RESEARCH REPORT - WORD COUNT: 2392

Bayesian Evidence Synthesis

Thom Benjamin Volker[1,2]

[1]Supervised by Prof. Dr. Irene Klugkist; [2]Utrecht University

ARTICLE HISTORY
Compiled April 7, 2022

KEYWORDS
Bayes factors, Evidence Aggregation, hypothesis evaluation

1. Introduction

In recent years, a meta-analytic way of thinking has been advocated in the scientific community. This approach is grounded in the belief that a single study is merely contributing to a larger body of evidence (e.g., Asendorpf et al. 2016; Cumming 2014; Goodman, Fanelli, and Ioannidis 2016; Schmidt 2009). In this context, the importance of replication has been legitimately supported (e.g., Baker 2016; Brandt et al. 2014; Munafò et al. 2017). However, most of the attention has been focused on studies that are highly similar, using an identical methodology and research design (e.g., Camerer et al. 2016, 2018; Klein et al. 2014; Nosek et al. 2021; Open Science Collaboration 2015). These studies, commonly referred to as exact, direct or close replications, are merely concerned with the statistical reliability of the results. If the results of the studies depend on methodological flaws, inferences from all studies will lead to suboptimal or invalid conclusions (Lawlor, Tilling, and Davey Smith 2017; Munafò and Smith 2018).

Conceptual replications, which primarily assess the validity of a study, protect against placing too much confidence in such flawed findings. Specifically, conceptual replications are a way of investigating whether the conclusions hold under alternative conditions, using varying measurement instruments or other operationalizations (Nosek, Spies, and Motyl 2012). Different methodologies may have different strengths and weaknesses, that may affect the final conclusion that can be drawn from the data. Combining evidence from multiple approaches mitigates the effect of these strengths and weaknesses, which enhances the validity and the robustness of the final conclusion (Lawlor, Tilling, and Davey Smith 2017; Lipton 2003; Mathison 1988; Munafò and Smith 2018; Nosek, Spies, and Motyl 2012).

In the conventional framework of direct replications, combining evidence over studies is straightforward, because established methods as (Bayesian) meta-analysis or Bayesian sequential updating can be applied to aggregate the results (Lipsey and Wilson 2001; Schönbrodt et al. 2017; Sutton and Abrams 2001). These methods pool the parameter estimates or effect sizes obtained in the individual studies (Cooper, Hedges,

CONTACT Thom Benjamin Volker. Email: t.b.volker@uu.nl

and Valentine 2009). However, if the studies differ considerably with regard to operationalizations of key variables or statistical models used, the parameter estimates or effect sizes will not be comparable. Accordingly, an aggregate of these estimates cannot be meaningfully interpreted, rendering the use of conventional methods unfeasible.

To overcome these difficulties, Kuiper et al. (2013) proposed a new method called *Bayesian Evidence Synthesis* (*BES*). At its core, *BES* quantifies the support for an overall scientific theory or an overarching hypothesis, by aggregating Bayes factors obtained at the level of individual studies. In every single study, a hypothesis can be formulated that reflects an overall theory, but that incorporates data characteristics and methodologies unique to that study. The support for each of these study-specific hypotheses can be expressed using a Bayes factor (BF), rendering the relative support for the hypothesis of interest over an alternative hypothesis (Kass and Raftery 1995). If the study-specific hypotheses represent the same underlying (i.e., latent) effect, the Bayes factors in favor of these hypotheses can be meaningfully combined.

Although *BES* has been applied successfully to substantive research questions (e.g., Kevenaar et al. 2021; Zondervan-Zwijnenburg et al. 2020a,b), hardly any empirical investigation into the method's performance has been conducted. As a consequence, researchers may be hesitant to implement the method in practice, because the required conditions to achieve good performance are relatively unknown. In this paper, it will be assessed how *BES* behaves empirically when combining evidence from studies that employ different statistical models, by focusing on whether *BES* indeed favors true hypotheses over alternatives. Given the fact that studies with low statistical power (i.e., 60%) are omnipresent in social science research (e.g., Ingre and Nilsonne 2018; Button et al. 2013), it will be assessed to what extent *BES* is vulnerable to power issues. Hence, this paper aims to assess how well *BES* performs when aggregating results of different analysis techniques, under varying sample sizes and effect sizes.

2. Methods

2.1. *Informative hypotheses*

At it's core, *BES* aims to quantify the evidence for a scientific theory, by aggregating the relative amount of evidence for the hypothesis of interest over multiple studies. A scientific theory is often best described by an informative hypothesis, which, contrary to the classical null hypothesis testing approach, allows to make scientific expectations explicit (e.g., see Hoijtink 2012; Van De Schoot, Hoijtink, and Romeijn 2011).[1] As an example, consider a research group investigating whether researchers who experience more publication pressure are more often involved in scientific misconduct (e.g., Gopalakrishna et al. 2021). When putting this theory to the test, the research group may measure the amount of publication pressure researchers experience and ask how often these researchers were involved in scientific misconduct, and quantify both measures on a continuous scale. Formalizing the theory as a testable informative hypothesis would yield

$$\mathcal{H}_1 : \beta_p > 0,$$

[1] Other advantages of evaluating informative hypotheses are numerous (e.g., see Béland et al. 2012; Hoijtink, Klugkist, and Boelen 2008; Klaassen 2020; Van de Schoot and Strohmeier 2011; Van de Schoot et al. 2011).

where the hypothesis of interest $\mathcal{H}_1$ states that the effect of publication pressure on scientific misconduct β_p is positive. Another researcher might have data in which the outcome, involvement in scientific misconduct, only consists of two categories (i.e., "yes" or "no"), but that contains two measures for publication pressure that are both expected to be positively related to scientific misconduct: the expected necessity of top-tier publications for successful grant applications and the expected necessity of top-tier publications for getting tenure. An expectation might then be that both sources of publication pressure are positively related to scientific misconduct, and that the effect of both is actually similar, because they are equally important to researchers. Capturing this expectation in an informative hypothesis would yield

$$\mathcal{H}_2 : \{\beta_{p_1} = \beta_{p_2}\} > 0,$$

where β_{p_1} and β_{p_2} indicate the effect of publication pressure through grant applications and getting tenure on scientific misconduct, respectively.

2.2. *Bayes factors*

The Bayes Factor (Kass and Raftery 1995) can be used to quantify the support for informative hypotheses (Béland et al. 2012; Hoijtink 2012; Hoijtink et al. 2019). Loosely speaking, a Bayes factor expresses the relative support from the data for an hypothesis of interest versus some alternative hypothesis. Returning to the example, the Bayes factor for hypothesis $\mathcal{H}_1$ versus an alternative hypothesis $\mathcal{H}_a$ is defined as the ratio of the marginal likelihoods of the data under these hypotheses

$$BF_{1,a} = \frac{P(D|\mathcal{H}_1)}{P(D|\mathcal{H}_a)}.$$

A Bayes factor that equals 7 indicates that after seeing the data, $\mathcal{H}_1$ is 7 times more likely than $\mathcal{H}_a$. When the alternative hypothesis is unconstrained (i.e., $\mathcal{H}_a = \mathcal{H}_u$), the expression of the Bayes factor boils down to

$$BF_{1,u} = \frac{f_1}{c_1},$$

where f_1 indicates the fit of the data to $\mathcal{H}_1$ and c_1 indicates the complexity of the hypothesis $\mathcal{H}_1$ (e.g., Béland et al. 2012; Klugkist, Laudy, and Hoijtink 2005). A similar approach can be used when the hypothesis of interest contains equality constraints (like hypothesis $\mathcal{H}_2$ in the example), but requires some additional steps that will not be discussed here (see e.g., Mulder, Hoijtink, and Klugkist 2010; Klugkist, Laudy, and Hoijtink 2005). When the alternative hypothesis of interest is another informative hypothesis $\mathcal{H}_3$ (e.g., $\beta_p < 0$), the same formulation of the Bayes factor can be applied to both hypotheses, which is generally easier than calculating the marginal likelihoods under both hypotheses (Klugkist, Laudy, and Hoijtink 2005), providing

$$BF_{1,3} = \frac{BF_{1,u}}{BF_{3,u}}.$$

Combining the obtained Bayes factors with prior model probabilities (i.e., the probability that the hypotheses under consideration are true *a priori*), allows to calculate

posterior model probabilities. The posterior model probabilities for hypothesis $\mathcal{H}_i$, with $i \in \{1, 2, \ldots, m\}$, are given by

$$PMP(\mathcal{H}_i) = \frac{\pi_i BF_{i,u}}{\sum_{i'=1}^{m} \pi_{i'} BF_{i',u}},$$

where π_i indicates the prior model probability of hypothesis H_i. The posterior model probabilities indicate the relative support for the hypotheses under consideration.

2.3. *Bayesian evidence synthesis*

Regardless of differences in study design and analysis plan, a Bayes factor can be calculated to express the support for the theory in each of the studies under consideration, as long as the hypotheses under consideration are conceptually similar. To aggregate the evidence for a theory over multiple studies, *BES* uses the possibility to specify the prior model probabilities. After every study that is conducted, the support for the hypothesis of interest can be expressed as a posterior model probability. These posterior model probabilities after study j can be used as prior model probabilities in study $j + 1$ (Kuiper et al. 2013). Independent of the order of updating, repeating this process for J studies yields

$$PMP(\mathcal{H}_i)^J = \frac{\pi_i^0 \prod_{j=1}^{J} BF_{i,u}^j}{\sum_{i'=1}^{m} \pi_{i'}^0 \prod_{j=1}^{J} BF_{i',u}^j},$$

where π_i^0 indicate the prior model probabilities for hypothesis $\mathcal{H}_i$ before a study has been conducted. Because one could argue that before observing any data all hypotheses are equally likely, π_i^0 is usually the same for all hypotheses (i.e., $\pi_i^0 = \frac{1}{m}$). If the initial prior model probabilities are the same for all hypotheses, the product of Bayes factors equals the aggregated posterior model odds (the numerator of $PMP(\mathcal{H}_i)^J$), which contains the same information as the aggregated posterior model probabilities. These measures then indicate the relative support for the theory (i.e., overall hypothesis) in all studies simultaneously, irrespective of the specifics of each study.

2.4. *Simulation*

A simulation study is conducted in `R` (R Core Team 2021, Version 4.1.0) to assess how the outcome of *BES* is affected by sample size and effect size. Because *BES* can be applied regardless of between-study heterogeneity, data to represent three different studies is generated using three models: ordinary least squares (OLS), logistic and probit regression. In all studies, the *true* hypothesis $\mathcal{H}_1 : \beta_1 < \beta_2 < \beta_3 < \beta_4$ is evaluated against an unconstrained alternative $\mathcal{H}_u : \beta_1, \beta_2, \beta_3, \beta_4$ using the `R`-package `BFpack` (Mulder et al. 2021, Version 0.3.2), and *BES* is used to pool the evidence over studies. Because the initial prior model probabilities are specified equally (i.e., $\pi_1^0 = \pi_u^0 = \frac{1}{2}$), and the product of Bayes factors is log-transformed for interpretability, a product that is greater than 0 yields more support for $\mathcal{H}_1$ than for $\mathcal{H}_u$.

Each study consists of a continuous or binary outcome Y and a predictor matrix X, with dimensions $n \times 4$. The sample sizes vary with $n \in \{25, 50, 100, 200, 400, 800\}$, while the effect sizes take on the values $R^2 \in \{0.02, 0.09, 0.25\}$, in accordance with small, medium and large effects as defined by Cohen (1988). For binary outcomes,

McKelvey and Zavoina's $R^2_{M\&Z}$ is used (McKelvey and Zavoina 1975), which closely resembles the conventional R^2 empirically (Hagle and Mitchell 1992; DeMaris 2002). Two sets of simulations will be considered: in the first, the sample sizes and effect sizes are consistent across studies, while in the second, the sample size of a single, randomly selected, study is kept fixed at $n = 25$. All simulation conditions are evaluated over 1000 iterations. All predictor variables are normally distributed with a mean of 0, a variance of 1 and a common covariance of 0.3, and their relation with the outcome Y is specified according to the weights-matrix $B = [1, 2, 3, 4]^T$, such that $\beta_4 = 4\beta_1, \beta_3 = 3\beta_1, \beta_2 = 2\beta_1$. Accordingly the exact coefficients can be defined as

$$\beta = B\left(\sqrt{\frac{\text{Var}(\hat{Y})}{G^T((BB^T) \odot \Sigma)G}}\right),$$

where G is a 4×1 column-vector of ones and $\text{Var}(\hat{Y}) = \text{Var}(X\beta)$ is defined as a function of the effect size.[2] The population-level regression coefficients are displayed in Table 1. When the outcome variable is continuous, Y is drawn from a normal distribution

$$Y \sim \mathcal{N}(X\beta, 1 - R^2),$$

with a mean vector of $X\beta$ and residual variance $\sigma^2 = 1 - R^2$. When Y is binary, Y is drawn from a Bernoulli distribution

$$Y \sim \mathcal{B}(p),$$

and p equals

$$p_{\text{logit}} = \frac{1}{e^{-X\beta}}, \qquad p_{\text{probit}} = \Phi(X\beta),$$

for logistic and probit regression, respectively, and with Φ indicating the cumulative normal distribution.

Table 1. Population-level regression coefficients for ordinary least squares (OLS), logistic and probit regression, given effect sizes of $R^2 \in \{0.02, 0.09, 0.25\}$.

R^2	OLS				Logistic				Probit			
	β_1	β_2	β_3	β_4	β_1	β_2	β_3	β_4	β_1	β_2	β_3	β_4
0.02	0.02	0.04	0.06	0.08	0.04	0.07	0.11	0.15	0.02	0.04	0.06	0.08
0.09	0.04	0.08	0.13	0.17	0.08	0.16	0.24	0.32	0.04	0.09	0.13	0.18
0.25	0.07	0.14	0.21	0.28	0.15	0.29	0.44	0.59	0.08	0.16	0.24	0.32

3. Results

The simulations show that the aggregated Bayes factors generally tend to increase with the sample size and effect size. In Simulation 1 in Figure 1, the log-transformed

[2]Note that $\text{Var}(\hat{Y}) = R^2$ when Y is continuous with a variance of $\sigma^2_Y = 1$, while $\text{Var}(\hat{Y}) = \frac{R^2\frac{\pi^2}{3}}{(1-R^2)}$ and $\text{Var}(\hat{Y}) = \frac{R^2}{1-R^2}$ for logistic and probit regression, respectively.

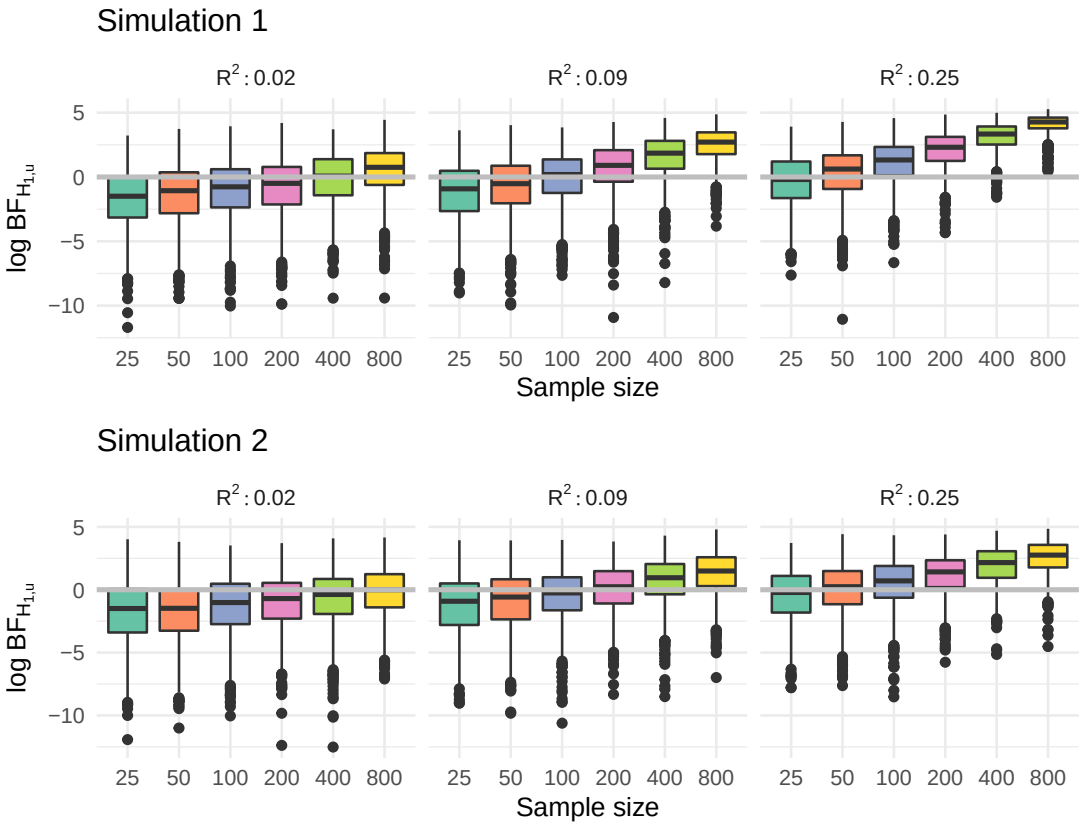

Figure 1. Logarithm of the product of Bayes factors for the true hypothesis of interest versus an unconstrained hypothesis over three studies (generated with ordinary least squares (OLS), logistic and probit regression), for multiple sample sizes and effect sizes. In Simulation 2 the sample size of one randomly selected study is kept fixed at $n = 25$.

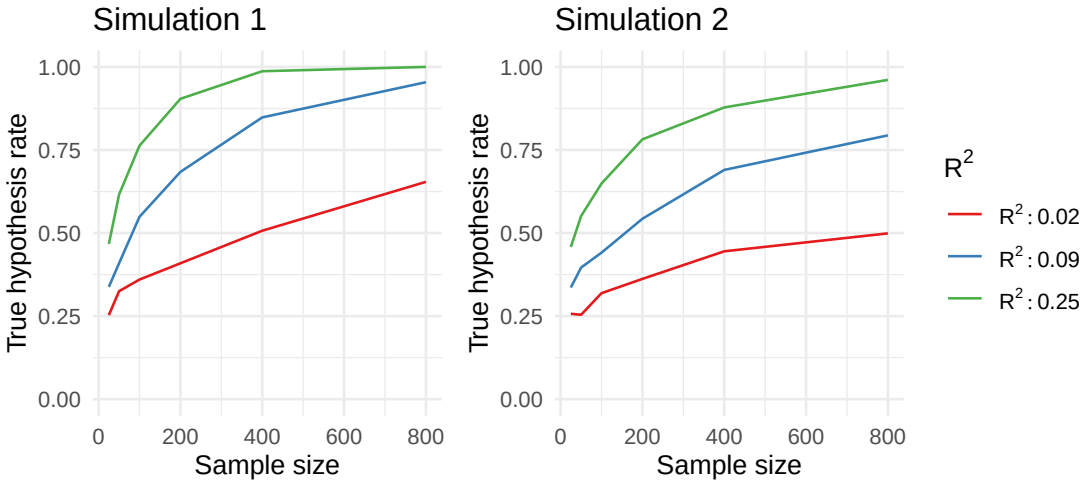

Figure 2. True hypothesis rate of *BES* over three studies (generated with ordinary least squares (OLS), logistic and probit regression), for multiple sample sizes and effect sizes. In Simulation 2 the sample size of randomly selected study is kept fixed at $n = 25$.

product of Bayes factor over the three studies $\log(BF_{1,u})$ is centered above zero only when the sample size equals $n = 800$ when the effect is small (i.e., $R^2 = 0.02$). Hence, with a small effect size, only the largest sample size considered renders more support for $\mathcal{H}_1$ than for $\mathcal{H}_u$ in the majority of the simulations. When $R^2 = 0.09$ or $R^2 = 0.25$, $\log(BF_{1,u})$ is centered around zero when $n = 100$ and $n = 25$, respectively, and shows convincing support for the true hypothesis $\mathcal{H}_1$ in nearly all simulations for large sample sizes.

In Simulation 2 in Figure 1, where one of the three studies is randomly selected to have a sample size of only $n = 25$ observations, $\log(BF_{1,u})$ is somewhat lower overall as compared to Simulation 1 (ignoring those simulations in which the sample size of all studies equals $n = 25$). This is to be expected, because with $n = 25$ the unconstrained hypothesis generally obtains more support than hypothesis $\mathcal{H}_1$, for all effect sizes. Apart from this, a similar trend can be observed, in the sense that the aggregated Bayes factors tend to increase with sample size and effect size. However, the point from which the true hypothesis $\mathcal{H}_1$ receives most support in the majority of the simulations (i.e., the point from which $\log(BF_{1,u})$ is centered above zero) is generally reached later than in Simulation 1, and $\mathcal{H}_1$ is convincingly supported in most simulations only for large sample sizes (i.e., $n \geq 400$) and a large effect (i.e., $R^2 = 0.25$).

A similar picture is shown in Figure 2, which displays the true hypothesis rates for the hypothesis of interest $\mathcal{H}_1$ using the product of Bayes factors for both simulation set-ups. The true hypothesis rate increases with the sample size and effect size. In Simulation 1, when $n = 25$ the true hypothesis receives most support in only a minority of the simulations for all effect sizes, while for $n = 100$ only the studies with medium or large effects render most support for $\mathcal{H}_1$ in the majority of the simulations (i.e., about 50% and 80%, respectively). For $n = 800$ all effect sizes yield most support for the true hypothesis in the majority of simulations. When the set of studies contains a single study with a small sample size (i.e., $n = 25$; in Simulation 2), the increase in the true hypothesis rate is less steep and $\mathcal{H}_1$ is supported less often. Apart from this, the same general patterns can be observed as in Simulation 1.

4. Discussion and conclusion

Overall, *BES* showed to have satisfactory properties, in the sense that it tends to select the true hypothesis when the power of the studies is sufficiently large. In that sense, it proves to be a welcome addition to conventional methods of research synthesis, like Bayesian updating of parameter estimates or (Bayesian) meta-analysis. In contrast to these methods, *BES* can be applied when the study designs differ (e.g., through the use of different statistical models, as presented here). However, rather than pooled parameter estimates, *BES* provides the relative support for the hypotheses under consideration over the entire set of studies simultaneously. That is, *BES* quantifies to what extent the theory of interest is supported in each of these studies.

Consequently, unlike meta-analysis and Bayesian sequential updating, which assess to what extent all studies together provide evidence for a hypothesis, *BES* cannot solve power issues. In fact, if the set of studies contains a single, severely underpowered, study, this might already affect the performance of *BES*. Our second simulation indeed suggests that including a single underpowered study, which is the case for studies with a sample size of $n = 25$ and a regression model containing 4 predictors, lowers the performance of *BES* substantially. This finding warrants further research into the

severity of the effect of statistical power on the performance of *BES*. Moreover, it provides a cautionary note for researchers who want to apply *BES*: a synthesis of studies that lack statistical power might yield invalid conclusions.

The current study focused exclusively on the effects of the sample size and effect size on the performance of *BES*. However, other factors may likewise affect the outcome of *BES*. When aggregating results of different studies, the complexities of the study-specific hypotheses may differ due different operationalizations of key variables. As Bayes factors reflect a trade-off between the fit of the data to the hypothesis and the complexity of the hypothesis, differences in complexity may affect the outcome of *BES*. Additionally, in our simulations, we compared our hypothesis of interest with an unconstrained alternative hypothesis. However, evaluating against the classical null hypothesis, another informative hypothesis or the complement hypothesis can be reasonable choices, too. To what extent such factors will influence the performance of *BES* is not yet understood.

References

Asendorpf, Jens B., Mark Conner, Filip de Fruyt, Jan De Houwer, Jaap J. A. Denissen, Klaus Fiedler, Susann Fiedler, et al. 2016. *Recommendations for increasing replicability in psychology.* Methodological issues and strategies in clinical research, 4th ed. Washington, DC, US: American Psychological Association. `https://doi.org/10.1037/14805-038`.

Baker, Monya. 2016. "Reproducibility crisis." *Nature* 533 (26): 353–66.

Béland, Sébastien, Irene Klugkist, Gilles Raîche, and David Magis. 2012. "A short introduction into Bayesian evaluation of informative hypotheses as an alternative to exploratory comparisons of multiple group means." *Tutorials in Quantitative Methods for Psychology* 8 (2): 122–126.

Brandt, Mark J, Hans IJzerman, Ap Dijksterhuis, Frank J Farach, Jason Geller, Roger Giner-Sorolla, James A Grange, Marco Perugini, Jeffrey R Spies, and Anna Van't Veer. 2014. "The replication recipe: What makes for a convincing replication?" *Journal of Experimental Social Psychology* 50: 217–224.

Button, Katherine S, John PA Ioannidis, Claire Mokrysz, Brian A Nosek, Jonathan Flint, Emma SJ Robinson, and Marcus R Munafò. 2013. "Power failure: why small sample size undermines the reliability of neuroscience." *Nature reviews neuroscience* 14 (5): 365–376.

Camerer, Colin F, Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, et al. 2016. "Evaluating replicability of laboratory experiments in economics." *Science* 351 (6280): 1433–1436.

Camerer, Colin F, Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, et al. 2018. "Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015." *Nature Human Behaviour* 2 (9): 637–644.

Cohen, J. 1988. *Statistical power analysis for the behavioral sciences.* 2nd ed. Lawrence Erlbaum Associates.

Cooper, Harris, Larry Vernon Hedges, and Jeffrey C Valentine. 2009. *The handbook of research synthesis and meta-analysis 2nd edition.* Russell Sage Foundation.

Cumming, Geoff. 2014. "The new statistics: Why and how." *Psychological science* 25 (1): 7–29.

DeMaris, Alfred. 2002. "Explained Variance in Logistic Regression: A Monte Carlo Study of Proposed Measures." *Sociological Methods & Research* 31 (1): 27–74. `https://doi.org/10.1177/0049124102031001002`.

Goodman, Steven N., Daniele Fanelli, and John P. A. Ioannidis. 2016. "What does research reproducibility mean?" *Science Translational Medicine* 8 (341): 341ps12–341ps12. `https://www.science.org/doi/abs/10.1126/scitranslmed.aaf5027`.

Gopalakrishna, Gowri, Gerben t Riet, Gerko Vink, Ineke Stoop, Jelte M Wicherts, and Lex

Bouter. 2021. "Prevalence of questionable research practices, research misconduct and their potential explanatory factors: a survey among academic researchers in The Netherlands." Jul. osf.io/preprints/metaarxiv/vk9yt.

Hagle, Timothy M., and Glenn E. Mitchell. 1992. "Goodness-of-Fit Measures for Probit and Logit." *American Journal of Political Science* 36 (3): 762–784. http://www.jstor.org/stable/2111590.

Hoijtink, Herbert. 2012. *Informative hypotheses: Theory and practice for behavioral and social scientists*. CRC Press.

Hoijtink, Herbert, Irene Klugkist, and Paul A. Boelen. 2008. *Bayesian evaluation of informative hypotheses*. Springer. https://link.springer.com/book/10.1007/978-0-387-09612-4.

Hoijtink, Herbert, Joris Mulder, Caspar van Lissa, and Xin Gu. 2019. "A tutorial on testing hypotheses using the Bayes factor." *Psychological methods* 24 (5): 539.

Ingre, Michael, and Gustav Nilsonne. 2018. "Estimating statistical power, posterior probability and publication bias of psychological research using the observed replication rate." *Royal Society Open Science* 5 (9): 181190. https://royalsocietypublishing.org/doi/abs/10.1098/rsos.181190.

Kass, Robert E., and Adrian E. Raftery. 1995. "Bayes Factors." *Journal of the American Statistical Association* 90 (430): 773–795. https://www.tandfonline.com/doi/abs/10.1080/01621459.1995.10476572.

Kevenaar, Sofieke T., Maria A.J. Zondervan-Zwijnenburg, Elisabet Blok, Heiko Schmengler, M. (Ties) Fakkel, Eveline L. de Zeeuw, Elsje van Bergen, et al. 2021. "Bayesian evidence synthesis in case of multi-cohort datasets: An illustration by multi-informant differences in self-control." *Developmental Cognitive Neuroscience* 47: 100904. https://www.sciencedirect.com/science/article/pii/S1878929320301535.

Klaassen, Fayette. 2020. "The latest update on Bayesian informative hypothesis testing." PhD diss., Utrecht University. https://fayetteklaassen.github.io/files/dissertation.pdf.

Klein, Richard A., Kate A. Ratliff, Michelangelo Vianello, Reginald B. Adams, Štěpán Bahník, Michael J. Bernstein, Konrad Bocian, et al. 2014. "Investigating Variation in Replicability." *Social Psychology* 45 (3): 142–152. https://doi.org/10.1027/1864-9335/a000178.

Klugkist, Irene, Olav Laudy, and Herbert Hoijtink. 2005. "Inequality constrained analysis of variance: a Bayesian approach." *Psychological methods* 10 (4): 477–493.

Kuiper, Rebecca M., Vincent Buskens, Werner Raub, and Herbert Hoijtink. 2013. "Combining Statistical Evidence From Several Studies: A Method Using Bayesian Updating and an Example From Research on Trust Problems in Social and Economic Exchange." *Sociological Methods & Research* 42 (1): 60–81. https://doi.org/10.1177/0049124112464867.

Lawlor, Debbie A., Kate Tilling, and George Davey Smith. 2017. "Triangulation in aetiological epidemiology." *International Journal of Epidemiology* dyw314. Accessed 2020-07-30. https://academic.oup.com/ije/article-lookup/doi/10.1093/ije/dyw314.

Lipsey, Mark W, and David B Wilson. 2001. *Practical meta-analysis*. SAGE publications, Inc.

Lipton, Peter. 2003. *Inference to the best explanation*. Routledge.

Mathison, Sandra. 1988. "Why triangulate?" *Educational researcher* 17 (2): 13–17.

McKelvey, Richard D., and William Zavoina. 1975. "A statistical model for the analysis of ordinal level dependent variables." *The Journal of Mathematical Sociology* 4 (1): 103–120. https://doi.org/10.1080/0022250X.1975.9989847.

Mulder, Joris, Herbert Hoijtink, and Irene Klugkist. 2010. "Equality and inequality constrained multivariate linear models: Objective model selection using constrained posterior priors." *Journal of Statistical Planning and Inference* 140 (4): 887–906. https://www.sciencedirect.com/science/article/pii/S0378375809003127.

Mulder, Joris, Caspar van Lissa, Donald R. Williams, Xin Gu, Anton Olsson-Collentine, Florian Boeing-Messing, and Jean-Paul Fox. 2021. *BFpack: Flexible Bayes Factor Testing of Scientific Expectations*. R package version 0.3.2, https://CRAN.R-project.org/package=BFpack.

Munafò, Marcus R., Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware, and John P. A. Ioannidis. 2017. "A manifesto for reproducible science." *Nature Human Behaviour* 1 (1): 0021. https://doi.org/10.1038/s41562-016-0021.

Munafò, Marcus R., and George Davey Smith. 2018. "Robust research needs many lines of evidence." *Nature* 553 (7689): 399–401. Number: 7689 Publisher: Nature Publishing Group, Accessed 2020-08-24. https://www.nature.com/articles/d41586-018-01023-3.

Nosek, Brian A., Tom E. Hardwicke, Hannah Moshontz, Aurélien Allard, Katherine S. Corker, Anna Dreber, Fiona Fidler, et al. 2021. "Replicability, Robustness, and Reproducibility in Psychological Science." *Annual Review of Psychology* 73 (1): null. PMID: 34665669, https://doi.org/10.1146/annurev-psych-020821-114157.

Nosek, Brian A., Jeffrey R. Spies, and Matt Motyl. 2012. "Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability." *Perspectives on Psychological Science* 7 (6): 615–631. PMID: 26168121, https://doi.org/10.1177/1745691612459058.

Open Science Collaboration. 2015. "Estimating the reproducibility of psychological science." *Science* 349 (6251). Publisher: American Association for the Advancement of Science _eprint: https://science.sciencemag.org/content/349/6251/aac4716.full.pdf, https://science.sciencemag.org/content/349/6251/aac4716.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Schmidt, Stefan. 2009. "Shall we Really do it Again? The Powerful Concept of Replication is Neglected in the Social Sciences." *Review of General Psychology* 13 (2): 90–100.

Schönbrodt, Felix D, Eric-Jan Wagenmakers, Michael Zehetleitner, and Marco Perugini. 2017. "Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences." *Psychological methods* 22 (2): 322.

Sutton, Alex J, and Keith R Abrams. 2001. "Bayesian methods in meta-analysis and evidence synthesis." *Statistical methods in medical research* 10 (4): 277–303.

Van De Schoot, Rens, Herbert Hoijtink, and Jan-Willem Romeijn. 2011. "Moving Beyond Traditional Null Hypothesis Testing: Evaluating Expectations Directly." *Frontiers in Psychology* 2: 24.

Van de Schoot, Rens, Joris Mulder, Herbert Hoijtink, Marcel A. G. Van Aken, Judith Semon Dubas, Bram Orobio de Castro, Wim Meeus, and Jan-Willem Romeijn. 2011. "An introduction to Bayesian model selection for evaluating informative hypotheses." *European Journal of Developmental Psychology* 8 (6): 713–729. https://doi.org/10.1080/17405629.2011.621799.

Van de Schoot, Rens, and Dagmar Strohmeier. 2011. "Testing informative hypotheses in SEM increases power: An illustration contrasting classical hypothesis testing with a parametric bootstrap approach." *International Journal of Behavioral Development* 35 (2): 180–190. https://doi.org/10.1177/0165025410397432.

Zondervan-Zwijnenburg, M. A. J., Sabine A.M. Veldkamp, Alexander Neumann, Stefania A. Barzeva, Stefanie A. Nelemans, Catharina E.M. van Beijsterveldt, Susan J.T. Branje, et al. 2020a. "Parental Age and Offspring Childhood Mental Health: A Multi-Cohort, Population-Based Investigation." *Child Development* 91 (3): 964–982.

Zondervan-Zwijnenburg, M.A.J., J.S. Richards, S.T. Kevenaar, A.I. Becht, H.J.A. Hoijtink, A.J. Oldehinkel, S. Branje, W. Meeus, and D.I. Boomsma. 2020b. "Robust longitudinal multi-cohort results: The development of self-control during adolescence." *Developmental Cognitive Neuroscience* 45: 100817.