# Tutorial: Aggregate evidence from heterogeneous replication studies using the product Bayes factor

Caspar J. Van Lissa[1,2] & Eli-Boaz Clapper[1]

[1] Utrecht University, dept. Methodology & Statistics

[2] Open Science Community Utrecht

## Author Note

This is a preprint paper, generated from Git Commit # b0c70e6. This work was funded by a NWO Veni Grant (NWO Grant Number VI.Veni.191G.090), awarded to the lead author.

The authors made the following contributions. Caspar J. Van Lissa: Conceptualization, Formal Analysis, Funding acquisition, Methodology, Project administration, Software, Supervision, Writing – original draft, Writing – review & editing; Eli-Boaz Clapper: Formal Analysis, Writing – original draft, Writing – review & editing.

Correspondence concerning this article should be addressed to Caspar J. Van Lissa, Padualaan 14, 3584CH Utrecht, The Netherlands. E-mail: c.j.vanlissa@uu.nl

Abstract

TODO

*Keywords:* bayes factor, evidence synthesis, bayesian, meta-analysis

Word count: 5356

# Tutorial: Aggregate evidence from heterogeneous replication studies using the product Bayes factor

Recent years have seen a crisis of confidence over the reliability of published results in psychology, and science more broadly [REF]. Replication research has come into focus as a solution to this crisis and a way to derive knowledge that will stand the test of time (see Lavelle, 2021). In step with this interest in replication research, research synthesis methods have blossomed [REF]. These methods aggregate the findings of multiple studies, and thus enable drawing overarching conclusions across multiple studies. Both quantitative (e.g., Van Lissa & van Erp, 2021) and qualitative (e.g., van Lissa, 2021) research synthesis methods exist. The present paper focuses on quantitative methods.

A key challenge in quantitative research synthesis is dealing with between-studies heterogeneity (Higgins, Thompson, & Spiegelhalter, 2009). Heterogeneity appears when, for example, studies examine the same research question in different laboratories, use idiosyncratic methods, and sample from distinct populations. Meta-analysis is the most common research synthesis method (Borenstein, Hedges, Higgins, & Rothstein, 2009). In meta-analysis, heterogeneity can be accounted for in four ways (see Van Lissa, 2020). First, if studies are exact replications, one can assume that no heterogeneity exists and conduct a fixed-effect meta-analysis to estimate their common population effect. Second, when differences between studies can be assumed to be random, random-effects meta-analysis can be used to estimate the mean of a distribution of population effects. Third, when there are a few systematic differences between studies, these can be accounted for using meta-regression. And finally, when there are many potential differences and it is not known beforehand which are relevant, exploratory techniques like random forest meta-analysis and penalized meta-regression can be used to identify relevant moderators. Each of these approaches requires making different assumptions about the nature of heterogeneity.

An alternative approach that does not impose such assumptions is Bayesian evidence

synthesis (BES, Kuiper, Buskens, Raub, & Hoijtink, 2013). BES focuses on the aggregation of evidence in favor of an informative hypothesis $H_i$ across studies. The amount of evidence for this hypothesis is expressed as a Bayes factor, BF. A Bayes factor can be interpreed as the ratio of evidence in favor of $H_i$ divided by evidence against it. Thus, BF = 10 means that the data provide ten times more support in favor of the hypothesis than against it. These Bayes factors can be synthesized across studies by taking their product. The resulting product Bayes factor PBF summarizes the total evidence for the hypothesis. Note that other approaches to BES exist (see "Bayesian Evidence Synthesis" in Heck et al., 2022).

This tutorial paper introduces the first implementation of BES in userfriendly open source software. A function `pbf()` was contributed to the `bain` R-package for Bayesian informative hypothesis evaluation, version 0.2.8. This paper presents a simulation study to validate the method and benchmark it against alternative evidence synthesis methods, and illustrates several use cases through reproducible examples.

## Simulation study

The present simulation study set out to validate the PBF algorithm. For each iteration of the simulation, we simulated a correlation coefficient in multiple samples. The informative hypothesis was kept constant at $H_i : \rho > .1$. All simulation conditions were evaluated once in the presence of a true population effect, defined as $\rho = .2$, and once in the presence of a null effect, defined as $\rho = .1$. A PBF $> 3$ was used as a decision criterion to conclude that $H_i$ was supported. As a benchmark for comparison, we used several other algorithms that might feasibly be used by researchers who intend to examine whether a hypothesis is true across several independent samples. The first was a "vote count" algorithm, which is common but not considered to be good practice [REF]. This approach used one-sided z-tests to examine whether a null hypothesis corresponding to the informative hypothesis was rejected in the majority of samples, i.e.: $H_0 : \rho = .1$. The

second was a random-effects meta-analysis (RMA), which is the standard in the field. For this algorithm, the null-hypothesis was rejected if a 90% confidence interval for the overall effect size excluded $H_0$. Third was an individual participant data (IPD) meta-analysis. Like classic meta-analysis, IPD is a multilevel model, clustered by sample. In contrast to classic meta-analysis, however, IPD freely estimates variance at the first level, because raw data are available. The PBF can be estimated using either sufficient statistics (as in meta-analysis) or using raw data (as in IPD). It is thus informative to compare the PBF to both these methods. IPD was also evaluated using a 90% confidence interval for the overall effect size.

## Performance indicators

For each algorithm, a confusion matrix was obtained by tabulating inferential decisions made using the criteria described above against the population status of the hypothesis (true or false). This matrix gives the number of decisions that were true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). These quantities were summarized as sensitivity, $1 - \frac{FN}{TP+FN}$, the probability of concluding the hypothesis is false given that it was truly false in the population, and specificity, $1 - \frac{FP}{FP+TN}$, the probability of concluding the hypothesis is true given that it was true in the population. The overall performance was captured by the accuracy, which represents the total proportion of correct (true positive/true negative) decisions, $\frac{TP+TN}{TP+TN+FP+FN}$.

## Design factors

To examine performance in a range of realistic scenarios, several design factors were manipulated: The sample size per group $n \in (20, 80, 200, 500)$, the number of groups $k \in (2, 3, 10)$, where 2 is the minimum number of groups and 10 is a very large number of groups, and the reliability of the two correlated variables, $\alpha \in (0.6, 0.8, 1.0)$, where 0.6 is the lowest reliability conventionally considered to be acceptable, and 1 represents perfect

reliability, as is the assumed when analyzing correlations between observed items or scale scores. The design factors combined to produce unique conditions. For all simulation conditions, 1000 data sets were generated.

## Results

First, we examined overall model performance across conditions. These results indicate that all algorithms except PBF had low sensitivity to detect a true effect. In contrast, specificity was very high for all algorithms except PBF. This suggests that the other algorithms classified most conditions as negatives (no effect found), regardless of the existence of a population effect. The PBF trades a loss of specificity for increased sensitivity, and average levels of both were approximately equal. As the PBF vastly outperformed other BFs, these were omitted from further analysis.

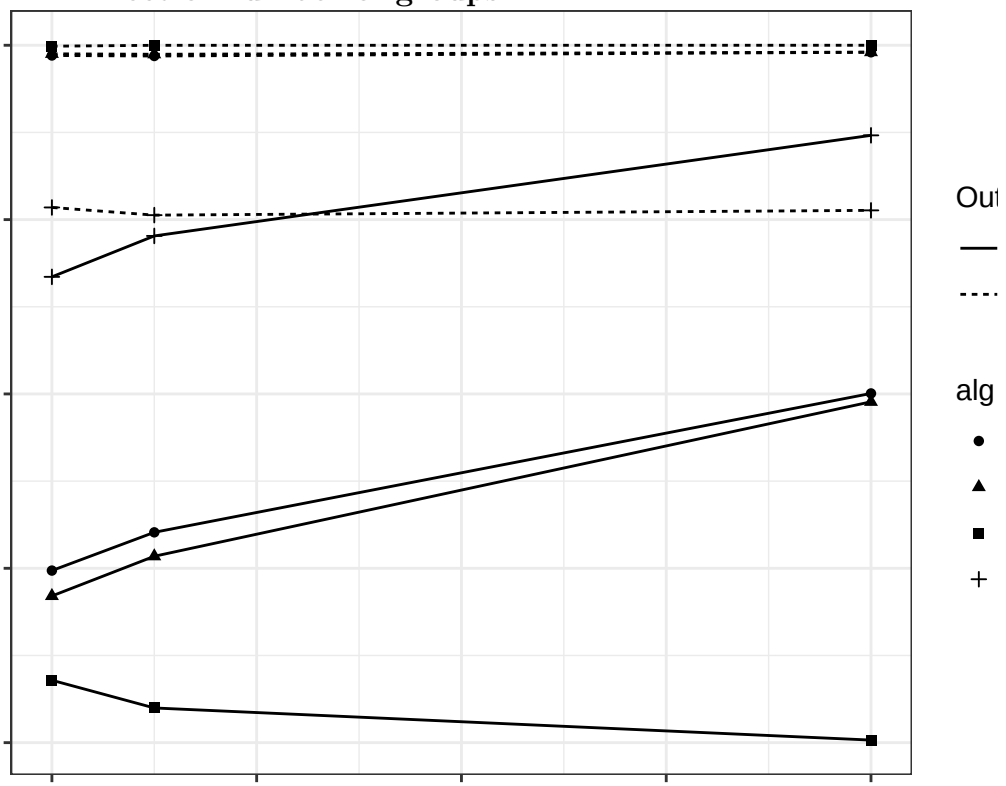
### Effect of simulation conditions

We examined the effect of simulation conditions on overall accuracy.

PBF performance was most impacted by sample size $n$, followed by the number of groups $k$, and reliability.

**Effect of sample size.** Figure 1 indicates that for PBF, both sensitivity and specificity increase with sample size. The other algorithms also show increasing sensitivity, but not specificity, which is at a ceiling. This difference explains the effect of reliability on the diference between algorithms (see Table 2).

**Effect of number of groups.**




The figure indicates that, for PBF at higher levels of k, lower sensitivity is exchanged for greater specificity. The other algorithms do not show this pattern, as their specificity is at a ceiling. This difference in pattern of effects explains why number of groups has a moderate effect on the difference between algorithms (see Table 2). Only VC shows decreasing sensitivity with an increasing number of groups; this is because the probability of obtaining any false negatives increases with the number of groups.

**Effect of reliability.** The figure indicates that for PBF, at higher reliability, lower sensitivity is exchanged for greater specificity. The other algorithms do not show this pattern, as their specificity is at a ceiling. Their sensitivity decreases with lower reliability, and therefore, so does their overall performance. This difference in pattern of effects explains why reliability has a moderate effect on the difference between algorithms (see Table 2). Note that, in contrast to other algorithms, the performance of PBF is less susceptible to reliability.

## Discussion

Within the scope of this simulation, PBF had relatively better inferential properties than the other algorithms under consideration. Although other algorithms had superior specificity, PBF had the highest levels of sensitivity. Thus, PBF balanced the ability to accept an informative hypothesis in the presence of a true population effect with the ability to reject the hypothesis when there was no effect.

An important caveat is that none of the algorithms met conventional criteria for power (80%) and type I error (5%). Overall, PBF had the best performance, with power of 76% and type I error of 24%. Thus, if researchers intend to synthesize evidence for an informative hypothesis across heterogeneous studies, PBF may be the most suitable method - but its limitations should be acknowledged in applied research.

The overall performance of PBF increased most with increasing sample size. With an increasing number of groups, slightly decreasing specificity was traded off for increasing sensitivity. With increasing reliability, increasing specificity was traded off for decreasing sensitivity.

## Results

## Discussion

148 **References**

149 Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009).
150 *Introduction to Meta-Analysis*. John Wiley & Sons, Ltd.
151 https://doi.org/10.1002/9780470743386

152 Heck, D. W., Boehm, U., Böing-Messing, F., Bürkner, P.-C., Derks, K., Dienes, Z.,
153 ... Hoijtink, H. (2022). A review of applications of the Bayes factor in
154 psychological research. *Psychological Methods*.
155 https://doi.org/10.1037/met0000454

156 Higgins, J. P. T., Thompson, S. G., & Spiegelhalter, D. J. (2009). A re-evaluation of
157 random-effects meta-analysis. *Journal of the Royal Statistical Society. Series A,*
158 *(Statistics in Society)*, *172*(1), 137–159.
159 https://doi.org/10.1111/j.1467-985X.2008.00552.x

160 Kuiper, R. M., Buskens, V., Raub, W., & Hoijtink, H. (2013). Combining
161 Statistical Evidence From Several Studies: A Method Using Bayesian Updating
162 and an Example From Research on Trust Problems in Social and Economic
163 Exchange. *Sociological Methods & Research*, *42*(1), 60–81.
164 https://doi.org/10.1177/0049124112464867

165 Lavelle, J. S. (2021). When a Crisis Becomes an Opportunity: The Role of
166 Replications in Making Better Theories. *The British Journal for the Philosophy*
167 *of Science*, 714812. https://doi.org/10.1086/714812

168 van Lissa, C. J. (2021). Mapping Phenomena Relevant to Adolescent Emotion
169 Regulation: A Text-Mining Systematic Review. *Adolescent Research Review*.
170 https://doi.org/10.1007/s40894-021-00160-7

171 Van Lissa, C. J. (2020). Small sample meta-analyses: Exploring heterogeneity using
172 MetaForest. In R. Van De Schoot & M. Miočević (Eds.), *Small Sample Size*
173 *Solutions (Open Access): A Guide for Applied Researchers and Practitioners*.
174 CRC Press.

Van Lissa, C. J., & van Erp, S. (2021). *Select relevant moderators using Bayesian regularized meta-regression* (Preprint). PsyArXiv. https://doi.org/10.31234/osf.io/6phs5

Table 1

*Marginal confusion matrix metrics.*

| Metric | PBF | IPD | RMA | VC |
|---|---|---|---|---|
| sensitivity | 0.76 | 0.35 | 0.32 | 0.05 |
| specificity | 0.76 | 0.99 | 0.99 | 1.00 |
| accuracy | 0.76 | 0.67 | 0.66 | 0.52 |

Table 2

*Partial eta squared of the effect of each design factor on accuracy for each algorithm and for the difference between PBF and all other algorithms (e.g., vs RMA).*

| condition | IPD | RMA | VC | PBF | vs IPD | vs RMA | vs VC |
|---|---|---|---|---|---|---|---|
| k | 0.35 | 0.40 | 0.13 | 0.32 | 0.01 | 0.02 | 0.23 |
| n | 0.60 | 0.58 | 0.29 | 0.62 | 0.01 | 0.00 | 0.19 |
| reliability | 0.62 | 0.61 | 0.23 | 0.04 | 0.27 | 0.25 | 0.01 |

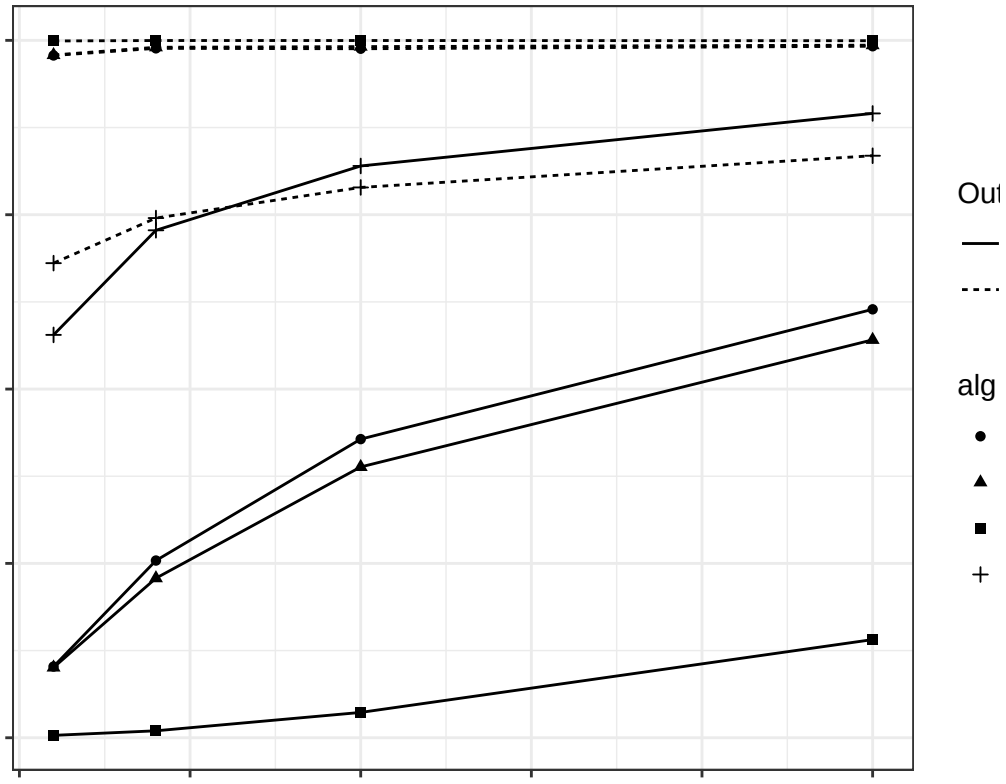



*Figure 1*. Mean performance by sample size

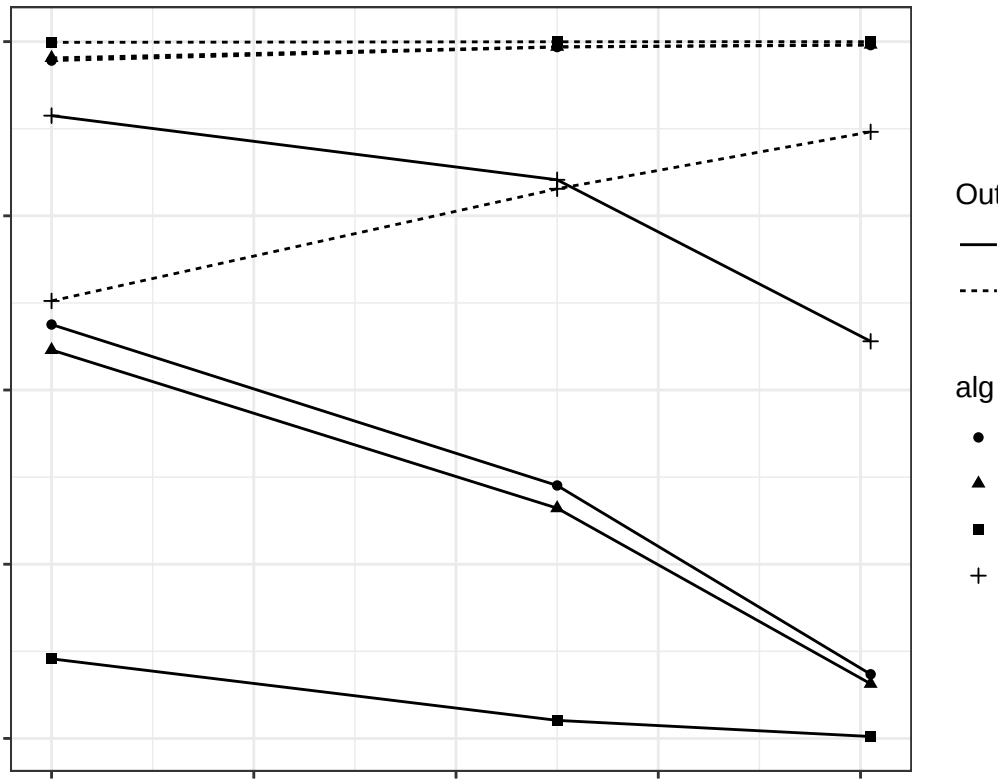



*Figure 2*. Mean performance by reliability