally convenient solutions for Bayes factors in mixed settings remains timely and topical.

To make the computation of Bayes factor convenient for the models at hand, we follow the development by Haaf and Rouder (2017). The key is the use of g-priors, and this form allows for a symbolic integration of all parameters in the general model up to the set of g parameters (Zellner & Siow, 1980). This symbolic-integration-solution is developed in Rouder et al. (2012) and implemented in the *BayesFactor* package for R (Morey & Rouder, 2015). The *BayesFactor* package ist most useful for comparison between models without order constraint.

Figure 4 provides an overview of computation. In the center are three models, serial, general, and a new model, the one-effect model. In this model all $\gamma_i = \nu_\gamma$, but unlike parallel-1 and coactive-1, there is no order restriction. The specification of ν_γ is $\nu_\gamma \sim \text{Normal}(0, g_\gamma, \sigma^2)$. The Bayes factors between these three models may be computed in the BayesFactor package. The remaining models have order constraints. To compare these to their unconstrained analogs, we use the *encompassing approach* from Klugkist, Hoijtink and colleagues to calculate Bayes factors (Klugkist et al., 2005; Klugkist & Hoijtink, 2007). Haaf and Rouder (2017) provide an overview of how the encompassing approach works in

this context. Figure 4 shows comparisons made with this encompassing approach.

Experiment 1

In Experiment 1 we assess whether chunking affects processing by comparing the aforementioned models in a perception condition and in a working-memory condition. In the perception condition, participants were presented two screwhead stimuli that may vary in the size of the screw and the orientation of the slot (see Figure 1). They had to decide if both features differed or if at least one feature was the same. The working-memory task consisted of the same stimuli, but instead of comparing two simultaneously presented screwheads, participants compared a presented screwhead to one presented one second previously and available only from memory.

Method

Participants. A total of 64 University of Missouri students participated in exchange for course credit. Two were discarded for below-chance performance and another for excessively long response times that averaged over 5 seconds.

Stimuli & Design. Stimuli were pairs of screwheads that varied in two features: size and orientation. Each feature could either be the same, differ by a small amount, or differ by a large amount. Crossing these three levels yields nine possible combinations as shown in Table 1.

In the perception condition, screwheads were presented in white on a black background. The screwhead on the left served as the standard. It had a radius that varied between 54 and 180 pixels (chosen randomly from a uniform distribution) and had an orientation that varied across all possible angles (again, chosen randomly from a uniform distribution). The screwhead on the right had either the same size radius (no change), a radius that was 15% larger or smaller (small change) or a radius that was 30% larger or smaller (large change). Likewise the screwhead on the right had either the same orientation, a $\pm 20^\circ$ orientation difference (small change) or a $\pm 60^\circ$ orientation difference (large

change). Changes in size were equally likely to be an enlargement or a reduction in radius; changes in orientation were equally likely to be clockwise or counterclockwise in direction.

In the memory condition, screwheads were white presented on a grey background, and this change of background was needed to reduce the formation of after images of the first stimulus. Unfortunately, participants in a pilot study were unable to perform the memory task at sufficiently high performance with these parameters. To provide for high accuracy, we increased the differences in features across the screwheads for this condition. The radius changes were 30% and 50% and the orientation changes were $\pm 35^{\circ}$ and $\pm 75^{\circ}$ of difference.

Task condition, whether memory or perception, was manipulated in a between-subject manner with thirty and thirty-two participants performing in the memory and perception conditions, respectively.

Procedure. A trial consisted of the events shown in Figure 1A and 1B for the perception and working memory conditions, respectively. There were 9 types of trials comprised of the crossing of feature levels as shown in Table 1. The five negative trial types, $(0, 0)$, $(0, 1)$, $(0, 2)$, $(1, 0)$, and $(2, 0)$, each occurred with probability .1. The four positive types, $(1, 1)$, $(1, 2)$, $(2, 1)$, and $(2, 2)$ each occurred with probability .125. There were 360 experimental trials in a session, and these were preceded by 18 practice trials. Participants were given a pleasant doubled beep for correct responses and a less pleasant buzz for wrong ones. Trials were blocked in groups of 60, and participants were given a self-paced break between blocks. Trials were self paced, and participants started each by pressing a space bar. Positive and negative responses were made by pressing the '/' and 'z' key, respectively.

Results & Discussion

The key target for analysis is the interaction between the factors. One worry in systems factorial research is that the manipulations are ineffective, and if so, then a null interaction might be expected. With ineffective manipulations a null interaction could be misdiagnosed as serial processing. Hence, we first check the overall effectiveness of the

manipulations.

Figure 2A-B show the empirical results for Experiment 1A, the perception condition. At the aggregate level, there are reasonably large main effects of angle difference (0.14 s) and radial size difference (0.227 s). Error bars denote within-subject pooled-variance 95% CIs (Masson & Loftus, 2003; Rouder & Morey, 2005). From the graph it is clear that both manipulations were effective. This fact licenses the ensuing analyses of the interactions.

Inspection of individual observed MICs seemingly provides support for a serial architecture for most participants. Figure 2C-D show the same for Experiment 1B, the memory condition. The main effects of size and angle differences are a bit smaller (0.043 s and 0.079 s for angle and size, respectively), and inspection of individual MICs in Figure 2D seemingly provides similar support for the serial conclusion.

We fit the general model to Experiments 1A and 1B using the BayesFactor package. Whereas the model is linear and well identified, it is not surprising that mixing was excellent as confirmed by parameter trace plots and autocorrelation functions. Figure 5A-B show the resulting parameter estimates. The open points are the observed MICs; the colored points are posterior means of γ_i, the interaction parameters. The shaded area is the 95% credible intervals for the model estimates. The estimates of γ_i order as the observed MICs, and in particular seemingly provide support for the serial conclusion. Notably, there is much shrinkage, which is an indication that the variation in observed MICs reflects sample noise to a large extent rather than true differences in individuals.

The definitive answers to the critical questions about processing architecture come from comparing the six aforementioned models. Table 2 provides the Bayes factors for both perception and memory conditions. The "*" indicates the most preferred model, and the serial model is most preferred for both the perception and memory conditions. The remaining values indicate the BF between the most preferred model and the remaining models. For example, for Experiment 1A, the listed value under the general model is 1-to-24. This value indicates that the serial model has 24 times the probability density at

the observed data than does the general model. Values of ≈ 0 indicate that the winning model has Bayes factors several orders of magnitude greater than the indicated model. From the table, the following results are clear: 1. the serial model performs notably well for both conditions. 2. The three models that specify individual differences—parallel-2, coactive-2, and the general model—fared poorly. Everyone seemingly processed these screwheads in serial in both the perception and working-memory tasks.

These results are somewhat surprising to us. We expected that recall from working memory would rely on a different architecture than perception, perhaps through consolidation, grouping, or chunking. Yet, we found both tasks were mediated by serial processing of features. We also expected that there might be noticeable individual differences. These expectation were guided by previous results in systems factorial technology where analysis of individuals reveals variability in processing (e.g., Little et al., 2011). Yet, we found evidence against individual variation in true interactions. Models with individual variation were heavily penalized for this flexibility, and models with a single true value fared much better. We discuss qualifications of these findings subsequently.

Experiment 2

Experiment 1 revealed serial processing in both the perception and working-memory tasks. In Experiment 2 we attempt to encourage chunking by using two-digit number stimuli instead of screwheads. An example stimulus was the number “46." We treat each digit as a feature, and refer to the 10s feature and the 1s feature. A difference in the 10s feature is seen in the difference between the numbers “46" and “56"; a difference in the 1s feature is seen in the difference between the numbers “46" and “47"; and a difference in both features is seen in the difference between the numbers “46" and “57." These differences in digits could be small, ± 1, as in the previous examples, or large, ± 3. For example "46" and "19" differ by a large amount in both features. Experiment 2 followed the same structure of Example 1; the main difference was the type of stimuli.

Whether the size of changes, from ±1 to ±3, affects responses deserves further scrutiny. Digit change size matters if digits are processed as magnitudes rather than as abstract symbols. Evidence for magnitude processing comes from the well-known distance-from-five effect (Moyer & Landauer, 1967). Participants in distance-from-five tasks must identify whether a single-digit number is less-than or greater-than five. Rouder, Lu, Speckman, Sun, and Jiang (2005) found that responses to digits far from five, e.g. 2 and 8, are responded to 50 ms faster than digits close to five, e.g. 4 and 6. In Experiment 2, we find similar effects across the change-size manipulation as is discussed subsequently.

Method

Participants. A total of 56 University of Missouri students participated in exchange for course credit. One was discarded for excessive errors.

Stimuli. The left-hand number served as the standard, and the digits that comprised it varied between 4 and 6, inclusively. In the small change condition, the digits of a common feature varied by ±1; in the large change condition digits in the common feature varied by ±3. Changes were equally likely to be positive or negative. The same nine combinations as Experiment 1 were used in the same frequencies.

Procedure. The procedure for Experiment 2 was identical to Experiment 1.

Results

We first check to make sure the manipulations were effective, especially since the digit changes, from ±1 to ±3, is not so large. Figure 6A and 6B show the empirical results for Experiment 2A, the perception condition. Figure 6C and 6D show the same for Experiment 2B, the memory condition. The effect across all conditions is about 109 ms, which is large for these type of digit effects (Moyer & Landauer, 1967; Rouder et al., 2005). Inspection also indicates the possibility of an interaction where the decrease in RT is disproportionately bigger when there is a large change in both digits. This type of interaction corresponds to a negative MIC. In contrast, inspection of individual MICs

seemingly provides support for a serial architecture for many participants. A minority of participants in the memory condition have a negative MIC, which is consistent with parallel processing for these participants.

Parameter estimates for each γ_i are shown in Figure 5. There is the familiar degree of shrinkage, but, importantly, there is also some evidence that several people exhibit parallel processing. The estimates for Experiment 2A, however, are ambiguous. It is difficult to tell if processing is parallel or serial, or if the outlying estimate is enough to favor the general model, which is the only model that is flexible enough to account for true dispersion across zero. Similar ambiguities are present for Experiment 2B; it is unclear if the trend toward parallel is influential or if the presence of several near-zero estimates are more compatible with serial processing.

These ambiguities are resolved with model comparison, and the Bayes factors are shown in Table 2. For Experiment 2A, the one with the sole outlier, the data are fairly equivocal between parallel processing and the general model. The slight preference for the general model reflects the leverage of a single participant with a large positive interaction that is too large to be due to sample noise. The Bayes factor for Experiment 2B reveals that the data are most compatible with the parallel-1 model, the parallel processing model without individual differences.

Effects of Prior Specifications

Bayesian analysis is predicated on specifying prior distributions on parameters. Analysts should be familiar with how these specifications affect model comparison. A few points of context are helpful. It seems reasonable as a starting point to require that if two researchers run the same experiment and obtain the same data, they should reach the same if not similar conclusions. Yet, almost all Bayesians note that priors have effects on inference. To harmonize Bayesian inference with the above starting point, many Bayesian analysts actively seek to minimize these effects by picking likelihoods, prior parametric

forms, and heuristic methods of inference so that variation in prior settings have marginal effects (Aitkin, 1991; Gelman, Carlin, Stern, & Rubin, 2004; Kruschke, 2012; Spiegelhalter, Best, Carlin, & van der Linde, 2002). In the context of these views, the effect of prior settings on inference is viewed negatively; not only is it something to be avoided, it is a threat to the validity of Bayesian analysis.

We reject the starting point above including the view that minimization of prior effects is necessary or even laudable. Rouder et al. (2016) argue that the goal of analysis is to add value by searching for theoretically-meaningful structure in data. Vanpaemel (2010) and Vanpaemel and Lee (2012) provide a particularly appealing view of the prior in this light. Accordingly, the prior is where theoretically important constraint is encoded in the model. In our case, the prior provides the critical constraint on the sign of the interaction term. The choice of prior settings are important because they unavoidably affect the predictions about data for the models (Figure 3). Therefore, these settings necessarily affect model comparison. We think it is best to avoid judgments that Bayes factor model comparisons depend too little or too much on priors. They depend on it to the degree they do. Whatever this degree, it is the degree resulting from the usage of Bayes rule, which in turn mandates that evidence for competing positions are the degree to which they improve predictive accuracy.

This call to embrace the prior as part of the model leads immediately to the realization that different researchers may reach different conclusions from the same data. Rouder et al. (2016) argue that this variation is not problematic. They recommend that so long as various prior settings are justifiable, the variation in results should be embraced as the legitimate diversity of opinion. We take this view. Our goal in understanding the dependence of Bayes factors on prior settings is to understand the diversity of opinions that may be drawn.

The critical prior settings are the scales on variability that determine the differences in the competing models, r_γ and r_{ν_γ}. We set these to 10% and 15%, respectively. The

interpretation is that in the general model, average interaction effects have a scale in variability that is 15% of the standard deviation σ, and the scale of individual variation around this average is 2/3rds of this value. These values reflect our past experiences with unidimensional strength variables. To explore the diversity of opinion, we took the limit of what we would consider reasonable values for these settings. We used four conditions. In one we doubled the scales ($r_\gamma = .2, r_{\nu_\gamma} = .3$), and these values represent an upper limit on effects in most psychological experiments. We also halved the scales ($r_\gamma = .05, r_{\nu_\gamma} = .075$), and these values are a lower limit on such effects. We also explored the ratio $r_{\nu_\gamma}/r_{\nu_\gamma}$, with upper and lower limits of 1 and .5, respectively.

We chose to explore the effect of prior settings in Experiment 1b and Experiment 2b because the observed MICs follow a difficult-to-interpret pattern. In Experiment 1b, it is not clear from inspection whether the slight negative effect is enough to support a parallel or serial model. In Experiment 2B, it is not clear if there is enough support for parallel or serial model, or if there is so much dispersion to support the general model. Because the data are middling between these positions, the possible effects of prior settings are enhanced. Table 3 shows the results of changing prior settings. This table follows the same format as Table 2—the most preferred model is indicated with a "*" and the remaining values in a row are relative to it. In our main analysis with $r_\gamma = .10$ and $r_{\nu_\gamma} = .15$, the best model was serial and parallel-1 for Experiments 1b and 2b, respectively. These conclusions remain unchanged across the different prior settings. That said, the Bayes factor values do vary appreciably and readers may use the range of these values as context in assessing the processing question.

The results in Table 3 may be understood in the context of the different constraints. One of the defining features of parallel-2, coactive-2, and the general model is that they allow for individual differences in true MIC. Hence, these models have much higher dimensionality than parallel-1, coactive-1, and the serial model as these have no individual differences in true MIC. One of our approaches to cope with this difference is to use

hierarchical specifications for parallel-2, coactive, and the general model (see Figure 3). Hierarchical models are becoming increasingly popular in psychology because they allow for both more accurate estimation of parameters and better generalization outside the sample to the population. In our case, the hierarchical specification reduces the dimensionality of individual differences by incurring correlation among individuals. The parameter that controls this model dimensionality is r_γ. Large settings imply much individual variation and a higher dimensional model. Small settings imply small degree of individual variation and a lower dimensional model. The degree of dimensional difference is a function of the number of participants, and the effect of r_γ becomes quite extreme with large numbers of people. Researchers using this approach to understand the nature of individual differences must be mindful of these dynamics (Haaf & Rouder, 2017).

Misspecification Revisited: A Simulation Study

The RT data in this report are skewed with a small degree of heterogeneity across conditions. The models, however, assume a normal, homogenous form across replicates. We argued previously that this misspecification is of limited concern because the main targets of inference are the expected values of cell means rather than the distribution of the replicates. Yet, some readers may worry about the usefulness of our normal homogeneous models with skewed heterogeneous RT data.

To understand what constitutes the usefulness of a model, it is helpful to have a sense of what models are and what they do. Models at their core are abstractions that exist in the platonic rather than real world (de Finetti, 1974). In this view they are neither true nor false—they are just models (Morey et al., 2016; Rouder et al., 2016). Perhaps a good metaphor is a subway map. Subway maps capture important constraints, say the order of stops on a line and the intersection of the lines. They do not capture all aspects of the subway, for example they distort the distances between stops and the color of the tracks. This combination of capturing important constraints while distorting other elements for

convenience and tractability is what we strive for as well. Our models capture theoretically important constraints among the individuals, say that all use one processing mode or another, while distorting for convenience and tractability nuisance elements. This approach can be compared to that in Houpt and Fific (2017) who take care to model the skew of response times—a nuisance element—while missing the constraints that all individuals may use the same processing.

The usefulness of the model, in our view, is its ability to capture the theoretically important constraints. To show that the treatment of the nuisance elements is inconsequential, we ran a small simulation study. In it, we generate simulated data from various known ground truths and see how the Bayes factors perform even when the models are misspecified. It is important, however, to recognize that these simulations assess the *frequentist properties* of model comparison. Once we condition on known truths, we have moved outside the Bayesian paradigm and entered the frequentist one where long-run error rates are central. We use the following simulations to show that the models have reasonable frequentist properties even when the ground truths do not match the nuisance assumptions.

To make the simulated data skewed and hetergeneous, we sampled RTs from an ex-Gaussian distribution (Heathcote, Popiel, & Mewhort, 1991). The effect of the manipulations was in the scale of the exponential component. Figure 7A shows three distributions: the quickest and most symmetric distribution (dotted lines) is for when both factors have large changes; the intermediate one (dashed lines) is for when one factor has a large change and the other has a small change; the slowest and most skewed one (solid lines) is for when both factors have small changes. Figure 7A shows a case with no interaction. To generate interactions, we added or subtracted an additional amount to the scale of the exponential for the (2,2) cell, the one where both factors have a small change.

To generate different interaction ground truths, we varied the distribution on this additional amount. For the serial ground truth, where each individual has no interaction, nothing is added. This setup is shown in Figure 7B by the arrow at zero. For the coactive

ground truth, the additional amount is distributed as a lognormal that is localized around 80 ms, a value chosen after observing the experimental data. Each individual's interaction parameter is sampled from this distribution on each run of the simulation. The distribution for the parallel ground truth is the negative of the coactive ground truth, and it is omitted to keep the figure less cluttered. Finally, for the unconstrained ground truth, the interaction distribution across individuals was widely spread. Simulation runs were modeled on Experiment 2B where there are 4697 observations distributed across 27 participants. For each run, all six Bayesian models were analyzed. Overall, there were 50 runs for each of the four different ground truth scenarios.

Figure 7 shows the Bayes factors for all truths, models, and runs. The center-left panel, labeled "Serial" shows the case where the serial case served as ground truth. The Bayes factors are normalized on each run so that all comparisons are made to the serial Bayesian model. Values below 1.0 mean that the serial model is preferred, and, indeed, the serial model won for 94% of the runs. One facet of the graph is that many values cluster at .0001. We rounded up all Bayes factors below this value for convenience in visualizing the simulation results. The observed values were often much smaller than this lower limit. The remaining plots show the same analysis for the other ground truths. Again, we used the same normalization where comparisons are relative to the model that best corresponded to the ground truth. The Bayes factor recovered the correct model with rates of 100%, 98%, and 82%, for the parallel, coactive, and unconstrained truths. These values indicate that the Bayes factor model-comparison has good frequentist properties even when misspecified.

There is no magic here. We designed the models to capture theoretically important truths on cell means. And that is why they are useful.

General Discussion

Systems factorial technology is an exciting methodology for addressing fundamental questions about processing architecture. We address some of the real-world statistical

difficulties in analysis. Our contribution here is the proposal and evaluation of models where all individuals have the same architecture. This type of "everyone does" formulation is the direct way to assess lawful regularity that undergirds theoretical propositions about the automaticity and universality of certain information-processing structures. In our case, we ask whether everyone uses serial, parallel, or coactive processing for simple perception and memory tasks. The alternative proposition is that the type of architecture varies across people. If so, then the architecture is not universal, and, perhaps, may even be under strategic control. To date, we know of no other attempt to conceptualize system factorial technology as universal or variational.

We compared models with Bayes factors to understand whether chunking in working memory changes the architecture of processing. Our experiments consisted of matched perception and memory tasks, and it was expected that this change in task would perhaps be associated with a change in processing. Plausibly, the chunking in working memory could have consolidated the features into a single whole that could be processed more efficiently than in serial. Such results, however, were not observed. For the screwhead task, where the features are highly separable, the usual serial result held not only for the perception version (Little et al., 2011), but for the memory version as well. For the digits task, where the stimuli are separable, but have common holistic interpretations, we observed that processing was parallel in both the perception and memory versions of the task. Hence, while the degree of separability of the features affected processing architecture, wether the task was perceptual or mnemonic did not. This combination of results provides a form of discriminative validity for the statistical methods. They are sensitive to the structure in the data in a reasonable way.

There was a marked lack of heterogeneity of processing architecture. The general model fared poorly overall and was competitive only in Experiment 2A. Here, we do observe an extreme value of MIC from a single individual that could not be regularized had outsized leverage. Outside of this observation, processing strategy seems to be the same for

all people within a given task. This result indicates that in a given setting architecture may be universal rather than under strategic control. Methodological approaches that rely on categorization (e.g., Houpt & Fific, 2017) implicitly assume heterogeneity and may understate the evidence for these substantively important regularities.

Perhaps the invariance of processing across perception and memory tasks reflects the low mnemonic demands. In most working memory tasks, participants are asked to hold more than one item. Indeed, the goal of most experiments is to push participants past their limits so that errors may be observed. It may be that binding or chunking of features is a consequence of having higher mnemonic demands, and for these stimuli presented in these configurations, the mnemonic demands were simply too low. Therefore, we treat the invariance result tentatively—it is noteworthy but certainly not yet definitive evidence against chunking or binding. How to increase mnemonic demands in this type of investigation to provide a more convincing test of chunking remains an open topic.

The final result and issue for discussion is a comparison of models without individual variability (serial, parallel-1, coactive-1) vs. those with it (general, parallel-2, coactive-2). The models without individual differences outperformed those that had them. This constellation might strike some as counterintuitive. It is nearly ubiquitous that individual variation is reported and expected across most tasks of human performance. And yet, we found evidence against such individual variation in the interaction parameter. The case is similar to our previous findings with reading speeds. Rouder, Tuerlinckx, Speckman, Lu, and Gomez (2008) provided a hierarchical Bayesian analysis of how lexical decision times varied with word frequency. The found that all individuals had the same 11% decline per doubling of word frequency. There was no evidence of deviation from this 11% value across the 50 participants. Likewise, Haaf and Rouder (2017) found that much of the variability across individual in Stroop, Simon, and Eriksen flanker tasks was due to trial noise rather than from true individual variability.

Some readers may find it difficult to consider models without true individual

variation. These models are probably good descriptions when the sample sizes (trials per participant) are not too large. For the resolution afforded by the data at hand, there is no need to consider individual variation. However, it may be the case with larger sample sizes, that models with individual variation are warranted. In this spirit, we recommend that models with constant effects be retained and taken seriously because at a minimum they indicate whether the data are sufficiently numerous to resolve individual variation. In this case, they do not provide such resolution. More generally, the results show that it takes evidence to document individual differences. Given the relative small size of individual variation in common cognitive tasks such as the ones here, it may take much larger sample sizes per individual to obtain this evidence than it is commonly assumed.

References

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

Atkinson, R. C. & Shiffrin, R. N. (1968). Human memory: a proposed system system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation: advances in research and theory. vol. 2* (pp. 89–195). New York: Academic Press.

Bayarri, M. J. & Garcia-Donato, G. (2007). Extending conventional priors for testing general hypotheses in linear models. *Biometrika*, *94*, 135–152.

Burbeck, S. L. & Luce, R. D. (1982). Evidence form auditory simple reaction times for both change and level detectors. *Perception & Psychophysics*, *32*, 117–133.

Cowan, N. (1995). *Attention and memory: An integrated framework*. Oxford Psychology Series. Oxford University Press.

Danker, J. F. & Anderson, J. R. (2010). The ghosts of brain states past: remembering reactivates the brain regions engaged during encoding. *Psychological Bulletin*, *136*(1), 87.

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

Estes, W. (1997). Processes of memory loss, recovery, and distortion. *Psychological Review*, *104*, 148–169.

Fific, M., Nosofsky, R. M., & Townsend, J. T. (2008). Information-processing architectures in multidimensional classification: A validation test of the systems factorial technology. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(2), 356–375.

Fific, M., Little, D. R., & Nosofsky, R. M. (2010). Logical-rule models of classification response times: a synthesis of mental-architecture, random-walk, and decision-bound approaches. *Psychological Review*, *117*(2), 309–348.

Garner, W. R. & Felfoldy, G. L. (1970). Integrality of stimulus dimensions in various types of information processing. *Cognitive Psychology*, *1*(3), 225–241.

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

Haaf, J. M. & Rouder, J. N. (2017). *Developing constraint in bayesian mixed models*. Revision submitted 3/17.

Heathcote, A., Popiel, S. J., & Mewhort, D. J. (1991). Analysis of response time distributions: An example using the Stroop task. *Psychological Bulletin*, *109*, 340–347.

Hebb, D. O. & Foord, E. N. (1945). Errors of visual recognition and the nature of the trace. *Journal of Experimental Psychology*, *35*(5), 335.

Houpt, W., J. & Fific, M. (2017). *A hierarchical bayesian approach to distinguishing serial and parallel processing*. submitted.

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

Kary, A., Taylor, R., & Donkin, C. (2016). Using bayes factors to test the predictions of models: a case study in visual working memory. *Journal of Mathematical Psychology*, *72*, 210–219.

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

Klugkist, I. & Hoijtink, H. (2007). The Bayes factor for inequality and about equality constrained models. *Computational Statistics & Data Analysis*, *51*(12), 6367–6379.

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

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

Little, D. R., Nosofsky, R. M., & Denton, S. (2011). Response time tests of logical-rule-based models of categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 1–27.

Little, D. R., Nosofsky, R. M., Donkin, C., & Denton, S. E. (2013). Logical rules and the classification of integral-dimension stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(3), 801–820.

Liu, Y. (1996). Queueing network modeling of elementary mental processes. *Psychological Review*, *103*, 116–136.

Luce, R. D. (1986). *Response times*. New York: Oxford University Press.

Maddox, W. T. & Ashby, F. G. (1993). Comparing decision bound and exemplar models of categorization. *Perception & Psychophysics*, *53*, 49–70.

Mandler, G. (1980). Recognizing: The judgment of previous occurence. *Psychological Review*, *87*, 252–271.

Masson, M. E. J. & Loftus, G. R. (2003). Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology*, *57*, 203–220.

McKinley, S. C. & Nosofsky, R. M. (1995). Investigations of exemplar and decision bound models in large, ill-defined category structures. *Journal of Experimental Psychology: Human Perception and Performance*, *21*, 128–148.

Miller, G. A. (1956). The magical number seven plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81–97.

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

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

Moyer, R. S. & Landauer, T. K. (1967). Time required for judgements of numerical inequality. *Nature*, *215*, 1519–1520.

Mulder, J., Klugkist, I., van de Schoot, R., Meeus, W. H. J., & Hoijtink, H. (2009). Bayesian model selection of informative hypotheses for repeated measurements. *Journal of Mathematical Psychology*, *54*(530-546).

Rissman, J. & Wagner, A. D. (2012). Distributed representations in memory: insights from functional brain imaging. *Annual Review of Psychology*, *63*, 101–128.

Robertson, T., Wright, F., & Dykstra, R. (1988). *Order restricted statistical inference.* Wiley, New York.

Rogan, J. C. & Keselman, H. (1977). Is the anova f-test robust to variance heterogeneity when sample sizes are equal?: an investigation via a coefficient of variation. *American Educational Research Journal*, *14*(4), 493–498.

Rouder, J. N. (2016). The what, why, and how of born-open data. *Behavioral Research Methods*, *48*, 1062–1069. Retrieved from 10.3758/s13428-015-0630-z

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

Rouder, J. N. & Morey, R. D. (2005). Relational and arelational confidence intervals: A comment on Fidler et al. (2004). *Psychological Science*, *16*, 77–79.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Pratte, M. S. (2007). Detecting chance: A solution to the null sensitivity problem in subliminal priming. *Psychonomic Bulletin and Review*, *14*, 597–605.

Rouder, J. N., Morey, R. D., & Wagenmakers, E.-J. (2016). The interplay between subjectivity, statistical practice, and psychological sciencecollabra. *Collabra*, *2*, 6. Retrieved from http://doi.org/10.1525/collabra.28

Rouder, J. N. & Morey, R. D. (2012). Default Bayes factors for model selection in regression. *Multivariate Behavioral Research*, *47*, 877–903. Retrieved from http://dx.doi.org/10.1080/00273171.2012.734737

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

Rouder, J. N., Province, J. M., Morey, R. D., Gomez, P., & Heathcote, A. (2015). The lognormal race: a cognitive-process model of choice and latency with desirable psychometric properties. *Psychometrika*, *80*, 491–513.

Rouder, J. N., Tuerlinckx, F., Speckman, P. L., Lu, J., & Gomez, P. (2008). A hierarchical approach for fitting curves to response time measurements. *Psychonomic Bulletin & Review*, *15*(1201-1208).

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

Schweikert, R. (1978). A critical path generalization of the additive factor method: Analysis of a Stroop task. *Journal of Mathematical Psychology*, *18*, 105–139.

Schweikert, R. & Townsend, J. T. (1989). A trichotomy: Interactions of factors prolonging sequential and concurrent mental processes in stochastic discrete mental (PERT) networks. *Journal of Mathematical Psychology*, *33*, 328–347.

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

Sternberg, S. (1969). The discovery of prossesing stages: Extensions of Donder's method. In W. G. Kosner (Ed.), *Attention and performance ii* (pp. 276–315). Amsterdam: North-Holland.

Townsend, J. T. (1990). Serial vs. parallel processing: sometimes they look like Tweedledum and Tweedledee but they can (and should) be distinguished. *Psychological Science*, *1*, 46–54.

Townsend, J. T. & Ashby, F. G. (1982). Experimental test of contemporary mathematical models of visual letter recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *8*, 834–864.

Townsend, J. T. & Nozawa, G. (1995). On the spatio-temporal properties of elementary perception: An investigation on parallel, serial, and coactive theories. *Journal of Mathematical Psychology*, *39*, 321–359.

Townsend, J. T. & Wenger, M. J. (2004). The serial-parallel dilemma: a case study in a linkage of theory and method. *Psychonomic Bulletin & Review*, *11*, 391–418.

Van Zandt, T. (2000). How to fit a response time distribution. *Psychonomic Bulletin and Review*, *7*, 424–465.

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

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

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*, 830–841.

Wenger, M. J. & Gibson, B. S. (2004). Assing hazard functions to assess changes in processing capacity in an attentional cuing paradigm. *Journal of Experimental Psychology: Human Perception and Performance*, *30*, 708–719.

Zellner, A. & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian statistics: Proceedings of the First International Meeting held in Valencia (Spain)* (pp. 585–603). University of Valencia.

Table 1
The Systems Factorial Task and Contrast

	Size		
Orientation	no change (0)	small change (1)	large change (2)
	Response Mapping		
no change (0)	-	-	-
small change (1)	-	+	+
large change (2)	-	+	+
	Cell Mean Notation		
no change (0)	$\bar{Y}_{00}$	$\bar{Y}_{01}$	$\bar{Y}_{02}$
small change (1)	$\bar{Y}_{10}$	$\bar{Y}_{11}$	$\bar{Y}_{12}$
large change (2)	$\bar{Y}_{20}$	$\bar{Y}_{21}$	$\bar{Y}_{22}$
	Interaction Contrast		
no change (0)	0	0	0
small change (1)	0	+	-
large change (2)	0	-	+

Table 2
Bayes factor values.

	Serial	Parallel-1	Parallel-2	Coative-1	Coactive-2	General
Expt. 1A	*	1-to-9.3	≈ 0	1-to-14.5	≈ 0	1-to-24
Expt. 1B	*	1-to-3.7	1-to-11696.1	1-to-26.2	≈ 0	1-to-7.4
Expt. 2A	1-to-2.7	1-to-1.1	1-to-196.6	1-to-129.8	≈ 0	*
Expt. 2B	1-to-6.5	*	1-to-145.2	1-to-383.6	≈ 0	1-to-6.1

Table 3
Effects of Prior Specification on Bayes Factors.

	Serial	Parallel-1	Parallel-2	Coative-1	Coactive-2	General
Experiment 1B						
Our Choice (.15,.10)	*	1-to-4.5	1-to-5972.5	1-to-30.1	≈ 0	1-to-7.6
Double (.30,.15)	*	1-to-6.6	≈ 0	1-to-49.4	≈ 0	1-to-113.3
Half(.075,.05)	*	1-to-2.1	1-to-94.8	1-to-15.3	≈ 0	1-to-2.6
Large Ratio (.15,.15)	*	1-to-3.8	1-to-22606	1-to-26.8	≈ 0	1-to-19.7
Small Ratio (.15,.05)	*	1-to-4	1-to-859.6	1-to-28.6	≈ 0	1-to-5.1
Experiment 2B						
Our Choice (.15,.10)	1-to-2.3	*	1-to-165.5	1-to-276.8	≈ 0	1-to-6.4
Double (.30,.15)	1-to-3.5	*	1-to-64685	1-to-399	≈ 0	1-to-82.3
Half(.075,.05)	1-to-11	*	1-to-5.4	1-to-343.8	≈ 0	1-to-2.4
Large Ratio (.15,.15)	1-to-6.6	*	1-to-2271.8	1-to-433.8	≈ 0	1-to-23.5
Small Ratio (.15,.05)	1-to-6	*	1-to-33.7	1-to-255.4	≈ 0	1-to-3.8

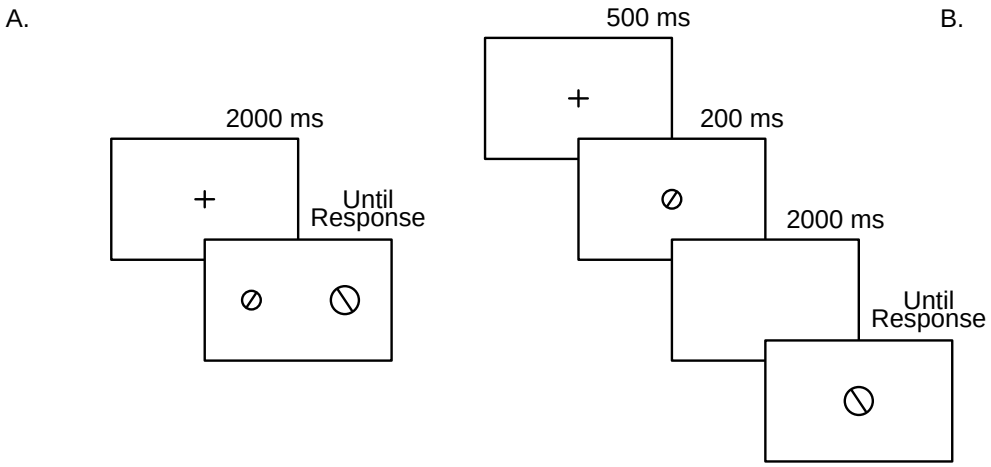

Figure 1. Paradigm for Experiment 1. **A.** Schematic of trials in the perception task. The participant decides if the screwheads differ in both size and slot orientation. **B.** Schematic of trials in the memory task.

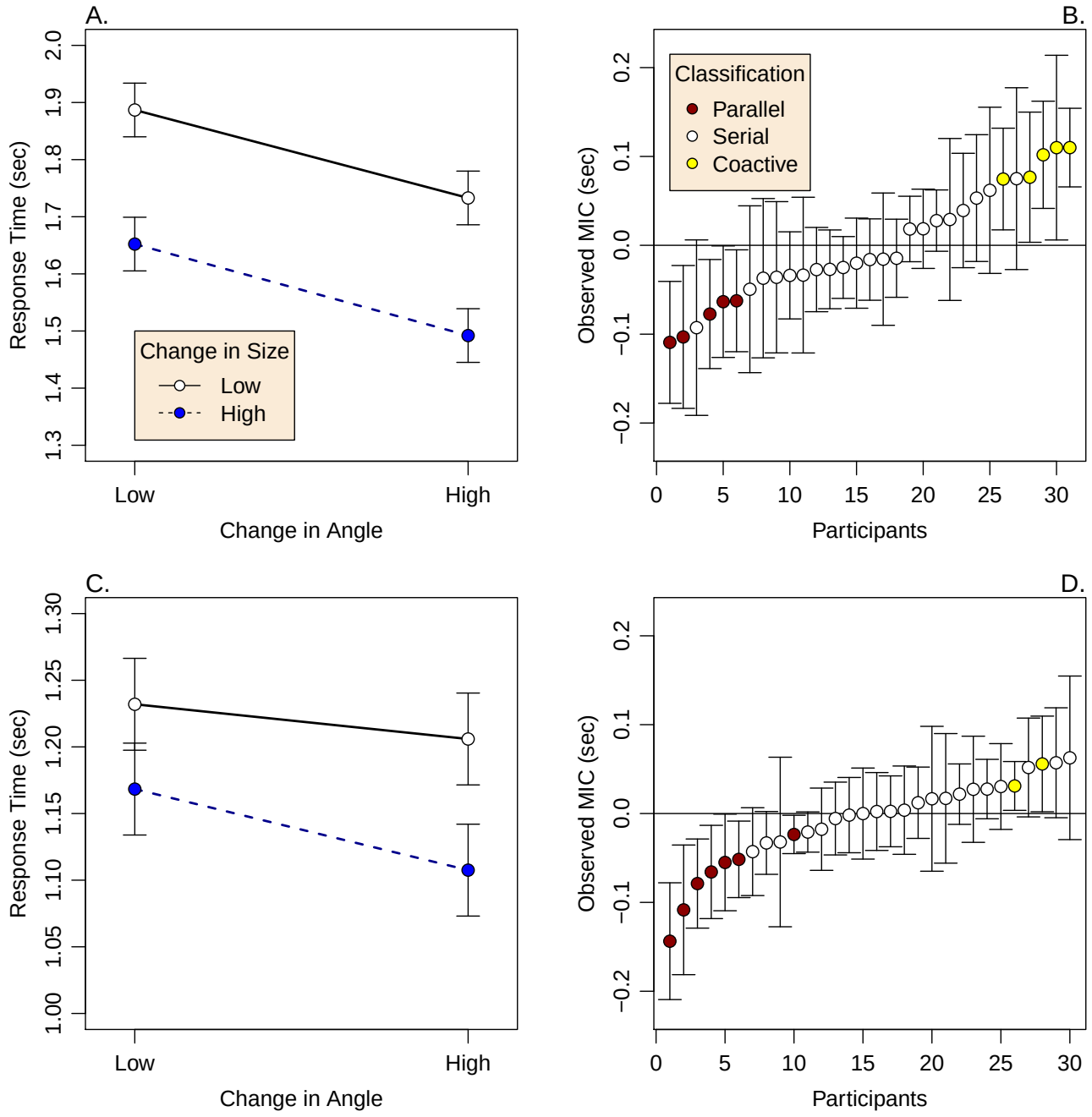

Figure 2. Results from Experiment 1: **A.,C.** Observed mean response times for Experiments 1A (perception) and 1B (memory), respectively. **B., D.** Observed mean interaction contrasts (MICs) with 80% confidence intervals for each individual. The CIs with open circles contain zero (serial processing); those with colored circles are either entirely above zero (coactive processing) or entirely below zero (parallel processing).

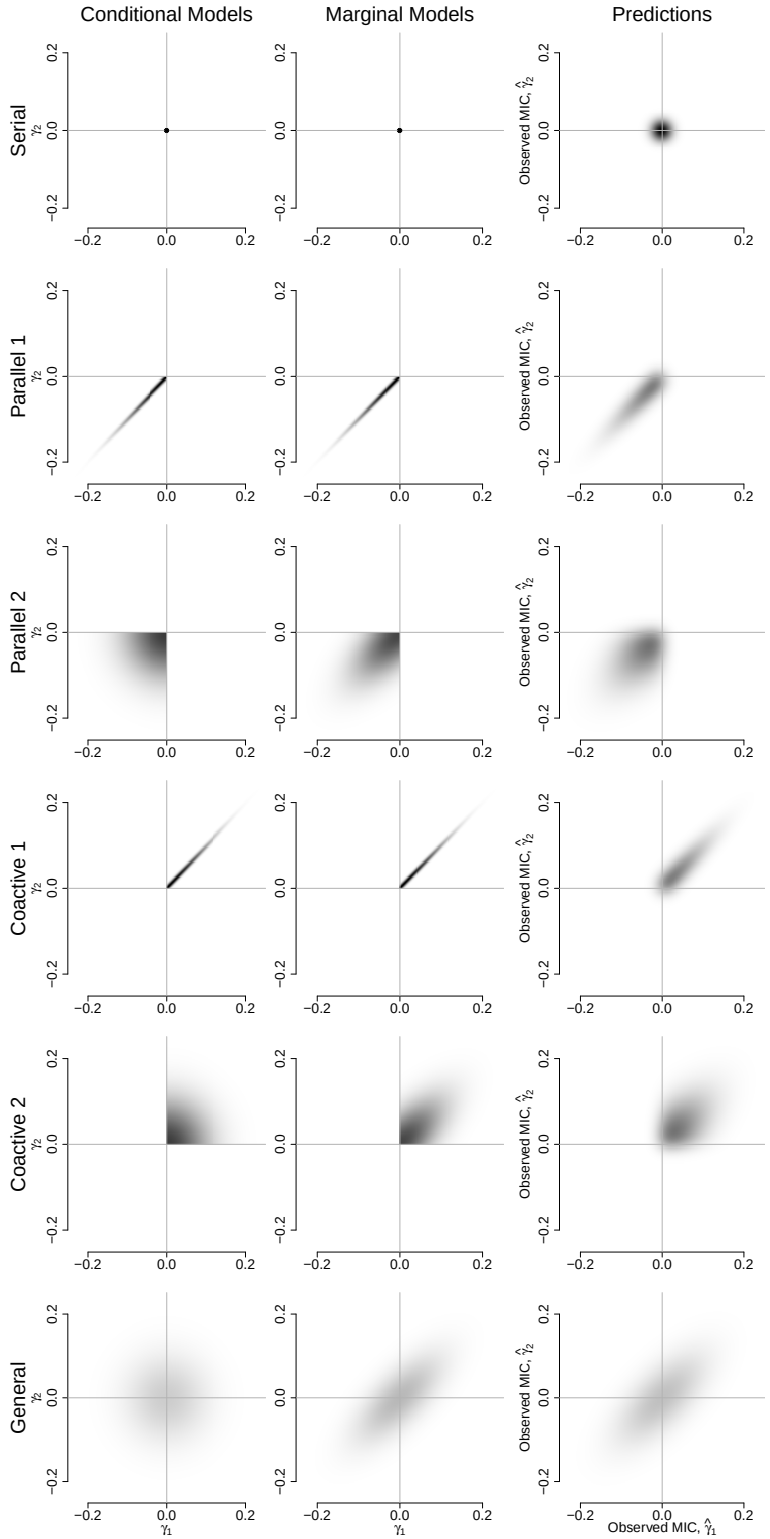

Figure 3. Model specification and predictions as shown for two individuals' parameters γ_1 and γ_2. Darker areas denote greater probability density. **Condition Model.** Specifications are made conditional on group parameters (mean=0, sd=.07). **Marignal Model.** The specification is marginalized over the priors on group parameters. The hierarchical structure induces correlations between γ_1 and γ_2. **Predictions.** Models yield predictions on observed MICs, and the role of sample noise is to smear the structure in the model. Bayesian model comparison consists of comparing these predictions at the observed data.

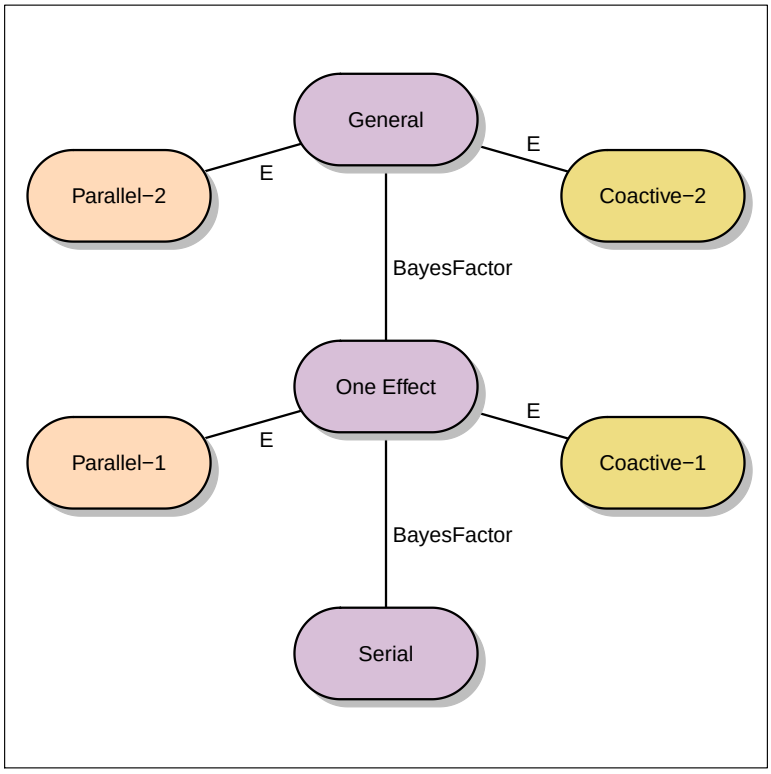

Figure 4. Bayes factor computation approaches for comparing the six critical models. The one-effect model serves as a critical bridge. The label “BayesFactor" refers to symbolic integration up to g parameters (Rouder, Morey, Speckman, & Province, 2012) as performed in the BayesFactor package (Morey & Rouder, 2015). The label “E" refers to the encompassing approach (Klugkist, Kato, & Hoijtink, 2005).

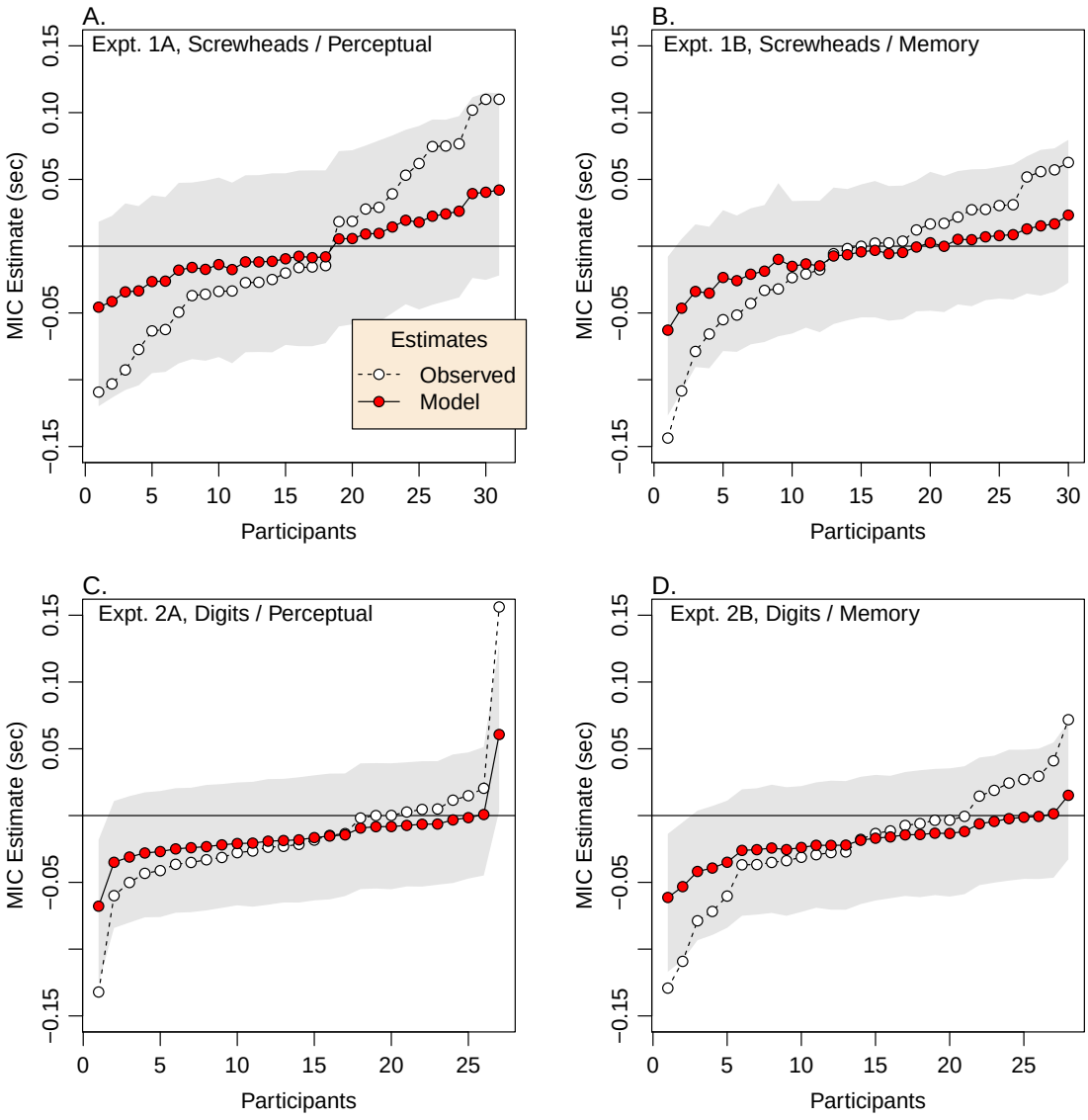

Figure 5. Estimates of the interaction parameters $\boldsymbol{\gamma}$ from the general model for all experiments. The filled points and the shaded area are the posterior means and 95% credible regions, respectively. The open points are the observed MIC values and are included to show the amount of shrinkage obtained by modeling trial and individual variability jointly.

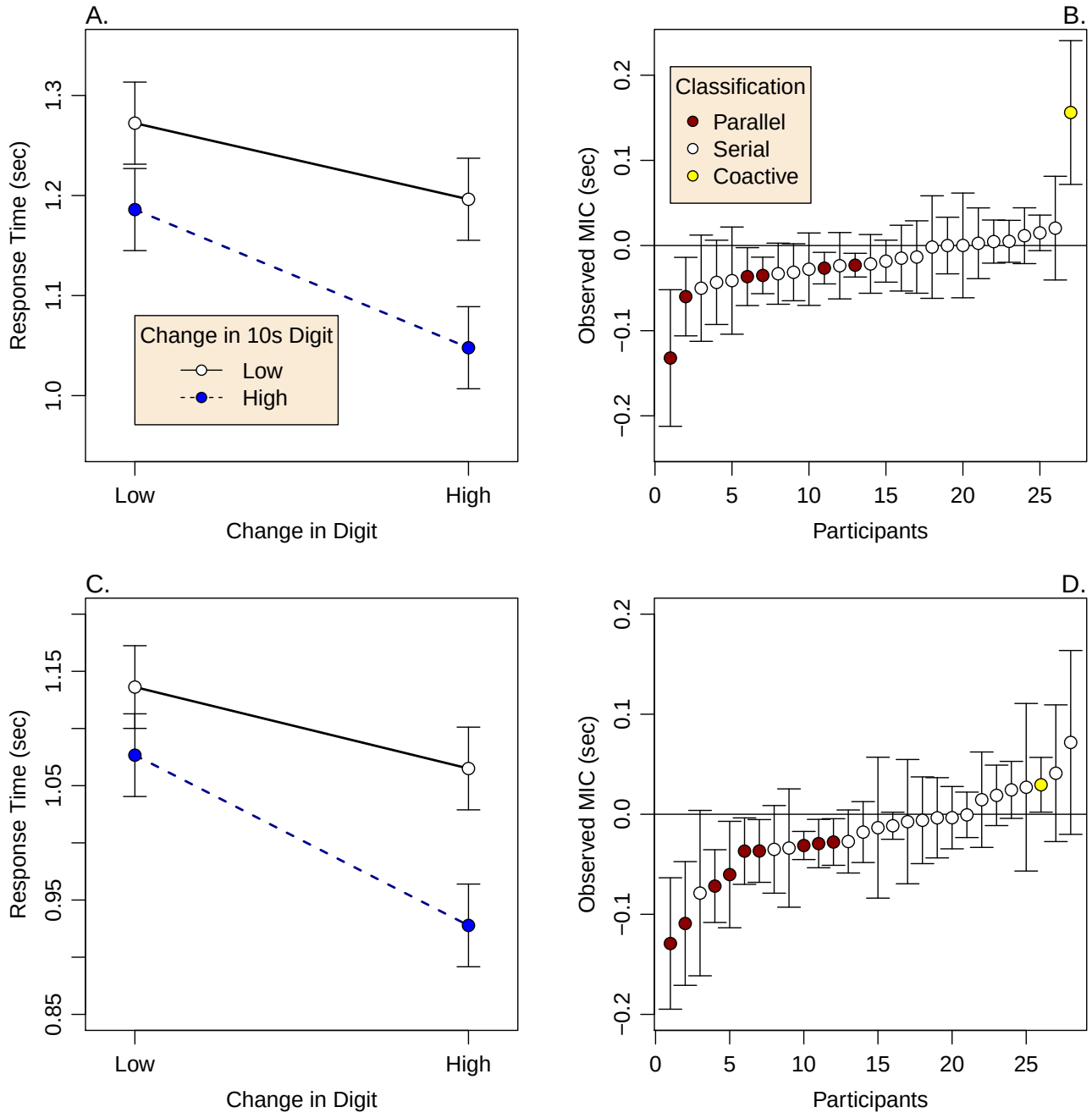

Figure 6. Results from Experiment 2: **A.,C.** Observed mean response times for Experiments 2A (perception) and 2B (memory), respectively. **B., D.** Observed mean interaction contrasts (MICs) with 80% confidence intervals each individual. The CIs with open circles contain zero (serial processing); those with colored circles are either entirely above zero (coactive processing) or entirely below zero (parallel processing).

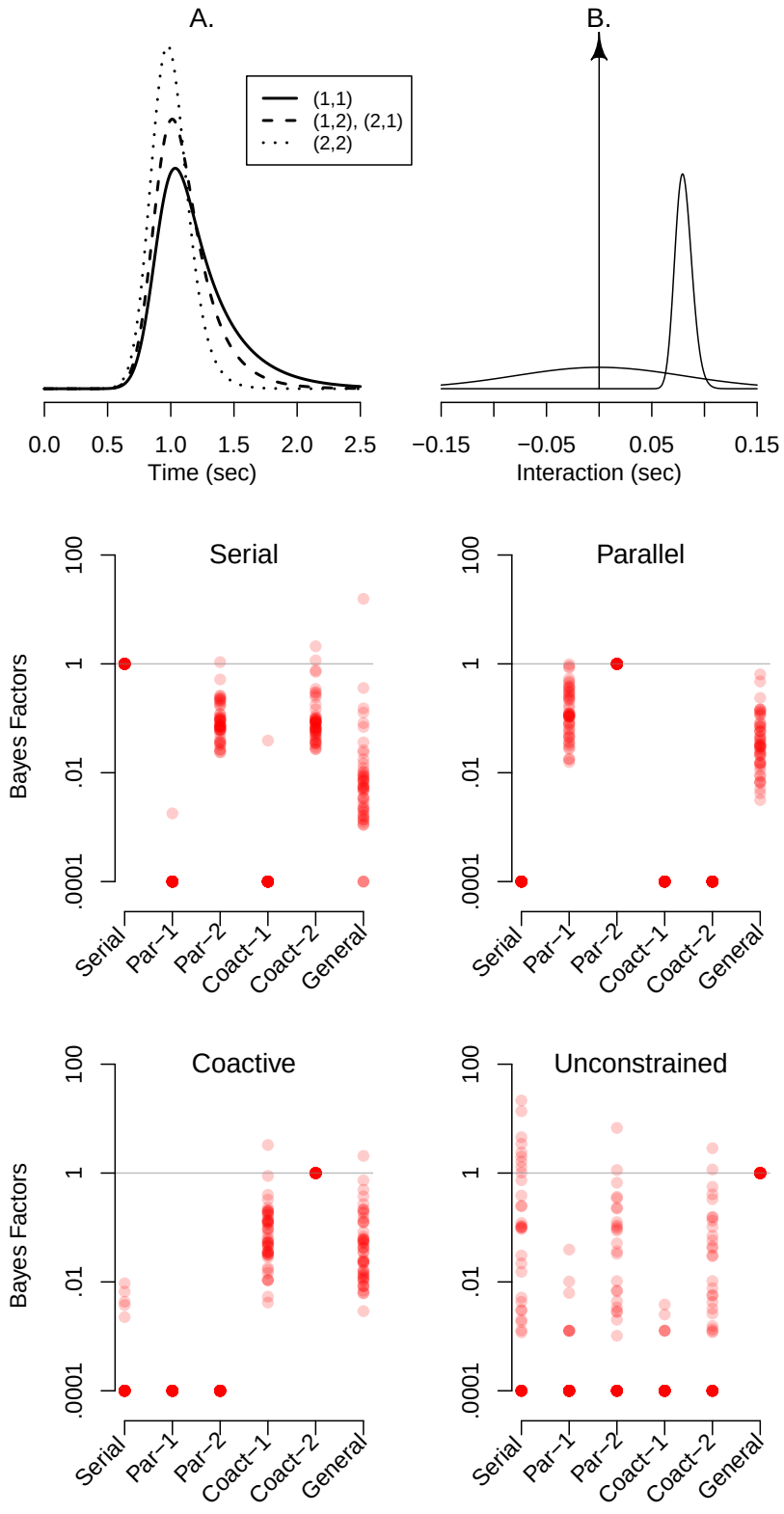

Figure 7. Simulation with misspecifications. **A.** Simulated data were drawn from an ex-Gaussian; the scale of the exponential component varied with condition. **B.** Ground truth interaction distributions across hypothetical participants. The serial truth is the spike at zero, the coactive truth is the positive distribution, the unconstrained truth follows the broad normal. **Middle and Bottom Row**. Bayes factor results for various ground truths. The true model was set to 1.0 on each run, and for the vast majority of runs, the true model was most preferred.