Bayesian modeling of the latent structure of individual differences in the numerical size-congruity effect

Thomas J. Faulkenberry and Kristen A. Bowman

Tarleton State University

Author Note

This paper was written in R-Markdown with code for data analysis integrated into the text. The RMarkdown file is available for download at https://git.io/vAEE8. This work was supported by a Faculty-Student Research Grant from Tarleton State University awarded to TJF.

Correspondence concerning this article should be addressed to Thomas J. Faulkenberry, Department of Psychological Sciences, Box T-0820, Tarleton State University, Stephenville, TX 76401. E-mail: faulkenberry@tarleton.edu

Abstract

When people are asked to choose the physically larger of a pair of numerals, they are often slower when relative physical size is incongruent with numerical magnitude. This size-congruity effect has applications across the field of numerical cognition, not only informing our understanding of mental representations of number, but also serving as an index for numerical ability in individuals. In this paper, we apply the methods of Haaf and Rouder (2017) to look at the size-congruity effect through the lens of individual differences. Here, we simply ask whether everyone exhibits the effect. We develop a class of hierarchical Bayesian mixed models with varying levels of constraint on the individual size-congruity effects. The models are then compared via Bayes factors, telling us which model best predicts the observed data. We then apply this modeling technique to three data sets. In all three data sets, the winning model was one in which the size-congruity effect was constrained to be positive. This indicates that, at least in the context of a physical comparison task with Arabic numerals, everyone exhibits a positive size-congruity effect. We discuss these results in the context of measurement fidelity and theory-building in numerical cognition.

Keywords: size-congruity effect, individual differences, hierarchical Bayesian model, Bayes factors

Bayesian modeling of the latent structure of individual differences in the numerical size-congruity effect

The numerical size-congruity effect is a classic phenomenon in numerical cognition in which people are slower to choose the physically larger of two presented numbers when the numbers are presented in a configuration where physical size is incongruent with their relative numerical magnitude (Henik & Tzelgov, 1982). For example, consider the stimuli in Figure 1. Compared to congruent trials, where the physically larger digit is also numerically larger (e.g., the left panel), people are slower to choose the physically larger digit on incongruent trials, where the physically larger digit is numerically smaller (e.g., the right panel). This simple phenomenon has been well studied in the fields of decision making and numerical cognition (Faulkenberry, Cruise, Lavro, & Shaki, 2016; Henik & Tzelgov, 1982; e.g., Paivio, 1975; Santens & Verguts, 2011; Schwarz & Heinze, 1998).

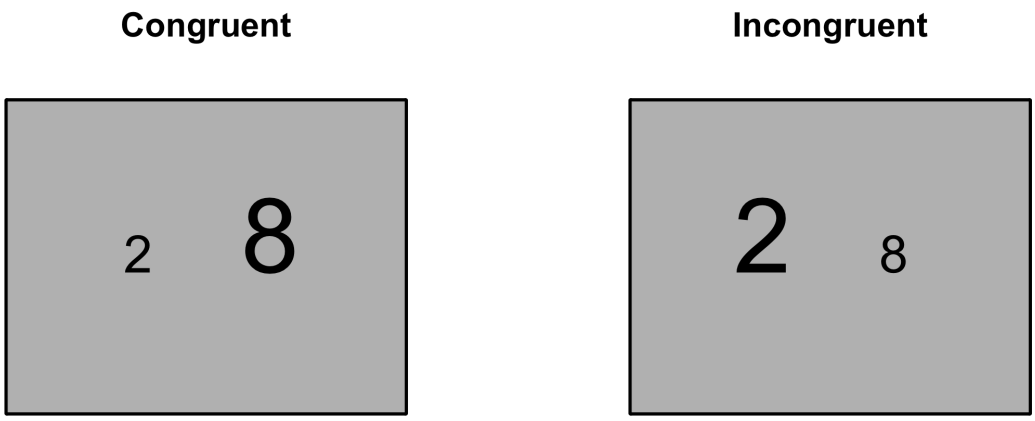

Figure 1. Example stimuli in a physical size comparison task. The left panel depicts a congruent trial, where the physically larger digit (8) is also the numerically larger digit. The right panel depicts an incongruent trial, where the physically larger digit (2) is the numerically smaller one.

The size-congruity effect is remarkable because it shouldn't have to occur. The task can be completed by simply monitoring the visual template and choosing the symbol which displaces the most visual area on the screen. If this is what people did, then there would be no interference from numerical magnitude on incongruent trials, and thus no size-congruity effect. The presence of the effect, then, implies that people are somehow unable to suppress this interference. As such, the size-congruity effect has historically been taken as an index for automatic processing of numerical magnitude in symbolic tasks (but see Fitousi, 2022, for a recent criticism of this view). Additionally, individual variation in the effect is often used as a tool to index various abilities in numerical cognition (Ashkenazi, Rubinsten, & Henik, 2009; Bugden & Ansari, 2011). Our goal in the present study is to develop and test models of various structures on these individual differences. Are there individual differences in the size-congruity effect? If so, what is their nature?

Modeling constraint on individual differences

Specifically, our aim is to model constraint on individual differences in the size congruity effect. We will argue throughout the paper that this is an important aim, but we think it is easy to demonstrate the need for such work with a quick "thought experiment". In a typical experiment, we usually take a sample mean as a central estimate of a population parameter. For context, let us consider an observed effect size (i.e., Cohen's d) as an index of the aggregate size-congruity effect obtained from a sample. We use this sample estimate d to predict the population-level "true" effect δ. Of course, this estimate comes with uncertainty – what exactly is the structure of true effects δ? Figure 2 displays two possibilities. In the left column, the distribution of true effects is normal, centered on some positive value of δ. In this model, most individuals' true effect δ will be close to this center ($\delta \approx 0.6$ in the figure), but some will be larger, and some will be smaller. Importantly, some will be negative. In the right column, the distribution of true effects is a *truncated* normal distribution with mean set equal to $\delta = 0.6$, as in the previous model. In this model, we will still see variation in

individuals' true effect δ, but the key difference is that *none* of the true effects will be negative. That is, this model *constrains* individuals' size-congruity effects to be positive.

So which model is correct? Unfortunately, we cannot determine this from an aggregate observed effect. To see why, consider the sampling distributions from each of these hypothetical populations, displayed in the bottom row of Figure 2. By the central limit theorem, both distributions of sampled effect sizes will be approximately normal and centered on the same value (e.g., $d = 0.6$). Unfortunately, when we carry out an experiment, we can only observe this sampled effect – the structure of the underlying true effects δ can only be inferred. Here, the mapping from population to sample is a 2-to-1 function – both structures map to the same distribution of sampled effect sizes. Hence, the inverse mapping is not well-defined – given an observed effect size d, we have no way to pinpoint the nature of the underlying population of true effects δ.

Though this question of structure may seem esoteric at first, it is of great importance for our understanding of the size-congruity effect. If the distribution of true effects δ is constrained (i.e., the right column of Figure 2), then the size-congruity effect is robust across individuals. Simply put, *everybody* exhibits it. In this situation, individual differences in the size-congruity effect are *quantitative* (Haaf & Rouder, 2017); that is, they are all positive and only vary in magnitude. If this is indeed the case, then the mechanisms underlying the size-congruity effect should be relatively straightforward to explain. On the other hand, if the distribution of true effects δ is unconstrained (i.e., the left column of Figure 2), then the size-congruity effect is more complex than originally thought and malleable (perhaps by many different possible mechanisms that vary by individual). Individual differences in this case are *qualitative*; that is, some people exhibit positive effects, whereas others exhibit negative effects.

Uncovering this latent structure of individual differences is not a trivial problem. Suppose we observe individuals with negative size-congruity effects, where responses are

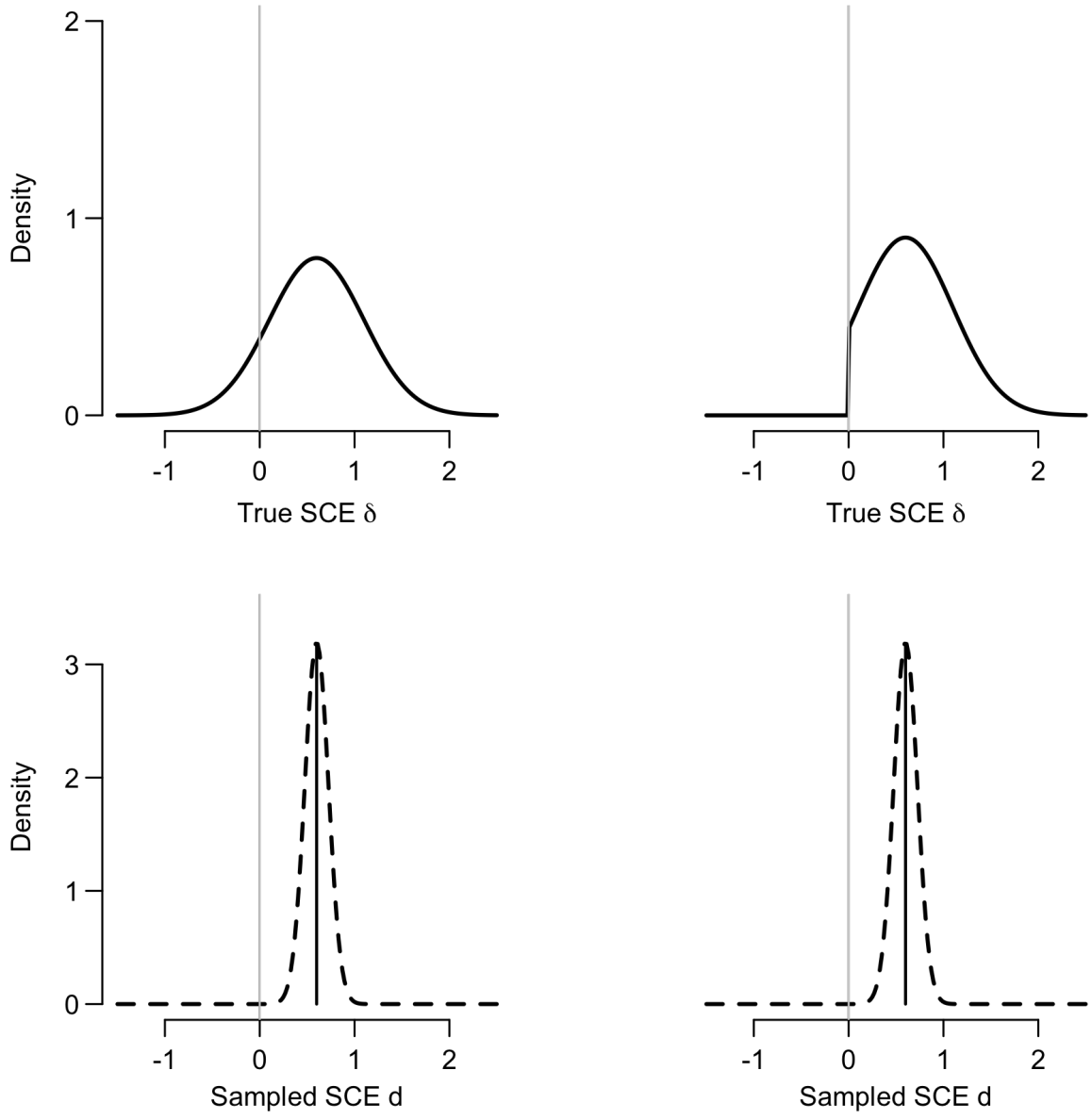

Figure 2. Possible scenarios for population-level size-congruity effects (SCE) and their respective sampling distributions. Figure adapted from Haaf and Rouder (2017).

faster when physical and numerical size are incongruent than when they are congruent. Does this suffice to conclude that individual differences in the size-congruity effect are qualitative? Not necessarily. Indeed, the observed negative effect may simply reflect sampling noise. The problem is even more difficult since there are at least two levels of sampling that occur here – at one level, we consider an individual as a random draw from the population of true effects, and at another level, the observed individual's aggregate effect size comes from a random draw of trials centered at that individual's true effect. Fortunately, Haaf and Rouder (2017) provided a methodological innovation that allows us to simultaneously model these multiple levels of individual variation. Their approach uses hierarchical Bayesian modeling to propose four different possibilities for how individuals' effect sizes may be constrained. These four proposed models are then compared using Bayesian model comparison, allowing us to index the evidence for each model from our observed data.

In the next section, we will describe these Bayesian mixed models in detail and more fully describe the methods used for model comparison. Then we will apply the model to three data sets from our lab. Two of these data sets have been previously published from our lab, and one is a new data set. Finally, we will answer the primary question of this paper: Are individual differences in the size-congruity effect constrained to be positive? That is, does everyone[1] exhibit the size-congruity effect? The answer turns out to be "yes".

Bayesian mixed models

We now describe our implementation of the Bayesian mixed model approach developed by Haaf and Rouder (2017). Before going into detail, the main aim of this approach is to build a (hypothetical) generative process for *each* observed response time (i.e., no aggregation of trials at the individual or group level). Each response time is assumed to be built from four components: (1) a grand mean; (2) a subject-specific adjustment to the

[1] Throughout the manuscript, we are assuming a population of people who are familiar with Arabic numerals and thus have a semantic association association with the underlying quantities that they represent.

grand mean (that is, a random intercept); (3) an subject-specific effect term; and (4) a noise term. The hierarchical model is then built by assuming each of these variable components (excluding the grand mean) is randomly drawn from some probability distribution which needs to be described. Later, we will pay specific attention to the distribution that generates each subject's effect term – it is this distribution on which we instantiate our models of individual difference structure.

Let Y_{ijk} denote the response time (in milliseconds) for the k^{th} replicate of the i^{th} subject in the j^{th} experimental condition ($j = 1, 2$). We place a random effects linear model on the vector of response times Y_{ijk}:

$$Y_{ijk} \sim \text{Normal}(\mu + \alpha_i + x_j \cdot \delta_i, \sigma^2).$$

Here, μ denotes the grand mean intercept and α_i represents the specific intercept adjustment for subject i. The term x_j is a binary variable coding the congruity condition of each trial in our experimental tasks; for congruent trials ($j = 1$), we set $x_1 = 0$, and for incongruent trials ($j = 2$), we set $x_2 = 1$. Under this specification, δ_i represents the size-congruity effect for subject i. Finally, σ^2 represents the latent sampling variance of the observed response times.

The critical mechanism of this modeling approach is that we need to propose a structure for the distribution from which each subject's size-congruity effect δ_i is randomly drawn. Our question about the latent structure of individual differences in the size-congruity effect then becomes one of *model selection* – which of these possible structure models best predicts our observed data? We define four possible generative models for these δ_i, each of which instantiate four possible theoretical positions about the distribution of size-congruity effects.

The unconstrained model

The unconstrained model, denoted $\mathcal{M}_u$, places no constraint on the individual size-congruity effects δ_i. We define this model as

$$\mathcal{M}_u : \delta_i \sim \text{Normal}(\nu, \eta^2),$$

where ν and η^2 represent the mean and variance, respectively, of the distribution of individual size-congruity effects δ_i. The values of ν and η^2 will be estimated from the observed data. In this model, we allow subjects' size-congruity effects to vary among all possible values (positive or negative), so we use this model to capture the framework of *qualitative* individual differences.

The positive-effects model

The positive-effects model places constraint on the distribution of size-congruity effects by assuming that all individual effects δ_i are positive. That is,

$$\mathcal{M}_+ : \delta_i \sim \text{Normal}_+(\nu, \eta^2),$$

where Normal_+ denotes a truncated normal distribution with lower bound 0 (i.e., upper right plot in Figure 2). This model assumes that there is variation in subjects' individual size-congruity effects, but since all $\delta_i > 0$, these effects are assumed to be in the same direction. As such, this model captures the framework of *quantitative* individual differences.

The common-effect model

The common-effect model places even more constraint on the distribution of size-congruity effects, as it specifically assumes that each individual has the *same* effect. That is,

$$\mathcal{M}_1 : \delta_i = \nu,$$

and thus any observed variation in the observed size-congruity effects would be due purely to sampling noise. We use this model for the same reason as Haaf and Rouder (2017). While

we think that it is highly unlikely that all individuals have the same size-congruity effect δ_i, it is important to propose such a model in order to benchmark our claim of individual differences. For example, if the common-effect model ended up being the best predictor of our observed data, we would need to question the efficiency of our experimental design as a test to elicit individual differences in the size-congruity effect.

The null model

Finally, the null model is the most constrained of the four, as it specifies that each subject's size-congruity effect is zero:

$$\mathcal{M}_0 : \delta_i = 0.$$

In this model, any observed variation in the observed response times would be due completely to sampling noise. Like the common-effect model, the null model also serves as a critical index of efficiency for our experimental design. Indeed, if the null model is the best predictor of our observed data, then we really need to question the efficiency of our experimental design to capture size-congruity effects of any sort.

Prior specifications

As our modeling is done within a Bayesian framework, we need to specify priors on the parameters in the model. Largely, we follow the recommendations of Haaf and Rouder (2017) to set these priors, but the parameters that vary across our models (i.e., δ_i, ν, η^2) deserve special attention. We use the g-prior approach (Rouder, Morey, Speckman, & Province, 2012; Zellner, 1986), which re-encodes these parameters in terms of *effect size*. To see how this works, consider the collection of individual effect parameters δ_i. We define $g_\delta = \eta^2/\sigma^2$, giving us a hyperparamter that casts the variability of δ_i in terms of the ratio of signal to noise – that is, true variability η^2 relative to sampling variability σ^2. This allows us to re-write our unconstrained model as

$$\mathcal{M}_u : \delta_i \sim \text{Normal}(\nu, g_\delta \sigma^2).$$

Similarly, we may scale the mean size-congruity effect ν in terms of sampling variability and get a new hyperparameter g_ν. Since we've introduced new (hyper)parameters into our model, we must place priors on them as well. The default method (Zellner, 1986) is to specify these priors as Inverse-χ^2 distributions with one degree of freedom and scale r^2.

Though the g-prior setup may appear complicated at first, it is actually quite convenient for us. By re-casting these critical parameters in terms of sampling variability σ^2, we convert the problem of specifying priors on δ_i, ν, and η^2 (which we think is generally difficult to do in any systematic manner) into one where we need only specify the expected variability of our size-congruity effects relative to the expected overall variability of the obsered response times. In general, we believe $\sigma = 300$ milliseconds is a reasonable prior expectation for the variability of response times in these size-congruity tasks (see also Luce, 1986).

So, after this setup, how do we actually set our priors? Let us first consider g_ν, the g-prior on the mean size-congruity effect. With the g-prior setup, we assume that $\nu \sim \text{Normal}(0, g_\nu\sigma^2)$, where $g_\nu \sim \text{Inverse-}\chi^2(r_\nu^2)$. The scale parameter r_ν should reflect our prior belief about the relative magnitude of size-congruity effects in numerical cognition. We think it is reasonable to expect such effects to be, on average, around 50 milliseconds, which is 1/6 of our expected overall trial-by-trial variability ($\sigma = 300$ milliseconds). Thus, we set $r_\nu = 1/6$.

Second, we consider g_δ. As we described above, this parameter specifies the *a priori* variability of individual size-congruity effects around the mean size-congruity effect. With the g-prior setup, we assume that $g_\delta \sim \text{Inverse-}\chi^2(r_\delta^2)$. We need to set the scale parameter r_δ – but, we have less guidance here. After all, estimating the variability of individual differences in the size-congruity effect is the point of this paper. However, like Haaf and Rouder (2017), we believe this variability should (1) be on the same scale as the expected effect, and (2) should be no larger than the expected effect. So, we are comfortable in specifying this

variability to be about 30 milliseconds, which is 1/10 of $\sigma = 300$ millseconds. For this reason, we set $r_\delta = 1/10$.

Model comparison

Since our goal is to capture the latent structure of individual differences in the size-congruity effect, our problem is first and foremost one of *model comparison*. That is, we ask which of the four competing models defined above is the most adequate as a predictor of our observed data? To answer this question, we use *Bayes factors* (Jeffreys, 1961; Kass & Raftery, 1995). Bayes factors index the relative predictive adequacy of two models by comparing the marginal likelihood of observed data under one model to another (Faulkenberry, Ly, & Wagenmakers, 2020). For example, a Bayes factor of 10 indicates that the observed data are 10 times more likely under one model compared to another. Techniques for computing Bayes factors among three of the four models above ($\mathcal{M}_u$, $\mathcal{M}_1$, $\mathcal{M}_0$) were previously developed by Rouder et al. (2012) and are implemented in the BayesFactor (Morey & Rouder, 2018) package in R (R Core Team, 2020). The Bayes factor between the constrained positive effects model $\mathcal{M}_+$ and the unconstrained model $\mathcal{M}_u$ is computed by the *encompassing prior* method (Faulkenberry, 2019; Klugkist, Kato, & Hoijtink, 2005), which is based on counting the number of posterior samples of $\mathcal{M}_u$ which obey the constraint placed by $\mathcal{M}_+$, then comparing this to the number of prior samples which obey the same constraint.

Description of data sets

We used the modeling approach described above to analyze three data sets from our lab. Two of the data sets (Data Sets 1 and 2) have already been reported in the literature (Bowman & Faulkenberry, 2020; Faulkenberry, Vick, & Bowman, 2018). Data Set 3 is an unpublished data set that has not previously been reported. With a few exceptions (noted below), all three data sets used the same experimental task (a physical comparison task), which we now describe in detail.

Participants

For all three data sets, participants were undergraduate psychology students who participated in exchange for partial course credit. Data Set 1 was generated by 23 subjects, Data Set 2 was generated by 53 subjects, and Data Set 3 was generated by 35 subjects.

Method and design

The task was implemented via the OpenSesame software package (Mathôt, Schreij, & Theeuwes, 2011). At the beginning of the task, subjects were told that they would be presented with pairs of numbers, with each number being displayed in a different font size. Furthermore, they were asked to quickly and accurately choose (via a keypress) the physically larger digit, pressing the "A" key if the number on the left was larger, and pressing the "L" key if the number on the right was larger. The number pairs were constructed from the single-digit Arabic numerals 2, 3, 4, 5, 6, 7, and 8. Pairs were chosen in order to balance the numerical distance between numerals (this is necessary because numerical distance modulates the magnitude of the size-congruity effect, Faulkenberry et al., 2016). Ignoring order, there were 12 possible pairs of numbers: 2-3, 3-4, 4-5 (distance 1); 2-4, 3-5, 4-6 (distance 2); 2-5, 3-6, 4-7 (distance 3); 2-6, 3-7, 4-8 (distance 4).

The size-congruity manipulation was implemented by varying the font size of each digit in the number pair. On each trial, the physically smaller digit was presented in 28 point font, whereas the physically larger digit was presented in 36 point font (these font sizes are identical to those originally used by Faulkenberry, Vick, and Bowman, 2018). This resulted in two different congruity conditions. In *congruent* trials, the numerically larger digit was also physically larger, and in *incongruent* trials, the numerically larger digit was physically smaller. Each pair was also presented in two different left-right orders and two different font configurations (i.e., configuration 1 = smaller font on left and larger font on right; configuration 2 = smaller font on right and larger font on left). In all, this resulted in $12 \times 2 \times 2 \times 2 = 96$ experimental trials per block. For Data Sets 1 and 2, subjects completed

four blocks of these 96 experimental trials (384 trials total). For Data Set 3, participants completed only two blocks, giving 192 trials total.

Each experimental trial began with a fixation cross displayed for 500 milliseconds, followed immediately by a number pair. One number was positioned 12.5 degrees to the left of the center of the screen, whereas the other number was positioned 12.5 degrees to the right of center (resulting in a visual angle between numbers of approximately 25 degrees). For each trial, the number pair remained on the screen until a response was made. If the response was correct, no feedback was given, and the next trial began immediately. If the response was incorrect, a red "X" was presented in the center of the screen for 1 second, after which the next trial began.

Results

We will now present the results of our modeling. For each data set, we break this analysis into three components:

1. *Aggregate size-congruity effect*: to test for an aggregate size-congruity effect, we performed a Bayesian paired-samples t-test (Rouder, Speckman, Sun, Morey, & Iverson, 2009) on the collection of mean response times, collapsed by subject and condition (congruent, incongruent). For each test, we compute a Bayes factor to compare the two models $\mathcal{H}_0 : \delta = 0$ and $\mathcal{H}_1 : \delta > 0$, where δ represents the population-level mean size-congruity effect. In the event that the data are evidential for $\mathcal{H}_1$ (which we fully expect), we give a 95% central credible interval for the raw effect (in ms).

2. *Model estimates*: using Markov chain Monte Carlo (MCMC) sampling functions from the BayesFactor package, we computed posterior distributions for all parameters in the unconstrained model. These posterior samples were then used to estimate size-congruity effects δ_i for each subject. Specifically, we computed the mean of the posterior samples as well as central 95% credible intervals for each subject, giving us

not only a central estimate of each subject's size-congruity effect, but also an interval containing, with probability 0.95, the true subject-level size-congruity effect δ_i.

3. *Model comparisons and sensitivity analysis*: using the methods described above, we computed Bayes factors among the four different models. We then assess the sensitivity of our model comparisons to our prior specifications (i.e., the *a priori* scale on the mean size-congruity effect ν and effect variability η^2) by recomputing the Bayes factors for other reasonable settings of r_ν and r_δ.

Note that all modeling is done on correct responses only. This resulted in removing 433 of 8832 trials in Data Set 1 (an error rate of 4.90%), 636 of 20352 trials in Data Set 2 (an error rate of 3.12%), and 160 of 6720 trials in Data Set 3 (an error rate of 2.38%)

Data Set 1

Aggregate size-congruity effect. Subjects in Data Set 1 exhibited an aggregate size-congruity effect of 64.82 ms (see Figure 3). Mean response times were faster on congruent trials ($M = 571$ ms) compared to incongruent trials ($M = 638$ ms), $\text{BF}_{10} = 520599$, 95% CrI = [46.66 ms, 82.16 ms].

Model estimates. Individual size-congruity effect estimates from the unconstrained model are displayed in the left column of Figure 4. We first notice that the *observed* size-congruity effects for each subject (denoted by black crosses) span from -1.30 ms to 176.61 ms. Here, observed effects are computed by subtracting each subject's mean response time for congruent trials from the mean response time for incongruent trials. With the exception of one subject, the observed size-congruity effects are all positive. Estimates from the hierarchical Bayesian model are displayed as black dots with shaded 95% credible intervals. These estimates are computed as means of the posterior samples for each δ_i, and the 95% credible intervals are computed as the central 95% of the posterior samples (i.e., ranging between the 2.5% and 97.5% quantiles of the samples). The black dashed line

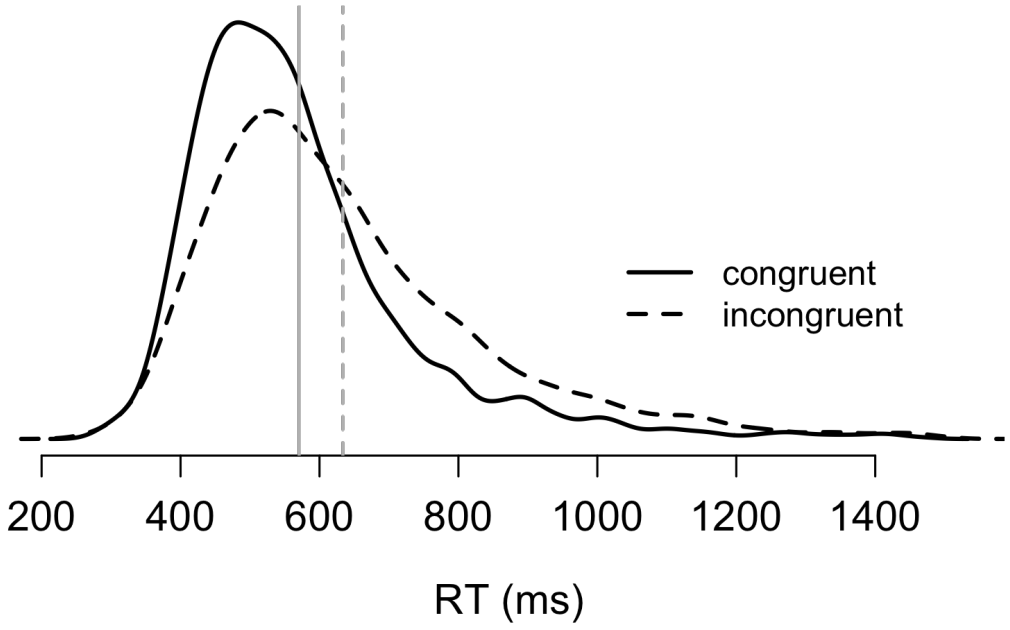

Figure 3. Density plot for observed response times in Data Set 1 split by congruity condition (congruent versus incongruent). Solid line represents congruent trials, and dashed line represents incongruent trials. The gray vertical lines (solid and dashed) represent mean response times for congruent and incongruent trials, respectively.

represents an (posterior) estimated mean size-congruity effect of $\nu = 63$ ms.

As is usual in hierarchical modeling, we see a fair amount of *shrinkage* (or regularization, see Davis-Stober, Dana, & Rouder, 2018) in our estimates. This refers to the phenomenon where the estimated effects (the black dots) extend over a smaller range (12.44 ms to 121.43 ms) than the observed effects (the black crosses; -1.30 ms to 176.61 ms). This shrinkage reflects how the hierarchical model accounts for sampling variability at all levels, thus providing effect estimates with smaller range after accounting for this noise.

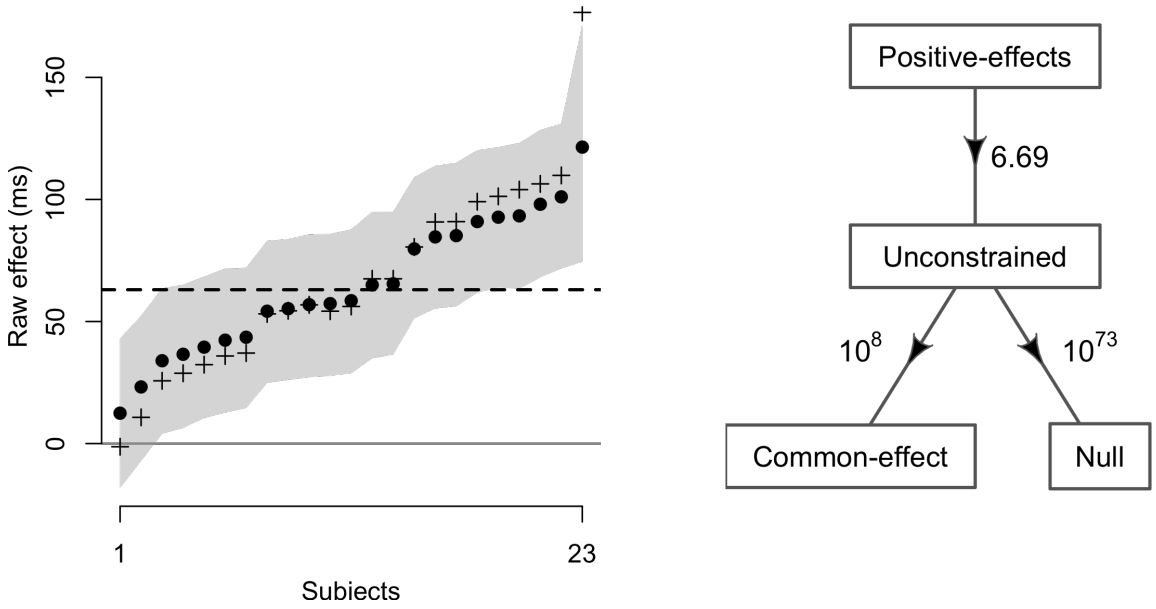

Figure 4. Individual size-congruity effect estimates (left column) and Bayes factor model comparisons (right column) for Data Set 1. Posterior means and 95% credible intervals for δ_i are represented by black dots and gray band, respectively. The + symbols represent the observed size-congruity effect for each subject. The dashed-line represents the estimated mean size-congruity effect ν. For the model comparisons, Bayes factors are displayed beside each arrow.

Model comparison and sensitivity analysis. In the previous section, we saw that all model estimates for the individual size-congruity effects δ_i were positive. This apparent pattern is supported by our Bayes factor computations (see right column of Figure 4). The observed data were 6.69 times more likely under the positive-effects model $\mathcal{M}_+$ than under the unconstrained model $\mathcal{M}_u$. If we assume 1-to-1 prior odds for $\mathcal{M}_+$ and $\mathcal{M}_u$, this means that our posterior odds in favor of $\mathcal{M}_+$ have increased to 6.69-to-1, which is equivalent to a posterior probability of $p(\mathcal{M}_+ \mid \text{data}) = 0.87$. These models were massively preferred over the common-effect model $\mathcal{M}_1$ and the null model $\mathcal{M}_0$, as $\mathcal{M}_u$ was more likely to have predicted the observed data by factors of 10^8-to-1 and 10^{73}-to-1, respectively. In all, these data provide positive evidence for *quantitative* individual differences in the size-congruity

Table 1

Sensitivity of Bayes factors to prior settings for Data Set 1

Scale on ν	Scale on δ_i	Null	Common-effect	Positive-effects	Unconstrained
$\frac{1}{6}$ (50 ms)	$\frac{1}{10}$ (30 ms)	6.53e −75	2.62e −9	*	0.15
$\frac{1}{12}$ (25 ms)	$\frac{1}{20}$ (15 ms)	1.94e −74	4.32e −9	*	0.15
$\frac{1}{12}$ (25 ms)	$\frac{1}{5}$ (60 ms)	2.38e −75	0.54e −9	*	0.04
$\frac{1}{3}$ (100 ms)	$\frac{1}{20}$ (15 ms)	1.99e −74	1.05e −8	*	0.35
$\frac{1}{3}$ (100 ms)	$\frac{1}{5}$ (60 ms)	3.27e −75	1.77e −9	*	0.14

Note. The first row contains the Bayes factors from the original prior settings. The asterisks mark the preferred model for each prior setting, and Bayes factors are computed against this model.

effect. That is, everybody exhibits a positive size-congruity effect.

To assess the robustness of this claim, we performed a sensitivity analysis. Here, we recomputed Bayes factors with some other reasonable choices of prior scale on mean size-congruity effect ν and effect variability η^2. Specifically, we halved and doubled each of the original prior scales $r_\nu = 1/6$ and $r_\delta = 1/10$, giving four combinations of prior settings. The resulting Bayes factors are displayed in Table 1.

Across this range of prior scales, we see the same outcome as our original analysis. The positive-effects model $\mathcal{M}_+$ is always preferred over the unconstrained model $\mathcal{M}_u$, and these models are collectively very much preferred over the common-effect and null models. At a minimum, the degree of preference for $\mathcal{M}_+$ over $\mathcal{M}_u$ is 2.86-to-1 for $r_\nu = 1/3$ and $r_\delta = 1/20$. At its maximum, the odds ratio is 24.55-to-1, and the other two prior settings result in Bayes factors that are approximately equal to the original.

Data Set 2

Aggregate size-congruity effect. Subjects in Data Set 2 exhibited an aggregate size-congruity effect of 59.68 ms (see Figure 5). Mean response times were faster on congruent trials ($M = 550$ ms) compared to incongruent trials ($M = 611$ ms), $BF_{10} = 159118497976284384$, 95% CrI = [50.97 ms, 67.80 ms].

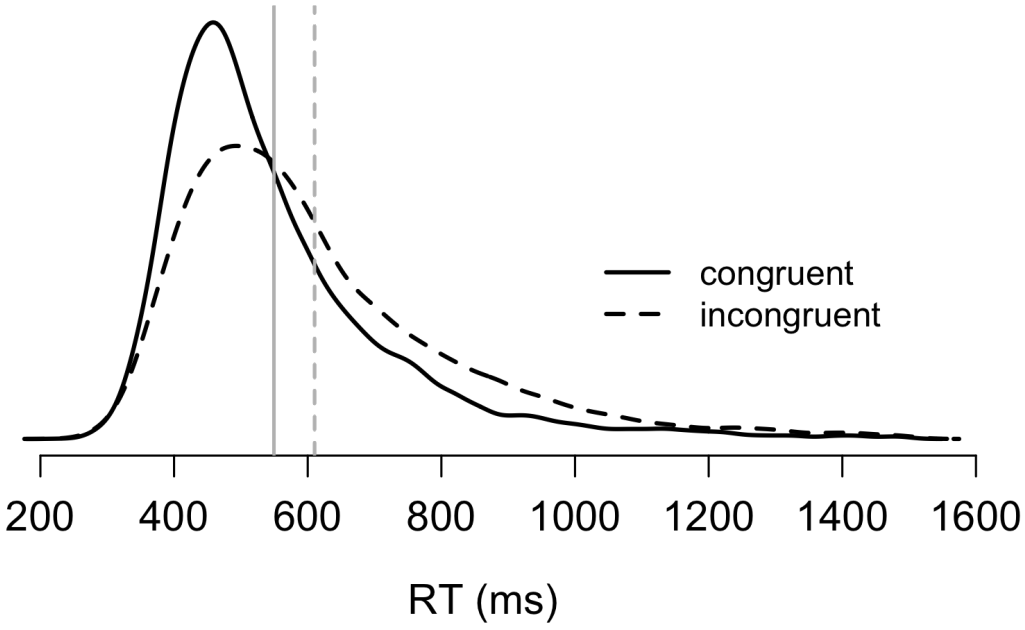

Figure 5. Density plot for observed response times in Data Set 2 split by congruity condition (congruent versus incongruent). Solid line represents congruent trials, and dashed line represents incongruent trials. The gray vertical lines (solid and dashed) represent mean response times for congruent and incongruent trials, respectively.

Model estimates. Individual size-congruity effect estimates from the unconstrained model are displayed in the left column of Figure 6. The *observed* size-congruity effects for each subject (denoted by black crosses) spanned from -14.59 ms to 142.10 ms. As in Data Set 1, all but one of the observed size-congruity effects were positive. The hierarchical model

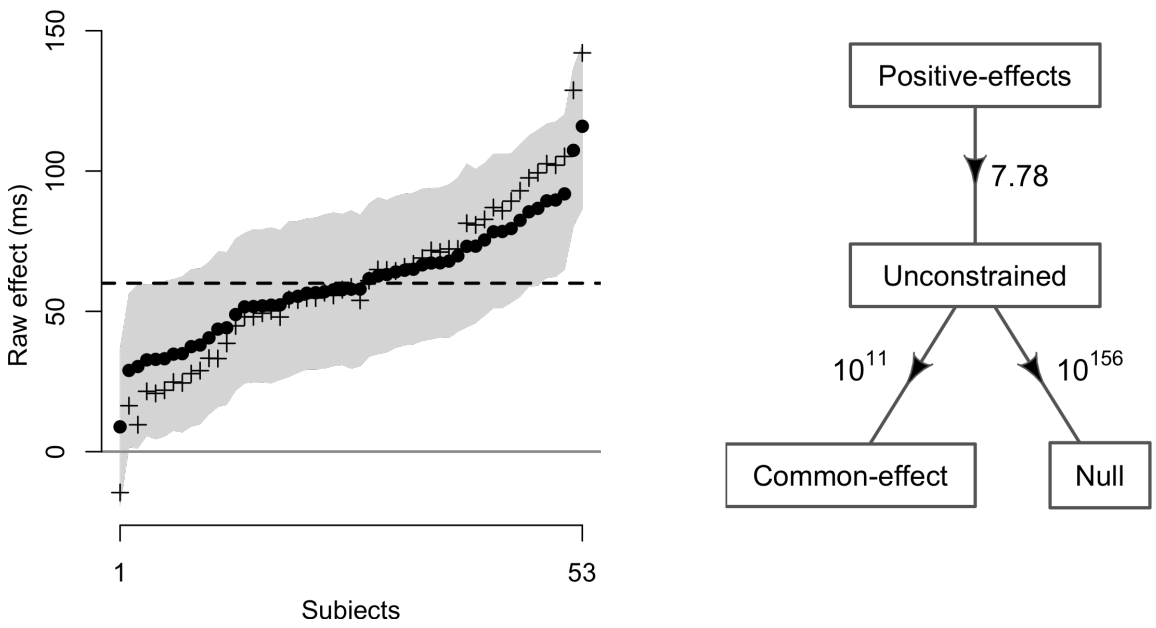

Figure 6. Individual size-congruity effect estimates (left column) and Bayes factor model comparisons (right column) for Data Set 2. Posterior means and 95% credible intervals for δ_i are represented by black dots and gray band, respectively. The + symbols represent the observed size-congruity effect for each subject. The dashed-line represents the estimated mean size-congruity effect ν. For the model comparisons, Bayes factors are displayed beside each arrow.

estimates for δ_i also followed a similar pattern to Data Set 1 and exhibited similar shrinkage. The estimated effects (the black dots) extended over a smaller range (8.84 ms to 115.94 ms) than the observed effects (the black crosses; -14.59 ms to 142.10 ms). Finally, the posterior estimated mean size-congruity effect was $\nu = 60$ ms.

Model comparison and sensitivity analysis. The apparent positive constraint displayed in the model estimated effects δ_i was confirmed in our model comparison (see right column of Figure 6). The observed data were 7.78 times more likely under the positive-effects model $\mathcal{M}_+$ than under the unconstrained model $\mathcal{M}_u$. Assuming 1-to-1 prior odds for $\mathcal{M}_+$ and $\mathcal{M}_u$, the observed data increased the posterior odds in favor of $\mathcal{M}_+$ to 7.78-to-1, which is equivalent to a posterior probability of $p(\mathcal{M}_+ \mid \text{data}) = 0.89$. Both

Table 2

Sensitivity of Bayes factors to prior settings for Data Set 2

Scale on ν	Scale on δ_i	Null	Common-effect	Positive-effects	Unconstrained
$\frac{1}{6}$ (50 ms)	$\frac{1}{10}$ (30 ms)	2.66e −157	4.29e −13	*	0.13
$\frac{1}{12}$ (25 ms)	$\frac{1}{20}$ (15 ms)	8.01e −157	7.51e −13	*	0.13
$\frac{1}{12}$ (25 ms)	$\frac{1}{5}$ (60 ms)	1.39e −157	1.27e −13	*	0.04
$\frac{1}{3}$ (100 ms)	$\frac{1}{20}$ (15 ms)	8.72e −157	1.87e −12	*	0.34
$\frac{1}{3}$ (100 ms)	$\frac{1}{5}$ (60 ms)	2e −157	4.37e −13	*	0.14

Note. The first row contains the Bayes factors from the original prior settings. The asterisks mark the preferred model for each prior setting, and Bayes factors are computed against this model.

models accounted for nearly all posterior model probability, as $\mathcal{M}_u$ was more likely to have predicted the observed data over the common-effect model and the null model by factors of 10^{11}-to-1 and 10^{156}-to-1, respectively. In all, Data Set 2 provided more positive evidence for *quantitative* individual differences in the size-congruity effect.

As with Data Set 1, we performed a similar sensitivity analysis. The resulting Bayes factors are displayed in Table 2. Across the range of reasonable prior scales for the mean effect ν and variability of the δ_i, we see that the positive-effects model $\mathcal{M}_+$ is always preferred over the unconstrained model $\mathcal{M}_u$. At a minimum, the degree of preference for $\mathcal{M}_+$ over $\mathcal{M}_u$ is 2.94-to-1 for $r_\nu = 1/3$ and $r_\delta = 1/20$. At its maximum, the odds ratio is 25.28-to-1, and the other two prior settings result in Bayes factors that are approximately equal to the original.

Data Set 3

Aggregate size-congruity effect. Subjects in Data Set 3 exhibited an aggregate size-congruity effect of 51.57 ms (see Figure 7). Mean response times were faster on

congruent trials ($M = 618$ ms) compared to incongruent trials ($M = 671$ ms), $BF_{10} = 97372213$, 95% CrI = [38.99 ms, 63.81 ms].

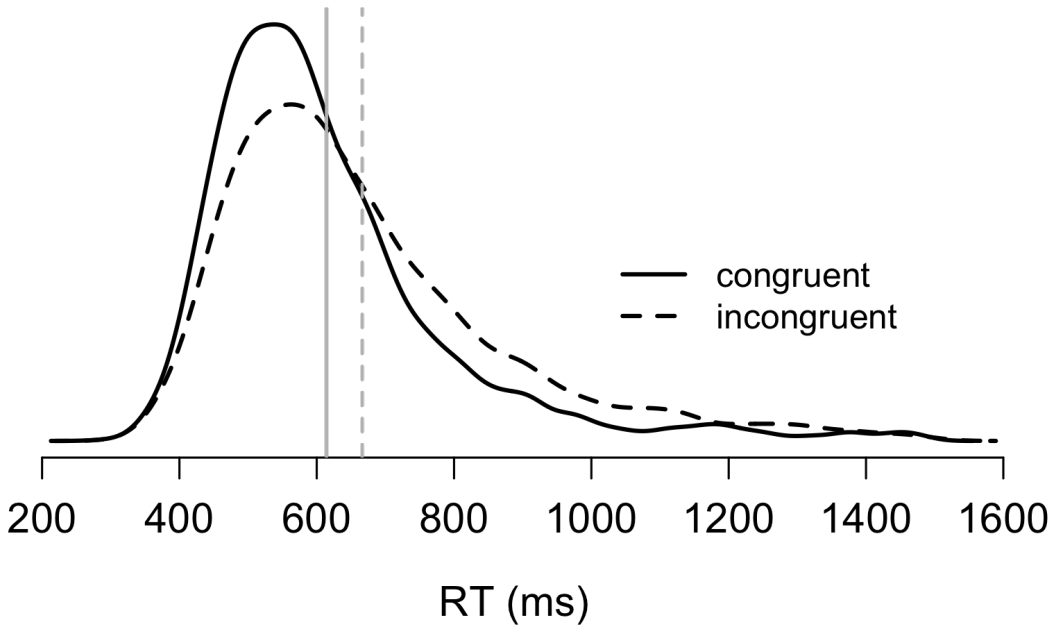

Figure 7. Density plot for observed response times in Data Set 3 split by congruity condition (congruent versus incongruent). Solid line represents congruent trials, and dashed line represents incongruent trials. The gray vertical lines (solid and dashed) represent mean response times for congruent and incongruent trials, respectively.

Model estimates. Individual size-congruity effect estimates from the unconstrained model are displayed in the left column of Figure 8. The *observed* size-congruity effects for each subject spanned from -15.85 ms to 152.78 ms. We see from Figure 8 that three observed effect estimates were negative, but the rest were positive. The hierarchical model estimates for δ_i also followed a similar pattern to Data Sets 1 and 2 and exhibited similar shrinkage. The estimated effects again extended over a smaller range (16.12 ms to 102.75 ms) than the observed effects (-15.85 ms to 152.78 ms). Finally, the posterior estimated mean

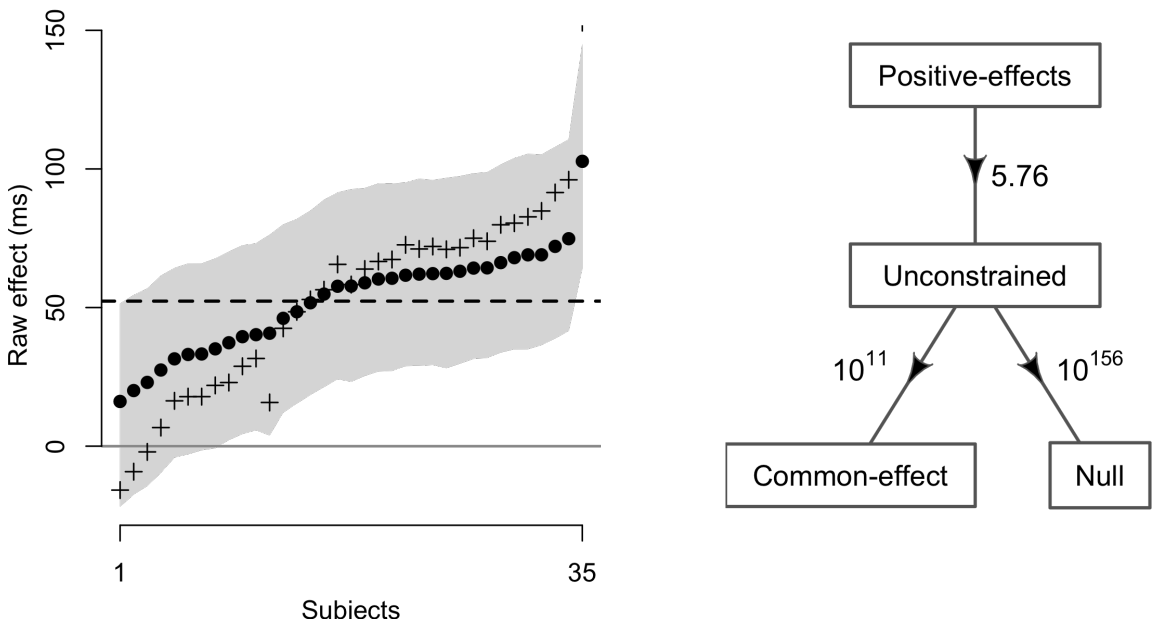

Figure 8. Individual size-congruity effect estimates (left column) and Bayes factor model comparisons (right column) for Data Set 3. Posterior means and 95% credible intervals for δ_i are represented by black dots and gray band, respectively. The + symbols represent the observed size-congruity effect for each subject. The dashed-line represents the estimated mean size-congruity effect ν. For the model comparisons, Bayes factors are displayed beside each arrow.

size-congruity effect was $\nu = 52$ ms.

Model comparison and sensitivity analysis. Though we did observe three negative size-congruity effects, the model-estimated effects were all positive. This positive constraint on the δ_i was confirmed in our model comparison (see right column of Figure 8). The observed data were 5.76 times more likely under the positive-effects model $\mathcal{M}_+$ than under the unconstrained model $\mathcal{M}_u$. Assuming 1-to-1 prior odds for $\mathcal{M}_+$ and $\mathcal{M}_u$, the observed data increased the posterior odds in favor of $\mathcal{M}_+$ to 5.76-to-1, which is equivalent to a posterior probability of $p(\mathcal{M}_+ \mid \text{data}) = 0.85$. As with both Data Sets 1 and 2, the positive-effects and unconstrained models were far preferred to either the common-effect or null models. In all, Data Set 3 provided even more positive evidence for *quantitative*

Table 3

Sensitivity of Bayes factors to prior settings for Data Set 3

Scale on ν	Scale on δ_i	Null	Common-effect	Positive-effects	Unconstrained
$\frac{1}{6}$ (50 ms)	$\frac{1}{10}$ (30 ms)	1.64e −40	3.58e −4	*	0.17
$\frac{1}{12}$ (25 ms)	$\frac{1}{20}$ (15 ms)	4.63e −40	6.06e −4	*	0.17
$\frac{1}{12}$ (25 ms)	$\frac{1}{5}$ (60 ms)	7.63e −41	9.9e −5	*	0.05
$\frac{1}{3}$ (100 ms)	$\frac{1}{20}$ (15 ms)	5.1e −40	1.38e −3	*	0.44
$\frac{1}{3}$ (100 ms)	$\frac{1}{5}$ (60 ms)	1.47e −40	3.99e −4	*	0.21

Note. The first row contains the Bayes factors from the original prior settings. The asterisks mark the preferred model for each prior setting, and Bayes factors are computed against this model.

individual differences in the size-congruity effect.

Finally, we performed the same sensitivity analysis we did with Data Sets 1 and 2. The resulting Bayes factors are displayed in Table 3. Across the range of reasonable prior scales, we see that the positive-effects model $\mathcal{M}_+$ is always preferred over the unconstrained model $\mathcal{M}_u$. At a minimum, the degree of preference for $\mathcal{M}_+$ over $\mathcal{M}_u$ is 2.28-to-1 for $r_\nu = 1/3$ and $r_\delta = 1/20$. At its maximum, the Bayes factor is 19.04-to-1, and the other two prior settings result in Bayes factors that are approximately equal to the original.

Model mis-specification?

One criticism of the Haaf and Rouder (2017) method is the assumption that the observed response times are drawn from a normal distribution. Such criticism is particularly salient here, as response times generally exhibit a distinct positive skew. This bears out with the three data sets tested in this paper, as can easily be seen in Figures 3, 5, and 7. While there are many methods for modeling response times with skewed distributions (i.e.,

ex-Gaussian, inverse Gaussian / Wald, etc.), the Haaf and Rouder (2017) implementation does not include these distributions. One simple approach that might prove to be easily implemented is to assume that the observed response times follow a (shifted) lognormal distribution; then, we may simply transform the observed response times by first shifting by a fixed amount to remove the leading edge of the distribution (e.g., 200 milliseconds) and then taking the (natural) logarithm of the shifted RTs. The resulting distribution (now on the log scale) is then approximately normal and may be "fed into" the Haaf and Rouder (2017) method with little difficulty.

To assess the impact of the default normal specification on our modeling outcomes, we re-analyzed Data Set 1 using the aforementioned shifted log transform. To do this, we first shifted the distribution of RTs by subtracting 200 ms from each observed response time. This serves to remove the "leading edge" of the RT distribution. Then, we took the natural logarithm of the shifted RTs; the result of this log transformation on the distribution can be seen in Figure 9.

We then applied our hierarchical modeling workflow to these log transformed response times, the results of which can be seen in Figure 10. The overall similarity of these results with the original analysis (Figure 4) for Data Set 1 is striking. We see very similar patterns of observed effects, estimated effects, and shrinkage. For the log transformed data, we see a posterior estimated common effect (black dashed line) of $\nu = 0.14$. If we back-transform this back to the original response time scale (by exponentiating this effect estimate as $\exp(\nu)$), we get an estimated common effect of 1.16. Because the data are on a logarithmic scale, this effect is multiplicative, so an estimated effect of 1.16 is a 16% increase in response times. For these data, this is roughly equivalent to a response time increase of 89 ms.

The similarity persists with the Bayes factor comparisons. In the right column of Figure 10) we can see the observed data were 6.65 times more likely under the positive-effects model $\mathcal{M}_+$ than under the unconstrained model $\mathcal{M}_u$. Further, these models

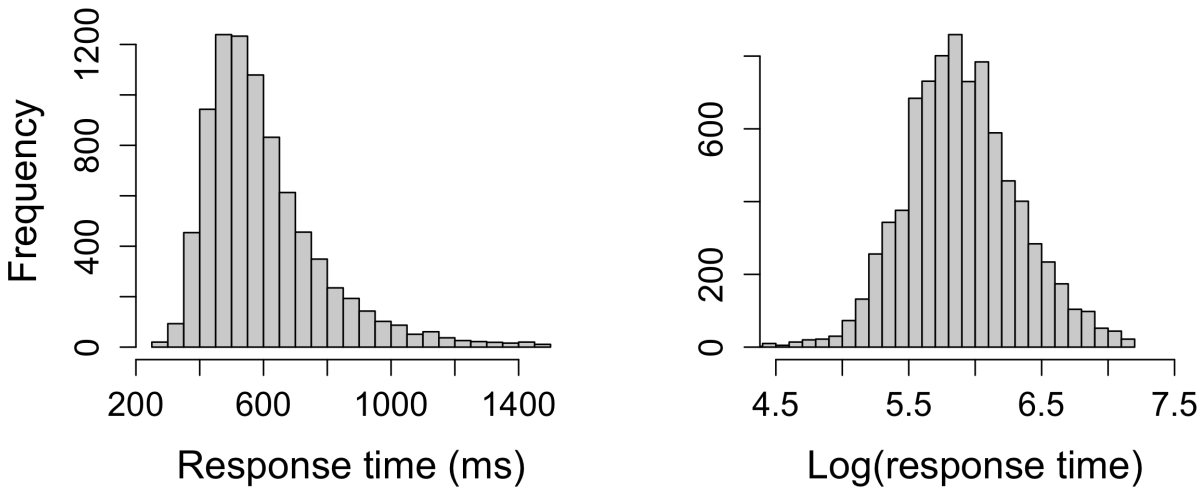

Figure 9. Distributions of observed response times in Data Set 1. The left panel displays the original observed response times, whereas the right panel displays the shifted-log-transformed response times.

were again overhelmingly preferred over the common-effect model $\mathcal{M}_1$ and the null model $\mathcal{M}_0$. In all, the inferences we obtain from using a shifted lognormal model on observed response times is very similar to that when we use the default normal specifications recommended by Haaf and Rouder (2017). In both cases, the positive effects model is preferred over the unconstrained model. Note that we see a similar outcome with Data Sets 2 and 3, but these analyses are omitted here. The interested reader can download our RMarkdown file at https://git.io/vAEE8 and do these analyses for themselves. As a result, we are confident that applying the default normal specification for response times is sufficient to model individual difference structures from these size congruity data.

Discussion

The purpose of the present study was to uncover the latent structure of individual differences in the numerical size-congruity effect. We did this by applying the techniques of

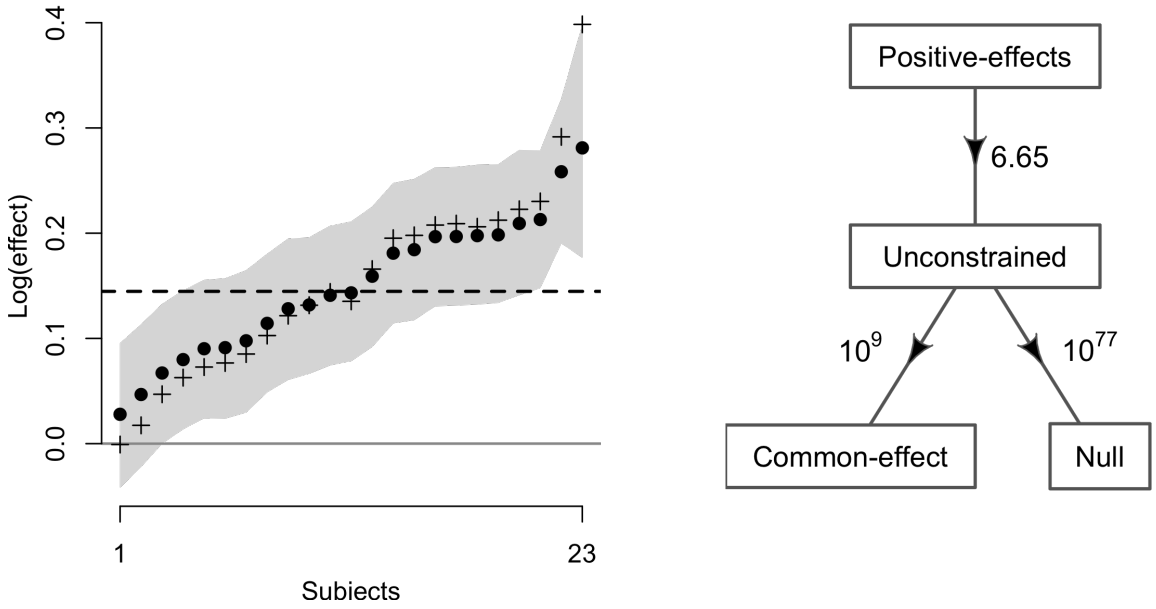

Figure 10. Individual effect estimates (left column) and Bayes factor model comparisons (right column) for Data Set 1 under a lognormal specification. Posterior means and 95% credible intervals for δ_i are represented by black dots and gray band, respectively. The + symbols represent the observed size-congruity effect for each subject. The dashed-line represents the estimated mean size-congruity effect ν. For the model comparisons, Bayes factors are displayed beside each arrow.

Haaf and Rouder (2017) to build a hierarchical Bayesian model of response times in a physical comparison task. We built four competing models which differed in the amount of constraint that was placed on the population-level size-congruity effect: (1) an unconstrained model, where the size-congruity effect was allowed to be any magnitude, either positive or negative (i.e., qualitative individual differences); (2) a positive-effects model, where the size-congruity effect was constrained to be positive (i.e., quantitative individual differences); (3) a common-effect model, where the size-congruity effect was assumed to be the same value for all individuals; and (4) a null model, which assumed that all variability in response times was reflective purely of sampling noise. We used Bayesian model comparison to adjudicate the four models against three different data sets. Across all three data sets, the

positive-effects model was the best predictor of the observed data. This model constrains the population-level size-congruity effect to be positive. This lends support to a "quantitative individual differences" model, where everyone exhibits a positive size-congruity effect.

Such results may seem counterintuitive; after all, in all three data sets, we indeed observed negative size-congruity effects in a small number of subjects. Shouldn't this serve as a counterexample to any claims of positive constraint on the size-congruity effect? We argue that it does not. Indeed, as explained above, any observed size-congruity effects consist also of sampling noise inherent in their measurements, so simply relying on the mean difference between congruent and incongruent trials is going to exaggerate the range of true effects. By building in this sampling noise as part of the hierarchical model specification, we effectively "subtract out" this noise from our model-estimated true effects. This shrinkage, or regularization, is a natural consequence of hierarchical modeling. After accounting for this sampling noise in our data sets, all subjects' model-estimated size-congruity effects were positive.

So why does this matter? We think our work adds value in two primary ways. The first is the practical application of our methods to measure the size-congruity effect. Researchers in numerical cognition often use the size-congruity effect as a marker of some aspect of numerical processing. Girelli, Lucangeli, and Butterworth (2000) found the size-congruity effect emerges gradually as numerical skill progresses. Similarly, Rubinsten, Henik, Berger, and Shahar-Shalev (2002) showed that the size-congruity effect appears later than other markers of numerical processing (e.g., the numerical distance effect, Moyer & Landauer, 1967). Mussolin, Mejias, & Noël (2010) demonstrated that children with developmental dyscalculia exhibit a reduced size-congruity effect compared to non-dyscalculic controls. All of these studies have a common need to measure the size-congruity effect in individuals. As discussed earlier, typical measurements of the size-congruity effect (e.g., difference in mean response times) are embedded with

measurement error in the form of sampling noise. The hierarchical Bayesian model of Haaf and Rouder (2017) that we applied is well-suited to estimating these effects while removing this sampling noise. To our knowledge, this method has been previously applied in a numerical cognition context only once (Vogel, Faulkenberry, & Grabner, 2021). Given that the size-congruity effect (and other performance measures in numerical tasks) are often used as predictors of various mathematical abilities, the measurement fidelity of the effect estimates that come from the hierarchical Bayesian approach are highly desirable.

The second way in which our work adds value is more theoretical. In much of our earlier work, we have sought to uncover the mechanisms that are responsible for the size-congruity effect. To date, there has been much conflicting evidence concerning the locus of this interference. Santens and Verguts (2011) proposed that the size congruity effect stems from either early representational overlap (an *early interaction account*) or a *late interaction* account reflecting response competition. A variety of different techniques have been used to test between these competing accounts, including electrophysiological techniques (Schwarz & Heinze, 1998; Szűcs & Soltész, 2007, 2008), neuroimaging (Cohen Kadosh et al., 2007), computer mouse tracking (Faulkenberry et al., 2016), visual search (Krause, Bekkering, Pratt, & Lindemann, 2017; Sobel, Puri, & Faulkenberry, 2016; Sobel, Puri, Faulkenberry, & Dague, 2017), and mathematical modeling (Bowman & Faulkenberry, 2020; Faulkenberry et al., 2018). However, consistent conclusions have remained elusive; some studies support the early interaction account, whereas others support the late interaction account. We think the present work may help to provide some insight into this debate. Indeed, our finding that everyone exhibits a positive size-congruity effect places some much-needed constraint onto theories of the size-congruity effect. In our recent mathematical modeling work (Bowman & Faulkenberry, 2020; Faulkenberry et al., 2018), we have found that the size-congruity effect occurs in the decision components of the elapsed response and not in the nondecision components (i.e., initial representation formation). As such, the positive constraint on the effect would implicate a common mechanism of response competition across all individuals,

supporting a late interaction account.

Before moving further with some limitations to our current work, it is worth pointing out that support for the positive-effects model does not necessarily imply that the size-congruity effect is automatic. Indeed, in their original paper, Haaf and Rouder (2017) posited that a positive-effects model could be used to say that an effect "is automatic and beyond strategic control" (p. 780). But they also described other interpretations of a positive-effects model, including correlation between stimulus strength and some neurological parameter. This must be underscored, because it would be entirely easy to conclude from the present work that the size-congruity effect reflects an automatic representation of number magnitude, a viewpoint which is commonly accepted in numerical cognition (e.g., Henik & Tzelgov, 1982; Dehaene, 1997). However, recent work by Fitousi and colleagues (e.g., Fitousi & Algom, 2006, 2018; Fitousi, 2022) has demonstrated that the size-congruity effect does not arise because of the automatic activation of numerical magnitude, but rather because of attentional contributions (e.g., Risko, Maloney, & Fugelsang, 2013). Thus, we must be careful when interpreting the present support of the positive-effects model. Rather than accepting these results as wholesale confirmation of an automaticity theory in a "knee-jerk" fashion, we simply take them as another step in the increasing mathematical specification of theories of the size-congruity effect. We believe strongly that all competing accounts of the size-congruity effect would benefit from this additional mathematical specification. Overall, we think this hierarchical Bayesian modeling approach will be an exciting technique for future developments on the architecture of the size-congruity effect, as well as other effects in numerical cognition.

One thing that we did not consider for this study is the possibility that "some do, some don't" (e.g., Haaf & Rouder, 2018). Here, we would need to consider another model on top of the four proposed models in this paper – namely, a model where some have a positive effect, yet others have a null effect. This is called a *spike-and-slab* mixture model, and Haaf

& Rouder (2018) found that it was the most predictive model for the location Stroop effect (Pratte, Rouder, Morey, & Feng, 2010). It could be the case that a spike-and-slab model might be useful for modeling the size-congruity effect, especially given past work that some subjects with developmental dyscalculia do not seem to exhibit a size-congruity effect (e.g., Rubinsten et al., 2002). However, we note that the three datasets here likely do not follow a spike-and-slab model of individual differences[2]. The first reason is if the latent size congruity effect distribution truly follows a spike-and-slab model, then there should some subjects who truly exhibit no size congruity effect (i.e., the spike). In such a case, these subjects should have an estimated effect close to 0 (see figure 5c in Haaf & Rouder, 2018). We do not see this in any of our datasets; indeed, even the smallest individual effect estimates are well off the floor in Figures 4, 6, and 8. The second reason that we do not believe a spike-and-slab model would not be relevant here is that the shape of the individual effect estimate curves (the black dots in Figures 4, 6, and 8) are somewhat S-shaped, which is perfectly reasonable if the distribution of effects is indeed a normal distribution. The shape of the effect estimate curve from a spike-and-slab model would not exhibit such a shape, as the left side of the curve would be mostly flat [reflecting the estimates around 0; again, see Figure 5c of Haaf & Rouder (2018)].

Another potential limitation concerns the model assumption that response times are normally distributed. It is well known that response time distributions are positively skewed rather than symmetric (Luce, 1986). Because of this skew in the underlying distribution, one might argue that individual effect estimates may be skewed as well. The argument works as follows. Because response times are necessarily positive, the impact of any manipulation which would potentially decrease response times is bounded so that the negative effect can only be so large (i.e., the response times can only decrease so much). On the other hand, any manipulation which *increases* response times (i.e., positive effects) is theoretically

---

[2] We thank Julia Haaf for suggesting this argument.

unbounded in its impact; individual size congruity effects can be of any magnitude. Thus, the magnitudes present in the distribution of positive effects would outweigh those from negative effects, thus potentially inflating the magnitude of the estimated common effect. However, as we demonstrated above, this argument does not bear out with the present data. Certainly, the pattern of inference we saw when applying a shifted lognormal model (see also Faulkenberry, 2022) to the observed response times did not differ from the inference we obtained from applying the original "default" recommendations of Haaf and Rouder (2017). Moreover, the effect estimate obtained from the normal specification was less than that obtained from the shifted lognormal specification; this contradicts the prediction that such an effect estimate would be inflated. One one hand, it is certainly principled to use a skewed distribution at the first level of the model – for example, the aforementioned lognormal model (Rouder, Province, Morey, Gomez, & Heathcote, 2014) or a shifted Wald model (Anders, Alario, & Maanen, 2016; Faulkenberry, 2017). However, we note that Haaf and Rouder (2018) have addressed this concern in detail, arguing that the normal model provides computational convenience that we feel outweighs the minimal (if any) penalties that we realize by assuming a normal specification at the first level.

Overall, we argue that the size-congruity effect is something that everyone exhibits (at least in the context of a physical comparison task). The methods we employed and results are immediately useful in two main ways: (1) the methods provide a way for researchers to compute better, noise-free estimates of individual size-congruity effects (and more broadly, any type of interference effect in numerical cognition); and (2) the positive-effects model of the latent structure the size-congruity effect constrains future theory building about the size-congruity effect, which can possibly lead to new insights about the architecture of numerical cognition.

Data Availability Statement

The data that support the findings of this study are openly available for download from Github. Datasets 1 and 3 can be downloaded from https://github.com/tomfaulkenberry/physNumComparisonTask, and Dataset 2 can be downloaded from https://github.com/Kbow27/Thesis.

References

Anders, R., Alario, F.-X., & Maanen, L. V. (2016). The shifted Wald distribution for response time data analysis. *Psychological Methods*, *21*(3), 309–327. https://doi.org/10.1037/met0000066

Ashkenazi, S., Rubinsten, O., & Henik, A. (2009). Attention, automaticity, and developmental dyscalculia. *Neuropsychology*, *23*(4), 535. https://doi.org/10.1037/a0015347

Bowman, K., & Faulkenberry, T. J. (2020). Modeling response times in the size-congruity effect: Early versus late interaction. *PsyArXiV*. https://doi.org/10.31234/osf.io/dns4t

Bugden, S., & Ansari, D. (2011). Individual differences in children's mathematical competence are related to the intentional but not automatic processing of arabic numerals. *Cognition*, *118*(1), 32–44. https://doi.org/10.1016/j.cognition.2010.09.005

Cohen Kadosh, R., Cohen Kadosh, K., Linden, D. E. J., Gevers, W., Berger, A., & Henik, A. (2007). The brain locus of interaction between number and size: A combined functional magnetic resonance imaging and event-related potential study. *Journal of Cognitive Neuroscience*, *19*(6), 957–970. https://doi.org/10.1162/jocn.2007.19.6.957

Davis-Stober, C. P., Dana, J., & Rouder, J. N. (2018). Estimation accuracy in the psychological sciences. *PLOS ONE*, *13*(11), e0207239. https://doi.org/10.1371/journal.pone.0207239

Dehaene, S. (1997). *The number sense: How the mind creates mathematics*. New York: Oxford University Press.

Faulkenberry, T. J. (2017). A single-boundary accumulator model of response times in an addition verification task. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.01225

Faulkenberry, T. J. (2019). A tutorial on generalizing the default Bayesian t-test via posterior sampling and encompassing priors. *Communications for Statistical Applications and Methods*, *26*(2), 217–238. https://doi.org/10.29220/csam.2019.26.2.217

Faulkenberry, T. J. (2022). A note on the normality assumption for Bayesian models of constraint in behavioral individual differences. *arXiv*. Retrieved from https://arxiv.org/abs/2112.05503

Faulkenberry, T. J., Cruise, A., Lavro, D., & Shaki, S. (2016). Response trajectories capture the continuous dynamics of the size congruity effect. *Acta Psychologica*, *163*, 114–123. https://doi.org/10.1016/j.actpsy.2015.11.010

Faulkenberry, T. J., Ly, A., & Wagenmakers, E.-J. (2020). Bayesian inference in numerical cognition: A tutorial using JASP. *Journal of Numerical Cognition*, *6*(2), 231–259. https://doi.org/10.5964/jnc.v6i2.288

Faulkenberry, T. J., Vick, A. D., & Bowman, K. A. (2018). A shifted Wald decomposition of the numerical size-congruity effect: Support for a late interaction account. *Polish Psychological Bulletin*, 391–397. https://doi.org/10.24425/119507

Fitousi, D. (2022). Conjoint measurement of physical size and numerical magnitude: Numerals do not automatically activate their semantic meaning. *Psychonomic Bulletin and Review*, *29*, 134–144. https://doi.org/10.3758/s13423-021-01990-1

Fitousi, D., & Algom, D. (2006). Size congruity effects with two-digit numbers: Expanding the number line? *Memory & Cognition*, *34*(2), 445–457. https://doi.org/10.3758/BF03193421

Fitousi, D., & Algom, D. (2018). A system factorial technology analysis of the size congruity effect: Implications for numerical cognition and stochastic modeling. *Journal of Mathematical Psychology*, *84*, 57–73. https://doi.org/10.1016/j.jmp.2018.03.006

Girelli, L., Lucangeli, D., & Butterworth, B. (2000). The development of automaticity

in accessing number magnitude. *Journal of Experimental Child Psychology*, *76*(2), 104–122. https://doi.org/10.1006/jecp.2000.2564

Haaf, J. M., & Rouder, J. N. (2017). Developing constraint in Bayesian mixed models. *Psychological Methods*, *22*(4), 779–798. https://doi.org/10.1037/met0000156

Haaf, J. M., & Rouder, J. N. (2018). Some do and some don't? Accounting for variability of individual difference structures. *Psychonomic Bulletin & Review*, *26*(3), 772–789. https://doi.org/10.3758/s13423-018-1522-x

Henik, A., & Tzelgov, J. (1982). Is three greater than five: The relation between physical and semantic size in comparison tasks. *Memory & Cognition*, *10*(4), 389–395. https://doi.org/10.3758/bf03202431

Jeffreys, H. (1961). *The Theory of Probability (3rd ed.)*. Oxford, UK: Oxford University Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773. https://doi.org/10.2307/2291091

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

Krause, F., Bekkering, H., Pratt, J., & Lindemann, O. (2017). Interaction between numbers and size during visual search. *Psychological Research*, *81*(3), 664–677. https://doi.org/10.1007/s00426-016-0771-4

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. New York: Oxford University Press.

Mathôt, S., Schreij, D., & Theeuwes, J. (2011). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*(2), 314–324. https://doi.org/10.3758/s13428-011-0168-7

Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of Bayes factors for common designs*. Retrieved from

https://CRAN.R-project.org/package=BayesFactor

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgements of numerical inequality. *Nature*, *215*(5109), 1519–1520. https://doi.org/10.1038/2151519a0

Mussolin, C., Mejias, S., & Noël, M.-P. (2010). Symbolic and nonsymbolic number comparison in children with and without dyscalculia. *Cognition*, *115*(1), 10–25. https://doi.org/10.1016/j.cognition.2009.10.006

Paivio, A. (1975). Perceptual comparisons through the mind's eye. *Memory & Cognition*, *3*(6), 635–647. https://doi.org/10.3758/bf03198229

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

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*(5), 356–374. https://doi.org/10.1016/j.jmp.2012.08.001

Rouder, J. N., Province, J. M., Morey, R. D., Gomez, P., & Heathcote, A. (2014). The lognormal race: A cognitive-process model of choice and latency with desirable psychometric properties. *Psychometrika*, *80*(2), 491–513. https://doi.org/10.1007/s11336-013-9396-3

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 & Review*, *16*(2), 225–237. https://doi.org/10.3758/pbr.16.2.225

Rubinsten, O., Henik, A., Berger, A., & Shahar-Shalev, S. (2002). The development of internal representations of magnitude and their association with arabic

numerals. *Journal of Experimental Child Psychology*, *81*(1), 74–92. https://doi.org/10.1006/jecp.2001.2645

Santens, S., & Verguts, T. (2011). The size congruity effect: Is bigger always more? *Cognition*, *118*(1), 94–110. https://doi.org/10.1016/j.cognition.2010.10.014

Schwarz, W., & Heinze, H. J. (1998). On the interaction of numerical and size information in digit comparison: A behavioral and event-related potential study. *Neuropsychologia*, *36*(11), 1167–1179. https://doi.org/10.1016/s0028-3932(98)00001-3

Sobel, K. V., Puri, A. M., & Faulkenberry, T. J. (2016). Bottom-up and top-down attentional contributions to the size congruity effect. *Attention, Perception, & Psychophysics*, *78*(5), 1324–1336. https://doi.org/10.3758/s13414-016-1098-3

Sobel, K. V., Puri, A. M., Faulkenberry, T. J., & Dague, T. D. (2017). Visual search for conjunctions of physical and numerical size shows that they are processed independently. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(3), 444–453. https://doi.org/10.1037/xhp0000323

Szűcs, D., & Soltész, F. (2007). Event-related potentials dissociate facilitation and interference effects in the numerical Stroop paradigm. *Neuropsychologia*, *45*(14), 3190–3202. https://doi.org/10.1016/j.neuropsychologia.2007.06.013

Szűcs, D., & Soltész, F. (2008). The interaction of task-relevant and task-irrelevant stimulus features in the number/size congruency paradigm: An ERP study. *Brain Research*, *1190*, 143–158. https://doi.org/10.1016/j.brainres.2007.11.010

Vogel, S. E., Faulkenberry, T. J., & Grabner, R. H. (2021). Quantitative and qualitative differences in the canonical and the reverse distance effect and their selective association with arithmetic and mathematical competencies. *Frontiers in Education*, *6*. https://doi.org/10.3389/feduc.2021.655747

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In P. K. Goel & A. Zellner (Eds.), *Bayesian inference*

and decision techniques: Essays in honor of Bruno de Finetti (pp. 233–243). Elsevier.