Received: 9 April 2024 | Revised: 20 June 2024 | Accepted: 20 August 2024

DOI: 10.1002/jrsm.1765

TUTORIAL

Research Synthesis Methods WILEY

A tutorial on aggregating evidence from conceptual replication studies using the product Bayes factor

Caspar J. Van Lissa[1] | **Eli-Boaz Clapper**[2] | **Rebecca Kuiper**[2]

[1]Department of Methodology & Statistics, Tilburg University, Tilburg, The Netherlands

[2]Department of Methodology & Statistics, Utrecht University, Utrecht, The Netherlands

Correspondence
Caspar J. Van Lissa, Cobbenhagenlaan 125, 5037 DB Tilburg, The Netherlands.
Email: c.j.vanlissa@tilburguniversity.edu

Funding information
Nederlandse Organisatie voor Wetenschappelijk Onderzoek, Grant/Award Number: VI.Veni.191G.090

Abstract

The product Bayes factor (PBF) synthesizes evidence for an informative hypothesis across heterogeneous replication studies. It can be used when fixed- or random effects meta-analysis fall short. For example, when effect sizes are incomparable and cannot be pooled, or when studies diverge significantly in the populations, study designs, and measures used. PBF shines as a solution for small sample meta-analyses, where the number of between-study differences is often large relative to the number of studies, precluding the use of meta-regression to account for these differences. Users should be mindful of the fact that the PBF answers a qualitatively different research question than other evidence synthesis methods. For example, whereas fixed-effect meta-analysis estimates the size of a population effect, the PBF quantifies to what extent an informative hypothesis is supported in all included studies. This tutorial paper showcases the user-friendly PBF functionality within the bain R-package. This new implementation of an existing method was validated using a simulation study, available in an Online Supplement. Results showed that PBF had a high overall accuracy, due to greater sensitivity and lower specificity, compared to random-effects meta-analysis, individual participant data meta-analysis, and vote counting. Tutorials demonstrate applications of the method on meta-analytic and individual participant data. The example datasets, based on published research, are included in bain so readers can reproduce the examples and apply the code to their own data. The PBF is a promising method for synthesizing evidence for informative hypotheses across conceptual replications that are not suitable for conventional meta-analysis.

KEYWORDS

Bayes factor, Bayesian, evidence synthesis, meta-analysis

This is a preprint paper, generated from Git Commit # 2b145001. This work was funded by a NWO Veni Grant (NWO Grant Number VI.Veni.191G.090), awarded to the lead author.

Highlights

What is already known

- Meta-analysis make strong assumptions about between-studies heterogeneity that are often violate.
- The PBF aggregates evidence for an informative hypothesis across conceptual replication studies without imposing assumptions about heterogeneity.

What is new

- This Tutorial introduces a user-friendly implementation of the PBF in the `bain` R-package.
- A simulation study showed favorable performance for PBF relative to random effects meta-analysis, individual participant data meta-analysis, and vote counting.
- Five tutorial examples, based on published research, illustrate distinct use cases of the method.

Potential impact

- Meta-analysts can now easily pool evidence for an informative hypothesis when assumptions about between-studies heterogeneity are violated.

1 | INTRODUCTION

Recent years have seen a crisis of confidence over the reliability of published results in psychology, and science more broadly.[1] This crisis has led to an increase in replication research as a way to derive knowledge that will stand the test of time.[2] In step with this interest in replication research, research synthesis methods have become increasingly popular. Such methods aggregate research findings, and thus enable drawing overarching conclusions across multiple replication studies. Meta-analysis is the most popular quantitative research synthesis method. It estimates an overall population effect size based on several close replication studies. One limitation is that its assumptions about the nature of between-studies heterogeneity may be untenable in some practical applications. The product Bayes factor (PBF) is a valid alternative in such cases.[3] Instead of estimating a pooled effect size across studies, the PBF computes a Bayes factor that quantifies the amount of support the studies provide for a common informative hypothesis. This tutorial introduces a novel user-friendly implementation of this existing method in the `bain` R-package. Others have recently validated the performance of the PBF using a simulation study,[4] and the present study includes Appendix S1 with an additional simulation study to benchmark the present implementation. We demonstrate the use of the PBF in a series of reproducible Tutorial examples, based on reanalysis of data from published studies that previously used the PBF.[3,5]

Differences between studies present a fundamental challenge to research synthesis methods.[6] Even when studies test the same hypothesis, they often do so in different laboratories, with idiosyncratic methods, and sample from distinct populations. These between-study differences can introduce heterogeneity in the effect sizes found. Meta-analysis is the most common quantitative research synthesis method. It aggregates results of different studies by estimating an overall effect size.[7] There are many ways to account for heterogeneity in meta-analysis. We provide four examples of common practices.[8] First, if studies are exact replications, researchers might assume that they share a single true effect and estimate its value using fixed-effect meta-analysis. Second, if heterogeneity between studies can be assumed to be normally distributed, random-effects meta-analysis can be used to estimate the mean and variance of this distribution. This assumption holds, for example, when the effect size is influenced in a minor way by many uncorrelated factors. Third, when there are a few moderators that cause larger systematic differences between studies, their influence can be accounted for using meta-regression. Note that it must be known beforehand which moderators are relevant, and there must be sufficient studies relative to the number of moderators. Fourth, when there are many moderators that could cause systematic differences, but it is not known beforehand which of them are relevant, exploratory techniques like random forest meta-analysis and penalized meta-regression can be used to identify relevant moderators.[9] Each of these approaches makes

different assumptions about the nature of heterogeneity ("Models for meta-analysis" in Ref. [8]).

A significant challenge arises in situations where none of the aforementioned assumptions are tenable. This occurs when studies all assess the same informative hypothesis, but differ in fundamental ways that preclude their effect sizes from being aggregated. Consider, for example, the situation that arises when the number of moderators is large relative to the number of studies. This issue is exemplified in the meta-analytic dataset from Kuiper et al.[3] in Table 1. Each of the four studies is unique in its combination of moderator values, making it impossible to control for their influence using meta-regression.[3] The problem of multicollinear moderators is immediately evident in small datasets like this, but also occurs in larger datasets, where it may be less conspicuous (e.g.,[14]). Another example illustrates that studies may be fundamentally incomparable even when they have sampled from the same populations, used the same designs, the same measurement instruments, and the same statistical models. At first glance, these studies might seem ideally suited for (fixed-effect) meta-analysis. The situation changes, however, if each study controlled for a different set of covariates. This would render the estimands of the effect size of interest fundamentally incomparable (Table 2).[15] Another example occurs when transforming different effect sizes to a common metric (e.g., odds ratio to correlation coefficient). While this is a common practice, researchers may be unaware that conversions often involve strong assumptions and may result in estimates with unknown distributional properties.[16] These examples illustrate the general problem that it may not always be possible to estimate a pooled effect size across studies. In such cases, the PBF can still be used to determine whether all studies support a common informative hypothesis.

1.1 | What are informative hypotheses?

Informative hypotheses relate the parameters of interest (e.g., β) to each other and/or to numerical constants by means of (in)equality constraints. The conventional null hypothesis H_0 can be thought of as a special case of an informative hypothesis, which specifies that the parameter is equal to the value zero, $H_i : \beta = 0$. But informative hypotheses can be much more complex, and ideally, should closely represent a theoretical expectation about the relations between parameter values. For example, if one is confident that a new parenting intervention will have a clinically relevant effect, one might hypothesize that a parameter is larger than a smallest effect size of interest, $H_i : \beta > 0.1$.[17] If a dose–response effect is expected for a certain drug, one might hypothesize that three treatments of increasing intensity will have effects of increasing size, and that all these treatments will exceed the smallest effect size of interest, $H_i : (\beta_1 < \beta_2 < \beta_3) > 0.1$.

1.2 | Bayesian informative hypothesis tests

The amount of evidence for an informative hypothesis H_i relative to another hypothesis can be expressed as a Bayes factor, or BF. Bayes factors for one hypothesis are thus always defined in relation to another hypothesis. In this tutorial paper, all informative hypotheses H_i are

TABLE 1 Data from Kuiper et al.[3]

Study	Beta	vi	*n*	Type	Sample	Analysis
Batenburg et al. (2003)[10]	0.09	0.00	971	Survey	Managers	Linear regression
Buskens and Raub (2002)[11]	0.14	0.00	348	Experiment	Managers	Linear regression
Buskens and Weesie (2000)[12]	1.09	0.01	1249	Experiment	Students	Probit regression
Buskens et al. (2010)[13]	1.78	0.03	2160	Experiment	Students	Multilevel model

TABLE 2 Example of incomparable studies.

Study	Sample	Design	Outcome	yi	vi
1	Europe	Experiment	Reaction time	0.03	0.00
2	Europe	Survey	Self-report	0.04	0.00
3	USA	Experiment	Self-report	0.04	0.00
4	China	Survey	Reaction time	0.02	0.00

compared to their complement $H_{!i}$.[18] This complement $H_{!i}$ can be interpreted as "not H_i." This Bayes factor, which we will refer to as BF_c, represents the ratio of evidence for H_i relative to the evidence against it. A value of $BF_c = 10 = \frac{10}{1}$ means that the data provide 10 times more support for the hypothesis than against it; a value of $BF_c = 0.1 = \frac{1}{10}$ means that the data provide 10 times less support for the hypothesis than against it.

Testing hypotheses using the Bayes factor deviates in important ways from conventional null-hypothesis significance testing. It involves thinking about probability in a more subjective way, which allows one to express confidence in the truth of a hypothesis. This deviates from frequentist thinking, where probability is defined in objective terms, as a function of the frequency with which an event would occur if a procedure would be repeated many times. As a result of these differences in interpretation, Bayes factors provide a continuous measure of evidence for the informative hypothesis, rather than a binary decision to (not) reject a typically uninformative null hypothesis. Bayes factors can also be used to evaluate relative support for competing informative hypotheses. Moreover, they are robust to multiple- and sequential testing. Importantly, the Bayes factor is not inherently better or worse than null-hypothesis significance testing. It is just an alternative inferential procedure, which—under some conditions—performs similarly to null-hypothesis significance tests.[19] While the convention of using a decision criterion of $p < 0.05$ for null-hypothesis significance testing is well-established in some fields, this is an arbitrary threshold.[20] Unlike the *p*-value, the Bayes factor is a continuous measure of evidence. Nonetheless, it is similarly possible to define decision criteria. Some have suggested cut-offs of $BF > 3$ or $BF > 10$ as conclusive evidence for a hypothesis of interest.[21] Since these decision criteria are somewhat arbitrary and conventions are not as well-established as for the *p*-value, preregistration is a useful way to demonstrate that the decision criteria were determined before the results were known.[22]

1.3 | How are Bayes factors calculated?

This tutorial uses the *Approximated adjusted fractional Bayes factor* (AAFBF) first introduced by Mulder.[19] AAFBFs are computed by integrating so-called posterior and prior distributions with respect to the parameter constraints specified in the hypothesis, and taking the ratio of these quantities. Instead of computing an analytic solution, the AAFBF uses Monte Carlo integration, sampling random values from the normal prior and posterior distributions. This introduces slight random fluctuations in the results. To make the analyses reproducible, it is therefore important to set a "random seed," which ensures that the same pseudo-random numbers will be generated upon repeated evaluation of the code. This is done throughout the tutorial examples. For a more detailed technical description of the AAFBF, readers can turn to Gu et al.[18]; for mathematical derivations, see Mulder.[19]

The term *approximated* in AAFBF refers to the fact that the prior and posterior are both approximated by normal distributions, which makes it possible to compute these Bayes factors from sufficient statistics. The (multivariate) normal posterior is centered around the maximum likelihood estimate of the parameter(s), and its covariance consists of the asymptotic (co)variance matrix of the parameters. For a single parameter, this simplifies to the parameter estimate and its sampling variance (the squared standard error).

Any Bayesian method is sensitive to the specification of the prior distribution. However, the AAFBF is a pragmatic Bayesian method: its priors are constructed to provide relatively uninformative sensible defaults.[19] In contrast to subjective Bayesian methods, where specifying an informative prior distribution for the estimated parameters is an important part of model specification, the AAFBF does not offer full control over the prior. The term *adjusted* in AAFBF refers to the fact that the center of the prior distribution is adjusted based on the set of informative hypotheses, to give each hypothesis a fair chance. The term *fractional* refers to the fact that the prior covariance matrix is based on a fraction of the information contained in the posterior covariance matrix. By default, this fraction is based on the notion of a minimal training sample.[19] It can be adjusted to change the scale of the prior distribution. Other tutorials illustrate how to perform sensitivity analyses to determine whether the results are sensitive to the prior scale.[23] Importantly, adjusting the prior scale does not affect the AAFBF when all hypotheses are composed of only inequality constraints.[24] To understand why, imagine that inequality-constrained hypotheses slice the prior distribution like a pie. Changing the scale affects the total size of the pie, but not the relative size of the slices (e.g., one slice is twice as large as the other). In contrast, when at least one hypothesis contains at least one equality constraint, adjusting the prior scale does affect the AAFBF. The equality constraint cuts a fixed-width sliver out of the pie. Changing the size of the pie while keeping this width fixed means that the sliver may be relatively larger or smaller to the remaining part.[23] Thus, when using the AAFBF, prior sensitivity checks are not necessary when the set of hypotheses contains only inequality constrained hypotheses. If the set does contain equality constrained hypotheses, performing sensitivity analyses is prudent.

1.4 | The product Bayes factor

The PBF is a Bayesian evidence synthesis (BES) technique that aggregates the evidence for a theoretical relationship across studies, without imposing assumptions about heterogeneity. As explained in Kuiper et al.,[3] the PBF aggregates evidence by taking the product of Bayes factors from individual studies. Bayes factors are defined as the ratio of evidence in favor of one hypothesis over another hypothesis. The PBF assumes that these two hypotheses are a priori equally likely. 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.[3] The resulting 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.[25] Such applications are out of scope of the present paper, which addresses the PBF approach to BES.

1.5 | Differences between meta-analysis and the PBF

Although meta-analysis and PBF 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. If all assumptions are met, this is a consistent estimate of the population effect size. Meta-analysis thus answers questions like: Given certain assumptions about between-studies heterogeneity, what is the average population effect size? The PBF, on the other hand, aggregates evidence for an informative hypothesis at the study-level. It is *not* a consistent estimator of the amount of support the hypothesis would have received based on data of all individual participants in the aggregated studies. Instead, the PBF answers the question: Do all studies support the hypothesis of interest? Both methods thus answer different research questions, and provide complementary information.

1.6 | Validity of the PBF

The validity of the PBF can be justified in several ways. One approach is to refrain from inference to a broader population, and interpret the PBF as a descriptive statistic of the amount of consistent support for an informative hypothesis provided by the available literature. Another justification is provided by the literature on triangulation, which calls for conceptual (rather than exact) replications of effects to demonstrate their robustness to method bias.[26] Any study design has inherent method bias, and exact replications replicate this bias. While exact replications meet the assumptions for estimating a population effect via meta-analysis, this estimate is affected by compounding method bias. For example, if a particular study design of the effect of X on Y uses a relatively weak manipulation of X, then it will underestimate the effect size of X on Y. This negative bias will accrue when meta-analyzing several close replications of studies with the same design, and the population effect will be underestimated. Conversely, if it is possible to combine evidence from studies with different designs, some of which underestimate, and others overestimate the effect of X on Y, then these biases will partly cancel each other out.

It should be noted that some criticisms have been raised about triangulation.[27] The first criticism is that conceptual replication is incentivized more strongly than exact replication by the pressure to publish novel findings. While this claim might be true, this would only affect the relative prevalence of both types of replications in the literature, not diminish the evidentiary value of either type. The second criticism is that failed triangulation attempts might not be published—but publication bias affects all research synthesis methods, not just triangulation. The third criticism is that researchers might construct a narrative to explain away failed replications by appealing to hidden moderators or boundary conditions. If such post-hoc explanations are presented as conclusions, then this amounts to questionable research practices.[28] However, if these explanations are presented as hypotheses for future testing, then this is standard research practice. These criticisms are thus not specific to triangulation, and all else being equal, finding consistent evidence with different methods should increase our confidence in a finding. The PBF offers an alternative to the narrative synthesis of triangulated findings; a valid numeric method to quantify evidence in favor of a shared hypothesis across heterogeneous studies.

1.7 | Are studies' sample sizes taken into account?

Conventional meta-analysis can be conceptualized as a weighted average of study effect sizes.[7] Each study's sample size indirectly affects its weight, via its influence on the sampling variance (the squared standard error). The approximate fractional Bayes factors also accounts for

sample size to some extent, as the sampling variance is part of the posterior distribution used to calculate the Bayes factors. The prior distribution is, in turn, informed by the posterior, and thus also affected by sample size.[29] The resulting Bayes factor for a true hypothesis increases monotonically with increasing sample size. Thus, the overall support for a true hypothesis will be greater if it is based on larger, rather than smaller, samples.

1.8 | Simulation study

This tutorial paper introduces a new implementation of the existing PBF method. A simulation study validating this new implementation is available in Appendix S1, and the reproducible code is available at 10.5281/zenodo.11615354. We simulated a PBF analysis of correlation coefficients and manipulated the presence or absence of a true population effect, the sample size per study, the number of replication studies, and the reliability of the correlated variables. We compared the performance of the PBF against vote counting,[30] random-effects meta-analysis,[31] and individual participant data meta-analysis.[32] PBF had the highest overall accuracy, primarily due to its greater sensitivity to detecting a true effect. However, PBF had lower specificity than all other algorithms, suggesting a trade-off between sensitivity and specificity (Table 1, p. 11 of Appendix S1). 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.

1.9 | This tutorial paper

This paper introduces the first implementation of the PBF in user-friendly free open source software. This tutorial requires the `bain` R-package for Bayesian informative hypothesis evaluation, version `0.2.11`, which contains the `pbf()` function and all tutorial data. The tutorials focus on different applications of the PBF, illustrated in reproducible examples. This tutorial was, itself, made reproducible using the Workflow for Open Reproducible Code in Science (WORCS,[33]). The code archive is available at 10.5281/zenodo.11615354.

2 | TUTORIALS

These tutorials demonstrate how to synthesize evidence for an informative hypothesis across heterogeneous replications using the PBF. We assume that users have installed the free open source statistical programming language R.[34] The R-package `bain` can be installed by running `install.packages("bain")` in the R console. The datasets used in this tutorial are all included in the `bain` package. The `kuiper2013` dataset is based on Kuiper et al.[3] and the original publications meta-analyzed therein. Its documentation is accessed by running `?kuiper2013` in the R console. The other datasets were simulated based on data presented in Ref. [5]. Their documentation is accessed by running `?synthetic_us`, `?synthetic_dk` or `?synthetic_nl`. While we introduce the basic functionality of the `bain` package, we direct the interested reader to Hoijtink et al.[29] for a general introduction to `bain`, and to Van Lissa et al.[23] for an in-depth tutorial on informative hypothesis tests for Structural Equation Models using `bain`. The `pbf()` function can be used to compute the PBF for any model object for which `bain()` methods exist (see `?bain`), and for sufficient statistics (see Tutorials 1 and 5).

2.1 | Tutorial 1: When meta-analysis falls short

The PBF was first introduced by Kuiper et al.[3] for cases where random-effects meta-analysis would otherwise be used, but is deemed inappropriate because its assumptions are likely to be violated. This Tutorial illustrates the use of the PBF in such cases, and compares it to meta-analysis to show that one can straightforwardly perform a PBF analysis with a dataset prepared for meta-analysis. As explained in the introduction, if the hypothesis of interest pertains to only one parameter, then the Bayes factor can be computed using the estimate and its standard error.[29] It is also possible to formulate more complex hypotheses that involve multiple parameters; for example, in structural equation models.[23] This requires access to the parameter covariance matrix. The present example focuses on sufficient statistics to illustrate how one might complement a conventional meta-analysis with a PBF analysis and report both in the same paper. This way, readers can determine whether they trust the assumption of normally distributed heterogeneity and interpret the random-effects estimate, or doubt the assumption and interpret the PBF instead.

The data for this example is shown in Table 1. Run `?kuiper2013` to view its documentation. Kuiper et al.[3] set out to aggregate evidence for the effect of prior interactions between partners on trust in (economic) exchange relations across four heterogeneous replication studies. All studies investigated the informative hypothesis that past (experience) with a seller has a positive effect on trust. Batenburg et al.[10] analyzed survey data using

linear regression with covariates; Buskens and Raub[11] analyzed experimental data using linear regression; Buskens and Weesie[12] used an experimental design with a binary outcome, analyzed using probit regression; and Buskens, Raub, and Van der Veer[13] used a longitudinal experimental design, analyzing the data with a three-level logistic regression. These studies each provide a regression coefficient assessing the effect of past experience on trust, and its estimated sampling variance (squared standard error). However, because the studies differ in terms of design (survey vs. experiment), operationalization of variables, measurement level of variables, and statistical model used, these regression coefficients are not directly comparable and should not be pooled using meta-analysis. The authors developed the PBF to determine whether all studies support the informative hypothesis.

Although there is one clear informative hypothesis, we follow the original study and estimate the evidence for three competing hypotheses: $H_1 : \beta = 0$ (the null hypothesis that the effect of past on trust is zero), $H_2 : \beta > 0$ (a directional hypothesis that the effect of past on trust is positive), and $H_3 : \beta < 0$ (a directional hypothesis that the effect of past on trust is negative). Note that Kuiper and colleagues computed the Bayes factors and posterior model probabilities by hand, using custom priors for the first study, and using the posterior of the first study as prior for subsequent studies. The present tutorial instead uses approximated adjusted fractional Bayes factors computed using `bain` with default settings. Consequently, the numerical results differ from Kuiper and colleagues, although the conclusions remain the same.

We first conduct a random-effects meta-analysis using the function `rma()` from the `metafor` package. Then, we perform a PBF analysis using the `pbf()` function in the `bain` package. This allows us to compare the interface of both functions and their results. First, we will load the required packages. This also makes the `kuiper2013` available. Printing it to the console should give the same result as Table 1.

```
library(metafor)
library(bain)
kuiper2013
```

To perform a random-effects meta-analysis, run the following code:

```
rma(yi = kuiper2013$beta, vi = kuiper2013$vi)
```

```
## Random-Effects Model (k = 4; tau^2 estimator: REML)
## tau^2 (estimated amount of total heterogeneity):
0.64 (SE = 0.53)
## Test for Heterogeneity:
## Q(df = 3) = 185.22, p-val < .01
## Model Results:
## estimate se zval pval ci.lb ci.ub
##  0.76 0.40 1.90 0.06 -0.03 1.55 .
```

Note that the pooled effect size is $\beta = 0.76$. Considering our hypothesis is one-sided, we can divide the p-value by two, and report $p = 0.03$. Thus, there is a significant positive effect of prior interactions on trust. However, this analysis assumes a normal distribution of population effect sizes. We doubt this assumption, because the studies are all qualitatively different. While empirical evidence for this violation of assumptions is given by a significant heterogeneity test, we do not need to make a purely data-driven decision.[35] We can support the assumption that population effect sizes are *not* normally distributed on theoretical grounds.

We now perform the PBF analysis, using the `pbf()` method for numeric input (see `?pbf`). This interface is very similar to `rma()`, and is specifically designed for applications where PBF is applied to meta-analytic datasets. Because PBF relies on Monte Carlo estimation (i.e., randomly sampling values), however, it is advisable to set a seed to make the analysis reproducible. Throughout this tutorial we use the value `set.seed(1)`, but in your analyses, make sure to select a different unique value. Run the following code:

```
set.seed(1)
pbf(yi = kuiper2013$beta,
 vi = kuiper2013$vi,
 ni = kuiper2013$n,
 hypothesis = "y = 0; y > 0; y 0")
```

##			PBF	Sample.1	Sample.2	Sample.3	Sample.4
##	H1:	y=0	0.00	0.25	0.65	0.00	0.00
##	H2:	y>0	1.15e+32	1044.59	208.96	2.29e+13	2.29e+13
##	H3:	y<0	0.00	0.00	0.00	0.00	0.00

Note that the `yi` and `vi` arguments are the same as those of `rma()`. Additional argument `ni` is used to construct the prior for the approximate Bayes factors.[29] Importantly, the `hypothesis` argument determines which informative hypotheses are tested. Its default value of `y = 0` is similar to a null hypothesis test. Here, we override this default value to test our three informative hypotheses. The resulting output shows Bayes factors for each of the three hypotheses (PBF column), as well as the study-specific evidence for each hypothesis (remaining columns). Note that the PBF for hypotheses H_1 and

H_3 are approximately zero; there is essentially no support in the data that the effect of prior interaction on trust is equal to, or smaller than, zero. Support for H_2 is overwhelming, however. Thus, we can conclude that all four included studies support the informative hypothesis that there is a positive effect of prior interaction on trust.

2.2 | Tutorial 2: Computing a Bayes factor

The preceding example used a simplified interface to the PBF, designed to be similar to `rma()` to facilitate analyzing data initially prepared for meta-analysis. The simplified interface internally performs several intermediate steps. These next tutorials go through those steps one by one, to help users understand the calculations involved. As these calculations require individual participant data, we cannot use the sufficient statistics from Tutorial 1. For these tutorials we use the synthetic data based on Van Leeuwen and colleagues.[5] They conducted a theory-driven, preregistered study of the informative hypothesis that higher self-reported moral dispositions would be associated with a more conservative socio-political orientation. Data were collected in three countries: the United States, 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 informative hypothesis. This tutorial follows the same rationale, but uses only one effect size per sample. We intentionally vary the way this effect size is computed to illustrate the use of the PBF in cases when the same informative hypothesis has been studied in different ways in multiple studies. Specifically, we will examine the informative hypothesis that self-reported importance of family morality is positively associated with a conservative socio-political orientation.

We must estimate a model suitable for evaluating this informative hypothesis. Because both scales consist of multiple items, we can use structural equation modeling (SEM) to perform latent variable regression[23]:

```
# 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 =
synthetic_nl)
```

Whereas the previous Tutorial used a test value of zero, similar to conventional frequentist null hypothesis testing, this tutorial uses a smallest effect size of interest.[17] Thus, our informative hypothesis is $H_i: \beta > .1$, where β (beta) is the standardized (partial) regression coefficient, and we hypothesize that its value will be at least .1.

The code below illustrates how to use the `bain()` function to obtain a Bayes factor for this informative hypothesis. In the `hypothesis` argument, we can refer to the parameter `beta` by name because we labeled it in the `lavaan` syntax above. If we had not labeled the parameter, we could inspect all parameter names and their values by running `get_estimates(results_nl, standardize = TRUE)`.

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

## Bayesian informative hypothesis testing for an
object of class lavaan:
##
##  Fit Com BF.u BF.c PMPa PMPb PMPc
## H1 0.96 0.50 1.92 23.25 1.00 0.66 0.96
## Hu      0.34
## Hc 0.04 0.50 0.08    0.04
##
## Hypotheses:
## H1: beta>.1
```

The results indicate that the informative hypothesis receives about 23 times as much support from the data relative to its complement, which is convincing evidence.

2.3 | Tutorial 3: Aggregating Bayes factors

As mentioned before, suitable data were collected to evaluate the informative 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 synthetic_dk and synthetic_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 =
synthetic_dk)
results_us <- sem(model = model_us, data =
synthetic_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
set.seed(1)
pbf(results, hypothesis = "beta > .1", standardize =
TRUE)

##   PBF Sample.1 Sample.2 Sample.3
## H1: beta>.1 1.01e+27 23.25 1.90e+12 2.29e+13
```

2.4 | Tutorial 4: 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 ($H_i : \beta_{\text{fam}} > 0.1$), but the Dutch sample now contains a sample-specific hypothesis regarding the effect of group morality, namely that $\beta_{\text{grp}} < 0.1$. Note that controlling for the effect of group morality affects the effect of family morality on conservatism, which is now a partial effect. This also affects the Bayes factor for this sample. 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 individual `bain` objects.

```
# Add the additional predictor to the model, label the
effect beta2
model_nl2 <- c(model_nl, "group =~ grp_1 + grp_2 + grp_3
    con ~ beta2 * group")
# Estimate the model in lavaan
results_nl2 <- sem(model = model_nl2, data =
synthetic_nl)

# Obtain BF for each sample
# Note that the Dutch sample has two hypotheses:
set.seed(1)
bf_nl2 <- bain(results_nl2,
  hypothesis = "beta > .1; beta2 < .1",
  standardize = TRUE)
bf_dk <- bain(results_dk, hypothesis = "beta > .1",
standardize = TRUE)
bf_us <- bain(results_us, hypothesis = "beta > .1",
standardize = TRUE)

# Bind bain objects into a list
bfs <- list(bf_nl2, bf_dk, bf_us)

# Call pbf on that list
pbf(bfs)

##   PBF Sample.1 Sample.2 Sample.3
## H1: beta>.1 5.68e+28 1301.18 1.90e+12 2.29e+13
```

In the output of this analysis, hypotheses common to all `bain` objects are retained and aggregated, but the sample-specific hypothesis of the first object is omitted. If there are no common hypotheses across all objects, `pbf()` throws an error. As can be seen, the results are somewhat different due to the additional control variable—but the conclusions are equivalent to the previous tutorial. Note the location of the `set.seed()` command: it is called before the initial `bain()` call, because the `bain()` function relies on random sampling—but `pbf()` does not. Compare this to Tutorial 3, where we called `set.seed()` before `pbf()`. In that case, `pbf()` called the `bain()` function internally.

2.5 | Tutorial 5: Using sufficient statistics

Bringing our examples full circle, we once again illustrate how to compute the PBF from sufficient statistics—but in this case, we will calculate them ourselves instead of extracting them from published papers as in Tutorial 1. In this Tutorial, we use the default interface of `bain`, as explained in Ref. [29]. 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.

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 synthetic_nl data; using Cohen's *D* gives $H_i^{\text{synthetic}_n l} : \delta_{\text{conservative} > \text{liberal}} > 0$, where δ is the mean difference between groups.
2. A bivariate regression coefficient was calculated using the synthetic_dk data, giving $H_i^{\text{synthetic}_d k} : \beta_{\text{fam}} > 0$.
3. A correlation coefficient was calculated using the synthetic_us data, giving $H_i^{synthetic_u s} : \rho_{\text{fam,con}} > 0$, 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
synthetic_nl <- data.frame(
 family = rowMeans(synthetic_nl[c("fam_1", "fam_2",
"fam_3")]),
 conservative = rowMeans(synthetic_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")]))
synthetic_dk <- data.frame(
 family = rowMeans(synthetic_dk[c("fam_1", "fam_2",
"fam_3")]),
 conservative = rowMeans(synthetic_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")]))
synthetic_us <- data.frame(
 family = rowMeans(synthetic_us[c("fam_1", "fam_2",
"fam_3")]),
 conservative = rowMeans(synthetic_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")]))

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

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

# synthetic_us: Correlation coefficient
US_est <- cor(synthetic_us)[1, 2]
US_var <- (1 - US_est^2)^2 / (nrow(synthetic_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 informative hypothesis on each parameter estimate. The `pbf()` function can be called on a list of the resulting bain objects.

```
# Use bain.default() to obtain BF for the central
hypothesis
NL_bain <- bain(x = NL_est,
   Sigma = matrix(NL_var, 1, 1),
   n = nrow(synthetic_nl),
   hypothesis = "parameter > 0",
```

```
  joint_parameters = 1)
DK_bain <- bain(x = DK_est,
  Sigma = matrix(DK_var, 1, 1),
  n = nrow(synthetic_dk),
  hypothesis = "parameter > 0",
  joint_parameters = 1)
US_bain <- bain(x = US_est,
  Sigma = matrix(US_var, 1, 1),
  n = nrow(synthetic_us),
  hypothesis = "parameter > 0",
  joint_parameters = 1)

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

##    PBF Sample.1 Sample.2 Sample.3
## H1: parameter>0 9.87e+20 2.29e+13 5.41e+05 79.66
```

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()`.

3 | CONCLUSION

The PBF aggregates evidence for a common informative hypothesis across conceptual replication studies that are too heterogeneous to meet the assumptions of conventional meta-analysis. We introduced a new user-friendly implementation of the existing PBF method in the `bain` R-package. A simulation study, available in Appendix S1, demonstrated that the PBF's overall accuracy compared favorably to other commonly used research synthesis methods. However, the PBF traded increased sensitivity for decreased specificity (see Appendix S1). Several tutorial examples demonstrated the use of the PBF in different scenarios. The first Tutorial illustrated how the PBF can be used to aggregate evidence based on sufficient statistics commonly collected to perform conventional meta-analysis. In this context, the PBF can replace, or complement, conventional meta-analyses when their assumptions about between-studies heterogeneity are (thought to be) violated. Other Tutorials demonstrated the use of the PBF with individual participant data, and exposed the intermediate steps of formulating informative hypotheses about model parameters and computing Bayes factors. The final example demonstrated how to extract sufficient statistics from individual participant data analyses and compute a PBF, thus circling back to the first Tutorial.

One important distinction between meta-analysis and PBF is that both methods answer different research questions. Meta-analysis estimates the value of an overall effect size and its heterogeneity, whereas the PBF expresses support for an informative hypothesis in terms of the Bayes factor. While this may be perceived as a loss of information, the PBF can be used in situations where the assumptions of meta-analysis are violated, and consequently, pooled effect size estimates would not be informative. These two evidence synthesis methods thus represent different approaches to inference and answer different research questions. 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, the available information, and assumptions about between-studies heterogeneity.

Another important distinction is that meta-analysis provides ever-more precise estimates of the population effect size as more studies are added. This is not the case for the PBF. Only if all aggregated studies support the informative hypothesis does the PBF increase monotonically with the number of studies. However, if studies offer mixed evidence of the informative hypothesis, the PBF can be inconclusive, even for a large number of studies. This is because the PBF essentially answers the question: do these studies all support the informative hypothesis? If the answer is no, the results will reflect that.

One final limitation of the interpretation of the PBF worth considering is that the PBF (as implemented here) renders support for one specific informative hypothesis versus its complement. If the informative hypothesis is supported, this does not 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 wrong (the earth is spherical). If we would have evaluated another hypothesis, for example, the earth is shaped like an American football, it would likely have received much more support, for example $BF = 1000$, even though it is also wrong—but less egregiously so. If both hypotheses were compared against the hypothesis that the earth is round, their relative support would be infinitesimal. But if the true hypothesis is not in the set, other comparisons might be misleading. 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 by the data, but does not mean that the hypothesis is definitively false. The conclusion is also affected by factors like sample size, measurement error, and sampling variance.

In sum, the present paper makes the PBF method more broadly accessible by implementing it in the `bain` R-package, validating it with a simulation study, and illustrating its use in several reproducible tutorial examples. Researchers should be mindful of the fact that the

PBF does answer a different research question than meta-analysis, and uses a different inferential procedure (Bayesian instead of frequentist). Consequently, the choice of research synthesis method and interpretation of the results should be made with care and aligned with the research question. Our simulation study suggests that PBF is a useful evidence synthesis method, which can be used when the assumptions of meta-analysis are (likely) violated.

AUTHOR CONTRIBUTIONS

Caspar J. Van Lissa: Conceptualization; methodology; software; validation; formal analysis; supervision; funding acquisition; project administration; writing – original draft; writing – review and editing. **Eli-Boaz Clapper:** Software; validation; formal analysis; writing – original draft; writing – review and editing. **Rebecca Kuiper:** Conceptualization; writing – original draft; writing – review and editing.

ACKNOWLEDGMENT

None.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

All analysis code is available in a version-controlled repository at https://doi.org/10.5281/zenodo.11615354.

ORCID

Caspar J. Van Lissa https://orcid.org/0000-0002-0808-5024

Rebecca Kuiper https://orcid.org/0000-0002-0267-5197

REFERENCES

1. Brembs B. Prestigious science journals struggle to reach even average reliability. *Front Hum Neurosci*. 2018;12. doi:10.3389/fnhum.2018.00037
2. Lavelle JS. When a crisis becomes an opportunity: the role of replications in making better theories. *Br J Philos Sci*. 2021;714:965-986. doi:10.1086/714812
3. Kuiper RM, Buskens V, Raub W, Hoijtink H. Combining statistical evidence from several studies: a method using Bayesian updating and an example from research on trust problems in social and economic exchange. *Sociol Methods Res*. 2013;42(1):60-81. doi:10.1177/0049124112464867
4. Klugkist I, Volker TB. Bayesian evidence synthesis for informative hypotheses: an introduction. *Psychol Methods*. 2023. doi:10.1037/met0000602
5. Van Leeuwen F, Van Lissa CJ, Papakonstantinou T, Petersen MB, Curry OS. Morality as cooperation, politics as conflict. *Soc Psychol Bull*. 2024;19:e10157. doi:10.32872/spb.10157
6. Higgins JPT, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *J R Stat Soc Ser A (Stat Soc)*. 2009;172(1):137-159. doi:10.1111/j.1467-985X.2008.00552.x
7. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. *Introduction to meta-analysis*. John Wiley & Sons, Ltd; 2009. doi:10.1002/9780470743386
8. Van Lissa CJ. Small sample meta-analyses: exploring heterogeneity using MetaForest. In: van de Schoot R, Miočević M, eds. *Small Sample Size Solutions (Open Access): A Guide for Applied Researchers and Practitioners*. CRC Press; 2020.
9. van Lissa CJ, van Erp S, Clapper E-B. Selecting relevant moderators with Bayesian regularized meta-regression. *Res Synth Methods*. 2023;14(2):301-322. doi:10.1002/jrsm.1628
10. Batenburg RS, Raub W, Snijders C. Contacts and contracts: temporal embeddedness and the contractual behavior of firms. *Res Sociol Organ*. 2003;20:135-188.
11. Buskens V, Raub W. Embedded trust: control and learning. *Adv Group Process*. 2002;19:167-202.
12. Buskens V, Weesie J. An experiment on the effects of embeddedness in trust situations: buying a used car. *Ration Soc*. 2000;12:227-253.
13. Buskens V, Raub W, van der Veer J. Trust in triads: an experimental study. *Soc Netw*. 2010;32:301-312.
14. Kwee CMB, Leen NA, van der Kamp RC, et al. Anxiolytic effects of endocannabinoid enhancing compounds: a systematic review and meta-analysis. *Eur Neuropsychopharmacol*. 2023;72:79-94. doi:10.1016/j.euroneuro.2023.04.001
15. Pearl J. The seven tools of causal inference, with reflections on machine learning. *Commun ACM*. 2019;62(3):54-60. doi:10.1145/3241036
16. van Assen MALM, Stoevenbelt AH, van Aert RCM. The end justifies all means: questionable conversion of different effect sizes to a common effect size measure. *Religion Brain Behav*. 2023;13(3):345-347. doi:10.1080/2153599X.2022.2070249
17. Lakens D. Equivalence tests: a practical primer for t tests, correlations, and meta-analyses. *Soc Psychol Person Sci*. 2017;8(4):355-362. doi:10.1177/1948550617697177
18. Gu X, Mulder J, Hoijtink H. Approximated adjusted fractional Bayes factors: a general method for testing informative hypotheses. *Br J Math Stat Psychol*. 2018;71(2):229-261. doi:10.1111/bmsp.12110
19. Mulder J. Prior adjusted default Bayes factors for testing (in) equality constrained hypotheses. *Comput Stat Data Anal*. 2014;71:448-463. doi:10.1016/j.csda.2013.07.017
20. Benjamin DJ, Berger JO, Johannesson M, et al. Redefine statistical significance. *Nat Hum Behav*. 2017;2(1):6-10. doi:10.1038/s41562-017-0189-z
21. Rouder JN, Speckman PL, Sun D, Morey RD, Iverson G. Bayesian t tests for accepting and rejecting the null hypothesis. *Psychon Bull Rev*. 2009;16(2):225-237. doi:10.3758/PBR.16.2.225
22. Peikert A. *Towards Transparency and Open Science (doctoral thesis, Humboldt-Universität zu Berlin)*. Humboldt-Universität zu Berlin; 2023. doi:10.18452/27056
23. Van Lissa CJ, Gu X, Mulder J, Rosseel Y, Zundert CV, Hoijtink H. Teacher's corner: evaluating informative hypotheses using the Bayes factor in structural equation models. *Struct Equ Model Multidiscip J*. 2020;28(2):1-10. doi:10.1080/10705511.2020.1745644

24. Hoijtink H. Prior sensitivity of null hypothesis Bayesian testing. *Psychol Methods*. 2022;27(5):804-821. doi:10.1037/met0000292
25. Heck DW, Boehm U, Böing-Messing F, et al. A review of applications of the Bayes factor in psychological research. *Psychol Methods*. 2022;28:558-579. doi:10.1037/met0000454
26. Munafò MR, Davey Smith G. Robust research needs many lines of evidence. *Nature*. 2018;553(7689):399-401. doi:10.1038/d41586-018-01023-3
27. Ioannidis JPA. The reproducibility wars: successful, unsuccessful, uninterpretable, exact, conceptual, triangulated, contested replication. *Clin Chem*. 2017;63(5):943-945. doi:10.1373/clinchem.2017.271965
28. Zwaan RA, Etz A, Lucas RE, Donnellan MB. Making replication mainstream. *Behav Brain Sci*. 2018;41:e120. doi:10.1017/S0140525X17001972
29. Hoijtink H, Mulder J, van Lissa C, Gu X. A tutorial on testing hypotheses using the Bayes factor. *Psychol Methods*. 2019;24(5):539-556. doi:10.1037/met0000201
30. Hedges LV, Olkin I. Vote-counting methods in research synthesis. *Psychol Bull*. 1980;88:359-369. doi:10.1037/0033-2909.88.2.359
31. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Softw*. 2010;36(3):1-48. http://www.jstatsoft.org/v36/i03/
32. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. *Br Med J*. 2010;340(7745):521-525. https://www.jstor.org/stable/25674217
33. Van Lissa CJ, Brandmaier AM, Brinkman L, et al. WORCS: a workflow for open reproducible code in science. *Data Sci*. 2021;4(1):29-49. doi:10.3233/DS-210031
34. Team, R. C. (2022). *R: A Language and Environment for Statistical Computing*. R Project. https://www.r-project.org
35. Wicherts JM, Veldkamp CLS, Augusteijn HEM, Bakker M, van Aert RCM, van Assen MALM. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: a checklist to avoid p-hacking. *Front Psychol*. 2016;7:1832. doi:10.3389/fpsyg.2016.01832

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Van Lissa CJ, Clapper E-B, Kuiper R. A tutorial on aggregating evidence from conceptual replication studies using the product Bayes factor. *Res Syn Meth*. 2024;15(6):1231-1243. doi:10.1002/jrsm.1765