Does Every Study? Implementing Ordinal Constraint in Meta-Analysis

Abstract

Enter abstract here. Each new line herein must be indented, like this line.

Keywords: keywords

Word count: X

Does Every Study? Implementing Ordinal Constraint in Meta-Analysis

The main focus of the previous chapters is to develop models of individual differences for cognitive tasks. These models, however, can be easily applied to other areas of psychology. We may address questions like: Does *everyone* show reduced symptoms after a clinical intervention? Are there some people who truly show no implicit racial bias, or even an opposite bias? Whenever it is possible to model several observations per individual, this information can be used to compare theoretically informed models of ordinal and equality constraints (Haaf, Klaassen, & Rouder, In preparation).

Another possible application is to, instead of modeling individuals, modeling *studies* in meta-analysis. We may ask questions such as: Does every study show an effect of a clinical intervention? Are there some studies that show a truly opposite racial bias effect? Questions of this type were first raised by Rouder, Haaf, Stober, and Hilgard (submitted) who proposed a modeling approach based on Haaf and Rouder (2017). This chapter provides an extension of their approach that widens its applicability to many common meta-analysis settings.

3.1 Introduction

One of the most important tools for literature review in psychology is meta-analysis. There are two main goals when conducting a meta-analysis: 1. To estimate the size of population effect from a collection of studies in the literature; 2. To estimate the variability of that effect across studies and identify the sources of this variability. Sources that may be identified are different samples, manipulations, or situations. A common approach to meet both of these goals is to conduct a random effects meta-analysis (Borenstein, Hedges, Higgins, & Rothstein, 2010). In random effects meta-analysis, it is assumed that each study's true effect comes from a distribution, typically a normal distribution. The most common meta-analytic model is as follows: Let T_i be the ith study's observed effect,

$i = 1, \ldots, I$. Then,

$$T_i \sim \text{Normal}(\theta_i, \sigma^2), \tag{1}$$

where θ_i is the ith study's true effect and σ^2 is the within-study variability. Note that σ^2 is the same for each study. In a random-effects model, θ_i are assumed to have some variability. Here, θ_i are typically modeled as coming from a normal distribution:

$$\theta_i \sim \text{Normal}(\mu, \tau^2), \tag{2}$$

where μ is the true average effect and τ^2 is the true study variability. To meet the first of the above goals, μ serves as an estimate of the population effect. To meet the second goal, τ^2 serves as an estimate of true variability after accounting for sample noise, σ^2. The random effects model therefore allows to distinguish between noise variability and true variability. Once the amount of true variability is established, researchers may seek for covariates that explain the variability of the effect in the literature, such as different samples, contexts or manipulations.

Yet, how shall we interpret mean effect and variance estimates? For example, we may want to meta-analyse studies on extinction in classical conditioning. In this area, experiments typically assess various measures including heart rate, reaction time and skin conductance. What is a mean effect across these different dependent measures with different units? And how can we interpret the true variance? Likewise, the assumption of equal variance in Equation (1) is more than questionable in this case. The useful application of this modeling approach is if T_i are not the observed effects, but instead some sort of transformation that consolidates all these different effects onto the same scale. In meta-analysis, this is typically done using effect size measures such as Fisher's Z or Cohen's d. Yet, the equal variances assumption remains problematic. Also problematic are the distributional properties assumed by Borenstein et al. (2010). Take, for example, the effect of a clinical intervention on self-reported well-being. We may compare two hypothetical distributions of true study effects in Figure 1. Both distributions have the same mean intervention effect and the same variance. Yet, they lead to entirely different conclusions about the efficacy of the treatment.

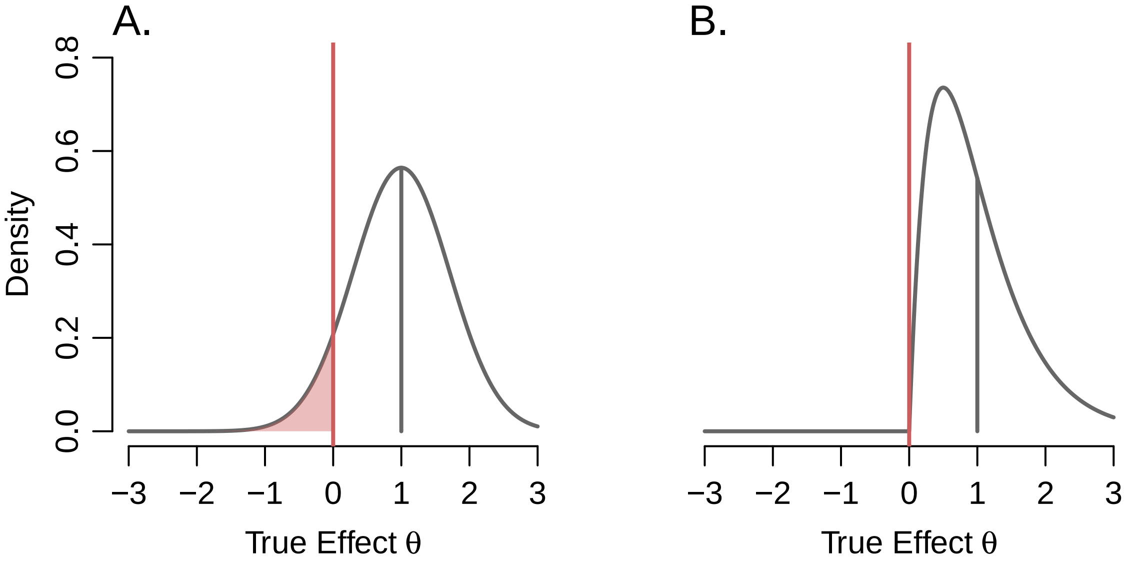

Figure 1. Hypothetical distributions of true study effects with mean of one and variance of one half. A. A normal distribution with a domain of all real numbers. The shaded area represents the probability of truly negative effects. B. A gamma distribution with domain of all positive reals. All true effects have to be positive.

Panel A in Figure 1 shows a normal distribution, the typical distribution assumed in random effects meta-analysis (see Equation (2)). The majority of true study effects are assumed to be positive, that is, well-being is enhanced after the intervention. Yet, the shaded area shows that there is some probability of a negative effect, that is, well-being is reduced after the intervention. This result would be highly problematic when arguing for a broader application of the intervention. Clearly, further investigation is needed to show under which circumstances the intervention reduces well-being, and we may limit the recommendation to circumstances under which the intervention is truly helpful.

Panel B in Figure 1 shows a gamma distribution. This distribution does not support negative true effects. Even though the variability is the same for both distributions, the different domain of the gamma leads to a less ambiguous explanation of the efficacy of the clinical intervention. Some circumstances may lead to a bigger or smaller effect on well-being, but the intervention may be recommended in all settings without limitations.

Rouder et al. (submitted) introduce a modeling approach that includes versions of the two models in Figure 1. The authors propose four meta-analytic models: An *unconstrained model* just like the one in Figure 1A, a *positive-effects model* similar to the one in Figure 1B, a *common-effect model* with no true study variability, and, finally, a *null model* where all studies have a true zero effect. The models are compared using a Bayes factor model comparison approach utilizing both an analytic Bayes factor solution (Rouder, Morey, Speckman, & Province, 2012; Zellner & Siow, 1980) and the encompassing approach (Haaf & Rouder, 2017; Hoijtink, Klugkist, & Boelen, 2008; Klugkist & Hoijtink, 2007). Rouder et al. (submitted) applied their approach to several meta-analytic data sets. They found evidence for the null model in Wagenmakers and colleagues' (2016) large-scale registered replication of the facial feedback effect; evidence for the common-effect model in a many labs study of moral credentialism (Ebersole et al., 2016); evidence for the positive-effects model in the reanalysis of several Stroop experiments reported in Haaf and Rouder (2017); and evidence for the unconstrained model in a reanalysis of Big Five personality data with samples from numerous universities (Corker, Donnellan, Kim, Schwartz, & Zamboanga, 2017).

Yet, there are a few catches with the application of Rouder et al.'s approach to meta-analysis (submitted). First, the development is for observed effects on the same scale, such as, for example, effects on a Likert scale. Second, Rouder et al.'s approach is best suited for experimental effects. The analysis of Corker et al.'s (2017) personality traits data set already proves to be difficult. Correlational data, however, cannot be analyzed with this approach. Third, the model estimation and model comparison approaches in Rouder et al.

(submitted) require data from individual participants for each study. For the applications in Rouder et al. (submitted), none of these catches are too problematic. Most of the applications are planned many-labs endeavors with the exact same experimental designs realized in different labs, and with the commitment to open data. In the typical meta-analytic setting, however, these catches lead to real issues. Observed effects are oftentimes on different scales and units; in psychology, meta-analyses are conducted on experimental and correlational effects; and raw data are almost never available to the meta-analyst.

To address these issues, the goal of this chapter is to expand Rouder and colleagues' (submitted) approach. To do so, we propose to model effect size measures as the unitless abstraction of the naive, observed effects. This change allows us to analyze effects on different units, to expand the development for the application to correlational designs, and to only use data that are typically available in meta-analysis: Effect size estimates and sample size. Additionally, this approach allows us to drop the assumption of equal within-study variances in Equation (1). Unfortunately, using effects size estimates instead of raw data introduces the assumption that the observed within-study variance is the true within-study variance. This assumption, though unattractive, is not too problematic with increasing sample size.

The first decision needed is which effect size measure to use. Meta-analysis is not a uniform method, and there are many approaches available using different unitless abstractions of observed effects. Probably the most common abstractions are the effect size measures Cohen's d for experimental designs and Fisher's Z for correlational designs. A transformation between these and other effect size measures is possible (Borenstein et al., 2010), so model development for one of the measures may suffice. We decided to use Fisher's Z for model development here. We explain our choice and describe the implementation in the next section. We then introduce the models on Fisher's Z and expand the Bayes factor model comparison approach for this setup. Finally, we apply the approach to a meta-analysis by Anderson et al. (2010) and discuss the development in light of the results.

3.2 Modeling Fisher's Z

Fisher's Z is typically used to assess the size of a correlation coefficient. The measure is a variance stabilizing transformation of the bivariate correlation coefficient r that maps r onto the real number space. The formula for Fisher's Z is

$$Z = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right) = \text{arctanh}(r), \tag{3}$$

and the asymptotic distribution of Z is a normal distribution:

$$Z \sim \text{Normal}(\theta, \frac{1}{N-3}), \tag{4}$$

where θ is the true effect size and N is the number of observations. Importantly, the size of the correlation does not affect the variance of Z; this variance is only dependent on N. This property makes Fisher's Z an attractive target for the modeling approach.

The asymptotic distribution of Z is well established for the correlation between two independent and identically distributed normal random variables (Ferguson, 1996). In the case of correlation, Fisher's Z serves as normalized correlation measure. Yet, how does Fisher's Z translate to experimental data? Let's assume an experimental setup with j conditions, $j = 1, 2$. The ith participant is assigned to one of the two conditions, $i = 1, \ldots, 2n$, where n is the number of participants in each condition and $2n$ is the total number of participants. Then Y_i denotes the ith person's observation with

$$Y_i \sim \begin{cases} \text{Normal}(\mu_1, \sigma^2) & \text{if } i = 1, \ldots, n, \\ \text{Normal}(\mu_2, \sigma^2) & \text{if } i = n + 1, \ldots, 2n. \end{cases} \tag{5}$$

Here, μ_j denotes the true mean of the jth condition and σ^2 denotes the within-group variance. We may first calculate the effect size measure Cohen's d as $d = (\bar{Y}_2 - \bar{Y}_1)/s_{pooled}$ with the condition means $\bar{Y}_j$ and the pooled standard deviation $s_{pooled} = \sqrt{\frac{\sum_1^n (Y_i - \bar{Y}_1)^2 + \sum_{n+1}^{2n} (Y_i - \bar{Y}_2)^2}{2n}}$.[1] If we wish to calculate Fisher's Z as an estimate of the size of the effect between the two conditions, we first need to transform Cohen's d to r using

$$r = \frac{d}{\sqrt{d^2 + 4)}}.$$

To understand what this correlation coefficient represents we may define a new variable, X_i, that denotes ith participant's condition. X_i is defined as

$$X_i = \begin{cases} -1 & \text{if } i = 1, \ldots, n, \\ 1 & \text{if } i = n + 1, \ldots, 2n. \end{cases} \tag{6}$$

The correlation coefficient r is the point-biserial correlation between $\mathbf{X}$ and $\mathbf{Y}$, and Fisher's Z may be calculated using Equation (3). Yet, is the asymptotic distribution of Fisher's Z for the biserial correlation the same as the asymptotic distribution of Fisher's Z

[1] Note that this formula is based on the biased variance estimator using n in the denominator instead of $n - 1$.

for the bivariate normal? Before modeling experimental and correlational data, we need to establish distributional properties for the biserial case. Especially the asymptotic stable variance of Fisher's Z is crucial for the modeling approach. For this purpose, we take a digression, conduct a simulation study and then provide a proof. Readers who are more interested in the application rather than the fundamentals of the current modeling approach may skip ahead.

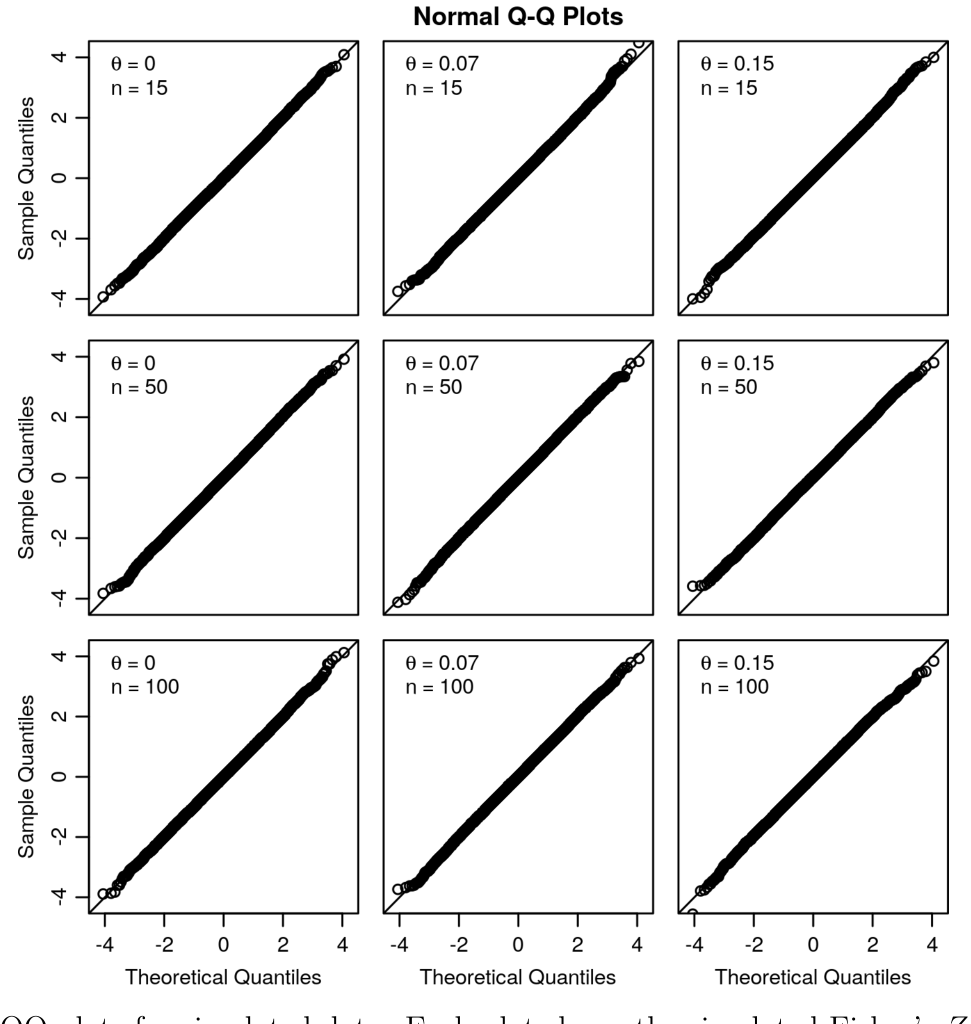

Figure 2. QQ-plots for simulated data. Each plot shows the simulated Fisher's Z for a true effect by sample size combination.

Simulation. To get a first insight in whether Fisher's Z has a stable variance for a point-biserial correlation, we conduct a small simulation with varying sample sizes and effect sizes. Data were simulated as coming from the distribution

Table 1

Simulated variances of Fisher's Z

	Per-group sample size n_ℓ		
	15	50	100
True effect δ_k			
0	0.0366	0.0104	0.0050
0.3	0.0366	0.0104	0.0051
0.6	0.0363	0.0102	0.0050
Predicted			
	0.0370	0.0103	0.0051

$$Y_i \sim \begin{cases} \text{Normal}(0,4) & \text{if } i = 1,\ldots,n_\ell, \\ \text{Normal}(\delta_k,4) & \text{if } i = n_\ell+1,\ldots,2n_\ell, \end{cases}$$

where δ_k, the true effect, may take values 0, 0.3, or 0.6, and n_ℓ, the condition sample size may take values 15, 50, or 100. All combinations $k \times \ell$ were realized $M = 10000$ times, and Z is calculated for each iteration. If the distribution of Z was the same as for a biserial correlation coefficient, we would expect the simulated Zs to come from a normal distribution. The means of these distributions are given by calculating the point-biserial correlation and applying Equation (3), and they are $\theta = 0$, 0.07, and 0.15 for $\delta_k = 0$, 0.3, and 0.6, respectively. The means are assumed to be stable across levels of sample size. The variances are given by Equation (4), and they are 0.0370, 0.0103, and 0.0051 for $n_\ell = 15$, 50, and 100, respectively. The variances are assumed to be stable across levels of effect sizes.

The simulation results are summarized in Figure 2 and Table 1. Figure 2 shows the resulting QQ-plots from the simulation with each plot representing one of the effect-by-sample-size combinations. As can be seen, the points map well onto the diagonal line. Table 1 shows the variances calculated from the 10000 simulated Z values per $\delta_k \times n_\ell$ combination. As predicted, the variances are stable across effects and only vary with sample size.

Proof. In addition to the simulation, we provide a proof that the asymptotic variance is stable and does not depend on the size of the effect. Consider again random variables Y_i and X_i as defined in Equations (5) and (6). Let $\bar{Y}$ be the grand mean of $\mathbf{Y}$, $\bar{Y} = 1/2(\bar{Y}_1 + \bar{Y}_2)$. Here, $\bar{Y}_1$ denotes the mean of Y_i for $i = 1, \dots, n$, and $\bar{Y}_2$ denotes the mean of Y_i for $i = n+1, \dots, 2n$. Let S_{YY} denote the variance of $\mathbf{Y}$, $S_{YY} = \sum_1^n (Y_i - \bar{Y}_1)^2 + \sum_{n+1}^{2n} (Y_i - \bar{Y}_2)^2 + n/2(\bar{Y}_1 - \bar{Y}_2)^2 = SS_1 + SS_2 + n/2(\bar{Y}_1 - \bar{Y}_2)^2$, where SS_1 is the sum of squares for $i = 1, \dots, n$, and SS_2 is the sum of squares for $i = n+1, \dots, 2n$. Let $S_{XX} = 2n$ denote the variance of $\mathbf{X}$ and let $S_{XY} = n(\bar{Y}_1 - \bar{Y}_2)$ denote the covariance of $\mathbf{X}$ and $\mathbf{Y}$. We can now insert these quantities into the formula for the correlation coefficient r:

$$\begin{aligned} r &= \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}} \\ &= \frac{n(\bar{Y}_1 - \bar{Y}_2)}{\sqrt{2n(SS_1 + SS_2 + n/2(\bar{Y}_1 - \bar{Y}_2)^2)}} \\ &= \frac{\sqrt{1/2}(\bar{Y}_1 - \bar{Y}_2)}{\sqrt{\frac{2(SS_1+SS_2)}{2n} + 1/2(\bar{Y}_1 - \bar{Y}_2)^2)}}. \end{aligned}$$

We may define a new random variable $m_n = \bar{Y}_1 - \bar{Y}_2$. Then, based on the Central Limit Theorem:

$$\sqrt{n}(m_n - \Delta) \xrightarrow{\mathscr{D}} \text{Normal}(0, 2\sigma^2),$$

where Δ is the true effect. The pooled variance of $\mathbf{Y}$ is $S_n = \frac{SS_1+SS_2}{2n}$.[2] The Central Limit Theorem gives:

[2] Note that S_n is not the variance of $\mathbf{Y}$, but the pooled variance of Y_i if $i = 1, \dots, n$ and Y_i if $i = n+1, \dots, 2n$

$$\sqrt{n}(S_n - 2\sigma^2) \overset{\mathscr{D}}{\to} \text{Normal}(0, 4\sigma^2).$$

Using the two asymptotic distributions, $r \to \frac{\sqrt{1/2}(\Delta)}{\sqrt{4\sigma^2+\Delta^2/2}}$. Let Σ be the variance-covariance matrix of $\binom{m_n}{S_n}$:

$$\Sigma = \begin{bmatrix} 2\sigma^2 & 0 \\ 0 & 4\sigma^2 \end{bmatrix}.$$

And let $h(a, b) = \frac{\sqrt{1/2}(a)}{\sqrt{2b+a^2/2}}$. The partial derivatives of $h(a, b)$ are:

$$\frac{\partial h}{\partial a} = \frac{\sqrt{2}b}{(2b + a^2/2)^{3/2}}, \tag{7}$$

$$\frac{\partial h}{\partial b} = -\frac{a}{\sqrt{2}(2b + a^2/2)^{3/2}}. \tag{8}$$

Using the Delta-method (Ferguson, 1996), we may express the asymptotic distribution of $h(m_n, S_n)$ as

$$\sqrt{n}(h(m_n, S_n) - h(\Delta, 2\sigma^2) \overset{\mathscr{D}}{\to} \text{Normal}(0, \Sigma^*),$$

where $h(m_n, S_n)$ is the correlation coefficient, the mean, $h(\Delta, 2\sigma^2)$, is the true correlation coefficient, ρ, and the asymptotic variance is

$$\begin{aligned}\Sigma^* &= \Big(\frac{2\sqrt{2}\sigma^2}{(4\sigma^2+\Delta^2/2)^{3/2}}\Big)^2 2\sigma^2 + \Big(\frac{-\Delta}{\sqrt{2}(4\sigma^2+\Delta^2/2)^{3/2}}\Big)^2 4\sigma^2\\ &= \frac{16}{\left(8+\frac{\Delta^2}{\sigma^2}\right)^2}.\end{aligned}$$

We may apply Fisher's Z transformation to stabilize the variance. Let $g(r) = 1/2\log\left(\frac{1+r}{1-r}\right)$. Then $g'(r) = \frac{1}{1-r^2}$. Substituting $h(\Delta, 2\sigma^2)$ for ρ, $g'(\rho) = \frac{8+\frac{\Delta^2}{\sigma^2}}{8}$. We may again use the Delta-method to express the asymptotic distribution of Z:

$$\sqrt{n}(g(r) - g(\rho) \xrightarrow{\mathscr{D}} \text{Normal}(0, (g'(\rho))^2\Sigma^*),$$

where $g(r) = Z$ and $g(\rho) = \theta$. The asymptotic variance is

$$(g'(\rho))^2\Sigma^* = \frac{\cancel{(8+\frac{\Delta^2}{\sigma^2})^2}}{8^2}\frac{16}{\cancel{\left(8+\frac{\Delta^2}{\sigma^2}\right)^2}} = 1/4.$$

Crucially, this value, 1/4, does not depend on ρ.

3.3 Models of Constraint in Meta-Analysis

After establishing the distributional properties of Fisher's Z, we may develop models on the collection of Zs. We are interested in models with ordinal and equality constraints on naive observed effects, and we may first inspect the relationship between such an observed effect, d, and the unitless abstraction of an effect, the calculated Z-value. Figure 4A shows Fisher's Z as a function of observed effects. In this case, the observed effects are for reaction time and the unit is milliseconds (ms). As can be seen, there is a relationship between observed effects and Z that larger effects lead to larger Z-values. Importantly, the ordinal descriptions of less than zero, equal to zero, and greater than zero are the same between d and Z. Because the ordinal properties are preserved, whenever a model restricts Z to be positive, the underlying naive effect is also restricted to be positive. We are now ready to develop models on Fisher's Z.

For the following models, let i denote the study in the meta-analysis, $i = 1, \ldots, I$, let j denote the condition in the study, $j = 1, \ldots, 2$, and let k denote the participants, $k = 1, \ldots, N_i$. To learn about the sign and variability of effects, we develop a set of models on the collection of Z_i. The basic model here is

$$Z_i \sim \text{Normal}\left(\theta_i, \frac{1}{N_i - 3}\right), \tag{9}$$

where θ_i is the ith study's true Z-value. We may place models with ordinal and equality constraints on the collection of these true study effects θ_i.

The General Model. The general model is a standard linear model without constraints, and it corresponds to the typical random-effects meta-analytic model in Equation (2). Here, the collection of study effects simply follows a normal distribution:

$$\mathcal{M}_g: \qquad \theta_i \overset{iid}{\sim} \text{Normal}(\nu, \eta^2),$$

where ν (equivalent to μ in Equation (2)) is the mean effect and η^2 (equivalent to τ^2 in Equation (2)) is the variance of effects. No constraints are placed on the collection of θ_i such that effects for some studies may truly be positive while effects for other studies may be truly negative.

The Positive-Effects Model. The positive-effects model corresponds to the hypothesis that every study has a true effect in the same direction:

$$\mathcal{M}_p: \qquad \theta_i \overset{iid}{\sim} \text{Normal}^{+}(\nu, \eta^2),$$

where Normal$^+$ is a normal distribution truncated below at zero. This model statement implies multiple order constraints, one for each study in the data set, reducing model complexity drastically compared to the general model. Assuming the same mean and variance as the general model, the positive model has higher density for (small) positive effects and no mass below zero.

The Null Model. The last model proposed is a null model. Here, all θ_i are exactly zero:

$$\mathcal{M}_0: \qquad \theta_i = 0.$$

This model is the most constrained model of the three, and it is a strict null in that it does not allow for true variability around zero. Instead, all variation is assumed to be noise. If an effect truly does not exist, then all studies truly should have a null result. True variation around zero would imply that there has to be a trade-off between the studies such that all true effects sum to zero, which is highly unlikely.

Prior Settings. For Bayesian analysis, priors are needed on the mean and variance parameters in the models. Here, hierarchical prior distributions are placed on ν, the true mean effect size, and on η^2, the true variability of effect sizes. I chose conjugate prior distributions:

$$\begin{aligned}\nu &\sim \text{Normal}(0, \sigma_\nu^2),\\ \eta^2 &\sim \text{Inverse-Gamma}(1, .01).\end{aligned}$$

Another inverse-gamma prior is placed on the variance of ν:

$$\sigma_\nu^2 \sim \text{Inverse-Gamma}(1, .01).$$

The priors on ν and η^2 are critical for model comparison because they are constrained for some of the models. Therefore, the prior settings, especially the shape and scale settings of the inverse-gamma distributions have to be chosen with care. Typically, these settings should be made by researchers familiar with the scaling of the variable (Haaf & Rouder, 2017; Rouder, Haaf, & Aust, 2018). Yet, we have limited experience with the application of Fisher's Z, and therefore have limited confidence in the prior settings. Therefore, a sensitivity analysis is crucial to assess the variability of the Bayes factor within a reasonable range of prior settings. Such a sensitivity analysis is subsequently provided for one of the applications.

3.4 Model Comparison

The main purpose of the modeling approach is to learn about the distribution of effects in a meta-analytic set. To do so, models with theoretically motivated constraints may be compared in a Bayesian setting. We propose a Bayes factor model comparison account to compare the general model, the positive-effects model and the null model. Here, we provide an informal discussion of Bayes factors, and we briefly present the Bayes factor estimation approaches. A more extensive discussion may be found in Jeffreys (1961), Kass and Raftery

(1995), Morey, Romeijn, and Rouder (2016), and Rouder, Morey, and Wagenmakers (2016).

In Bayes factor model comparison, the main target of interest is the relative evidence for a model compared to another. The Bayes factor is this relative evidence, and it results directly from Bayes rule. Bayes rule for two models, $\mathcal{M}_1$ and $\mathcal{M}_2$ is

$$\frac{P(\mathcal{M}_1 \mid \boldsymbol{Y})}{P(\mathcal{M}_2 \mid \boldsymbol{Y})} = \frac{P(\boldsymbol{Y} \mid \mathcal{M}_1)}{P(\boldsymbol{Y} \mid \mathcal{M}_2)} \times \frac{P(\mathcal{M}_1)}{P(\mathcal{M}_2)}, \tag{10}$$

where $\frac{P(\mathcal{M}_1|\boldsymbol{Y})}{P(\mathcal{M}_2|\boldsymbol{Y})}$ are the posterior odds, $\frac{P(\mathcal{M}_1)}{P(\mathcal{M}_2)}$ are the prior odds, and $\frac{P(\boldsymbol{Y}|\mathcal{M}_1)}{P(\boldsymbol{Y}|\mathcal{M}_2)}$ is the Bayes factor. The Bayes factor is also referred to as the updating factor because it is the amount by which the prior odds need to be updated in the light of the data to get to the posterior odds. The Bayes factor is therefore the evidence for $\mathcal{M}_1$ relative to $\mathcal{M}_2$.

We may also view the Bayes factor as *predictive accuracy* of Model 1 relative to the predictive accuracy of Model 2. In this sense, the Bayes factor denotes how well $\mathcal{M}_1$ predicted the data compared to $\mathcal{M}_2$. Figure 3 illustrates this point. The left panel in the top row of the figure shows the model specifications for the positive-effects and the general model for any one study. The predictions for data from these models are shown in the right panel, top row. As can be seen, the positive-effects model best predicts small positive effects while the general model predicts both small positive and negative effects to the same degree. As a result, if a small positive effect is observed, the positive-effects model will be preferred as it has more density for positive effects than the general model. In contrast, if a small negative effect is observed the general model will be preferred. Yet, the positive-effects model *can* predict small observed negative effects despite the ordinal constraint on true effects.

Figure 3 shows multivariate model specifications for any two studies for both the general and the positive-effects model. The middle row shows model specification and model predictions for the positive-effects model. The effect size for study 1 is specified on the x-axis; the effect size for study 2 is specified on the y-axis. The correlation between the two effects is introduced by the hierarchical nature of the models, and is a desirable feature of hierarchical modeling. This correlation is also preserved in the predictions, and the positive-effects model best predicts small, similar, positive effects. The same correlation is introduced in the general model for two effects, shown in the bottom row of Figure 3. The model specification and predictions are relatively diffused as they again cover more of the parameter space than the positive-effects model.

To quantify the relative predictive accuracy, Bayes factors can be estimated in several ways. Here, we use the encompassing approach to estimate the Bayes factor between the general model and the positive-effects model (Haaf & Rouder, 2017; Hoijtink, 2012; Klugkist, Laudy, & Hoijtink, 2005), and an analytic approach to assess the Bayes factor between the general model and the null model (Rouder et al., 2012). The Bayes factor between the positive-effects model and the null model can be obtained using the transitivity property of Bayes factors.[3]

The encompassing approach (Hoijtink, 2012) may be used to estimate Bayes factors between nested models, where one model is an order-constrained version of a more general model. This is the case with the general and positive-effects models. The Bayes factor between the general and the positive-effects model is

$$B_{g+} = \frac{P(\boldsymbol{\theta} > \mathbf{0} | \mathcal{M}_{\mathbf{g}})}{P(\boldsymbol{\theta} > \mathbf{0} | \mathbf{Y}, \mathcal{M}_{\mathbf{g}})},$$

where $\boldsymbol{\theta}$ is the collection of θ_i, and $\boldsymbol{Y}$ are the observed data. Using the encompassing approach, this Bayes factor can be estimated using samples from the posterior and prior distributions of the general model $\mathcal{M}_g$. We may sample M samples from the prior and

[3] Using transitivity, the Bayes factor between the positive-effects model and the null model, B_{+0}, may be obtained as $B_{+0} = \frac{B_{g0}}{B_{g+}}$.

posterior distributions of θ_i with m indicating a sample, $m = 1, \ldots, M$. The mth sample is evidential of the positive-effects model if all θ_i in the sample are positive. Let n_{0+} indicate the frequency of evidential samples from the prior, and let n_{1+} indicate the frequency of evidential samples from the posterior. Then, the Bayes factor between the general and the positive-effects models is approximately

$$B_{g+} \approx \frac{n_{0+}}{n_{1+}}.$$

The analytic approach employed here is taken from Rouder et al. (2012), and it is used to estimate the Bayes factor between the general model and the null model. The main targets of the approach are the probability of data conditional on the two models. This conditional probability may generically be expressed using The Law of Total Probability as

$$P(\boldsymbol{Y} \mid \mathcal{M}) = \int_{\boldsymbol{\xi} \in \Xi} P(\boldsymbol{Y}|\boldsymbol{\xi}) P(\boldsymbol{\xi}) d\boldsymbol{\xi}, \tag{11}$$

where $\boldsymbol{\xi}$ is a vector of parameters from parameter space Ξ. For the null model, this quantity is straight-forward to compute. The likelihood function, $P(\boldsymbol{Y}|\boldsymbol{\xi})$, is given by Equation (9), and it is the product of normal densities with mean θ_i and variance $\frac{1}{n_i - 3}$ evaluated for the data. For the null model, the parameter space of θ_i is reduced to a point, to zero. So, the integral in (9) is simply the likelihood of the data when $\theta_i = 0$ for all i. For the general model, the integral in (9) may be simplified by integrating out the collection of θ_i. The marginal likelihood of Z_i when integrating out θ_i and ν is

$$Z_i | \eta^2, \sigma_\nu^2 \sim \text{Normal}(0, \frac{1}{n_i - 3} + \eta^2 + \sigma_\nu^2),$$

where η^2 is the variance of θ_i and σ_ν^2 is the variance of ν. The integral is now reduced to two dimensions, and may be evaluated for possible values of η^2 and σ_ν^2 to estimate $P(\boldsymbol{Y} \mid \mathcal{M}_g)$.

3.5 Application I: Simulation study

The proposed model comparison approach is new, and modeling Fisher's Z instead of observed effects or raw data may change model estimation and model comparison substantially. I therefore decided to compare this new approach with Rouder et al.'s (submitted) approach using a simple simulated meta-analytic data set. The data set contains 100 studies with two conditions each. For each study, the sample size is 100 with 50 participants in each condition. In total, 10000 observations are analysed. Data are simulated from the following hierarchical model:

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

where the grand mean is $\mu = 1$, the study baseline was sampled from $\alpha_i \sim \text{Normal}(0, 0.07^2)$, the study effect was sampled from a normal distribution truncated below at zero, $\delta_i \sim \text{Normal}^+(0.065, 0.03^2)$, and the within-study variance is $\sigma^2 = .3^2$. The seed in `R` was set to `set.seed(123)` for the sampling.

The data are first analysed using the `nWayAOV()` function from the `BayesFactor` package in `R` (similar to Rouder et al., submitted). Here, observed effects are analyzed. In a second analysis, Fisher's Z values are calculated from Equation (3), and these values are submitted to the current-approach analysis.

Observed effects and observed Fisher's Z values may be compared in Figure 4A. The biggest change is a larger range of Z values than observed effect values. This relationship, however, is a function of the units and scale of the submitted data, it is not a function of

Fisher's Z. Panels B and C in Figure 4 show the estimation results from the general model using the all-data approach and the current approach, respectively. Both approaches lead to a considerable amount of hierarchical shrinkage which is both expected and desirable (Efron & Morris, 1977). Critically, even though 13 observed study effects are negative, all estimated posterior mean effects from both approaches are positive, just as the true effects in the simulation.

To assess evidence for the "does every study show an effect?" question, Bayes factors are computed, again using the Rouder et al. approach and the current approach. Both approaches provide evidence for the positive-effects model over the other models, but to different degrees. Using the Rouder et al. approach, the Bayes factor for the positive-effects model over the general model is 4.41 to one; the Bayes factor for the positive-effects model over the null model is 3.7×10^{25} to one. Using the current approach, the Bayes factor for the positive-effects model over the general model is 3.35 to one; the Bayes factor for the positive-effects model over the null model is 4×10^{9} to one.

The Bayes factors from the two approaches differ, but the crucial difference, the one between the positive-effects model and the general model, is fairly small. This difference is most likely due to prior settings. Rouder et al. (submitted) use a g-prior approach, and the settings on observed effects are based on experience with these types of experimental effects. The settings from the new approach may not be as well-tuned. To assess the sensitivity of the Bayes factors from the current approach to these prior settings, we conducted a sensitivity analysis. Therefore, we repeated the analysis for varying, reasonable prior settings. The critical settings for the Bayes factors are the prior settings on η^2, the variance of study effect sizes θ_i, and σ_ν^2, the variance of the mean effect size ν. As a reminder, the priors are

Table 2
Bayes factors for simulation study

	Priors on η^2		Priors on σ_ν^2			
	a	b	c	d	B_{pg}	B_{p0}
1*	1	0.01	1	0.01	3.4	4×10^9
2	2	0.01	2	0.01	7.6	2.4×10^{10}
3	2	0.01	1	0.01	2.7	5.5×10^9
4	0.5	0.01	0.5	0.02	1.3	1.2×10^8
5	2	0.01	0.5	0.02	1.1	4.9×10^8
6	0.5	0.01	2	0.02	010	6.2×10^9

Note. First row shows Bayes factors. Other rows show sensitivity analysis with different prior settings.

$$\eta^2 \sim \text{Inverse-Gamma}(a, b),$$

$$\sigma_\nu^2 \sim \text{Inverse-Gamma}(c, d),$$

where a and c are set to 1, and b and d are set to .01. These settings, however, may be varied in a reasonable range. The results of this variation are shown in Table ??. The top row shows the results for the chosen prior settings; rows 2 and 3 show more informed priors favoring less variability; row 4 shows a less informative prior allowing for larger variability; and the last two rows show variations where one of the settings on one of the variances is more informed and settings on the other are less informed. The resulting Bayes factors vary only to a small degree. Throughout, the positive-effects model is more or less favored with the smallest Bayes factor for less informed priors. This result seems reasonable: If the priors on the general and positive-effects models are similarly broad, and observed effect sizes are mostly positive, none of them can be favored to a high degree.

3.6 Application II: Anderson et al.

Modeling Fisher's Z is feasible from a statistical point of view, and it holds up in comparison with Rouder and colleagues' approach (submitted) when applied to a simulated data set. What can we learn from this approach when applied to an actual meta-analytic data set? Here, we apply the modeling approach to a meta-analysis by Anderson et al. (2010) on violent video games and aggression.

The meta-analysis by Anderson et al. (2010) investigated effects of violent video games on six different dependent variables: Aggressive affect, aggressive behavior, aggressive cognition, and physiological arousal, pro-social behavior, and empathy. In the meta-analysis, studies with experimental, cross-sectional, and longitudinal design are included resulting in a data set consisting of 381 effect size estimates from a total of 130,296 participants. About 60% of these effect size estimates met "best practices" criteria set by Anderson et al. (2010). The results yielded support for the hypothesis that violent video games increase aggressive affect, behavior and cognition while reducing empathy and pro-social behavior. Additionally, Anderson et al. (2010) report that overall effect size estimates were even higher for "best practices" studies compared to methodologically weaker studies.

The gray line in Figure 5 shows the ordered Fisher's Z effect size measures for all studies from Anderson et al. (2010) that have aggressive affect, aggressive behavior, aggressive cognition, or physiological arousal as dependent variables (302 in total). From the figure, it is apparent why Anderson and colleagues conclude there is evidence for an effect: The naive average of these Z-values is 0.19. Yet, considerable critique on the meta-analysis emerged, and this criticism was mostly of methodological nature. Hilgard, Engelhardt, and Rouder (2017) identified several sources of biases that may be partially responsible for Anderson and colleagues' results: Publication bias, questionable research practices, and selection bias. The selection bias may be most pronounced in the criteria set for "best practices". To address these issues, Hilgard et al. (2017) reanalyzed the meta-analysis data set using improved correction for publication bias, and the results suggested a much more nuanced picture. In fact, the largest amount of publication bias was detected for the "best practices" studies.

The full data set is provided in the Open Science Framework repository accompanying Hilgard and colleagues' article. We applied the modeling approach to the subset of 302 studies in Figure 5. The blue points show the posterior estimates from the general model. There is a substantial amount of hierarchical shrinkage to the posterior mean effect size. Note that, in contrast to the simulation data set, the sample sizes of the studies vary. This variability has an effect on the posterior estimates of θ_i including the credible intervals. The sample sizes vary substantially between 9 and 7137 participants. The smaller the sample size, the more shrinkage toward the mean, and the wider the credible interval. Conversely, the larger the sample size the less shrinkage toward the mean, and the tighter the credible interval. Due to this effect of sample size, the posterior estimates appear to be less related to the observed Fisher's Z than for the simulation study in Figure 4. Yet, smaller Z-values still lead to smaller θ-estimates, and larger Z-values lead to larger θ-estimates. Additionally, two of the observed negative Z-values have a negative θ-estimate.

In model comparison for this large meta-analysis the general model is preferred by a Bayes factor of 16.15 to one over the positive-effects model. The Bayes factor comparing the general and the null model could be estimated because the marginal likelihood for the null model was out of computer precision (essentially zero). This result may be interpreted as extremely high evidence against the null model.

Why was the general model preferred? The direct interpretation is that some studies truly have negative effects. Yet, *why*, and *which ones* are questions immediately following this interpretation. To investigate this question further, we selected two subsets from the full data set to apply the Fisher's Z modeling approach. First, we selected all experiments with aggressive behavior as dependent variable. Hilgard et al. (2017) describe this subset as the most evidential for the effect of violent video games. Second, we selected all non-published experiments in the set. These studies are from dissertations, and Hilgard et al. (2017) argue that they are less likely to suffer from publication bias as compared to journal articles.

Violent Video Games and Aggressive Behavior. The subset of aggressive behavior studies consists of 39 experiments. The gray line in Figure 6 shows the observed Z-values. Six of these studies showed an observed negative Z-value. Yet, the posterior estimates of θ_i, depicted by the points, are all positive. The reason is again hierarchical shrinkage to the mean, ν. The posterior estimate of ν is depicted by the dashed line and is 0.17, 95% CI [0.13 , 0.22]. The influence of shrinkage is again dependent on sample size, and sample sizes vary between 14 and 515 participants.

The estimation results show that, even though some observed Z-values are negative, there may be no evidence for the general model. All θ_i estimates are positive. Yet, we have to be careful with inference from these hierarchical model estimates due to the built-in dependency between the estimates (Haaf et al., In preparation). For inference, Bayes factors are the more appropriate. For this subset of Anderson and colleagues' (2010) meta-analysis there is some evidence in favor of the positive-effects model over the other models: The Bayes factor between the positive-effects model and the runner-up, the general model, is 9.37 to one; the Bayes factor between the positive-effects model and the null model is 1.2×10^{12} to one.

In summary, there is a large amount of evidence that there is an effect of violent video games on aggressive behavior, and there is also evidence that the effects from all studies are truly positive. Yet, this subset of studies may suffer from a high degree of publication bias. Hilgard et al. (2017) report drastically reduced effect size estimates after adjusting for publication bias. For a less biased subset, we analyse unpublished experimental data subsequently.

Unpublished Experiments. The set of unpublished experimental data in the Anderson et al. (2010) meta-analysis comes from dissertation theses, and it contains 26 effect size estimates. Here, studies with aggressive affect, aggressive cognition, aggressive behavior, and physiological arousal as dependent variables are included. The observed Z-values are depicted by the gray line in Figure 7. Eight of the Z-values are negative. The sample sizes range from 24 to 123 participants, resulting in relatively smooth posterior estimates of θ_i (points) and credible intervals (gray area). Despite a large amount of hierarchical shrinkage, one of the posterior estimates are negative. One reason is that the estimates are shrunken to a mean effect size, ν, that is estimated to be close to zero, $\hat{\nu} = 0.05$, 95% CI [-0.01 , 0.11].

For the subset of unpublished studies, there is some evidence in favor of the null model. The Bayes factor of the null model over the general model is 24.45 to one; the Bayes factor of the null model over the positive-effects model is 108.98 to one. In summary, the analysis of a less biased sample of unpublished studies provides evidence *against* the effect of violent video games on aggression.

3.7 Discussion

When combining results from multiple studies using meta-analysis, researchers are faced with many problems. First, studies of a phenomenon or effect can be heterogeneous in many aspects including sample, study design and variable operationalization, and the units the effects are measured in. When applying meta-analytic models, however, we often make distributional assumptions to estimate an overall mean that may not be appropriate (Rouder et al., submitted). Second, oftentimes we may obtain only surface statistics such as effect size measures and sample size for each individual study in the set. This lack of raw data leads to a limited set of appropriate modeling approaches. Third, the meta-analyst has only little impact on the quality of the studies in the set. Especially recently, publication bias is discussed as one of the main reasons to question the usefulness of meta-analysis (Corker, 2018).

The current approach addresses the first two issues. Rather than solely focusing on an overall effect size measure, we may ask whether every study in the data set shows an effect in the expected direction or not. We develop a set of three models: A general model much like the conventional meta-analytic random-effects model, a positive-effects model, and a null model. To assess the relative strength of evidence between these models, we propose a Bayes factor model comparison approach adapted from Haaf and Rouder (2017) and Rouder et al. (submitted). Critically, the new approach can handle the data availability problem by modeling the collection of the studies' Fisher's Z values. We apply the approach to a simulated data set and find comparable results to Rouder et al.'s (submitted) approach for modeling raw data. Additionally, we apply the approach to Anderson and colleagues' (2010) meta-analysis on violent video games and aggression, a subset of this meta-analysis where Hilgard et al. (2017) suspect publication bias, and a subset of unpublished studies where publication bias is unlikely. The results seem well-calibrated and reasonable.

Limitations and Future Directions. A main limitation of the current approach is that it does not allow for the correction of publication bias. Publication bias and questionable research practices are intimately tied to the question of the usefulness of meta-analysis (Carter, Schönbrodt, Hilgard, & Gervais, 2017; Corker, 2018). If the studies in the meta-analytic set are heavily biased, using them to learn about the population is difficult. A number of more or less successful corrections for publication bias have been proposed (Carter et al., 2017). One possible extension of the current modeling approach is to add publication bias correction measures. In principle, this extension should not be too difficult. Yet, assessing the success of these corrections can be problematic, as neither the process of study censorship nor the amount of publication bias are known in any real meta-analytic set (Guan & Vandekerckhove, 2016). There is an additional problem with adding publication bias correction to the current modeling approach. To illustrate the problem, consider the right panel in the first row of Figure 3. If all studies indeed have a true positive effect, then the resulting distribution of observed effects should be similar to the slightly skewed, mostly positive distribution. However, the distribution that is assumed by most publication bias correction methods is the on resulting from the general model, which is symmetric, normal-shaped. Most, if not all correction methods use a skewed distribution as an indicator for publication bias. It is unclear how to resolve this issue.

Another limitation comes from the prior specification. The hyperpriors on ν and η^2 are critical for the Bayes factor. Yet, setting them is difficult without any experience of the typical scaling of Z_i. As of now, we are not sure what size of the overall mean and variance of θ_i may be a reasonable expectation, and whether these expectations should substantially differ across applications. To assess the variability of Bayes factors based on reasonable prior settings, it is useful to conduct a sensitivity analysis (e.g. Haaf & Rouder, 2017; Heycke, Gehrmann, Haaf, & Stahl, 2018). We did so for the simulated data set in Application 1. Here, we assessed the effect of prior settings we deemed reasonable on the Bayes factors. Fortunately, all of the settings lead to the same order of preference for the models, and the sizes of the Bayes factors only varied moderately.

In addition, with every model development come model assumptions that need to be taken into account. An issue with Rouder and colleagues' (submitted) development was the underlying assumption of homogeneity-of-variance across the studies in the meta-analysis. For many-labs studies as the ones analyzed by Rouder et al. (submitted), this assumption may be reasonable. For conventional meta-analysis, however, the homogeneity-of-variance assumption is difficult where studies with vary in scale and design. This assumption is not needed with the current approach. Using Fisher's Z as target of the models removes this necessity. However, modeling Fisher's Z introduces the assumption that the observed within-study variance is the true variance. With larger sample sizes, however, the analysis may be robust against violations of this assumption. In the future, an examination of the effect of violations of this assumption may be useful

Conclusion. In summary, the current approach allows to investigate questions of ordinal constraint in a meta-analytic setting. We think that answering the “does every study” question is important. If, indeed, every study shows a *true* effect in the same, expected direction, developing or maintaining an underlying theory for the target phenomenon seems reasonable. If, however, the ordinal constraint is violated and some studies indeed show true negative effects, researchers need to investigate the underlying mechanism that lead to opposite effects. These mechanisms may be found in design or sample considerations.

The current development also shows how limited the options for meta-analysis are due to the problem of data availability. It is much easier to learn a lot more from a meta-analysis if all the data are provided. Just providing condition means and standard deviations in an article enables meta-analysts to answer more interesting and more complicated questions.

Of course, there are many other possible expansions of the approach. A first example may be the addition of a mixture model to assess whether some studies truly do show a true effect in the expected direction while others truly show no effect. Such a mixture model would be useful in cases when the investigated effect is sample- or design-specific. Second, we may want to explore the possibility of modeling Cohen’s d for experimental data sets instead of Fisher’s Z. Modeling Cohen’s d may lead to more accessible prior settings for experimental researchers than Fisher’s Z-based models. Third, the current models could be extended to allow for the inclusion of covariates. Testing the influence of covariates is a common goal in meta-analysis. All these paths show that developing flexible and reasonable modeling approaches in meta-analysis remains timely and topical.

Anderson, C. A., Shibuya, A., Ihori, N., Swing, E. L., Bushman, B. J., Sakamoto, A., . . . Saleem, M. (2010). Violent video game effects on aggression, empathy, and prosocial behavior in eastern and western countries: A meta-analytic review. *Psychological Bulletin*, *136*(2), 151–173. Retrieved from http://psycnet.apa.org/doi/10.1037/a0018251

Borenstein, M., Hedges, L. V., Higgins, J., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, *1*(2), 97–111.

Carter, E. C., Schönbrodt, F. D., Hilgard, J., & Gervais, W. M. (2017). *Correcting for bias in psychology: A comparison of meta-analytic methods.* Retrieved from https://osf.io/preprints/psyarxiv/9h3nu/

Corker, K. S. (2018). *Strengths and weaknesses of meta-analyses.*

Corker, K. S., Donnellan, M. B., Kim, S. Y., Schwartz, S. J., & Zamboanga, B. L. (2017). College student samples are not always equivalent: The magnitude of personality differences across colleges and universities. *Journal of Personality*, *85*(2), 123–135.

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., … Nosek, B. A. (2016). Many labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68–82. Retrieved from http://ezid.cdlib.org/id/doi:10.17605/OSF.IO/QGJM5

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

Ferguson, T. S. (1996). *A course in large sample theory.* Chapman & Hall Ltd.

Guan, M., & Vandekerckhove, J. (2016). A Bayesian approach to mitigation of publication bias. *Psychonomic Bulletin and Review*, *23*(1), 74–86. Retrieved from http://www.cidlab.com/prints/guan2015bayesian.pdf

Haaf, J. M., & Rouder, J. N. (2017). Developing constraint in Bayesian mixed models.

Psychological Methods, *22*(4), 779–798.

Haaf, J. M., Klaassen, F., & Rouder, J. N. (In preparation). *Using systems of orders to capture theoretical constraint in psychological science.*

Heycke, T., Gehrmann, S., Haaf, J. M., & Stahl, C. (2018). Of two minds or one? A registered replication of rydell et al.(2006). *Cognition and Emotion*, *0*(0), 1–20.

Hilgard, J., Engelhardt, C. R., & Rouder, J. N. (2017). Overstated evidence for short-term effects of violent games on affect and behavior: A reanalysis of anderson et al. (2010). *Psychological Bulletin*, *143*, 757–774.

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.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795. Retrieved from http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1995.10476572

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.

Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, *72*, 6—18. Retrieved from http://www.sciencedirect.com/science/article/pii/S0022249615000723

Rouder, J. N., Haaf, J. M., & Aust, F. (2018). From theories to models to predictions: A Bayesian model comparison approach. *Communication Monographs*, *85*, 41–56. Retrieved from https://doi.org/10.1080/03637751.2017.1394581

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., 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

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

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., R. B. Adams, J., … Zwaan, R. A. (2016). Registered replication report: Strack, martin, & stepper (1988). *Perspectives on Psychological Science*, *11*(6), 917–928. Retrieved from https://doi.org/10.1177/1745691616674458

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

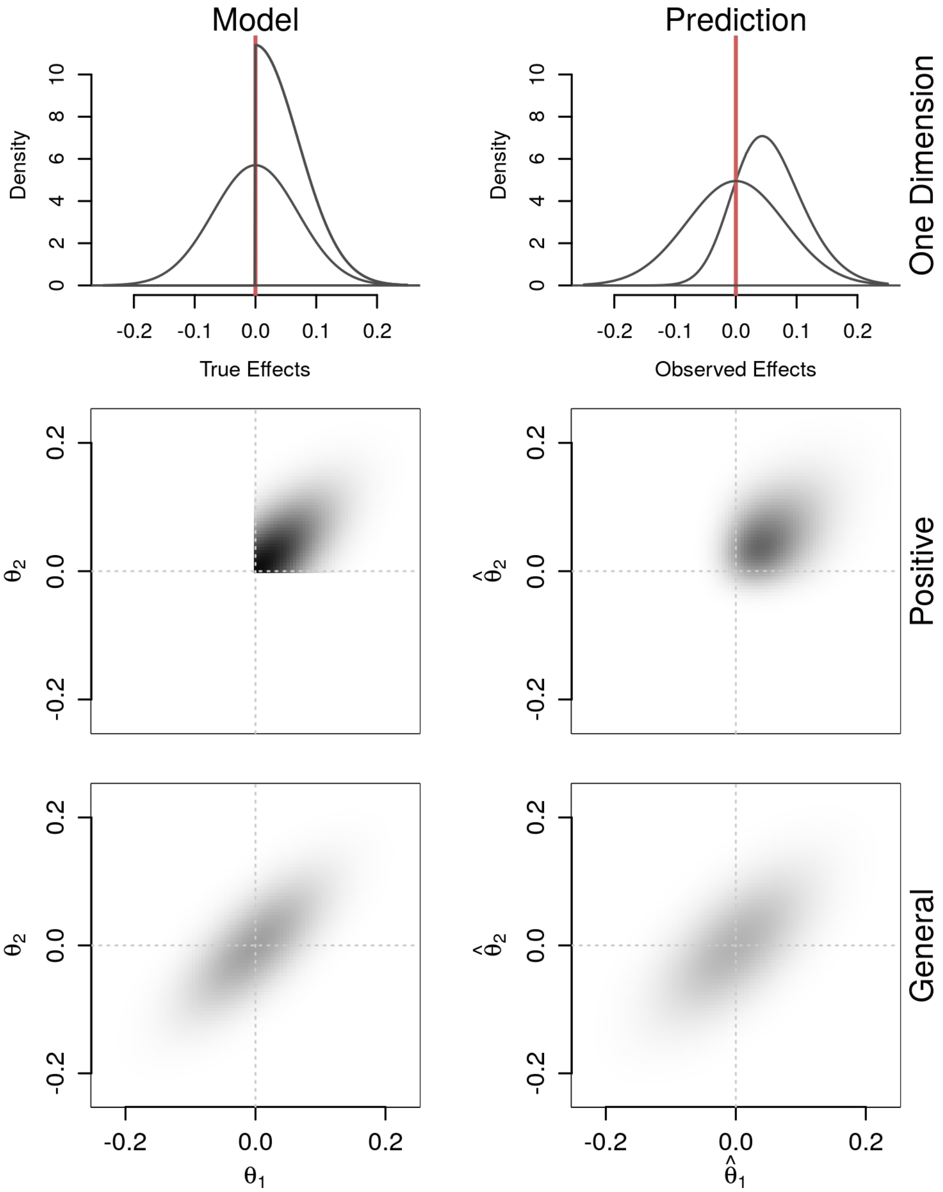

Figure 3. Model specification and predictions for the positive-effects model and the general model. The models are on the left hand-side, and the predictions for data are on the right hand-side. Top row: Models for one study. Even though the positive-effects model is restricted to positive values in specification, it can predict small negative observed values. Middle row: Positive-effects model for two studies. Effects for any two studies are predicted to be correlated due to the hiearchical nature of the model, and mostly positive. Bottom row: General model for two studies. Effects are still correlated, but may be positive or negative. The predictions for any specific effect combination is weaker for the general model as is covers more parameter space.

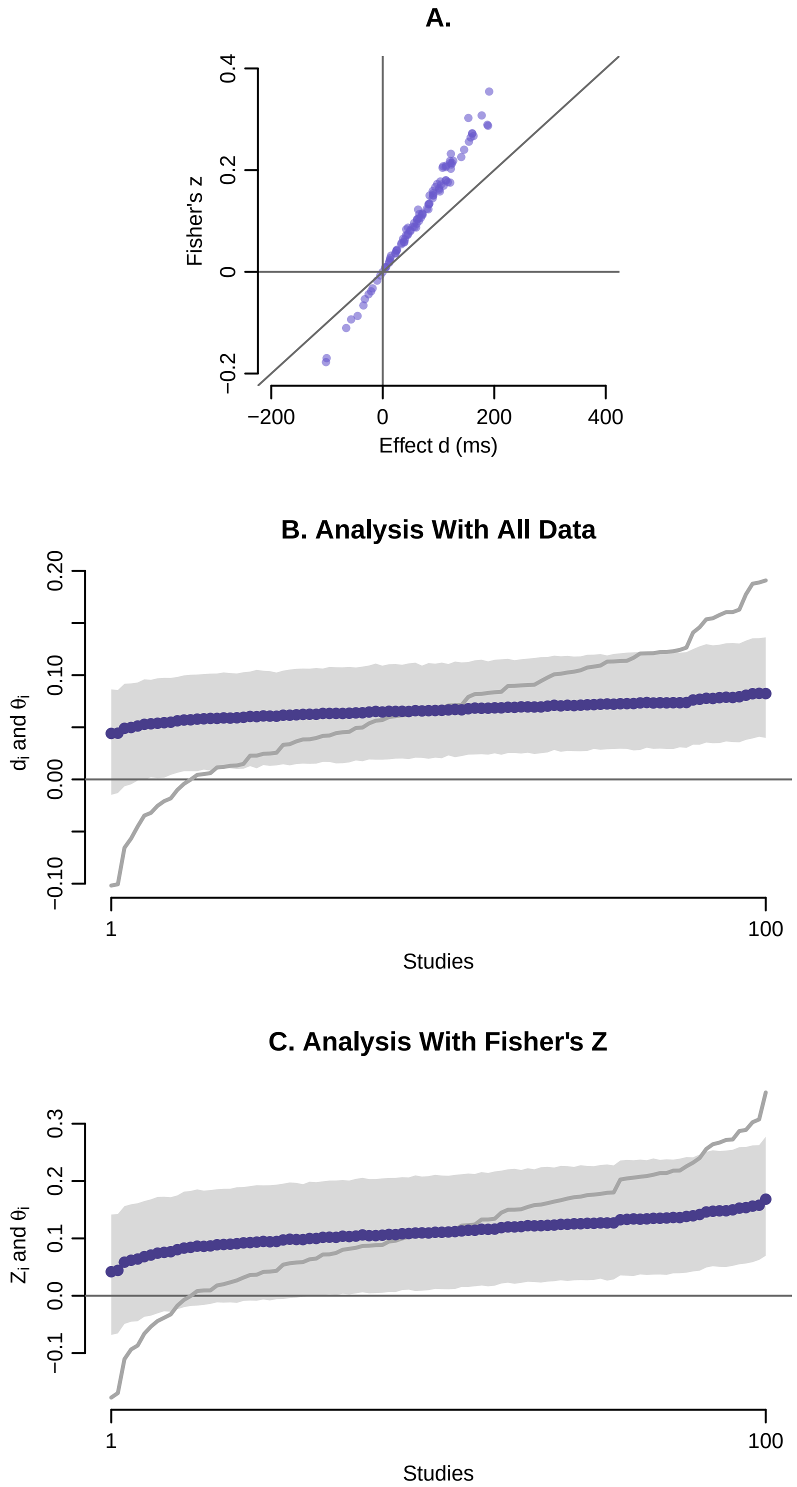

Figure 4. Results for the simulated data set. A. Fisher's z as a function of the observed effect. Importantly, the direction of the effect is preserved. B. Estimation results from the raw-data analysis (Rouder et al., submitted). Observed effects are depicted by the gray line, posterior estimates are given by the points, credible intervals are the gray areas around the points. There is some hierarchical shrinkage to the overall effect. C. Estimation results from the Fisher's Z analysis. The Z values have a larger range than the observed effects. There is again some hierarchical shrinkage so that all posterior estimates of θ_i are postitive in value.

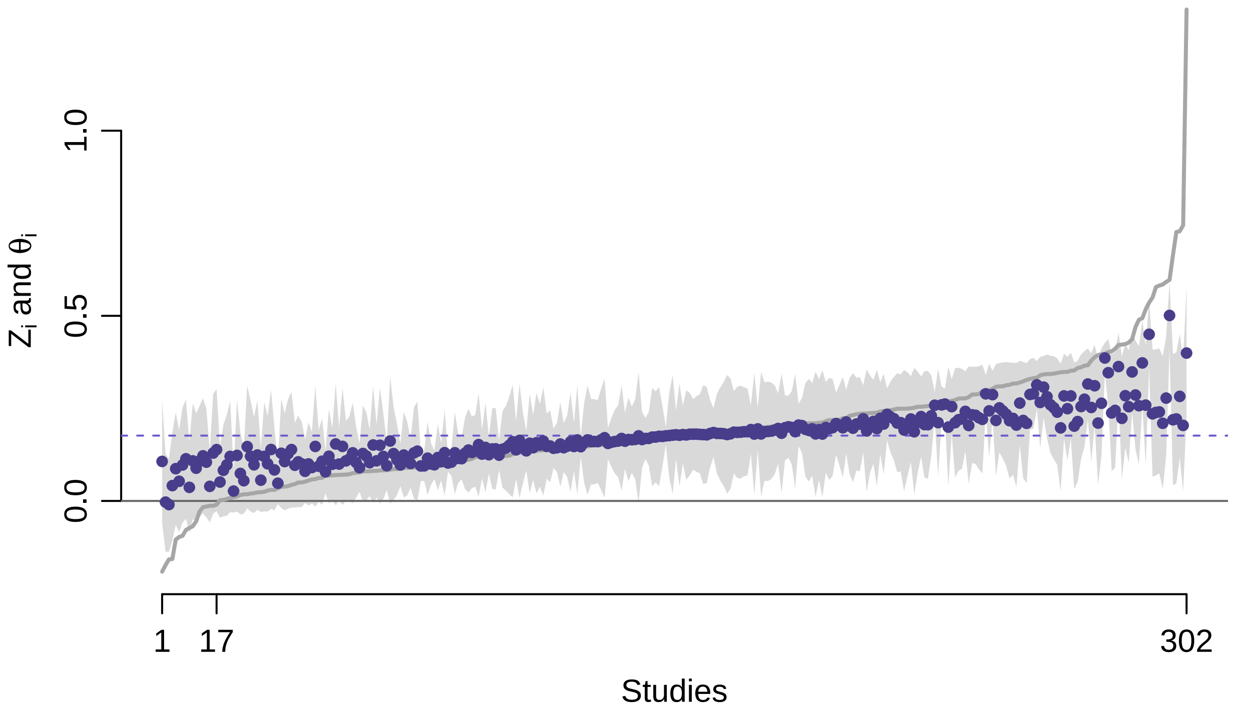

Figure 5. Model estimation for the Anderson et al. (2010) meta-analysis. Observed Fisher's Z values are given by the gray line. Posterior estimates of θ_i are depicted by the points, credible intervals are depicted by the gray area. The posterior estimate of the overall effect is given by the dashed line. The estimates are shrunken to the posterior mean, and they are very variable. The amount of shrinkage is determines by the sample size.

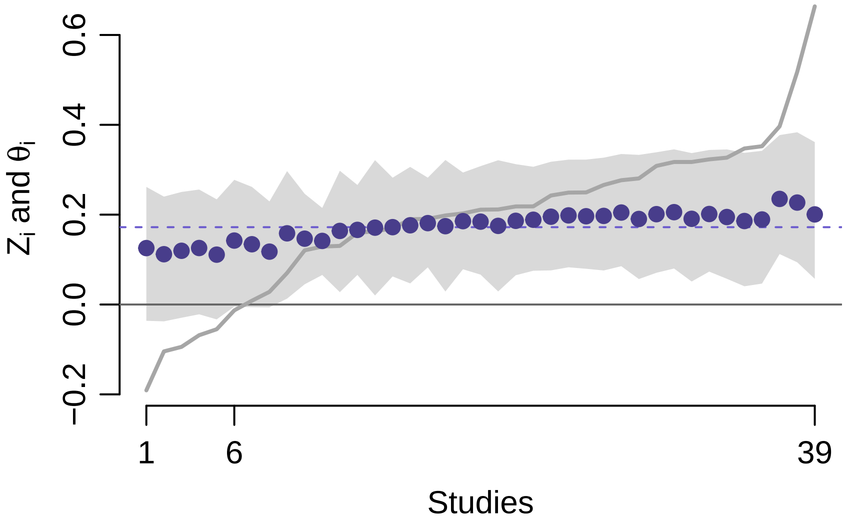

Figure 6. Results for the Anderson et al. (2010) subset of experimental data with aggressive behavior as dependent variable. Observed Fisher's Z values are given by the gray line. Posterior estimates of θ_i are depicted by the points, credible intervals are depicted by the gray area. The posterior estimate of the overall effect is given by the dashed line. All posterior estimates of θ_i are above zero, even for the observed negative Z values.

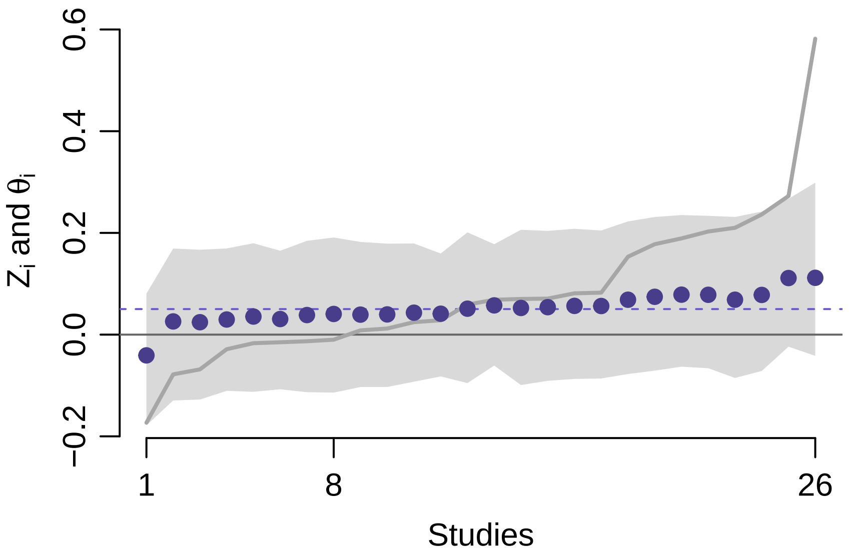

Figure 7. Results for the Anderson et al. (2010) subset of experimental data from unpublished studies. Observed Fisher's Z values are given by the gray line. Posterior estimates of θ_i are depicted by the points, credible intervals are depicted by the gray area. There is a large amount of hiearchical shrinkage to the posterior estimate of the overall effect is given by the dashed line.