


Enriching Meta-Analytic Models of Summary Data: A Thought Experiment and Case Study

**Blakeley B. McShane**  and **Ulf Böckenholt**

Marketing Department, Kellogg School of Management, Northwestern University

Advances in Methods and
Practices in Psychological Science
2020, Vol. 3(1) 81–93
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2515245919884304
www.psychologicalscience.org/AMPPS


Abstract

Meta-analysis typically involves the analysis of summary data (e.g., means, standard deviations, and sample sizes) from a set of studies via a statistical model that is a special case of a hierarchical (or multilevel) model. Unfortunately, the common summary-data approach to meta-analysis used in psychological research is often employed in settings where the complexity of the data warrants alternative approaches. In this article, we propose a thought experiment that can lead meta-analysts to move away from the common summary-data approach to meta-analysis and toward richer and more appropriate summary-data approaches when the complexity of the data warrants it. Specifically, we propose that it can be extremely fruitful for meta-analysts to act as if they possess the individual-level data from the studies and consider what model specifications they might fit even when they possess only summary data. This thought experiment is justified because (a) the analysis of the individual-level data from the studies via a hierarchical model is considered the “gold standard” for meta-analysis and (b) for a wide variety of cases common in meta-analysis, the summary-data and individual-level-data approaches are, by a principle known as statistical sufficiency, equivalent when the underlying models are appropriately specified. We illustrate the value of our thought experiment via a case study that evolves across five parts that cover a wide variety of data settings common in meta-analysis.

Keywords

meta-analysis, hierarchical model, multilevel model, random effects, heterogeneity, between-study variation, open data, open materials

Received 6/19/18; Revision accepted 10/2/19

Meta-analysis typically involves the analysis of summary data (e.g., means, standard deviations, and sample sizes) from a set of studies via a statistical model that is a special case of a hierarchical (or multilevel) model. The common summary-data approach to meta-analysis is the basic random-effects meta-analytic model, and this approach is overwhelmingly dominant in psychological research (e.g., in meta-analyses published in *Psychological Bulletin*; Stanley, Carter, & Doucouliagos, 2018; van Erp, Verhagen, Grasman, & Wagenmakers, 2017). However, recent work demonstrates that meta-analytic practice in psychology is in need of improvement (Tipton, Pustejovsky, & Ahmadi, 2019), and this is consistent with the fact that this model is often employed in settings where the complexity of the data warrants richer and more appropriate approaches.

Specifically, the basic random-effects meta-analytic model is a univariate, two-level model. Consequently, it is suitable only when there is a single group of subjects (or a single experimental condition), a single dependent measure, and a single effect of interest in each study. This is seldom the case in contemporary psychological research studies, and when there is more than one of any of these, the approach can be problematic, and more extensive results can be obtained via richer and more appropriate summary-data approaches.

Corresponding Author:

Blakeley B. McShane, Marketing Department, Kellogg School of Management, Northwestern University, 2211 Campus Dr., Evanston, IL 60208
E-mail: b-mcshane@kellogg.northwestern.edu

In this article, we propose a thought experiment that can lead meta-analysts to move away from the common summary-data approach to meta-analysis and toward richer and more appropriate summary-data approaches when the complexity of the data warrants it. Specifically, we propose that it can be extremely fruitful for meta-analysts to act as if they possess the individual-level data from the studies and consider what model specifications they might fit even when they possess only summary data.

This thought experiment is justified by considering the following two facts in tandem. First, although meta-analysis is often equated to the analysis of summary data, the analysis of the individual-level data from the studies via a hierarchical model is considered the “gold standard” for meta-analysis (Cooper & Patall, 2009; Haidich, 2010; Simmonds et al., 2005; Stewart & Tierney, 2002). Second, however, for a wide variety of cases common in meta-analysis, the summary-data and individual-level-data approaches are, by a principle known as statistical sufficiency, equivalent when the underlying models are appropriately specified.

The value of our thought experiment is twofold. First, it makes clear whether or not there is a single group of subjects (or a single experimental condition), a single dependent measure, and a single effect of interest in each study. Second, when there is more than one of any of these, it suggests not only that one should move away from the common summary-data approach to meta-analysis but also how one might move away from it and toward richer and more appropriate summary-data approaches.

Such summary-data approaches to meta-analysis are by no means new. Indeed, they were introduced and applied in noted research articles (Berkey, Hoaglin, Antczak-Bouckoms, Mosteller, & Colditz, 1998; Kalaian & Raudenbush, 1996; Raudenbush, Becker, & Kalaian, 1988; Rosenthal & Rubin, 1986), have been covered in classic textbooks (Hedges & Olkin, 1985; Raudenbush & Bryk, 2002) and handbooks (Becker, 2000; Cooper & Hedges, 1994; Cooper, Hedges, & Valentine, 2009), and remain the subject of research (Cheung, 2015; McShane & Böckenholt, 2017, 2018b).

It is not our intent to provide a full review of these approaches. Instead, we illustrate via case study how our proposed thought experiment, based on the equivalence of the summary-data and individual-level-data approaches, can lead to increased use of these approaches when the complexity of the data warrants it. However, we note that these richer and more appropriate summary-data approaches generalize the common summary-data approach along four principal dimensions, to accommodate (a) an arbitrary number of groups of subjects arising

from discrete individual-level covariates (or, alternatively, an arbitrary number of experimental conditions arising from the manipulation of one or more discrete experimental factors), (b) an arbitrary number of levels that account for the variation and covariation in effect sizes (also known as heterogeneity) induced by the fact that observations are nested (e.g., individuals nested within demographic groups nested within countries, subjects nested within study conditions nested within studies nested within articles, and students nested within classrooms nested within schools), (c) an arbitrary number of dependent measures, and (d) study-level covariates (or moderators). Consequently, we organize our case study to evolve across these four dimensions.

Following our case study, we conclude by discussing these four dimensions with respect to which our thought experiment can be beneficial, using the case study for illustration, two potential objections to our thought experiment, and some advantages that individual-level data offer over summary data despite the fact that the two may be equivalent for the purpose of conceptualizing the model specification.

Disclosures

The data and code for our case study are available both as Supplemental Material (<http://journals.sagepub.com/doi/suppl/10.1177/2515245919884304>) and at the Open Science Framework (<https://osf.io/ua9h4/>).

Case Study

This case study illustrates how it can be extremely fruitful for meta-analysts to act as if they possess the individual-level data and consider what model specifications they might fit even when they possess only summary data. The case study evolves across five parts that cover a wide variety of data settings common in meta-analysis. Specifically, the data sequentially build in complexity across the parts, which in turn necessitates increasingly complex model specifications that also sequentially build upon one another. In each part, we show how the results produced by a model fit to the individual-level data can be matched by an appropriately specified model fit to the summary data, thereby illustrating the equivalence of the summary-data and individual-level-data approaches to meta-analysis—and thus the value of our thought experiment—in a straightforward manner across a sequence of increasingly complex data settings.

In Part I, there are a single group of subjects and a single dependent measure in each study, two levels in the nesting structure of the data, and no study-level

covariates. Interest centers on the overall average of the dependent measure. The common summary-data approach to meta-analysis is fully appropriate in this setting, and we illustrate its equivalence to the individual-level-data approach.

In Part II, the case study evolves to include two groups of subjects in each study and thus three levels in the nesting structure. Even though this is the canonical setting for the common summary-data approach presented in introductory meta-analysis textbooks (see, e.g., Borenstein, Hedges, Higgins, & Rothstein, 2009) and this approach is overwhelmingly dominant in practice in this setting and beyond, we illustrate how it is less appropriate than the individual-level-data approach in that the latter provides more extensive results, in particular with regard to heterogeneity. We then illustrate how the individual-level-data approach can lead one to consider a richer and more appropriate summary-data approach to meta-analysis that is equivalent to the individual-level-data approach.

In Parts III through V, the case study evolves to include several (four) groups of subjects in each study, study-level covariates, and multiple (three) dependent measures, respectively. We illustrate further deficiencies of the common summary-data approach and how individual-level-data and richer summary-data approaches are equivalent, are more appropriate, and provide more extensive results.

We summarize the evolution of the case study in Table 1, which shows the four dimensions across which the five parts vary. There are important distinctions regarding the levels of each dimension. The primary distinction regarding the number of groups is, as discussed in Parts I through III, one versus two versus several (i.e., more than two). An important empirical distinction regarding the number of levels is also one versus two versus several. One-level models do not allow for between-study variation in effect sizes, two-level models allow for between-study variation but not covariation, and three- and higher-level models allow for both between-study variation and covariation; given recent empirical work documenting that between-study variation is rife in psychological research (McShane, Tackett, Böckenholt, & Gelman, 2019; Stanley et al., 2018; van Erp et al., 2017), one-level models are seldom appropriate, and the choice between two-level models and three- and higher-level models will, as discussed in Part II, depend on the nesting structure of the data. The primary distinction regarding the number of dependent measures is, as discussed in Part V, one versus multiple (i.e., more than one). Finally, the primary distinction regarding study-level covariates is obviously, as discussed in Part IV, their absence or presence.

Table 1. Case-Study Schema

Part	Number of groups	Number of levels	Number of dependent measures	Study-level covariates
Part I	One	Two	One	No
Part II	Two	Three	One	No
Part III	Four	Three	One	No
Part IV	Four	Three	One	Yes
Part V	Four	Three	Three	Yes

Data

The original data on which our principal data are based are from Johnson (2014), who administered the IPIP-NEO-120 to 619,150 individuals; these original data are available at the Open Science Framework (<https://osf.io/wxvth>). The IPIP-NEO-120 is a 120-item inventory designed to yield assessments of each of the five broad domains of the five-factor model of personality, as well as each of the six narrower facets of each of the five broad domains. It features 24 items per domain, and 4 items per facet. Responses are on the 1- to 5-point integer scale. The data contain the response of each individual to each item, as well as the sex, age, and country of each individual.

We defined college-age individuals as those ages 18 to 21 inclusive and adults as individuals over the age of 21. We restricted our analysis to the 47 countries with at least 25 individuals in each of the four groups defined by sex and age (i.e., (female, male) \times (adult, college)); in the case of countries with more than 100 individuals in a given group, we randomly subsampled 100 individuals from the group.

In the first three parts of our case study, we treat the data from each country simply as a separate study. In the fourth and fifth parts, we model differences among countries. To do so, we augment our principal data with the latitude measurement of the capital city of each country, available at <https://www.latlong.net>.

Part I: one group and two levels

We begin with the data setting depicted in the first row of Table 1. Suppose a personality psychologist was interested in conducting a meta-analysis of extraversion in adult males. Toward this end, he gathered data from 47 studies, in all of which extraversion was measured via the simple average of the 24 extraversion items included in the IPIP-NEO-120.

The common summary-data approach to meta-analysis of these data involves the following. Let y_i denote the mean of the individual-level data (i.e., extraversion scores) in study i . The model specification for the y_i is given by

$$y_i = \alpha + \beta_i + \varepsilon_i,$$

where α is treated as a fixed effect that models overall average extraversion, the β_i are treated as random effects for each study, and the ε_i are random errors for each study. The model further assumes that the β_i are independent and identically normally distributed with mean zero and variance τ^2 , the ε_i are independent normally distributed with mean zero and variance v_i , and there is zero covariation among the β_i and ε_j . This model is sometimes called the basic random-effects meta-analytic model; it is also sometimes called the two-level meta-analytic model because it allows for heterogeneity in effect sizes across studies (Level 2) via τ^2 , as well as sampling error (Level 1) via the v_i . We note that throughout this article, (a) terms denoted by a τ^2 (or a \mathbf{T} in Part V) model heterogeneity, (b) terms denoted by a v or a σ^2 (or a \mathbf{V} or a $\mathbf{\Sigma}$, respectively, in Part V) model sampling error, and (c) terms denoted by a v (or a \mathbf{V} in Part V) are, despite being estimates, treated as known, as is standard in the summary-data approach to meta-analysis.

A common approach to estimation of this model involves (a) estimating the sampling variances v_i in each study using the conventional formula (i.e., the variance of the individual-level data in each study divided by the sample size), (b) estimating τ^2 using restricted (or residual or reduced) maximum likelihood (REML) conditional on the estimates of the v_i , and (c) estimating α and its standard error using the standard best-linear-unbiased-prediction (BLUP) estimators conditional on the estimates of the v_i and τ^2 (Harville, 1977; Robinson, 1991). This estimation approach or an analogue of it is followed in all summary-data approaches to meta-analysis in this case study.

The individual-level-data approach to meta-analysis of these data involves the following. Let y_{ik} denote the individual-level data (i.e., extraversion score) for individual k in study i . The model specification for the y_{ik} is given by

$$y_{ik} = \alpha + \beta_i + \varepsilon_{ik},$$

where α is treated as a fixed effect that models overall average extraversion, the β_i are treated as random effects for each study, and the ε_{ik} are random errors for each individual. The model further assumes that the β_i are independent and identically normally distributed with mean zero and variance τ^2 , the ε_{ik} are independent and identically normally distributed with mean zero and variance σ^2 , and there is zero covariation among the β_i and ε_{ik} . This model is sometimes called the two-level hierarchical model because it allows for heterogeneity

Table 2. Results for Case Study Part I

Effect (α) or level (τ)	Summary Data I	Summary Data II	Individual-level data
α estimates			
Adult male	3.3461 (0.0175)	3.3438 (0.0178)	3.3438 (0.0178)
τ estimates			
Study	0.1049	0.1070	0.1070

Note: Values inside parentheses are estimates of standard errors. Results for the common summary-data approach (see the text) are given in the Summary Data I column, and results for the common summary-data approach using the estimate of σ^2 from the individual-level-data approach in place of the variance of the individual-level data in each study are given in the Summary Data II column.

in effect sizes across studies (Level 2) via τ^2 , as well as sampling error (Level 1) via σ^2 .

A common approach to estimation of this model involves (a) estimating τ^2 and σ^2 using REML and (b) estimating α and its standard error using the standard BLUP estimators conditional on the estimates of τ^2 and σ^2 . This estimation approach or an analogue of it is followed in all individual-level-data approaches to meta-analysis in this case study.

Results for the summary-data approach (Summary Data I column) and individual-level-data approach (Individual-Level Data column) are presented in Table 2. We note that throughout this article, we provide excess digits in our tables of results to facilitate the ability of the reader to verify the reproducibility of the results via our Supplemental Material. Overall average extraversion is estimated to be about 3.35 on the 5-point integer scale. More interesting, however, is the considerable variability in this average across the studies: The estimate of heterogeneity of about 0.10 indicates that average extraversion typically ranged between 3.15 and 3.55 in these studies (throughout this article, estimates of heterogeneity are presented as standard deviations unless otherwise noted, so that they are on the same scale as the effect estimates).

The table shows that the results for the two approaches are extremely similar but not identical. This is due to the fact that they make slightly different assumptions about the variance of the random errors of the individual-level data. Specifically, the summary-data approach assumes that this variance is known and may differ across studies, whereas the individual-level-data approach assumes that it is unknown and common across studies.

Although the assumption of known variance is necessary in the summary-data approach (i.e., because the variances of the random effects and random errors are confounded), the assumption of differing variance is

not. In fact, when the two approaches are forced to make the same assumption (e.g., by using the estimate of σ^2 from the individual-level-data approach in place of the variance of the individual-level data in each study), then they truly are equivalent and yield identical results, as indicated by the Summary Data II and Individual-Level-Data columns in the table.

Part II: two groups and thus three levels

We proceed to the data setting depicted in the second row of Table 1. Suppose a personality psychologist was interested in conducting a meta-analysis of extraversion in adult females and males. Toward this end, he gathered data from 47 studies, in all of which extraversion was measured via the simple average of the 24 extraversion items included in the IPIP-NEO-120.

The common summary-data approach to meta-analysis of these data proceeds as in Part I; however, y_i is now defined as some contrast or other statistic that collapses across the groups of subjects arising from discrete individual-level covariates such as sex (or, alternatively, across the experimental conditions arising from the manipulation of one or more discrete experimental factors).

A common choice for y_i is the standardized mean difference (or Cohen’s d), that is, the difference between the means of the two groups (or experimental conditions) divided by the pooled standard deviation. However, it is preferable in meta-analysis, as in statistical analysis more generally, to model the data on the original measurement scale when possible (Baguley, 2009; Bond, Wiitala, & Richard, 2003; Greenland, Schlesselman, & Criqui, 1986; Tukey, 1969; Wilkinson, 1999); therefore, when all studies use the same measurement scale for the dependent measure (or measures) of interest (as here), the raw mean difference is preferable to the standardized mean difference. Regardless, and as noted, this setting is the canonical setting for the common summary-data approach presented in introductory meta-analysis textbooks, and this approach is overwhelmingly dominant in practice in this setting and beyond.

Results for the summary-data approach using the standardized mean difference and raw mean difference are presented in Table 3 (Summary Data I column and Summary Data II column, respectively); for the reasons just mentioned, we discuss only the latter. Overall average extraversion is estimated to be about 0.04 lower in males than in females on the 5-point integer scale. Again, however, more interesting is the considerable variability in this difference across the studies: The estimate of heterogeneity of about 0.10 indicates that this difference typically ranged between -0.16 and 0.24 in these studies.

Table 3. Results for Case Study Part II: Common Summary-Data Approaches

Effect (α) or level (τ)	Summary Data I	Summary Data II
α estimates		
Contrast	-0.0718 (0.0325)	-0.0405 (0.0182)
τ estimates		
Study	0.1699	0.0957

Note: Values inside parentheses are estimates of standard errors. Results for the common summary-data approach using the standardized mean difference are given in the Summary Data I column, and results for the common summary-data approach using the raw mean difference are given in the Summary Data II column.

The individual-level-data approach to meta-analysis of these data involves the following. Let y_{ijk} denote the individual-level data for individual k in group (i.e., sex) j in study i . The model specification for the y_{ijk} is given by

$$y_{ijk} = \alpha_j + \beta_i + \gamma_{ij} + \epsilon_{ijk},$$

where the α_j are treated as fixed effects that model overall average extraversion for each group, the β_i are treated as random effects for each study, the γ_{ij} are treated as random effects for each study group, and the ϵ_{ijk} are random errors for each individual. The model further assumes that the β_i are independent and identically normally distributed with mean zero and variance τ_3^2 , the γ_{ij} are independent and identically normally distributed with mean zero and variance τ_2^2 , the ϵ_{ijk} are independent and identically normally distributed with mean zero and variance σ^2 , and there is zero covariation among the β_i , γ_{ij} , and ϵ_{ijk} . This model is sometimes called the three-level hierarchical model because it allows for heterogeneity in effect sizes across studies (Level 3) via τ_3^2 and heterogeneity in effect sizes across study groups within studies (Level 2) via τ_2^2 —and thus variation and covariation in effect sizes—as well as sampling error (Level 1) via σ^2 .

Before proceeding to the results, we note an important distinction between the two-level model discussed earlier and this three-level model that arises when there is, as in this setting, more than one group of subjects per study. Specifically, because the two-level model is fit to a contrast such as a mean difference, it is limited in the degree to which it can account for heterogeneity. In particular, any heterogeneity that is common across the groups in a given study cannot be accounted for, and the only heterogeneity that can be accounted for is that which is idiosyncratic to each group; in other words, heterogeneity involving between-study differences in levels is not identified, and only heterogeneity involving between-study differences in differences (i.e., contrasts) is identified. In contrast, this three-level model can

account for both—accounting for differences in levels via τ_3^2 and differences in differences via τ_2^2 —and thus is preferable when applicable, as in this setting. Given this, it is perhaps surprising that this setting is the canonical setting for the common summary-data approach presented in introductory meta-analysis textbooks and that this approach is overwhelmingly dominant in practice in this setting and beyond.

To make this concern more concrete in the context of this data setting, consider three scenarios. In the first scenario, assume that the bulk of the individual-level extraversion scores—regardless of subjects' sex—cluster toward the low part of the 5-point integer scale in some studies, toward the middle in others, and toward the high part in still others; further assume that the average difference between the two sexes is relatively stable across the studies. Thus, in this scenario, the studies tend to differ in levels but not in differences between groups; to put it differently, the heterogeneity that is common across the groups in a given study is relatively high, whereas the heterogeneity that is idiosyncratic to each group is relatively low. In this case, because the two-level model can account only for the latter but not the former, it would incorrectly assess heterogeneity to be quite low—despite the fact that the individual-level extraversion scores differ considerably across studies.

Next, consider a second scenario that is the opposite of the first scenario. Specifically, assume that the bulk of the individual-level extraversion scores—regardless of subjects' sex—tend to cluster toward the same part of the 5-point integer scale (say, the middle) in all the studies; further assume that the average difference between the two sexes is rather different across the studies. Thus, in this scenario, the studies tend to differ in differences between groups but not in levels; to put it differently, the heterogeneity that is common across the groups in a given study is relatively low, whereas the heterogeneity that is idiosyncratic to each group is relatively high. In this case, because the two-level model can account for the latter, it would correctly assess heterogeneity to be quite high.

Finally, consider a third scenario that combines the first and second scenarios. Specifically, assume that the individual-level extraversion scores—regardless of subjects' sex—cluster toward the low part of the 5-point integer scale in some studies, toward the middle in others, and toward the high part in still others; further assume that the average difference between the two sexes is rather different across the studies. Thus, in this scenario, the studies tend to differ both in levels and in differences between groups; to put it differently, both the heterogeneity that is common across the groups in a given study and the heterogeneity that is idiosyncratic to each group are relatively high. In this case, because

Table 4. Results for Case Study Part II: Richer Approaches

Effect (α) or level (τ)	Summary Data III	Individual-level data
α estimates		
Adult female	3.3864 (0.0163)	3.3870 (0.0165)
Adult male	3.3470 (0.0165)	3.3438 (0.0165)
τ estimates		
Study	0.0703	0.0697
Study group	0.0673	0.0688

Note: Values inside parentheses are estimates of standard errors. Results for the richer summary-data approach (see the text) are given in the Summary Data III column.

the two-level model can account only for the latter but not the former, it would incorrectly assess heterogeneity to be substantially lower than it is.

In sum, the two-level model can only partially account for heterogeneity in these three scenarios. In contrast, because the three-level model can account for both heterogeneity that is common across the groups and heterogeneity that is idiosyncratic to each group, it can fully account for heterogeneity and thus would correctly assess it in all three scenarios.

Results for the individual-level-data approach in this part of the case study are presented in Table 4 (Individual-Level Data column). Overall average extraversion is again estimated to be about 0.04 lower in males than in females. More interesting is the considerable heterogeneity. In particular, the total heterogeneity from one group of subjects in one study to another group of subjects in another study is estimated to be about 0.10 (i.e., $\sqrt{0.0697^2 + 0.0688^2}$). Further, about half (i.e., $0.0697^2 / (0.0697^2 + 0.0688^2)$) of this heterogeneity is common to the groups within a given study. Thus, we have gleaned something from the individual-level-data approach that we could not have gleaned from the common summary-data approach, namely, that there is not only variation but also covariation in effect sizes: If the males in a given study tended to be more extraverted than the males in other studies, the females in that study also tended to be more extraverted than the females in other studies.

Before proceeding, we note that it is simply a coincidence that the estimate of total heterogeneity obtained via the individual-level-data approach to these data is similar to that obtained via the common summary-data approach. In general, this will not be the case. However, the estimate of study-group heterogeneity obtained via the individual-level-data approach will, in general, be half the estimate of heterogeneity obtained via the common summary-data approach when presented as a variance and when applied to a simple contrast between two groups, as we have done here (i.e., $0.0688^2 \approx$

0.0957²/2; strict equality would hold but for the slightly different assumptions that the two approaches make about the variance of the random errors); more generally, this relationship is determined by the specific contrast vector employed.

The more extensive results provided by the individual-level-data approach illustrate how it can be fruitful for meta-analysts to act as if they possess the individual-level data and consider what model specifications they might fit even when they possess only summary data. For example, such a meta-analyst might seek to mimic the individual-level-data approach via a richer and more appropriate summary-data approach. Specifically, let y_{ij} denote the mean of the individual-level data in group (i.e., sex) j in study i . A summary-data model specification for the y_{ij} equivalent to the individual-level-data model specification given earlier is

$$y_{ij} = \alpha_j + \beta_i + \gamma_{ij} + \varepsilon_{ij},$$

where the α_j are treated as fixed effects that model overall average extraversion for each group, the β_i are treated as random effects for each study, the γ_{ij} are treated as random effects for each study group, and the ε_{ij} are random errors for each study group. The model further assumes that the β_i are independent and identically normally distributed with mean zero and variance τ_3^2 , the γ_{ij} are independent and identically normally distributed with mean zero and variance τ_2^2 , the ε_{ij} are independent normally distributed with mean zero and variance v_{ij} , and there is zero covariation among the β_i , γ_{ij} , and ε_{ij} . This model is sometimes called the three-level meta-analytic model because it allows for heterogeneity in effect sizes across studies (Level 3) via τ_3^2 and heterogeneity in effect sizes across study groups within studies (Level 2) via τ_2^2 —and thus variation and covariation in effect sizes—as well as sampling error (Level 1) via the v_{ij} .

We note that the model just presented is the special case of our (McShane and Böckenholt, 2017) model specification that is used when all studies follow a between-subjects study design, as in the present case (i.e., the study groups are females and males, and thus the individuals within one group are distinct from the individuals in the other group). It can easily be generalized to accommodate any mix of between-subjects and within-subjects study designs (see McShane & Böckenholt, 2017, for details). An easy-to-use website that implements this model and that is discussed in our Supplemental Material is available at <https://blakemcshane.shinyapps.io/spmeta/>.

Results for this richer summary-data approach are presented in Table 4 (Summary Data III column). As

the table shows, the results are for all intents and purposes identical to those obtained via the individual-level approach (they would be strictly identical but for the slightly different assumptions the two approaches make about the variance of the random errors).

Part III: several groups

We proceed to the data setting depicted in the third row of Table 1. Suppose a personality psychologist was interested in conducting a meta-analysis of extraversion in college-age and adult females and males. Toward this end, he gathered data from 47 studies, in all of which extraversion was measured via the simple average of the 24 extraversion items included in the IPIP-NEO-120.

With more than two groups, there are multiple potential effects of interest (e.g., in this case, the simple contrasts between all pairs of the four groups, the main effects of sex and age, and the interaction effect). As a result, the common summary-data approach to meta-analysis of these data is even more problematic than in Part II. As in that setting, study heterogeneity is not identified with this approach because of the differencing involved in forming a contrast. However, in addition, the common summary-data approach is suited for analyzing only a single effect of interest. To analyze multiple effects of interest, it can be applied to each one separately. However, this is problematic because (a) it falsely assumes independence of the effects and thus, *inter alia*, fails to provide estimates of the covariance of effect-size estimates, (b) it is statistically inefficient because each analysis makes use of only a subset of the data, and (c) it yields estimates of heterogeneity that have poor statistical properties and that can vary across effects. Consequently, we do not consider it.

Instead, we consider the richer summary-data approach discussed in Part II, which proceeds as there (except that there are now four study groups rather than two).

Because of the fact that the richer summary-data approach and the individual-level-data approach yield essentially identical results (differing only because of the slightly different assumptions the two approaches make about the variance of the random errors), we do not present results for the individual-level-data approach for the remainder of our case study; however, those results can be reproduced using the data and code in our Supplemental Material. We note that, just as does the richer summary-data approach, the individual-level-data approach proceeds as in Part II (except that there are now four study groups rather than two).

Results for the richer summary-data approach are presented in Table 5. Overall average extraversion is

Table 5. Results for Case Study Part III

Effect (α) or level (τ)	Estimate
α estimates	
Adult female	3.3866 (0.0165)
Adult male	3.3471 (0.0166)
College female	3.3735 (0.0173)
College male	3.2903 (0.0180)
τ estimates	
Study	0.0639
Study group	0.0747

Note: Values inside parentheses are estimates of standard errors.

estimated to be highest in adult and college-age females, next highest in adult males, and lowest in college-age males. Again, however, more interesting is the considerable heterogeneity. In particular, the total heterogeneity from one group of subjects in one study to another group of subjects in another study is estimated to be about 0.10 (i.e., $\sqrt{0.0639^2 + 0.0747^2}$). Further, about 40% (i.e., $0.0639^2 / (0.0639^2 + 0.0747^2)$) of this heterogeneity is common to the groups within a given study. Again, if a given group of subjects tended to be more extraverted than the same group in other studies, the other groups in that study also tended to be more extraverted than the corresponding groups in other studies.

Part IV: study-level covariates

We proceed to the data setting depicted in the fourth row of Table 1. Suppose a personality psychologist was interested in studying how extraversion varied in college-age and adult females and males across different countries. Toward this end, he gathered data from studies conducted in 47 countries, in all of which extraversion was measured via the simple average of the 24 extraversion items included in the IPIP-NEO-120. His interest centered on whether the climate of the country was associated with overall average extraversion in each of the four groups; climate was operationalized via a single study-level (or, equivalently, country-level) covariate, namely, the absolute value of the degree of latitude of the country's capital city.

The common summary-data approach to meta-analysis of these data is again problematic, for the reasons discussed in Part III, namely, because (a) study (or country) heterogeneity is not identified with this approach because of the differencing involved in forming a contrast and (b) there are multiple effects of interest. Consequently, we do not consider this approach.

The richer summary-data approach and individual-level-data approach to meta-analysis with study-level covariates proceed as in Part II but with one exception:

Table 6. Results for Case Study Part IV

Effect (α) or level (τ)	Estimate
α estimates	
Adult female	3.3535 (0.0395)
Adult male	3.4773 (0.0397)
College female	3.3399 (0.0411)
College male	3.3762 (0.0422)
Adult Female \times Latitude	0.0009 (0.0010)
Adult Male \times Latitude	-0.0037 (0.0010)
College Female \times Latitude	0.0010 (0.0011)
College Male \times Latitude	-0.0024 (0.0011)
τ estimates	
Country	0.0658
Country group	0.0653

Note: Values inside parentheses are estimates of standard errors.

The α_j are replaced by α_{ij} , which are parameterized as follows:

$$\alpha_{ij} = \alpha_{0j} + \alpha_{1j}x_{1i} + \dots + \alpha_{pj}x_{pi},$$

where x_{qi} is the value of covariate q in study i and the α_{qj} are treated as fixed effects.

Results for the richer summary-data approach are presented in Table 6. Overall average extraversion decreased among males in countries farther from the equator. Nonetheless, substantial country heterogeneity and country-group heterogeneity remained.

Part V: multiple dependent measures

We proceed to the data setting depicted in the fifth row of Table 1. Suppose a personality psychologist was interested in studying how neuroticism, extraversion, and openness varied in college-age and adult females and males across different countries. Toward this end, he gathered data from studies conducted in 47 countries, in all of which each trait was measured via the simple average of the 24 corresponding items included in the IPIP-NEO-120.

When there is more than one dependent measure, there is the possibility (indeed, the near certainty) of covariation among them. This results in the common summary-data approach to meta-analysis of these data being even more problematic than in Part II and Part III. As in those parts, study (or country) heterogeneity is not identified with this approach because of the differencing involved in forming a contrast, and there are multiple effects of interest. However, in addition, the common summary-data approach is suited for analyzing only a single dependent measure of interest. To analyze multiple dependent measures of interest, it can be

applied to each one separately. However, this is problematic because it assumes independence of the measures, and thus, *inter alia*, it fails to provide estimates of the covariance of effect-size estimates across measures, and it assumes zero covariation in effect sizes across measures. Consequently, we do not consider this approach.

A richer and more appropriate summary-data approach to meta-analysis of the three dependent measures of interest proceeds as in Part II. However, \mathbf{y}_{ij} now denotes the vector of the means of the individual-level data (i.e., neuroticism, extraversion, and openness scores) in group j in study i . Similarly, the $\boldsymbol{\alpha}_j$, $\boldsymbol{\beta}_i$, $\boldsymbol{\gamma}_{ij}$, and $\boldsymbol{\varepsilon}_{ij}$ are now vectors, and thus the $\boldsymbol{\beta}_i$ are independent and identically multivariate normally distributed with mean zero and variance-covariance matrix \mathbf{T}_3 , the $\boldsymbol{\gamma}_{ij}$ are independent and identically multivariate normally distributed with mean zero and variance-covariance matrix \mathbf{T}_2 , the $\boldsymbol{\varepsilon}_{ij}$ are independent multivariate normally distributed with mean zero and variance-covariance matrix \mathbf{V}_{ij} , and there is zero covariation among the $\boldsymbol{\beta}_i$, $\boldsymbol{\gamma}_{ij}$, and $\boldsymbol{\varepsilon}_{ij}$. Study-level (or country-level) covariates are accommodated as in Part IV; the approach taken there is applied separately to each element of the $\boldsymbol{\alpha}_j$. This model is sometimes called the three-level multivariate meta-analytic model because it allows for variation and covariation in effect sizes across studies (Level 3) via \mathbf{T}_3 and variation and covariation in effect sizes across study groups within studies (Level 2) via \mathbf{T}_2 , as well as sampling error (Level 1) via the \mathbf{V}_{ij} .

We note that this is a special case of the highly general multilevel multivariate compound-symmetry model we specified in previous work (McShane & Böckenholt, 2018b). That model introduces multilevel multivariate meta-analysis methodology that simultaneously accommodates (a) an arbitrary number of dependent measures, (b) an arbitrary number of study groups (or an arbitrary number of experimental conditions), and (c) an arbitrary number of levels that account for the variation and covariation induced by the fact that the observations are nested (e.g., within countries and country groups, as here, or within articles, studies, and study conditions, as in much experimental work). Although we do not explore them here, we note that some of the more parsimonious special cases of the multilevel multivariate compound-symmetry model specification (i.e., cases that put restrictions on the \mathbf{T}_k variance-covariance matrices; see McShane & Böckenholt, 2018b) are more appropriate in many applied settings than the unrestricted version examined here. An easy-to-use website that implements the multilevel multivariate compound-symmetry model and that is discussed in our Supplemental Material is available at <https://blakemcshane.shinyapps.io/mlmvmeta/>.

An individual-level-data approach to meta-analysis of these data proceeds as in Part II, with the model specification presented there generalized along the lines just discussed. Specifically, \mathbf{y}_{ijk} now denotes the vector of the individual-level data (i.e., neuroticism, extraversion, and openness scores) for individual k in group j in study i . Similarly, the $\boldsymbol{\alpha}_j$, $\boldsymbol{\beta}_i$, $\boldsymbol{\gamma}_{ij}$, and $\boldsymbol{\varepsilon}_{ijk}$ are now vectors, and thus the $\boldsymbol{\beta}_i$ are independent and identically multivariate normally distributed with mean zero and variance-covariance matrix \mathbf{T}_3 , the $\boldsymbol{\gamma}_{ij}$ are independent and identically multivariate normally distributed with mean zero and variance-covariance matrix \mathbf{T}_2 , the $\boldsymbol{\varepsilon}_{ijk}$ are independent and identically multivariate normally distributed with mean zero and variance-covariance matrix $\boldsymbol{\Sigma}$, and there is zero covariation among the $\boldsymbol{\beta}_i$, $\boldsymbol{\gamma}_{ij}$, and $\boldsymbol{\varepsilon}_{ijk}$. Study-level (or country-level) covariates are accommodated as in Part IV; the approach taken there is applied separately to each element of the $\boldsymbol{\alpha}_j$. This model is sometimes called the three-level multivariate hierarchical model because it allows for variation and covariation in effect sizes across studies (Level 3) via \mathbf{T}_3 and variation and covariation in effect sizes across study groups within studies (Level 2) via \mathbf{T}_2 , as well as sampling error (Level 1) via $\boldsymbol{\Sigma}$.

Results for the richer summary-data approach are presented in Tables 7 and 8. Overall average neuroticism is unrelated to distance from the equator, overall average extraversion decreases among males in countries farther from the equator, and overall average openness increases in all groups in countries farther from the equator. Nonetheless, substantial country heterogeneity and country-group heterogeneity remain. Further, although the total heterogeneity from one group of subjects in one country to another group of subjects in another country is estimated to be about 0.10 for all three dependent measures, country heterogeneity is larger than country-group heterogeneity for neuroticism and openness but roughly equal to country-group heterogeneity for extraversion. This means that across countries, the four groups covary more strongly with respect to neuroticism and openness than with respect to extraversion. In addition, the correlation of this heterogeneity is moderately positive for extraversion and openness at the country level, rather negative for neuroticism and extraversion at the country-group level, and moderately positive for neuroticism and openness at the country-group level.

Discussion

Although the summary-data approach to meta-analysis is so widespread that it is often equated with meta-analysis, the analysis of the individual-level data from the studies via a hierarchical model is considered the

Table 7. Results for Case Study Part V: Principal Estimates

Dependent measure and effect (α) or level (T)	Estimate
α estimates	
Neuroticism	
Adult female	2.8433 (0.0444)
Adult male	2.6661 (0.0443)
College female	2.9858 (0.0448)
College male	2.8143 (0.0453)
Adult Female \times Latitude	-0.0010 (0.0011)
Adult Male \times Latitude	-0.0014 (0.0011)
College Female \times Latitude	-0.0002 (0.0012)
College Male \times Latitude	-0.0015 (0.0012)
Extraversion	
Adult female	3.3538 (0.0389)
Adult male	3.4749 (0.0392)
College female	3.3440 (0.0404)
College male	3.3747 (0.0415)
Adult Female \times Latitude	0.0009 (0.0010)
Adult Male \times Latitude	-0.0036 (0.0010)
College Female \times Latitude	0.0009 (0.0010)
College Male \times Latitude	-0.0024 (0.0011)
Openness	
Adult female	3.4198 (0.0403)
Adult male	3.4306 (0.0403)
College female	3.4085 (0.0415)
College male	3.3163 (0.0415)
Adult Female \times Latitude	0.0061 (0.0010)
Adult Male \times Latitude	0.0032 (0.0010)
College Female \times Latitude	0.0053 (0.0011)
College Male \times Latitude	0.0053 (0.0011)
T (standard deviation) estimates	
Neuroticism	
Country	0.0859
Country group	0.0579
Extraversion	
Country	0.0635
Country group	0.0653
Openness	
Country	0.0888
Country group	0.0465

Note: Values inside parentheses are estimates of standard errors.

gold standard for meta-analysis. However, as we have illustrated in this article, the summary-data and individual-level-data approaches are, for a wide variety of cases common in meta-analysis, equivalent when the underlying models are appropriately specified.

This equivalence is due to the fact that the mean and variance are sufficient statistics for the normal distribution. Consequently, this equivalence holds more broadly than in the examples presented here. For example, when studies employ the same measurement scale for

Table 8. Results for Case Study Part V: **T** Correlation Matrix Estimates

Trait	Neuroticism	Extraversion	Openness
Country			
Neuroticism	1.0000	.0893	.0141
Extraversion	.0893	1.0000	.2787
Openness	.0141	.2787	1.0000
Country group			
Neuroticism	1.0000	-.5544	.3624
Extraversion	-.5544	1.0000	.1400
Openness	.3624	.1400	1.0000

the dependent measure (or measures) of interest and interest centers on one or more contrasts of means of groups of subjects (or experimental conditions), the two approaches are equivalent regardless of (a) the number of groups of subjects (or the number of experimental conditions), (b) the number of levels in the nesting structure of the data, (c) the number of dependent measures, and (d) the number of discrete or continuous study-level covariates; it is only when continuous covariates at the individual level are of interest that the two approaches are no longer equivalent and the individual-level data are necessary. Similar considerations hold also, for example, in the meta-analysis of regression coefficients.

Given this equivalence, we have proposed that it can be extremely fruitful for meta-analysts to act as if they possess the individual-level data and consider what model specifications they might fit even when they possess only summary data. This thought experiment can lead them to move away from the common summary-data approach to meta-analysis—that is, the basic random-effects meta-analytic model that is overwhelmingly dominant in practice—and toward richer and more appropriate summary-data approaches when the complexity of the data warrants it.

Specifically, the basic random-effects meta-analytic model is a univariate, two-level model. Consequently, it is suitable only when there is a single group of subjects (or a single experimental condition), a single dependent measure, and a single effect of interest in each study. This is seldom the case in contemporary psychological research studies, and when there is more than one of any of these, the approach can be problematic, and more extensive results can be obtained via richer and more appropriate summary-data approaches.

Our thought experiment can be beneficial with respect to the four principal dimensions along which these richer and more appropriate summary-data approaches generalize the common summary-data approach and across which our case study evolved,

namely, (a) the number of study groups (or experimental conditions), (b) the number of levels, (c) the number of dependent measures, and (d) study-level covariates. For example, in Part II of our case study, the thought experiment made it clear that there were three levels in the nesting structure and therefore the common summary-data approach was inadequate—even though this is the canonical setting for it presented in introductory meta-analysis textbooks and this approach is overwhelmingly dominant in practice in this setting and beyond. Instead, we avoided the differencing involved in forming a contrast required by the common summary-data approach and obtained, via a three-level model, more extensive results, namely, not only an estimate of heterogeneity involving differences in differences (i.e., contrasts) across studies but also an estimate of heterogeneity involving differences in levels.

Similarly, in Part III of our case study, the thought experiment made it clear that there were not only three levels in the nesting structure but also multiple potential effects of interest arising from the fact that there were four groups. Therefore, applying the common summary-data approach to each effect separately was inadequate. Instead, we analyzed the effects simultaneously via one coherent and more appropriate three-level model.

The purpose of Part IV of our case study was to illustrate a setting with study-level covariates. The thought experiment is also useful in such settings. In particular, because the thought experiment makes clear the number of levels, it brings to the fore the fact that each level can have its own set of covariates, including those aggregated from lower levels. Appreciating this can also lead to richer and more appropriate summary-data approaches that provide more extensive results. For example, insofar as it suggests as yet unconsidered discrete individual-level covariates, it could lead to a model with more groups at the second level.

Finally, in Part V of our case study, the thought experiment made it clear not only that there were three levels in the nesting structure and that there were multiple potential effects of interest arising from the fact that there were four groups, but also that there were multiple dependent measures. Therefore, applying the common summary-data approach to each effect separately was inadequate. Instead, we analyzed the effects simultaneously via one coherent and more appropriate three-level multivariate model.

In sum, the value of our thought experiment is twofold. First, it makes clear whether or not there is a single group of subjects (or a single experimental condition), a single dependent measure, and a single effect of interest in each study. Second, when there is more than one of any of these, it suggests not only that one should

move away from the common summary-data approach to meta-analysis but also how one might move away from it and toward richer and more appropriate summary-data approaches.

Two potential objections to our thought experiment are that it requires (a) the dependent measure (or measures) to be on the same scale across studies and (b) statistical sufficiency such that the summary-data and individual-level-data approaches are equivalent. We disagree and believe that the thought experiment can be fruitful when either or both of these conditions fail to hold.

First, when a dependent measure is not on the same scale in all the studies, this does not preclude the thought experiment. For example, one can still act as if one possesses the individual-level data and consider what model specifications one might fit were the dependent measure on the same scale across the studies. This is likely to move one toward richer and more appropriate summary-data approaches, even if one ultimately chooses to adjust for the differences prior to modeling (e.g., by converting the data to a standardized scale, such as Cohen's d). Further, the thought experiment could lead one to consider how one might adjust for these differences within the context of a model for the individual-level data (i.e., rather than prior to modeling), which in turn might lead one to consider an analogous approach that adjusts for these differences within the context of a model for the summary data. For instance, consider Part III of our case study and suppose that the dependent measure had not been on the same scale in all the studies; the thought experiment might still move one away from applying the common summary-data approach to each of the multiple potential effects of interest separately and toward approaches that analyze them simultaneously (see, e.g., Gleser & Olkin, 1994, 2009, for an approach that does so by adjusting for the differences prior to modeling and McShane & Böckenholt, 2018a, for an approach that does so by adjusting for the differences within the context of a model).

Second, we argue that our thought experiment can prove extremely fruitful when the equivalence between the summary-data and individual-level-data approaches does not strictly hold. For example, consider a meta-analysis in which the dependent measure is binary. Meta-analysts who possess the individual-level data would likely fit some form of a hierarchical generalized linear model, such as a hierarchical logistic regression. In contrast, meta-analysts who possess summary data such as proportions or odds ratios would likely fit some form of a hierarchical normal model, as discussed in this article. We argue that it would be far preferable to seek to mimic the former approach in the context of a similarly specified hierarchical normal model insofar as

possible, rather than to apply the common summary-data approach; indeed, doing so would offer many of the same benefits demonstrated in this article.

In addition, the thought experiment is likely to make it clear that possessing the summary data for a binary dependent measure is—at least sometimes—actually equivalent to possessing the individual-level data. For example, often when one possesses proportions and sample sizes, one can multiply them to obtain counts and thus re-create the individual-level data. Consequently, in such cases, the thought experiment can lead one to move away from the common summary-data approach to meta-analysis and toward richer and more appropriate individual-level-data approaches even when one possesses only summary data.

In conclusion, although it is our hope that our thought experiment will lead to increased use of richer and more appropriate summary-data approaches to meta-analysis when the complexity of the data warrants it, we by no means wish to suggest that possessing summary data is generally equivalent to or as advantageous as possessing individual-level data. Although they may be equivalent for the purpose of conceptualizing the model specification, individual-level data offer numerous advantages over summary data. For example, individual-level data allow for the evaluation of distributional, functional-form, and other model-specification assumptions; the investigation and imputation of missing data; and the analysis of continuous covariates at the individual level. Consequently, the analysis of individual-level data remains the gold standard for meta-analysis.

Transparency

Action Editor: Frederick L. Oswald

Editor: Daniel J. Simons

Author Contributions

B. B. McShane wrote the analysis code and analyzed the data. U. Böckenholt verified the accuracy of the analyses. B. B. McShane wrote the first draft of the manuscript. Both authors critically edited the manuscript and approved the final version for submission.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Open Practices

Open Data: <https://osf.io/ua9h4/>

Open Materials: <https://osf.io/ua9h4/>


Preregistration: not applicable

All data and materials have been made publicly available at the Open Science Framework and can be accessed at <https://osf.io/ua9h4/>. The complete Open Practices Disclosure for this article can be found at <http://journals.sagepub.com/doi/suppl/10.1177/2515245919884304>. This article has received badges for Open Data and Open Materials. More information about the Open Practices badges

can be found at <http://www.psychologicalscience.org/publications/badges>.



ORCID iD

Blakeley B. McShane  <https://orcid.org/0000-0002-4839-266X>

Acknowledgments

This article is dedicated to Murray McShane.

Supplemental Material

Additional supporting information can be found at <http://journals.sagepub.com/doi/suppl/10.1177/2515245919884304>

References

- Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology*, *100*, 603–617.
- Becker, B. J. (2000). Multivariate meta-analysis. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 499–526). San Diego, CA: Academic Press.
- Berkey, C. S., Hoaglin, D. C., Antczak-Bouckoms, A., Mosteller, F., & Colditz, G. A. (1998). Meta-analysis of multiple outcomes by regression with random effects. *Statistics in Medicine*, *17*, 2537–2550.
- Bond, C. F., Jr., Wiitala, W. L., & Richard, F. D. (2003). Meta-analysis of raw mean differences. *Psychological Methods*, *8*, 406–418.
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, England: Wiley.
- Cheung, M. W.-L. (2015). *Meta-analysis: A structural equation modeling approach*. Chichester, England: Wiley.
- Cooper, H., & Hedges, L. V. (Eds.). (1994). *The handbook of research synthesis*. New York, NY: Russell Sage Foundation.
- Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York, NY: Russell Sage Foundation.
- Cooper, H., & Patall, E. A. (2009). The relative benefits of meta-analysis conducted with individual participant data versus aggregated data. *Psychological Methods*, *14*, 165–176.
- Gleser, L. J., & Olkin, I. (1994). Stochastically dependent effect sizes. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 339–355). New York, NY: Russell Sage Foundation.
- Gleser, L. J., & Olkin, I. (2009). Stochastically dependent effect sizes. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 357–376). New York, NY: Russell Sage Foundation.
- Greenland, S., Schlesselman, J. J., & Criqui, M. H. (1986). The fallacy of employing standardized regression coefficients and correlations as measures of effect. *American Journal of Epidemiology*, *123*, 203–208.

- Haidich, A. (2010). Meta-analysis in medical research. *Hippokratia*, *14*, 29–37.
- Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, *72*, 320–338.
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.
- Johnson, J. A. (2014). Measuring thirty facets of the Five Factor Model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *Journal of Research in Personality*, *51*, 78–89.
- Kalaian, H. A., & Raudenbush, S. W. (1996). A multivariate mixed linear model for meta-analysis. *Psychological Methods*, *1*, 227–235.
- McShane, B. B., & Böckenholt, U. (2017). Single paper meta-analysis: Benefits for study summary, theory-testing, and replicability. *Journal of Consumer Research*, *43*, 1048–1063.
- McShane, B. B., & Böckenholt, U. (2018a). *Meta-analysis of studies with multiple effects of interest and differences in measurement scales: Generalizing the Cohen's d approach*. Unpublished manuscript, Northwestern University, Kellogg School of Management, Marketing Department.
- McShane, B. B., & Böckenholt, U. (2018b). Multilevel multivariate meta-analysis with application to choice overload. *Psychometrika*, *83*, 255–271.
- McShane, B. B., Tackett, J. L., Böckenholt, U., & Gelman, A. (2019). Large-scale replication projects in contemporary psychological research. *The American Statistician*, *73*(Suppl. 1), 99–105.
- Raudenbush, S. W., Becker, B. J., & Kalaian, H. A. (1988). Modeling multivariate effect sizes. *Psychological Bulletin*, *103*, 111–120.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.
- Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects. *Statistical Science*, *6*, 15–32.
- Rosenthal, R., & Rubin, D. B. (1986). Meta-analytic procedures for combining studies with multiple effect sizes. *Psychological Bulletin*, *99*, 400–406.
- Simmonds, M. C., Higgins, J. P., Stewart, L. A., Tierney, J. F., Clarke, M. J., & Thompson, S. G. (2005). Meta-analysis of individual patient data from randomized trials: A review of methods used in practice. *Clinical Trials*, *2*, 209–217.
- Stanley, T., Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin*, *144*, 1325–1346.
- Stewart, L. A., & Tierney, J. F. (2002). To IPD or not to IPD? Advantages and disadvantages of systematic reviews using individual patient data. *Evaluation & the Health Professions*, *25*, 76–97.
- Tipton, E., Pustejovsky, J. E., & Ahmadi, H. (2019). Current practices in meta-regression in psychology, education, and medicine. *Research Synthesis Methods*, *10*, 180–194.
- Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist*, *24*, 83–91.
- van Erp, S., Verhagen, J., Grasman, R. P. P. P., & Wagenmakers, E.-J. (2017). Estimates of between-study heterogeneity for 705 meta-analyses reported in *Psychological Bulletin* from 1990–2013. *Journal of Open Psychology Data*, *5*, Article 4. doi:10.5334/jopd.33
- Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604.