Received: 10 October 2018 | Revised: 24 April 2019 | Accepted: 7 June 2019

DOI: 10.1002/jrsm.1365

WILEY

TUTORIAL

A comparison of Bayesian synthesis approaches for studies comparing two means: A tutorial

Han Du[1] | **Thomas N. Bradbury**[1] | **Justin A. Lavner**[2] | **Andrea L. Meltzer**[3] | **James K. McNulty**[3] | **Lisa A. Neff**[4] | **Benjamin R. Karney**[1]

[1]Psychology, University of California, Los Angeles, California

[2]Psychology Department, University of Georgia, Athens, Georgia

[3]Department of Human Development and Family Sciences, Florida State University, Tallahassee, Florida

[4]Educational Psychology, University of Texas at Austin, Austin, Texas

Correspondence
Han Du, Psychology, University of California,1285 Franz Hall, Box 951563, Los Angeles, CA 90095-1563.
Email: hdu@psych.ucla.edu

Researchers often seek to synthesize results of multiple studies on the same topic to draw statistical or substantive conclusions and to estimate effect sizes that will inform power analyses for future research. The most popular synthesis approach is meta-analysis. There have been few discussions and applications of other synthesis approaches. This tutorial illustrates and compares multiple Bayesian synthesis approaches (i.e., integrative data analyses, meta-analyses, data fusion using augmented data-dependent priors, and data fusion using aggregated data-dependent priors) and discusses when and how to use these Bayesian synthesis approaches to combine studies that compare two independent group means or two matched group means. For each approach, fixed-, random-, and mixed-effects models with other variants are illustrated with real data. R code is provided to facilitate the implementation of each method and each model. On the basis of these analyses, we summarize the strengths and limitations of each approach and provide recommendations to guide future synthesis efforts.

KEYWORDS
Bayesian statistics, data fusion, integrative data analysis, meta-analysis

1 | INTRODUCTION

Researchers seek to combine data from multiple studies for many reasons. For example, when several studies have addressed the same research question, researchers may want to combine their results to draw an overall conclusion. When researchers have conducted pilot studies, rather than ignoring their results, they often want to merge the results of those studies with the results from formal studies. This process has also been called data or research synthesis,[1] data integration,[2] or data fusion.[3]

Such data syntheses offer several benefits. First, they capitalize on the money, time, and resources invested in the existing studies, allowing investigators to address new questions. Second, data synthesis is an efficient way to conduct sample size planning.[4] To determine sample size, power analyses,[5] rely on estimated parameters or estimated effect sizes, but these estimates are uncertain because they are not the true/population parameters. Different estimates lead to different planned sample sizes, and researchers face questions about which planned sample size should be used. Data synthesis combines pertinent information from existing studies and pilot studies to provide overall estimates of population parameters/effect sizes and thereby determine the sample size in future research. Because the overall sample size is increased with data synthesis, standard errors of the estimates are decreased, and the uncertainty of sample size planning is reduced. Third, despite being largely ignored in power analysis research, Bayesian data synthesis provides a way to calculate statistical power taking uncertainty into account. Even though sample size planning can be

conducted based on multiple studies, because of the uncertainty of estimates, we still cannot predict what estimates will be observed in a formal study. Accordingly, even with sample size planning, conventional statistical power is risky because it could be too low with the planned sample size if the sample effect size in the future study is too small.[6] Bayesian power analysis that combines multiple studies and also considers uncertainty in parameter estimation can be used for sample size planning to claim statistical power with a certain assurance level. In this paper, we will outline how to conduct such analyses.

This paper focuses on four different approaches to data synthesis. First, in meta-analysis, the most popular data synthesis approach, aggregated study-specific results, such as standardized group mean differences, are analyzed. Second, in integrative data analysis, multiple data sets are merged into a single data set, allowing researchers to conduct subsequent data analyses on all of the raw data simultaneously; thus, all the original information is kept and the influence of covariates at different levels can be examined. Although the first and second approaches can be employed in both frequentist and Bayesian frameworks, when the model is complex, Bayesian modeling is more feasible and less mathematically challenging because Bayesian statistics have algorithms such as Gibbs sampling and Metropolis–Hastings algorithm. More specifically, Bayesian modeling has advantages in dealing with high-dimensional data and/or nonlinear functions because Bayesian algorithms can avoid calculating integration and obtaining analytical closed form. In comparing two means, since the models are usually not very complex, both frequentist and Bayesian estimation are applicable. This paper presents Bayesian integrative data analysis and Bayesian meta-analysis in the special case of comparing two means. Illustrating the Bayesian methods in this special case allows the paper to introduce important fundamental concepts without mathematical complications inherent in more complex models (eg, multidimensional logistic regression and multidimensional item response theory model). The complex models would interfere with developing a conceptual understanding of the methods. Illustrating Bayesian synthesis approaches in a simple model establishes a foundation for generalizing these Bayesian methods to more complex models. The third approach, data fusion using augmented data-dependent priors (AUDPs), is a pure Bayesian method, in which each study's information contributes to the inference sequentially and the contribution of each study is summarized. The fourth approach, data fusion using aggregated data-dependent priors (AGDP), is also within the Bayesian framework and uses aggregated informative priors constructed by multiple studies. In meta-analysis and integrative data analysis, information from all data sets enters the models simultaneously, whereas in the latter two approaches, the information of data sets is either entered sequentially (AUDP) or summarized as priors (AGDP). The first two approaches and latter two approaches are not unrelated: AUDP and AGDP could either use the models for aggregated data as in meta-analysis or use the models for raw data as in integrative data analysis. To provide an overall picture of these methods, the strengths and limitations of each method are presented in Table 1 and will be elaborated further in the following sections. In contrast to the meta-analysis, the other three approaches are seldom mentioned in the literature or used in practice, despite their strengths.

In the current paper, we present when and how to use these synthesis approaches in the Bayesian framework to combine studies that compare two independent or matched group means, and we illustrate our work with real data.[1] R code is provided to facilitate the implementation of each method and each model. We begin by introducing our data, which comes from research on marital satisfaction in couples. When introducing the Bayesian synthesis approaches, we start by introducing different models for integrative data analysis. Then, we apply different Bayesian integrative data analyses to the aforementioned real data. Next, different models for meta-analysis are presented, and the Bayesian meta-analyses are illustrated with real data. Because the models for integrative data analysis and meta-analysis can also be used in AUDP and AGDP, we will discuss each model with details. Then, the AUDP and AGDP approaches are introduced and illustrated with real data. Later, Bayesian estimation coupled with Bayesian power is presented for sample size planning. Finally, we offer recommendations for future synthesis efforts.

[1]We did not use a simulation study to compare these approaches for several reasons. First, no matter which approach or prior is used, the posterior mode is a consistent estimator for a coefficient when the model is correctly specified.[7] Second, if we use informative priors that center around the true values (equivalent to increasing prior sample size), there is no doubt that the estimation will become better than the one using informative priors that deviate away from the true values or using noninformative priors. We can make such conclusions without simulation. Third, in practice, the true values are unknown. Accordingly, researchers need to decide whether using informative priors and how informative the priors are based on their understanding and experience on the specific topic. Fourth, as illustrated later, different approaches rely on different study information. Some approaches use raw data; some use aggregated data, and some use characteristics of studies to weight studies. It is unfair to compare methods using richer information with the methods using sparse information. In practice, we choose the most appropriate methods based on the available information, and such guidelines are provided in Table 1.

TABLE 1 Summary of integrative data analysis, meta-analysis, data fusion using augmented data-dependent priors, data fusion using aggregated data-dependent priors, fixed-effects model, random-effects model, and mixed-effects model

Method	Strengths	Limitations
Integrative data analysis	Straightforward; retain all the original information; examine the influence of the study-level, pair-level, and individual-level covariates	Raw data are required; labor-intensive, time-consuming, and costly; the measurements used across studies should be the same or can be equated
Meta-analysis	Only require sample effect sizes and sample sizes; Allow different measurements across studies; less labor-intensive, less time-consuming, and cheaper	The normality approximation is good only when the per-study sample size is relatively large; no individual- or pair-level covariates; less powerful
Data fusion using augmented data-dependent priors (AUDPs)	The contribution of each study is clearly summarized; accommodate raw data or aggregated data	The order of the data sets in entering the analysis could influence results, thus sensitivity analysis is needed
Data fusion using aggregated data-dependent priors (AGDPs)	An intuitive way to construct prior	Lose precision; results depend on which studies to construct the priors and which study to serve as the formal study; thus, it is fundamentally flawed
Model	**Strengths**	**Limitations**
Fixed-effects models	A simple model	Fail to consider between-study heterogeneity
Random-effects models	Take between-study heterogeneity into account	The estimate of between-study heterogeneity will have poor precision when the number of studies is very small
Mixed-effects models	Consider covariates to further explain population discrepancies	The estimate of parameters will have poor precision and nonconvergence may occur when the number of studies is very small and/ or number of covariates is large

2 | OUR REAL DATA AND SUBSTANTIVE RESEARCH QUESTIONS

2.1 | Data information

The analyses that follow illustrate various ways of synthesizing 11 independent data sets. Characteristics of the data sets are presented in Table 2. All 11 data sets consist of heterosexual, married couples, most of whom were contacted within 6 months of their wedding date. All 11 studies administered a common measure of marital satisfaction: a version of the Semantic Differential.[8,9] Spouses were asked to rate their marriage on a 7-point scale between each pair of opposing adjectives (eg, hopeful–discouraging and bad–good). Scores on each item were reverse coded as appropriate, and the sum of the ratings was treated as an index of marital satisfaction with a potential range of 15 to 105. Coefficient alpha was above 0.90 for both spouses in all studies.

Before we address our substantive question, it is worth noting that combining these 11 data sets creates a data set with a three-level structure, in which individuals (level 1) are nested within couples (level 2) who are nested within studies (level 3). Because there are not enough degrees of freedom, the discussed data synthesis approaches use paired difference scores within couples.[4] In this case, the individual level disappears, and the data have a two-level structure (i.e., couple level and study level). While we present analyses that treat spouses as matched pairs, we also illustrate how to conduct Bayesian synthesis for independent groups where we assume husbands and wives are independent

[4]This paper considers fixed-effects models that constrain between-study heterogeneity to be 0 and random-effects models that freely estimate between-study heterogeneity. However, the extremely small level 1 sample size (only two individuals per couple) fails to provide enough pieces of information to allow a three-level model, which means that the model will be unidentified. If we simply constrain the variances at the second level at 0 (i.e., the between-couple variances), the between-couple covariances are the same as the within-couple covariances, which ignores within-couple interdependence. An exception is the fixed-effects model where there is no between-study heterogeneity, for which a two-level model that considers an individual-level residual variance and a couple-level variance can be applied to the raw scores, and the model is identified. The limited degrees of freedom is a classic problem in dyadic data. We refer the interested reader to Du and Wang[10] and Kenny et al.[11]

TABLE 2 Characteristics of 11 independent samples of married couples

PI	N	Location	Year Initiated	Compensation	Eligibility Criteria	Funding Source
Thomas N. Bradbury	60	Los Angeles, CA	1991	$50	First married, childless, newlyweds married less than 6 months	University of California, Los Angeles
Thomas N. Bradbury	172	Los Angeles, CA	1993	$75	First married, childless, newlyweds married less than 6 months	NIMH[a]
Benjamin R. Karney	82	Gainesville, FL	1998	$50	First married, childless, newlyweds married less than 6 months	University of Florida
Benjamin R. Karney	169	Gainesville, FL	2001	$70	First married, childless, newlyweds married less than 6 months	NIMH[a]
James K. McNulty	72	Mansfield, OH	2003	$60	First married, children allowed, newlyweds married less than 6 months	University of Ohio
Lisa A. Neff	61	Toledo, OH	2005	$70	First married, childless, newlyweds married less than 6 months	University of Toledo
James K. McNulty	135	Knoxville, TN	2006	$80	First married, childless, newlyweds married less than 6 months	NICHD[a]
Andrea L. Meltzer	113	Dallas, TX	2013	$100	First marriage, newlyweds married less than 6 months	Southern Methodist University
James K. McNulty	119	Tallahassee, FL	2013	$100	Remarriages allowed, children allowed, newlyweds married less than 3 months	NSF[a]
Andrea L. Meltzer	99	Tallahassee, FL	2016	$100	Remarriages allowed, newlyweds married less than 6 months	Florida State University
James K. McNulty	143	Tallahassee, FL	2016	$50	Married couples (not newlyweds), remarriages allowed, children allowed	DoD[a]

Abbreviations: N, the number of couples; NICHD, Eunice Kennedy Shriver National Institute of Child Health and Human Development; NIMH, National Institute of Mental Health; NSF, National Science Foundation; DoD, Department of Defense.
[a]Extramural funding.

within each study. In this case, couple level disappears, and the data structure changes to two levels (i.e., individual level and study level). We want to emphasize that assuming independent husbands and wives is only for pedagogical purposes and the related models and methods are only appropriate when the two groups are really independent; in actual research, husbands and wives should be considered as interdependent pairs as illustrated in the paper.

2.2 | Substantive question

Within each study, the presence of marital satisfaction data provided by each spouse allows us to evaluate a common substantive question: on average, do husbands and wives differ in their evaluations of their relationship? Hundreds of studies of married couples have collected data relevant to this issue, but to date, attempts at synthesizing this literature have relied exclusively on meta-analysis.[12] In the analyses that follow, we illustrate several Bayesian data synthesis approaches including Bayesian meta-analytic techniques.

3 | INTEGRATIVE DATA ANALYSIS MODELS

When the original data from individual studies are available, it is straightforward to conduct an integrative data analysis, also referred to as individual participant-level data meta-analysis, pooled analysis, or mega-analysis.[13,14] In an integrative data analysis, all the raw data from different studies are merged into a large data set and are then analyzed as a whole. Integrative data analysis is not widely used in psychology, with some exceptions.[15-17] One prerequisite of integrative data analysis is that the measurements (i.e., questionnaire or survey) used across studies are the same or can be equated. When studies to be combined use different measurements, equating is necessary using either item response theory or classical test theory to ensure that scores are comparable across studies.[16,18] Because this paper uses studies that all administered the same measure of marital satisfaction, equating will not be discussed in detail here. With this prerequisite, a traditional analysis can be applied based on the pooled data.

When pooled data come from multiple studies, analysis options include fixed-effects, random-effects, and mixed-effects models. The strengths and limitations of each model are provided in Table 1. Fixed-effects models assume all studies have the same fixed true parameters, whereas random-effects models assume true parameters are a random sample from the population of parameters, and thus, different studies differ in their true parameters.[19,20] If we incorporate study-level covariates (eg, sampling and design characteristics) and/or individual-level covariates (eg, age, education level, and socioeconomic status), population discrepancies could be explained by the covariates to some degree. This model is referred to as a mixed-effects model.[20,21] In addition, study-level covariates such as birth cohort and year of experiments provide a way to distinguish cohort effects from age effects. Compared with fixed-effects integrative data analysis, random-effects integrative data analysis and mixed-effects integrative data analysis are seldom discussed.[13]

In practice, for both integrative data analysis and meta-analysis, researchers can test for between-study variance to help choose between random-/mixed-effects and fixed-effects models. If the between-study variance is statistically significant, a random-effects model can be used, and a mixed-effect model can be further considered to explore the influence of covariates; otherwise, a fixed-effects model should be used. However, the test for between-study variance may not be powerful enough to detect heterogeneity. The setup of the random- and mixed-effects model is more general and flexible. In addition, from a practical perspective, it would be unlikely that real-life studies are homogeneous, making $\tau^2=0$ an implausible assumption.[22-24] Therefore, even when the test for between-study variance is significant, researchers may still prefer to use the random- or mixed-effects model. We will introduce each model when the two groups are independent or matched by pairs, separately.

3.1 | Testing mean differences for independent groups

Suppose there are J studies comparing two independent group means. For study j, group 1 has a sample size of n_{1j}, and group 2 has a sample size of n_{2j}. The null hypothesis is that the population mean of group 1 (μ_1) equals the population mean of group 2 (μ_2) averaging over studies, H_0:$\mu_1-\mu_2=0$.

3.1.1 | Fixed-effects integrative data analysis

A fixed-effects integrative data analysis assumes that there is no between-study heterogeneity. After pooling the data sets, the subscript j is not needed; the individual i's score from group 1 is denoted as y_{1i}, and the individual i's score from group 2 is denoted as y_{2i}. Thus, it is just an independent t test based on the pooled data, y_{1i}s and y_{2i}s, which also can be written as a simple regression.[5] When the data from two groups are normally distributed with equal variance, we assume that $y_{1i}\sim N(\mu_1,\sigma^2)$ and $y_{2i}\sim N(\mu_2,\sigma^2)$. When data from two groups are normally distributed but with unequal variances, we assume that $y_{1i}\sim N(\mu_1,\sigma_1^2)$ and $y_{2i}\sim N(\mu_2,\sigma_2^2)$.

3.1.2 | Random-effects integrative data analysis

For random-effects integrative data analysis, multilevel modeling is performed to take the between-study heterogeneity into account. Let G_{ij} be an indicator variable for the two groups, whereby $i=1,\ldots,(n_{1j}+n_{2j})$ indicates participant in each study, and $j=1,\ldots,J$ indicates study. When a participant is from group 1 and study j, $G_{ij}=1$; otherwise, $G_{ij}=0$. A two-level model is defined as follows:

$$\begin{aligned} y_{ij} &= \beta_{0j}+\beta_{1j}G_{ij}+\epsilon_{ij} \\ \beta_{0j} &= \beta_{00}+u_{0j} \\ \beta_{1j} &= \beta_{10}+u_{1j} \end{aligned}, \qquad (1)$$

where y_{ij} is the dependent variable, ϵ_{ij} is the level 1 residual with distribution $\epsilon_{ij}\sim N\left(0,\sigma_\epsilon^2\right)$, β_{0j} is the mean of group 1 in the jth study, β_{00} is the overall mean of group 1 across studies (i.e., population grand mean of group 1), β_{1j} is the group mean difference in the jth study, β_{10} is the group mean difference between groups 1 and 2 across studies (i.e., population group mean difference), u_{0j} and u_{1j} are the level 2 residuals, and $\begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix}\sim MN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix},\sum_s=\begin{pmatrix} \sigma_{u0}^2 & \sigma_{u01} \\ \sigma_{u01} & \sigma_{u1}^2 \end{pmatrix}\right)$ represents the between-study heterogeneity in the means of group 1 and in the mean differences between groups 1 and 2.

3.1.3 | Mixed-effects integrative data analysis

For mixed-effects integrative data analysis, we consider both study-level covariates (eg, the year of survey and site) and individual-level covariates (eg, education level) to further explain population discrepancies at different

[5]We can assume a simple regression with a binary predictor (when a participant is from group 1, $G_i=1$; otherwise, $G_i=0$), $y_i=\beta_0+\beta_1G_i+\epsilon_i$. β_1 is the group mean difference (i.e., $\mu_1-\mu_2$).

levels (i.e., within-study variance and between-study variance), compared with random-effects integrative data analysis, which is a multilevel empty model (i.e., no predictors). Thus, Model (1) changes to the following:

$$\begin{aligned} y_{ij} &= \beta_{0j} + \beta_{1j}G_{ij} + \beta_{2j}X_{ij} + \epsilon_{ij} \\ \beta_{0j} &= \beta_{00} + \beta_{01}Z_j + u_{0j} \\ \beta_{1j} &= \beta_{10} + \beta_{11}Z_j + u_{1j} \\ \beta_{2j} &= \beta_{20} + \beta_{21}Z_j + u_{2j} \end{aligned}, \quad (2)$$

where β_{00} is the overall mean of group 1 after controlling for an individual-level covariate X_{ij} and a study-level covariate Z_j by fixing them at 0, β_{01} is the effect of the study-level covariate Z_j on the scores of group 1, β_{10} is the group mean difference between groups 1 and 2 after controlling for X_{ij} and Z_j by fixing them at 0, β_{11} is the effect of Z_j on the differences between groups 1 and 2, β_{20} is the effect of the individual-level covariate X_{ij}, β_{21} is the cross-level interaction effect of the two covariates X_{ij} and Z_j, u_{0j}, u_{1j}, and u_{2j} are the level 2 residuals, and $\begin{pmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \end{pmatrix} \sim MN\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \sum_s = \begin{pmatrix} \sigma^2_{u0} & \sigma_{u01} & \sigma_{u02} \\ \sigma_{u01} & \sigma^2_{u1} & \sigma_{u12} \\ \sigma_{u02} & \sigma_{u12} & \sigma^2_{u2} \end{pmatrix}\right)$ represents the between-study heterogeneity in the intercepts of group 1, in the intercept differences between groups 1 and 2, and in the effects of X_{ij}. Different ways of centering X_{ij} and Z_j will change the research question that β_{10} answers. Without centering, β_{10} is the group mean difference when X_{ij} and Z_j are zero; with centering, β_{10} is the group mean difference when X_{ij} and Z_j are equal to the values that they are centered at. There are different ways to center X_{ij} and Z_j. For example, we can center X_{ij} and Z_j at their minimal values, in which case β_{10} is the group mean difference between groups 1 and 2 when X_{ij} and Z_j are at their minimal values. We can center X_{ij} at its grand mean across individuals and studies and center Z_j at its mean, in which case β_{10} is the group mean difference between groups 1 and 2 when X_{ij} and Z_j are at their grand means. In addition, we can center X_{ij} at its study-level mean across individuals within each study and center Z_j at its mean, in which case β_{10} is the group mean difference between groups 1 and 2 when X_{ij} is at the study-level mean and Z_j is at the grand mean.

Model (2) assumes that the individual-level covariate X_{ij} has the same effect on both groups. To allow X_{ij} to have different effects on different groups, an interaction term $G_{ij}X_{ij}$ can be added to the level 1 equation in Model (2). Then, we have two more fixed effects: β_{30} is the effect of X_{ij} on the group difference, and β_{31} is the interaction effect of X_{ij} and Z_j on the group difference.

3.2 | Testing mean differences for matched groups

The examined groups are not always independent when comparing two group means. There are cases in which participants in two groups are matched in some way such as twins and couples or when each individual is measured twice under different experimental conditions. The correlation between the two members within pairs influences the results in the matched group tests. Suppose there are J studies of comparing two matched group means. For study j, the number of pairs is n_j, and the paired difference is computed as the difference within the dth pair, $y_{dj}=y_{1dj}-y_{2dj}$. The null hypothesis is that the population group mean difference between groups 1 and 2 averaging over studies is 0.

3.2.1 | Fixed-effects integrative data analysis

A fixed-effects integrative data analysis that assumes the same matched group difference across studies employs a paired t test. After pooling the data from multiple studies, the subscript j is not needed; the difference scores are denoted as y_d, and the total number of pairs is $n=\sum_{j=1}^{J} n_j$. The assumption of paired difference data is $y_d \sim N(\mu_{raw.d}, \sigma^2)$.

3.2.2 | Fixed-effects integrative data analysis with between-pair variance

A fixed-effects integrative data analysis assumes no between-study heterogeneity. Thus, we have enough degrees of freedom to estimate both the between-pair variance and individual-level residual variance. In contrast to the other models for matched groups, we can directly use the raw score for each individual instead of the paired differences. Because no between-study heterogeneity is assumed, the subscript i indicates individual, and the subscript d indicates dyad/pair, and there is no need to distinguish among different studies. Then, the model is as follows:

$$\begin{aligned} y_{id} &= \beta_{0d} + \beta_{1d}G_{id} + \beta_{2d}X_{id} + \beta_{3d}G_{id}X_{id} + \epsilon_{id} \\ \beta_{0d} &= \beta_{00} + \beta_{01}W_d + u_{0d} \\ \beta_{1d} &= \beta_{10} + \beta_{11}W_d \\ \beta_{2d} &= \beta_{20} + \beta_{21}W_d \\ \beta_{3d} &= \beta_{30} + \beta_{31}W_d \end{aligned}, \quad (3)$$

where β_{00} is the overall mean of group 1 after controlling for an individual-level covariate X_{id} and a pair-level covariate W_d by fixing them at 0, β_{01} is the effect of the pair-level covariate W_d on the scores of group 1, β_{10} is the group mean difference between groups 1 and 2 conditional on specific levels of X_{id} and W_d, β_{11} is the effect of W_d on the differences between groups 1 and 2, β_{20} is the effect of the

individual-level covariate X_{id} on group 1, β_{21} is the impact of the pair-level covariate W_d on the effect of X_{id} in group 1, β_{30} is the effect of X_{id} on the group difference, and β_{31} is the interaction effect of X_{id} and W_d on the group difference. Centering X_{id} and W_d and different ways of centering will change the meaning of coefficients and answer different research questions. Because of the extremely small level 1 sample size (i.e., 2), we do not have enough degrees of freedom to estimate all the level 2 variances. Thus, only $u_{0d} \sim N(0, \sigma^2_{u0})$, which represents the between-pair heterogeneity in the pair-specific intercepts of group 1, is estimated.

3.2.3 | Random-effects integrative data analysis

A random-effects integrative data analysis that considers between-study heterogeneity uses multilevel modeling for the paired difference data:

$$\begin{aligned} y_{dj} &= \beta_{0j} + \epsilon_{dj} \\ \beta_{0j} &= \beta_{00} + u_{0j} \end{aligned}, \tag{4}$$

where y_{dj} is the pooled paired differences after combining studies, such that the subscript d indicates dyad/pair and the subscript j indicates study, ϵ_{dj} is the level 1 residual with distribution $\epsilon_{dj} \sim N(0, \sigma^2_{\epsilon})$, β_{00} is the overall group mean difference across studies, and u_{0j} is the level 2 residual with distribution $u_{0j} \sim N(0, \sigma^2_{u0})$, which represents the between-study heterogeneity in group mean differences.

3.2.4 | Mixed-effects integrative data analysis

A mixed-effects integrative data analysis does not consider the individual-level covariates for matched pairs because difference scores are directly used. Thus, with a pair-level covariate X_{dj} and a study-level covariate Z_j, a multilevel model is as follows:

$$\begin{aligned} y_{dj} &= \beta_{0j} + \beta_{1j} X_{dj} + \epsilon_{dj} \\ \beta_{0j} &= \beta_{00} + \beta_{01} Z_j + u_{0j} \\ \beta_{1j} &= \beta_{10} + \beta_{11} Z_j + u_{1j} \end{aligned}, \tag{5}$$

where β_{00} is the overall group mean difference after controlling for X_{dj} and Z_j by fixing them at 0, β_{01} is the effect of Z_j on the group difference, β_{10} is the effect of X_{dj}, β_{11} is the cross-level interaction effect of X_{dj} and Z_j, u_{0j} and u_{1j} are level 2 residuals, and $var\begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} = \begin{pmatrix} \sigma^2_{u0} & \sigma_{u01} \\ \sigma_{u01} & \sigma^2_{u1} \end{pmatrix}$ represents the between-study heterogeneity in group mean differences and effects of X_{dj}. Like in the independent groups case, centering X_{dj} and Z_j and different ways of centering will change the meaning of coefficients and answer different research questions.

4 | BAYESIAN INTEGRATIVE DATA ANALYSIS WITH OUR REAL DATA

The fixed-effects integrative data analyses, random-effects integrative data analyses, and mixed-effects integrative data analyses can be implemented in both frequentist and Bayesian frameworks. In the frequentist framework, parameters (eg, β_{10} in Equation (2)) are treated as fixed constants, whereas in the Bayesian framework, parameters are treated as random variables. Thus, in Bayesian modeling, prior distributions for unknown parameters need to be specified based on the prior knowledge of the parameters ($f(\theta)$), and likelihood function given the observed data is constructed based on models ($L(\theta|y)$).[7] With both the likelihood and the priors defined, the posterior distributions of parameters are derived via Bayes' theorem, $f(\theta|y) = \frac{L(\theta|y)f(\theta)}{f(y)} \propto L(\theta|y)f(\theta)$.

To sample from posterior distributions without deriving the analytical closed form, Markov chain Monte Carlo (MCMC) sampling is usually used by software such as BUGS and R packages such as **`rjags`**.[25] In MCMC sampling, a number of early iterations before convergence need to be discarded since they are not representative samples of the target distribution.[7,26] This period is called a burn-in period. In some software and R packages (eg, rjags), there is an initial adaptive period during which MCMC sampling is tuned to maximize the efficiency.[25] Additionally, posterior samples are autocorrelated within chains because of the iterative sampling.[26] To save computation storage, a thinning period is used by taking every kth sampled value (i.e., a thinning period of k). The convergence of the MCMC algorithm can be assessed by visual inspection (i.e., trace plots) as well as statistical tests for each parameter. Gelman and Rubin's test is one of the widely used diagnostic tests and requires more than one independent MCMC chains with different starting values.[27] In Gelman and Rubin's test, a potential scale reduction fact (PSRF) value close to 1 and the upper limit of an interval estimate smaller than 1.1 usually indicate convergence.

In all of the discussed methods, we can calibrate posterior inferences to the frequentist framework: the frequentist statistics provide a useful approach for evaluating the properties of Bayesian inferences.[7,28] That is, posterior samples allow us to compute the posterior mean, posterior mode, quantile-based probability (QBP) interval, and highest posterior density (HPD) interval. The posterior mean and posterior mode provide point estimates of unknown parameters. In particular, the posterior mode is the value that provides the

TABLE 3 Results of the Bayesian integrative data analyses and meta-analyses

Integrative Data Analyses					
Bayesian integrative data analyses for independent groups					
Fixed-effects integrative data	δ_{raw}	μ_1	μ_2	σ^2	
analysis with equal variance	**1.08 (0.08–1.81)**	95.17 (94.57–95.82)	94.21 (93.59–94.84)	122.6 (116.55–130.54)	
Fixed-effects integrative data	δ_{raw}	μ_1	μ_2	σ_1^2	σ_2^2
analysis with unequal variances	**0.95 (0.09–1.87)**	95.2 (94.57–95.79)	94.25 (93.55–94.82)	121.51 (112.35–131.44)	123.96 (115.47–135.37)
Random-effects integrative	β_{00}	β_{10}	ρ	σ_{u0}^2	σ_{u1}^2
data analysis	94.17 (92.44–95.73)	**0.87 (−0.12 to 2.05)**	0.82 (−0.43 to 0.97)	4.39 (1.61–14.17)	0.44 (0.08–2.92)
	σ_e^2				
	115.66 (108.9, 122.06)				
Mixed-effects integrative	β_{00}	β_{01}	β_{10}	β_{11}	β_{20}
data analysis	96.28 (92.82–100.26)	−0.22 (−0.42 to 0)	**2.1 (0.17–4.1)**	−0.1 (−0.21 to 0.03)	−0.23 (−0.92 to 0.45)
	β_{21}	ρ_{01}	ρ_{02}	ρ_{12}	σ_{u0}^2
	0.02 (−0.02 to 0.06)	0.49 (−0.77 to 0.93)	0.01 (−0.66 to 0.66)	0.01 (−0.68 to 0.66)	0.39 (0.08–6.36)
	σ_{u1}^2	σ_{u2}^2	σ_e^2		
	0.29 (0.08–2.37)	0.16 (0.06–0.5)	114.9 (109.17–122.21)		
Mixed-effects integrative	β_{00}	β_{01}	β_{10}	β_{11}	β_{20}
data analysis with	97.65 (93.64–102.74)	−0.23 (−0.51 to 0.01)	**−1.68 (−7.62 to 5.64)**	−0.04 (−0.36 to 0.39)	−0.41 (−1.16 to 0.39)
heterogeneous effect of	β_{21}	β_{30}	β_{31}	ρ_{01}	ρ_{02}
individual-level covariate	0.02 (−0.01 to 0.08)	0.38 (−0.62 to 1.35)	−0.01 (−0.07 to 0.04)	0.24 (−0.83 to 0.95)	0.01 (−0.69 to 0.66)
	ρ_{03}	ρ_{12}	ρ_{13}	ρ_{23}	σ_{u0}^2
	−0.11 (−0.65 to 0.69)	0.01 (−0.64 to 0.72)	−0.16 (−0.79 to 0.51)	−0.11 (−0.63 to 0.57)	0.38 (0.06–4.43)
	σ_{u1}^2	σ_{u2}^2	σ_{u3}^2	σ_e^2	
	0.34 (0.06–3.77)	0.16 (0.06–0.49)	0.12 (0.05–0.41)	115.41 (108.94–121.97)	
Bayesian integrative data analyses for matched groups					
Fixed-effects integrative	$\mu_{raw.d}$	σ^2			
data analysis	**1.03 (0.44–1.67)**	116.46 (107.63–126.65)			
Fixed-effects integrative	β_{00}	β_{10}	β_{20}	β_{30}	σ_{u0}^2
data analysis with	94.36 (92.48–96.19)	**−0.31 (−2.46 to 2.32)**	−0.03 (−0.23 to 0.21)	0.12 (−0.16 to 0.42)	64.53 (56.63–72.29)
between-pair variance	σ_e^2				
	58.4 (54.09–63.75)				

(Continues)

TABLE 3 (Continued)

Integrative Data Analyses				
Random-effects integrative	β_{00}	σ^2_{u0}	σ^2_e	
data analysis	**1.07 (0.21–1.81)**	0.04 (0–2.88)	116.28 (107.45–126.34)	
Mixed-effects integrative	β_{00}	β_{01}	σ^2_{u0}	σ^2_e
data analysis	**2.23 (0.82–3.54)**	−0.08 (−0.17 to 0)	0.04 (0–1.49)	116.14 (107.76–126.4)
Meta-Analyses				
Bayesian meta-analyses for independent groups				
Fixed-effects	μ_δ			
meta-analysis	**0.10 (0.02–0.18)**			
Random-effects	μ_δ	τ^2		
meta-analysis	**0.11 (0.01–0.20)**	0 (0–0.03)		
Mixed-effects	μ_δ	β	τ^2	
meta-analysis	**0.23 (0.04–0.41)**	−0.01 (−0.02 to 0)	0 (0–0.02)	
Fixed-effects meta-	μ_δ			
analysis with power prior	**0.11 (0.03–0.21)**			
Random-effects meta-	μ_δ	τ^2		
analysis with power prior	**0.12 (0–0.22)**	0 (0–0.02)		
Mixed-effects meta-	μ_δ	β	τ^2	
analysis with power prior	**0.25 (0.05–0.46)**	−0.01 (−0.02 to 0)	0 (0–0.02)	
Bayesian meta-analyses for matched groups				
Fixed-effects	μ_δ			
meta-analysis	**0.09 (0.04–0.15)**			
Random-effects	μ_δ	τ^2		
meta-analysis	**0.09 (0.01–0.19)**	0 (0–0.03)		
Mixed-effects	μ_δ	β	τ^2	
meta-analysis	**0.22 (0.07–0.39)**	−0.01 (−0.02, 0)	0 (0–0.02)	
Fixed-effects meta-	μ_δ			
analysis with power prior	**0.11 (0.04–0.17)**			
Random-effects meta-	μ_δ	τ^2		

(Continues)

TABLE 3 (Continued)

Meta-Analyses				
analysis with power prior	**0.11 (0.01–0.2)**	0 (0–0.02)		
Mixed-effects meta-	μ_δ	β	τ^2	
analysis with power prior	**0.27 (0.11–0.43)**	−0.01 (−0.02 to 0)	0 (0–0.01)	

Note: Posterior modes are provided. And 95% HPD intervals are inside the parentheses. The parameters of interest are highlighted as bold.

maximum posterior probability. The QBP interval and HPD interval are credible intervals, and they provide interval estimates within which an unknown parameter value falls with a certain probability (eg, 95%[26]). More specifically, the QBP interval is constructed by assuming equal probability in each tail, and the HPD interval is the narrowest interval that contains the values of highest posterior probability density based on the posterior samples. In addition, the posterior standard deviation of the posterior samples is calculated, which can be used to calculate Monte Carlo standard error. The R code for summarizing all the aforementioned statistics and conducting all the following analyses is provided in the Supporting Information (some representative code is presented in Appendix A). All results of integrative data analyses are summarized in Table 3.

Our goal is to explore the difference between husbands' and wives' marital satisfaction using data from 11 studies. First, assuming husbands and wives are independent, Bayesian integrative data analyses for independent groups are illustrated. Then, considering husbands and wives are paired as couples, Bayesian integrative data analyses for matched groups are illustrated. We also conduct the corresponding traditional frequentist analyses (see the Supporting Information for details), and since noninformative priors are used, the results from the Bayesian analyses and frequentist analyses are consistent. We illustrate the analyses with fixed-effects, random-effects, and mixed-effects models. Although we present all of the results, it is not very meaningful to compare the estimated coefficients across models because the meaning of coefficients varies in different models and they answer different research questions. Therefore, we do not anticipate that the results will be exactly the same across the models. Although researchers typically choose a model based on their assumptions or theories, another way to think about these models is to treat them as a sensitivity analysis. We conduct a sensitivity analysis with different models to investigate whether and how the estimate of the group mean difference varies under different assumptions.

4.1 | Testing mean differences for independent groups

As mentioned in the preceding section, the general process in practice is to begin with conducting a random-effects model. If the between-study variances are significant, we adopt the mixed-effects model to explore the influence of covariates; if the between-study variances are not significant, we adopt the fixed-effects model. One may directly use a fixed-effects model if there is a strong assumption for homogeneity. The illustration of different models in the following sections does not follow the general process; we

begin with the fixed-effects model because it is the simplest model.

4.1.1 | Fixed-effects integrative data analysis

For a Bayesian fixed-effects integrative data analysis that compares means from husbands and wives with equal variance, $y_{1i} \sim N(\mu_1, \sigma^2)$ (wives' scores) and $y_{2i} \sim N(\mu_2, \sigma^2)$ (husbands' scores), we employ mutually independent priors on μ_1 and μ_2 and the common variance σ^2:

$$f(\mu_1) \text{ or } f(\mu_2) \sim \mathrm{N}(0,\, a), \tag{6}$$

$$f(\sigma^2) \tilde{} \mathrm{Inv-Gamma}(b,\, b). \tag{7}$$

The specified priors are the semi-conjugate priors for normal likelihood, which are widely used in Bayesian modeling. The priors are called semi-conjugate because the posterior distribution of each of the parameters would be in the same family as the prior distribution given that all the other parameters are known.[29] When a is large (eg, 10 000) and b is small (eg, 0.001), these prior distributions are considered noninformative as long as the population variance, σ^2, is not very close to 0.[7] To answer our research question whether there is a group difference, a new parameter equal to the group mean difference, $\delta_{raw} = \mu_1 - \mu_2$, is created in order to make direct inferences from the posterior samples.

In real data, missing data are unavoidable due to many reasons. For ignorable missingness in the outcome variable, it is not necessary to specify a model and priors for the missing data. If the values are indicated as missing (i.e., NA), BUGS, and rjags will generate predictive missing data from the posterior predictive distributions.[30] Thus, we denote the missing data in y_{1i} and y_{2i} as NA.

The R code using rjags[25] to conduct the Bayesian fixed-effects integrative data analysis, construct the plots, and summarize the results is displayed in the Supporting Information. Two MCMC chains for the unknown model parameters (μ_1, μ_2, and σ^2) and the parameter that is created to answer our research question (δ_{raw}) are generated with different sets of starting values. Note that in rjags and BUGS, prior distributions are specified for precision parameters (the inverse of variance or the inverse of variance-covariance matrix) instead of for variance parameters. For each chain, the total number of MCMC iterations was 3000 after an adaptive period of 100, a burn-in period of 1000, and a thinning period of 5. Gelman and Rubin's test[27] indicates convergence for all parameters. Therefore, the posterior samples from the two chains are combined to calculate the posterior mean, posterior mode, posterior standard deviation, QBP interval, and HPD interval. In this example, we found the posterior modes and posterior means are similar, and the HPD intervals and QBP intervals are similar. The posterior modes and 95% HPD intervals are presented in Table 3. Since the HPD interval does not contain 0, we can conclude that wives have significantly higher marital satisfaction than husbands, $\hat{\delta}_{raw} = 1.08$ with an HPD interval of (0.08–1.81).

Without assuming equal variances for the two groups, $y_{1i} \sim N(\mu_1, \sigma_1^2)$ and $y_{2i} \sim N(\mu_2, \sigma_2^2)$, prior distributions need to be specified for σ_1^2 and σ_2^2 separately. We use the same prior as in Equation (7) for both σ_1^2 and σ_2^2. Similar to the results with equal variance, wives have significantly higher marital satisfaction than husbands, $\hat{\delta}_{raw} = 0.95$ with an HPD interval of (0.09–1.87).

4.1.2 | Random-effects integrative data analysis

In this paper, G_{ij}=0 for wives, and G_{ij}=1 for husbands. For the random-effects integrative data analysis model in Equation (1), the parameter indicating the average difference between husbands and wives across studies is β_{10}. We use an unstructured covariance structure for Σ_s and infer the between-study variance in the means of husbands (σ_{u0}^2), the between-study variance in the mean differences between husbands and wives (σ_{u1}^2), and the correlation $\left(\rho_{01} = \frac{\sigma_{u01}}{\sqrt{\sigma_{u0}^2 \sigma_{u1}^2}}\right)$. We specify normal priors for the fixed effects (β_{00} and β_{10}) as in Equation (6), an inverse-gamma prior for the level 1 residual variance (σ_ϵ^2) as in Equation (7), and an inverse-Wishart prior for the level 2 variance-covariance matrix (Σ_s), $f(\Sigma_s) \sim \mathrm{Inv-Wishart}(m, V)$. The specified priors are the widely used semi-conjugate priors for multilevel models with normal likelihood. a is set at 10^4; b is set at 0.001; m is set as the number of dimension of the level 2 covariance matrix Σ_s (2 in this example), and V is specified as an identity matrix. The HPD interval contains 0, $\widehat{\beta_{10}} = 0.87$ with HPD interval (−0.12 to 2.05).[7] The results indicate that after considering between-study heterogeneity, there is not enough evidence to reject the null hypothesis that husbands and wives have the same marital satisfaction.

4.1.3 | Mixed-effects integrative data analysis

We consider covariates at different levels in the mixed-effects integrative data analysis. To address the possibility of cohort effects on gender differences in marital satisfaction,

[7] In the discussed analyses, the confidence intervals of the fixed-effects models are slightly narrower than those from the random-effects models because when the random-effects model is true, fixed-effects models would tend to underestimate standard errors of fixed effects and yield narrower confidence intervals. From a practical perspective, it is unlikely that real-life studies are homogeneous (i.e., fixed-effect models) because it is almost impossible that the between-study variance would be exactly 0.

we treat the year that each study was initiated as a study-level covariate. To address the possibility that gender differences are associated with social class or differences in class between spouses, we treat each spouse's years of education as an individual-level covariate. To better interpret the coefficients, the year of survey is centered at the earliest year (1991), and the years of education is centered at the minimum value (8) in all of the following analyses. We also can center the year of survey and the years of education at different values, which will change the meaning of coefficients and answer different research questions. When there are missing data in covariates, the simplest solution is listwise deletion by deleting the individuals who have missing covariates, but this requires ignorable missingness; otherwise, it will cause biased estimation. Another solution in Bayesian modeling is to set missing covariates as NA, and because covariates are exogenous variables, we need to (a) specify a model to capture why the covariates are missing and how the covariates can be generated, (b) specify priors for the parameters in the missingness model, and (c) estimate this missingness model. In this example, we use listwise deletion. The generalization to the latter solution is straightforward by adding the needed missingness model based on the assumption and adding the corresponding priors (for more details, refer to Lunn et al[30] and Enders et al[31]).

First, we consider Model (2) where years of education are assumed to have the same effect on husbands and wives. The fixed effects (β_{00} to β_{21}) are specified to have normal priors, the level 2 variance-covariance matrix (Σ_s) has an inverse Wishart prior, and the level 1 residual variance (σ_ϵ^2) has an inverse-gamma prior. The parameter of interest is β_{10}, the group mean difference between husbands and wives conditional on the specific levels of the covariates. We found $\widehat{\beta_{10}} = 2.10$ with an HPD interval of (0.17–4.10), indicating that wives have significantly higher marital satisfaction than husbands after controlling for the years of education and the year of survey. Although this conclusion is similar to the conclusions of the fixed-effects integrative data analyses, the difference between husbands' and wives' marital satisfaction when taking the years of education, the year of survey, and between study heterogeneity into account (i.e., 2.10) is larger than the ones in fixed-effects model (i.e., 1.08 and 0.95), but as mentioned above, it is not surprising that the results are not the same. More specifically, the marital satisfaction difference is larger when we are conditioning on the couples with 8 years of education and taking survey in 1991, compared with averaging over the years of education and the year of survey in the random-effects model. We can also center the year of survey at the latest year (2016) and center the years of education at the maximum value (24). Then, β_{10} becomes the group mean difference between husbands and wives when we are conditioning on the couples with 24 years of education and taking survey in 2016. And $\widehat{\beta_{10}} = -0.11$ with an HPD interval of (−1.67 to 1.61), which is not significant.

Second, we allow the effect of years of education to differ between husbands and wives. We specify the same priors as in Model (2). The group mean difference between husbands and wives after controlling for the covariates, β_{10}, is estimated to be −1.68 with an HPD interval of (−7.62 to 5.64). Thus, after considering the heterogeneous effect of the years of education by adding more interaction terms, wives and husbands no longer differ in marital satisfaction. The wider credible interval of marital satisfaction difference is probably due to the collinearity issue caused by the interaction, $G_{ij}X_{ij}$; therefore, the estimation is less efficient because of the large standard error. Instead of centering the covariates at their minimal values, we can consider centering the covariates at their means to solve the collinearity issue to some degree. We first center the years of education at its grand mean (16.07) and center the year of survey at its mean (2005). The group mean difference between husbands and wives when their year of education is 16.07 and the year of survey is 2005 is estimated to be 0.96 with an HPD interval of (−0.10 to 1.98). Then, we center the years of education at its study-level mean and center the year of survey at its mean. The group mean difference between husbands and wives when their years of education is the mean of years of education in each study and the year of survey is 2005 is estimated to be 0.92 with an HPD interval of (−0.08 to 1.97).

In summary, depending on whether allowing between-study heterogeneity, whether considering the effects of covariates, how to center the covariates, and how to specify the influence of covariates, the inference of difference between husbands' and wives' marital satisfaction varies. This implies that researchers need to choose the suitable model based on their research interests and assumptions because the relatively widely used fixed-effects integrative data analysis[13] may provide misleading results.

4.2 | Testing mean differences for matched groups

4.2.1 | Fixed-effects integrative data analysis

Considering that husbands and wives are from the same families, paired difference data are computed as the wives' scores minus their husbands' scores within couples (y_d). The results are presented in Table 3. $\hat{\mu}_{raw.d}$ is 1.03 with an HPD interval of (0.44–1.67), which indicates that wives have significantly higher marital satisfaction than their husbands.

4.2.2 | Fixed-effects integrative data analysis with between-pair variance

We use Model (3) to estimate the between-pair variance and assume that there is no between-study heterogeneity. We consider years of education as an individual-level covariate. Since there is no couple-level covariate in the current real data, only β_{00}, β_{10}, β_{20}, and β_{30} are estimated. The parameter that indicates the overall difference between husbands and wives across studies and pairs is β_{10}. $\widehat{\beta_{10}}$ is −0.31 with an HPD interval of (−2.46 to 2.32). Therefore, we fail to find significantly different marital satisfaction between husbands and wives when conditioning on the year of survey in 1991, which is different from the result without considering between-pair variance.

4.2.3 | Random-effects integrative data analysis

For the random-effects integrative data analysis model in Equation (4), the parameter that indicates the overall difference between husbands and wives across studies is β_{00}. $\widehat{\beta_{00}}$ is 1.07 with an HPD interval of (0.21–1.81), which indicates that wives have significantly higher marital satisfaction than their husbands when considering between-study heterogeneity.

4.2.4 | Mixed-effects integrative data analysis

There is no couple-level covariate in the data, and we consider year that each study was initiated as a study-level covariate. $\widehat{\beta_{00}}$is 2.23 with an HPD interval of (0.82–3.54). That is, wives have significantly higher marital satisfaction than their husbands when conditioning on the year of survey in 1991, which is similar to the conclusions from the fixed- and random-effects analyses.

In summary, similar to the independent group case, the inference of the difference in marital satisfaction between husbands and wives varies as function of whether allowing between study heterogeneity, whether allowing between pair heterogeneity, and whether considering the effects of covariates.

5 | META-ANALYSIS MODELS

Sometimes the original raw data are inaccessible, but instead, the aggregated data such as effect sizes are presented in the literature. In this case, meta-analyses can be applied to gather information from effect sizes. There has been considerable discussion comparing integrative data analysis and meta-analysis.[13,32-35] Integrative data analysis requires more efforts in contacting authors, collecting data, and cleaning data than meta-analysis, making it more labor-intensive, time-consuming, and costly. But on the other hand, Cooper and Patall[13] and Lambert et al[36] found that meta-analysis has smaller power compared with integrative data analysis. Similar to the integrative data analyses, depending on whether between-study heterogeneity and covariates are considered, there are fixed-effects, random-effects, and mixed-effects models. We will present these three models when comparing means from independent groups and when comparing means from matched groups, separately. The null hypothesis is that the overall population effect size is 0, $H_0 : \mu_\delta = 0$.

5.1 | Testing mean differences for independent groups

For studies that test mean differences for two independent groups, a widely used effect size is the standardized mean difference. For study j, the observed effect size is $g_j = \frac{\bar{y}_{1j} - \bar{y}_{2j}}{s_j}$, where $\bar{y}_{1j}$ and $\bar{y}_{2j}$ are the sample means of the two groups respectively, s_j is the sample standard deviation calculated by $s_j = \sqrt{\frac{(n_{1j}-1)s_{1j}^2 + (n_{2j}-1)s_{2j}^2}{n_{1j}+n_{2j}-2}}$, s_{1j}^2 and s_{2j}^2 are the sample group variances of two groups respectively, and n_{1j} and n_{2j} are the sample sizes of two groups, respectively. And the observed effect size g_j needs to be corrected to obtain an unbiased estimate of the true effect size δ_j, which is $d_i = \left(1 - \frac{3}{4(n_{1j}+n_{2j})-9}\right) g_i$.[19]

5.1.1 | Fixed- and random-effects meta-analysis

Suppose there are J unbiased effect size estimates d_j (j=1,…, J). A typical random-effects meta-analysis model[19,37] is as follows:

$$\begin{aligned} d_j &= \delta_j + e_j \\ \delta_j &= \mu_\delta + u_j \\ e_j &\sim N\left(0, \sigma_j^2\right) \\ u_j &\sim N(0, \tau^2) \end{aligned}, \tag{8}$$

where e_j is the deviation of the observed study effect size d_i from the true/population study effect size δ_j, and u_j is the deviation of the true study effect size δ_j from the overall population effect size μ_δ. The variance of e_j represents within-study sampling variability of study j, $\sigma_j^2 = \frac{n_{1j}+n_{2j}}{n_{1j}n_{2j}} + \frac{\delta_j^2}{2(n_{1j}+n_{2j})}$. The normal distribution of e_j is based on a large sample size approximation, and the normality approximation is good when the true effect size is small and

the per-study sample size is relatively large.[19,38] The variance of u_j and τ^2 represents between-study heterogeneity. When τ^2 is 0, the model in Equation (8) becomes a fixed-effects model, meaning that the studies are homogeneous with the same true effect size $\delta_i=\mu$.[19]

5.1.2 | Mixed-effects meta-analysis

A mixed-effects meta-analysis model (also called meta-regression model) with a study-level covariate Z_j[21]) is as follows:

$$\begin{aligned} d_j &= \delta_j + e_j \\ \delta_j &= \mu_\delta + \beta Z_j + u_j \\ e_j &\sim N\left(0, \sigma_j^2\right) \\ u_j &\sim N(0, \tau^2) \end{aligned}, \quad (9)$$

where β is the regression coefficient for Z_j. Centering Z_j and different ways of centering will change the meaning of coefficients and answer different research questions.

5.2 | Testing mean differences for matched groups

For studies that test mean differences for two matched groups, the standardized matched mean difference in study j is computed as $g_j=\frac{\bar{y}_{dj}}{s_{dj}}\sqrt{2(1-r_i)}$, where $\bar{y}_{dj}$ is the sample mean of paired differences, s_{dj} is the sample standard deviation of paired difference data, and r_j is the sample correlation within pairs in study j.[39] The corrected unbiased estimate of the true effect size δ_j is $d_j=\left(1-\frac{3}{4n_j-9}\right)g_i$.[19]

5.2.1 | Fixed-, random-, and mixed-effects meta-analyses

A random-effects meta-analysis model is the same as Equation (8) except that the within study sampling variance[40] is $\sigma_j^2=\frac{2(1-\rho_j)}{n_j}+\frac{\delta_j^2}{2n_j}$, where ρ_j is the population correlation within pairs. In practice, we can substitute d_j and r_j for δ_j and ρ_j, respectively. Similar to the independent groups case, a mixed-effects meta-analysis model can be applied to incorporate study-level covariates.

Comparing meta-analysis that is based on aggregated data with integrative data analysis that is based on original raw data, there are three noticeable differences. First, individual-level covariates cannot be considered in meta-analysis. When the observed effect size (i.e., aggregated data) is calculated in meta-analysis, the original information for individuals is abandoned and cannot enter the later steps. Second, meta-analysis does not require that the measurements that are used for outcomes are the same across studies or can be equated. Although different measurements have different reliabilities, which lead to different levels of attenuation of the true effect sizes, in meta-analysis each effect size can be individually corrected for the attenuation based on the measurement, and thus, the effect of measurement error can be eliminated (for more details, refer to Schmidt and Hunter[41]). Or in the Bayesian framework, a power prior can adjust the estimation of the population effect size by down weighting the sample effect sizes with low-scale reliabilities.[42] Third, meta-analysis models are based on a large sample size normality approximation. That is, the analytical form of the normal distribution of residuals is only good when the population effect size is small and per-study sample size is relatively large.[19,38] Different from meta-analysis, the distributions of residuals in the models of integrative data analysis are assumed to be normal, and the variances will be freely estimated.

5.3 | Power prior

A power prior is proposed to control the contribution of the information from data, which is accomplished by giving different studies different weights. In particular, weight is given based on the study characteristics that are not the same across studies, such as study quality indicators. Mathematically, we raise the likelihood of data to a power a (i.e., $L(\theta|D)^a$)[43-46] The power prior enables the previously entered data to generate a data-dependent prior that is weighted based on the power. The general form of power prior is given by the following:

$$f(\theta|D) \propto L(\theta|D)^a f(\theta), \quad (10)$$

where D indicates a specific historical data set, θ indicates the parameters of interest, and $a \in [0,1]$ controls the importance of historical data or researchers' confidence about the quality of the data. If a=1, Equation (10) is the traditional Bayes' theorem, and all the information from the historical data contributes to the posterior distribution. If a=0, $L(\theta|D)^0=1$, and the information from the historical data is not used at all. Thus, the power prior is a posterior distribution based on the historical data by controlling the information of the data.

In this paper, we use the power prior focusing on aggregated data. In a meta-analysis, there are multiple historical data sets (i.e., estimated study effect sizes). For each data set, we would like to control its contribution based on its study quality. In fitting a normal distribution, raising the likelihood for study j, $L\left(\theta|D_j\right)=N\left(\theta,\sigma_j^2\right)$, to a power a_j is almost equivalent to scaling the likelihood to

$N\left(\theta, \frac{\sigma_j^2}{a_j}\right)$.[30,42,43] This conclusion can be applied to standardized mean differences and standardized matched mean differences. It means that we modify Equation (8) to be $e_i \sim N\left(0, \frac{\sigma_j^2}{a_j}\right)$ and $u_i \sim N(0, \frac{\tau^2}{a_j})$.[8] How to specify power coefficient to control study quality is flexible. For example, to down weight unreliable studies, smaller power coefficients are specified for less reliable studies.[42] Because the specification of power coefficients has many options (eg, based on different variables and different weights), we echo the suggestion by Ibrahim and Chen[43] that power prior can be considered as a method for a sensitivity analysis rather than using only one set of power prior.

6 | BAYESIAN META-ANALYSIS WITH OUR REAL DATA

The R code for conducting all of the following analyses is provided in the Supporting Information (some representative code is presented in Appendix A), and all of the results are summarized in Table 3. In addition to the Bayesian meta-analyses, we conducted traditional frequentist analyses (see Supporting Information for details), and found that the results from the frequentist analyses were consistent with the ones in the Bayesian framework with noninformative priors.

6.1 | Testing mean differences for independent groups

6.1.1 | Fixed-effects meta-analysis

The between-study heterogeneity τ^2 is set at 0 in the fixed-effects meta-analysis by assuming that the 11 studies have the same true effect size. We specify a normal prior for the overall true effect size μ_δ as in Equation (6). μ_δ is estimated in Bayesian modeling at 0.10 with an HPD interval of (0.02–0.18) (see Table 3). Therefore, based on the fixed-effects meta-analysis, wives have significantly higher marital satisfaction than husbands. Compared with the result obtained in integrative data analysis ($\hat{\delta} = 1.08$), the estimated overall true effect size seems smaller (i.e., $\widehat{\mu_\delta} = 0.10$), but the former is the raw group difference and the latter is the standardized group difference.

6.1.2 | Random-effects meta-analysis

A random-effects meta-analysis is shown in Equation (8). The between-study heterogeneity τ^2 is freely estimated in the random-effects meta-analysis while assuming that the 11 studies have different true effect sizes. In addition to the normal prior for the overall true effect size μ_δ, an inverse-gamma prior is specified for the between-study heterogeneity τ^2. μ_δ is estimated at 0.11 with an HPD interval of (0.01–0.20) (see Table 3). Therefore, the overall true effect size is significantly different from 0 (i.e., wives are more satisfied with their marriage than husbands), which is the same conclusion from the fixed-effect meta-analysis since τ^2 is estimated to be almost 0.

6.1.3 | Mixed-effects meta-analysis

In a mixed-effects meta-analysis model in Equation (9), we control the year of survey as a study-level covariate and center it at the earliest year, 1991. A normal prior is specified for β, the effect of the year of survey. μ_δ is estimated at 0.23 with an HPD interval of (0.04–0.41), which indicates that the overall true effect size is still significantly different from 0 (i.e., wives are more satisfied with their marriage than husbands) when we are conditioning on the year of conducting the studies in 1991.

6.1.4 | Meta-analysis with power prior

We use a mixed-effect meta-analysis to illustrate how to incorporate power priors. This procedure can easily be generalized to fixed- and random-effects meta-analyses. The year of survey is the study-level covariate in the illustrated example. We considered two factors in determining power *a.* The first is the compensation of participants and the second is whether the study was supported by an extramural funding (listed in Table 2). When the compensation is higher, we can predict that participants will devote more time to respond to questionnaires carefully. When a study is funded by an extramural grant, we can predict that the funding may support purchasing resources that improve a study's quality. Thus, for a funded study (1 for funded

[8]The conditional posterior distribution of the between-study variance (τ^2) that is calculated by scaling the likelihood is slightly different from the one directly calculated by raising the likelihood to the specific power, based on our derivation. In a random-effects meta-analysis with priors Inv −Gamma(α,β) for τ^2 and $N(0,\varphi_0)$ for μ_δ, by scaling the likelihood to $N\left(\theta, \frac{\sigma_j^2}{a_j}\right)$, the conditional posterior distribution of τ^2 is $f(\tau^2|\cdot) = \mathrm{IG}\left(\alpha + \frac{J}{2}, \beta + \frac{\sum_j^J a_j(\delta_j - \mu_\delta)^2}{2}\right)$. With directly raising the likelihood to the power of a_j, $f(\tau^2|\cdot) = \mathrm{IG}\left(\alpha + \frac{\sum_j^J a_j}{2}, \beta + \frac{\sum_j^J a_j(\delta_j - \mu_\delta)^2}{2}\right)$. The conditional posterior distributions of the overall true effect size μ_δ and the study-specific true effect size δ_j are the same between the two approaches.

studies and 0 for the ones without extramural funding) with higher compensation, the power is closer to 1. Specifically, a=*fund*×0.5+*compensation*×0.005, and the compensation is adjusted based on the Consumer Price Index (CPI-U) that is provided by the US Department of Labor Bureau of Labor Statistics. For the 11 studies, the power values are 0.37, 0.97, 0.44, 1.12, 0.52, 0.5, 0.43, 0.39, 0.98, 1.02, and 0.75, respectively. The choice recommended here can be used as a starting point based on which sensitivity analyses can be conducted.

The R code using rjags is presented in the Supporting Information. Since the code is almost the same as the one for the illustrated standard Bayesian mixed-effects meta-analysis, only the model specification section is presented. After controlling the information of each study, the overall true effect size μ_δ is estimated to be 0.25, which is slightly larger than the estimated μ_δ in the standard Bayesian mixed-effects meta-analysis (see Table 3). And the HPD interval, (0.05–0.46), does not contain 0. Thus, after controlling the year of study and the information for each study based on study quality, we still find that wives are more satisfied with their marriage than husbands.

We also present the results of the fixed- and random-effects meta-analyses with power priors in Table 3. Compared with the results without power priors, we still reach the same conclusion in the fixed-effects and random-effects meta-analyses that wives have significantly higher marital satisfaction than husbands.

6.2 | Testing mean differences for matched groups

6.2.1 | Fixed-effects meta-analysis

A normal prior is specified for the overall true effect size μ_δ, and μ_δ is estimated at 0.09 with an HPD interval of (0.04–0.15). Thus, by considering that husbands and wives are matched in families, we found that wives are more satisfied with their marriage than their husbands.

6.2.2 | Random-effects meta-analysis

Priors are specified as in the aforementioned random-effects meta-analysis of independent groups. μ_δ is estimated at 0.09 with an HPD interval of (0.01–0.19). Therefore, after considering that husbands and wives are matched in families and there is between-study heterogeneity, wives have significantly higher marital satisfaction than their husbands, which is the same as the conclusion from fixed-effect meta-analysis since τ^2 is estimated to be almost 0.

6.2.3 | Mixed-effects meta-analysis

We use the year of survey as a study-level covariate and center it at the earliest year, 1991. Priors are specified as in the independent groups case. μ_δ is estimated at 0.22 with an HPD interval of (0.07–0.39); thus, wives still have significantly higher marital satisfaction than their husbands when we are conditional on the year of survey in 1991. We can also use power prior with compensation and extramural funding as criteria for study quality (see Table 3 for the results).

In summary, different meta-analyses lead to similar inferences in the current example. Thus, the results are relatively robust to the assumed models. One reason is that the estimated between study heterogeneity is almost 0.

7 | DATA FUSION USING AUGMENTED DATA-DEPENDENT PRIORS

We can conduct the integrative data analysis and meta-analysis with both frequentist and Bayesian modeling. We have used our real data to illustrate how to use Bayesian modeling to conduct the aforementioned analyses. Now, we move to methods that are completely within the Bayesian framework. As a starting point, we briefly summarize Bayesian integrative data analysis and Bayesian meta-analysis. They both begin with noninformative priors and adopt the likelihood from either the whole raw data (i.e., integrative data analysis) or the whole aggregated data (i.e., meta-analysis). Overall, these two methods can be presented as follows:

$$f(\theta) \underset{L(\theta|D_1,D_2,\ldots,D_J)}{\rightarrow} f(\theta|D_1,D_2,\ldots,D_J), \qquad (11)$$

where θ is a set of unknown parameters such as regression coefficients or overall population effect size, D_1,D_2, …,D_J can be J sets of raw data or J observed effect sizes (i.e., aggregated data), $f(\theta)$ represents the prior information of θ, $L(\theta|D_1,D_2,\ldots,D_J)$ represents the likelihood from J studies, and $f(\theta|D_1,D_2,\ldots,D_J)$ represents the posterior distribution of θ. In Equation (11), the prior is only used once, and the likelihood of all data sets enters the model simultaneously. As an alternative approach, we can allow the information of each data set to enter the model sequentially. Marcoulides[3] refers to this approach as data fusion using AUDPs. In this way, it is intuitive to monitor the influence of each study. That is, the contribution of each study can be clearly summarized and presented in terms of parameter estimation and statistical power. For example, we order the study based on the year of survey, and we want to know how the conclusion will change when a subsequently conducted study provides more

information. When the first study enters the analysis, posterior distributions are first derived with noninformative priors, and then, the obtained posterior distributions serve as the prior distributions for the parameters in the second data set. All the data sets enter the model in turn, and the posterior distributions at the final step are based on all of the data sets. The algorithm is presented as follows:

$$f(\theta) \underset{L(\theta|D_1)}{\rightarrow} f(\theta|D_1) \underset{L(\theta|D_2)}{\rightarrow} f(\theta|D_1,D_2)\ldots \underset{L(\theta|D_J)}{\rightarrow} f(\theta|D_1,D_2,\ldots,D_J) \quad . \tag{12}$$

Besides using a noninformative prior $f(\theta)$ to initialize AUDP, AUDP can start with an informative prior. The informative prior is constructed based on one of the available studies (an example is presented in the Supporting Information). For example, the point estimates and standard error estimates for parameters in the first study could construct the initial priors for the unknown parameters in AUDP. Thus, the algorithm is as follows:

$$f(\theta|D_1) \underset{L(\theta|D_2)}{\rightarrow} f(\theta|D_1,D_2)\ldots \underset{L(\theta|D_J)}{\rightarrow} f(\theta|D_1,D_2,\ldots,D_J) \quad . \tag{13}$$

For both AUDP starting with a noninformative prior and starting with an informative prior, at the last step, the Algorithms (12) and (13) are equivalent to the Algorithm (11) with noninformative prior. That is, the posterior distribution is $f(\theta|D_1,D_2,\ldots,D_J)$. Thus, theoretically, the order of the data sets in entering the analysis in AUDP does not influence the final outcome because if we analytically calculate the posterior mean and variance based on $f(\theta|D_1,D_2,\ldots,D_J)$, the posterior mean and variance stay the same regardless of the order. However, we also need to note that posterior distributions have stochastic property. Without analytically deriving the posterior mean and variance, calculating the posterior mean and variance by finite posterior samples will not lead to exactly the same results even when the order stays the same, although the variation could be so tiny as to be ignorable. When the number of posterior samples goes to infinity, the posterior mean and variance are consistent with the analytical results, and the order has no influence. With finite posterior samples, the cumulative order effect of repeatedly calculating the posterior mean and variance may exist, especially when per-study sample size is not large. We will explore this with our real data.

The AUDP approach can be applied to both raw data and aggregated data. The fixed-effects, random-effects, and mixed-effects models (eg, Equations (1), (2), (8), and (5)) in the integrative data analysis and meta-analysis, which have been introduced with details, can be analyzed by the AUDP approach. When the AUDP approach is applied to aggregated data (i.e., effect size), it is a Bayesian cumulative meta-analysis.[9]

The AUDP approach can also be coupled with power priors. In this way, it is easy to control the contribution of each study and view how the results change when each study enters the analysis. A power prior $f(\theta|D_1)$ is derived based on a noninformative prior of θ and the first entered study by controlling the power of the likelihood, $f(\theta|D_1) \propto f(\theta)L(\theta|D_1)^a$. Using a power prior $f(\theta|D_1)$, the posterior distribution of θ is updated based on the second entered data D_2, $f(\theta|D_1,D_2) \propto f(\theta|D_1)L(\theta|D_2)^{a_2}$. Overall, the algorithm is presented as follows:

$$f(\theta) \underset{L(\theta|D_1)^{a_1}}{\rightarrow} f(\theta|D_1) \underset{L(\theta|D_2)^{a_2}}{\rightarrow} f(\theta|D_1,D_2)\ldots \underset{L(\theta|D_J)^{a_J}}{\rightarrow} f(\theta|D_1,D_2,\ldots,D_J) \quad .$$

7.1 | AUDP using power priors with our real data

We use the aggregated data (observed effect sizes) with independent groups as an example. In this example, AUDP is applied to the 11 estimated standardized mean differences for a fixed-effects model where the overall true effect size μ_δ is unknown. Random- and mixed-effects models can also be analyzed in AUDP. However, when the first few studies enter the analyses, the models are not identified with a noninformative prior on τ^2 because there are not enough studies to estimate the between-study variance τ^2. Therefore, with between-study variance, a very informative prior should be specified for τ^2 to initialize the AUDP process. In addition, we consider the compensation of participants and extramural funding as criteria to determine the power of the likelihoods as for the aforementioned meta-analyses with power priors. The posterior mean and posterior variance of μ_δ are used to

[9]Although in the frequentist framework, we also can allow each data set to enter the analysis sequentially, such as frequentist cumulative meta-analysis[47] and the frequentist cumulative analysis should provide similar estimation to AUDP at the last step, we do not refer to this process as frequentist AUDP. In AUDP, our knowledge of parameters keeps updating when studies enter the analysis. Accordingly, the posterior distributions of the parameters or the true effect size, which represent the updated knowledge, will be the data-dependent priors in the next step. On the other hand, in frequentist cumulative meta-analysis, although the inferences of parameters are updated after each study enters the analysis, such updated knowledge (i.e., the estimated overall effect size) will not be used in the next step. Instead, the next step uses all the raw data from the data sets that have entered the analysis. Thus, frequentist cumulative meta-analysis is equivalent to the Bayesian process where a Bayesian meta-analysis is conducted with a noninformative prior when a new data set is merged with the previously entered data sets.

summarize the posterior samples of $f(\mu_\delta|D_1)$ and get recorded. A normal power prior for D_2 is specified with a mean equal to the posterior mean of $f(\mu_\delta|D_1)$ and a variance equal to the posterior variance of $f(\mu_\delta|D_1)$. In the same way, we can construct the power prior for D_3 to D_J and obtain posterior distributions of $f(\mu_\delta|D_1,D_2,D_3)$to $f(\mu_\delta|D_1,D_2,\ldots,D_J)$. The code for the AUDP approach with power priors is presented in the Supporting Information.

For the 11 studies, there are numerous ways of ordering the studies. Even though we expect that the order of the data sets in entering the analysis will not significantly impact the final results, we find different orders could provide slightly different final outcomes but completely different decisions of rejecting the null hypothesis. As mentioned above, the difference is due to the fact that the posterior distribution is a random process; thus, the inferences that are calculated empirically with a finite sample size are also random. We illustrate results from three sets of order in Table 4, with the first order based on the year of survey. Although the hypothesis testing conclusions based on the different orders are

TABLE 4 Results of the fixed-effects model from the AUDP approach with power prior

	Study	Posterior Mean	Posterior Mode	Posterior SD	QBP Interval	HPD Interval
1	3	−0.08	−0.05	0.29	−0.65 to 0.5	−0.63 to 0.51
	4	0.24	0.24	0.1	0.05–0.43	0.05–0.43
	1	0.23	0.23	0.09	0.05–0.41	0.05–0.41
	2	0.22	0.22	0.07	0.09–0.36	0.09–0.36
	8	0.22	0.22	0.07	0.09–0.36	0.09–0.35
	7	0.22	0.22	0.07	0.09–0.36	0.09–0.36
	9	0.17	0.17	0.06	0.05–0.3	0.05–0.3
	5	0.16	0.16	0.06	0.05–0.28	0.04–0.27
	10	0.14	0.14	0.05	0.03–0.25	0.04–0.25
	6	0.12	0.14	0.05	0.03–0.22	0.03–0.21
	11	0.11	0.11	0.05	0.02–0.2	0.02–0.2
2	3	−0.08	−0.05	0.29	−0.65 to 0.5	−0.63 to 0.51
	5	0.02	0	0.16	−0.28 to 0.33	−0.27 to 0.33
	8	0.07	0.08	0.14	−0.2 to 0.34	−0.19 to 0.35
	7	0.12	0.12	0.13	−0.13 to 0.36	−0.14 to 0.35
	2	0.17	0.17	0.08	0.01–0.34	0.01–0.33
	11	0.13	0.13	0.07	0–0.27	0–0.27
	10	0.11	0.12	0.06	−0.01 to 0.23	−0.01 to 0.23
	1	0.11	0.11	0.06	0–0.23	−0.01 to 0.22
	6	0.09	0.09	0.06	−0.03 to 0.2	−0.03 to 0.2
	4	0.14	0.15	0.05	0.03–0.24	0.03–0.23
	9	0.12	0.13	0.05	0.03–0.21	0.04–0.22
3	4	0.28	0.25	0.1	0.08–0.49	0.07–0.47
	9	0.18	0.19	0.08	0.02–0.33	0.02–0.33
	8	0.18	0.18	0.08	0.04–0.34	0.03–0.33
	3	0.16	0.16	0.08	0.01–0.31	0.01–0.32
	10	0.13	0.13	0.07	−0.01 to 0.27	−0.01 to 0.26
	2	0.15	0.15	0.06	0.04–0.27	0.03–0.26
	1	0.15	0.16	0.06	0.03–0.27	0.04–0.27
	7	0.16	0.15	0.06	0.04–0.27	0.05–0.27
	11	0.13	0.12	0.06	0.03–0.25	0.03–0.25
	5	0.12	0.12	0.06	0.01–0.24	0.01–0.24
	6	0.1	0.1	0.06	−0.01 to 0.21	−0.01 to 0.22

Abbreviations: AUDP, augmented data-dependent prior; HPD, highest posterior density; QBP, quantile-based probability.

different, the values of the posterior summary statistics at the last step from different orders are, as expected, almost the same. Because the posterior mean and mode of the parameter of interest are close to 0, with different orders, sometimes the lower bound of the credible interval is slightly smaller than 0 and sometimes it is slightly larger than 0. Consequently, the hypothesis testing results are different.[10] Table 4 illustrates how each study contributes to the final conclusion. Take the order of time as an example. The study in 1991 (the first entered study) has an observed effect size of −0.09; therefore, the posterior mean and mode based on only this study are negative. The study in 1993 (the second entered study) has an observed effect size of 0.06 and a sample size of 343, which increases the posterior mean and mode to be positive. The study in 1998 (the third entered study) has an observed effect size of 0.14, which maintains the posterior mean and mode to be about 0.23. With adding information from more later studies, the posterior point estimation (i.e., posterior mean and posterior mode) becomes more stable, and the newly entered negative effect size only influences the overall estimation to a small degree. The study in 2016 (the 10th entered study) has an observed effect size of −0.15 and a sample size of 237, but it barely influences the posterior estimation. In addition, with more studies, the posterior standard deviation becomes smaller, and the QBP and HPD intervals generally become narrower, regardless of the order of the studies.

8 | DATA FUSION USING AGGREGATED DATA-DEPENDENT PRIORS

The fourth Bayesian synthesis approach is intuitively appealing. It chooses one study to serve as the formal study and to provide likelihood, and all of the other studies are used to construct a data-dependent prior. We call this approach data fusion using AGDPs. After deciding the form of the prior distribution of the parameter θ (i.e., a normal distribution or a beta distribution), the estimated θs in the data sets are assumed to be a random sample from the prior. Thus, the most intuitive way to construct the prior is to summarize the estimated θs. Then, the hyperparameters in the prior distribution can be calculated based on these estimated θs. Using the aforementioned models for raw data or aggregated data, we can estimate the parameters. The algorithm is as follows:

$$f(\theta|D_1,D_2,\ldots,D_{J-1}) \underset{L(\theta|D_J)}{\rightarrow} f(\theta|D_1,D_2,\ldots,D_J).$$

For example, when the prior $f(\theta)$ is a normal distribution and there are several $\hat{\theta}$s in the literature, $f(\theta)$ can be specified as $N\left(mean\left(\hat{\theta}\right), var\left(\hat{\theta}\right)\right)$. When the prior $f(\theta)$ is an inverse-gamma distribution (Inv−Gamma(shape=α,scale=β)), the hyperparameters can be computed based on $mean\left(\hat{\theta}\right)$and $var\left(\hat{\theta}\right)$. More specifically, $E(\theta)=\frac{\beta}{\alpha-1}$ is specified at $mean\left(\hat{\theta}\right)$, and $var(\theta)=\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$ is specified at $var\left(\hat{\theta}\right)$; thus, $\alpha=\frac{mean\left(\hat{\theta}\right)^2}{var\left(\hat{\theta}\right)}+2$and $\beta=\left(\frac{mean\left(\hat{\theta}\right)^2}{var\left(\hat{\theta}\right)}+1\right)\times mean\left(\hat{\theta}\right)$.

8.1 | AGDP using our real data

We use a t test assuming unequal variances as an example. We select one study as the formal study, which provides likelihood. For the other 10 studies, the sample mean ($\bar{y}_{1j}$ and $\bar{y}_{2j}$) and sample variance (s_{1j}^2 and s_{2j}^2) of each group are computed within each study. Then, we calculate $mean\left(\bar{y}_{1j}\right)$, $mean\left(\bar{y}_{2j}\right)$, $mean\left(s_{1j}^2\right)$, $mean\left(s_{2j}^2\right)$, $var\left(\bar{y}_{1j}\right)$, $var\left(\bar{y}_{2j}\right)$, $var\left(s_{1j}^2\right)$, and $var\left(s_{2j}^2\right)$ across studies. To specify normal priors for μ_1 and μ_2, and inverse-gamma priors for σ_1^2 and σ_2^2, the hyperparameters are computed as described in the previous section based on $mean\left(\bar{y}_{1j}\right)$, $mean\left(\bar{y}_{2j}\right)$, $mean\left(s_{1j}^2\right)$, $mean\left(s_{2j}^2\right)$, $var\left(\bar{y}_{1j}\right)$, $var\left(\bar{y}_{2j}\right)$, $var\left(s_{1j}^2\right)$, and $var\left(s_{2j}^2\right)$. Besides the setting of the hyperparameters, the code is the same as for the Bayesian integrative analysis.

There are 11 choices for the formal study. Depending on which studies are used to construct the priors and which study serves as the formal study, the results vary noticeably. In some cases, there is a significant gender difference as in the Bayesian integrative data analysis, while in the other cases, the effect is nonsignificant. We present the inference of group mean difference $\delta=\mu_1-\mu_2$ from all 11 choices in Table 5. Furthermore, the credible intervals in AGDP are wider than those in the t test assuming unequal variance in the integrative analysis, which means that precision is smaller in AGDP. The main cause of the diverse results with different formal studies and wide credible intervals in AGDP is that the sample size information of the studies that are used to construct the priors is not considered. Although the sample mean and sample variance are sufficient statistics for normal distributions, they do not carry the information of sample size. Thus, when constructing the priors, we treat the sample mean and sample variance from a small sample size (eg, 2) and those from a large sample size (eg, 1000) equally. The prior distributions can be regarded as from 10 studies, each of which only has a sample size of 2, though in reality the sample size of each study is much larger than 2. As a consequence, the prior information from the 10 studies is relatively little compared with the information in the formal study since the sample size is considered in the formal study when providing likelihood. Therefore, the

TABLE 5 Inference of group mean difference $\delta=\mu_1-\mu_2$ in the t test assuming unequal variances from the AGDP approach

Formal Study	d_j of the Formal Study	$n_{1j}+n_{2j}$ of the Formal Study	Posterior Mean	Posterior Mode	Posterior SD	QBP Interval	HPD Interval
1	−0.09	114	−0.73	−0.57	2.04	−4.75 to 3.32	−4.72 to 3.33
2	0.28	343	2.44	2.56	0.97	0.51–4.28	0.51–4.28
3	0.14	163	1.29	1.61	1.47	−1.6 to 4.15	−1.58 to 4.15
4	0.21	338	2.08	2.21	1.07	−0.03 to 4.15	0.02–4.19
5	0.22	142	2.23	1.92	1.77	−1.22 to 5.76	−1.37 to 5.57
6	0.29	122	2.7	2.79	1.71	−0.68 to 6.05	−0.69 to 6.02
7	0.02	268	0.33	0.5	1.03	−1.7 to 2.32	−1.67 to 2.34
8	0.06	220	0.69	0.79	1.5	−2.26 to 3.63	−2.29 to 3.56
9	0.05	237	0.55	0.6	1.27	−1.98 to 2.98	−1.91 to 3.04
10	−0.15	198	−1.23	−1.33	1.39	−3.97 to 1.42	−3.97 to 1.42
11	0.01	283	−0.43	−0.27	1.46	−3.34 to 2.4	−3.4 to 2.33

Abbreviations: HPD, highest posterior density; QBP, quantile-based probability.

information in the formal study dominates the inferences, and different formal studies yield different results. As shown in Table 5, when the observed effect size in the formal study (d_j) is larger, the posterior point estimates tend to be larger; when the formal study has a larger sample size ($n_{1j}+n_{2j}$), the credible intervals tend to be narrower. Therefore, although AGDP is intuitively appealing, it is fundamentally flawed. On the other hand, if we calculate the weighted average of sample means or standardized sample means where the weight is the inverse of the sample size, the process is the same as the aforementioned Bayesian meta-analysis. That is, we conduct a Bayesian meta-analysis based on the 10 studies. A small difference is that the posterior distributions serve as the prior distributions for the 11th study.

9 | BAYESIAN POWER FOR SAMPLE SIZE PLANNING

All of the discussed data synthesis approaches can be implemented using Bayesian modeling. One important advantage of Bayesian data syntheses is that they allow researchers to compute statistical power and plan sample size more scientifically and cautiously than conventional frequentist power analysis. The conventional process plugs the estimates of parameters or effect sizes from the literature or pilot studies into the power calculation formula. This process ignores the fact that the estimates have uncertainty and treats the point estimates as if they were the population parameters, which may result in the planned study being underpowered.[6,48] A paradox is that if the population parameters were actually known, there would be no need to conduct sample size planning for any future studies. In contrast, Bayesian modeling can naturally consider uncertainty in the parameter estimates. The previous sections focus on Bayesian point estimation and credible intervals from posterior distributions, but the posterior distributions themselves model the uncertainty. Using any of the Bayesian integrative data analyses, Bayesian meta-analyses, or AUDP, the posterior distributions at the final step combine the information from multiple studies, which should represent the uncertainty of parameter estimation better than the posterior distributions from any single study. Bayesian power, power considering uncertainty, and assurance are relatively new concepts, but they provide a more scientific way to conduct sample size planning and deserve more attention.

After obtaining the posterior distributions of parameters, we can draw potential parameters from their posterior distributions and simulate data based on each set of the drawn parameters. For example, in the random-effects meta-analysis (Equation (8)), the overall population effect size μ_δ and the between-study variance τ^2 are drawn from the posterior distribution $p(\mu_\delta,\tau^2|d)$. Given each set of drawn μ_δ and τ, we simulate J study-level predicted true effect sizes (i.e., $\delta.pre_j$) by $\delta.pre_j \sim N(\mu_\delta,\tau^2)$ (J=11 in our real study). $\delta.pre_j \sim N(\mu_\delta,\tau^2)$ is the so-called posterior predictive distribution. In **rjags**, we can add one more line to the original code to specify the posterior predictive distribution of $\delta.pre_j$: **td.pre[j]~dnorm(d_mu,pre.phi)** is specified after **td[j]~dnorm(d_mu,pre.phi)** within the model part to indicate the distribution of the predicted true effect size $\delta.pre_j$. Then, in the output of the MCMC chain, J $\delta.pre_j$s are simulated at each iteration, which represents a new data set. As another example, in the mixed-effects integrative data analysis (Equation (2)), the fixed effects β_{00} to β_{12}, the level 1 residual variance σ_ϵ^2, and the level 2 variance-covariance matrix Σ_s are drawn by MCMC sampling in each iteration. Given the drawn values of parameters, $y.pre_{ij}$s are simulated with a planned sample size (nplan) that will be used in the

future study. In rjags, in addition to the original code, we need to specify another loop where `y.pre[p]`$\sim$ `dnorm(mu[p],pre.phi)` and p is from 1 to nplan, and we also need to specify how many individuals are in each group and their individual-level covariates. Then, the predicted observations $y.pre_{ij}$ are provided in the output of the MCMC chain.

After obtaining a number of sets of predicted observations, there are two options for Bayesian power calculation: the hybrid Bayesian power approach[6,49] and the full Bayesian power approach.[48,50] The definition of power is different in the two approaches. In the hybrid Bayesian power approach, power is calculated in the frequentist way based on whether the p value of the test is smaller than the α level (eg, 0.05). Thus, the definition of power is consistent with conventional frequentist power except that we consider uncertainty in power. For example, for Model (2), each set of predicted observations is analyzed by frequentist hypothesis testing, and we count how many times the test of β_{10} has a significant result. The proportion of significant results is the Bayesian power. Du and Wang[6] illustrated the hybrid Bayesian power computation process based on meta-analysis, and R functions were provided. In the full Bayesian power approach, the Bayesian analysis is conducted again with a noninformative prior for each set of the predicted data. Each of the constructed posterior distributions determines Bayesian significance, and there can be different criteria (eg, the 95% PHD interval should be narrower than a specific range). And we record the number of times a specific criterion is met, which is the Bayesian power. We refer to Kruschke[48] for the different criteria and detailed process.

Regardless of using the hybrid Bayesian power approach or the full Bayesian power approach, if we keep simulating predicted data sets and calculating Bayesian power repeatedly, in the end, we have a distribution of power instead of a single value of power. Assurance level and expected power then can be computed from the power distribution. Assurance level is the probability of the power values larger than the target power,[6] and expected power is the average power in the power distribution.[49] We use the random-effects meta-analysis with independent groups for an example. The R functions for calculating the assurance level and the expected Bayesian power are in Du and Wang.[6] On the basis of the 11 studies, even with 400 participants per group, we are only 6% certain of achieving the target power of 0.8 or higher in a future study, and the expected power is 0.3.

10 | CONCLUSION AND RECOMMENDATION

When synthesizing data from multiple studies, researchers overwhelmingly turn to meta-analysis, possibly because they are not aware of other options and therefore cannot appreciate their advantages. To broaden the options considered by researchers, the goal of this paper has been to introduce several Bayesian synthesis approaches for comparing two group means. Specifically, we present the algorithms for different approaches and introduce the models for meta-analysis and integrative data analysis, which also can be used in data fusion using AUDPs and data fusion using AGDPs. To facilitate the practical application of these methods, R code is provided. The results from the same model but with different approaches are presented and compared in Tables 3 to 5. The strengths and limitations of each method and model are summarized in Table 1.

Integrative data analysis is the most straightforward approach. After combining multiple data sets into a large pooled data set, a fixed-effects integrative data analysis is just the traditional data analysis as if we only had one large data set. The fixed-effects integrative data analysis is also the most commonly used integrative data analysis in the existing literature.[15-17] But because data are from different studies, researchers could monitor between-study heterogeneity in a random-effects integrative data analysis. Furthermore, because all the original information is retained in integrative data analysis, it offers a means of examining the influence of study-level, pair-level, and/or individual-level covariates in a mixed-effects integrative data analysis. Note that because of the extremely small level 1 sample size in the data sets described here, we could only fit a two-level model. In particular, for comparing independent group means, because members are not matched by pairs, we only estimated the individual-level residual variance and/or the between-study variance. For comparing matched group means, the between-pair variance is crucial; otherwise, we would have had to ignore within-pair interdependence. Thus, in this case, to estimate both the between-pair variance and the between-study variance, paired difference scores were used. To estimate both the individual-level residual variance and the between-pair variance, raw scores were used. Despite the strengths of integrative data analysis, it also has limitations. Most notably, pooling studies requires access to all the raw data from each study, but raw data are not always available. Moreover, accessing the raw data is often labor-intensive, time-consuming, and costly. In addition, the measurements used across studies must be the same or at least directly comparable (Table 1). In our real data example, the results vary across different integrative data analysis models (Table 3).

Among the discussed approaches, meta-analysis is used most widely. One strength of meta-analysis is that it requires only sample effect sizes (aggregated data), which are usually reported. Therefore, meta-analysis is less labor-intensive, less time-consuming, and cheaper compared with integrative

data analysis. Another strength of meta-analysis is that it allows for different measurement instruments across studies, as long as attenuation because of measurement error is corrected.[41] Meta-analysis also allows for fixed-, random-, and mixed-effects models. Nonetheless, in the mixed-effects meta-analysis, no individual-level or pair-level covariates can be incorporated because each study is treated as a unit. Another limitation is that the sampling distribution of the observed effect sizes is built on the normality approximation, which is valid only when the per-study sample size is relatively large (Table 1). In our real data example, the estimation of the overall population effect size is different across models depending on whether we were conditioning on specific levels of covariates or controlled for the power of the likelihoods (Table 3). Meta-analysis usually provides very similar estimation with integrative data analysis (data in integrative data analysis are standardized) when integrative data analysis does not consider individual-level covariates, pair-level covariates, and between-pair heterogeneity, and both use the same measurement instruments.[13,36] But it is difficult to draw the conclusion that meta-analysis and integrative data analysis always provide the same hypothesis testing results even when the aforementioned criteria are met, since meta-analysis is based on aggregated data and loses information to some degree compared with integrative data analysis. Although Cooper and Patall[13] and Lambert et al[36] found that in the long run, meta-analysis has smaller power than integrative data analysis, in each specific study, integrative data analysis may yield nonsignificant results when meta-analysis yields significant results. For example, in our real data example, the random-effects integrative data analysis for independent groups does not find a significant difference between husbands and wives in marital satisfaction, but the random-effects meta-analysis for independent group finds a significant difference.

In contrast to integrative data analysis and meta-analysis, most researchers are not familiar with the data fusion using AUDPs approach, with few exceptions.[3] AUDP can be applied to both raw data and aggregated data (effect sizes). Noninformative priors or informative priors that come from one of the studies can be used to initialize the process. The major strength of AUDP is that the contribution of each study is clearly summarized, and it is easy to see how the results are updated when each study enters the analysis. The limitation of AUDP is that the order in which the data enter the analysis may impact the final decision of rejecting the null hypothesis in some cases, as shown in our real data example. But it is also worth noting that, in our example, the posterior mean and mode of the parameter of interest are close to 0. When the posterior point estimate deviates more from 0, the number of studies is larger, and per-study sample size is larger, the Bayesian statistical conclusion will be more consistent across different orders since the final inferences are drawn based the posterior distribution $f(\theta|D_1,D_2,\ldots,D_J)$. $f(\theta|D_1,D_2,\ldots,D_J)$ captures the information of all observations regardless of the order. To ensure that different orders in AUDP do not impact the final results substantially and support different decisions of rejecting the null hypothesis, researchers using this approach should conduct a sensitivity analysis with different orders to investigate whether and how the estimation varies. One issue not explored in the current paper is how the order may influence convergence. If the first entered data set has a small sample size, nonconvergence may occur when the model is complex.

If researchers are familiar with the moments of different distributions (eg, mean and variance), data fusion using AGDPs seems an intuitive way to construct priors. But AGDP has two major drawbacks (Table 1). The first one is that the credible intervals are much wider than those from the same models in the Bayesian integrative data analyses. Although AGDP uses the results from multiple studies to construct priors, it only uses limited information from each study. In our real data example, we only used the sample group mean and sample group variance of each study but did not consider the sample size information of each study. Although meta-analysis also only uses the observed effect size from each study, the sampling variance of the observed effect sizes is calculated based on the per-study sample size. Thus, the final inferences from meta-analysis consider the sample sizes of all studies. As a consequence of losing information, researchers lose the precision of estimation in AGDP; therefore, the credible intervals are relatively wide. Another drawback is that both the point estimates and credible intervals greatly depend on which study serves as the formal study to provide likelihood. The reason is that the sample size is considered in the formal study but not in the priors. As a result, the priors do not provide much information compared with the formal study and have limited influence on the posterior distributions. Therefore, AGDP is not recommended, unless we calculate the weighted average of sample means in AGDP, which makes it the same as Bayesian meta-analysis.

Given the strengths and limitations of each approach (Table 1), we offer several recommendations. AGDP has technical problems as summarized above, and thus, it is not recommended. When the raw data from each study are available, researchers can use integrative data analysis, meta-analysis, or AUDP; but an integrative data analysis or AUDP using the raw data is ideal, because analysis with raw data is more powerful than analysis with aggregated data.[13,36] If the influence of the individual-level or pair-level covariates is important to the research question, an

integrative data analysis is necessary because it keeps the raw data. If researchers want to further examine the contribution of each study, AUDP using the raw data should be considered. If only the effect sizes of studies are available, a meta-analysis and AUDP using the observed effect sizes are the available options. When researchers plan to control for study quality, power prior can be paired with any of these approaches. When the measurements of different studies are neither the same nor can be equated, even if the raw data are accessible, a meta-analysis with correcting attenuation because of measurement errors should be used. In term of the model, if between-study heterogeneity is documented in the literature or is anticipated by researchers, random- and mixed-effects models that freely estimate the between-study variance should be used; if researchers plan to use covariates to further explain within-study variance and/or between-study variance, mixed-effects models should be used. Overall, researchers should choose the most appropriate method and model based on the available information and their specific research questions.

Bayesian synthesis approaches not only illustrate a way to estimate the parameters of interest based on the current information of multiple studies but also provide a way to cautiously plan sample size for a future study, which is another refined goal of statistical analysis. More specifically, Bayesian synthesis approaches compute statistical power while accounting for the uncertainty in the parameter estimation. Each Bayesian synthesis approach could provide posterior distributions, based on which we can use either the hybrid Bayesian power approach or the full Bayesian power approach to calculate the assurance level and the expected power. Bayesian power considers that the uncertainty of parameter estimation leads to the uncertainty of power. Thus, instead of claiming that we are 100% certain of achieving power of 0.8 in a future study (i.e., the traditional power concept), achieving power of 0.8 or higher is treated as a probability event and can be expressed by the assurance level.

One future direction is to explore the performance of each method with simulation. Although meta-analysis is widely used, simulation studies of examining the performance of Bayesian meta-analysis are rare. And the performance of Bayesian integrative analysis and AUDP is not widely studied except few papers.[3,15] Therefore, examining the performance of each method under different conditions (eg, different sample size and prior) should be done in the future.

In conclusion, each method has its own pros and cons. Researchers should make the decisions based on their research questions, data structures, and the features of Bayesian synthesis approaches summarized in this work. Furthermore, all of the Bayesian synthesis approaches can provide Bayesian power with assurance level and expected power, facilitating sample size planning.

DATA AVAILABILITY STATEMENT

All the effect sizes and raw data are not provided in Appendix A and supporting information. The code in Appendix A and supporting information can be used to analyze data that are similar to the data used in the paper. Interested readers can email the corresponding author for more information on the data used in the paper.

CONFLICT OF INTEREST

The author reported no conflict of interest.

ORCID

Han Du https://orcid.org/0000-0002-3174-6935

REFERENCES

1. Cooper H, Hedges LV, Valentine JC. *The Handbook of Research Synthesis and Meta-analysis*. New York, NY: Russell Sage Foundation; 2009.
2. Curran PJ, Hussong AM. Integrative data analysis: the simultaneous analysis of multiple data sets. *Psychol Methods*. 2009;14(2):81–100. https://doi.org/10.1037/a0015914.
3. Marcoulides KM. A Bayesian synthesis approach to data fusion using augmented data-dependent priors. *Ph.D. Thesis*; 2017.
4. Maxwell SE, Kelley K, Rausch JR. Sample size planning for statistical power and accuracy in parameter estimation. *Annu Rev Psychol*. 2008;59(1):537–563. https://doi.org/10.1146/annurev.psych.59.103006.093735.
5. Cohen J. *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum; 2013.
6. Du H, Wang L. A bayesian power analysis procedure considering uncertainty in effect size estimates from a meta-analysis. *Multivar Behav Res*. 2016;51(5):589–605. https://doi.org/10.1080/00273171.2016.1191324.
7. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis*. London: Chapman & Hall; 2014.
8. Karney BR, Bradbury TN. Neuroticism, marital interaction, and the trajectory of marital satisfaction. *J Pers Soc Psychol*. 1997;72(5):1075–1092. https://doi.org/10.1037/0022-3514.72.5.1075.
9. Osgood CE, Suci GJ, Tannenbaum PH. *The Measurement of Meaning*. Urbana, IL: University of Illinois Press; 1957.
10. Du H, Wang L. The impact of the number of dyads on estimation of dyadic data analysis using multilevel modeling. *Methodology*. 2016;23:21–31. https://doi.org/10.1027/1614-2241/a000105.
11. Kenny DA, Kashy DA, Cook WL, Simpson JA. *Dyadic data analysis (methodology in thessocial sciences)*. New York, NY: Guilford; 2006.
12. Jackson JB, Miller RB, Oka M, Henry RG. Gender differences in marital satisfaction: a meta-analysis. *J Marriage Fam*. 2014;76(1):105–129. https://doi.org/10.1111/jomf.12077.

13. Cooper H, Patall EA. The relative benefits of meta-analysis conducted with individual participant data versus aggregated data. *Psychol Methods*. 2009;14(2):165–176. https://doi.org/10.1037/a0015565.
14. Hofer SM, Piccinin AM. Integrative data analysis through coordination of measurement and analysis protocol across independent longitudinal studies. *Psychol Methods*. 2009;14(2):150–164. https://doi.org/10.1037/a0015566.
15. Marcoulides KM, Grimm KJ. Data integration approaches to longitudinal growth modeling. *Educ Psychol Meas*. 2016;77(6):971–989. https://doi.org/10.1177/0013164416664117.
16. McArdle JJ, Hamagami F, Meredith W, Bradway KP. Modeling the dynamic hypotheses of Gf–Gc theory using longitudinal lifespan data. *Learn Individ Differ*. 2000;12(1):53–79. https://doi.org/10.1016/S1041-6080(00)00036-4.
17. Sternberg KJ, Baradaran LP, Abbott CB, Lamb ME, Guterman E. Type of violence, age, and gender differences in the effects of family violence on children's behavior problems: A mega-analysis. *Dev Rev*. 2006;26(1):89–112. https://doi.org/10.1016/j.dr.2005.12.001.
18. Kolen MJ, Brennan RL. *Test equating, scaling, and linking: methods and practices*. New York, NY: Springer; 2004.
19. Hedges LV, Olkin I. *Statistical Method for Meta-analysis*. London: Academic Press; 1985.
20. Hedges LV, Vevea JL. Fixed-and random-effects models in meta-analysis. *Psychol Methods*. 1998;3(4):486–504. https://doi.org/1037/1082-989x.3.4.486.
21. Card NA. *Applied Meta-analysis for Social Science Research*. New York, NY: Guilford Publications; 2015.
22. Borenstein M, Hedges LV, Higgins J, Rothstein HR. *Introduction to Meta-analysis*. West Sussex, UK: Wiley; 2009.
23. Chung Y, Rabe-Hesketh S, Choi I. Avoiding zero between-study variance estimates in random-effects meta-analysis. *Stat Med*. 2013;32(23):4071–4089. https://doi.org/10.1002/sim.5821.
24. Thompson SG. Controversies in meta-analysis: the case of the trials of serum cholesterol reduction. *Stat Methods Med Res*. 1993;2(2):173–192. https://doi.org/10.1177/096228029300200205.
25. Plummer M. rjags: Bayesian graphical models using MCMC. R package version 4–6; 2016.
26. Lynch SM. *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*. New York, NY: Springer Science & Business Media; 2007.
27. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. *Stat Sci*. 1992;7(4):457–472. https://doi.org/10.1214/ss/1177011136.
28. Rubin DB. Bayesianly justifiable and relevant frequency calculations for the applies statistician. *Ann Stat*. 1984;12(4):1151–1172. https://doi.org/10.1214/aos/1176346785.
29. Cowles MK. *Applied Bayesian Statistics: with R and Open BUGS Examples*. New York, NY: Springer Science & Business Media; 2013.
30. Lunn D, Jackson C, Best N, Thomas A, Spiegelhalter D. *The BUGS Book: A practical Introduction to Bayesian Analysis*. Boca Raton, FL: CRC press; 2012.
31. Enders CK, Du H, Keller BT. Amodel-based imputation procedure for multilevel regression models with random coefficients,interaction effects, and nonlinear terms. *Psychol Methods*. in press. http://doi.org/10.1037/met0000228.
32. Jeng GT, Scott JR, Burmeister LF. A comparison of meta-analytic results using literature vs individual patient data: paternal cell immunization for recurrent miscarriage. *JAMA*. 1995;274(10):830–836. https://doi.org/10.1001/jama.1995.03530100070037.
33. Olkin I, Sampson A. Comparison of meta-analysis versus analysis of variance of individual patient data. *Biometrics*. 1998:317–322. https://doi.org/10.2307/2534018.
34. Steinberg KK, Smith SJ, Stroup DF, et al. Comparison of effect estimates from a meta-analysis of summary data from published studies and from a meta-analysis using individual patient data for ovarian cancer studies. *Am J Epidemiol*. 1997;145(10):917–925. https://doi.org/10.1093/oxfordjournals.aje.a009051.
35. Stewart LA, Parmar MKB. Meta-analysis of the literature or of individual patient data: is there a difference? *The Lancet*. 1993;341(8842):418–422. https://doi.org/10.1016/0140-6736(93)93004-K.
36. Lambert PC, Sutton AlJ, Abrams KR, Jones DR. A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. *J Clin Epidemiol*. 2002;55(1):86–94. https://doi.org/10.1016/s0895-4356(01)00414-0.
37. Hedges LV. A random effects model for effect sizes. *Psychol Bull*. 1983;93(2):388–395. https://doi.org/10.1037/0033-2909.93.2.388.
38. Johnson NL, Welch BL. Applications of the non-central t-distribution. *Biometrika*. 1940;31(3/4):362–389. https://doi.org/10.2307/2332616.
39. Dunlap WP, Cortina JM, Vaslow JB, Burke MJ. Meta-analysis of experiments with matched groups or repeated measures designs. *Psychol Methods*. 1996;1(2):170–177. https://doi.org/10.1037/1082-989x.1.2.170.
40. Becker BJ. Synthesizing standardized mean-change measures. *Br J Math Stat Psychol*. 1988;41(2):257–278. https://doi.org/10.1111/j.2044-8317.1988.tb00901.x.
41. Schmidt FL, Hunter JE. *Methods of meta-analysis: correcting error and bias in research findings*. Thousand Oaks, CA: Sage publications; 2014.
42. Zhang Z, Jiang K, Liu H, Oh I. Bayesian meta-analysis of correlation coefficients through power prior. *Communications in Statistics-Theory and Methods*. 2017;46(24):11988–12007. https://doi.org/10.1080/03610926.2017.1288251.
43. Ibrahim JG, Chen M. Power prior distributions for regression models. *Stat Sci*. 2000;15(1):46–60. https://doi.org/10.1214/ss/1009212673.
44. Ibrahim JG, Chen M, Gwon Y, Chen F. The power prior: theory and applications. *Stat Med*. 2015;34(28):3724–3749. https://doi.org/10.1002/sim.6728.
45. Neuenschwander B, Branson M, Spiegelhalter DJ. A note on the power prior. *Stat Med*. 2009;28(28):3562–3566. https://doi.org/10.1002/sim.3722.
46. Rietbergen C, Klugkist I, Janssen KJM, Moons KGM, Hoijtink HJA. Incorporation of historical data in the analysis of randomized therapeutic trials. *Contemp Clin Trials*. 2011;32(6):848–855. https://doi.org/10.1016/j.cct.2011.06.002.
47. Cohn LD, Becker BJ. How meta-analysis increases statistical power. *Psychol Methods*. 2003;8(3):243–253. https://doi.org/10.1037/1082-989x.8.3.243.
48. Kruschke JK. Bayesian estimation supersedes the t test. *J Exp Psychol Gen*. 2013;142(2):573–603. https://doi.org/10.1037/a0029146.

49. Liu F. An extension of Bayesian expected power and its application in decision making. *J Biopharm Stat*. 2010;20:941–953. https://doi.org/10.1080/10543401003618967.
50. Spiegelhalter DJ, Abrams KR, Myles JP. *Bayesian approaches to clinical trials and health-care evaluation*. Chichester, UK: Wiley; 2004.

SUPPORTING INFORMATION

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

How to cite this article: Du H, Bradbury TN, Lavner JA, et al. A comparison of Bayesian synthesis approaches for studies comparing two means: A tutorial. *Res Syn Meth*. 2020;11:36–65. https://doi.org/10.1002/jrsm.1365

APPENDIX A: | LEFT HEART GEOMETRY

R code is provided for all the aforementioned analyses in the Supporting Information. In Appendix A, the R code for mixed-effects integrative data analysis, mixed-effects meta-analysis (with and without a power prior), and AUDP using a power prior is illustrated as representative examples.

```
##################
####MIXED-EFFECTS INTEGRATIVE DATA ANALYSIS FOR INDEPENDENT GROUPS
##################
model= "
model {
##Level-2
for (j in 1:J){
beta.j[j, 1:3]~dmnorm(mub[j,1:3], pre.S[1:3,1:3])
mub[j,1]<-beta[1]+beta[2]*s_covariate[j]
mub[j,2]<-beta[3]+beta[4]*s_covariate[j]
mub[j,3]<-beta[5]+beta[6]*s_covariate[j]
}
##Level-1
for (i in 1:N){
y[i]~dnorm(muy[i],pre.phi)
muy[i]<-beta.j[sample[i],1]+beta.j[sample[i],2]*role[i]
+beta.j[sample[i],3]*i_covariate[i]
}
#Priors
for (i in 1:6){
beta[i]~dnorm(0, a)
}
pre.phi ~ dgamma(b,b)
pre.S[1:3,1:3]~dwish(V[1:3,1:3],m)
V[1,1]<-1
V[2,2]<-1
V[3,3]<-1
V[1,2]<-0
V[1,3]<-0
V[2,3]<-0
V[2,1]<-V[1,2]
V[3,1]<-V[1,3]
V[3,2]<-V[2,3]
##Define coefficients of interest
sigma2<-1/pre.phi
cov[1:3,1:3]<-inverse(pre.S[1:3,1:3])
sig.u0<-cov[1,1]
sig.u1<-cov[2,2]
sig.u2<-cov[3,3]
rho.u01<-cov[1,2]/sqrt(cov[1,1]*cov[2,2])
rho.u02<-cov[1,3]/sqrt(cov[1,1]*cov[3,3])
rho.u12<-cov[2,3]/sqrt(cov[3,3]*cov[2,2])
}
"
# Save model
writeLines(model, con="model.txt")
#- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
```

```
# Load data
dataList = list(
J = 11,
y = y, # Read in the data as long formart
sample = sample,
i_covariate=i_covariate,
s_covariate=s_covariate,
N = N,
role = role,
m = 3,
b = 0.001,
a = 1/10000
#- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
# Specifying starting values in two independent chains
beta <- c(0,0,0,0,0,0)
pre.S <- matrix(c(1,0,0,0,1,0,0,0,1),nrow=3)
pre.phi <- 1
initsList1 = list( beta=beta, pre.S=pre.S, pre.phi=pre.phi,
.RNG.name="base::Wichmann-Hill",.RNG.seed=2018)
beta <- c(1,1,1,1,1,1)
pre.S <- matrix(c(0.5,0.2,0.2,0.2,0.5,0.2,0.2,0.2,0.5),nrow=3)
pre.phi <- 0.5
initsList2 = list( beta=beta, pre.S=pre.S, pre.phi=pre.phi,
.RNG.name="base::Wichmann-Hill",.RNG.seed=2016)
#- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
parameters = c("beta", "sigma2","sig.u0","sig.u1","sig.u2","rho.u01",
"rho.u02","rho.u12") # Specify the estimated parameters
adaptSteps =500  # Adaptive period
burnInSteps = 2000  # Burn-in period
nChains = 1
thinSteps=100  #Thinning period
numSavedSteps=3000  # The number of kept iterations
nIter = ceiling( numSavedSteps * thinSteps )
jagsModel1 = jags.model( "model.txt" , data=dataList , inits=initsList1,
n.chains=nChains , n.adapt=adaptSteps )
jagsModel2 = jags.model( "model.txt" , data=dataList , inits=initsList2,
n.chains=nChains , n.adapt=adaptSteps )
update( jagsModel1 , n.iter=burnInSteps)
update( jagsModel2 , n.iter=burnInSteps)
codaSamples1 = coda.samples( jagsModel1 , variable.names=parameters,
n.iter=nIter , thin=thinSteps)
codaSamples2 = coda.samples( jagsModel2 , variable.names=parameters,
n.iter=nIter , thin=thinSteps)
mcmcChain1 = as.matrix( codaSamples1 )
mcmcChain2 = as.matrix( codaSamples2 )
mcmcChain<-rbind(mcmcChain1,mcmcChain2)
##################
####MIXED-EFFECTS META-ANALYSIS FOR INDEPENDENT GROUPS
##################
model = "
model {
for (j in 1:J){
for (j in 1:J){
```

```
d[j]~dnorm(delta[j],a0[j]/((n1[j]+n2[j])/n1[j]/n2[j]+ delta[j]^2/(2*(n1[j]+n2[j]))))
delta[j] ~ dnorm(mu[j], pre.phi)
mu[j]<-mu.d+beta*s_covariate[j]
}
pre.phi ~ dgamma(b,b)
tau2<-1/pre.phi
mu.d~dnorm(0,a)
beta~dnorm(0,a)
}
"
# Save model
writeLines(model, con="model.txt")
#- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
# Load data
dataList = list(
d=d,
n1=n1,
n2=n2,
s_covariate=s_covariate,
J = 11,
b = 0.001,
a = 1/10000
)
#- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
# Specifying starting values in two independent chains
mu.d<-0
pre.phi<-1
beta<-0
initsList1 = list(mu.d=mu.d,pre.phi=pre.phi,beta=beta,
.RNG.name="base::Wichmann-Hill",.RNG.seed=2018)
mu.d<-0.2
pre.phi<-0.5
beta<-0.2
initsList2 = list( mu.d=mu.d,pre.phi=pre.phi,beta=beta,
.RNG.name="base::Wichmann-Hill",.RNG.seed=2016)
#- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
parameters = c( "mu.d",""tau2","beta") # Specify the estimated parameters
adaptSteps =500  # Adaptive period
burnInSteps = 1000  # Burn-in period
nChains = 1
thinSteps=80  # Thinning period
numSavedSteps=3000  # The number of kept iterations
nIter = ceiling( numSavedSteps * thinSteps )
jagsModel1 = jags.model(
"model.tx" , data=dataList , inits=initsList1,
n.chains=nChains , n.adapt=adaptSteps )
\onecolumn
jagsModel2 = jags.model( "model.txt" , data=dataList , inits=initsList2,
n.chains=nChains , n.adapt=adaptSteps )
update( jagsModel1 , n.iter=burnInSteps)
update( jagsModel2 , n.iter=burnInSteps)
codaSamples1 = coda.samples( jagsModel1 , variable.names=parameters,
n.iter=nIter , thin=thinSteps)
```

```
codaSamples2 = coda.samples( jagsModel2 , variable.names=parameters,
n.iter=nIter , thin=thinSteps)
mcmcChain1 = as.matrix( codaSamples1 )
mcmcChain2 = as.matrix( codaSamples2 )
mcmcChain<-rbind(mcmcChain1,mcmcChain2)
##################
####MIXED-EFFECTS META-ANALYSIS FOR INDEPENDENT GROUPS WITH POWER PRIOR
##################
model = "
model {
for (j in 1:J){
d[j] ~ dnorm(td[j],a0[j]/((n1[j]+n2[j])/n1[j]/n2[j]+ td[j]^2/(2*(n1[j]+n2[j]))))
td[j] ~ dnorm(delta[j], a0[j]*pre.phi)
delta[j] <- d_mu+beta*s_covariate[j]
}
pre.phi ~ dgamma(0.001,0.001)
sigma <- 1/pre.phi
d_mu ~ dnorm(mu0,tau2)
beta ~ dnorm(mu0,tau2)
}
"
##################
####AUDP USING POWER PRIOR (FIXED-EFFECTS MODEL WITH EFFECT SIZE ESTIMATES FROM INDEPENDENT GROUPS)
##################
J=11
out.table<-matrix(NA, nrow = J, ncol =9)
colnames(out.table)<-c('',`length`,`estimate.mean`,`estimate.mode`,`SD`,
`CI.L`,`CI.U`,`HPD.L`,`HPD.U`)
p.mean<-c()
p.sd<-c()
set.seed(2004)
order<-sample(1:J,J,replace=FALSE)
d<-d[order]
n1<-n1[order]
n2<-n2[order]
a0<-a0[order]
for (j in 1:J){
model = "
model {
d[j]~dnorm(mu.d,a0[j]/((n1[j]+n2[j])/n1[j]/n2[j]+ mu.d^2/(2*(n1[j]+n2[j]))))
mu.d~dnorm(mu0,a)
}
"
writeLines(model, con="model.txt")
#- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
if (j==1){
mu0<-0
a<- 1/10000
} else {
mu0<- p.mean[j-1]
a<-1/(p.sd[j-1]^2)}
dataList = list(
d=d,
```

```
n1=n1,
n2=n2,
a0=a0,
j = j,
mu0 = mu0,
a = a
)
#- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
# Specifying starting values in two independent chains
mu.d<-0
initsList1 = list( mu.d=mu.d,
.RNG.name="base::Wichmann-Hill",.RNG.seed=2018)
mu.d<-0.2
initsList2 = list( mu.d=mu.d,
.RNG.name="base::Wichmann-Hill",.RNG.seed=2016)


#- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
parameters = c( "mu.d") # Specify the estimated parameters
adaptSteps =500  # Adaptive period
burnInSteps = 2000  # Burn-in period
nChains = 1
thinSteps=10  # Thinning period
numSavedSteps=3000  # The number of kept iterations
nIter = ceiling( numSavedSteps * thinSteps )
jagsModel1 = jags.model("model.txt" , data=dataList , inits=initsList1,
n.chains=nChains , n.adapt=adaptSteps )
jagsModel2 = jags.model("model.txt" , data=dataList , inits=initsList2,
n.chains=nChains , n.adapt=adaptSteps )
update( jagsModel1 , n.iter=burnInSteps)
update( jagsModel2 , n.iter=burnInSteps)
codaSamples1 = coda.samples( jagsModel1 , variable.names=parameters,
n.iter=nIter , thin=thinSteps)
codaSamples2 = coda.samples( jagsModel2 , variable.names=parameters,
n.iter=nIter , thin=thinSteps)
mcmcChain1 = as.matrix( codaSamples1 )
mcmcChain2 = as.matrix( codaSamples2 )
mcmcChain<-rbind(mcmcChain1,mcmcChain2)
for (i in 1:1){
p.mean[j]<- sum.stat(mcmcChain[,i])[3]
p.sd[j]<- sum.stat(mcmcChain[,i])[4]
out.table[j,]<-c(j,sum.stat(mcmcChain[,i]))
}
}
out.table<-cbind(order,d,out.table)
```