Single-Paper Meta-Analysis: Benefits for Study Summary, Theory Testing, and Replicability

BLAKELEY B. MCSHANE
ULF BÖCKENHOLT

A typical behavioral research paper features multiple studies of a common phenomenon that are analyzed solely in isolation. Because the studies are of a common phenomenon, this practice is inefficient and forgoes important benefits that can be obtained only by analyzing them jointly in a single-paper meta-analysis (SPM). To facilitate SPM, we introduce meta-analytic methodology that is user-friendly, widely applicable, and specially tailored to the SPM of the set of studies that appear in a typical behavioral research paper. Our SPM methodology provides important benefits for study summary, theory testing, and replicability that we illustrate via three case studies that include papers recently published in the *Journal of Consumer Research* and the *Journal of Marketing Research*. We advocate that authors of typical behavioral research papers use it to supplement the single-study analyses that independently examine the multiple studies in the body of their papers as well as the "qualitative meta-analysis" that verbally synthesizes the studies in the general discussion of their papers. When used as such, this requires only a minor modification of current practice. We provide an easy-to-use website that implements our SPM methodology.

Keywords: meta-analysis, between-study variation, heterogeneity, hierarchical, multilevel, random effects

Meta-analysis is a well-established statistical technique that synthesizes two or more studies of a common phenomenon. Because multiple studies provide more information about the common phenomenon than any single one of them, meta-analysis can offer a number of benefits. For example, insofar as the studies measure the common phenomenon with some degree of error, a meta-analysis, which pools the results from the studies via a weighted average, will yield an estimate that is on average more accurate than that of any individual study. In addition, the uncertainty in the meta-analytic estimate will typically be smaller than the uncertainty in the estimate of any individual study, thereby, *inter alia*, increasing statistical power relative to individual studies and providing a means of resolution when individual studies yield so-called conflicting results. Further, meta-analysis allows for the investigation of differences among studies, for example by quantifying the impact of study-level covariates or the degree of between-study variation.

These benefits have been widely realized in behavioral research in traditional meta-analyses of studies that appear in multiple papers. However, they have only very seldom been realized in meta-analyses of studies that appear in a single paper. Indeed, a typical behavioral research paper features multiple studies of a common phenomenon that are analyzed solely in isolation. Because the studies are of a common phenomenon, this practice is inefficient and

Blakeley B. McShane (b-mcshane@kellogg.northwestern.edu) is an associate professor of marketing and Ulf Böckenholt (u-bockenholt@kellogg.northwestern.edu) is the John D. Gray Professor of Marketing, both at the Kellogg School of Management, Northwestern University, Evanston, IL 60208. Correspondence concerning this article should be addressed to Blakeley B. McShane. Supplementary materials for this article are available online.

This tutorial was invited by editors Darren Dahl, Eileen Fischer, Gita Johar, and Vicki Morwitz but was peer-reviewed by editor Vicki Morwitz, associate editor Praveen Kopalle, and three anonymous reviewers.

Advance Access publication January 23, 2017

• Vol. 43 • 2017
DOI: 10.1093/jcr/ucw085

forgoes important benefits that can be obtained only by analyzing them jointly in a single-paper meta-analysis (SPM).

Authors contemplating conducting an SPM can choose from among a host of meta-analytic methodological options that, although originally developed for the traditional meta-analysis of studies that appear in multiple papers, could also be used for SPM. For instance, they might consider the approaches discussed by Rosenthal (1978), such as Fisher's method of adding log *p*s (Fisher 1925, 1948) and Stouffer's method of adding *z*s (Mosteller and Bush 1954; Stouffer et al. 1949). They might also consider the standardized effect approach, which involves converting their observed effects to a common standardized scale such as the Cohen's *d* scale and modeling the standardized effects via basic meta-analytic models (see, for example, Borenstein et al. 2009).

In this paper, we introduce meta-analytic methodology that—in contradistinction to these approaches—is specially tailored to the SPM of the set of studies that appear in a typical behavioral research paper. Our SPM methodology is user-friendly because it requires only basic summary information (e.g., means, standard deviations, and sample sizes); importantly, despite requiring only this basic summary information, the model underlying our SPM methodology is equivalent, by a principle known as statistical sufficiency, to that underlying the "gold standard" meta-analytic approach—namely, an appropriately specified hierarchical (or multilevel) model fit to the individual-level observations (Cooper and Patall 2009; Haidich 2010; Simmonds et al. 2005; Stewart and Tierney 2002). In addition, our SPM methodology is widely applicable; indeed, a literature review reveals that it could have been used in 86% of the behavioral research papers published in the three most recent volumes of the *Journal of Consumer Research* (volumes 40–42).

Our SPM methodology provides important benefits for study summary, theory testing, and replicability that we illustrate via three case studies that include papers recently published in the *Journal of Consumer Research* and the *Journal of Marketing Research* and that, as we further note in our discussion, are either not provided by or are provided only in part by alternative approaches. Specifically, our SPM methodology provides a graphical and quantitative summary of the studies. The graphical summary facilitates the communication and comparison of results within and across studies, thus simplifying assessments of convergence, while the quantitative summary provides a more precise estimate of each effect of interest as well as the uncertainty in this estimate. This increased precision is important for theory testing because it allows for more powerful tests of posited effects. Further, these more powerful tests can deepen theory testing by motivating new decompositions of the effects that investigate alternative explanations. Additionally, our SPM methodology provides an estimate of and accounts for between-study variation. This estimate of between-study variation can suggest unaccounted-for moderators that have the potential to enrich theory, while accounting for between-study variation improves calibration of Type I and Type II error. Finally, our SPM methodology provides sample size analyses for future studies and future sets of studies that account for the uncertainty associated with effect estimates as well as between-study variation, thus enhancing replicability.

Our SPM methodology has two additional benefits. First, because it requires only basic summary information and this information is often reported in papers, it allows readers as well as authors to conduct an SPM and obtain the benefits for study summary, theory testing, and replicability discussed above. Second, because the reporting of an SPM is extremely concise, it allows authors to, if they desire, include in the SPM studies they have that are related to those reported in the paper but which themselves were not reported in the paper; this allows authors to provide further evidence about the phenomenon of interest without taking up a great deal of journal space and thus should enhance replicability.

Because our SPM methodology is user-friendly, is widely applicable, and provides these manifold benefits, we advocate that authors of typical behavioral research papers use it to supplement the single-study analyses that independently examine the multiple studies in the body of their papers as well as the "qualitative meta-analysis" that verbally synthesizes the studies in the general discussion of their papers. When used as such, this requires only a minor modification of current practice.

To facilitate this, we provide an easy-to-use website that implements our SPM methodology (http://www.singlepapermetaanalysis.com/). It includes a detailed tutorial that shows how to replicate the case studies presented in this paper and how to apply it to new papers.

In the remainder of this paper, we discuss the prevalence of between-study variation in behavioral research studies–a topic that has been largely overlooked to date–and the implications of this for SPM. We then describe the model underlying our SPM methodology, illustrate it via three case studies, and discuss the sample size analyses it provides. Finally, we summarize the principal benefits of our SPM methodology, remark on some of its features, and conclude with a brief discussion.

HETEROGENEITY IN BEHAVIORAL RESEARCH STUDIES

The purpose of meta-analysis is to synthesize a set of studies of a common phenomenon. This task is complicated in behavioral research by the fact that behavioral research studies can never be direct or exact replications of one another (Brandt et al. 2014; Fabrigar and Wegener 2016; Rosenthal 1991; Stroebe and

Strack 2014; Tsang and Kwan 1999). Instead, studies differ at minimum in their method factors—that is, anything pertaining to the implementation of the study that is not directly related to the theory under investigation. Method factors can include seemingly major factors such as the operationalization of the dependent measure, the operationalization of the experimental manipulation(s), or unaccounted-for moderators, but also seemingly minor factors such as the social context, the subject pool, or the time of day (for a comprehensive list, see Brown et al. 2014). Differences in method factors result in true effects that vary from study to study (i.e., between-study variation also known as heterogeneity), necessitating hierarchical meta-analytic models that account for this by decomposing the variation in observed effects into a between-study component that results from differences in method factors and a within-study component that results from sampling error (i.e., measuring only a subset, or sample, of the population; Borenstein et al. 2009; Cooper, Hedges, and Valentine 2009; Hedges and Olkin 1985; Hunter and Schmidt 2014).

While accounting for heterogeneity has long been regarded as important in meta-analyses of sets of studies that consist of general (i.e., systematic or conceptual) replications, there is mounting evidence and growing appreciation that this is also the case in meta-analyses of sets of studies that consist entirely of close replications (i.e., studies that use identical or similar materials). For example, consider the Many Labs project, which provides 16 estimates of 13 classic and contemporary behavioral research effects from 36 independent samples totaling 6,344 subjects (Klein et al. 2014). Each of the 36 laboratories involved in the project used identical materials, and these materials were administered through a web browser in order to minimize the effect of laboratory-specific method factors (i.e., the studies were close replications of one another). Nonetheless, meta-analyses of these studies conducted by the Many Labs authors yielded nonzero estimates of heterogeneity for all 14 of the effects they found to be non-null; further, 40% of the total variability on average across these effects was due to heterogeneity resulting from laboratory-specific method factors (Klein et al. 2014, table 3).

Among the 6,344 Many Labs subjects were 1,000 recruited via Amazon.com's Mechanical Turk. The study materials were administered to these 1,000 subjects over seven unique days beginning on August 29, 2013, and ending on September 11, 2013 (i.e., seven consecutive days excluding Fridays, weekends, and the Labor Day holiday). Restricting attention to only these subjects and treating each unique day as a separate sample yields seven extremely close replications of each effect. Again, however, despite the extreme degree of closeness, heterogeneity is nontrivial: meta-analyses yield nonzero estimates of heterogeneity for nine of the 14 non-null effects, and across these effects 21% of the total variability on average was due to heterogeneity resulting from method factors.

Given the degree of heterogeneity present in the Many Labs studies (where the only difference among the studies was the location of the laboratory) and in the Mechanical Turk subsample of these studies (where the only difference among the studies was the day on which the study materials were administered), it seems reasonable to conclude that some degree of heterogeneity is likely to be present in much behavioral research, as more typical sets of studies—even those that appear in a single paper—tend to have more variation in their method factors than do the Many Labs studies (e.g., they seldom use identical materials). This suggests it is critical to account for heterogeneity when analyzing behavioral research studies. Importantly, this is not possible when studies are analyzed solely in isolation but is possible when they are analyzed jointly via hierarchical meta-analytic models.

MODEL DESCRIPTION

Our SPM methodology is specially tailored to account for the complex patterns of variation and covariation among the observations of a dependent measure from a set of studies that appear in a typical behavioral research paper. In particular, it extends prior hierarchical meta-analytic models to accommodate (i) an arbitrary number of study conditions that result from the variation of one or more experimental factors and give rise to multiple dependent effects of interest (e.g., simple effects and interaction effects); (ii) a mix of study designs (e.g., two-condition vs. 2 × 2, between-subjects vs. within-subjects); (iii) the variation (or heterogeneity) resulting from differences in method factors among the observations; and (iv) the covariation induced by the fact that the observations are nested within studies and study conditions (or, when one or more studies follow a within-subjects design, nested within studies, subject groups, and study conditions).

To achieve this, our SPM methodology decomposes each observation—that is, a statistic (e.g., the mean, the proportion of successes) that summarizes the individual-level observations in each condition of each study—into three components: (i) an overall condition average component, (ii) a study-condition component that reflects method factors, and (iii) a study-condition component that reflects sampling error. It allows the second component to reflect method factors specific to each study as a whole as well as to each condition of each study (and, when one or more studies follow a within-subjects design, to each subject group). It assumes method factors and sampling error operate independently; it further assumes they have zero mean as the mean is captured by the overall condition average components.

As noted in our introduction, despite requiring only basic summary information, the model underlying our

SPM methodology is equivalent to that underlying the gold-standard meta-analytic approach. Specifically, the same results are obtained when our SPM methodology is fit to the basic summary information and when an appropriately specified hierarchical model (i.e., one that accommodates the four features discussed in the opening paragraph of this section) is fit to the individual-level observations. This equivalence—and the widespread applicability of these equivalent models for the meta-analysis of the set of studies that appear in typical behavioral research papers—has not been noted or exploited in the meta-analytic literature.

We refer the reader interested in the full details regarding our SPM model specification and estimation procedure to our supplementary materials.

CASE STUDIES

In this section, we illustrate our SPM methodology by applying it to two recently published papers (Maimaran and Fishbach 2014; Shah et al. 2014) as well as one yet-to-be-published work. Each paper is rather different in terms of the study designs, the effects of interest, and the results. Consequently, we begin with the most straightforward paper and work toward more complicated ones. Our focus here is on describing the input data required by our SPM methodology (and website), interpreting the SPM estimates relative to the single-study estimates, and demonstrating the additional benefits the SPM provides relative to the single-study analyses. After doing so, we offer a summary of the SPM that would add to the general discussion of the papers.

The first case study serves primarily as an example of our SPM methodology in a setting that is simple both theoretically and empirically. Nonetheless, even in this simple setting the SPM provides benefits: it yields better estimates of the effects of interest and the uncertainty in them as well as an estimate of heterogeneity. The second case study shows how the more precise estimates provided by the SPM can deepen theory testing. In particular, the SPM detects effects that the single-study analyses did not, thereby motivating new decompositions that investigate alternative explanations. Finally, the third case study shows that the estimate of heterogeneity provided by the SPM can suggest unaccounted-for moderators that drive this heterogeneity and have the potential to enrich theory.

We selected Maimaran and Fishbach (2014) and Shah et al. (2014) because they were exemplary in their data reporting practices. In particular, Maimaran and Fishbach (2014) report not only the mean but also the standard deviation of the dependent measure for all conditions of all studies in one concise table (Maimaran and Fishbach 2014, table 2). Similarly, Shah et al. (2014) report the proportion of successes for all conditions of all studies in tables throughout the text (Shah et al. 2014, tables 1–4). Total sample sizes for all studies were also clearly reported by both sets of authors, and in our analysis we make the further assumption that the subjects were split evenly across the conditions of each study; this assumption is not required by our SPM methodology but is not unreasonable to employ when, as here and as is typical, the sample size of each condition of each study was not reported.

Case Study I: Maimaran and Fishbach (2014)

Maimaran and Fishbach (2014) "propose[d] that preschoolers infer that if food is instrumental to achieve a goal, it is less tasty, and therefore they consume less of it." They tested this hypothesis across five studies that used similar stimuli.[1] The primary dependent measure in these studies was food consumption measured in morsels, and the primary hypothesis concerned the difference between the control and instrumental conditions; a secondary hypothesis motivated by practical parental considerations concerned the difference between the control and yummy conditions.

We present the summary information for the series of studies conducted by Maimaran and Fishbach (2014) in table 1, which serves as the primary input data for our SPM methodology. The contrasts studied by the authors are given by (1 –1 0) for the control versus instrumental

TABLE 1

MAIMARAN AND FISHBACH (2014) SUMMARY INFORMATION

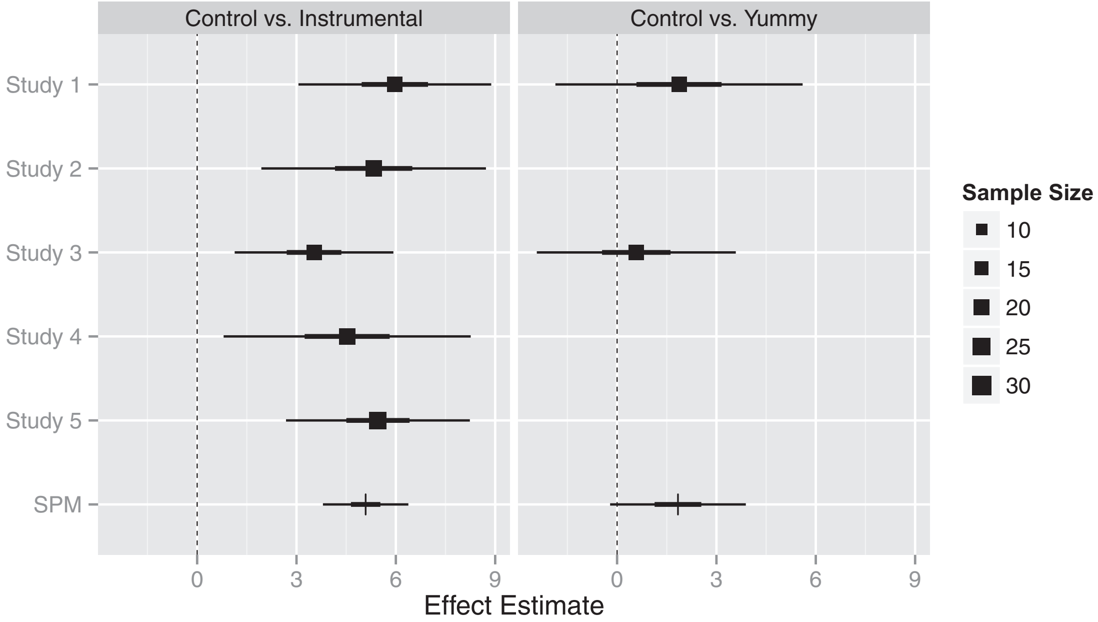
Study	Factor 1	$\bar{x}$	s	n
Study 1	Control	9.07	5.60	19
	Instrumental	3.10	3.25	19
	Yummy	7.20	6.13	19
Study 2	Control	10.00	5.93	22
	Instrumental	4.67	5.54	22
	Yummy			
Study 3	Control	7.11	4.77	19
	Instrumental	3.58	2.38	19
	Yummy	6.53	4.68	19
Study 4	Control	8.14	7.40	21
	Instrumental	3.61	4.62	21
	Yummy			
Study 5	Control	10.78	4.80	24
	Instrumental	5.32	5.01	24
	Yummy			

NOTE.—This table reproduces table 2 of Maimaran and Fishbach (2014) except the measurements for study 4 have been converted from grams to morsels using the conversion rate of 42 grams per 20 morsels reported in Maimaran and Fishbach (2014).

[1] Studies 1 and 3 followed a three-condition design (control vs. instrumental vs. yummy), while studies 2, 4, and 5 followed a two-condition design (control vs. instrumental). All studies followed a between-subjects design.

FIGURE 1

RESULTS FROM MAIMARAN AND FISHBACH (2014)



NOTE.—Effect estimates are given by the squares for single-study estimates and the vertical bars for SPM estimates; 50% and 95% intervals are given by the thick and thin lines, respectively. The average sample size per condition in each study is given by the size of the squares. The SPM estimates are much more precise thereby giving greater support to the authors' conclusions.

contrast and (1 0 –1) for the control versus yummy contrast.

We present the single-study estimates from Maimaran and Fishbach (2014) in figure 1. The point estimates are given by the squares, and 50% and 95% intervals are given by the thick and thin lines, respectively. The figure can be thought of, *inter alia*, as providing a graphical t-test where 95% intervals that overlap the dashed vertical line at zero represent a failure to reject the null hypothesis significance test of zero effect.

The figure is consistent with the theory presented in Maimaran and Fishbach (2014) that subjects in the control condition consume more than those in the instrumental condition. In particular, all five single-study estimates of the control versus instrumental contrast are positive and none of the intervals overlap zero; in other words, all five estimates are positive and attain statistical significance. Similarly, both estimates of the control versus yummy contrast overlap zero; in other words, both fail to attain statistical significance. In sum, these estimates and intervals depict graphically the results of the null hypothesis significance tests reported in the text of Maimaran and Fishbach (2014).

We also present the SPM estimates in figure 1. The estimates are given by the vertical bars, and again 50% and 95% intervals are given by the thick and thin lines, respectively. By combining information across studies, the SPM estimates have uncertainty intervals that are much narrower than those of the single-study estimates. Consequently, one can be more confident in the estimates—both in terms of size and sign—and thus in the authors' conclusion that subjects in the control condition consume more than those in the instrumental condition and about the same amount as those in the yummy condition. In particular, the narrower uncertainty interval implies the SPM provides a more powerful test of the posited null effect of the control versus yummy contrast as compared to any single study and demonstrates that any effect is likely to be small even if nonzero.

The single-study estimates displayed in figure 1 vary from 3.53 to 5.97 morsels for the control versus instrumental contrast and .58 to 1.87 morsels for control versus yummy contrast. The amount of this variation that can be attributed to sampling error versus method factors is quantified by the SPM heterogeneity estimate, which is .58 on the variance scale (or $\sqrt{.58} = .76$ on the standard deviation scale). While this can be interpreted by comparing it to the standard deviations and sample sizes displayed in table 1, a statistical measure known as I^2 (Higgins et al. 2003; Higgins and Thompson 2002)—which gives the percentage of the variation in the observations (beyond that attributable to the experimental manipulations) that is due to heterogeneity—can sometimes provide a cleaner interpretation. I^2 is estimated at 19%, suggesting that method factors account for about one-fifth of the variation in the observations beyond that attributable to the experimental manipulations. To put this in perspective, Pigott (2012) provides guidance on the typical size of I^2 in behavioral research; in particular, she defines low heterogeneity as $I^2 = 25\%$, medium heterogeneity as $I^2 = 50\%$, and high heterogeneity as $I^2 = 75\%$ (see also Higgins and Thompson 2002). Thus, according to this schema, heterogeneity is estimated to be low in these studies; this is unsurprising, as the five studies were reasonably close replications of one another (e.g., they used similar stimuli and were conducted in the same preschool). However, the uncertainty interval for I^2 is (0%–59%), suggesting the data are consistent with there being anywhere from zero to medium heterogeneity. In other words, the estimate of heterogeneity is imprecise; this is also unsurprising, as five studies with small sample sizes (which is understandable given the subject population) are simply unable to provide a precise estimate of heterogeneity.

To summarize this SPM, we suggest the following:

> Across five studies, we showed that preschoolers consume less food relative to a control condition when it is presented as instrumental to achieve a goal, but about the same amount of food when it is presented as yummy. An SPM of our studies estimates the first effect at 5.09 morsels (95% CI: 3.80–6.38) and the second effect at 1.84 (95% CI: −.21–3.89), indicating that presenting food as instrumental reduces consumption substantially. I^2 was estimated at 19% (95% CI: 0%–59%), suggesting heterogeneity is low, with method factors accounting for only about one-fifth of the variation in the observations beyond that attributable to the experimental manipulations; however, the width of the interval suggests that heterogeneity is not estimated precisely.

Case Study II: Shah et al. (2014)

Shah et al. (2014) proposed that restaurant menus that combine a price surcharge with an unhealthy label can "reduce the demand for unhealthy food." They tested this hypothesis across three laboratory studies and one field study that used similar stimuli.[2] The primary dependent measure in these studies was whether or not the subject ordered an unhealthy entrée, and the authors' analysis focused on three contrasts—the difference in the proportion of unhealthy entrées ordered between a control menu (i.e., no surcharge and no label) and each of three intervention menus (i.e., surcharge only, label only, both surcharge and label).

TABLE 2

SHAH ET AL. (2014) SUMMARY INFORMATION

Study	Factor 1	Factor 2	*p*	*n*
Study 1A	No Label	No Surcharge	0.422	300
	No Label	Surcharge	0.358	300
	Label	No Surcharge	0.333	300
	Label	Surcharge	0.252	300
Study 1B (No Information)	No Label	No Surcharge	0.426	149
	No Label	Surcharge	0.349	149
	Label	No Surcharge		
	Label	Surcharge	0.262	149
Study 1B (Information)	No Label	No Surcharge	0.389	149
	No Label	Surcharge	0.366	149
	Label	No Surcharge		
	Label	Surcharge	0.257	149
Study 2 (Alone)	No Label	No Surcharge	0.454	248
	No Label	Surcharge	0.430	248
	Label	No Surcharge	0.390	248
	Label	Surcharge	0.322	248
Study 2 (Friend)	No Label	No Surcharge	0.487	248
	No Label	Surcharge	0.463	248
	Label	No Surcharge	0.414	248
	Label	Surcharge	0.272	248
Study 3	No Label	No Surcharge	0.499	116
	No Label	Surcharge	0.457	116
	Label	No Surcharge	0.297	116
	Label	Surcharge	0.291	116

NOTE.—This table reproduces the data found in tables 1–4 of Shah et al. (2014).

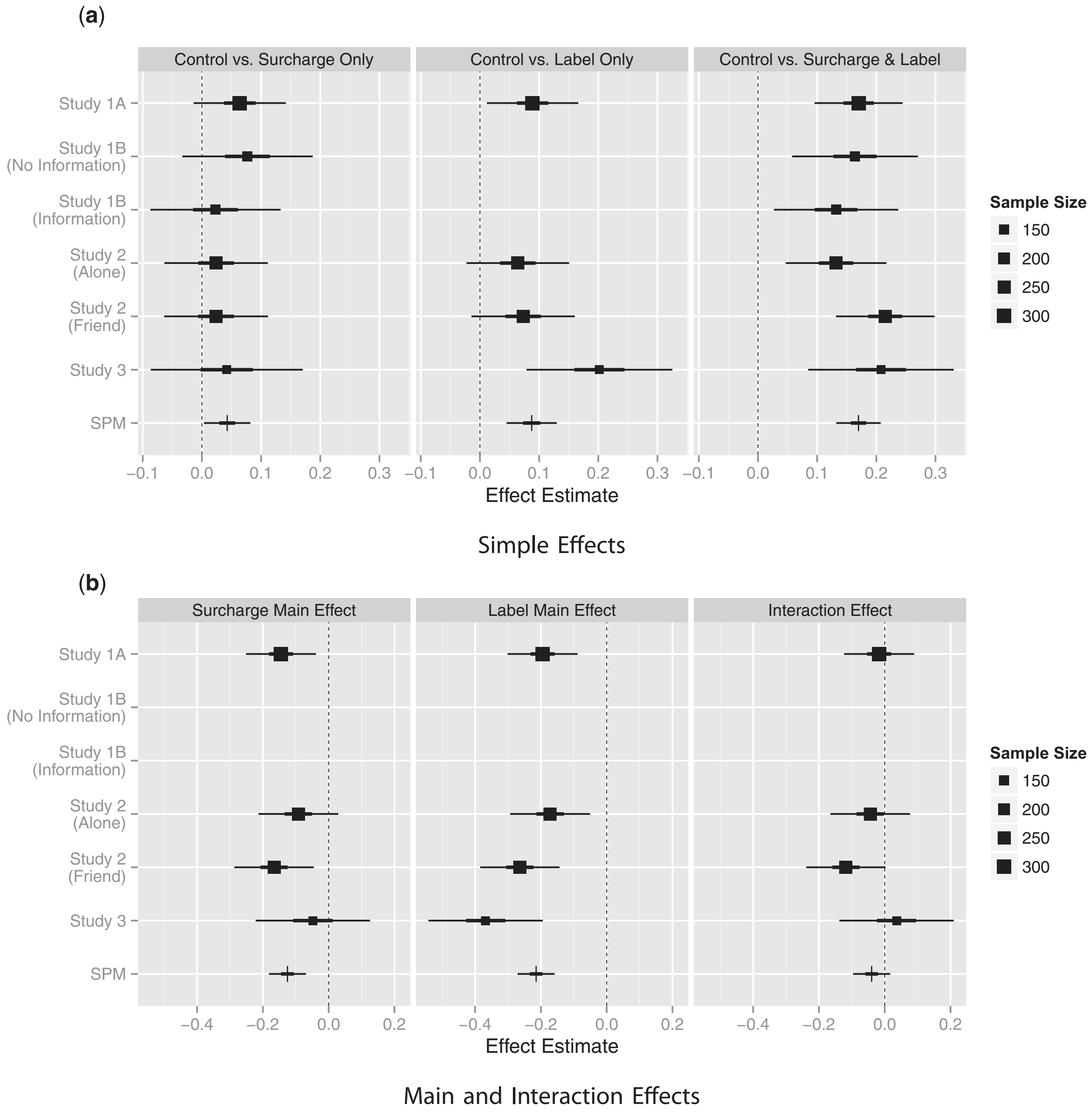
We present the summary information for the series of studies conducted by Shah et al. (2014) in table 2, which serves as the primary input data for our SPM methodology; when the dependent measure is binary as here, only proportions and sample sizes (as opposed to means, standard deviations, and sample sizes) are required. The contrasts studied by the authors are given by (1 –1 0 0) for the control versus surcharge-only contrast, (1 0 –1 0) for the control versus label-only contrast, and (1 0 0 –1) for the control versus surcharge and label contrast.

We present single-study estimates from Shah et al. (2014) in figure 2(a). As per the authors' theory, the contrast between the control menu and the menu with both a surcharge and a label consistently attains statistical significance. In addition, the estimates of all three contrasts are remarkably consistent

[2] Study 1A followed a 2 × 2 design (absence or presence of a price surcharge; absence or presence of an unhealthy label). Study 1B dropped the surcharge-only menu but added an additional experimental factor (absence or presence of calorie and health information); in all other studies, this information was absent. Study 2 returned to the design of study 1A but added an additional experimental factor (dining alone vs. dining with a friend). Finally, study 3 was a field study that followed the design of study 1A. All studies followed a between-subjects design. As there was no main effect of calorie and health information in study 1A, there was no main effect of dining partner in study 2, and these additional experimental factors were not primary to the authors' theory, we treat these conditions as additional observations of the two main experimental factors. This is a modeling decision we made in analyzing the data, and we do not mean to exclude other decisions or suggest this is the most appropriate one for all purposes. Since our main purpose is to illustrate our SPM methodology, this decision seems reasonable here.

FIGURE 2

RESULTS FROM SHAH ET AL. (2014)



Simple Effects

Main and Interaction Effects

NOTE.—Effect estimates are given by the squares for single-study estimates and the vertical bars for SPM estimates; 50% and 95% intervals are given by the thick and thin lines, respectively. The average sample size per condition in each study is given by the size of the squares. The SPM suggests all three contrasts shown in the top subfigure attain statistical significance, thereby motivating the decomposition of the effects shown in the bottom subfigure. This shows the simple contrasts can be decomposed into main effects and any interaction effect is likely to be negligible in size.

across studies—whether or not they attain statistical significance. This suggests that the findings are substantially stronger than indicated by "vote-counting" the number of statistically significant results. It is also interesting because it shows that the results of the field study (study 3) are consistent not only in sign with but also in size with the laboratory studies, thereby strengthening the authors' contribution in terms of robustness and generalizability.

Due to the consistency of the effect estimates across studies, the SPM estimates, which we also present in figure 2(a), are estimated with uncertainty intervals that are much narrower than those of the single-study estimates. Consequently, while the single-study contrasts involving the surcharge-only menu and the label-only menu generally do not attain statistical significance, the SPM indicates a small but nonzero effect on the order of 5 to 10 percentage points for each. Further, the effect of the menu with both a surcharge and a label is estimated to be about 15 to 20 percentage points.

That all three SPM estimates attain statistical significance raises the question of how the effects arise. Understanding this can be instructive for theory, and thus we decompose the effects into two main effects and an interaction effect and present results in figure 2(b).[3] For the single-study estimates, the surcharge main effect sometimes attains and sometimes fails to attain statistical significance, the label main effect consistently attains statistical significance, and the interaction effect consistently fails to attain statistical significance. Nonetheless, the estimates are highly consistent across studies. Consequently, the SPM estimates of the two main effects attain statistical significance; on the other hand, the SPM estimate of the interaction effect fails to attain statistical significance.

The SPM heterogeneity estimate is .0008 on the variance scale (or $\sqrt{.0008} = .0277$ on the standard deviation scale); this means that the proportion of unhealthy entrées ordered in a given condition varies by nearly three percentage points from study to study due to method factors (i.e., even in the absence of sampling variation). I^2 is estimated at 36% with an uncertainty interval of (0%, 63%), suggesting that heterogeneity is low but could range from zero to medium.

To summarize this SPM, we suggest the following:

> Across three laboratory studies and one field study, we found that consumers presented with a menu that incorporates either a price surcharge or an unhealthy label order unhealthy food about as frequently as consumers presented with a no-intervention menu that incorporates neither, while consumers presented with a menu that incorporates both order unhealthy food less frequently. However, an SPM of our studies shows that all three menus do in fact reduce the frequency with which unhealthy food is ordered relative to the no-intervention menu: a menu with a price surcharge by 4.28% (95% CI: .36%–8.19%), a menu with an unhealthy label by 8.74% (95% CI: 4.50%–12.98%), and a menu with both by 16.99% (95% CI: 13.23%–20.74%). While these effects range in size, all are potentially large enough to be practically important for combatting obesity.
>
> As a follow-up, we decomposed the effects into main and interaction effects. This showed that the strongest effect is driven by the label, that there is a weaker effect of the surcharge, and most importantly that these two effects appear to operate independently of each other. I^2 was estimated at 36% (95% CI: 0%–63%), suggesting that heterogeneity is low but could range from zero to medium; this estimate, along with the visual convergence of effects demonstrated in figure 2(a), is particularly encouraging, as it shows our field study was consistent not only in sign with but also in size with the laboratory studies, thereby demonstrating the robustness and generalizability of our findings.

Case Study III: Anonymous Consumer Behavior Researcher

A consumer behavior researcher investigated how satisfaction with a product chosen by a consumer varies as a function of the choice task difficulty as well as the choice set size across five studies that used similar stimuli.[4] The primary dependent measure in these studies was satisfaction measured on a seven-point integer scale and the hypotheses concerned (i) the simple effect of the choice set size when choice task difficulty was low, (ii) the simple effect of choice task difficulty when the choice set size was small, and (iii) the interaction effect.

As the researcher wishes to remain anonymous, we present a masked version of the summary information from the studies in table 3, which serves as the primary input data for our SPM methodology.[5] The contrasts studied by the researcher are given by (–1 1 0 0) for the first simple effect, (–1 0 1 0) for the second simple effect, and (1 –1 –1 1) for the interaction effect.

[3] The contrasts for these effects are given by (–1 1 –1 1) for the surcharge main effect, (–1 –1 1 1) for the label main effect, and (1 –1 –1 1) for the interaction effect.

[4] Study 1 followed a two-condition design (small vs. large choice set with low choice task difficulty), while the remaining four studies followed a 2 × 2 design (low vs. high choice task difficulty; small vs. large choice set). Studies 1, 2, and 5 followed a between-subjects design; study 3 followed a partially within-subjects design; and study 4 followed a fully within-subjects design.

[5] Individual-level observations were simulated according to the study data such that the individual-level observations underlying the summary information presented in table 3 have zero mean and unit variance across all conditions of all studies. When the study designs follow a mix of between-subjects and within-subjects designs as here, information on the designs and observed covariances may optionally be provided as input data for our SPM methodology. For full details see our website.

TABLE 3

ANONYMOUS CONSUMER BEHAVIOR RESEARCHER SUMMARY INFORMATION

Study	Factor 1	Factor 2	$\bar{x}$	*s*	*n*
Study 1	Low	Small	–.227	.868	50
	Low	Large	.330	.939	50
	High	Small			
	High	Large			
Study 2	Low	Small	.167	.881	100
	Low	Large	.747	.858	100
	High	Small	–.145	1.039	100
	High	Large	–.009	1.003	100
Study 3	Low	Small	–.258	1.072	75
	Low	Large	.175	.918	75
	High	Small	–.145	.908	75
	High	Large	–.540	1.005	75
Study 4	Low	Small	.234	.894	125
	Low	Large	.706	.961	125
	High	Small	–.026	.917	125
	High	Large	.082	.840	125
Study 5	Low	Small	–.277	.922	150
	Low	Large	.196	1.037	150
	High	Small	–.324	.911	150
	High	Large	–.581	.861	150

We present single-study estimates from the five studies in figure 3(a). As can be seen, consistent with the researcher's theory, the estimates of the first simple effect are all positive and attain statistical significance, and the estimates of the interaction effect are all negative and attain statistical significance. On the other hand, the estimates of the second simple effect are mixed in terms of sign and statistical significance.

We also present the SPM estimates in figure 3(a). As is typical, these estimates lie near the midpoint of the single-study estimates, and the uncertainty in them is small relative to the uncertainty in the single-study estimates. However, even the SPM estimate of the second simple effect fails to attain statistical significance. This would seem to suggest the researcher's theory is false or requires refinement.

The SPM heterogeneity estimate is .06 on the variance scale (or $\sqrt{.06} = .24$ on the standard deviation scale). While this can be interpreted by comparing it to the standard deviations and sample sizes displayed in table 3, the I^2 statistic, as noted above, can yield a cleaner interpretation. I^2 is estimated at 87% with an uncertainty interval of (81%, 92%), suggesting that heterogeneity is very high.

When heterogeneity is high, it is worthwhile to consider whether the studies differ in terms of stimuli, social context, subject pool, or other potentially important method factors. In this case, the researcher considered this and discovered the products used in studies 1, 3, and 5 were utilitarian (e.g., pens), while those used in studies 2 and 4 were hedonic (e.g., chocolates). We then suggested, for exploratory purposes, that the researcher account for this method factor in the SPM; this can be done by treating the studies as if they had followed a 2 × 2 × 2 (rather than 2 × 2) design where the additional factor reflects the utilitarian versus hedonic distinction.

We present results from this analysis in figure 3(b). The single-study estimates are of course the same as those in figure 3(a). However, the SPM estimates have changed substantially and can show dramatic differences depending on whether the product is utilitarian or hedonic. In particular, the second simple effect is negative for hedonic products but about zero for utilitarian products. Further, the interaction effect is larger in absolute value for utilitarian products relative to hedonic products. On the other hand, the first simple effect appears unaffected by whether the product is utilitarian or hedonic. Thus, whether the product is utilitarian or hedonic seems theoretically and practically important in this domain.

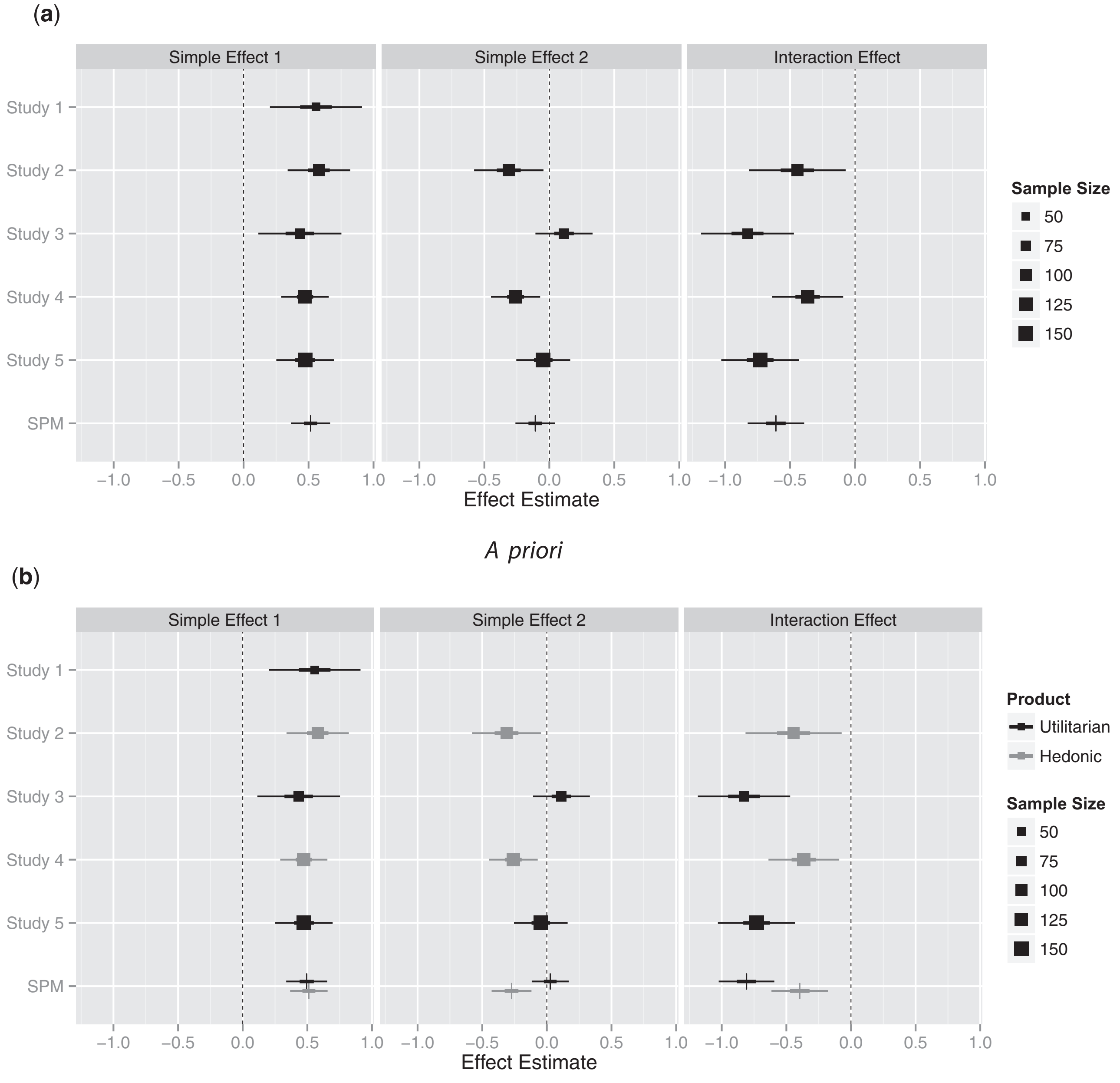
When the product type is accounted for in the SPM, the estimate of heterogeneity drops to zero and I^2 is estimated at 0% with an uncertainty interval of (0%, 22%), suggesting heterogeneity is zero to low. In sum, the utilitarian versus hedonic method factor seems to have been driving the high heterogeneity.

To summarize this SPM, we suggest the following:

> Across five laboratory studies, we found that satisfaction is lower for small choice sets relative to large ones when choice task difficulty is low and that the choice set size interacts with choice task difficulty. However, contrary to our theory, we found that satisfaction is about the same for small choice sets regardless of whether choice task difficulty is high or low. An SPM of our studies estimates the first simple effect at .52 (95% CI: .37–.67) and the interaction effect at $-.61$ (95% CI: $-.83 - -.39$), while it estimates the second simple effect at $-.11$ (95% CI: $-.26$–.05). I^2 was estimated at 87% (95% CI: 81%–92%), suggesting that heterogeneity is very high.
>
> This large heterogeneity prompted us to reflect upon differences among the studies. We discovered the products used in studies 1, 3, and 5 were utilitarian (e.g., pens), while those used in studies 2 and 4 were hedonic (e.g., chocolates). As an exploratory analysis, we conducted another SPM that accounted for this method factor. In this SPM, the estimate of I^2 dropped to 0% (95% CI: 0%–22%), suggesting the utilitarian versus hedonic method factor seems to have been driving the high heterogeneity. Further, this SPM suggested the effect of the choice set size when choice task difficulty is low is similar for utilitarian and hedonic products; the effect of choice task difficulty when the choice set size is small is approximately zero for utilitarian products but negative for hedonic products; and the interaction effect is larger for utilitarian products than for hedonic products.

FIGURE 3

RESULTS FROM ANONYMOUS CONSUMER BEHAVIOR RESEARCHER



A priori

A posteriori

NOTE.—Effect estimates are given by the squares for single-study estimates and the vertical bars for SPM estimate; 50% and 95% intervals are given by the thick and thin lines, respectively. The average sample size per condition in each study is given by the size of the squares. When the utilitarian versus hedonic method factor is accounted for, the SPM estimates vary depending on the product type.

We believe this exploratory analysis suggests a fruitful avenue for future research. In particular, we plan to incorporate the product type as an experimentally manipulated factor in future studies of this phenomenon in order to assess whether the results of our exploratory analysis hold more generally. We also plan to determine whether other differences in products (e.g., luxury vs. mass market products, durable vs. nondurable products, products bought for the self vs. for others) interact with our effects of interest in future work.

SAMPLE SIZE ANALYSES

Our SPM methodology provides a sample size analysis so that future studies of each effect of interest are adequately powered. This analysis is based principally on the SPM effect estimates. However, because optimistic assessments of power—and thus sample sizes that are too small—result if either the uncertainty associated with estimates (McShane and Böckenholt 2016) or the heterogeneity resulting from method factors (McShane and Böckenholt 2014) is ignored, our SPM sample size analysis is novel and unique in that it accounts for both of these important factors.

As an illustration, consider a simple effect that is estimated at .4. Suppose that the estimated standard error is .1, the estimate of heterogeneity is .1, and, without loss of generality, the standard deviation of the individual-level observations is 1. Our SPM sample size analysis suggests that 101 subjects per condition are required for a one-sided single-study null hypothesis significance test of zero effect with size $\alpha = .05$ to be adequately powered (i.e., at 80%; Cohen 1992). In contrast, a standard sample size analysis (i.e., one that assumes the effect is known to be .4 with complete certainty and that heterogeneity is zero) suggests that only 78 subjects per condition are required for adequate power. Thus, properly accounting for the uncertainty associated with the estimate as well as heterogeneity results in a larger requisite sample size, but one that provides adequate power.

Our SPM methodology also provides a sample size analysis so that future sets of studies are adequately powered. This analysis accounts for heterogeneity and uncertainty as above. However, it is again novel and unique in that it is a meta-analytic sample size analysis: the sample size does not necessarily yield individual studies that are adequately powered, but rather a meta-analysis of the set of studies taken as a whole that is adequately powered. For example, using the assumptions in the prior paragraph, our SPM sample size analysis suggests only 35 subjects per condition per study are required for a meta-analysis of three studies to be adequately powered. Clearly, each study is underpowered, as the sample size is less than the 101 subjects per condition required for a single study to be adequately powered. Nonetheless, when taken as a whole, the three individually underpowered studies are adequately powered. Similarly, our SPM sample size analysis suggests only 25 subjects per condition per study are required for a meta-analysis of four studies to be adequately powered. As $25 \times 4 < 35 \times 3$, this reveals an interesting insight: when heterogeneity is nonzero, a given level of power is maintained when one splits fewer subjects across a larger number of studies.

In addition to allowing for fewer subjects in total, there are several additional benefits associated with conducting multiple small studies of a given phenomenon rather than one large study when heterogeneity is nonzero (McShane and Böckenholt 2014). Multiple studies allow researchers to estimate heterogeneity via meta-analysis. This can suggest method factors including unaccounted-for moderators, as illustrated in Case Study III, and it allows for better calibration of Type I and Type II error, as illustrated in the appendix. Multiple studies also yield more efficient estimates—and thus greater power for a fixed total sample size—of overall average effects.

In sum, our SPM sample size analyses help ensure that researchers avoid sample sizes that are too small (large) and thus power that is less (more) than adequate. Therefore, they help researchers use their resources in an efficient manner and enhance replicability.

We refer the reader interested in the full details regarding our SPM sample size analyses to our supplementary materials.

BENEFITS

In this section, we discuss several benefits provided by our SPM methodology. While many of these were touched on in prior sections, we here elaborate on how it aids in study summary, theory testing, and replicability.

Study Summary

Our SPM methodology facilitates the summary of a set of studies in a variety of ways. Most notably, the intuitive graphical summary (e.g., figures 1–3) allows for the easy and rapid communication and comparison of results both at the single-study level and in aggregate. One can quickly examine the plot to see if, for example, there are one or more outlying studies or whether there is convergence across the set of studies.

Further, it provides estimates for all effects of interest as well as the uncertainty in them (i.e., as depicted at the bottom of figures 1–3), thus heeding recent calls for effect size estimates (Cohen 1990; Eich 2014; Iacobucci 2005; Lindsay 2015). These estimates are more reliable than those based on single studies and generally have narrower uncertainty intervals. This increases the quality of reported results, diverts attention away from single-study estimates (which can be noisy), and focuses attention on both SPM estimates and the convergence of results across studies. In sum, SPM shifts the focus away from the demonstration of deviations from questionably meaningful sharp point null hypothesis significance tests of zero effect and toward estimation.

Theory Testing

Our SPM methodology improves theory testing in several ways. For example, because the SPM uncertainty intervals are narrower than the single-study intervals, hypothesized effects—both null and non-null—can be tested more powerfully. For posited non-null effects, the findings will be stronger than indicated by vote-counting the number of statistically significant results. For posited null effects, the more precise estimate and narrower uncertainty interval provides a stronger test.

These more powerful tests of hypothesized effects can also deepen theory testing. When authors conduct only weaker single-study tests, they may fail to detect non-null effects (i.e., commit Type II error). However, when they conduct stronger meta-analytic tests, they may detect these effects; this may in turn motivate them to consider new decompositions of the effects that inform theory and investigate alternative explanations, as illustrated in Case Study II.

Further, because behavioral research studies can never be direct or exact replications of one another (Brandt et al. 2014; Fabrigar and Wegener 2016; Rosenthal 1991; Stroebe and Strack 2014; Tsang and Kwan 1999), our SPM methodology estimates and accounts for heterogeneity, which has been shown to be important in a wide variety of behavioral research settings (Hedges and Pigott 2001; Klein et al. 2014; Pigott 2012). This estimate not only is important in its own right but also can enrich theory by suggesting unaccounted-for moderators, as illustrated in Case Study III.

Accounting for heterogeneity has another important benefit for theory testing. When heterogeneity is falsely assumed to be zero, as, for example, when studies are analyzed in isolation, the Type I error of null hypothesis significance tests is larger than the nominal size α and the Type II error is higher (and thus statistical power is lower) than suggested by standard formulae. Thus, our SPM methodology, which accounts for heterogeneity, provides not only more powerful tests of posited effects but also, as illustrated in the appendix (see also McShane and Böckenholt 2014), better calibrated tests.

Replicability

Our SPM methodology enhances replicability in a number of ways. Most notably, it provides sample size analyses so that future studies and sets of studies of each effect of interest are adequately powered; importantly, these analyses account for both uncertainty and heterogeneity.

In addition, because the primary input data required by our SPM methodology can be reported in an extremely concise table (e.g., tables 1–3), it encourages and facilitates an important level of reporting that is sometimes lacking in published work and which enhances replicability.

An additional benefit of this concision is that it allows authors to, if they desire, report studies they have that are related to those reported in the paper but which themselves were not reported in the paper. In particular, authors can simply add the data from these studies to the table and include it in the SPM. This allows them to provide further evidence about the phenomenon of interest without taking up a great deal of journal space and is generally superior to entirely omitting the studies. Further, it helps avoid the upwardly biased estimates of effect sizes and downwardly biased estimates of heterogeneity that generally result from the selective reporting of studies that attain statistical significance (Gelman and Carlin 2014; Gelman and Weakliem 2009; McShane, Böckenholt, and Hansen 2016). Consequently, it enhances replicability.

REMARKS

In this section, we remark on a number of features of our SPM methodology that are important for application. First, strictly speaking, our SPM methodology accommodates a single dependent measure that is measured on the same measurement scale across studies. Nonetheless, insofar as the dependent measures employed across studies assess different but related constructs or are measured on different but similar measurement scales, the studies may be analyzed via our SPM methodology; however, these differences will tend to increase heterogeneity. Further, when secondary dependent measures that assess unrelated constructs are measured and are of interest, they may be analyzed via an additional separate SPM.

Second, our SPM methodology accommodates only discrete covariates, such as the discretely manipulated experimental factors used in the vast majority of behavioral research studies, but it does not accommodate continuous covariates such as continuous measured variables. This is a conscious design choice with genuine benefits: it allows for a model that requires only basic summary information rather than individual-level observations yet is equivalent to that underlying the gold-standard meta-analytic approach, thus providing greater ease of use and allowing readers as well as authors to conduct an SPM. Models that accommodate continuous covariates require access to the individual-level observations, and we encourage authors who possess individual-level observations with continuous covariates to model this data directly via a hierarchical model.

Third, while our SPM methodology can accommodate (discrete) study-level covariates as illustrated in Case Study III, because single papers feature a relatively small number of studies, assessing the impact of study-level covariates is generally a task best left to traditional meta-analyses of large numbers of studies. We note that, regardless, meta-analysis can establish any such impact as a mere association rather than as a cause.

Fourth, our SPM methodology may in some cases be applicable to only a subset of the studies that appear in a typical behavioral research paper. In this case, we believe an SPM of the subset is still valuable and advocate this practice. However, we also note that an SPM of the full set may still be possible. For example, consider a set of studies in which one of the studies has a continuous covariate. If this covariate is, say, an experimental factor that was manipulated discretely in the other studies but continuously in the one study, then we would advocate an SPM of the subset. On the other hand, if this covariate is, say, a measured variable used as a control variable, then an SPM of the full set but omitting the variable is also possible.

Fifth, and as previously noted, our SPM methodology may in some cases be applicable to a superset of the studies that appear in a typical behavioral research paper. In particular, authors may, if they desire, include in the SPM studies they have that are related to those reported in the paper but which themselves were not reported in the paper.

Finally, because the purpose of SPM is to summarize sets of studies that appear in a typical behavioral research paper, our SPM methodology does not attempt to assess or adjust for publication bias or any other forms of bias in the studies (McShane et al. 2016; Rothstein, Sutton, and Borenstein 2005). Any bias in the studies is—as we believe is proper given the purpose of SPM—incorporated into the SPM estimates.

DISCUSSION

The current practice of analyzing the multiple studies of a common phenomenon that appear in a typical behavioral research paper solely in isolation is inefficient and forgoes benefits for study summary, theory testing, and replicability that can be obtained only by analyzing them jointly. Consequently, we advocate that authors of typical behavioral research papers include a table of summary information from each study (e.g., tables 1–3), conduct and discuss an SPM of the studies, and provide the intuitive graphical summary (e.g., figures 1–3). This will supplement the single-study analyses that independently examine the multiple studies as well as the qualitative meta-analysis that verbally synthesizes the studies and requires only a minor modification of current practice.

To facilitate this, we make four principal contributions. First, we introduce meta-analytic methodology that is user-friendly, widely applicable, and—most importantly—specially tailored to account for the complex patterns of variation and covariation among the observations from a set of studies that appear in a typical behavioral research paper. Second, we note that the model underlying our SPM methodology is equivalent to that underlying the gold-standard meta-analytic approach and exploit that this equivalence is widely applicable for the meta-analysis of the set of studies that appear in typical behavioral research papers. Third, we introduce two novel and unique sample size analyses; our sample size analysis for future studies integrates the approaches of McShane and Böckenholt (2014, 2016) to account for both uncertainty and heterogeneity, while our sample size analysis for future sets of studies extends this to multiple studies. Fourth, we provide a website that implements our SPM methodology and sample size analyses and provides the intuitive graphical summary in a single, easy-to-use package.

Our SPM methodology is advantageous relative to the alternative approaches noted in our introduction, such as those discussed by Rosenthal (1978) and the standardized effect approach. For instance, the approaches discussed by Rosenthal (1978) produce only a *p*-value and no estimate of effect size or uncertainty, while the preprocessing of the observed effects required by the standardized effect approach (i.e., to convert them to a common scale) is laborious and results in a lack of correspondence between the single-study analyses and the meta-analysis. In addition, these approaches accommodate only a single effect of interest (i.e., because they do not account for the dependence between the multiple effects of interest typical in behavioral research studies); make it difficult to accommodate a mix of between-subjects and within-subjects study designs; and fail to accommodate studies that provide only partial information about the effect of interest (i.e., studies that omit one or more conditions relevant for an effect of interest must be omitted from the meta-analysis). Further, the approaches discussed by Rosenthal (1978) fail to account for heterogeneity and, in practice, the same holds for the standardized effect approach (i.e., because, when there are few studies as in SPM, the standardized effect approach often estimates heterogeneity at zero when it is in fact nonzero; Chung et al. 2013); failing to account for heterogeneity can, as illustrated in the appendix, result in miscalibrated Type I and Type II error. Finally, these approaches do not produce an intuitive graphical summary, sample size analyses, or other benefits of our SPM methodology.

In closing, we note our SPM methodology has benefits for three constituents not discussed in depth in this paper. First, and as previously noted, because our SPM methodology requires only basic summary information, it allows readers as well as authors to conduct an SPM and obtain the benefits for study summary, theory testing, and replicability (provided this information is reported). Consequently, all benefits provided by our SPM methodology are available for a large corpus of previously published work.

Second, because our SPM methodology heeds Cohen's maxim that "the primary product of a research inquiry is one or more measures of effect size, not *p*-values"

(Cohen, 1990, 1310) as well as more recent calls for effect size estimates (Eich 2014; Iacobucci 2005; Lindsay 2015), it is particularly useful to those researchers who believe results should not be assessed solely or even primarily with reference to null hypothesis significance tests. This focus on effect sizes also helps avoid several well-known problems associated with sharp point null hypothesis significance testing, such as the arbitrariness of the standard size $\alpha = .05$ (Cowles and Davis 1982); the fact that statistical significance is not the same as practical importance (Freedman et al. 2007); and the tendency to view evidence dichotomously rather than continuously resulting in the dismissal of differences observed in practice (McShane and Gal 2016; Rosnow and Rosenthal 1989).

Third, we recognize that some researchers are opposed to meta-analysis on principle: they believe each study is entirely unique and therefore pooling data across studies involves combining things that are not alike. While we ultimately believe this to be a grim perspective because, taken seriously, it precludes generalization and, *inter alia*, suggests a return to single-study papers, we note that our SPM methodology provides at least two benefits even to researchers who hold this perspective: (i) the graphical summary can be used to assess convergence across studies without recourse to meta-analysis by simply ignoring the bottom portion of the graphic that reflects the SPM and (ii) the estimate of heterogeneity can be used to quantify the degree to which the studies are not alike. Thus, we are optimistic that even researchers skeptical of meta-analysis can find our SPM methodology to some degree useful.

Appendix

THE IMPACT OF HETEROGENEITY ON TYPE I AND TYPE II ERROR

In this appendix, we show that heterogeneity causes the Type I error of standard single-study null hypothesis significance tests to be greater than the nominal size α, but that tests based on a meta-analysis that accounts for heterogeneity preserve their nominal size α. We also show that heterogeneity impacts Type II error, in particular (i) that heterogeneity causes standard single-study power formulae, which assume heterogeneity to be zero, to overstate power (i.e., understate Type II error) and (ii) that, when heterogeneity is nonzero, splitting the same number of subjects across a larger number of studies results in greater power (or, equivalently, a given level of power is maintained when one splits fewer subjects across a larger number of studies).

Type I Error

Many researchers are interested in controlling the Type I error of their null hypothesis significance tests. However, when heterogeneity is nonzero and one possesses only a single study, controlling Type I error is not possible; nonzero heterogeneity implies that the true effect in a study is the sum of an overall average component and a study-specific component, but these two components are entirely confounded with only a single study—regardless of the sample size of the study. Consequently, while one wants to test whether the overall average component is zero, one can test only whether the sum of the overall average component and the study-specific component is zero, thus resulting in Type I error that is greater than the nominal size α.

Assuming the definitions of no, low, medium, and high heterogeneity given by Pigott (2012) apply at the level of the effect of interest, and assuming one is interested in conducting a two-sided, single-study, null hypothesis significance test of zero effect with size $\alpha = .05$, the Type I error of the test is .05 when heterogeneity is zero (i.e., the test maintains its nominal size α). However, the Type I error rises to .09, .17, and .33 when heterogeneity is respectively low, medium, and high; the Type I error is larger than the nominal size α because the single-study test that the sum of the overall average component and the study-specific component is zero does not properly test whether the overall average component is zero. On the other hand, when one conducts a meta-analysis that accounts for heterogeneity, the Type I error matches the nominal size $\alpha = .05$ of the test because the meta-analytic test properly tests whether the overall average component is zero.

Type II Error

Heterogeneity impacts Type II error and thus power. To illustrate this, suppose that one is interested in a simple effect of size .4 and, without loss of generality, the standard deviation of the individual-level observations is 1. Further, assume the definitions of no, low, medium, and high heterogeneity given by Pigott (2012) apply at the level of this effect and that within-study variation is quantified by a study with a sample size of 78 subjects per condition (as noted in the main text, a standard sample size analysis suggests that this sample size is required for adequate—i.e., 80%—power in this setting).

Now, assume one is interested in conducting a one-sided, single-study, null hypothesis significance test of zero effect with size $\alpha = .05$ and that one plans to use 156 subjects per condition (i.e., double the amount required by the standard sample size analysis) in the hope that this large sample size will provide more than adequate power. As shown in the prior subsection, the single-study test that the sum of the overall average component and the study-specific component is zero does not properly test whether the overall average component is zero when heterogeneity is nonzero, thus causing Type I error to increase above the nominal size α; for the same reason, heterogeneity also causes power to drop. In particular, power is .97 when heterogeneity is zero. However, power drops to .93, .86, and .76 when

TABLE 4

META-ANALYSIS POWER

Number of studies	Heterogeneity			
	Zero	Low	Medium	High
Three	.97	.94	.86	.65
Four	.97	.95	.89	.72
Five	.97	.95	.91	.77
Six	.97	.96	.92	.80

NOTE.—When heterogeneity is zero, meta-analysis and single-study analysis are equivalent. When heterogeneity is nonzero, splitting the same number of subjects across a larger number of studies results in greater power.

heterogeneity is respectively low, medium, and high. In sum, optimistic assessments of power—and thus sample sizes that are too small—result if heterogeneity is ignored; for example, even with double the sample size required by the standard sample size analysis, power is inadequate when heterogeneity is high.

Instead, suppose one is interested in taking those same 156 subjects per condition and splitting them across multiple studies to conduct a one-sided meta-analytic null hypothesis significance test of zero effect with size $\alpha = .05$. As shown in the prior subsection, the meta-analytic test properly tests whether the overall average component is zero when heterogeneity is nonzero and thus maintains its nominal size α; this has implications for Type II error and thus power as shown in table 4 (as the single-study test discussed in the prior paragraph and the meta-analytic test discussed in this paragraph test different hypotheses when heterogeneity is nonzero, power is not directly comparable between them). When heterogeneity is zero, conducting multiple studies is equivalent to conducting one large study because the study-specific components are all zero; thus, meta-analytic power is equivalent to single-study power. On the other hand, when heterogeneity is nonzero, a meta-analysis consisting of only a small number of studies can have relatively low power but, when the same number of subjects is split across a larger number of studies, power can increase considerably.

REFERENCES

Borenstein, Michael, Larry V. Hedges, Julian P.T. Higgins, and Hannah R. Rothstein (2009), *Introduction to Meta-Analysis*, Chichester, UK: Wiley.

Brandt, Mark J., Hans IJzerman, Ap Dijksterhuis, Frank J. Farach, Jason Geller, Roger Giner-Sorolla, James A. Grange, Marco Perugini, et al. (2014), “The Replication Recipe: What Makes for a Convincing Replication?” *Journal of Experimental Social Psychology*, 50 (1), 217–24.

Brown, Sacha D., David Furrow, Daniel F. Hill, Jonathon C. Gable, Liam P. Power, and W. Jake Jacobs (2014), “A Duty to Describe: Better the Devil You Know Than the Devil You Don’t,” *Perspectives on Psychological Science*, 9 (6), 626–40.

Chung, Yeojin, Sophia Rabe-Hesketh, Vincent Dorie, Andrew Gelman, and Jingchen Liu (2013), “A Nondegenerate Estimator for Hierarchical Variance Parameters via Penalized Likelihood Estimation,” *Psychometrika*, 78 (4), 685–709.

Cohen, Jacob (1990), “Things I Learned (So Far),” *American Psychologist*, 45 (12), 1304–12.

—— (1992), “A Power Primer,” *Psychological Bulletin*, 112 (1), 155–59.

Cooper, Harris and Erika A. Patall (2009), “The Relative Benefits of Meta-Analysis Conducted with Individual Participant Data versus Aggregated Data,” *Psychological Methods*, 14 (2), 165.

Cooper, H. M., Larry V. Hedges, and Jeffrey C. Valentine, eds. (2009), *The Handbook of Research Synthesis and Meta-Analysis*, New York: Sage.

Cowles, Michael and Caroline Davis (1982), “On the Origins of the .05 Level of Significance,” *American Psychologist*, 37 (5), 553–58.

Eich, Eric (2014), “Business Not as Usual,” *Psychological Science*, 25 (1), 3–6.

Fabrigar, Leandre R. and Dwayne T. Wegener (2016), “Conceptualizing and Evaluating the Replication of Research Results,” *Journal of Experimental Social Psychology*, 66 (September), 68–80.

Fisher, Ronald A. (1925), *Statistical Methods for Research Workers*, Edinburgh: Oliver and Boyd.

—— (1948), “Questions and Answers 14: Combining Independent Tests of Significance,” *The American Statistician*, 2 (5), 30.

Freedman, David, Robert Pisani, and Roger Purves (2007), *Statistics, 4th ed.*, New York: W. W. Norton and Company.

Gelman, Andrew and John Carlin (2014), “Beyond Power Calculations Assessing Type S (Sign) and Type M (Magnitude) Errors,” *Perspectives on Psychological Science*, 9 (6), 641–51.

Gelman, Andrew and David Weakliem (2009), “Of Beauty, Sex, and Power: Statistical Challenges in Estimating Small Effects,” *American Scientist*, 97 (August), 310–16.

Haidich, Anna-Bettina (2010), “Meta-Analysis in Medical Research,” *Hippokratia*, 14 (1), 29–37.

Hedges, Larry V. and Ingram Olkin (1985), *Statistical Methods for Meta-Analysis*, Orlando, FL: Academic Press.

Hedges, Larry V. and Therese D. Pigott (2001), “The Power of Statistical Tests in Meta-Analysis,” *Psychological Methods*, 6 (3), 203–17.

Higgins, Julian P. T. and Simon G. Thompson (2002), “Quantifying Heterogeneity in a Meta-Analysis,” *Statistics in Medicine*, 21 (11), 1539–58.

Higgins, Julian P. T., Simon G. Thompson, J. J. Deeks, and D. G. Altman (2003), “Measuring Inconsistency in Meta-Analyses,” *British Medical Journal*, 327 (7414), 557.

Hunter, John E. and Frank L. Schmidt (2014), *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*, Newbury Park, CA: Sage.

Iacobucci, Dawn (2005), “From the Editor,” *Journal of Consumer Research*, 32 (1), 1–6.

Klein, Richard A., Kate A. Ratliff, Michelangelo Vianello, Reginald B. Adams Jr., Štěpán Bahník, Michael J. Bernstein, Konrad Bocian, Mark J. Brandt, et al. (2014), “Investigating Variation in Replicability: A ‘Many Labs’ Replication Project,” *Social Psychology*, 45 (3), 142–52.

Lindsay, D. Stephen (2015), "Replication in Psychological Science," *Psychological Science*, 26 (12), 1827–32.

Maimaran, Michal and Ayelet Fishbach (2014), "If It's Useful and You Know It, Do You Eat? Preschoolers Refrain from Instrumental Food," *Journal of Consumer Research*, 41 (3), 642–55.

McShane, Blakeley B. and Ulf Böckenholt (2014), "You Cannot Step into the Same River Twice: When Power Analyses Are Optimistic," *Perspectives on Psychological Science*, 9 (6), 612–625.

—— (2016), "Planning Sample Sizes When Effect Sizes Are Uncertain: The Power-Calibrated Effect Size Approach," *Psychological Methods*, 21 (1), 47–60.

McShane, Blakeley B., Ulf Böckenholt, and Karsten T. Hansen (2016), "Adjusting for Publication Bias in Meta-Analysis: An Evaluation of Selection Methods and Some Cautionary Notes," *Perspectives on Psychological Science*, 11 (5), 730–49.

McShane, Blakeley B. and David Gal (2016), "Blinding Us to the Obvious? The Effect of Statistical Training on the Evaluation of Evidence," *Management Science*, 62 (6), 1707–18.

Mosteller, Frederick and Robert R. Bush (1954), "Selected Quantitative Techniques" in *Handbook of Social Psychology, Vol. 1*, ed. Gardner Lindzey, New York: Addison-Wesley, 289–334.

Pigott, Terri D. (2012), *Advances in Meta-Analysis*, New York: Springer.

Rosenthal, Robert (1978), "Combining Results of Independent Studies," *Psychological Bulletin*, 85 (1), 185–193.

—— (1991), "Replication in Behavioral Research," in *Replication Research in the Social Sciences*, ed. James W. Neulip, Newbury Park, CA: Sage, 1–30.

Rosnow, Ralph and Robert Rosenthal (1989), "Statistical Procedures and the Justification of Knowledge in Psychological Science," *American Psychologist*, 44 (10), 1276–84.

Rothstein, Hannah, Alexander J. Sutton, and Michael Borenstein, eds. (2005), *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*, Chichester, UK: John Wiley & Sons.

Shah, Avni M., James R. Bettman, Peter A. Ubel, Punam A. Keller, and Julie A. Edell (2014), "Surcharges Plus Unhealthy Labels Reduce Demand for Unhealthy Menu Items," *Journal of Marketing Research*, 51 (6), 773–89.

Simmonds, Mark C., Julian P. T. Higgins, Lesley A. Stewart, Jayne F. Tierney, Mike J. Clarke, and Simon G. Thompson (2005), "Meta-Analysis of Individual Patient Data from Randomized Trials: A Review of Methods Used in Practice," *Clinical Trials*, 2 (3), 209–17.

Stewart, Lesley A. and Jayne F. Tierney (2002), "To IPD or Not to IPD? Advantages and Disadvantages of Systematic Reviews Using Individual Patient Data," *Evaluation & the Health Professions*, 25 (1), 76–97.

Stouffer, Samuel A., Edward A. Suchman, Leland C. DeVinney, Shirley A. Star, and Robin M. Williams Jr. (1949), *The American Soldier: Adjustment during Army Life (Studies in Social Psychology in World War II, Vol. 1)*, Princeton, NJ: Princeton University Press.

Stroebe, Wolfgang and Fritz Strack (2014), "The Alleged Crisis and the Illusion of Exact Replication," *Perspectives on Psychological Science*, 9 (1), 59–71.

Tsang, Eric W. K. and K.-M. Kwan (1999), "Replication and Theory Development in Organizational Science: A Critical Realist Perspective," *Academy of Management Review*, 24 (4), 759–80.