# A Note on Using Systems of Orders to Capture Theoretical Constraint in Psychological Science

Julia M. Haaf[1], Fayette Klaassen[2], & Jeffrey N. Rouder[3]

[1] University of Missouri
[2] Utrecht University
[3] University of California, Irvine

Version 3, 3/2018

Most theories in the social sciences are verbal and provide ordinal-level predictions for data. For example, a theory might predict that performance is better in one condition than another, but not by how much. One way of gaining additional specificity is to posit multiple ordinal constraints simultaneously. For example a theory might predict an effect in one condition, a larger effect in another, and none in a third. We call such simultaneous constraints a 'system of orders' and show how common theoretical positions lead naturally to system-of-order predictions. We adopt a Bayesian model comparison approach to assess evidence for multiple, simultaneous order constraints, a difficult endeavor in a frequentist framework. The result is a statistical system custom tuned for the way social scientists conceptualize theory that is more intuitive and informative than current linear-model approaches.

*Keywords:* Theory specification, Order constrained inference, Bayesian Inference

At the core of science is the ability to use data to inform theoretical positions. In psychological science, these theoretical positions are often stated verbally and often lead to ordinal predictions for data. Consider the proposition that reading is fast, obligatory, and automatic (Kahneman, 1973). One ordinal implication of this theoretical statement is the usual Stroop effect (Stroop, 1935) where the colors of color terms (e.g. "red") are identified more quickly when they are congruent with the word identity (the word "red" written in red) than when they are incongruent with the identity (the word "green" written in red).

Psychologists often assess these ordinal predictions with a *t*-test, and the hypotheses underlying a one-sided *t*-test are direct tests of ordinal relations. For example, a *t*-test can be used to state whether there is a Stroop effect (congruent identified on average more quickly than incongruent) or not. One problem with the ordinal approach, however, is what we may call *intellectual inefficiency*. By positing coarse verbal theory that provides for only modest constraints on the data, we are neither risking nor learning much from the data.

An alternative approach is to focus on more complex models that seemingly provide greater constraints on data. Some of these models go by *psychological process models*, and they describe in more detail the processes and representations used in cognition. Examples of process models are sequential sampling models and race models (Lee & Wagenmakers, 2013; Lewandowsky & Farrell, 2011; Logan, 1988; Smith, 2000; Townsend & Ashby, 1983). Some researchers may argue that process models make metric predictions, and therefore provide for a more efficient analysis of data. We tend to disagree, however. While not a hard-and-fast rule, we find that metric predictions from psychological process models often reflect free parameters rather than deep structural commitments.

Take, for example, Cohen, Dunbar, & McClelland's (1990) neural network model of the Stroop effect. The metric value of the size of the effect reflects scaling parameters that convert network cycles to a time scale. The same type of free-parameter explanation holds say for a diffusion model account of the flanker task (Pe, Vandekerckhove, & Kuppens, 2013). In fact, we know of no theory that predicts the size of the Stroop effect be it 4 milliseconds or 4 seconds before the data are collected. In our view, even for complex models, there are few if any *a priori* metric predictions about psychological phenomena.

The question, then, is how shall psychological scientists gain more constraint in predictions? We advocate that, instead of focusing on the prediction of the size of an effect, researchers should honor the verbal theoretical tradition by

Correspondence concerning this article should be addressed to Julia M. Haaf, 210 McAlester Hall, Columbia, MO, USA, 65203. E-mail: jhaaf@mail.missouri.edu

focusing on many ordinal constraints simultaneously. When many ordinal constraints are considered simultaneously, we may call them a *system of order constraints*. Here are two examples of such systems:

### Example 1: Symbolic Distance

Moyer and Landauer (1967) developed a task where the observer decides if a presented digit is greater than five or less than five. The presented digit is either 2, 3, 4, 6, 7, or 8. Participants are highly accurate, and the dependent variable of interest is the time to make the comparison. The main question of interest is how does the time vary as a function of the digit, and in particular, does this time increase or decrease as the digit is further from five. Consider the following three theories, corresponding to three different sets of simultaneous orders:

*Analog-Representation Theory* posits that numbers are stored in an analog system as a uni-dimensional quantity much like length (Gallistel & Gelman, 1992). Just as comparing similar lengths is slower than disparate ones, this theory predicts that $\mu_4 > \mu_3 > \mu_2$ and $\mu_6 > \mu_7 > \mu_8$, where $\mu_i$ is the true mean response time for the $i$th digit. This ordering is shown in Figure 1A. The symbol ">" represents the relation "is faster than."

*Propositional Representation Theory* posits that numbers are represented as semantic propositions, much like in a computer. Accordingly, there should be no effect for distance-from-five.This prediction leads to the equalities $\mu_2 = \mu_3 = \mu_4$ and $\mu_6 = \mu_7 = \mu_8$. We avoid specifying all six condition means are equal in case there is a speed difference across the "less than" and "greater than" response. These equalities are represented in Figure 1B. Unconnected nodes have no relation; so even though the node "3" is above node "6", there is no implied ordering because there is no line connecting the nodes.

*Priming + Spreading Activation Theory* posits that familiar or anticipated items are responded to more quickly, that is, they are primed. Because participants need to keep the value of 5 in mind as part of the task demands, similar numbers are primed and responded to more quickly. Assuming a semantic spreading activation network (Collins & Loftus, 1975), numbers closer to five receive greater priming. Hence, this theory predicts $\mu_2 > \mu_3 > \mu_4$ and $\mu_8 > \mu_7 > \mu_6$. This ordering is shown in Figure 1C; it is the reverse of that in Figure 1A.

Note that the orders are weak (they may include equalities) and partial (not all nodes have to be connected with each other node). Figure 1D further illustrates this property. Here there are six cells, and there is an ordering where Cell 1 is greater than Cells 2, 3, 4, 5, which in turn are greater than Cell 6. There are equalities, Cells 2 and 3 are equal as are Cells 4 and 5. And there is partiality—there are no relations between select cells such as Cells 2 and 4. Partial weak orders are sufficiently general to include no order constraints, as shown in Figure 1E. *A system of order constraints is the conjunction and disjunction of several partial weak orders.*

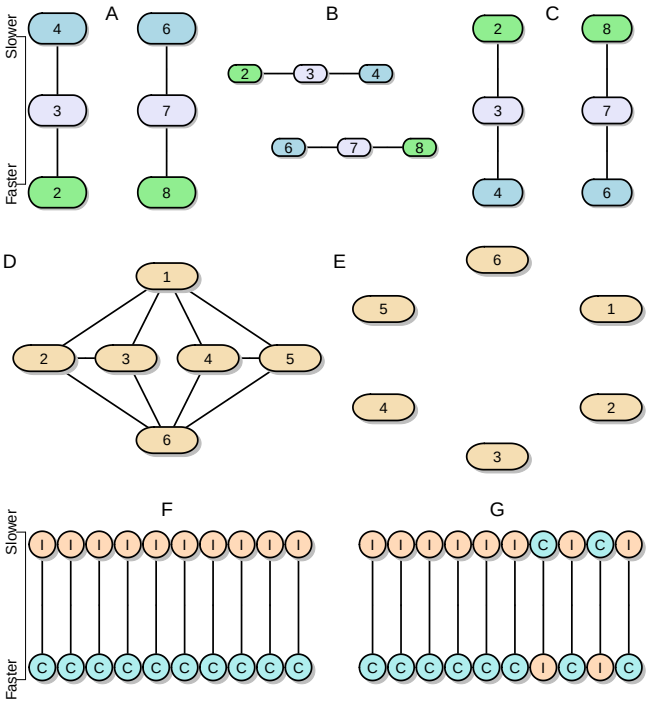


*Figure 1.* Examples of Orders. **A.** Analog representation in the symbolic distance task implies response times decrease with distance from five. **B.** Propositional representation implies no response-time effects of distance. **C.** The priming account implies response times increase with distance from five. **D.** Example of a partial weak order for six cells. **E.** A partial order may be unconstrained. **F.** Order constraints where all people show a true Stroop effect. **G.** An example of an order where some people have truly reversed Stroop effects (incongruent colors are identified faster than congruent ones).

### Example 2: Does Everybody?

In the above example, the focus is on the ordering of true population means, say whether the true mean response time for one condition is faster, slower, or the same as for another. Yet, the true mean across a population is a fairly removed abstraction, especially for the study of psychology. As an alternative we may ask about ordinal constraints that are preserved across all participants. We call this the *does everybody* question. Does everybody plausibly Stroop in the same direction, or are there some people who truly name incongruent colors faster than congruent ones? Does everybody plausibly detect loud tones more quickly than soft ones, or are there some people who truly detect soft tones more quickly than loud ones? Do all people plausibly throw a ball further with their right hand, or are there some people who truly throw

a ball further with their left?

It is important to note that the focus is on true effects rather than sample scores. Even when the true effects of everybody in a population are the same, we may observe some people who have scores that reverse the phenomena-of-interest due to sample noise. The posed questions are about what happens after sample noise is accounted for. Is it plausible that all individuals have true effects in the same direction, or, alternatively, is there evidence that some have true effects in the reverse direction.

If all people show an effect in the same direction, we may consider this behavior as lawful, automatic, perhaps biological, and largely outside of human variation. If not, say if some people have effects in one direction and others have them reversed, we would examine theories with multiple processes, perhaps with some volitional control. Moreover, the next set of questions would be to see whether the direction of effects is correlated with other characteristics, skills, or abilities.

This question can be addressed by aggregating multiple individual analyses (Klaassen, Zedelius, Veling, Aarts, & Hoijtink, 2017). Alternatively, Haaf and Rouder (2017) propose a hierarchical approach. The authors asked whether all readers truly identify congruent colors more quickly than incongruent ones, or alternatively, if there exist some readers who are Stroop reversed where they truly identify incongruent colors faster than congruent ones. Our setup may be expressed by the system in Figure 1F. The "everybody Stroops" model is shown as a consistency of ordering across all individuals; the negation is that some people, or at least one person, truly have reversed Stroop effects, and an example of this negation is shown in Figure 1G.

## Statistical Development

Encoding theories as systems of ordinal constraints raises a set of statistical considerations—how may evidence from data for competing systems be assessed. This assessment not only requires stating positive evidence for ordinal and inequality constraints, but doing so across many of them simultaneously in both conjunctive and disjunctive relations. To the best of our knowledge, a general frequentist solution is not known (cf. Robertson, Wright, & Dykstra, 1988; Silvapulle & Sen, 2011). Fortunately, Bayesian inference with inequality constraints is conceptually straightforward (Gelfand, Smith, & Lee, 1992). Here we follow the work of Klugkist and Hoijtink and colleagues who instantiate collections of equality and order constraints on linear model parameters (Klugkist & Hoijtink, 2007; Klugkist, Laudy, & Hoijtink, 2005; Mulder, Klugkist, Schoot, Meeus, & Hoijtink, 2009). Comparing systems in the Klugkist and Hoijtink framework becomes a matter of comparing the relative strength of evidence from the data for competing models. These relative strengths may be quantified with Bayes factors (Jeffreys, 1961), which themselves are the direct consequence of updating beliefs about models with Bayes' rule. Hoijtink, Klugkist, and Boelen (2008) provides a computationally convenient approach, further elaborated in Hoijtink (2012). We use the work of Haaf and Rouder (2017) to extend this approach for "does everyone" questions.

## Models for Orders: The Symbolic-Distance Effect Example

We demonstrate Bayesian inference on systems of ordinal constraints for the symbolic-distance effect example. Rouder, Lu, Speckman, Sun, and Jiang (2005) ran a standard symbolic distance experiment to assess how response times change with symbolic distance. Here, we consider the three order systems in Figure 1A-C: The Analog-Representation theory, the Propositional-Representation theory, and the Priming + Spreading-Activation theory. Moreover, we consider whether all participants have the same ordering. Let $Y_{ijk}$ denote the response time for the $i$th participant in the $j$th digit condition ($j = 2, 3, 4, 6, 7, 8$), and for the $k$th replicate:

$$Y_{ijk}|\nu_{ij}, \sigma^2 \overset{ind}{\sim} \text{Normal}(\nu_{ij}, \sigma^2),$$

where $\nu_{ij}$ is the $i$th person's true mean response time for the $j$th condition, and $\sigma^2$ is the trial-by-trial variation. To represent the theories it is useful to reparameterize the above model into relative differences between condition means for each individual. The following appears complicated, but it is just a matter of defining the relevant contrasts within a linear model using dummy variables. Let $\Delta_{im}$ be $m$th relative difference for the $i$th person. There are four differences implied by the theories defined as follows: $\Delta_{i1} = \nu_{i3} - \nu_{i2}$, $\Delta_{i2} = \nu_{i4} - \nu_{i3}$, $\Delta_{i3} = \nu_{i6} - \nu_{i7}$, and $\Delta_{i4} = \nu_{i7} - \nu_{i8}$. With these differences, the cell means $\nu_{ij}$ are given as

$$\nu_{ij} = \left(\nu_{i3} - x_{2j}\Delta_{i1} + x_{4j}\Delta_{i2}\right)^{s_j} + \left(\nu_{i7} - x_{8j}\Delta_{i4} + x_{6j}\Delta_{i3}\right)^{1-s_j},$$

where $s_j = 1$ if the digit condition $j < 5$, and $s_j = 0$ otherwise; and $x_{j'j} = 1$ if $j'$ indicates the current digit condition, $j$.

With this reparameterization models may be placed on the relative differences, $\Delta_{im}$. These differences are defined so that they are a subtraction of a digit further from 5 from a digit that is closer to 5. Hence, positive values of $\Delta_{im}$ are consistent with the Analog-Representation theory where response times are larger for digits closer to 5.

The theories then correspond to the following constraints: 1. The Analog-Representation theory holds for the $i$th individual if $\Delta_{im} > 0$ for each $m$. 2. The Priming + Spreading-Activation theory holds for the $i$th individual if $\Delta_{im} < 0$ for each $m$. 3. The Propositional-Representation model holds for the $i$th individual if $\Delta_{im} = 0$ for each $m$.

In the next step, we write these constraints as formalized models on the collection of individuals' relative differences.

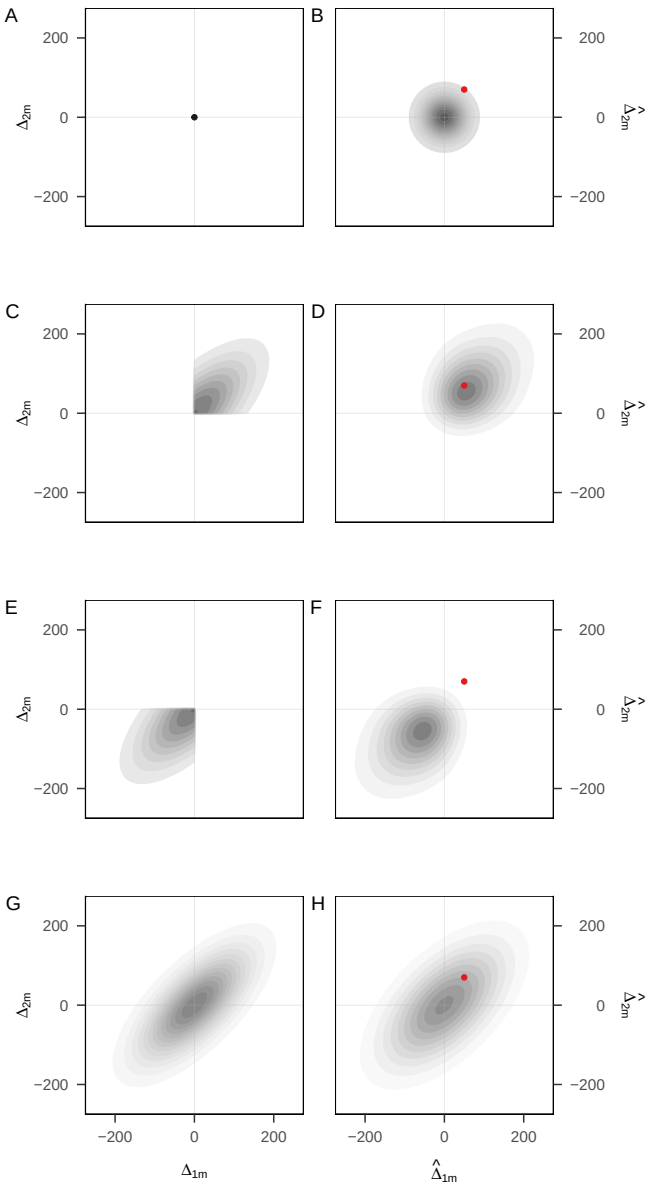


*Figure 2*. Models (left) and predictions (right) for the symbolic distance effect. Darker areas represent higher plausibility of $\Delta_{im}$ before the data are collected. Models are conditional on set values of $\mu_m = 0$ms and $\eta = 90$ms. Predictions take into account sampling noise and the correlation reflects the priors placed on $\mu_m$ and $\eta^2$. The red point represents a hypothetical observed data point for two individuals. The hypothetical data point is best predicted by the Analog-Representaion model.

Model $\mathcal{M}_0$ instantiates the statement that everybody uses propositional representation. It is given by

$$\mathcal{M}_0 : \quad \Delta_{im} = 0.$$

Figure 2A shows a graphical depiction of the Propositional-Representation model. The $x$-axis shows the true effect for one participant for the $m$th difference, $\Delta_{1m}$; the $y$-axis shows the true effect for a second participant, $\Delta_{2m}$. The only point with mass is $(0, 0)$ showing that each participants' relative differences between digit conditions must be identically zero.

Model $\mathcal{M}_+$ instantiates the statement that everybody uses analog representation, and it is given by:

$$\mathcal{M}_+ : \quad \Delta_{im}|\mu_m, \eta^2 \sim \text{Normal}_+(\mu_m, \eta^2),$$

where $\text{Normal}_+$ is a normal distribution truncated from below at zero, $\mu_m$ is the population mean of the $m$th digit condition difference, and $\eta^2$ is the variability of individuals around this mean. Figure 2C depicts this model for two participants (for set values of $\mu_m$ and $\eta^2$). For both participants, only positive values have mass.

Model $\mathcal{M}_-$ instantiates the statement that everybody uses priming + spreading activation, and it is given by:

$$\mathcal{M}_- : \quad \Delta_{im}|\mu_m, \eta^2 \sim \text{Normal}_-(\mu_m, \eta^2),$$

where $\text{Normal}_-$ is a normal truncated from above at zero. Figure 2E depicts this model for two participants. For both participants only negative values have mass.

Of course, it may be that not everyone uses the same number representation. We decided to implement a "none-of-the-above" model by placing no ordinal constraints on the relative differences. This model is termed the unconstrained model and is denoted $\mathcal{M}_u$:

$$\mathcal{M}_u : \quad \Delta_{im}|\mu_m, \eta^2 \sim \text{Normal}(\mu_m, \eta^2).$$

If this model is strongly preferred, the interpretation is that none of the everybody-does models are appropriate. Consequently, we may conclude different people use different representations. Figure 2G illustrates this model for two participants. Here, the participants' effects are not constrained to any of the quadrants.

Prior specifications are needed for $\sigma^2$, $\mu_m$, $\eta^2$, $\nu_3$ and $\nu_7$. We take a $g$-prior approach as discussed in Haaf and Rouder (2017), and the prior specifications for this application are provided in the Appendix.

In Figure 2, we show the model specification for any two participants on any one difference contrast (for any one value of $m$). For the symbolic-distance experiment by Rouder et al. (2005), however, four difference contrasts are specified, and 52 individuals participated in the experiment. The figure therefore understates the dramatic differences in constraint between the models. The constraints provided by the models must hold across all people and across all difference contrasts simultaneously.

## Bayesian Model Comparison

We use Bayes factors to assess the relative evidence for the above models. The Bayes factor for any two models,

Models $\mathcal{M}_a$ and $\mathcal{M}_b$, is given by:

$$B_{ab} = \frac{P(\boldsymbol{Y}|\mathcal{M}_a)}{P(\boldsymbol{Y}|\mathcal{M}_b)},$$

where $\boldsymbol{Y}$ is the collection of all observations. The numerator and denominator are marginal probabilities of the data under the respective models, and these quantities are best thought of as the model predictions of data. The right column of Figure 2 shows the predictions on observed relative differences for the four models applied here. One aspect of these predictions that is not obvious is the correlation across participants. This correlation comes from the hierarchical structure in these models, and is a direct result of variability of the population mean, $\mu_m$, as compared to the between-person variability, $\eta^2$. A formal discussion of both the Bayes factor computations and the hierarchical structure of the models is provided in Haaf and Rouder (2017).

Once the predictions are known, model comparison is simple. All we need to do is note where the data fall (cf. complexity and fit in Hoijtink (2012)). The red dots in the right column denote hypothetical observed sample differences, $\hat{\Delta}_{im}$, for both participants. As can be seen, this datum is best predicted by the Analog-Representation model. In computation, this comes down to comparing samples from the prior (predictions) to samples from the posterior (where the data fall).

## Results

The results of the analysis of Rouder et al.'s (2005) data are shown in Figure 3. Panel A depicts sample means across people as a function of digit condition, and it is obvious that, at an aggregated level, the Analog-Representation explanation is preferred. The next question is whether all participants use this analog representation. Panel B shows the participant-specific sample means, $\hat{\Delta}_{im}$ for all people and differences. As can be seen, there is a lot of variability. The variability in this example shows how difficult it is to answer the question of whether everybody uses an analog representation by just inspecting sample effects. A model-based analysis is needed.

Panel C shows model-based estimates from the unconstrained model. Here, the hierarchical structure in the prior results in the regularization of difference contrasts indicating that much of the variability in the sample means in Panel B reflects trial-to-trial noise. This type of regularization depends on the number of observations within a person, and it is common in within-subject designs, and it shows a key practical value of hierarchical models in analysis (Efron & Morris, 1977; Gelman & Carlin, 2017; Rouder et al., 2005). By inspection, all but one of the estimates are above zero. As a consequence, it may be tempting to use these estimates as evidence for the everybody-has-analog-representation constraint. Yet, much more caution is needed here. The individual estimates provided by the unconstrained model are not independent over persons, and as a consequence, observing which are above and below zero may be deceiving. For these data, estimation is useful, but it does not address the question of interest.

This question may be answered by computing Bayes factors for the four models. First, we compare the three *everybody does* models. The clear winner is the Analog-Representation model, and it beats the Propositional-Representation model by over $10^{55}$-to-1. The Priming + Spreading-Activation model performs so poorly that its decrement is outside of our numerical precision, and it is at least 100 orders worse than the Analog-Representation model. Second, we may also compare the Analog-Representation model to the unconstrained model, and the Bayes factor is 1.4-to-1 in favor of the former.

Practically, the data provides equivalent support for everybody-has-analog-representation as it does for the unconstrained model. Note that this unconstrained model was used as a none-of-the-above option. It does not exclude the options of the other models, but it is only preferred if none of the other models predicts the data well enough. We may use the estimates from the unconstrained model to diagnose the source of the equivalence. As can be seen in the sample means in Figure 3B, the overall trend is for smaller values for the first and fourth difference ($\hat{\Delta}_{i1}$ and $\hat{\Delta}_{i4}$), and, concordantly, many participants have negative observed values for these contrasts. In retrospect, this behavior is quite reasonable as there must be a fall off in the distance-from-five effect with increasing distances. For example, large numbers, say 21 and 22, are probably both classified as greater than five with the same speed. Indeed, this type of fall-off is prevalent with physical uni-dimensional quantities such as brightness and tone volumes (Luce, 1986). For future application, a more refined instantiation of the Analog-Representation model that incorporates this expected decrease may be useful.

We are sanguine about the utility and parsimony of the *everybody does* models. The data rule out the Propositional-Representation and Priming + Spreading-Activation models while retaining the plausibility that everybody uses analog representation in this task.

## Conclusions and Limitations

Given the coarse state of theory in the social sciences, a profitable avenue for gaining constraint and falsifiability is to propose multiple ordinal constraints simultaneously. Here, we develop principled and straightforward Bayesian inference that is broadly applicable and scales up well to questions such as "does everyone," "do some," or "do none." It is our hope that such tools will motivate researchers to better specify theory as systems of constraints.

In the beginning of this paper, we noted that models with metric relations, process models, rarely make metric

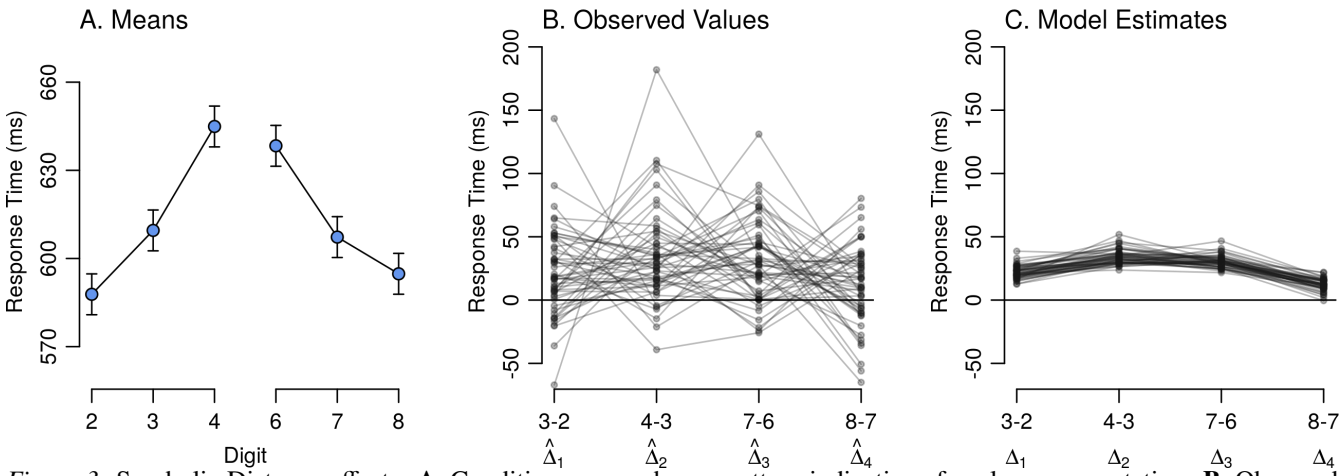


*Figure 3.* Symbolic Distance effects: **A.** Condition means show a pattern indicative of analog representation. **B.** Observed successive differences for all individuals. Positive values are indicative of analog representation. **C.** Model estimates of successive differences from the unconstrained model. There is much shrinkage indicating that much of the noise in the observed values is due to trial-by-trial noise.

predictions on data. Instead, much like physical models, they yield conservations. For example, the diffusion model of perception (Ratcliff & Rouder, 1998) predicts that speed-vs-accuracy instructions should affect a bound parameter and not a stimulus-strength parameter (drift rate). These predictions form a system of orders that may be analyzed with the aforementioned developments.

There are perhaps three limitations of the current approach: 1. the use of parametric models, 2. the need for prior specifications, and 3. the lack of implementation of these methods in popular software platforms. We take these in turn.

The current modeling approach is to place ordinal constraints on mean parameters in linear models with normally-distributed noise (Klugkist et al., 2005). One may wonder about the sensitivity of inference to the normal specification. In linear models, effects are assumed to shift the noise distribution without changing its scale or shape. The general consensus is that these specifications are not too problematic in classical inference (Hays, 1994). In our experience, both from simulation and from practice (Haaf & Rouder, 2018; Rouder, Haaf, Stober, & Hilgard, submitted; see, for example, Thiele, Haaf, & Rouder, 2017), the effects we explore are so small relative to trial noise that the shifted normal is a fine approximation.

The more concerning limitation is the need to specify prior distributions with tuning settings. The main parameters of concern are those that differentiate the models, and in the case above, these are $\mu_m$ and $\eta^2$. The prior on these parameters define *a priori* expectations about the overall size of an effect and *a priori* expectations about between-person variability around this overall mean. The choice of these tuning parameters is a substantive choice rather than a statistical one, and researchers who are substantive experts in their domain should be unafraid to add value here. We set ours as follows: We reasoned that any distance-from-five effect is roughly on the order of 50 ms and between-person variability is roughly on the order of 30 ms. This is not to say that $\mu_m$ and $\eta^2$ are these values, but they come from distributions that reflect these settings. Our advise is that researchers should use a range of scale settings they find reasonable and track how inference changes across these settings. Detailed discussion and examples of this approach may be found in Rouder, Morey, and Wagenmakers (2016) and Haaf and Rouder (2017).

The final limitation is that the current analyses are not yet available in popular packages. Perhaps the easiest software package to use for computing Bayes factor is `JASP` (Love et al., 2015). This package was reverse engineered to look and feel like `SPSS`, and users familiar with with `SPSS` can use `JASP` easily. `JASP` has some inequality constraints implemented, but is not yet set up to analyze systems of order constraints more generally. Users of `R` can get the same functionality out of the `BayesFactor` package (Morey & Rouder, 2015). Less well known is the package `BAIN` (Gu, Mulder, & Hoijtink, in press), which is also an `R` package specifically designed for computing Bayes factors with equality and inequality constraints. While the current software situation is fluid and there are turn-key solutions for only a handful of models, we suspect that easy-to-use packages which address a full range of order constraints will be available soon.

## Appendix

Prior specifications are needed for the parameters $\sigma^2$, the trial-by-trial variation, $\nu_{i3}$ and $\nu_{i7}$, the mean response times for the two contrasting conditions, $\mu_m$, the means of difference contrasts, and $\eta^2$, the variance of difference contrasts.

The full Bayesian specification of the model comes from

Haaf and Rouder (2017). Priors on $\sigma^2$, $\nu_{i3}$ and $\nu_{i7}$ are fairly broad and do not affect model comparison. The crucial prior specifications are on $\mu_m$ and $\eta^2$, and we place mildly informative priors on these parameters following Zellner's $g$-prior specification (Zellner, 1986). Let $g_\Delta = \eta^2/\sigma^2$. Then:

$$\mu_m \sim \text{Normal}(0, g_\mu \sigma^2),$$
$$g_\Delta \sim \text{inverse-}\chi^2(1, r_\Delta^2),$$
$$g_\mu \sim \text{inverse-}\chi^2(1, r_\mu^2),$$

where inverse-$\chi^2$(a,b) is a scaled inverse chi-squared distribution with $a$ degrees-of-freedom and a scale of $b$ (see Gelman, Carlin, Stern, & Rubin, 2004). The following settings are used: $r_\Delta = .1$, and $r_\mu = .16$.

It is important to know how these substantive choices for setting $r_\Delta$ and $r_\mu$ affect model comparison results. We address this issue in Conclusions and Limitations.

## References


Cohen, J., Dunbar, K., & McClelland, J. (1990). On the control of automatic processes: A parallel distributed processing account of the stroop effect. *Psychological Review*, *97*, 332–361.

Collins, A., & Loftus, E. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*, 407–428.

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

Gallistel, C. R., & Gelman, R. (1992). Preverbal and verbal counting and computation. *Cognition*, *44*, 43–74.

Gelfand, A. E., Smith, A. F. M., & Lee, T.-M. (1992). Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *Journal of the American Statistical Association*, *87*(418), 523–532. Retrieved from http://www.jstor.org/stable/2290286

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

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

Gu, X., Mulder, J., & Hoijtink, H. (in press). Approximated adjusted fractional bayes factors: A general method for testing informative hypotheses. *British Journal of Mathematical and Statistical Psychology*. Retrieved from http://dx.doi.org/10.1111/bmsp.12110

Haaf, J. M., & Rouder, J. N. (2017). Developing constraint in Bayesian mixed models. *Psychological Methods*, *22*(4), 779–798.

Haaf, J. M., & Rouder, J. N. (2018). *Some do and some don't? Accounting for varieties of individual difference structures.* Retrieved from https://psyarxiv.com/zwjtp/

Hays, W. L. (1994). *Statistics* (fifth.). Ft. Worth, T.X.: Harcourt Brace.

Hoijtink, H. (2012). *Informative Hypotheses. Theory and Practice for Behavioral and Social Scientists*. Boca Raton: Chapman & Hall/CRC.

Hoijtink, H., Klugkist, I., & Boelen, P. (2008). *Bayesian evaluation of informative hypotheses*. New York: Springer.

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

Kahneman, D. (1973). *Attention and effort*. Englewood Cliffs, NJ: Prentice-Hall.

Klaassen, F., Zedelius, C., Veling, H., Aarts, H., & Hoijtink, H. (2017). All for one or some for all? Evaluating informative hypotheses for multiple N=1 studies. *Behavior Research Methods*. Retrieved from

10.3758/s13428-017-0992-5

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

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: A bayesian approach. *Psychological Methods*, *10*(4), 477.

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Lewandowsky, S., & Farrell, S. (2011). *Computational modeling in cognition: Principles and practice*. Thousand Oaks, CA: Sage.

Logan, G. D. (1988). Towards an instance theory of automization. *Psychological Review*, *95*, 492–527.

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

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

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

Pe, M. L., Vandekerckhove, J., & Kuppens, P. (2013). A diffusion model account of the relationship between the emotional flanker task and rumination and depression. *Emotion*, *13*(4), 739–747.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for decisions between two choices. *Psychological Science*, *9*, 347–356.

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

Rouder, J. N., Haaf, J. M., Stober, C., & Hilgard, J. (submitted). *Beyond overall effects: A bayesian approach to finding constraints across a collection of studies in meta-analysis*. Retrieved from https://psyarxiv.com/zubr3/

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

Silvapulle, M. J., & Sen, P. K. (2011). *Constrained statistical inference: Order, inequality, and shape constraints* (Vol. 912). John Wiley & Sons.

Smith, P. L. (2000). Stochastic dynamic models of response time and accuracy: A foundational primer. *Journal of Mathematical Psychology*, *44*, 408–463.

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

Thiele, J. E., Haaf, J. M., & Rouder, J. N. (2017). Bayesian analysis for systems factorial technology. *Journal of Mathematical Psychology*, *81*, 40–54.

Townsend, J. T., & Ashby, F. G. (1983). *Stochastic modeling of elementary psychological processes*. London: Cambridge.

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