Aggregating evidence from conceptual replication studies using the product Bayes factor

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

[1] Tilburg University, dept. Methodology & Statistics

[2] Utrecht University, dept. Methodology & Statistics

Author Note

This is a preprint paper, generated from Git Commit # 878da062. 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; Rebecca Kuiper: Conceptualization, Writing – review & editing.

Correspondence concerning this article should be addressed to Caspar J. Van Lissa, Professor Cobbenhagenlaan 125, 5037 DB Tilburg, The Netherlands. E-mail: c.j.vanlissa@tilburguniversity.edu

Abstract

The product Bayes factor (PBF) can synthesize evidence for an informative hypothesis across heterogeneous replication studies. It is particularly useful when the number of studies is relatively low and conventional assumptions about between-studies heterogeneity are likely violated. The present paper introduces a user-friendly implementation of the PBF in the `bain` R-package. The method was validated in a simulation study that manipulated sample size, number of replication samples, and reliability. Several tutorial examples demonstrate the use of the method in distinct use cases. Results of the simulation study show that PBF had a higher overall accuracy when benchmarked against other evidence synthesis methods, including random-effects meta-analysis (RMA). This was primarily due to PBF's greater sensitivity in detecting a true effect. However, PBF had relatively lower specificity. The PBF showed increasing sensitivity and specificity with increasing sample size. With an increasing number of samples, lower sensitivity was traded for greater specificity. Although PBF's overall performance was less susceptible to reliability than the other algorithms, this masked a trade-off between reliability and specificity. PBF thus appears to be a promising method for meta-analysis of heterogeneous conceptual replication studies. Nonetheless, users should be aware of its lower specificity, and the fact that the Bayesian approach to inference addresses a qualitatively different research question than other evidence synthesis methods.

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

Word count: 5039

Aggregating evidence from conceptual 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 (Brembs, 2018). Replication research has emerged as one potential way to address this crisis and derive knowledge that will stand the test of time (see Lavelle, 2021). In step with this interest in replication research, research synthesis methods have become increasingly popular. These methods aggregate research findings, and thus enable drawing overarching conclusions across multiple (replication) studies. This paper addresses Bayesian evidence synthesis, a research synthesis method that aggregates evidence for an informative hypothesis, quantified by the Bayes factor, across multiple studies. This method has the potential to provide a more comprehensive and accurate picture of the state of the literature, and to identify areas of consensus and disagreement among studies. We describe the method in detail, benchmark its performance against other commonly used research synthesis methods, and demonstrate its application through a tutorial example analysis. To facilitate uptake of the method by applied researchers, we introduce an implementation of the method in the `bain` R-package. This implementation enables the use of Bayesian evidence synthesis for many commonly used statistical analyses in R.

A key challenge in quantitative research synthesis is dealing with between-studies heterogeneity (Higgins, Thompson, & Spiegelhalter, 2009). When studies examine the same research question in different laboratories, use idiosyncratic methods, and sample from distinct populations, these between-study differences can introduce heterogeneity in findings. The most common quantitative research synthesis method is meta-analysis, in which results of different studies are aggregated to estimate an aggregate effect size (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 may assume that no heterogeneity in the outcome exists and a fixed-effect meta-analysis can be conducted to estimate the common population effect. Second, when heterogeneity 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. Finally, when there are many potential variables that cause systematic 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 (Van Lissa, Van Erp, & Clapper, 2023). However, accounting for moderators requires a relatively high number of observations per moderator, which may not be available.

Each of the aforementioned approaches makes different assumptions about the nature of heterogeneity (see “Models for meta-analysis” in Van Lissa, 2020). A crucial shortcoming of existing research synthesis methods is that these assumptions may not be tenable when meta-analyzing studies that investigate the same informative hypothesis, but are otherwise very heterogeneous. The situation may arise where each study is uniquely identified by a combination of linearly dependent moderators. In this case, it is no longer possible to synthesize *effect sizes* while accounting for heterogeneity using statistical methods. It is still possible, however, to quantify the support these studies provide for the underlying informative hypothesis. To this end, Bayesian evidence synthesis (BES) aggregates the evidence for a theoretical relationship across studies, without imposing assumptions about heterogeneity (Kuiper, Buskens, Raub, & Hoijtink, 2013). Although this assumption is not necessary, for the sake of simplicity we assume that this theoretical relationship is evaluated via informative hypothesis H_i in all studies.

The amount of evidence for a hypothesis can be expressed as a Bayes factor, or BF. The BF can be interpreted as the ratio of evidence for one hypothesis relative to another hypothesis. Within the scope of this paper, all Bayes factors are the ratio of evidence for an informative hypothesis H_i relative to its complement $H_{!i}$ (see Gu, Mulder, & Hoijtink, 2018). The subscript $_{!}$ represents the negation operator; in other words, $H_{!i}$ means “not

H_i". This Bayes factor, which we will refer to as BF_c, represents the ratio of evidence for H_i divided by evidence against it. A value of $BF_c = 10$ means that the data provide ten times more support for the hypothesis than against it.

When multiple studies each provide evidence for H_i in the form of complement Bayes factors, these Bayes factors can be synthesized across studies by taking their product (Kuiper et al., 2013). The resulting product Bayes factor (PBF) summarizes the total evidence for the hypothesis. The only assumption of the PBF is that all study-specific hypotheses provide evidence about the same underlying theoretical relationship. Note that other approaches to BES exist; for instance, it is possible to use the posterior of one study as the prior for a replication study, and thus accumulate evidence across studies (see Heck et al., 2022). Such applications are out of scope of the present paper, which addresses the PBF approach to BES.

Although meta-analysis and BES are both research synthesis methods, they answer different research questions. Meta-analysis estimates the point estimate or distribution of a population effect size. It pools estimates of this effect size across multiple studies to obtain an overall estimate of the effect size. It thus answers questions like: Given certain assumptions about between-studies heterogeneity, what is the average population effect size? BES, on the other hand, aggregates evidence for an informative hypothesis across multiple studies. It thus answers the question: Do all these studies support the hypothesis of interest? Both methods are appropriate for different research questions, and provide complementary information.

This paper introduces the first implementation of BES in user-friendly free open source software. A function `pbf()` was contributed to the `bain` R-package for Bayesian informative hypothesis evaluation, version `0.2.9`. This paper presents a simulation study to validate the method and benchmark it against alternative evidence synthesis methods. It additionally illustrates several use cases through reproducible examples.

Simulation study

The present simulation study set out to validate the PBF algorithm and benchmark it against other evidence synthesis methods. We simulated a scenario where an informative hypothesis about a correlation between two variables was measured across several independent samples, and the resulting evidence was synthesized across samples using multiple methods. The informative hypothesis, set to be equal across studies, was $H_i : \rho > .1$. To examine the performance of the different evidence synthesis methods in a range of scenarios, several design factors were manipulated. First was the presence or absence of a true population effect. Given the informative hypothesis of $H_i : \rho > .1$, the presence of a true population effect was defined as $\rho = .2$ and a null effect was defined as $\rho = .1$. The second design factor was the number of observations per sample $n \in (50, 200, 500, 800)$. These values were chosen because they correspond to a statistical power to reject a false null hypothesis of $\beta \in (.10, .30, .60, .80)$ power, respectively, assuming $\alpha = .05$ and a known effect size of $\rho = .1$ (Cohen, 1988). Third, we manipulated the number of independent samples (or: replication studies), $k \in (2, 3, 10)$. Fourth, the reliability of the two correlated variables was varied between $\alpha \in (0.6, 0.8, 1.0)$ to range from questionable to perfect reliability (Nunnally & Bernstein, 2017). Questionable reliability is the lowest level considered to be acceptable in social scientific research, and perfect reliability is what is assumed when analyzing correlations between observed items or scale scores. For all unique combinations of these design factors, the simulation was repeated 1000 times.

Algorithms

The main algorithm of interest was the PBF. As a decision criterion to conclude that H_i was supported over its complement, we used $PBF > 3$ - a conventional threshold for inference using Bayes factors (Jeffreys, 1998).

As a benchmark for comparison, we included 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 benchmark was *vote counting*: counting the number of significant effects. Although this method is still in use for aggregating conceptual replications, it is considered bad practice. Three disadvantages are that vote counting disregards sample size, reduces statistical power, and does not quantify the strength of the evidence (Hedges & Olkin, 1980). Our vote counting algorithm summed the number of one-sided z-tests of a null hypothesis corresponding to the informative hypothesis, so $H_0 : \rho = .1$ and $H_a : \rho > .1$, which corresponds to H_i. The decision criterion was that the hypothesis was supported in the majority of samples. Thus, for example, if H_0 was rejected in three out of five samples, our vote counting algorithm would find overall support for H_a (and, by extension, H_i).

The second benchmarking algorithm was *random-effects meta-analysis* (RMA), which is the current gold standard for evidence synthesis (Viechtbauer, 2010). For this algorithm, the null-hypothesis was rejected if a 90% confidence interval excluded the hypothesized value under H_i. Note that a 90% confidence interval corresponds to a test at $\alpha = .05$, because all effects in the simulation are directional.

The third benchmarking algorithm was *individual participant data* (IPD) meta-analysis (Riley, Lambert, & Abo-Zaid, 2010). Like classic meta-analysis (RMA), IPD is a multilevel model, clustered by sample. IPD uses the raw data, which makes it possible to estimate variance at the first level. By contrast, RMA treats the first-level variance as known. Note that the PBF can be estimated using either sufficient statistics (as in meta-analysis) or using raw data (as in IPD). With this in mind, it is informative to benchmark it against both of these methods. Just as for RMA, a 90% confidence interval was used for inference.

Performance indicators

For each algorithm, inferential decisions made using the criteria described above were compared to the population status of the hypothesis (true or false). The resulting confusion 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, $\frac{TP}{TP+FN}$, the ability to detect an effect given that it was indeed true in the population, and specificity, $\frac{TN}{TN+FP}$, the ability to correctly conclude that the informative hypothesis is not supported, given that it was indeed false in the population. The overall performance was captured by the accuracy, which represents the total proportion of correct (true positive and true negative) decisions, $\frac{TP+TN}{TP+TN+FP+FN}$.

Results

We examined overall model performance across conditions. PBF had a higher overall accuracy than other algorithms followed by IPD, then RMA, and finally VC, see Table 1. This higher accuracy was primarily driven by PBF's greater sensitivity to detect a true effect compared to other algorithms. However, PBF had a lower specificity compared to all other algorithms. This suggests that the PBF trades a loss of specificity for increased sensitivity.

Effect of simulation conditions

We used ANOVAs to examine the effect of simulation conditions on overall accuracy. The differences between algorithms were analyzed in analyses that included two-way interactions between design factors and algorithm. As the sample size was very large, significance tests were uninformative. We thus focused on interpreting the effect sizes of the design factors. The performance of PBF was most impacted by sample size n, followed by the number of groups k, and reliability. The differences in the effects of sample size and number of groups were relatively small between PBF and the two best practice algorithms, RMA and IPD - but substantial between PBF and the suboptimal VC algorithm. The reverse pattern occurred for reliability, however: it showed a substantial difference in effect between PBF and the two best practice algorithms only.

Effect of sample size. Figure 1 indicates that for PBF, both sensitivity and specificity increased with sample size. The other algorithms showed only increasing sensitivity; specificity was limited by a ceiling effect. This difference explains the effect of sample size on the difference between algorithms (see Table 2).

Effect of number of samples. Figure 2 indicates that PBF showed increasing specificity at higher levels of k, while sensitivity was relatively unaffected. RMA and IPD showed a similar pattern, although their specificity was at a ceiling. Only VC showed decreasing sensitivity with an increasing number of groups; this is because the probability of obtaining false negatives increases with the number of groups (as also noted by Hedges & Olkin, 1980). This difference in pattern of effects explains why number of samples had a moderate effect on the difference between algorithms (see Table 2).

Effect of reliability. Figure 3 indicates that the PBF traded sensitivity for specificity. At low levels of reliability, specificity exceeded sensitivity; at high levels of reliability, this pattern was reversed. The other algorithms did not show this pattern, as their specificity was at a ceiling. Their sensitivity increased with higher reliability, however, and therefore, so did their overall performance. This difference in pattern of effects explains why reliability had a moderate effect on the difference between algorithms (see Table 2). Note that, whereas the overall accuracy of the PBF was found to be less susceptible to reliability as compared to other algorithms, this finding masked the trade-off between reliability and specificity.

Discussion

This simulation study examined the performance of the product Bayes factor (PBF) as compared against three other common methods for evidence synthesis (IPD, RMA, and VC). The results showed that PBF had a higher overall accuracy than the other algorithms, primarily due to its greater sensitivity to detect a true effect. PBF had lower specificity, however, suggesting that it trades specificity for increased sensitivity. PBF's performance was most impacted by sample size, followed by the number of groups and reliability. Even at the smallest sample size of $n = 20$, PBF had a superior sensitivity to the other algorithms. This suggests that PBF is more suitable than other methods as a small sample solution. When the number of samples increased, most algorithms showed approximately stable specificity and increasing sensitivity. A notable exception was vote counting; its sensitivity decreased with an increasing number of samples, as has been previously documented (Hedges & Olkin, 1980). With increasing reliability, PBF showed a substantially different pattern of results than the two other best practice algorithms (RMA and IPD). Although the overall effect of reliability on overall accuracy was smaller for PBF than for other algorithms, this masks a trade-off between reliability and specificity. Whereas most algorithms showed near-stable specificity and increasing sensitivity, PBF showed a clear trade-off between decreasing specificity and increasing sensitivity. This implies that the PBF is more conservative - less likely to detect an effect - in the presence of increasing measurement error.

These results have important implications for applied evidence synthesis. For example, the finding that PBF had a higher overall accuracy due to greater sensitivity suggests that it may be a better choice than the other algorithms, particularly when detecting true effects has high priority. However, researchers should be aware that this increased sensitivity comes at a loss of specificity, which incurs a greater risk of false positive results. If specificity is a higher priority, other algorithms such as IPD or RMA may thus be more appropriate.

The present study also has some limitations. First, the simulation study makes specific assumptions that may not generalize to all real-world applications. A second important caveat is that most of the algorithms did not reach a level of sensitivity that is considered acceptable from a perspective of statistical power (i.e., greater than .80, J. Cohen, 2013). PBF performed notably better than other algorithms, but its sensitivity still fell below .80 in many conditions. Low power increases the risk of false negatives, or failing to detect a true effect. One reason power was low is that, in conditions where a true effect was present, its value only exceeded the boundary value of the informative hypothesis by .1 points. Such small effects are hard to detect. All algorithms will likely perform better when the true effect is larger. The low sensitivity of all algorithms highlights the importance of reticence when interpreting evidence syntheses of studies with small samples and small effect sizes. It may be prudent to avoid generalizing such results to the population, and instead consider them as merely descriptive of the published research. Additionally, sensitivity analyses can be used to assess the robustness of the results to different modeling assumptions and methods.

A third limitation is that the evidence synthesis methods compared here represent different approaches to inference and answer different research questions. Since each of these methods is optimized for a different purpose, the present study should not be considered as a comprehensive assessment of their strengths and weaknesses. We nonetheless compare them because of their similar usage in evidence synthesis. It is up to individual researchers to choose an appropriate method, guided by the research question and the available information. For instance, when raw data is unavailable, IPD cannot be used, and when parameter estimates or effect sizes are not reported, only VC can be used.

Aside from the aforementioned fact that the PBF answers a different research question than the other algorithms, it is worth noting limitations of the interpretation of the PBF. The PBF renders support for one specific informative hypothesis versus its complement. If the informative hypothesis is supported, this does not necessarily mean that it is also true. Consider the hypothetical example that the informative hypothesis that the earth is flat was supported with $BF = 3.01$. Although the data support this hypothesis over its complement, the hypothesis is clearly wrong (the earth is spherical). If we would have evaluated another hypothesis, e.g., the earth is shaped like an American football, it would have received much more support, e.g. $BF = 1000$, even though it is also wrong. A high Bayes factor thus does not mean that the hypothesis is true. Conversely, a low Bayes factor merely indicates that the informative hypothesis is not supported, and does not provide information about the true state of affairs. A related limitation is that our simulation study used an arbitrary - albeit conventional - threshold for inference (Jeffreys, 1998). In applied research, it is more sensible to evaluate the weight of evidence, rather than resorting to a rule of thumb.

Tutorial

This tutorial demonstrates how to synthesize evidence for an informative hypothesis across heterogeneous replications using the Product Bayes Factor (PBF). We assume that users have installed the free open source statistical programming language R (R Core Team, 2021). The R-package `bain` version `0.2.9` or later is required, which can be installed by running `install.packages("bain")` in the R console. The data used in this tutorial are included in the `bain` package, and have been simulated based on the data presented in (Leeuwen, Van Lissa, Papakonstantinou, Petersen, & Curry, 2022). A more detailed description of the datasets is found in (Leeuwen et al., 2022); additionally, the dataset documentation is accessed by running `?synthetic_us`, `?synthetic_dk` or `?synthetic_nl` in the R console. Van Leeuwen and colleagues conducted a theory-driven, preregistered study to address the research question whether political orientation and moral dispositions are associated. Suitable data were collected in three countries: the United states of

America, Denmark, and the Netherlands. Each sample contained multiple measures of political orientation and moral dispositions. In the original publication, the PBF was used to aggregate evidence across scales and countries to obtain an overall measure of support for the central hypothesis. This tutorial follows the same rationale, but uses only one effect size per sample, and varies the way this effect size is computed to illustrate the more typical use case where the same informative hypothesis has been studied in different ways in multiple studies. We will examine the informative hypothesis that self-reported importance of family morality is positively associated with a conservative socio-political orientation. We load the `bain` library and assign the data to three objects with convenient names:

```
library(bain)
NL <- synthetic_nl
DK <- synthetic_dk
US <- synthetic_us
```

How to use bain. We briefly introduce the basic use of the `bain()` function, and how to interpret its output. We must estimate a model suitable for evaluating our informative hypothesis. Because both scales consist of multiple items, we can use structural equation modeling (SEM) to perform latent variable regression (see Van Lissa et al., 2020):

```
# Load lavaan package for SEM
library(lavaan)

# Specify SEM-model for latent variable regression
model_nl <- "
fam =~ fam_1 + fam_2 + fam_3
con =~ sepa_soc_1 + sepa_soc_2 + sepa_soc_3 + sepa_soc_4 + sepa_soc_5 +
       sepa_eco_1 + sepa_eco_2 + sepa_eco_3 + sepa_eco_4 + sepa_eco_5
```

```
con ~ beta * fam"

# Estimate the model in lavaan
results_nl <- sem(model = model_nl, data = NL)
```

The informative hypothesis in this tutorial is $H_i : \beta > .1$, where β (beta) is the standardized regression coefficient. Instead of a conventional null hypothesis, $H_0 : \beta = 0$, the value of .1 was used as a minimal effect size of interest. The code below illustrates how to obtain a Bayes factor for this informative hypothesis, using the output of the SEM analysis above. We can refer to the parameter `beta` by name because we labeled it in the `lavaan` syntax; if we had not done so, we could find the names of all model parameters by running `get_estimates(results_nl, standardize = TRUE)`. The results indicate that the hypothesis is supported when compared to its complement. For a more in-depth tutorial on `bain()`, see Hoijtink, Mulder, Lissa, and Gu (2019), and for further guidance on the use of `bain()` for SEM, see Van Lissa et al. (2020).

```
# Test that the effect labeled 'beta' is positive
bf_nl <- bain(results_nl, hypothesis = "beta > .1", standardize = TRUE)
bf_nl
```

Aggregating evidence across studies. As mentioned before, suitable data were collected to evaluate the substantive hypothesis in three countries. There are differences between countries that prevent analyzing these data as a multilevel model, however. For instance, conservatism was measured using different scales. This is an appropriate situation to use the PBF to aggregate evidence across countries. Below, we estimate a latent regression model for the remaining two countries, taking care to use the same label for the parameter of interest in all samples. Then, we bind all three SEM-models in a list, and call PBF to evaluate the hypothesis of interest on all models and aggregate the evidence. As the BF in all three samples is positive, the resulting PBF is very large. We can thus conclude that the central hypothesis receives overwhelming support across samples.

```
# Specify the models for DK and US
model_dk <- "
fam =~ fam_1 + fam_2 + fam_3
con =~
sepa_soc_1 + sepa_soc_2 + sepa_soc_3 + sepa_soc_4 + sepa_soc_5 +
sepa_eco_1 + sepa_eco_2 + sepa_eco_3 + sepa_eco_4 + sepa_eco_5
con ~ beta * fam"
model_us <- "
fam =~ fam_1 + fam_2 + fam_3
con =~
secs_soc_1 + secs_soc_2 + secs_soc_3 + secs_soc_4 + secs_soc_5 +
secs_soc_6 + secs_soc_7 +
secs_eco_1 + secs_eco_2 + secs_eco_3 + secs_eco_4 + secs_eco_5
con ~ beta * fam"

# Estimate the model in lavaan
results_dk <- sem(model = model_dk, data = DK)
results_us <- sem(model = model_us, data = US)

# Bind the models into a list
results <- list(results_nl, results_dk, results_us)
# Test the hypothesis that the effect size labeled 'beta' is positive
pbf(results, hypothesis = "beta > .1", standardize = TRUE)
```

```
##                       PBF Sample.1     Sample.2     Sample.3
## H1: beta>.1 1.013928e+27 23.24645 1.903063e+12 2.291908e+13
```

Using bain objects. The `pbf()` function also accepts multiple `bain` objects. This makes it possible to, for example, evaluate different sets of hypotheses on different data sets before using the resulting `bain` objects to aggregate the evidence for all common hypotheses across datasets. The example below illustrates this use case. As before, all analyses share one hypotheses in common ($H_i : \beta_{fam} > .1$), but the Dutch sample now contains a sample-specific hypothesis regarding the effect of group morality, namely that $\beta_{grp} < .1$. The `pbf()` function is called on a list of `bain` objects. Note that, in this case, `pbf()` does not require an argument `hypothesis`, as the hypotheses are contained in the `bain` objects.

```
# Add the additional predictor to the model, label the effect beta2
model_nl <- c(model_nl, "group =~ grp_1 + grp_2 + grp_3
                         con ~ beta2 * group")

# Estimate the model in lavaan
results_nl <- sem(model = model_nl, data = NL)

# Obtain BF for each sample; note that the Dutch sample has two hypotheses
bf_nl <- bain(results_nl, hypothesis = "beta > .1;
                                        beta2 < .1",
              standardize = TRUE)
bf_dk <- bain(results_dk, "beta > .1")
bf_us <- bain(results_us, "beta > .1")

# Bind bain objects into a list
bfs <- list(bf_nl, bf_dk, bf_us)
```

```
# Call pbf on that list
pbf(bfs)
```

As can be seen, the results are equivalent to the results in the previous example. The sample-specific hypothesis has been left out, and common hypotheses are retained and aggregated. If there are no common hypotheses across all objects, `pbf()` throws an error.

Using sufficient statistics. A third use case occurs when the raw data from different samples are not available. This may happen, for example, when aggregating findings from the published literature (similar to meta-analysis). In this case, one can use the default interface of `bain`, as explained in (Hoijtink et al., 2019). This function requires four arguments: A named vector of parameter estimates, their asymptotic covariance matrix, the original sample size, and the number of within-group and between-group parameters. Note that, when analyzing a single parameter per sample, the standard error is sufficient to construct the asymptotic covariance matrix. Thus, this method can be applied to data that have been prepared for classic meta-analysis (effect sizes and their sampling variances). Importantly, unlike meta-analysis, the present method is suitable for conceptual replications. It does not require uniform effect size measures across studies. The example below illustrates how to aggregate evidence for one hypotheses across three studies that each used different methods.

The present use case evaluates the following hypothesis: *There is a positive association between family morality and political conservatism.* This conceptual hypothesis is evaluated differently in the three samples, resulting in three different types of statistics and distinct sample-specific hypotheses:

1. A t-test was performed using the NL data; using Cohen's D gives $H_i^{NL} : \delta_{conservative>liberal} > 0$, where δ is the mean difference between groups.
2. A bivariate regression coefficient was calculated using the DK data, giving $H_i^{DK} : \beta_{fam} > 0$

3. A correlation coefficient was calculated using the US data, giving $H_i^{US} : \rho_{fam,con} > 0$, where ρ is the correlation between family morality and conservatism.

Note that we intentionally manipulate the data to illustrate these different analyses; for example, we compute mean scale scores and dichotomize the continuous conservatism scale to conduct a t-test. We do not advocate these practices for applied research.

First we obtain the relevant parameter estimates and their sampling variances, which allows us to evaluate the specific hypotheses in `bain`:

```
# Create mean scale scores
NL <- data.frame(
  family = rowMeans(NL[c("fam_1", "fam_2", "fam_3")]),
  conservative = rowMeans(NL[c("sepa_soc_1", "sepa_soc_2", "sepa_soc_3",
                               "sepa_soc_4", "sepa_soc_5", "sepa_eco_1",
                               "sepa_eco_2", "sepa_eco_3", "sepa_eco_4",
                               "sepa_eco_5")]))
DK <- data.frame(
  family = rowMeans(DK[c("fam_1", "fam_2", "fam_3")]),
  conservative = rowMeans(DK[c("sepa_soc_1", "sepa_soc_2", "sepa_soc_3",
                               "sepa_soc_4", "sepa_soc_5", "sepa_eco_1",
                               "sepa_eco_2", "sepa_eco_3", "sepa_eco_4",
                               "sepa_eco_5")]))

US <- data.frame(
  family = rowMeans(US[c("fam_1", "fam_2", "fam_3")]),
  conservative = rowMeans(US[c("secs_soc_1", "secs_soc_2", "secs_soc_3",
                               "secs_soc_4", "secs_soc_5", "secs_soc_6",
                               "secs_soc_7", "secs_eco_1", "secs_eco_2",
                               "secs_eco_3", "secs_eco_4", "secs_eco_5")]))

# NL: Conduct t-test using Cohen's D
NL$group <- cut(NL$conservative, breaks = 2,
                labels = c("liberal", "conservative"))
sample_sizes <- table(NL$group)
sds <- tapply(NL$family, NL$group, sd)
pooled_sd <- sqrt(sum((sample_sizes - 1) * sds) / (sum(sample_sizes) - 2))
NL_est <- diff(tapply(NL$family, NL$group, mean)) / pooled_sd
NL_var <- (sum(sample_sizes) / prod(sample_sizes)) +
  (NL_est^2 / (2*sum(sample_sizes)))

# DK: Conduct bivariate regression
DK_fit <- lm(conservative ~ family, data = DK)
DK_est <- coef(DK_fit)["family"]
DK_var <- vcov(DK_fit)["family", "family"]

# US: Correlation coefficient
US_est <- cor(US)[1, 2]
US_var <- (1 - US_est^2)^2 / (nrow(US) - 1)

# Name the estimates so hypotheses will be the same
names(NL_est) <- names(DK_est) <- names(US_est) <- "parameter"
```

Then, we use `bain.default()` to evaluate the central hypothesis on each parameter estimate. The `pbf()` function can be called on a list of the resulting bain objects.

```r
# Use bain.default() to obtain BF for the central hypothesis
NL_bain <- bain(x = NL_est,
                Sigma = matrix(NL_var, 1, 1),
                n = nrow(NL),
                hypothesis = "parameter > 0",
                joint_parameters = 1)
DK_bain <- bain(x = DK_est,
                Sigma = matrix(DK_var, 1, 1),
                n = nrow(DK),
                hypothesis = "parameter > 0",
                joint_parameters = 1)
US_bain <- bain(x = US_est,
                Sigma = matrix(US_var, 1, 1),
                n = nrow(US),
                hypothesis = "parameter > 0",
                joint_parameters = 1)

# Aggregate evidence using pbf()
pbf(list(US_bain, DK_bain, NL_bain))
```

The results suggest substantial evidence for the hypothesis that there is a positive association between family morality and political conservatism. Although each study used a different method to assess this hypothesis, their evidence can be synthesized using `pbf()`.

Conclusion

In conclusion, this study evaluated the performance of the product Bayes factor as a method for evidence synthesis, and compared it against other commonly used evidence synthesis methods under different simulation conditions. Compared to the other methods, PBF had the highest overall accuracy. This was primarily due to its greater sensitivity. However, PBF had lower specificity than all other algorithms, suggesting a trade-off between sensitivity and specificity. The other algorithms showed ceiling effects in specificity, limiting their sensitivity. The performance of the PBF was most strongly affected by sample size, followed by the number of samples and reliability. We introduced a user-friendly implementation of the PBF in the `bain` R-package, and demonstrated its use with various analysis techniques in R, as well as with sufficient statistics that are already routinely coded for meta-analysis (i.e., effect sizes and their sampling variance). This means that researchers can now use the PBF to aggregate evidence in situations where classic meta-analytic methods are less suitable. For example, when one informative hypothesis has been evaluated in several replication studies, but these replication studies are quite heterogeneous because they sample from different populations and use different methods or analysis techniques. Especially when the number of replication studies is too small to adequately account for these sources of between-study heterogeneity, the PBF may be a useful method to aggregate evidence for the common informative hypothesis. Researchers should be aware that the PBF trades off increased sensitivity for decreased specificity, and that it addresses a different research question than other research synthesis methods. This highlights the importance of careful interpretation of the results, and consideration of the research question when selecting an aggregation method. In sum, our results suggest that PBF is a useful evidence synthesis method, which is now broadly accessible due to its inclusion in the `bain` R-package.

Highlights

- Many research synthesis methods make strong assumptions about between-studies heterogeneity that are violated when studies are conceptually replicated.
- The product Bayes factor (PBF) aggregates evidence for an informative hypothesis across conceptual replication studies without imposing assumptions about heterogeneity.
- This paper introduces a user-friendly way to compute the PBF for a variety of widely used models via the `pbf()` function in the `bain` R-package.
- A simulation study shows favorable performance for PBF relative to random effects meta-analysis, individual participant data meta-analysis, and vote counting.
- Three tutorial examples illustrate distinct use cases of the method.

Data Availability Statement

All analysis code is available in a version-controlled repository at https://github.com/cjvanlissa/bayesynth.

Conflict of Interest Statement

The authors declare no conflict of interest.

References

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

Brembs, B. (2018). Prestigious science journals struggle to reach even average reliability. *Frontiers in Human Neuroscience*, *12*, 37.

Cohen. (1988). *Statistical Power Analysis for the Behavioral Sciences* (Second). NJ: Lawrence Erlbaum Associates.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge.

Gu, X., Mulder, J., & Hoijtink, H. (2018). Approximated adjusted fractional Bayes factors: A general method for testing informative hypotheses. *British Journal of Mathematical and Statistical Psychology*, *71*(2), 229–261. https://doi.org/10.1111/bmsp.12110

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

Hedges, L. V., & Olkin, I. (1980). Vote-counting methods in research synthesis. *Psychological Bulletin*, *88*, 359–369. https://doi.org/10.1037/0033-2909.88.2.359

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

Hoijtink, H., Mulder, J., Lissa, C. van, & Gu, X. (2019). A tutorial on testing hypotheses using the bayes factor. *Psychological Methods*, *24*(5), 539.

Jeffreys, H. (1998). *The theory of probability* (Third). OUP Oxford.

Kuiper, R. M., Buskens, V., Raub, W., & Hoijtink, H. (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

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

Leeuwen, F. van, Van Lissa, C. J., Papakonstantinou, T., Petersen, M. B., & Curry, O. S. (2022). *Morality as cooperation, politics as conflict.*

Nunnally, J. C., & Bernstein, I. H. (2017). *Psychometric theory* (Third). New York: McGraw-Hill.

R Core Team. (2021). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Riley, R. D., Lambert, P. C., & Abo-Zaid, G. (2010). Meta-analysis of individual participant data: Rationale, conduct, and reporting. *BMJ: British Medical Journal*, *340*(7745), 521–525. Retrieved from https://www.jstor.org/stable/25674217

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

Van Lissa, C. J., Gu, X., Mulder, J., Rosseel, Y., Zundert, C. V., & Hoijtink, H. (2020). Teacher's Corner: Evaluating Informative Hypotheses Using the Bayes Factor in Structural Equation Models. *Structural Equation Modeling: A Multidisciplinary Journal*, *0*(0), 1–10. https://doi.org/10.1080/10705511.2020.1745644

Van Lissa, C. J., Van Erp, S., & Clapper, E. B. (2023). Selecting relevant moderators with Bayesian regularized meta-regression. *Research Synthesis Methods*, *14*(2), 301–322. https://doi.org/10.1002/jrsm.1628

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48.

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

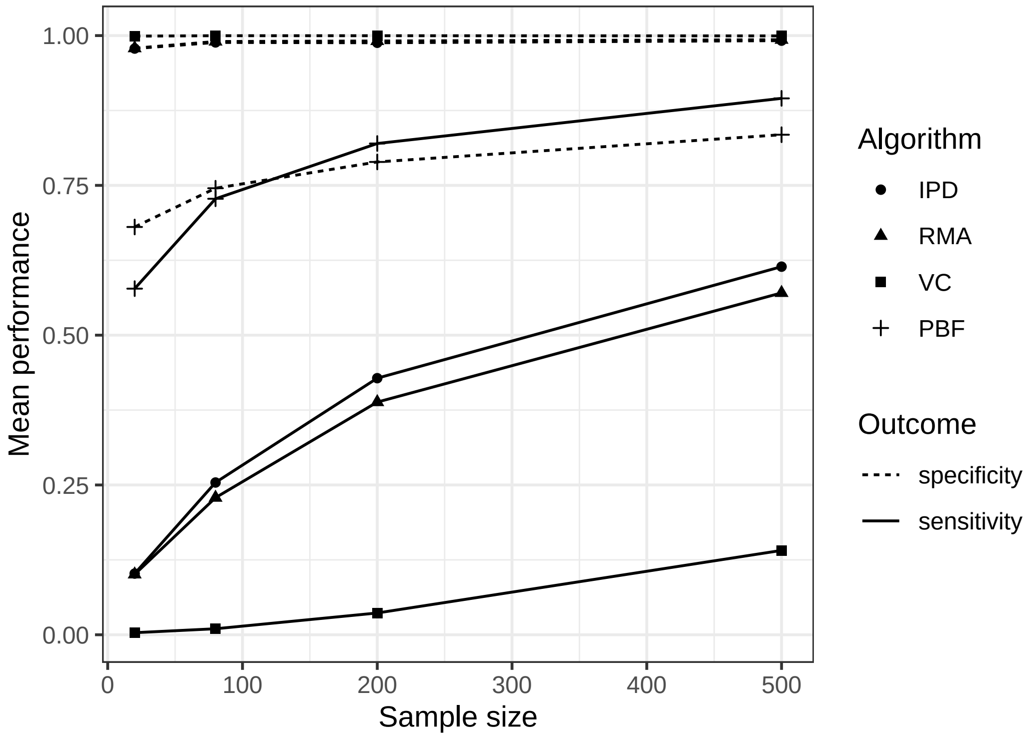

Figure 1. Mean performance by sample size

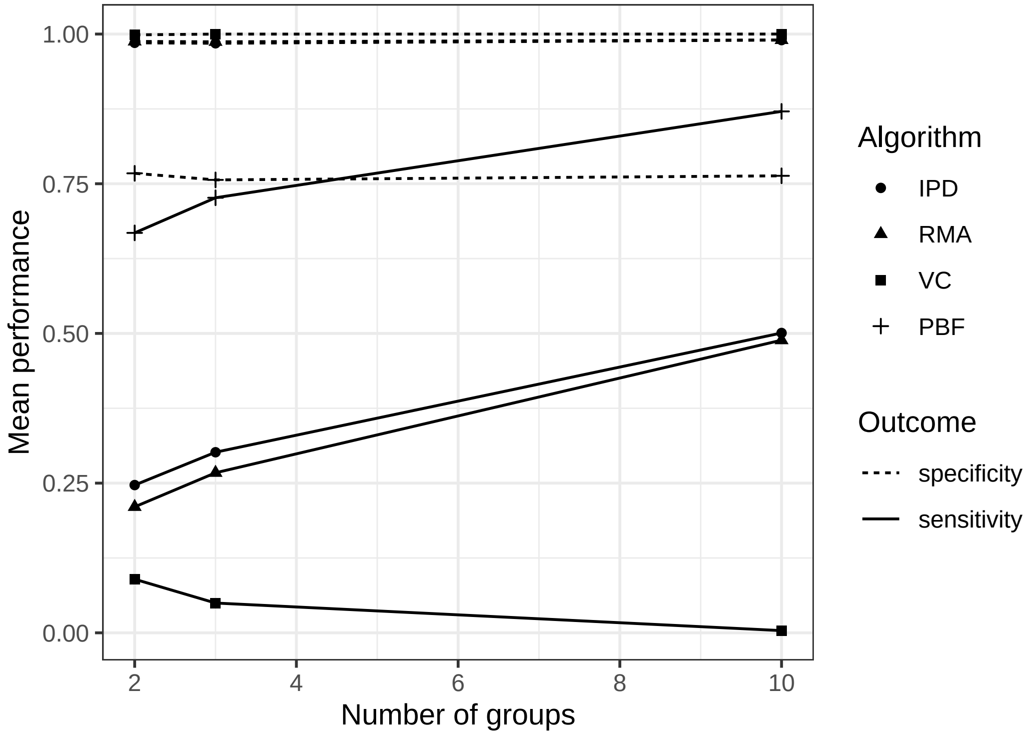

Figure 2. Mean performance by number of groups

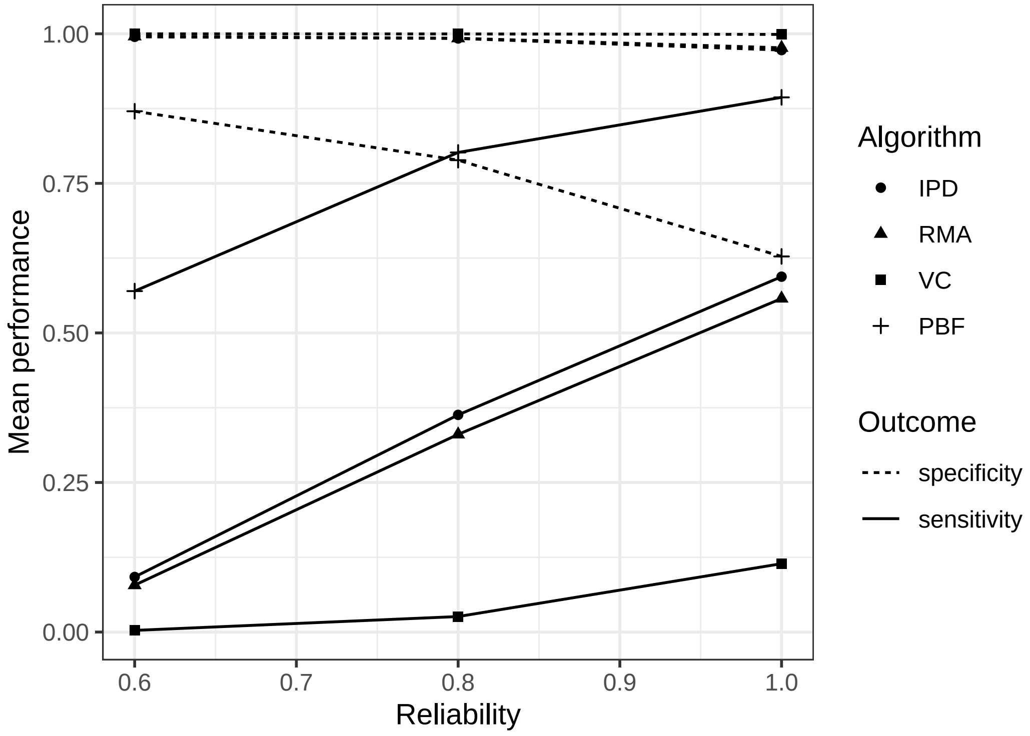

Figure 3. Mean performance by reliability