Research Master's programme Methodology and Statistics for the Behavioural, Biomedical and Social Sciences
Utrecht University, The Netherlands

MSc Thesis Anita Encheva Lyubenova[1] (3384022)
TITLE: Bayesian Evidence Synthesis: The Value of the Unconstrained Hypothesis
May 2023

Supervisor:
Prof. Dr. Irene Klugkist

Second grader:
Dr. Peter van de Ven

Preferred journal of publication**:** Methodology
Word count[2]: 6328

[1] ChatGPT was used to improve the quality of the text.
[2] Including references, appendices, tables, and figures as specified in the journal guidelines.

Abstract:

Bayesian evidence synthesis (BES) is a novel synthesis method that allows for aggregation of methodologically diverse studies that cannot be aggregated with traditional methods. However, when testing an informative hypothesis Hi against its complement Hc, BES provides inconsistent results in the presence of heterogeneity. This paper proposes conjoint testing, i.e., testing Hi, Hc, and the unconstrained hypothesis Hu simultaneously, as a potential solution for this issue, where Hu should be most supported in the presence of substantial heterogeneity. The performance of conjoint testing in BES is investigated in a simulation study for varying sample sizes, number of aggregated studies, and degrees of heterogeneity. The findings indicate that conjoint testing can overcome inconsistencies caused by testing against Hc, but 20 to 30 studies are required for reliable detection of Hu as the true hypothesis. Therefore, study sets with less than 20 studies may be underpowered to detect heterogeneity through conjoint testing.

Word count: 150

Keywords: Bayesian evidence synthesis, heterogeneity, posterior model probabilities, conceptual replications

Introduction

The replication crisis has brought attention to the need for exact replications (Open Science Collaboration, 2015, 2017; Schweinsberg et al., 2016). Additionally, it has stimulated a discussion on the significance of conceptual replications for building confidence in theory (Crandall & Sherman, 2016; Derksen & Morawski, 2022; Stroebe & Strack, 2014; Zwaan et al., 2017). Conceptual replications involve testing the same effect using different, theoretically meaningful methodologies. These replications assess the resilience of a theoretical proposition to alternative research designs, operational definitions, and samples. By doing so, they offer insights into the generalizability of the theory across different ways of operationalizing key constructs and different populations (Zwaan et al., 2017). To draw conclusions from a set of conceptual replications, evidence synthesis methods, such as Bayesian or traditional meta-analysis, are often applied. However, methodological heterogeneity, such as differences in measurement scales and research design, might hinder aggregation with traditional methods (Kuiper et al., 2013).

Bayesian evidence synthesis (BES) is a method for synthesizing evidence from diverse studies with varying methodologies (Kuiper et al., 2013). Rather than pooling effect sizes or raw data, BES combines the support for a hypothesis across studies. To achieve this, first, Bayes factors and posterior model probabilities (BFs and PMPs; Kass & Raftery, 1995) are used to quantify the strength of evidence for each hypothesis in each study. Then, BES integrates this evidence to determine the overall support across all studies. The final aggregated BFs or PMPs indicate the degree to which each tested hypothesis is supported in each study (Klugkist & Volker, 2022), thereby indicating which hypothesis is relatively most likely to be true in the population.

As BES is based on Bayesian hypothesis testing, it offers flexibility in the choice of hypotheses and allows researchers to test theoretically derived (informative) hypotheses (Baig, 2020; Klugkist et al., 2011; Kluytmans et al., 2012; Mulder & Olsson-Collentine, 2019; van de Schoot & Strohmeier, 2012; Vanbrabant et al., 2014). If one informative hypothesis Hi is of interest (e.g., H1: $\mu_1>\mu_2>\mu_3$), common alternative hypotheses are the unconstrained hypothesis (Hu) that

imposes no restriction on the parameters and the complement hypothesis Hc: “not Hi” (Hoijtink, 2011). Complement testing (Hi vs Hc) has been preferred over simple unconstrained testing (Hi vs Hu) in individual studies for three reasons. First, complement testing has a straightforward interpretation because it directly addresses the question of whether a hypothesis or theory is correct or not, which is often of primary interest in applied research (van de Schoot et al., 2012; van Rossum et al., 2013). Second, Hi and Hc are mutually exclusive. Thus, the sample sizes required to differentiate Hi from Hc will be smaller than those needed to distinguish Hi from Hu since Hi is nested within Hu (Hoijtink, 2013). Third, the BF from complement testing can become infinitely large in favor of either hypothesis as the amount of evidence for this hypothesis increases (Hoijtink, 2011). On the contrary, in a simple unconstrained test, the BF has maximum value, which makes it problematic to interpret it as a measure of strength of evidence (Huisman, 2022).

Despite its advantages, complement testing becomes challenging when BES is used to aggregate results across multiple studies. It relies on the assumption that among the tested hypotheses there is (at least) one that is commonly true for all studies. This assumption would be violated, for example, if for some studies H1 is true, while for others Hc is true. In such cases BES will indicate which hypothesis is relatively most likely in “the population” when there is, in fact, no common population for all studies. The problem has been illustrated in a simulation study by van Wonderen (2022). The hypothesis regarding a regression coefficient H1: $\beta_1>0$ was tested against Hc: $\beta_1<0$, while for some studies H1 was true, and for others Hc was true. When aggregating over the study set, H1 was considered to be most likely in the population 50% of the iterations while in the remaining 50% Hc was more supported (van Wonderen, 2022). In consequence, a conclusion drawn from a single iteration would be based on chance in such setting. This issue makes the interpretation of the results from BES ambiguous and might lead to erroneous inferences.

In reality, the assumption of a common true hypothesis might be unrealistic. A large survey of 150 psychological meta-analyses (Linden & Hönekopp, 2021) found high levels of

unexplainable heterogeneity. Thus, it is likely that a study set contains studies with different true hypotheses, invalidating the assumption of a common true hypothesis.

One potential solution would be to test Hu along with H1 and Hc (from now on referred to as *conjoint testing*). Because Hu is always true, conjoint testing ensures that there is at least one common true hypothesis for all studies. Thus, Hu would be preferred if neither Hi, nor Hc is true for all studies. Conversely, if Hi (or Hc) is true for all studies (and they are sufficiently powered), it would be most supported because each is more parsimonious than Hu (Volker, 2022). This is the case, because BES balances the fit of a hypothesis, i.e., the degree to which the data supports it, and its complexity, i.e., the degree to which it constrains the parameter space (with less constraints meaning higher complexity). Hu has a perfect fit but is the most complex hypothesis. Thus, a hypothesis with a lower complexity would be preferred, as long as its fit is good enough. Therefore, with conjoint testing the question we ask is "Which is the *most parsimonious common true hypothesis* (MPCTH) for all studies?".

The aim of this study was to scrutinize the utility of conjoint testing with BES by assessing its performance under realistic conditions and comparing it to alternative tests. First, we examined whether conjoint testing could reconcile inconsistent results arising from complement testing in heterogeneous study populations. Particularly, we compared how the PMPs for Hi and Hc were distributed with and without testing Hu, across populations with different heterogeneity levels. The second objective of our study was to investigate the performance of BES and conjoint testing for a range of sample sizes and number of aggregated studies. To this end, we assessed the probability of obtaining the most support for the MPCTH across different conditions. Third, we compared the performance of conjoint testing with simple unconstrained testing.

The paper is structured as follows: the next section outlines the implementation of BES. This is followed by an explanation of the simulation procedure and the data analysis. The results of the simulation study are then presented. Finally, the Discussion section reflects on the findings and provides advice for applied researchers.

Bayesian Evidence Synthesis

BES uses BFs to quantify the support for each hypothesis of interest relative to an alternative hypothesis. The main building block of an informative hypothesis test is the BF of a simple unconstrained test (Hi vs. Hu), denoted by BF_{iu} (Klugkist et al., 2005). BF_{iu} can be computed as a function of the fit (f_i) and the complexity (c_i) of Hi:

$$BF_{iu} = \frac{f_i}{c_i},$$

Where the fit quantifies the degree to which the data supports Hi, and the complexity quantifies the degree to which Hi constrains the parameter space. Note that the fit and the complexity of Hu are both equal to 1, therefore BF_{uu} for Hu is also always 1. Posterior model probabilities (PMPs) for each hypothesis can be used to compare the relative support across multiple hypotheses in each study. In individual studies the PMPs for each Hi can be computed as

$$PMP(Hi) = \frac{\pi_i^0 BF_{iu}}{\Sigma_{i'=1}^{m} \pi_{i'}^0 BF_{i'u}}$$

where π_i^0 is the prior model probability for hypothesis Hi, which indicates how likely the hypothesis is before looking at the study data; i' = 1, 2, …, m, where m is the total number of hypotheses. The higher the value of PMP(Hi), the more plausible it is relative to the remaining hypotheses given the data.

BES uses the prior model probabilities π_i^0 to inform the PMPs for each hypothesis from one study to the next (Kuiper et al., 2013). The initial prior model probabilities (before looking at the data from the first study) are commonly fixed to be equal for all hypotheses (Hoijtink, 2011). The PMPs resulting from the first study can be denoted by π_i^1 and are then used as prior model probabilities in the second study. Then, PMPs resulting from the second study (π_i^2) are used as prior model probabilities in the third study, and so on. The PMPs of Hi after aggregating T studies are denoted by π_i^T and are computed as follows:

$$\pi_i^T = \frac{\prod_{t=1}^{T} BF_{iu}^t}{\Sigma_{i'=1}^{m} \prod_{t=1}^{T} BF_{i'u}^t}$$

where BF_{iu}^t is the BF_{iu} from study t=1, 2, ..., T. Because of the multiplicative nature of the aggregation, the order in which the studies are aggregated is not relevant to the final value (Kuiper et al., 2013).

Methods

To address the aims of this study a simulation was performed. Data generation and analyses were performed in R (version 4.2.0; R Core Team, 2022). All materials to reproduce the results can be found on GitHub (https://github.com/anita-lyubenova/MasterThesis_BES). The study was approved by the Ethical Review Board of the Faculty of Social and Behavioural Sciences of Utrecht University (protocol number 22-1864).

Simulation study

Study sets were simulated to consist of t individual studies, where t = 1, ..., 30. Individual studies' data was generated using a multivariate ordinary least squares (OLS) regression model. A continuous outcome variable Y was regressed on 3 standard normally distributed predictor variables. We considered three hypotheses about the ordering of the regression coefficients: H1: $\beta_1 > \beta_2 > \beta_3$, Hc: “not H1”, Hu: $\{\beta_1, \beta_2, \beta_3\}$. The population (and the true hypothesis) of an individual study was determined by the ratios between the coefficients $\beta_1 : \beta_2 : \beta_3$, where larger ratio values corresponded to larger coefficients. The ratios of the individual studies were not directly manipulated; instead, to simulate heterogeneity within the study set, they were sampled from a log-normal distribution. The distribution had an arithmetic mean μ_k (k=1,2,3) and arithmetic standard deviation $\sigma_k = \mu_k cv$, where cv was the coefficient of variation (Lovie, 2005; for details on the log-normal distribution, see Appendix A). The sampled study-level ratio values were used

Table 1. Populations of the study sets.

Population	MPCTH	$\mu_1{:}\mu_2{:}\mu_3$	cv	(Average) proportion of studies that do not originate from H1[1]
1. Fixed H1	H1	3:2:1	.00[1]	.00
2. Fixed Hc	Hc	1:2:3	.00	1.00
3. mixed H1/Hc population (50% fixed H1 and 50% fixed Hc)	Hu	3:2:1 and 1:2:3	.00 and .00	.50
4. Heterogeneous H1 (average cv)	Hu	3:2:1	.86	.43
5. Heterogeneous H1 (small cv)	Hu	3:2:1	.50	.59

Hypotheses H1: $\beta_1 > \beta_2 > \beta_3$, Hc: "not H1", Hu: $\{\beta_1, \beta_2, \beta_3\}$

to compute population regression coefficients for each study. Then, outcome values Y were sampled from a normal distribution determined by the coefficients (Fu, 2022). The considered sample sizes were n ∈ {25,50,75,100,150,200,300,500} and were based on common sample sizes in existing meta-analyses (Appendix B, Linden & Hönekopp, 2021). The correlation between the predictors ρ was fixed to .30, and R^2 was fixed to .13, both representing medium effect sizes (Cohen, 1988).

In this study, cv was varied to achieve different levels of heterogeneity for two reasons. First, multiple investigations reported that large effects had higher heterogeneity than small effects (Klein et al., 2018; Linden & Hönekopp, 2021; Olsson-Collentine et al., 2020). Varying cv guaranteed that larger μ_k had larger standard deviation than small μ_k, mimicking empirical patterns. Second, because cv is unitless (Lovie, 2005), it was possible to determine realistic values for cv from existing meta-analyses irrespective of the scale units. The considered cv values were 0, .50, and .86 (based on Linden & Hönekopp, 2021, Appendix B)

The populations of the study sets (determined by μ_k and cv) were specified such that H1, Hc or Hu was the MPCTH. Table 1 presents the 5 populations discussed in this paper: the first two populations represented fixed effects populations for which H1 and Hc was true, respectively; the third population is a mix of the first two, where H1 was true for half of the studies, and Hc for

the other half. The fourth population was a heterogeneous H1-population with the average cv; the fifth population was a heterogeneous H1-population with the a common lower-than-average cv. In total, there were 5 populations x 8 sample sizes x 30 study set sizes = 1200 conditions. Each of them was iterated 1000 times.

Analytical Strategy

For each generated study, a Bayesian hypothesis test was performed for 3 sets of hypotheses– H1 vs Hu, complement testing (H1 vs Hc), and conjoint testing (H1 vs Hc vs Hu). For each of the tested hypotheses a BF was computed with the R package BFpack (Mulder et al., 2021). BFpack uses default Bayes factors that do not necessitate the incorporation of prior (subjective) knowledge. Instead, a minimal fraction of the information in the data is used to construct a fractional Cauchy prior, while the remaining fraction is used for hypothesis testing (for a detailed account of the procedure, see Mulder et al., 2021). The fractional prior is placed at the boundary of the constrained space to adhere to the rationale that small effects are more likely a priori than large effects and that positive and negative effects are equally plausible (Mulder, 2014; Mulder & Olsson-Collentine, 2019). Once the BFs were obtained, they were aggregated cumulatively for 30 studies via BES (with equal initial prior model probabilities). This resulted in aggregated PMPs for each hypothesis of interest after aggregating 1 through 30 studies.

First, we investigated whether testing along Hu resolves inconsistencies arising from complement testing in populations in which the assumption of a common true hypothesis is violated for complement testing: mixed H1/Hc population, heterogeneous H1 with c_v = .86, and heterogeneous H1 with cv = .50. The distributions of aggregate PMPs were compared across complement, simple unconstrained, and conjoint testing.

Second, we investigated the performance of conjoint testing across a range of sample sizes, study set sizes, and heterogeneity levels. The performance was measured as the probability of the test to support the true hypothesis, which is similar to the concept of statistical power in null

hypothesis testing. It has been previously applied by Kuiper et al. (2015) and from now on it would be referred to as true positive rate (TPR).

To obtain an overall measure of the performance of the test across the three populations, the average of the TPRs per condition was computed. This is equivalent to the concept of accuracy in classification problems, where the number of correct classifications is divided by the total number of classifications. Accuracy of 0.87 was considered to be sufficient test accuracy, because of an analogy to traditional null hypothesis testing. There, $\alpha = .05$, and power of .80 provides accuracy of (.95 + .80)/2 = .87 (given equal number of studies from the null and alternative populations. Other useful benchmarks for the interpretation of TPR and accuracy are: 0 = systematic decision against the true hypothesis; 0.5 = conclusion based chance; 0.7 = insufficient accuracy, 0.8 = almost sufficient; 0.87 = sufficient; and 1 = perfect accuracy.

Conjoint testing was compared to simple unconstrained testing by taking the difference in accuracies. This would indicate the extent to which the performance of a conjoint test is different than of a simple unconstrained test under the same conditions.

Results

Complement vs conjoint testing

The first objective of this study was to compare the distribution of aggregate PMPs for each hypothesis between complement testing and testing that includes Hu (both Hi vs Hu and conjoint testing)[3]. Figure 1A illustrates the results for a mixed H1/Hc-population, highlighting the issue of complement testing. The median PMP for Hc approached 1 after aggregating only 4 studies, even though only 50% of the studies originated from Hc. When Hu was included (Figure 1B and 1C), the support for Hc decreased in the first 10 studies, both for simple unconstrained test and conjoint

[3] Figures 1 and 2 show only the aggregation over 15 studies because adding additional studies did not contribute any additional insights.

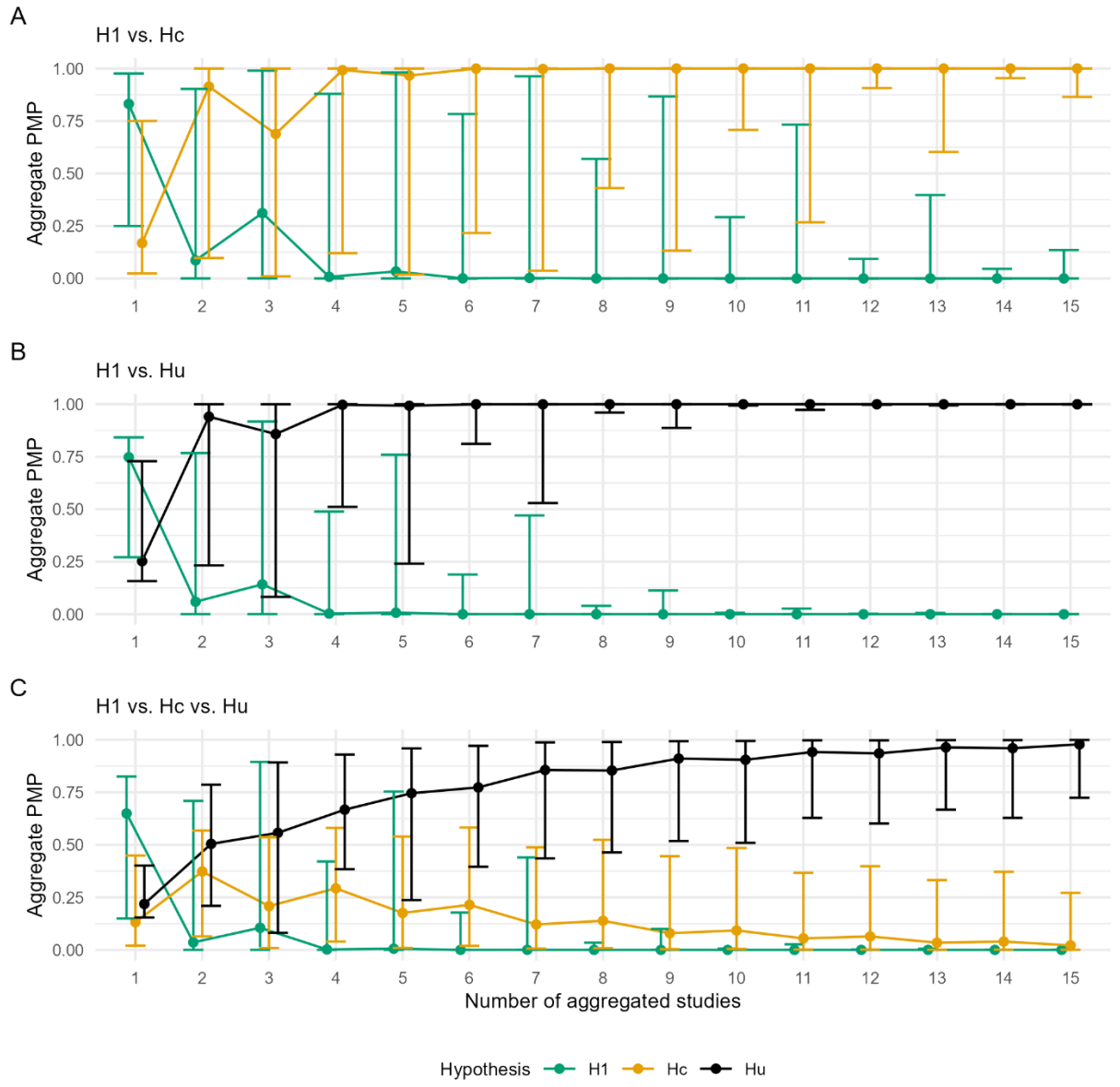

Figure 1. Aggregate posterior model probabilities (PMP) for each hypothesis for increasing number of aggregated studies in a mixed H1/Hc population. Points indicate the median PMP per hypothesis; the bars indicate the interval from the 2.5th and 97.5th percentile of the PMPs and contain 95% of the values; the hypotheses of interest are H1: β1 > β2 > β3, Hc: "not H1", and Hu: {β1, β2, β3}.

testing. Meanwhile, the support for Hu, the only common true hypothesis, increased over the studies.

In the heterogeneous H1-population with cv =.86, complement testing showed indecision between hypotheses as shown in Figure 2A. The PMPs for H1 were only slightly higher than those for Hc, and their values ranged from 0 to 1. When tested against Hu (Figures 2B and 2C), the

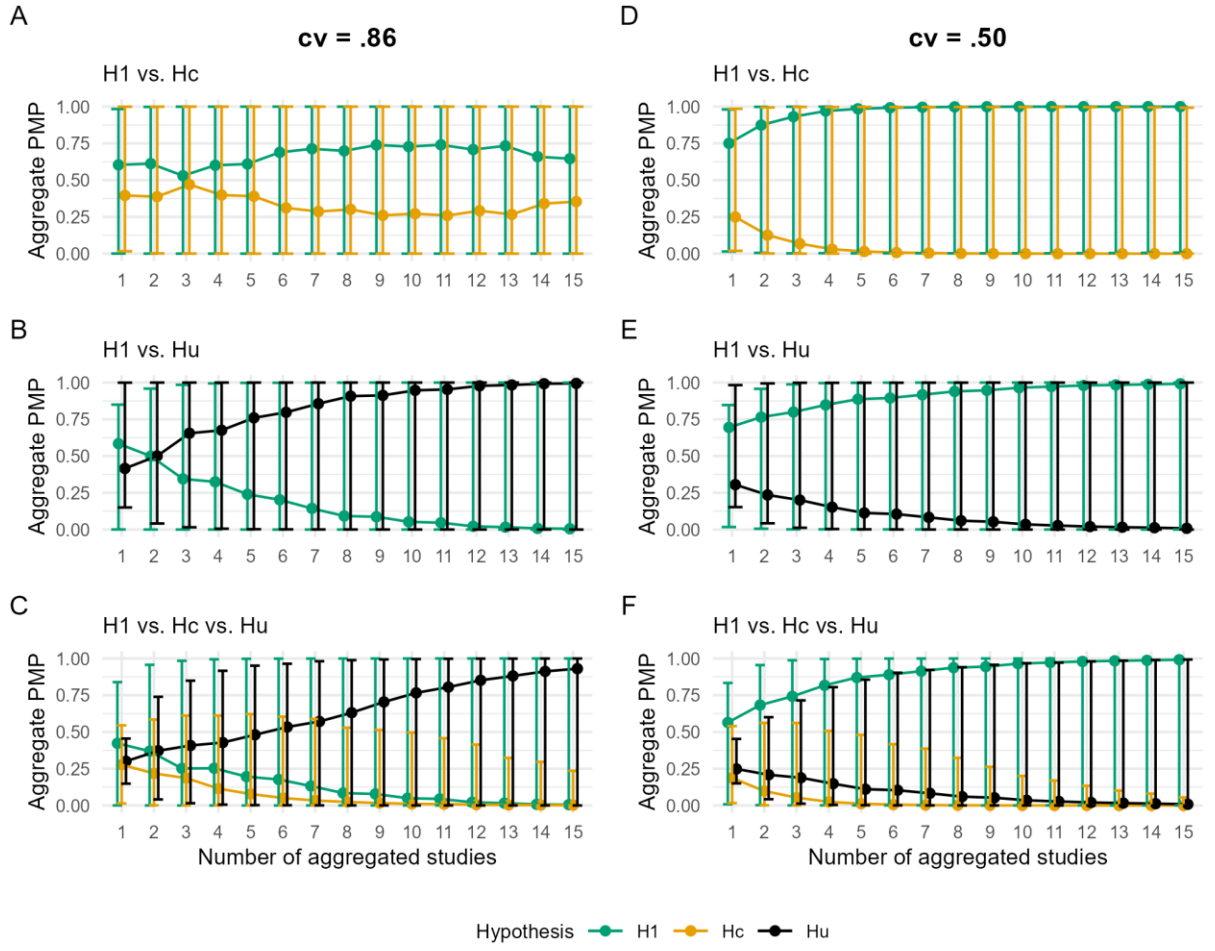

Figure 2. Aggregate posterior model probabilities (PMP) for each hypothesis for increasing number of aggregated studies originating from a heterogeneous H1 population with cv=.86 (A, B, C) and cv=.50 (D, E, F). Points indicate the median PMP per hypothesis; the bars indicate the interval from the 2.5th and 97.5th percentile of the PMPs and contain 95% of the values; the hypotheses of interest are H1: $\beta_1 > \beta_2 > \beta_3$, Hc: "not H1", and Hu: $\{\beta_1, \beta_2, \beta_3\}$.

results were similar to the previous condition: the support for Hu increased compared to H1 and Hc. However, the PMPs for Hu and H1 had wider distributions and could take on any value between 0 and 1. Despite this, for t=15 the median for the MPCTH (Hu) approached 1, indicating that 50% of the PMPs for Hu were approximately 1. Moreover, in conjoint testing 70% of the PMPs for Hu were above 0.5, while this was the case for 29% of the PMPs for H1 (again for t=15).

In the heterogeneous H1-population with cv =.50, complement testing resulted in PMPs supporting H1 over Hc (Figure 2D). The median PMP of H1 was approx. 1 after aggregating 4 studies, indicating that 50% of the PMPs were 1. Additionally, in this case 78% of the PMPs of H1

were larger than 0.5, indicating a strong skew and much support for H1 in complement testing. H1 also obtained more support than the alternative hypotheses both in simple unconstrained and conjoint testing (Figures 2E and 2F). For comparison, for t=4 in conjoint testing, the PMP distribution of H1 was less skewed, as the median was .81, and 67% of the PMPs were larger than 0.5.

True positive rates

To evaluate the probability of correctly identifying the MPCTH in conjoint testing, we examined the true positive rates (TPR) when each of H1, Hc, and Hu was the MPCTH (Figure 3). Figure 3A presents the results from a H1-population, which revealed an interaction effect between the sample size and the number of aggregated studies. Specifically, the support for H1 increased as more studies were aggregated, but only when the individual studies had large sample sizes. When the individual studies had small sample sizes, the support for H1 either remained unchanged (for n=50) or decreased (for n=25). Moreover, fewer studies to achieve a high TPRs when aggregating over larger studies.

On the other hand, when Hc was the true hypothesis, the support for Hc consistently increased with each sample size, as shown in Figure 3B. Even with a small number of aggregated studies (less than 10), the error probability was substantially reduced. Only a sample size of 25 resulted in a slower increase of TPR with increasing number of studies. A comparison between Figure 3A and 3B reveal different TPR patterns between H1 and Hc populations,particularly when aggregating small or few studies. This suggest that Hc had higher power than H1 in conjoint testing.

Figure 3C illustrates that the probability of correctly identifying Hu as the MPCTH was generally low in a heterogeneous H1-population with cv = .86 , and increased slowly with the number of studies. For small numbers of studies (<10), the TPRs were largely below .60, making

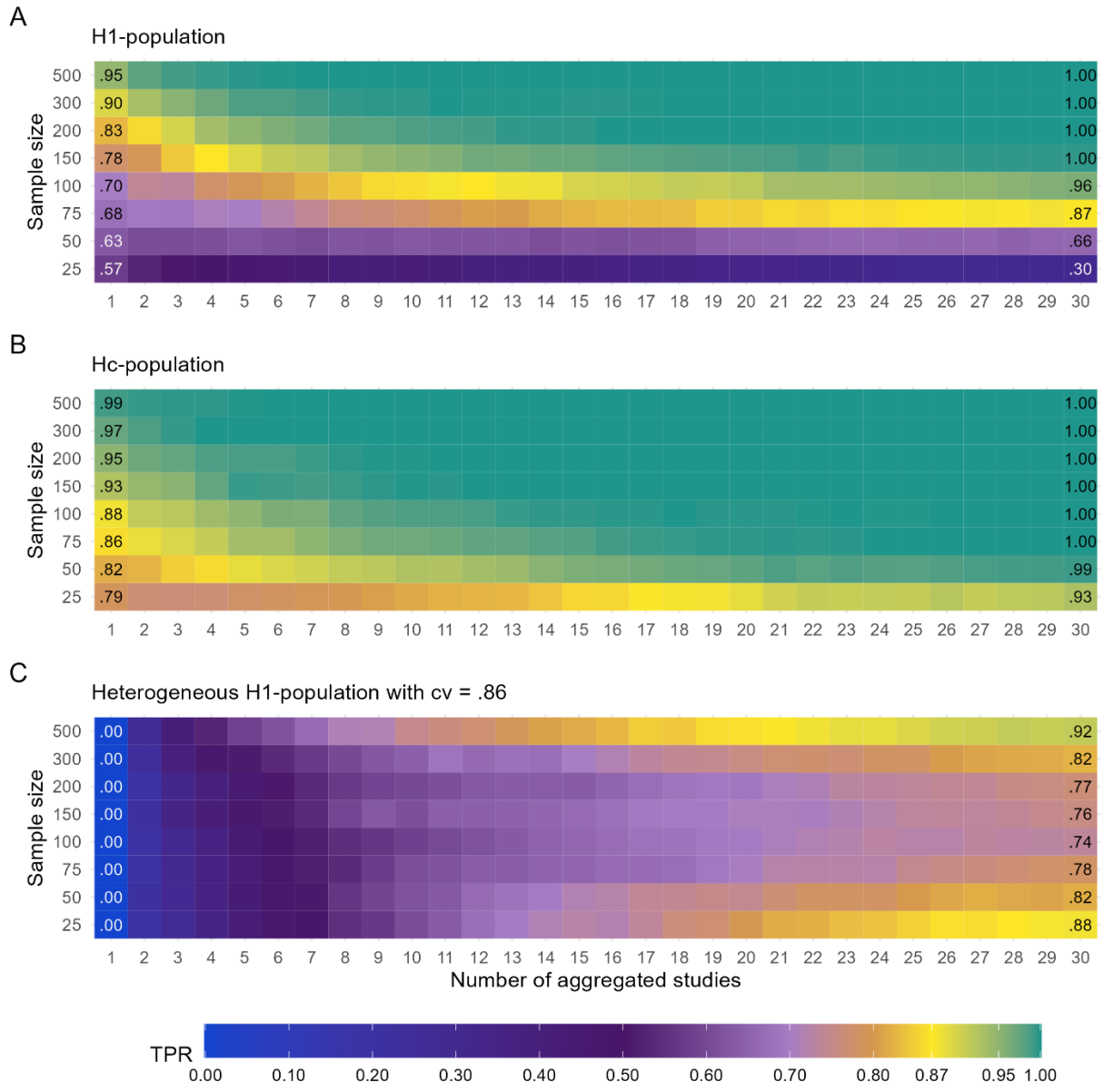

Figure 3. True positive rates (TPR) for each MPCTH. Based on populations H1 (A), Hc (B) or heterogeneous H1 with cv=.86 (C). The TPR in each cell is determined from 1000 iterations.

it unlikely that Hu was most supported. Even after aggregating 30 studies of large sample sizes (e.g., n=300), the TPR only reached .83. High TPRs only emerged when aggregating over studies with very large samples (n=500) after aggregating 25-30 studies.

Figure 3C shows a curvilinear pattern between sample size and number of aggregated studies. TPRs increase more quickly over studies for very small samples (n=25) than for larger studies (n=150). The increasing TPRs for Hu for n=25 in Figure 3C mirror the decreasing TPRs

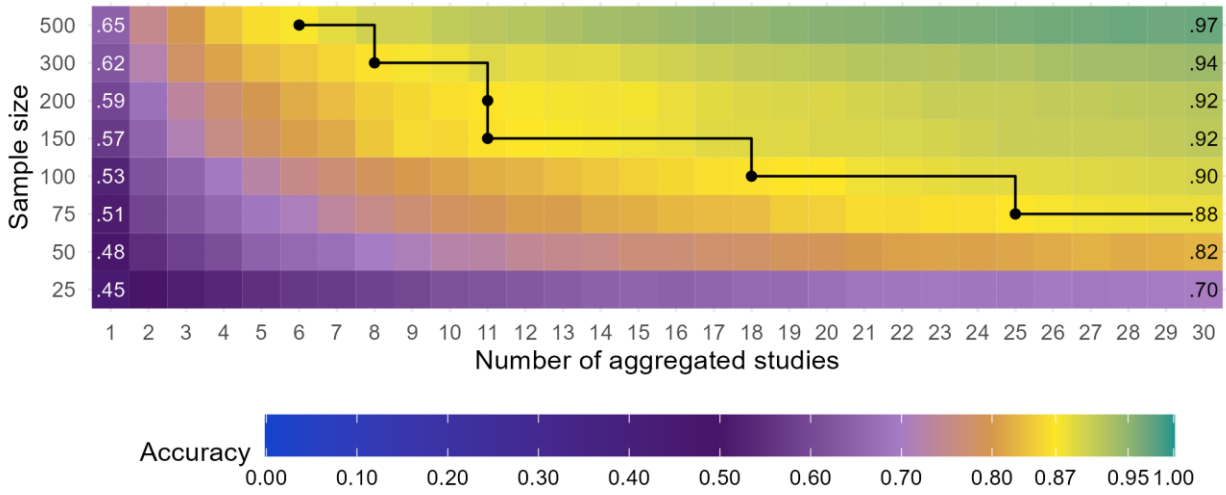

Figure 4. Accuracy of conjoint testing in a mix of three populations: H1, Hc and heterogeneous H1 with p=.86. The black dots indicate the first study in the row that reached accuracy of 0.87. The accuracy in each cell is determined from 3000 iterations (1000 per population).

for H1 for the same sample size (Figure 3A). Thus, when aggregating small-sampled studies, Hu is likely to be preferred over H1.

Accuracy

Figure 4 shows the accuracy of conjoint testing in a mix of three populations: H1, Hc, and heterogeneous H1 (cv = .86). The accuracy improved with larger number of studies irrespective of the sample size. However, aggregating large studies resulted in a rapid increase in accuracy; for instance, 8 studies with a sample size of 300 were needed to achieve an accuracy of .87. Contrarily, for small samples (n=25), the accuracy of conjoint testing increased slowly, and even after aggregating 30 studies, the accuracy remained low at .70.

The length of the line segments in Figure 4 reflects the advantage of larger sample sizes, with longer segments indicating greater advantages. However, the reduction in the required number of studies was not directly proportional to the increase in sample size. Specifically, increasing the sample size of small studies had greater benefit than increasing the sample size of large studies. For exapmle, a sample size increase from 75 to 100 reduced of necessary studies by 8, whereas a larger increase in sample size from 150 to 30 did not reduce the necessary number of studies.

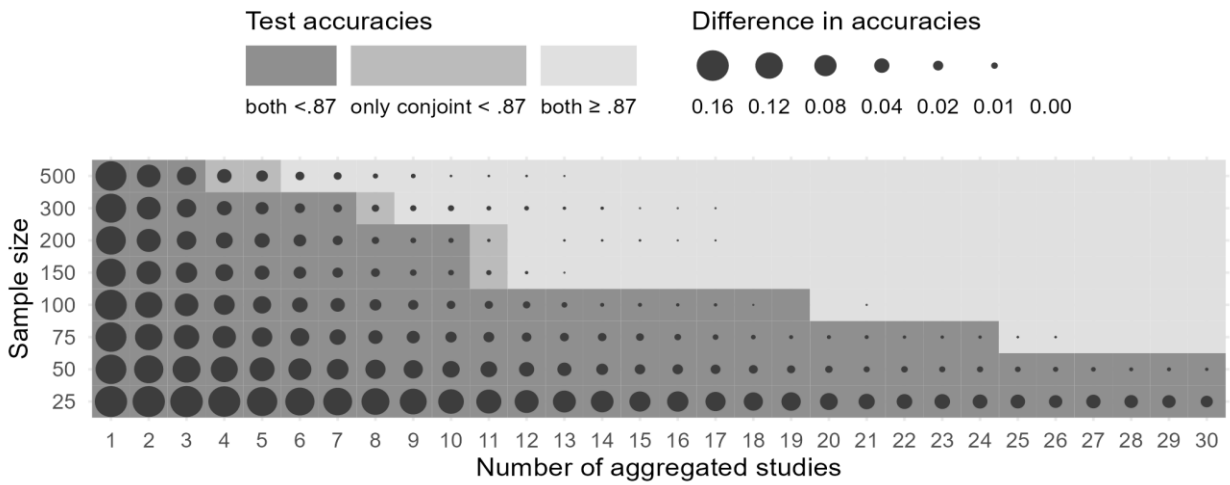

Figure 5. Accuracy of conjoint testing in a mix of three populations: H1, Hc and heterogeneous H1 with p=.86. The black dots indicate the first study in the row that reached accuracy of 0.87. The accuracy in each cell is determined from 3000 iterations (1000 per population).

Comparing the results in Figure 4 with those in Figure 3 shows that accuracies would have different main sources of error depending on the number of studies and their sample sizes. For instance, for n=300 and 8 studies (accuracy = .87), the only source of error is the incorrect rejection of Hu when it is true, as the TPR for H1 and Hc are both 1 (Figures 3A and 3B), while the TPR for Hu is only about .60 (Figure 3C). On the contrary, for n=25 and 30 (accuracy = .70) studies the main source of error is the incorrect rejection of H1, while the TPR for Hc and Hu are fairly high.

Simple unconstrained vs. conjoint testing

The comparison between simple unconstrained and conjoint testing in Figures 1B and 1C reveals the higher costs associated with the latter, as it requires a larger number of studies to achieve a similar distribution of PMPs. This pattern persists across different levels of heterogeneity, as shown in Figures 2B and 2C, and to a lesser extent with small amounts of heterogeneity (Figures 2E and 2F).

It was therefore of interest to compare the accuracies of conjoint and simple unconstrained testing. Figure 5 shows that the differences in accuracy between the two methods were not substantial. The largest difference was 0.16 in favor of simple unconstrained testing. The differences were relatively large in small study sets and is sets consisting of only small-sampled

studies. As the accuracies increased (with larger sample sizes and number of studies), the gap in accuracy between the two tests narrowed. A simple unconstrained test showed a practical advantage in cases where few large-sample studies (n=500) were aggregated, as it required two fewer studies to achieve an accuracy of .87 compared to conjoint testing.

Discussion

Previous research on BES has identified two underlying assumptions: (1) all target studies investigate the same underlying common theory, and (2) each study provides independent evidence for this theory (Klugkist & Volker, 2022). In this paper, we draw attention to a third implicit assumption: that at least 1 of the tested hypotheses is commonly true for all studies. This assumption would be violated, for instance, if there is a substantial heterogeneity in the individual study effects and each of the tested hypotheses is true in some studies, and not in others. Previous research has shown that BES provides unreliable aggregate support in such cases (van Wonderen, 2022). The present study introduced conjoint testing, which is the simultaneous test of an informative hypothesis Hi, its complement Hc, and the unconstrained hypothesis Hu, as a potential solution to this issue. The main aim of the study was to assess whether conjoint testing in the context of BES could provide reliable results under realistic conditions.

Summary & Interpretation of Findings

First, the study compared complement, conjoint and simple unconstrained testing in 3 populations with different heterogeneity levels. In the two more heterogenous populations (mixed H1/Hc and heterogeneous H1 with cv = .86) conjoint and simple unconstrained testing resolved inconsistencies observed in complement testing, as they successfully shifted the PMPs toward the only common true hypothesis Hu. In contrast, in the population with lower heterogeneity (heterogeneous H1 with cv = .50), all three tests supported H1. This can be considered a favourable behavor, as potentially negligible amounts of heterogeneity would not compromise the overall aggregate result. Importantly, in this case complement testing provided stronger support

for H1 than the other two tests (given the same number of aggregated studies). This is in line with the idea that complement testing can distinguish better between the hypotheses than a test including Hu, because H1 and Hc are mutually exclusive (Hoijtink, 2013). It also corroborates previous findings that complement testing necessitated fewer studies than simple unconstrained testing to support the correct H1 (van Wonderen, 2022; Volker, 2022). However, it is arguable whether here cv = .50 can be considered negligible, if the proportion of studies that originated from H1 was only 60% on average.

Another important finding was that in a H1 population with average heterogeneity levels (cv = .86) between 20 and 30 studies were necessary to correctly identify Hu as the MPCTH. This can be a daunting number for some fields. For instance, in medicine the median number of studies was found to be 3; in Psychology, it was 12 (van Erp et al., 2017). On the positive side, Linden and Hönekopp (2021) have found that the mean number of studies in psychological meta-analyses to be approximately 30. Thus, sufficiently large study sets are not unlikely. Note that the number of aggregated studies is a limiting factor also for traditional methods, such as random effects meta-analysis (Guolo & Varin, 2017), and the Q-test for heterogeneity (Higgins & Thompson, 2002).

Similarly to findings from previous research on BES, the size of the aggregated studies had a relatively small impact in correctly identifying Hu as the MPCTH (van Wonderen, 2022). That is, only very large studies (n=300,500) made a positive difference. Notably, aggregating only small studies also increased the probability of obtaining most support Hu. However this was a side effect of underpowered testing, rather than a sign of good performace. The reason is when aggregating underpowered studies, Hu would tended to obtain most support even when it is not true. Volker (2022) has shown this by performing a simple unconstrained test with BES on small-sampled studies: even though the correct hypothesis was H1: $\{\beta_1, \beta_2, \beta_3\} > 0$, the support accumulated for the false Hu. Conversely, when H1 was tested against Hc, H1 obtained larger aggregate support (Volker, 2022).

It was further investigated whether simple unconstrained testing had less stringent requirements regarding the size of the study set than conjoint testing. The rationale was that when testing more hypotheses the probability to support an incorrect hypothesis increases (Hoijtink et al., 2019). Here, the largest differences in accuracy between the two tests were observed when both had low accuracies, less than 0.87. However, as the accuracies increased, the differences between the two tests diminished. Notably, there was little advantage of simple unconstrained testing in terms of the necessary number of aggregated studies.

Implications

BES has demonstrated promise as an evidence synthesis method that can be utilized regardless of methodological differences among the studies. When selecting hypotheses to test, researchers should consider the number and power of individual studies. For a large number of sufficiently powered studies, conjoint testing is recommended due it has an interpretational benefit: it is possible to discern whether the hypothesis of interest is true (support for H1) or not (support for Hc), or if there is substantial heterogeneity (support for Hu). However, the power of individual studies is crucial for the consistency of the aggregate results. Thus, researchers are advised to obtain information about power analysis/sample size determination procedures of each study. If such information is not available, it is recommended to determine the expected power of a study given its sample size.

When the study set or the sample sizes of individual studies are small, complement testing can be considered, as it would have more power to support the MPCTH (if either Hi or Hc is the MPCTH). However, the assumption of a common true hypothesis should be made explicit, and the limitations of the conclusions should be stated in case the assumption would be violated.

The last implication discusses accuracy as an overall performance measure. In this study, accuracy was useful to compare the performance of conjoint to simple unconstrained testing. However, in practical applications, it is important to consider the individual true positive rates (TPR, as Figure 3) or error probabilities of each tested hypothesis, as previously also argued by Hoijtink

(2011) and Fu (2022). Relying solely on overall accuracy may mask the asymmetric TPRs of the individual hypotheses, some of which may be insufficient, leading to an overall unreliable test.

Limitations & future directions

The main strength of the present study was that it considered realistic simulation conditions based on a large sample of psychological meta-analyses. For brevity, only a limited number of conditions has been presented in this paper. More extensive simulation conditions can be found in the supplementing Shiny App (http://anita-lyubenova.shinyapps.io/BES_The_Value_Of_Hu). The Shiny App can be used to identify potential further research directions, such as varying the complexity of the hypotheses, or investigating the impact of heterogeneity on the results from BES in a more fine-grained fashion.

This study used ratio between the regression coefficient as the effect size measure, however, this is not a conventional measure used by applied researchers. The advantage of using the ratios, instead of other measures of predictor's relative importance (e.g., Gu, 2021) was that the results are comparable with previous BES investigations that also used ratios (Klugkist & Volker, 2022; Volker, 2022). However, future research might consider effect sizes closer to applied research.

Another consideration is the decision rule to determine the MPCTH. A hypothesis was considered accepted if its PMPs were the highest. This may be considered too lenient because usually it is necessary to obtain support above a certain threshold to accept a hypothesis. This was a pragmatic choice to be able to compare a test of 3 hypotheses with tests of 2 hypotheses. In such a case using the same threshold in both tests would have been inappropriate.

References

Baig, S. A. (2020). Bayesian Inference: Understanding Experimental Data With Informative Hypotheses. *Nicotine & Tobacco Research : Official Journal of the Society for Research on Nicotine and Tobacco*, *22*(11), 2118–2121. https://doi.org/10.1093/ntr/ntaa120

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Routledge. https://doi.org/10.4324/9780203771587

Crandall, C. S., & Sherman, J. W. (2016). On the scientific superiority of conceptual replications for scientific progress. *Journal of Experimental Social Psychology*, *66*, 93–99. https://doi.org/10.1016/j.jesp.2015.10.002

Derksen, M., & Morawski, J. (2022). Kinds of Replication: Examining the Meanings of “Conceptual Replication” and “Direct Replication”. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *17*(5), 1490–1505. https://doi.org/10.1177/17456916211041116

Fu, Q. (2022). *Sample Size Determination for Bayesian Informative Hypothesis Testing.* https://doi.org/10.33540/1221

Gu, X. (2021). Evaluating predictors’ relative importance using Bayes factors in regression models. *Psychological Methods.* Advance online publication. https://doi.org/10.1037/met0000431

Guolo, A., & Varin, C. (2017). Random-effects meta-analysis: The number of studies matters. *Statistical Methods in Medical Research*, *26*(3), 1500–1518. https://doi.org/10.1177/0962280215583568

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*(11), 1539–1558. https://doi.org/10.1002/sim.1186

Hoijtink, H. (2011). *Informative hypotheses: Theory and practice for behavioral and social scientists / Herbert Hoijtink*. *Chapman & Hall/CRC statistics in the social and behavioral sciences*. CRC Press. https://doi.org/10.1201/b11158

Hoijtink, H. (2013). Objective Bayes Factors for Inequality Constrained Hypotheses. *International Statistical Review*, *81*(2), 207–229. https://doi.org/10.1111/insr.12010

Hoijtink, H., Mulder, J., van Lissa, C. J., & Gu, X. (2019). *A tutorial on testing hypotheses using the Bayes factor*. Center for Open Science. https://doi.org/10.31234/osf.io/v3shc

Huisman, L. (2022). Are P-values and Bayes factors valid measures of evidential strength? *Psychonomic Bulletin & Review.* Advance online publication. https://doi.org/10.3758/s13423-022-02205-x

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., . . . Nosek, B. A. (2018). Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. https://doi.org/10.1177/2515245918810225

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: a Bayesian approach. *Psychological Methods*, *10*(4), 477–493. https://doi.org/10.1037/1082-989x.10.4.477

Klugkist, I., van Wesel, F., & Bullens, J. (2011). Do we know what we test and do we test what we want to know? *International Journal of Behavioral Development*, *35*(6), 550–560. https://doi.org/10.1177/0165025411425873

Klugkist, I., & Volker, T. B. (2022). *Bayesian Evidence Synthesis for Informative Hypotheses: An introduction*. Manuscript submitted for publication. https://github.com/thomvolker/bes-intro-paper/blob/main/BES%20intro%20paper.pdf

Kluytmans, A., van de Schoot, R., Mulder, J., & Hoijtink, H. (2012). Illustrating bayesian evaluation of informative hypotheses for regression models. *Frontiers in Psychology*, *3*, 2. https://doi.org/10.3389/fpsyg.2012.00002

Kuiper, R. M., Buskens, V., Raub, W., & Hoijtink, H. (2013). Combining Statistical Evidence From Several Studies. *Sociological Methods & Research*, *42*(1), 60–81. https://doi.org/10.1177/0049124112464867

Kuiper, R. M., Nederhoff, T., & Klugkist, I. (2015). Properties of hypothesis testing techniques and (Bayesian) model selection for exploration-based and theory-based (order-restricted) hypotheses. *The British Journal of Mathematical and Statistical Psychology*, *68*(2), 220–245. https://doi.org/10.1111/bmsp.12041

Linden, A. H., & Hönekopp, J. (2021). Heterogeneity of Research Results: A New Perspective From Which to Assess and Promote Progress in Psychological Science. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *16*(2), 358–376. https://doi.org/10.1177/1745691620964193

Lovie, P. (2005). Coefficient of Variation. In B. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science.* John Wiley. https://doi.org/10.1002/0470013192.bsa107

Mulder, J. (2014). Prior adjusted default Bayes factors for testing (in)equality constrained hypotheses. *Computational Statistics & Data Analysis*, *71*, 448–463. https://doi.org/10.1016/j.csda.2013.07.017

Mulder, J., & Olsson-Collentine, A. (2019). Simple Bayesian testing of scientific expectations in linear regression models. *Behavior Research Methods*, *51*(3), 1117–1130. https://doi.org/10.3758/s13428-018-01196-9

Mulder, J., Williams, D. R., Gu, X., Tomarken, A., Böing-Messing, F., Olsson-Collentine, A., Meijerink, M., Menke, J., van Aert, R., Fox, J.-P., Hoijtink, H., Rosseel, Y., Wagenmakers, E.-J., & van Lissa, C. (2021). BFpack : Flexible Bayes Factor Testing of

Scientific Theories in R. *Journal of Statistical Software*, *100*(18). https://doi.org/10.18637/jss.v100.i18

Olsson-Collentine, A., Wicherts, J. M., & van Assen, M. A. L. M. (2020). Heterogeneity in direct replications in psychology and its association with effect size. *Psychological Bulletin*, *146*(10), 922–940. https://doi.org/10.1037/bul0000294

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Open Science Collaboration. (2017). Maximizing the Reproducibility of Your Research. In *Psychological Science Under Scrutiny* (pp. 1–21). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119095910.ch1

Schweinsberg, M., Madan, N., Vianello, M., Sommer, S. A., Jordan, J., Tierney, W., Awtrey, E., Zhu, L. L., Diermeier, D., Heinze, J. E., Srinivasan, M., Tannenbaum, D., Bivolaru, E., Dana, J., Davis-Stober, C. P., Du Plessis, C., Gronau, Q. F., Hafenbrack, A. C., Liao, E. Y., . . . Uhlmann, E. L. (2016). The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *Journal of Experimental Social Psychology*, *66*, 55–67. https://doi.org/10.1016/j.jesp.2015.10.001

Stroebe, W., & Strack, F. (2014). The Alleged Crisis and the Illusion of Exact Replication. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *9*(1), 59–71. https://doi.org/10.1177/1745691613514450

van de Schoot, R., Hoijtink, H., Hallquist, M. N., & Boelen, P. A. (2012). Bayesian Evaluation of inequality-constrained Hypotheses in SEM Models using Mplus. *Structural Equation Modeling : A Multidisciplinary Journal*, *19*(4). https://doi.org/10.1080/10705511.2012.713267

van de Schoot, R., & Strohmeier, D. (2012). Aggressive Behaviour in Native, First- and Second-Generation Immigrant Youth: Testing Inequality Constrained Hypotheses. In *Migrations:*

Interdisciplinary Perspectives (pp. 89–98). Springer, Vienna. https://doi.org/10.1007/978-3-7091-0950-2_8

van Erp, S., Verhagen, J., Grasman, R. P. P. P., & Wagenmakers, E.-J. (2017). Estimates of Between-Study Heterogeneity for 705 Meta-Analyses Reported in Psychological Bulletin From 1990–2013. *Journal of Open Psychology Data*, *5*(1), 4. https://doi.org/10.5334/jopd.33

van Rossum, M., van de Schoot, R., & Hoijtink, H. (2013). “Is the Hypothesis Correct” or “Is it Not”. *Methodology*, *9*(1), 13–22. https://doi.org/10.1027/1614-2241/a000050

van Wonderen, E. (2022). *Comparing Bayesian evidence synthesis to meta-analysis: A simulation study and empirical application.*

Vanbrabant, L., van de Schoot, R., & Rosseel, Y. (2014). Constrained statistical inference: Sample-size tables for ANOVA and regression. *Frontiers in Psychology*, *5*, 1565. https://doi.org/10.3389/fpsyg.2014.01565

Volker, T. B. (2022). *Combining support for hypotheses over heterogeneous studies with Bayesian Evidence Synthesis: A simulation study*. https://github.com/thomvolker/bes_master_thesis_ms/blob/main/manuscript_VK/manuscript_VK.pdf

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2017). Making replication mainstream. *Behavioral and Brain Sciences*, *41*, e120. https://doi.org/10.1017/s0140525x17001972

Appendix A

Sampling of study-level ratio values

Study-level ratios between coefficients were captured in the column vector $\Gamma = [\gamma_1\ \gamma_2\ \gamma_3]'$, for which $\beta_1{:}\beta_2{:}\beta_3 = \gamma_1{:}\gamma_2{:}\gamma_3$. Heterogeneity in the population implied that studies within the set could have different Γ. To simulate heterogeneity across studies, for each study the ratio values γ_k were sampled from a log-normal distribution[4]. The log-normal distribution has a location and shape parameter, which are the mean and the variance of the natural logarithm of the distribution. To sample a ratio value γ_k from a log-normal distribution with specified arithmetic mean μ_k and arithmetic standard deviation $\sigma_k = \mu_k cv$, the location and shape parameters were parametrized as follows:

$$\gamma_k \sim Lognormal\left(location_k = \frac{\mu_k^2}{\sqrt{\mu_k^2 + (\mu_k cv)^2}},\qquad shape_k = \ln\left(1 + \frac{\sigma_k^2}{\mu_k^2}\right) \right)$$

[4] a log-normal distribution was selected to avoid sampling negative values for the ratios

Appendix B

Determine levels of N and cv

To determine realistic values for the sample sizes N and the coefficient of variation cv, their empirical distributions were obtained from a survey of 150 psychological meta-analyses (Linden & Hönekopp, 2021)[5]. The distribution of both N and cv had a very strong positive skew; therefore, the distributions were adjusted by winsorizing[6] 10% of the most extreme values (Figure B1 and Figure B2).

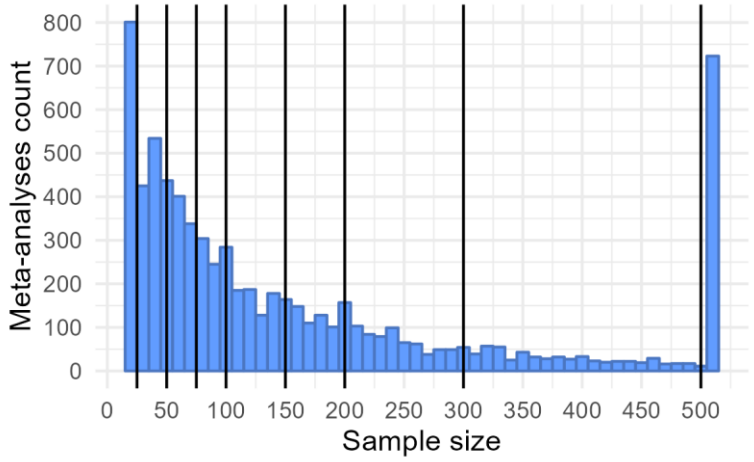

Figure B1. Distribution of 10% winsorized sample sizes N across 7227 studies from 150 meta-analyses included in the survey of Linden and Hönekopp (2021). The vertical lines indicate the sample sizes used in the simulation study.

[5] Consent from the authors has been obtained via e-mail.
[6] Winsorizing refers to the technique of replacing a certain percentage of the most extreme values with the value closest to them. It is an efficient way of dealing with skewed distributions and extreme outliers

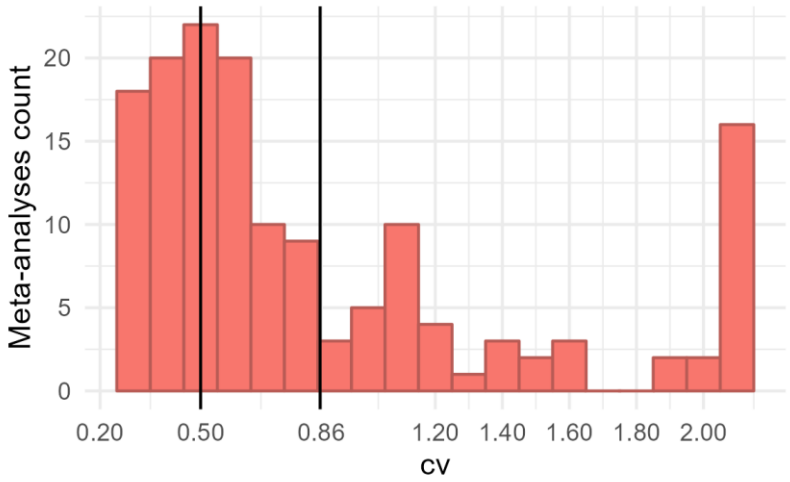

Figure B2. Distribution of 10% winsorized cv across 150 meta-analysies included in the survey of Linden and Hönekopp (2021). The vertical lines indicate the cv values used in the simulation study.