Universiteit Utrecht

Methodology & Statistics for the Behavioural, Biomedical and Social Sciences

Internship Report

Bayesian Evidence Synthesis: Aggregating evidence from partially overlapping hypotheses

Author:

Florian Metwaly

Student-Nr.: 0778265

Supervisor:

Irene Klugist

Duco Veen

Alfons Edmar

September 17, 2024

Contents

1 Introduction

Common practice in research is to conduct meta-analyses, which aggregate evidence over multiple studies at the level of effect sizes, to draw more reliable and generalizable conclusions about a particular phenomenon. By combining the results of several studies, meta-analyses increase statistical power and reduce the uncertainty associated with individual studies that may have small sample sizes or inconsistent results. But, this is only possible if multiple studies are conceptual replications of each other. Bayesian Evidence Synthesis (BES) is an alternative for meta-analyses, which aggregates evidence on the level of the Bayes Factor (BF). As has been shown, BES can reliably aggregate the evidence over multiple studies with diverse study set-ups, as long as the underlying hypotheses represent the same underlying effect. Kevenaar et al. (2021) used BES to aggregate the evidence over multiple studies, in which each study only contained information about part of an overarching hypothesis, while all studies together did cover the full range of the overarching hypothesis. This application of Bayesian Evidence Synthesis deems clear practical relevance, but the concept has yet to be proven. Therefore, this internship tries to answer the question: Is the synthesis of BFs from studies that evaluate only parts of the overarching hypothesis a good measure of joint support for the overarching hypothesis?

2 Bayesian Evidence Synthesis

The Bayes Factor quantifies the evidence for or against an informative hypothesis relative to either one other hypothesis or a set of hypotheses (Hoijtink, Mulder, van Lissa, & Gu, 2019). The Bayes Factor is calculated by taking the ratio of the marginal likelihoods m_1 and m_2 of two hypotheses H_1 and H_2 (Chib, 1995; Kass & Raftery, 1995)

$$BF_{1,2} = \frac{m_1}{m_2} = \frac{P(D|H_1)}{P(D|H_2)}. \tag{1}$$

By reflecting the ratio of evidence for the different hypotheses, the $BF_{1,2}$ is interpreted as the amount of evidence for H_1 over H_2. An $BF_{1,2}$ of 10 therefore is interpreted as ten times the amount of evidence for H_1 over H_2.

There are different alternative hypotheses that can be used in a Bayesian Hypothesis Testing

framework. Common choices include the complement hypothesis, the null hypothesis or the unconstrained hypothesis. In this report, the unconstrained hypothesis is further chosen as alternative hypothesis. The unconstrained hypothesis H_u imposes no constraints and therefore allows the parameters to vary freely, without any specific restrictions or predefined values. Testing a hypothesis H_1 against the unconstrained hypothesis H_u simplifies the calculation of the Bayes Factor to

$$BF_{1u} = \frac{m_1}{m_u} = \frac{f_1}{c_1} \tag{2}$$

with f_1 being the fit and c_1 the complexity of H_1, as shown in Klugkist, Laudy, and Hoijtink (2005). The fit f_1 is the proportion of the unconstrained posterior distribution that is in agreement with H_1, while the complexity c_1 is the proportion of the unconstrained prior distribution that is in agreement with H_1 (Hoijtink, 2011). By using the Bayes Factor for hypothesis testing, as alternative to null hypothesis significance testing (NHST), one is able to test different informative hypotheses against each other. Further, as the Bayes Factor quantifies the support for a hypothesis, this method allows to find direct support for a hypothesis, instead of support against the null hypothesis, as in the NHST framework (Hoijtink et al., 2019).

Bayesian Evidence Synthesis can be used to aggregate the support for a hypothesis based on the Bayes Factors aggregated over multiple studies. BES assumes that multiple studies with a diverse methodology exist, that all test the same hypothesis (Klugkist & Volker, 2023; Kuiper, Buskens, Raub, & Hoijtink, 2012). The evidence of these studies can then meaningfully be combined by combining the Bayes Factors of these studies, using BES, to aggregate support for the hypothesis over multiple studies that can not be combined by traditional meta analysis or data pooling approaches. This method has multiple advantages over traditional meta analysis approaches, especially concerning the flexibility of the design of the different studies. Even though there are meta analysis methods that do not require the included studies to be exact replications, like random-effect meta analyses, the studies do need to show a high methodological similarity (Klugkist & Volker, 2023). Through aggregating the evidence for a hypothesis, represented by the Bayes Factor, BES allows for a high methodological diversity between studies, as long as they test the same theoretical concept

(Klugkist & Volker, 2023).

What needs to be highlighted when using BES, is that it answers a different question than data pooling approaches like Bayesian Sequential Updating (BSU). Bayesian Evidence Synthesis is explicitly not a method to pool different data sources, and calculate the combined evidence over the pooled data, like it is done with BSU. The BF that results from BES reflects the support for a hypothesis in each study separately (Klugkist & Volker, 2023). Klugkist and Volker (2023) showcase in which cases this leads to different results. As data is not pooled, an important limitation of BES is also that it does not overcome power issues due to a small sample size in a study, which is an important application of BSU or meta-analysis (Klugkist & Volker, 2023).

An advantage of BES also is the simplicity, as the aggregation essentially happens by multiplying the Bayes Factors of the different studies

$$\left(\frac{P\left(H_1 \mid D\right)}{P\left(H_2 \mid D\right)}\right)^T = \frac{P\left(H_1\right)}{P\left(H_2\right)} \prod_{t=1}^{T} \left(\mathrm{BF}_{1,2}\right)^t, \tag{3}$$

where $t = 1, ...T$ is the number of studies. The prior $P(H_1)/P(H_2)$ is normally set to one (Klugkist & Volker, 2023). With equal prior model probabilities for each hypothesis H_t, when only two hypotheses are tested against each other and when posterior model probabilities are aggregated, Equation 3 can be rewritten to

$$PMP\left(H_i\right)^T = \frac{\prod_{t=1}^{T} PMP\left(H_i\right)^t}{\prod_{t=1}^{T} PMP\left(H_i\right)^t + \prod_{t=1}^{T} \left(1 - PMP\left(H_i\right)^t\right)} \tag{4}$$

(Van Wonderen, Zondervan-Zwijnenburg, & Klugkist, 2024).

3 The Keveenar Case

That Bayesian Evidence Synthesis is an appropriate method for combining evidence of multiple studies, that test the same hypothesis, has been shown in the literature (Klugkist & Volker, 2023; Kuiper et al., 2012; Van Wonderen et al., 2024; Volker & Klugkist, 2023). Kevenaar et al. (2021) applies BES to a different scenario. In the Keveenar case, studies do not vary by model or methodology, but by available data. Keveenar uses the data of four cohort studies on children's self-control problems, rated by different informants, to evaluate

the overarching hypothesis:

$$H_i : \mu_{self} > \mu_{mother} > \mu_{father} > \mu_{teacher}.$$

Each cohort study only contains data of a different subset of informants, e.g. only the child and teacher, or mother, father and teacher. Therefore, each study investigates only part of the overarching H_i. All studies together cover the full range of H_1, therefore Keveenar aggregates the resulting Bayes Factors of each study using Bayesian Evidence Synthesis. While this method does show clear practical relevance, there yet is no methodological work confirming that BES can be used in this case. This gab is explored in this report. This results in the question: Is the synthesis of BFs from studies that evaluate only parts of the overarching hypothesis a good measure of joint support for the overarching hypothesis? To investigate this question a simulation study is conducted.

To match the scope of this report, a simplified case is chosen for the simulation study. Here, only the correct hypothesis H_i is tested against the unconstrained hypothesis H_u, to investigate the behaviour of the BF_{iu}, or subsequently the PMP_{iu}. Further, to be able to isolate the effect of different parameters, the number of studies (and therefore also hypotheses) that are aggregated is reduced from four in Keveenar's case, to three in this simulation. The resulting hypotheses that will be aggregated reflect a case in which, with the least amount of available information, the whole range of the overarching hypothesis is covered:

$$H_1 : \mu_{self} > \mu_{mother}$$

$$H_2 : \mu_{mother} > \mu_{father}$$

$$H_3 : \mu_{father} > \mu_{teacher}$$

The simulation procedure is outlined in the following section.

4 Simulation

As for the calculation of the Bayes Factor only the mean μ, the sample size n and the covariance matrix Σ are needed, instead of simulating full datasets and calculating these effect sizes, the effect sizes are simulated directly.

First, the population parameters are defined. These consist of four mean values and a 4x4 covariance matrix. Further, the following sample sizes per study are defined: 25, 50, 100, 500, 5000. Those sample sizes are chosen to being able to replicate the conditions of the Keveenar study and to cover a range of popular sample sizes in social sciences, from low to large. The population μ values are chosen to reflect no effect and low (0.2), medium (0.5) and large (0.8) effects according to Cohen's d. Simultaneously for each effect size the covariance matrix is fixed with a variance of 1, and covariances are chosen to reflect no correlation, and low (0.2), medium (0.5) and large (0.8) correlation. All simulations are run with 10000 iterations.

The mean values are drawn from a multivariate normal distribution, using the defined population mean and covariance matrix. To account for sampling variability, the covariance matrix is corrected for the sample size of the study the value is drawn for

$$\Sigma_{sim} = \frac{\Sigma_{population}}{n}{}^{1}. \tag{5}$$

The sample covariance matrix $Sigma_{sample}$ is drawn from the Wishart distribution. The Wishart distribution allows for the drawing of covariance matrices given the degrees of freedom ($df = n - 1$) and the population covariance matrix $\Sigma_{population}$ (Wishart, 1928). Further, due to the properties of the distribution, the resulting covariance matrix has to be scaled down by the factor $1/m, where m = df$.

The Bayes Factors are calculated using the BFpack package in R (Mulder et al., 2021). Finally, Bayesian Evidence Synthesis is performed according to Equation 4.

All simulations and calculations were performed using R version 4.4.1 (R Core Team, 2024), BFpack package version 1.3.0 (Mulder et al., 2021) and the MASS package version 7.3-60.2 (Venables & Ripley, 2002). All scripts and simulated data sets are available in the Supplementary Material.

[1]This correction is based on the standard error of the mean. For a discussion of this see `https://math.stackexchange.com/questions/504288/what-situation-calls-for-dividing-the-standard-deviation-by-sqrt-n`, Barde and Barde (2012) and Lee, In, and Lee (2015).

5 Results

While the sample size of 5000 was computed, it is removed from all figures to keep them more readable, as there was almost no difference in the results of sample size 500 and 5000. Figure 1 shows an overview of the simulation results. The plots in the grid increase in effect size on the x-axis and in correlation on the y-axis. Each plot shows the sample size n on the x-axis and the posterior model probabilities against the unconstrained hypothesis PMP_{1u} on the y-axis. The full line shows the PMPs of the complete hypothesis, $H_i : \mu_{self} > \mu_{mother} > \mu_{father} > \mu_{teacher}$, tested in one study with sample size n. The dashed line shows the aggregated BES over three studies, each tested with sample size n/3. As the hypotheses are tested against the unconstrained hypothesis H_u, the Bayes Factors have an upper limit which is dependent on their complexity. This limit is 0.96 for the complete hypothesis and 0.889 for the BES posterior model probabilities. The grey dotted line shows the limit of the BES in Figure 1.

Overall, Figure 1 indicates that Bayesian Evidence Synthesis is an appropriate way to aggregate evidence, in the case where only partial data sources, that test parts of an overarching hypothesis, are available. In the case of no effect, as can be seen in column one, the BES posterior model probabilities do find slightly more support for H_1 than the PMP_{1u}, but there still is way larger support for H_u. Further, two clear trends can be noted. The effect size plays a big role, large effect sizes ($d \geq 0.5$) reach, or get very close to, the limit of the BES in almost every case. When there is only a small effect ($d = 0.2$), BES still finds good support of the overarching hypothesis, although in small sample sizes it is way smaller than the PMP_{1u} which test the complete hypothesis. When having a large sample ($n_{total} = 1500$), BES also reaches the limit. Higher correlation in the data leads to small improvements in the posterior model probabilities, but compared with the effect of the sample size and effect size, the effect of the correlation is rather small.

Figure 2 is picked out as an interesting case to look at the results in more detail. The boxplots show the results of the simulation over 10k iterations, for effect size $d = 0.2$ and correlation $\rho = 0.2$. Figure 2b shows that there is a higher variability in the results of the BES, compared to testing the complete hypothesis in Figure 2a. An increasing sample size leads to a steady reduction in the variability, but BES leads to more strong outliers. Overall,

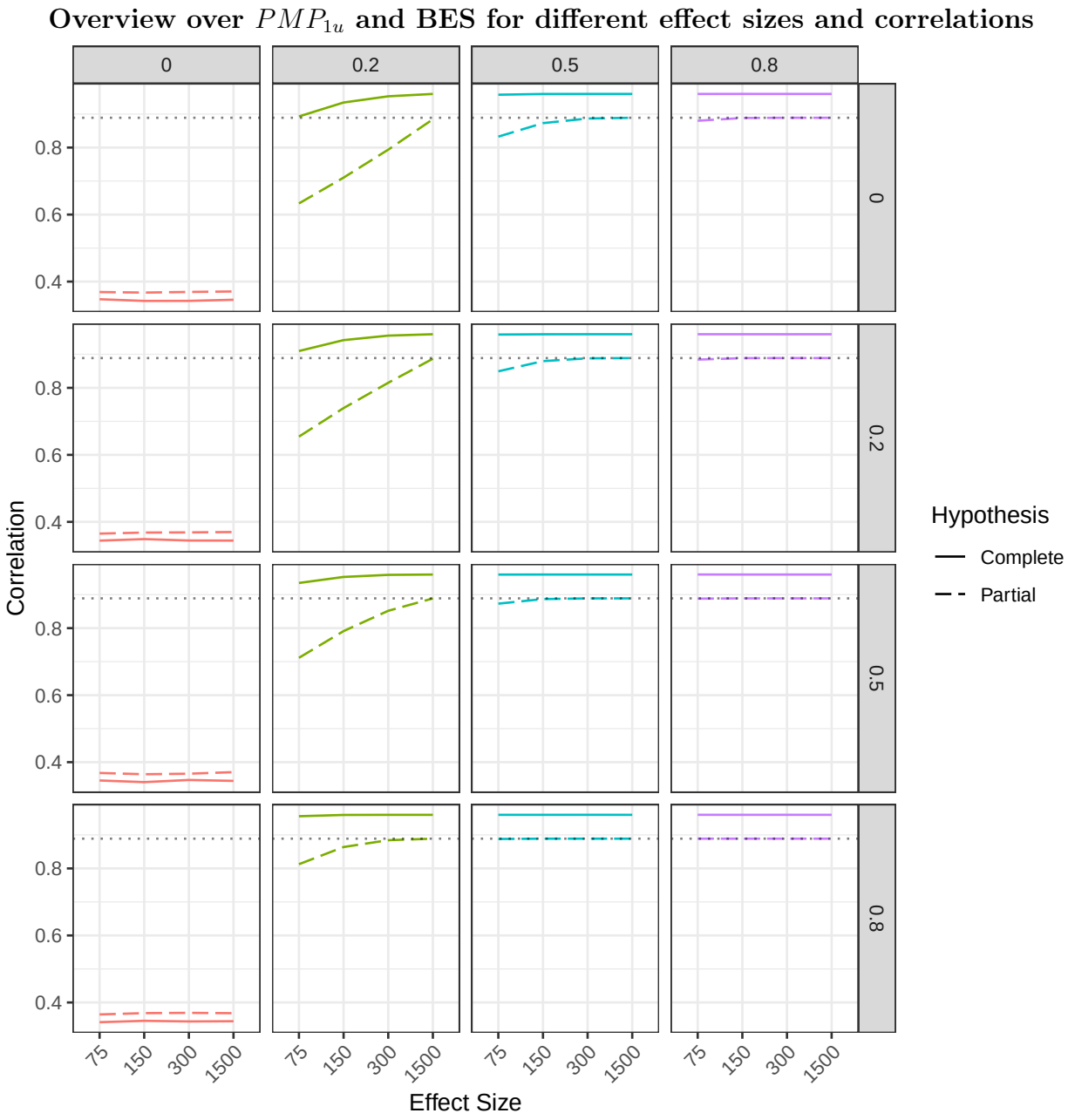
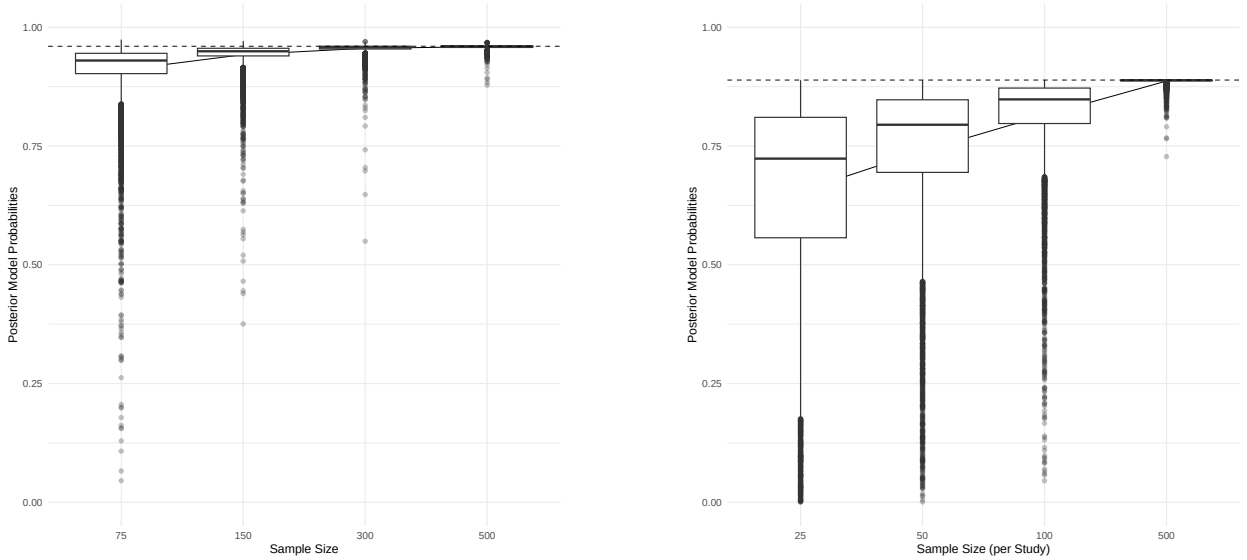


Figure 1: The figure shows the comparison of PMPs of a study testing the complete hypothesis ($H_i : \mu_{self} > \mu_{mother} > \mu_{father} > \mu_{teacher}$) against the BES of three studies with partial hypotheses ($H_1 : \mu_{self} > \mu_{mother}, H_2 : \mu_{mother} > \mu_{father}, H_3 : \mu_{father} > \mu_{teacher}$). The complete hypothesis was tested over a sample of size n, while the partial hypotheses each got tested over a sample of n/3. All hypotheses are tested against the unconstrained hypothesis H_u. The grey dotted line displays the maximum PMP that the BES can reach due to the limit that the BF_{1u} can reach.

Comparison of PMP_{1u} and BES for effect size $d = 0.2$ and correlation $\rho = 0.2$



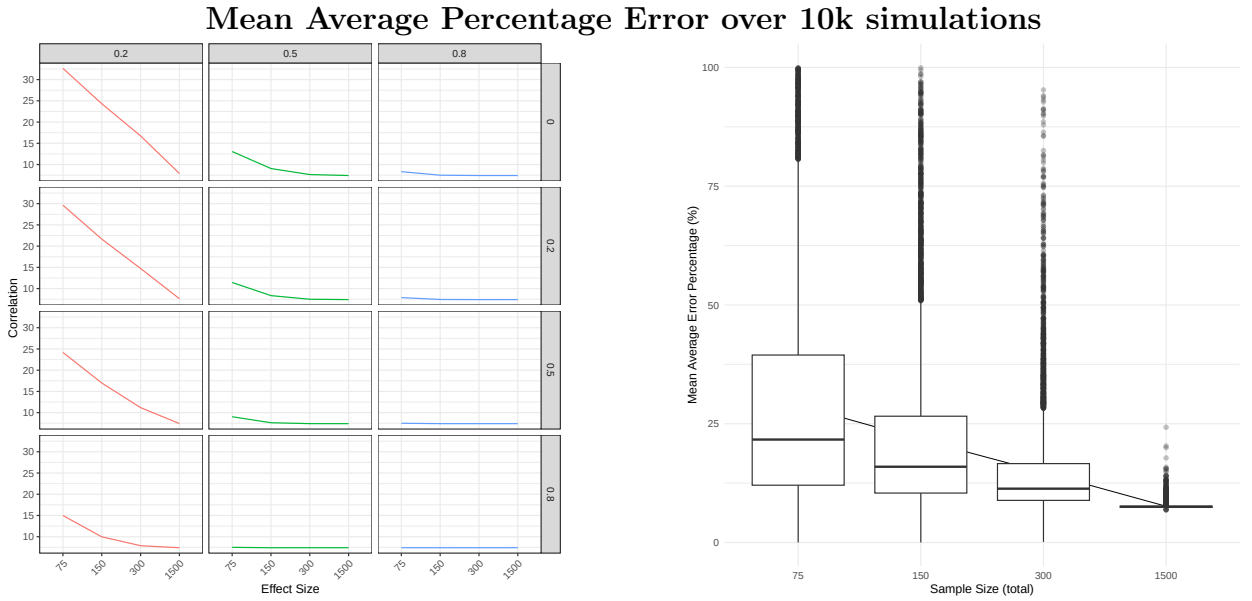
(a) PMP_{1u} of hypothesis $H_i : \mu_{self} > \mu_{mother} > \mu_{father} > \mu_{teacher}$.

(b) Aggregated BES over three studies with hypotheses $H_1 : \mu_{self} > \mu_{mother}; H_2 : \mu_{mother} > \mu_{father}; H_3 : \mu_{father} > \mu_{teacher}$.

Figure 2: The boxplots show the result of the simulations for effect size $d = 0.2$ and correlation $\rho = 0.2$, over 10k iterations. The full black line represents the mean value of the posterior model probability and the dashed black line shows the limit of the PMP_{1u} in Figure 2a and the limit of the BES in Figure 2b.

the results indicate that Bayesian Evidence Synthesis is an appropriate method to combine the evidence of partial hypotheses, to test an overarching common theory. The BES posterior model probabilities do show a larger variability and more outlier, but all in all, BES produces stable results in these simulation cases.

Lastly, the Mean Average Percentage Error (MAPE) is shown in Figure 3. The MAPE is a



(a) MAPE over different effect sizes (x-axis) and correlations (y-axis).

(b) Boxplot of MAPE for effect size $d = 0.2$ and correlation $\rho = 0.2$.

Figure 3: The plots show an overview of the MAPE (Figure 3a) and the MAPE for effect size $d = 0.2$ and correlation $\rho = 0.2$ (Figure 3b) over 10k simulations.

measure of accuracy and is calculated with the following equation:

$$MAPE = 100 * \frac{1}{n} * \sum_{i=1}^{n} * \left| \frac{PMP_{1u_i} - BES_i}{PMP_{1u_i}} \right| \tag{6}$$

(Kim & Kim, 2016).

Figure 3a shows that especially with a small effect size ($d = 0.2$), the MAPE tends to be large with up to around 33%. Increasing the sample size, or with a higher effect size, the MAPE reduces heavily down to about 7%. High correlation within the data also leads to a small reduction of the MAPE. Figure 3b shows a boxplot of the MAPE for effect size $d = 0.2$ and correlation $\rho = 0.2$. It is clear that increasing sample size leads to a reduction in the

variability of the results. While a sample size of 25 observations per study still leads to a large whisker which reaches around 80%, already increasing the sample size to 50 observations per study reduces the variability largely. While there are strong outliers, the large majority of results are stable. Again, these results indicate that Bayesian Evidence Synthesis is a feasible method to combine the evidence of multiple studies that test partial hypotheses of one overarching theory.

6 Conclusion

This report investigated the question, if the synthesis of BFs from studies that evaluate only parts of the overarching hypothesis is a good measure for joint support for the overarching hypothesis. To answer this question, a simulation study was conducted, which was oriented on Kevenaar et al. (2021) application of the BES in this way. The simulation conducted BES for studies of varying sample sizes, effect sizes and correlations in the data, but was limited to only testing true hypotheses with the same complexity against the unconstrained hypothesis. The resulting BES were compared with the posterior model probabilities of the overarching hypothesis $H_i : \mu_{self} > \mu_{mother} > \mu_{father} > \mu_{teacher}$, which was tested on a sample of the same size as the combined studies together.

Results show that, in this case, Bayesian Evidence Synthesis is an appropriate method to aggregate evidence. While there still is a rather large difference in posterior model probabilities between testing the complete hypothesis and doing BES when the sample size and the effect size are low, BES still finds correct support for the overarching hypothesis. Increasing sample size, or if the effect in the population is of a medium or large size, BES quickly reaches it's limiting value which is effected by the complexity of the hypotheses in each aggregated study. A further measure to compare the results was the Mean Average Percentage Error. The MAPE confirmed the results again. While the error gets as high as about 33% when the effect and sample sizes are low, increasing the sample size, or a medium to high effect size in the population reduces the error to around 7%.

While these results indicate that BES can be used to aggregate evidence of multiple partial studies to test an overarching hypothesis, there are still multiple aspects that need to be

explored to reliably use this method in applied research in the future. One of these aspects is to explore how BES behaves when only parts of the overarching hypothesis is true. An easy first extension to this work could be to switch two of the parameters in the hypothesis, for example to

$$H_i : \mu_{self} > \mu_{father} > \mu_{mother} > \mu_{teacher}.$$

This new setup could give new insights in the behaviour of Bayesian Evidence Synthesis.
As already mentioned, it is known that the complexity of a hypothesis influences the BF. This relationship should be further explored. In the presented scenario, the complexity of H_i was 24, while the complexity that the BES could reach was only 8. This resulted in different maximal values for the posterior model probabilities. This effect needs to be further investigated. Discussions here could be about scaling BES for the complexities of the partial hypotheses, in comparison to the overarching hypothesis. Another way to investigate the influence of complexity could be to vary the complexity of a hypothesis without changing the other parameters. But this is a challenging task, as it probably needs a more complex study setup with a different effect size measure. Lastly, in this study, the partial hypotheses were chosen in a way that the range of the complete hypothesis was covered with the least amount of information possible. Further research should investigate how BES behaves when the partial hypotheses overlap to a larger extent.

Supplementary Material

All scripts, data and figures are available in the supplementary material on GitHub: `https://github.com/flo1met/Internship_BES/`.

AI Statement

During the internship and while writing this report, generative AI models (ChatGPT, Grammarly) were solely used for stylistic and grammatical improvements on the text and for assistance during coding.

Bibliography

Barde, M. P., & Barde, P. J. (2012, September). What to use to express the variability of data: Standard deviation or standard error of mean? *Perspectives in Clinical Research*, *3*(3), 113. Retrieved 2024-09-16, from https://journals.lww.com/picp/fulltext/2012/03030/what_to_use_to_express_the_variability_of_data_.7.aspx doi: 10.4103/2229-3485.100662

Chib, S. (1995, December). Marginal Likelihood from the Gibbs Output. *Journal of the American Statistical Association*, *90*(432), 1313–1321. Retrieved 2024-09-13, from https://www.tandfonline.com/doi/abs/10.1080/01621459.1995.10476635 (Publisher: ASA Website _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1995.10476635) doi: 10.1080/01621459.1995.10476635

Hoijtink, H. (2011). *Informative Hypotheses: Theory and Practice for Behavioral and Social Scientists*. CRC Press. (Google-Books-ID: SzfOBQAAQBAJ)

Hoijtink, H., Mulder, J., van Lissa, C., & Gu, X. (2019). A tutorial on testing hypotheses using the Bayes factor. *Psychological Methods*, *24*(5), 539–556. (Place: US Publisher: American Psychological Association) doi: 10.1037/met0000201

Kass, R. E., & Raftery, A. E. (1995, June). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773–795. Retrieved 2024-09-13, from https://www.tandfonline.com/doi/abs/10.1080/01621459.1995.10476572 (Publisher: ASA Website _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1995.10476572) doi: 10.1080/01621459.1995.10476572

Kevenaar, S. T., Zondervan-Zwijnenburg, M. A., Blok, E., Schmengler, H., Fakkel, M. T., De Zeeuw, E. L., ... Oldehinkel, A. J. (2021, February). Bayesian evidence synthesis in case of multi-cohort datasets: An illustration by multi-informant differences in self-control. *Developmental Cognitive Neuroscience*, *47*, 100904. Retrieved 2024-08-16, from https://linkinghub.elsevier.com/retrieve/pii/S1878929320301535 doi: 10.1016/j.dcn.2020.100904

Kim, S., & Kim, H. (2016, July). A new metric of absolute percentage error for inter-

mittent demand forecasts. *International Journal of Forecasting*, *32*(3), 669–679. Retrieved 2024-09-16, from https://www.sciencedirect.com/science/article/pii/S0169207016000121 doi: 10.1016/j.ijforecast.2015.12.003

Klugkist, I., Laudy, O., & Hoijtink, H. (2005, December). Inequality Constrained Analysis of Variance: A Bayesian Approach. *Psychological Methods*, *10*(4), 477–493. Retrieved 2024-09-13, from http://doi.apa.org/getdoi.cfm?doi=10.1037/1082-989X.10.4.477 doi: 10.1037/1082-989X.10.4.477

Klugkist, I., & Volker, T. B. (2023, September). Bayesian evidence synthesis for informative hypotheses: An introduction. *Psychological Methods*. Retrieved 2024-08-16, from https://doi.apa.org/doi/10.1037/met0000602 doi: 10.1037/met0000602

Kuiper, R. M., Buskens, V., Raub, W., & Hoijtink, H. (2012). Combining Statistical Evidence From Several Studies.

Lee, D. K., In, J., & Lee, S. (2015, May). Standard deviation and standard error of the mean. *Korean Journal of Anesthesiology*, *68*(3), 220–223. Retrieved 2024-09-16, from https://synapse.koreamed.org/articles/1156109 (Publisher: The Korean Society of Anesthesiologists) doi: 10.4097/kjae.2015.68.3.220

Mulder, J., Williams, D. R., Gu, X., Tomarken, A., Böing-Messing, F., Olsson-Collentine, A., ... Van Lissa, C. (2021). **BFpack** : Flexible Bayes Factor Testing of Scientific Theories in *R*. *Journal of Statistical Software*, *100*(18). Retrieved 2024-08-22, from https://www.jstatsoft.org/v100/i18/ doi: 10.18637/jss.v100.i18

R Core Team. (2024). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Van Wonderen, E., Zondervan-Zwijnenburg, M., & Klugkist, I. (2024, March). Bayesian evidence synthesis as a flexible alternative to meta-analysis: A simulation study and empirical demonstration. *Behavior Research Methods*, *56*(4), 4085–4102. Retrieved 2024-08-16, from https://link.springer.com/10.3758/s13428-024-02350-2 doi: 10.3758/s13428-024-02350-2

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (Fourth ed.). New York: Springer. Retrieved from https://www.stats.ox.ac.uk/pub/MASS4/

Volker, T. B., & Klugkist, I. (2023, December). *Combining support for hypotheses over heterogeneous studies with Bayesian Evidence Synthesis: A simulation study.* arXiv. Retrieved 2024-08-16, from `http://arxiv.org/abs/2312.15032` (arXiv:2312.15032 [stat])

Wishart, J. (1928). The Generalised Product Moment Distribution in Samples from a Normal Multivariate Population. *Biometrika*, *20A*(1/2), 32–52. Retrieved 2024-09-15, from `https://www.jstor.org/stable/2331939` (Publisher: [Oxford University Press, Biometrika Trust]) doi: 10.2307/2331939